Title
A Study of High-dimensional Clustering and Statistical Inference on Networks

Author
Bhattacharyya, Sharmodeep

Publication Date
2013

Peer reviewed|Thesis/dissertation

A Study of High-dimensional Clustering and Statistical Inference of Networks

by

Sharmodeep Bhattacharyya

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

and the Designated Emphasis

in

Communication, Computation and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peter J. Bickel, Chair
Professor Bin Yu
Professor Jitendra Malik

Fall 2013

A Study of High-dimensional Clustering and Statistical Inference of Networks

Abstract

A Study of High-dimensional Clustering and Statistical Inference of Networks

by

Sharmodeep Bhattacharyya

Doctor of Philosophy in Statistics

and the Designated Emphasis in Communication, Computation and Statistics

University of California, Berkeley

Professor Peter J. Bickel, Chair

Clustering is an important unsupervised classification technique. In supervised classification, we are provided with a collection of labeled (pre-classified) patterns and the problem is to label a newly encountered, yet unlabeled, pattern. Typically, the given labeled (training) patterns are used to learn the descriptions of classes, which in turn are used to label a new pattern. In clustering, a set of unlabeled patterns are grouped into clusters in such a way that patterns in the same cluster are similar in some sense and patterns in different clusters are dissimilar in the same sense. In a sense, labels are associated with clusters also, but these category labels are data driven; that is, they are obtained solely from the data. The patterns which are to be classified in clustering can come from different sources, they can be vectors in a multi- dimensional space or nodes in discrete spaces. At first, we consider clustering in Euclidean space in large dimensions. Then, we delve into the discrete setting of networks. We at first go into issues related to network modeling and then into a specific method of clustering in networks.

In the first chapter, we consider the problem of estimation and deriving theoretical properties of the estimators for the elliptical distributions. The class of elliptical distributions have distributions with varied tail behavior. So, estimation under class of elliptic distributions lead to automatic robust estimators. The goal of the chapter is to propose efficient and adaptive regularized estimators for the nonparametric component, mean and covariance matrix of the elliptical distributions in both high and fixed dimensional situations. Semiparametric estimation of elliptical distribution has also been discussed in [27]. However, we wish to expand the model in two ways. First, study adaptive estimation methods with a novel scheme of estimating the nonparametric component and second, we perform regularized estimation of Euclidean parameters of the elliptical distribution such that high dimensional inference of the Euclidean parameters under certain additional structural assumption can be carried out. Some methods have already been developed. But we extend the work in [25] [24] [34] [57] [56]. The estimate of elliptical densities can also be used to approximately estimate certain sub-class of log-concave densities by using results from convex geometry.

The problem of estimation of mixture of elliptical distributions is also important in clustering, as the level sets produce disjoint elliptical components, which can be viewed as model of clusters of specific shape high dimensional space. A mixture of elliptical distributions is a natural generalization of mixture of Gaussian distributions, which have been extensively studied in the literature. So, an algorithm for regularized estimation of mixture of elliptical distributions will also lead to an algorithm for finding elliptical clusters in high dimensional space under highly relaxed tail conditions.

In clustering, one of the main challenges is the detection of number of clusters. Most clustering algorithms need the number of clusters to be specified beforehand. Previously, there has been some work related to choosing the number of clusters. We propose a new method of selecting number of clusters, based on hypothesis testing. One way to look at clustering is - getting hold of the most block-diagonal form of the similarity matrix. So, we test the hypothesis, whether the resulting similarity matrix after clustering is block-diagonal or not. The number of clusters for which we have the most block diagonal similarity matrix is considered to be the most suitable number of clusters for the data set. So, the method can be applied for any optimal partitioning algorithm (like k-means or spectral clustering or EM algorithm). We show that this method works well compared to currently used methods for both simulated and real data sets. We go into details on this work in Chapter 2.

The study of networks has received increased attention recently not only from the social sciences and statistics but also from physicists,computer scientists and mathematicians. [98]. But a proper statistical analysis of features of different stochastic models of networks is still underway. We give an account of different network models and then we analyze a specific nonparametric model for networks. We follow Bickel and Chen [22] by considering network modeling from a nonparametric point of view using a characterization of infinite graph models due to Aldous and Hoover [89], Bickel and Chen [22] studied this model primarily in the context of the community identification problem. Bickel, Chen and Levina [23] further considered inference using moment methods and generalized degrees in sparse graphs. We investigate the behavior of a histogram estimate of a canonical version of a function characterizing the model,

$$h((U,V)) = \mathbb{P}\left[A_{ij} = 1 | (\xi_i, \xi_j) = (U,V)\right]$$

where, ξ_i is the continuous latent variables corresponding to the i^{th} node of the network. We consider this estimate in dense social graphs in the context of network modeling and exploratory statistics. We apply the methods to an analysis of Facebook networks , given in [155]. We go into details on this work in Chapter 3.

We also propose bootstrap methods for finding empirical distribution of count features or 'moments' (Bickel, Chen & Levina, AoS, 2011) and smooth functions of these for the networks. Using these methods, we can not only estimate variance of count features but also get good estimates of such feature counts, which are usually expensive to compute numerically in large networks. In our paper, we prove theoretical properties of the bootstrap variance estimates of the count features as well as show their efficacy through simulation.

We also use the method on publicly available Facebook network data for estimate of variance and expectation of some count features. We go into details on this work in Chapter 4.

Lastly, we propose a clustering or community detection scheme for networks. One of the principal problem in networks is community detection. Many algorithms have been proposed for community finding [116] [140] but most of them do not have have theoretical guarantee for sparse networks and networks close to phase transition boundary proposed by physicists [50]. There are some exceptions but all have incomplete theoretical basis [44] [41] [100]. Here we propose an algorithm based on the graph distance of vertices in the network. We give theoretical guarantees that our method works in identifying communities for block models, degree-corrected block models [91] and block models with number of communities growing with number of vertices. We illustrate on a network of political blogs, Facebook networks and some other networks. We go into details on this work in Chapter 5.

The chapters 1, 2, 4 and 5 are written as self contained papers to be submitted, where as, chapter 2 is more of expository nature. Also, chapters 1, 4 and 5 are more of theoretical nature and chapters 2 and 3 are more of non-technical nature.

To my parents Pradip Kumar Bhattacharyya and Rita Bhattacharyya

For your constant support and encouragement throughout my life.

Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to thank my dissertation advisor Peter Bickel for his encouragement, guidance and inspiration throughout my graduate life. Apart from being an excellent advisor and providing invaluable inputs to my research, he also helped me grow as a person, enabling me to find a career path for myself. I learned a lot about life as a researcher and academician as much as I did about statistics from him. Even when I was in personal crisis, he came in to my support and helped me overcame my difficulties. Lessons on statistics as well as life that I have garnered from Peter will be my asset throughout my life.

I would also like to thank Bin Yu. She has been another source of constant encouragement for me. She also introduced me to challenges and opportunities in modern statistics through her Stat 215A course. She was always available whenever I need any help. Especially, her personal conversations with me greatly helped me in my research and career path.

I am grateful to all my teachers at Berkeley, especially Steve Evans, Elchanan Mossel, Jim Pitman, Alistair Sinclair, Philip Stark and Martin Wainwright, for their invaluable teachings which was extremely useful in my later research and for being generous to me with their advice and encouragement. Many of the other professors in the department have provided me with valuable advice during my PhD, among them are John Rice, Noureddine El Karoui and Adityanand Guntuboyina. Outside the department, Yoav Benjamini has also been another person who provided me with valuable research insight. He not only taught a great course on multiple comparisons, but also introduced me to several interesting problems in that subject area. I would like to thank Jitendra Malik for agreeing to be in my qualifying exams and dissertation committee. I would also like to thank Joshua Bloom and Joseph Richards from Center for Time-domain Informatics for introducing me to the rich data sets from astronomy. I would also like to thank professors under whom I did teaching assistantship, Hank Ibser and Jon Mcauliffe, for their valuable insight on teaching statistics. Besides that, I want to thank all other professors and staff in this small wonderful Statistics department, whose door was always open for me. I am grateful to Anindita Adhikari for her help and advice during the first year. Many thanks to Angie Fong and La Shana Polaris for answering all my important and not-so-important non-academic questions.

It is imperative that I thank all my fellow graduate students and postdocs who were part of my life in the last four years and without whose presence the path would not have been so smooth. I am thankful to Riddhipratim Basu, Derek Bean, Yuval Benjamini, Mu Cai, Tessa Childers, Dave Choi, Partha Dey, Subhroshekhar Ghosh, Brianna Hirst, Wayne Lee, Jing Lei, James Long, Joe Neeman, Nima Reyhani, Karl Rohe, Sean Ruddy, Arnab Sen, Sujayam Saha, Ying Xu and all other friends and colleagues for making the graduate life experience so enriching. Special thanks to my two roommates, Koushik Pal and Sayak Ray for always being with me through the highs and lows, for cooking so many nice dishes and for numerous nice discussions about life and research.

I am grateful to all of my teachers at Indian Statistical Institute, Kolkata (especially Arijit Chakraborty, Probal Chaudhuri, Alok Goswami, Gautam Mukherjee, Bimal Kumar

Roy, Debasis Sengupta, Tapas Samanta and all others) for teaching me what statistics and probability is really about.

Last but not the least, I thank the two persons without whom I would never have been the person I am today. They are my parents: Pradip Kumar Bhattacharyya and Rita Bhattacharyya, who gave me the freedom to explore new avenues, who encouraged me to think independently, who always supported me throughout my life and specially whose endless love provides me the strong support to go forward.

Chapter 1

Regularized estimation of elliptic distributions and high dimensional clustering

1.1 Introduction

We consider the estimation of semi parametric family of elliptic distributions for the purpose of data description and classification (regression and clustering). The class of elliptically contoured or elliptical distributions provide a very natural generalization of the class of Gaussian distribution. An elliptical density has elliptic contours like a Gaussian distribution, but can have either heavier or lighter tails than the Gaussian density. The class of elliptical distributions is also very attractive for statistical inference as it has the same location-scale Euclidean parameters as in Gaussian distribution with an additional univariate function parameter. There has been extensive work done on estimation of Euclidean parameters for elliptical distributions. Adaptive and efficient estimation of the Euclidean parameters were addressed by Bickel et.al. (1993) [28], Bickel (1982) [19] and Anderson et al (1986) [6]

It may be argued that semi parametric family is too restrictive and one instead should focus on the more general family of shape-constrained densities and their mixtures. This are has been heavily studied theoretically in different contexts since the seminal work of Grenander [70] on nonparametric maximum likelihood estimation of monotone univariate density. In particular the natural generalization of Gaussian and elliptical families the log-concave densities and their generalizations have received much attention (algorithmic [158] [142], theoretical [53] [46] and extensions [97] [8]). However, for all these problems, estimation of densities in large dimensions become a computationally challenging problem. The algorithm proposed by Cule et.al. (2010)[46] Koenker and Mizera [97] works for estimation of multivariate log-concave densities but is too slow to work in large dimensions. So, the application of such models to clustering and classification are very limited. That is why, we consider a smaller class of multivariate density functions, which can be estimated with

relative ease for large dimensions.

Semiparametric estimation of elliptic densities for fixed dimension were first addressed in Stute and Werner (1991) [148] by using kernel density estimators for estimating the function parameter. Cui and He (1995) [45] addressed a similar problem. Liebscher (2005) [112] used transformed data and kernel density estimators for estimating the unknown function parameter. Battey and Linton (2012) [12] used finite mixture sieves estimate for the function parameter by using the scale mixture of normal representation of consistent elliptic density. ML-Estimation of only the Euclidean parameters of the elliptical distributions were considered in the work of Tyler [156] and Kent and Tyler [94]. These works were extended to get shrinkage estimates of covariance matrix of the elliptical distributions by Chen, Wiesel and Hero (2011) [42] and Wiesel (2012) [163] with the shrinking being towards identity, diagonal or given positive-semi-definite matrix. In all of these works the theoretical properties of the estimators were also addressed. We focus on maximum likelihood estimation of the elliptic densities using penalized likelihood functions. The estimation consists of nonparametric estimation of a univariate function as well as parametric estimation of the location and scale parameters.

Recently, there has been a general focus on statistical inference in high-dimensional problems, with examples of high-dimensional data coming from biology especially genetics and neuroscience, imaging, finance, atmospheric science, astronomy and so on. In most of these cases, the number of dimensions of data is nearly of the same order or greater than the number of data points. So, the appropriate asymptotic framework is as both $n \to \infty$ and $p \to \infty$, where n is the number of data points and p is the dimension of data points.

Little is possible without some restriction on parameters. In case of regression restrictions are put on regression parameters such as sparsity and size and on the design matrix such as incoherence and related conditions. A variety of regularization methods have been studied such as [117] [119] [166] [26] [36] and effective algorithms proposed such as [55] [118]. A good reference book on all different forms of regularization and algorithms is Bühlmann and Van de Geer (2011) [32]. Another problem that has been considered is covariance or precision matrix estimation where again you need regularization if you has to have consistency as $p, n \to \infty$. Again there has been considerable theoretical work [25] [24] [34] [67] [141] [101] focussing on Gaussian and sub-Gaussian distributions, except some like [105], which are distribution-free. However, there has been little focus on tail behavior, except some on sample covariance matrix behavior [147] [49], although there has been earlier work in the robustness literature (Chapter 5 of [73], [84]). Here we consider elliptical distributions, which can have arbitrary tail behavior. For such distributions, we have attempted to estimate sparse mean and covariance matrices using penalized likelihood loss function. Thus, we have generalized the class of regularized covariance estimators, so that estimation of sparse covariance and precision matrix becomes possible under arbitrary tail behavior of the underlying distribution in high-dimensions. We have tried to provide a framework of semiparametric inference of elliptical distributions for Euclidean parameters as well as mixtures.

Contributions and Outline of the Chapter

So, in this chapter we have done the following.

1. We develop estimation procedure for the density generator function of the elliptical distribution in a log-linear spline form in Section 1.3 and derive respective error bounds.

2. We use the estimate of the density generator function of elliptical distribution to adaptively estimate Euclidean parameters of elliptical distribution in Section 1.4. We show how using appropriate regularization we can obtain, under conditions similar to those of [26], [141] and [101], consistent estimates of Euclidean parameters for both fixed dimensional case and when $p, n \to \infty$ in Section 4.4.

3. Develop feasible algorithms for all these methods in Section 1.5 and illustrate our method by simulation and one real data example in Section 4.5 and Section 1.8.

4. Extend the results to three special cases - (a) Estimation of Covariance and Precision matrix (b) Regression with Elliptical errors and (c) Clustering via mixtures of elliptical distribution in Section 1.6.

We give the main definitions and results in Section 1.2.

1.2 Elliptical Distributions and Main Results

The formal definition of *elliptically symmetric* or *elliptical* distributions is given in the following way in [60] -

Definition 1.2.1. *Let X be a p-dimensional random vector. X is said to be 'elliptically distributed' (or simply 'elliptical') if and only if there exist a vector $\mu \in \mathbb{R}^p$, a positive semidefinite matrix $\Omega \equiv \Sigma^{-1} \in \mathbb{R}^{p\times p}$, and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $t \mapsto \phi_{X-\mu}(t)$ of $X - \mu$ corresponds to $t \mapsto \phi(t^T\Sigma t)$, $t \in \mathbb{R}^p$.*

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu, \Omega)$. Then the density function $f(\cdot; \mu, \Omega)$ is of the form

$$f(x; \mu, \Omega) = |\Omega|^{1/2} g_p\left((x-\mu)^T\Omega(x-\mu)\right) \tag{1.1}$$

where $\theta = (\mu, \Omega) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and covariance parameters respectively with $\mu \in \mathbb{R}^p$ and $\Omega \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x)dx = 1$$

$\mathbb{R}^+ = [0, \infty)$. g_p is also called the *density generator* of the elliptical distribution in $\mathbb{R}^p$.

Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\mu_0||_0 = s_1$ and $||\Omega^-||_0 = s_2$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s_1 and s_2 indicates *sparsity*. We first consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n.

Now, if we define $Y = (X-\mu)^T\Sigma^{-1}(X-\mu)$, then, by transformation of variables, Y has the density

$$f_Y(y) = c_p y^{p/2-1} g_p(y) \quad y \in \mathbb{R}^+ \tag{1.2}$$

where, $c_p = \frac{\pi^{p/2}}{\Gamma(p/2)}$. So, we can now use estimate of f_Y from the data $Y_1, \ldots, Y_n$ to get an estimate of the non-parametric component g_p. From here onwards we shall drop the suffix and denote g_p by g.

We divide the family of density generators into two different classes - *monotone* and *non-monotone*. The following proposition gives the equivalence between *monotone* density generator and *unimodal* elliptical density.

Proposition 1.2.2. *The density generator g_p is monotonically non-increasing if and only if the elliptical density f with density generator g_p is unimodal. Also the mode of density f is at μ.*

Proof. If: If g_p is not monotonically increasing, then, there exists a mode of g_p which is not at zero. Let that mode be at ν. By symmetry of f now f has a mode at all points at in an ellipse around μ whose points are $(\nu-\mu)^T\Sigma^{-1}(\nu-\mu)$. So, f does not remain unimodal anymore. So, if f is unimodal, then, g_p has to be monotonically non-increasing.
Only If: This part is obvious. □

So, we divide the class of elliptical densities into two - *unimodal* and *multimodal*. Unimodal elliptical densities have monotone density generator, where as, multimodal elliptical density has non-monotone density generator. Examples of unimodal elliptical density include normal, t, logistic distributions, and examples of multimodal elliptical density include a subclass of Kotz type and multivariate Bessel type densities. See Table 3.1 (pp. 69) of [60] for more examples.

Another desirable property of the class of elliptical distributions is *consistency*.

Definition 1.2.3. *An elliptical distribution with density $f_p(\cdot; \mathbf{0}_p, \mathbf{I}_p)$ and density generator g_p is said to possess **consistency** property if and only if*

$$\int_{-\infty}^{\infty} g_{p+1}\left(\sum_{i=1}^{p+1} x_i^2\right) dx_{p+1} = g_p\left(\sum_{i=1}^{p} x_i^2\right) \tag{1.3}$$

This consistency property of elliptical distributions is quite desirable and natural, since it ensures that marginal distribution of the elliptical distributions also follow the elliptical distribution with same density generator. This property becomes indispensable if we go for

high-dimensional situation, since, in high-dimensions, we have to depend on the projection of the random variable in low-dimensions and if in the low-dimensions, we have a different density generator, we can not devise any adaptive estimator. Equivalent conditions for the consistency property is given in Theorem 1 of [90]. We just mention an excerpt of that Theorem in form of Lemma below

Lemma 1.2.4 ([90]). *Let g_p be the density generator for a p-variate random variable, X_p following elliptical distribution with mean and covariance matrix parameters $(\mathbf{0}_p, \mathbf{I}_p)$. Then, X_p follows consistent elliptical distribution if and only if $X_p \stackrel{d}{=} Z_p/\sqrt{\xi}$, where, Z_p is a p-variate normal random variable with parameters $(\mathbf{0}_p, \mathbf{I}_p)$ and $\xi > 0$ is some random variable unrelated with p and independent of Z_p.*

Examples of elliptical distributions with consistency property include multivariate Gaussian distributions, multivariate t-distributions, multivariate stable laws and such, where as, examples of elliptical distribution without consistency property include multivariate logistic distributions. For more discussion and insight on the issue see [90].

We shall first try to estimate density generator g_p with monotonicity constraint in Section 1.3. Unimodal elliptical density is more commonly seen in practice and is easier to handle. We shall only estimate Euclidean parameters for consistent elliptical distributions in high dimensions.

Some Notations

For any vector $x \in \mathbb{R}^p$, we define,

$$\begin{aligned} ||x||_2 &= \sqrt{\sum_{i=1}^{n} x_i^2}, \\ ||x||_1 &= \sum_{i=1}^{n} |x_i|, \\ ||x||_\infty &= \max\{x_1, \ldots, x_n\}, \\ ||x||_0 &= \sum i = 1^n \mathbf{1}(x_i \neq 0) \end{aligned}$$

For any matrix $M = [m_{ij}]$, we write $|M|$ for the determinant of M, $tr(M)$ for the trace of M, and $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ for the largest and smallest eigenvalues of M, respectively. We write $M^+ = \text{diag}(M)$ for a diagonal matrix with the same diagonal as M and $M^- \equiv M - M^+$. We will use $||M||_F$ to denote the Frobenius matrix norm and $||M|| \equiv \lambda_{\max}(MM^T)$ to denote the operator or spectral norm (also known as matrix 2-norm). We will also write $|\cdot|_1$ for the l_1 norm of a vector or matrix vectorized $|M|_1 = \sum_{i,j} |m_{ij}|$. $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form.

For two numerical sequences a_n and b_n, $a_n = o(b_n)$ means $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$ and $a_n = O(b_n)$ means there exists constant C such that $a_n \leq C b_n$ for $n \geq N$. Also, for two random or

numerical sequences X_n and Y_n, $X_n = o_P(Y_n)$ means that $\frac{X_n}{Y_n} \xrightarrow{P} 0$ and $X_n = O_P(Y_n)$ means that X_n is stochastically bounded by Y_n, that is, given $\varepsilon > 0$ there exists constant C and integer $N \geq 1$, such that $\mathbb{P}\left[\left|\frac{X_n}{Y_n}\right| \leq C\right] \leq \varepsilon$ for $n \geq N$.

Main Results

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. Then the density function $f(\cdot; \mu_0, \Omega_0)$ is of the form

$$f(x; \mu_0, \Omega_0) = |\Omega_0|^{1/2} g_p\left((x-\mu_0)^T \Omega (x-\mu_0)\right) \tag{1.4}$$

where $\theta_0 = (\mu_0, \Omega_0) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and inverse covariance or precision matrix parameters (Σ_0 is the covariance parameter) respectively with $\mu_0 \in \mathbb{R}^p$ and $\Omega_0 \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x) dx = 1$$

$\mathbb{R}^+ = [0, \infty)$.

We start with the following assumption -

(A1) Assume that we have an initial consistent estimators $\hat{\mu}$ and $\hat{\Omega}$, such that, $||\hat{\mu} - \mu_0||_2$ and $||\hat{\Omega} - \Omega_0||_F$ concentrates to zero with tail bounds given by functions $J_1(t, n, p)$ and $J_2(t, n, p)$ such that,

$$\mathbb{P}[||\mu_0 - \hat{\mu}||_2 > t] \leq J_1(t, n, p) \tag{1.5}$$
$$\mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] \leq J_2(t, n, p). \tag{1.6}$$

For fixed dimensions, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(n))$, where, $\omega(n) \to 0$ as $n \to \infty$ with given tail bounds. For high dimensions, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(p, n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(p, n))$, where, $\omega(p, n) \to 0$ as $p, n \to \infty$ with given tail bounds.

Now, let us consider the high-dimensional situation. So, in this case, the dimension of Euclidean parameters grows with n. In the high-dimensional situation we assume the additional structure of sparsity imposed on the Euclidean parameters $\theta_0 = (\mu_0, \Omega_0)$. Density generator g comes from a consistent family elliptical distributions and so by Lemma 1.2.4, $\Omega_0^{1/2}(X - \mu_0) \stackrel{d}{=} Z_p/\sqrt{\xi}$. Let us consider the probability density function of X_j to be h $(j = 1, \ldots, p)$ and

$$H(u) \equiv \int_u^\infty h(v) dv \tag{1.7}$$

We consider the following assumptions -

(A2) Suppose that $||\Omega^-||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*.

(A3) $\lambda_{\min}(\Sigma_0) \geq \underline{k} > 0$, or equivalently $\lambda_{\max}(\Omega_0) \leq 1/\underline{k}$. $\lambda_{\max}(\Sigma_0) \leq \overline{k}$.

(A4) The function $H(t)$ defined in Eq. (1.7) and $J_1(t), J_2(t)$ defined in Eq. (1.5), satisfies the following conditions -

(a) there exists a function $\sigma_1(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $c, d > 0$, for $t = O(\sigma_1(p,n))$,

$$p(c\exp(-dt)J_1(t,n,p)J_2(t,n,p))^n \to 0 \text{ as } p,n \to \infty \tag{1.8}$$

(b) there exists a function $\sigma_2(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $d_1, d_2, d_3 > 0$, for $t = O(\sigma_2(p,n))$,

$$d_1 p^2 (H(t)J_1(t,n,p)J_2(t,n,p))^n (\exp(-nd_2 t))(d_3 \exp(-nd_3 t^2)) \to 0 \text{ as } p,n \to \infty \tag{1.9}$$

Let us consider that we have obtained estimators of the Euclidean parameters $\tilde{\mu}$ and $\tilde{\Omega}$ and the nonparametric component $\hat{g}$ by following the estimation procedure of the elliptical density in Section 1.5. Let us consider first the fixed dimensional situation. So, in this case, the dimension of Euclidean parameters does not grow with n

Theorem 1.2.5. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (A1-A4),*

(a) $||\mu_0 - \tilde{\mu}||_2 = O_P(\sigma_1(n))$.

(b) $||\Omega_0 - \tilde{\Omega}||_F = O_p(\sigma_2(n))$

(c) $\hat{g}$ *is an uniform consistent estimator of* g *with rate* $O_p\left(\left(\frac{\log n}{n}\right)^{1/3}\right)$.

We consider the high-dimensional situation now, that means the number of dimensions p and the number of samples n both grow. We have the estimates, $\tilde{\mu}$ and $\tilde{\Omega}$ as stated in Section 1.5.

Theorem 1.2.6. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (A1)-(A4),*

(a) $||\mu_0 - \tilde{\mu}||_2 = O_P\left(\sqrt{p}\sigma_1(p,n)\right)$

(b) $||\Omega_0 - \tilde{\Omega}||_F = O_P\left(\sqrt{(p+s)}\sigma_2(p,n)\right)$ *for* $\nu = O\left(\sigma_2(p,n)\right)$.

(c) $\hat{g}$ *is an uniform consistent estimator of* g *with rate* $O_p\left(\left(\frac{\log n}{n}\right)^{1/3}\right)$.

We apply the semi parametric inference technique of estimating elliptical distributions to two parametric inference and one semi-parametric inference problems -

(a) In Section 1.6, we apply it for robust regularized covariance and precision matrix estimation in high-dimensions.

(b) In Section 1.6, we apply it for robust regularized regression in high-dimensions.

(c) In Section 1.6, we apply the semi parametric inference technique of estimating elliptical distributions to clustering by devising an inference scheme for mixtures of elliptical distributions.

1.3 Inference I: Estimation of Density Generator g_p

We try to find a maximum likelihood estimate for the semiparametric elliptical distribution. The main idea is using non-parametric maximum likelihood estimate (NPMLE) to estimate density generator g_p and then use that NPMLE estimate of g_p, to get a likelihood estimate of the Euclidean parameters. Throughout this section, we shall consider that the Euclidean parameters $\theta = (\mu, \Omega)$ are given and the dimension of data p is fixed.

We shall propose non-parametric maximum likelihood estimates (NPMLE) of density generator g_p under the monotonicity assumption. This is the most common situation for elliptical distributions as monotone density generators gives rise to unimodal elliptical distributions according to Proposition 1.2.2. We shall principally focus on this case. We consider this case in Section 1.3.

Maximum Likelihood Estimation of Monotone Density Generator

The likelihood for (θ, g) is

$$\mathcal{L}(\theta, g|X_1, \ldots, X_n) = \prod_{i=1}^{n} |\Sigma|^{-1/2} g\left((X_i - \mu)^T \Sigma^{-1} (X_i - \mu)\right)$$

The log-likelihood is

$$\ell(\theta, g|X_1, \ldots, X_n) = -\frac{n}{2}\log|\Sigma| + \sum_{i=1}^{n} \log g\left((X_i - \mu)^T \Sigma^{-1} (X_i - \mu)\right)$$

Let us start with an ideal case when the Euclidean parameters μ and Σ are known. Then, the non-parametric likelihood of g in terms of data Y_i, $i = 1, \ldots, n$, obtained by the transformation $Y_i = (X_i - \mu)^T \Sigma^{-1} (X_i - \mu)$ becomes

$$\mathcal{L}(g|\boldsymbol{Y}) \;=\; (c_p)^n \prod_{i=1}^{n} Y_i^{\frac{p}{2}-1} g\left(Y_i\right)$$

The log-likelihood ignoring the constants becomes

$$\ell(g|\boldsymbol{Y}) \;=\; \sum_{i=1}^{n} \left(\left(\frac{p}{2} - 1 \right) \log(Y_i) + \log g\left(Y_i\right) \right)$$

So, the NPMLE $\hat{g}_n$ can be written as

$$\hat{g}_n = \arg\max_g \sum_{i=1}^{n} \left(\left(\frac{p}{2} - 1 \right) \log(Y_i) + \log g\left(Y_i\right) \right) = \arg\max_g \sum_{i=1}^{n} \left(\log g\left(Y_i\right) \right)$$

Now, $g(y)$ is a monotonically non-increasing function, however, $f_Y(y)$ as defined in (1.2), which is the density of Y_i's, is not monotone. But, we can still formulate the problem as a generalized isotonic regression problem as done in Example 1.5.7 (pp. 38-39) in Robertson et.al. (1988) [138].

First note that the NPMLE $\hat{g}_n$ must be constant on intervals $(Y_{(i-1)}, Y_{(i)}]$, $i = 1, \ldots, n$ $(Y_0 = 0)$, where $Y_{(i)}$ is the i^{th} order statistic of Y_i, and $\hat{g}_n$ must be zero on $(Y_{(n)}, \infty)$. It follows by observing that if $\hat{g}_n$ is not constant on $(Y_{(i-1)}, Y_{(i)}]$, then, we can always construct another estimator $\tilde{g}_n = (Y_{(i)} - Y_{(i-1)})^{-1} \int_{Y_{(i-1)}}^{Y_{(i)}} \hat{g}_n(t)dt$ constant on $(Y_{(i-1)}, Y_{(i)}]$ which gives larger likelihood than $\hat{g}_n$. So, the NPMLE $\hat{g}_n$ has to be piecewise constant and left-continuous.

Hence the problem of finding NPMLE boils down to the optimization problem on $(g_1, \ldots, g_n)$, where $g_i = g(Y_i)$ for $i = 1, \ldots, n$. The optimization problem is defined as

$$\max_{(g_1, \ldots, g_n)} \sum_{i=1}^{n} \log g_i \tag{1.10}$$

such that

$$\sum_{i=1}^{n} c_p \int_{Y_{(i-1)}}^{Y_{(i)}} y^{p/2-1} g_i dy = \frac{2c_p}{p} \sum_{i=1}^{n} g_i \left(Y_{(i)}^{p/2} - Y_{(i-1)}^{p/2} \right) = 1 \tag{1.11}$$

and

$$g_1 \geq g_2 \geq \cdots \geq g_n \tag{1.12}$$

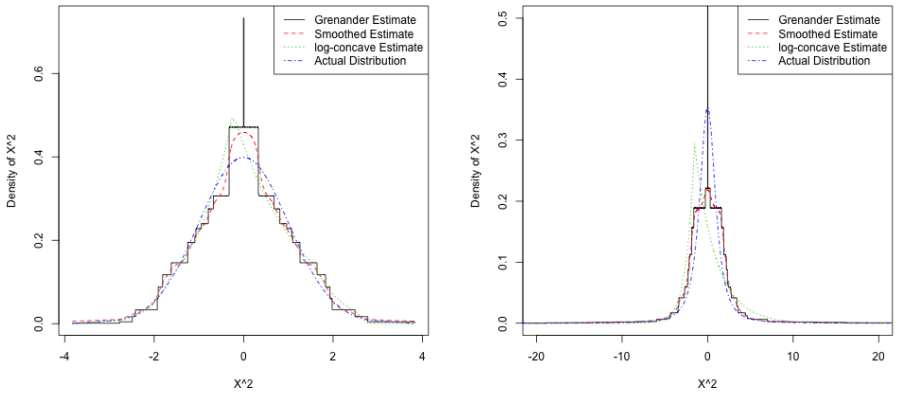
The above defined optimization problem can be solved in the similar way as described in Example 1.5.7 of [138] (pp. 38-39) and by following Theorem 1.4.4 of [138] the solution can be written as

$$\hat{g}_i = \frac{p}{2c_p} \min_{s \leq i-1} \max_{t \geq i} \frac{F_n(Y_{(t)}) - F_n(Y_{(s)})}{Y_{(t)}^{p/2} - Y_{(s)}^{p/2}} \tag{1.13}$$

where, F_n is the empirical cumulative distribution function (CDF) of the data $(Y_1, \ldots, Y_n)$. The NPMLE $\hat{g}_n$ is given by

$$\hat{g}_n(y) \quad = \quad \begin{cases} \hat{g}_i, & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{1.14}$$

Note that, the NPMLE $\hat{g}_n$ is quite related to the Grenander estimator [70] of monotonically non-increasing densities. The Grenander estimator is a piece-wise constant or histogram type density estimate, where the constant values come from the left-derivative of the least concave majorant of the empirical CDF function. Similarly, NPMLE $\hat{g}_n$ is also a piece-wise constant or histogram type density generator estimate, where the constant values come from the left derivative of the least concave majorant of the empirical CDF plotted against the abscissa of $\frac{2c_p}{p} y^{p/2}$ instead of y. Figure 1.1 gives an example of the NPMLE $\hat{g}_n$ for a simulated small sample case.



(a) (b)

Figure 1.1: (a) Estimated univariate normal density curve using several NPML techniques (b) Estimated univariate t with 2 degrees of freedom density curve using several NPML techniques.

The above description of the NPMLE $\hat{g}_n$ also provides us with a cautionary note while implementing this estimator. The transformation $y \mapsto \frac{2c_p}{p} y^{p/2}$ highly stretches the abscissa for large values of p, so for numerical implementation of the algorithm care should be taken so that machine precision problems does not hurt the computation of the estimator. However, in this discourse we shall not dwell on these numerical issues. Some thoughts on this issue is given in Section 4.5.

The asymptotic properties NPMLE $\hat{g}_n$ is provided in Lemma 1.3.1. The asymptotic properties are quite as expected of isotonic regression estimates. The proof borrows techniques

from Groeneboom (1985) [71], Jonker and Van der Vaart (2002) [88] and Example 3.2.14 of citeMR1385671.

Lemma 1.3.1. *Let g be the monotonically decreasing density generator of the elliptical distribution and the NPMLE of g is $\hat{g}_n$, whose definition is given in* (1.14)*. Suppose that g is continuously differentiable on the interval $(0,\infty)$ with derivative g' bounded away from 0 (from above) and $-\infty$. Then,*

(a) For any $y > 0$, as $n \to \infty$,

$$n^{1/3}\left(\hat{g}_n(y) - g(y)\right) \xrightarrow{w} \left|4g(y)g'(y)\right|^{1/3} arg\max_h \left\{W(h) - \sqrt{c_p y^{p/2-1}}h^2\right\}. \tag{1.15}$$

where, W is the Wiener process on $(0,1)$.

(b) For any $x_n \to \infty$, $\delta_n = O\left(n^{-1/3}(\log n)^{1/3}\right)$, $U > x_n\delta_n$ and $n \to \infty$

$$\mathbb{P}\left[\sup_{x_n\delta_n \le y \le U}\left(\frac{n}{\log n}\right)^{1/3}\left|\hat{g}_n(y) - g(y)\right| \ge x\right] \le O\left(\frac{1}{\sqrt{x}}\right). \tag{1.16}$$

so, we have,

$$\sup_{x_n\delta_n \le y \le U}\left|\hat{g}_n(y) - g(y)\right| = O_P\left(\left(\frac{\log n}{n}\right)^{1/3}\right). \tag{1.17}$$

(c) For any $U > 0$, if $||\hat{g}_n - g||_1 = \int_0^U |\hat{g}_n(y) - g(y)|dy$, then, as $n \to \infty$

$$n^{1/3}\mathbb{E}||\hat{g}_n - g||_1 \to \int_0^U |4\mathbb{E}|V_y|g'(y)g(y)|dy \tag{1.18}$$

where, $V_y = arg\max_h \left\{W(h) - \sqrt{c_p y^{p/2-1}}h^2\right\}$.

Proof. (a) Let us define a stochastic process $\{\hat{s}_n(a) : a > 0\}$ by

$$\hat{s}_n(a) = \arg\max_s \left\{F_n(s) - \frac{2ac_p}{p}s^{p/2}\right\}$$

where the largest value is chosen when multiple maximizers exist. It is easy to see that $\hat{g}_n(t) \le a$ if and only if $\hat{s}_n(a) \le t$. It follows that,

$$\mathbb{P}\left(n^{1/3}\left|\frac{g(t)g'(t)}{2}\right|^{1/3}\left(\hat{g}_n(t) - g(t)\right) \le x\right) = \mathbb{P}\left(\hat{s}_n\left(g(t) + \delta_n\right) \le t\right)$$

where, $\delta_n = xn^{-1/3}\left|\frac{g(t)g'(t)}{2}\right|^{1/3}$. By definition,

$$\hat{s}_n(a+\delta_n) = \sup\left\{s \geq 0 : F_n(s) - \frac{2c_p}{p}(a+\delta_n)s \text{ is maximal}\right\}$$

Hence, we can write,

$$\hat{s}_n(a+\delta_n) = \sup\left\{s \geq 0 : \sqrt{n}\left(F_n(s) - F(s)\right) + \sqrt{n}\left(F(s) - \frac{2c_p}{p}(a+\delta_n)s\right) \text{ is maximal}\right\}$$

By Hungarian embedding theorem [99],

$$\sqrt{n}\left(F_n(t) - F(t)\right) = B_n(F(t)) + O_P\left(n^{-1/2}\log n\right)$$

where, $(B_n, n \in \mathbb{N})$ is a sequence of Brownian bridges, constructed on the same space as F_n and where, F is the CDF of $Y = (X-\mu)^T\Sigma^{-1}(X-\mu)$. So by (1.2),

$$F'(t) = f(t) = c_p t^{p/2-1} g(t)$$
$$f'(t) = c_p t^{p/2-1} g'(t) + \frac{c_p(p-2)}{2} t^{p/2-2} g(t)$$

So, the limiting distribution of $n^{1/3}(\hat{s}_n(a+\delta_n) - t)$ will be the same as limiting distribution of $n^{1/3}(s_n(a+\delta_n) - t)$, where, $s_n(b)$ is the location of the maximum of the process $\left\{B(F(s)) + \sqrt{n}(F(s) - \frac{2c_p}{p}bs^{p/2}), s \geq 0\right\}$ and B is a standard Brownian bridge on $[0,1]$.

Now, location of the maximum of the process

$$\left\{B(F(s)) + \sqrt{n}(F(s) - \frac{2c_p}{p}(a+\delta_n)s^{p/2}), s \geq 0\right\}$$

behaves as $n \to \infty$ as the location of maximum of the process

$$\left\{B(F(t)+f(t)(s-t)) + \sqrt{n}(F(t)+f(t)(s-t)+\frac{f'(t)}{2}(s-t)^2 - \frac{2c_p}{p}(a+\delta_n)s^{p/2}), s \geq 0\right\}$$

Consider, $a = g(t)$, $c = -\frac{g'(t)}{2}$ and $h = \left(\frac{nc^2}{a}\right)^{1/3}(s-t)$, location of the maximum of above mentioned process behave as location of the maximum of following process as $n \to \infty$

$$\{B(F(t) + f(t)(s-t)) + \sqrt{n}(F(t) + f(t)(nc^2d/a)^{-1/3}h + \frac{f'(t)}{2}(nc^2/a)^{-2/3}h^2$$
$$-\frac{2c_p}{p}(a+\delta_n)\left(t + (nc^2/a)^{-1/3}h\right)^{p/2}), h \in \mathbb{R}\}$$

and as $n \to \infty$ it is equivalent to the location of the maximum of the process

$$\{B(F(t)+f(t)(s-t))+\sqrt{n}(F(t)+f(t)(nc^2/a)^{-1/3}h+\frac{f'(t)}{2}(nc^2/a)^{-2/3}h^2 \\ -\frac{2c_p}{p}(a+\delta_n)\left((p/2)t^{p/2-1}(nc^2/a)^{-1/3}h+\binom{p/2}{2}t^{p/2-2}(nc^2/a)^{-2/3}h^2\right), h\in\mathbb{R}\}$$

and as $n \to \infty$ it is equivalent to the location of the maximum of the process

$$\{B(F(t)+f(t)(s-t))-\sqrt{n}((c_p/2)t^{p/2-1}g'(t)(nc^2/a)^{-2/3}h^2 \\ +(c_p\delta_n t^{p/2-1}(nc^2/a)^{-1/3}h),\ h\in\mathbb{R}\}$$

Since, a Brownian bridge behaves locally as a Brownian motion in $(0,1)$, the limiting distribution of

$$\{W(c_p t^{p/2-1}g(t)(nc^2/a)^{-1/3}h)-\sqrt{n}(c_p t^{p/2-1}(g'(t)/2)(nc^2/a)^{-2/3}h^2 \\ +(c_p\delta_n t^{p/2-1}(nc^2/a)^{-1/3}h),\ h\in\mathbb{R}\}$$

where, W is the Wiener process on $(0,1)$. Now, by writing the values of a, c and $\delta_n = xn^{-1/3}\left(d(t)ac\right)^{1/3}$ and using Brownian scaling, we get that

$$\{\sqrt{c_p}t^{(p/2-1)/2}a^{2/3}(nc^2)^{-1/6}W(h)-c_p t^{(p/2-1)}a^{2/3}(nc^2)^{-1/6}h^2 \\ -c_p t^{(p/2-1)}a^{2/3}(nc^2)^{-1/6}xh,\ h\in\mathbb{R}\}$$

The location of maximum of the above process is equivalent to the location of maximum of the process

$$\left\{W(h)-\sqrt{c_p t^{p/2-1}}\left(h^2+xh\right),\ h\in\mathbb{R}\right\}$$

Let

$$V(a)\equiv\arg\max_h\left\{W(h)-\sqrt{c_p t^{p/2-1}}(h-a)^2, h\in\mathbb{R}\right\}$$

The, $\{V(a)-a : a\in\mathbb{R}\}$ is a stationary process and $\mathbb{P}(V(a)\le t)=\mathbb{P}(V(0)\le t-a)$. So, in summary, as $n\to\infty$

$$\begin{aligned}\mathbb{P}\left(n^{1/3}\left|\frac{g(t)g'(t)}{2}\right|^{-1/3}(\hat{g}_n(t)-g(t))\le x\right) &= \mathbb{P}\left(\hat{s}_n(a+\delta_n)-t\le 0\right)\\ &\to \mathbb{P}\left(V(-x/2)\le 0\right)=\mathbb{P}\left(2V(0)\le x\right)\end{aligned}$$

and we prove for each $y>0$ as $n\to\infty$,

$$n^{1/3}\left(\hat{g}_n(y)-g(y)\right)\xrightarrow{w}\left|4g(y)g'(y)\right|^{1/3}\arg\max_h\left\{W(h)-\sqrt{c_p t^{p/2-1}}h^2\right\}.$$

(b) Let us use the stochastic process $\hat{s}_n(a)$ again for this proof. Now,

$$\begin{aligned}\hat{s}_n(a) &= \arg\max_{s:s\geq 0}\left\{F_n(s) - \frac{2ac_p}{p}s^{p/2}\right\} \\ &= \arg\max_{h:h\geq -\delta_n^{-1}t}\left\{F_n(t+\delta_n h) - \frac{2ac_p}{p}(t+\delta_n h)^{p/2}\right\} \\ &= \arg\max_{h:h\geq -\delta_n^{-1}t}\left\{G_n(t+\delta_n h) + \sqrt{n}\left(F(t+\delta_n h) - F(t) - \frac{2ac_p}{p}(t+\delta_n h)^{p/2}\right)\right\}\end{aligned}$$

where, $G_n(s) = \sqrt{n}\,(F_n(s) - F(s))$ and F_n and F are as defined in part (a). Consider that $s \in (0, L)$. So, $h \in (\delta_n^{-1}t, \delta_n^{-1}(L-t))$. Let us take $a = g(t) + x\delta_n$ with $x > 0$ fixed. Now, by Taylor expansion,

$$\begin{aligned}&F(t+\delta_n h) - F(t) - \frac{2ac_p}{p}(t+\delta_n h)^{p/2} \\ &= f(t)\delta_n h + \frac{f'(t+\xi\delta_n h)}{2}\delta_n^2 h^2 - \frac{2c_p}{p}(g(t)+x\delta_n)(t+\delta_n h)^{p/2} \\ &= c_p t^{p/2-1}\delta_n^2\left(\frac{g'(t)}{2}h^2 - xh\right) + r_n(h) \\ &\begin{cases} \leq -c\delta_n^2 h^2 - \gamma_n x\delta_n^2 h \\ \geq -d\gamma_n\delta_n^2 h^2 - Cx\delta_n^2 h\end{cases}\end{aligned}$$

for certain $c, d > 0$ (since, $g(t)$ is monotonically decreasing) independent of δ_n, t, h. $\gamma_n > 0$ is a lower bound for t and $\gamma_n \to 0$ and $n \to \infty$. Now, if $(\hat{g}_n(t) - g(t)) > x\delta_n$, then, for any $h_0 \in (-t\delta_n^{-1}, 0)$,

$$\sup_{h>0}\left(G_n(t+\delta_n h) - \sqrt{n}\delta_n^2(ch^2 + \gamma_n xh)\right) \geq \left(G_n(t+\delta_n h_0) - \sqrt{n}\delta_n^2(d\gamma_n h_0^2 + Cxh_0)\right)$$

Choose, $h_0 \equiv -\frac{xC}{2d\gamma_n}$ and note that $ch^2 + \gamma_n xh \geq ch^2$ for $h \geq 0$. So, we can write,

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in(\max(\delta_n x/(2dC\gamma_n),\gamma_n),U)}(\hat{g}_n(t) - g(t)) > x\delta_n\right) \\ &\leq \mathbb{P}\left(\sup_{t\in(0,U)}\left(G_n(t+\delta_n h) - \sqrt{n}\delta_n^2 ch^2 - G_n(t+\delta_n h_0)\right) \geq \sqrt{n}C^2\delta_n^2\frac{x^2}{4d\gamma_n}\right) \\ &\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sup_{t\in(0,U),j\leq h\leq j+1}(G_n(t+\delta_n h) - G_n(t+\delta_n h_0)) \geq \sqrt{n}\delta_n^2\left(cj^2 + \frac{x^2C^2}{4d\gamma_n}\right)\right)\end{aligned}$$

We can define the class of functions

$$\mathcal{G}_{n,j} = \{\mathbf{1}(((t+\delta_n' h_0'), (t+\delta_n h)]) : t \in (0, U), j \leq h \leq j+1\}$$

where, $\delta'_n = \delta_n/\gamma_n$ and $h'_0 = -Cx/(2d)$ and using Markov inequality we get that

$$\mathbb{P}\left(\sup_{t\in(\delta_n x/(2dC\gamma_n),L)}(\hat{g}_n(t)-g(t)) > x\delta_n\right) \leq \text{const}\sum_{j=0}^{\infty}\frac{\mathbb{E}||G_n||_{\mathcal{G}_{n,j}}}{\sqrt{n}\delta_n^2(j^2+x^2/\gamma_n)}$$

Now, using bracketing integral entropy bounds from [157] and $\gamma_n = O((\log n)^{-1/3})$, we get that,

$$\begin{aligned}\text{const}\sum_{j=0}^{\infty}\frac{\mathbb{E}||G_n||_{\mathcal{G}_{n,j}}}{\sqrt{n}\delta_n^2(j^2+x^2/\gamma_n)} &\leq \sum_{j=0}^{\infty}\frac{1}{\sqrt{n}\delta_n^{3/2}(j+x)^{3/2}}\\ &= o\left(\frac{1}{\sqrt{x}}\right)\end{aligned}$$

with the last equality coming by taking $\delta_n = O(n^{-1/3})$ and the RHS goes to zero for$x = x_n \to \infty$. Thus, we can combine all the arguments to get that, for any $\epsilon > 0$, there exists a $x_n > 0$ for sufficiently large n with $\delta_n = n^{-1/3}$ and $\gamma_n = (\log n)^{-1/3}$ such that

$$\mathbb{P}\left(\sup_{x_n\delta_n/\gamma_n\leq t\leq U}|\hat{g}_n(t)-g(t)| > x\delta_n\right) \leq O\left(\frac{1}{\sqrt{x}}\right) < \epsilon$$

(c) Follows from (a) and arguments of Groeneboom (1985) [71].

□

Spline approximation of NPMLE $\hat{g}_n$

In the previous section, we constructed an isotropic regression based NPMLE $\hat{g}_n$ for density generator g. One of the main problems with NPMLE $\hat{g}_n$ is that it is piece-wise constant and thus discontinuous. This is in general a problem with isotonic estimators. A number of works has been done to address this issue and obtain isotonic continuous or smooth estimators. Some of the approaches are - (a) kernel or spline smoothing of isotonic estimators [68], (b) finding isotonic estimator for smoothed empirical CDF, (c) kernel and spline fitting with additional isotonic constraints on the fitted models. For each of these approaches several different methods have been proposed with proper theoretical justification for most of them. But there still exist some unanswered questions in this domain. However, we shall not go into those questions in this discourse. We present approach (a) version here only.

We have already found the NPMLE $\hat{g}_n$ and proved some of its properties in Lemma 1.3.1. Now, consider the density generator

$$g(y) \equiv \exp(\phi(y)) \text{ and } (\phi_1,\ldots,\phi_n) \equiv (\phi(Y_1),\ldots,\phi(Y_n)) \tag{1.19}$$

and the NPMLE in the form

$$\exp(\hat{\phi}_n(y)) \equiv \hat{g}_n(y) \quad = \quad \begin{cases} \hat{g}_i, & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\hat{\phi}_n(y) \quad \equiv \quad \begin{cases} \hat{\phi}_i \equiv log(\hat{g}_i), & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{1.20}$$

Now, consider $\varphi(y)$ to be a twice continuously differentiable monotonically non-increasing function with bounded second derivative and

$$(\varphi(Y_1), \ldots, \varphi(Y_n)) \equiv (\hat{\phi}_1, \ldots, \hat{\phi}_n).$$

We consider the problem of estimating monotonically decreasing $\varphi(y)$ with the help of $(\phi_1, \ldots, \phi_n)$. We have $((Y_1, \phi_1), \ldots, (Y_n, \phi_n))$ as the data and we want to solve regression problem

$$\hat{\phi}_i = \varphi(Y_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{1.21}$$

where, ϵ_i are mean zero random variables with variance σ^2 and exponentially decaying tails. Now, we solve the regression problem by finding the monotone continuous function $\psi(y)$ which minimizes the penalized least-squares loss function

$$L(\psi) = \sum_{i=1}^{n} \left(\hat{\phi}_i - \psi(Y_i)\right)^2 + \lambda \int_0^U (\psi'(t))^2 dt \tag{1.22}$$

Note that the true regression function is $\varphi(y)$. We choose the above smoothing spline loss function in order to get a natural linear spline estimate for the function $\varphi(y)$ and in turn $\phi(y)$ on design points $(Y_1, \ldots, Y_n)$ and thus get a log-linear spline estimate for the density generator $g_p(y)$ on design points $(Y_1, \ldots, Y_n)$.

The algorithm to solve the minimization problem (1.22) was given in [150]. Let us denote the resultant linear spline estimate by $\hat{\psi}_n(y)$. So, our estimate of $\phi(y)$ is a linear spline estimate $\hat{\psi}(y)$ of the form

$$\hat{\psi}_n(y) \quad \equiv \quad \begin{cases} a_i y + b_i, & \text{if } Y_{(i-1)} < y \leq Y_{(i)} \\ 0, & \text{otherwise} \end{cases} \tag{1.23}$$

where, $(a_i, b_i))_{i=1}^n$ are estimated by solving the optimization problem in Eq (1.22).

Pal and Woodroofe (2007) [135] provided the asymptotic properties of the estimator $\hat{\phi}_n(y)$ in the Theorem 2 of their paper [135], which we restate in following lemma

Lemma 1.3.2.

$$\hat{\psi}(y) = \tau_\lambda(y) + \frac{\nu}{n} \sum_{i=1}^{n} \exp(-\nu(y - Y_i))\epsilon_i + O_P\left(n^{-2/3} \log n\right) \nu + \exp(-\nu y(U - y)) O_P(\nu) \tag{1.24}$$

uniformly in λ *and in* $y \in (0, U)$ *with* $\nu = \lambda^{-1/2}$ *and* $\tau_\lambda(y) = \varphi(y) + \lambda\varphi''(y) + o(\lambda)$.

Now, we can use the above lemma and some extensions of it based on [135] to get concentration of $\hat{\psi}_n(y)$ for a special case.

Lemma 1.3.3. *Let $\lambda = O\left(\left(\frac{\log n}{n}\right)^{2/3}\right)$ and we minimize loss function in Eq. (1.22) with the λ considered to get $\hat{\psi}(y)$. Suppose also that $\phi(y)$ defined in Eq. (1.19) is twice continuously differentiable with bounded second derivative and ϕ' bounded away from 0 (from above) and $-\infty$. Then as $n \to \infty$ for each Y_i, for some constants, $c_1, c_2 > 0$, we have,*

$$\mathbb{P}\left[\left|\hat{\psi}_n(Y_i) - \phi(Y_i)\right| \geq t\right] \leq c_1 \exp(-c_2 t). \tag{1.25}$$

Proof. From Lemma 1.3.2 using $\lambda = O\left(\left(\frac{\log n}{n}\right)^{2/3}\right)$ and substituting λ in Eq. (1.24), we have that, for $y \in (0, U)$,

$$\begin{aligned}
\hat{\psi}(y) &= \varphi(y) + o(n^{-2/3}) + \frac{\nu}{n}\sum_{i=1}^{n} \exp(-n^{1/3}(y - Y_i))\epsilon_i \\
&\quad + \nu O_P\left(n^{-1/3}\log n\right) + O_P\left(n^{1/3}\exp(-n^{2/3}y(U-y))\right) \\
\hat{\psi}(y) - \varphi(y) &= \frac{\nu}{n}\sum_{i=1}^{n} \exp(-n^{1/3}(y - Y_i))\epsilon_i \\
&\quad + \nu O_P\left(n^{-1/3}\log n\right) + O_P\left(n^{1/3}\exp(-n^{2/3}y(U-y))\right) + o(n^{-2/3})
\end{aligned}$$

Now, the first term of the RHS has sub-Gaussian concentration with rate $O_P(n^{-1/3})$ following Hoeffding's inequality, since ϵ_i are iid sub-Gaussian random variables. The second term is bounded by $\nu||\hat{\Phi} - \Phi||_\infty$ and $o(1/n)\nu$, where, $\hat{\Phi}(y) = \frac{1}{n}\sum_{i:Y_i \leq y} \hat{\phi}(Y_i)$ and $\Phi(y) = \int_0^y \phi(y)$. Now, $||\hat{\Phi} - \Phi||_\infty$ has sub-Gaussian tails by Marshall's Lemma and Dvoretsky-Kiefer-Wolfowitz Theorem [96], we have that, $||\hat{\Phi} - \Phi||_\infty = O_P\left((\log n/n)^{2/3}\right)$ with sub-Gaussian concentration. So, the second term has sub-Gaussian concentration with rate $O_P(n^{-1/3})$. The third term is bounded by $\frac{\nu}{n}\exp(-\nu(U-y))\sum_{i=1}^{n}\epsilon_i + \nu\exp(-\nu y(U-y))$ and thus has sub-Gaussian concentration with rate $o_P(n^{-1/3})$.

Now, $(Y_1, \ldots, Y_n) \in (0, U)$. So, given $(Y_1, \ldots, Y_n) \in (0, U)$, we have, for some constants $c_1, c_2 > 0$,

$$\mathbb{P}\left[\left(\frac{n}{\log n}\right)^{1/3}\left|\hat{\psi}_n(Y_i) - \hat{\phi}(Y_i)\right| \geq t\right] \leq c_1 \exp(-c_2 t^2). \tag{1.26}$$

From, Lemma 1.3.1(b), we have that,

$$\mathbb{P}\left[\left(\frac{n}{\log n}\right)^{1/3}\left|\exp(\phi(Y_i)) - \exp(\hat{\phi}(Y_i))\right| \geq \exp(t)\right] = O\left(\frac{1}{\sqrt{\exp(t)}}\right).$$

So, we have,

$$\mathbb{P}\left[\left(\frac{n}{\log n}\right)^{1/3}\left(\frac{\exp(\phi(Y_i))}{\exp(\hat{\phi}(Y_i))}-1\right)\geq \exp(t)\right] = O\left(\frac{1}{\sqrt{\exp(t)}}\right),$$

which implies,

$$\mathbb{P}\left[\log n\left|\phi(Y_i)-\hat{\phi}(Y_i)\right|\geq C\log(\exp(t)\pm 1)\right] = O\left(\exp(-t/2)\right),$$

for some constant $C>0$. So, for large $t>0$, we have, for some constants $c_1>0$ and $c_2>0$,

$$\mathbb{P}\left[\log n\left|\phi(Y_i)-\hat{\phi}(Y_i)\right|\geq t\right] = c_1\exp(-c_2 t),$$

Now, combining the above equation with Eq. (1.26), we get that, for each Y_i and large t,

$$\mathbb{P}\left[\left|\hat{\psi}(Y_i)-\phi(Y_i)\right|\geq t\right] \leq c_1\exp(-c_2 t),$$

and thus the Lemma follows. □

1.4 Inference II: Estimation of Euclidean Parameters

The estimation of Euclidean parameters is carried in an iterative fashion. We start with a consistent estimate of the Euclidean parameters and then by using the estimate of density generator in Section 1.3, we try to get better estimates of the Euclidean parameters. In this section, we shall try to devise the estimation procedure for improving the initial estimate of the Euclidean parameters.

Initial Estimates of Euclidean Parameters

We have $X_1,\ldots,X_n$ where $X_i\in\mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot;\mu_0,\Omega_0)$, where, $\Omega_0\equiv\Sigma_0$. We shall try to give different initial estimates of Euclidean parameters for fixed dimension and high-dimensional cases.

Fixed dimensional case

There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [73] is a good source. We can also use sample mean and covariance estimates, as they are also consistent estimates of mean and covariance parameters for the class of Euclidean distributions. We suggest using Stahel-Donoho robust estimator of

multivariate location and scatter [114]. Stahel-Donoho estimators $(\hat{\mu}, \hat{\Sigma})$ of (μ_0, Σ_0) are also weighted mean and covariance matrix estimators, which are of the form

$$\hat{\mu} = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \tag{1.27}$$

$$\hat{\Sigma} = \frac{\sum_{i=1}^n w_i (X_i - \hat{\mu})(X_i - \hat{\mu})^T}{\sum_{i=1}^n w_i} \tag{1.28}$$

where, the weight, w_i is a function on "oulyingness" of a data point X_i from the center $(i = 1, \ldots, n)$. See [73] for more details on weight function $\boldsymbol{w}$. From Theorem 1 of [73], we get $\sqrt{n}$ consistency of the estimators $(\hat{\mu}, \hat{\Sigma})$. So, we have,

$$\sqrt{n}|\hat{\mu}_i - (\mu_0)_i| = O_P(1) \text{ for all } i = 1, \ldots, p \tag{1.29}$$

$$\sqrt{n}|\hat{\Sigma}_{ij} - (\Sigma_0)_{ij}| = O_P(1) \text{ for all } i, j = 1, \ldots, p \tag{1.30}$$

Another alternative is Tyler's M-estimate of Multivariate scatter given in [156], which also gives a $\sqrt{n}$ consistent estimate of Σ.

High-dimensional case

(a) Sample mean, thresholded mean or LASSO estimator can be used to estimate μ

(b) Ledoit-Wolf estimator of covariance and Precision matrix [105], which gives distribution-free consistent estimators of covariance and precision matrix in high-dimensions as $p, n \to \infty$ or graphical lasso estimators [67] can be used to estimate covariance, Σ and precision matrix, Ω.

Estimation of Density Generator Using Estimates $\hat{\mu}_n$ and $\hat{\Sigma}_n$

The difference between the approach in Section 1.3 and this one is that $Y_i = (X_i - \mu)^T \Sigma^{-1} (X_i - \mu)$ is replaced by $\hat{Y}_i = (X_i - \hat{\mu}_n)^T \hat{\Sigma}_n^{-1} (X_i - \hat{\mu}_n)$. But, if we use $\hat{Y}_i$ instead of Y_i in finding the estimate of $g_p(y)$, then, we shall show that we have a new rate of convergence depending on behavior of $||\Omega - \hat{\Omega}||$ and $||\mu - \hat{\mu}||$.

Lemma 1.4.1. *Under conditions of Lemma 1.3.3 and ϕ' being bounded and*

$$\mathbb{P}[||\mu - \hat{\mu}||_2 > t] \leq J_1(t, n, p)$$
$$\mathbb{P}[||\Omega - \hat{\Omega}||_F > t] \leq J_2(t, n, p).$$

Then as $n \to \infty$, for some constants $c, d > 0$

$$\mathbb{P}\left[\left|\hat{\phi}_n(\hat{Y}_i) - \phi(Y_i)\right| > t\right] \leq c \exp(-dt) J_1(t, p, n) J_2(t, p, n) \tag{1.31}$$

Proof. Let us consider $\hat{\phi}_{(\hat{\mu},\hat{\Omega})}$ as the estimate of log-density generator using $(\hat{Y}_i)_{i=1}^n$ as the data and $\hat{\phi}_{(\mu,\Omega)}$ as the estimate of log-density generator using $(Y_i)_{i=1}^n$ as the data. By applying Lemma 1.3.1 and 1.3.3, we get, for some constants, $k_1, k_2, k_3, k_4 > 0$,

$$\begin{aligned}\mathbb{P}\left[\left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i)\right| \geq t\right] &\leq k_1 \exp(-k_2 t)\\ \mathbb{P}\left[\left|\hat{\phi}_{(\mu,\Omega)}(Y_i) - \phi_{(\mu,\Omega)}(Y_i)\right| \geq t\right] &\leq k_3 \exp(-c_4 t)\end{aligned}$$

Now, $\hat{\phi}_n(\hat{Y}_i) \equiv \hat{\phi}_{(\hat{\mu},\hat{\Omega})}$, we want to prove, that for some constants $c, d > 0$,

$$\mathbb{P}\left[\left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi(Y_i)\right| > t\right] \leq c \exp(-dt)$$

Now,

$$\begin{aligned}\left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right| &\leq \left|\hat{\phi}_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i)\right|\\ &\quad + \left|\hat{\phi}_{(\mu,\Omega)}(Y_i) - \phi_{(\mu,\Omega)}(Y_i)\right|\\ &\quad + \left|\phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right|\end{aligned}$$

Since, we already have bounds for first and second term, we only have to bound the third term. Now,

$$\begin{aligned}\left(\phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right) &= \phi\left((X_i-\hat{\mu})^T\hat{\Omega}(X_i-\hat{\mu})\right) - \phi\left((X_i-\mu)^T\Omega(X_i-\mu)\right)\\ &\leq |\phi'|\Big((X_i-\hat{\mu})^T\hat{\Omega}(X_i-\hat{\mu}) - (X_i-\mu)^T\Omega(X_i-\mu)\Big)\end{aligned}$$

$$\begin{aligned}(X_i-\mu)^T\Omega(X_i-\mu) &= (X_i-\hat{\mu}+\hat{\mu}-\mu)^T\Omega(X_i-\hat{\mu}+\hat{\mu}-\mu)\\ &= (X_i-\hat{\mu})^T\Omega(X_i-\hat{\mu}) + 2(X_i-\hat{\mu})^T\Omega(X_i-\hat{\mu}) + (\mu-\hat{\mu})^T\Omega(\mu-\hat{\mu})\\ &= (X_i-\hat{\mu})^T\hat{\Omega}(X_i-\hat{\mu}) + (X_i-\hat{\mu})^T(\Omega-\hat{\Omega})(X_i-\hat{\mu})\\ &\quad +2(X_i-\hat{\mu})^T\Omega(X_i-\hat{\mu}) + (\mu-\hat{\mu})^T\Omega(\mu-\hat{\mu})\end{aligned}$$

So, from the assumptions on the estimators $\hat{\mu}$ and $\hat{\Omega}$ and if ϕ' is bounded, we get that for some constant $k_5, k_6 > 0$,

$$\mathbb{P}\left[\left|\phi_{(\hat{\mu},\hat{\Omega})}(\hat{Y}_i) - \phi_{(\mu,\Omega)}(Y_i)\right| \geq t\right] \leq k_5 \exp(-k_6 t) J_1(p,n,t) J_2(p,n,t)$$

and so the lemma follows. □

Maximum Likelihood Estimation of μ and Ω

From the Section 1.4, we have the data in the form $\hat{Y}_i \equiv (X_i - \hat{\mu})^T \hat{\Omega}(X_i - \hat{\mu})$, which we use to get an estimate of the density generator function g_p in a log-linear spline form like in Eq(1.23) in Section 1.3. So, the linear spline estimate of $\log g$ takes the form

$$\log \hat{g_p}(x) = -\sum_{i=1}^{n-1}(a_i x + b_i)\mathbf{1}\left((\hat{Y}_{(i)}, \hat{Y}_{(i+1)}]\right) - (a_{n+1}x + b_{n+1})\mathbf{1}\left((\hat{Y}_{(n)}, \hat{Y}_{(n+1)})\right) \quad (1.32)$$

where, $\hat{Y}_{(i)}$ is the i^{th} order statistic for $\{\hat{Y}_i\}_{i=0}^{n+1}$ with $\hat{Y}_0 \equiv -\infty$ and $\hat{Y}_{n+1} \equiv \infty$ and $(a_i, b_i)_{i=1}^n$ as define in Eq. (1.23).

Now, if we define $Y_i = (X_i - \mu)^T \Omega (X_i - \mu)$, then, the likelihood function of $\theta = (\mu, \Omega)$ given data $(X_1, \ldots, X_n)$ and density generator g_p becomes -

$$\mathcal{L}(\theta|\boldsymbol{Y}, g_p) = \prod_{i=1}^{n} |\Omega|^{1/2} g_p\left((X_i - \mu)^T \Omega (X_i - \mu)\right) = \prod_{i=1}^{n} |\Omega|^{1/2} g_p(Y_i) \quad (1.33)$$

and the log-likelihood function of $\theta = (\mu, \Omega)$ given data $(X_1, \ldots, X_n)$ and density generator g_p becomes -

$$\ell(\theta|\boldsymbol{Y}, g_p) = \frac{n}{2}\log|\Omega| + \sum_{i=1}^{n} \log g_p(Y_i)$$

Now, we can plug-in the a variant of estimate of $\log g_p$ from Eq (1.32), in the form

$$\log \tilde{g_p}(x) = -\sum_{i=1}^{n-1}(a_i x + b_i)\mathbf{1}\left((Y_{(i)}, Y_{(i+1)}]\right) - (a_{n+1}x + b_{n+1})\mathbf{1}\left((Y_{(n)}, Y_{(n+1)})\right)$$

where, $Y_{(i)}$ is the i^{th} order statistic for $\{Y_i\}_{i=0}^{n+1}$ with $Y_0 \equiv -\infty$ and $Y_{n+1} \equiv \infty$ and plug in $\log \tilde{g}_p$ in place of $\log g_p$, to get the approximated log-likelihood - Then, we can write

$$\begin{aligned} \ell(\theta|\boldsymbol{X}) &= \frac{n}{2}\log|\Omega| - \sum_{i=1}^{n} a_i\left((X_i - \mu)^T \Omega (X_i - \mu)\right) + Constant \\ &= \frac{n}{2}\log|\Omega| - \mathrm{tr}\left(S^*\Omega\right) + Constant \end{aligned}$$

where, $S^* = \sum_{i=1}^n a_i (X_i - \mu)(X_i - \mu)^T$. By maximizing the approximated log-likelihood $\tilde{\ell}(\theta)$ -

$$\tilde{\ell}(\theta|\boldsymbol{X}) = \frac{n}{2}\log|\Omega| - \mathrm{tr}\left(S^*\Omega\right) \quad (1.34)$$

we will get estimates of μ and Ω as

$$(\tilde{\mu}, \tilde{\Omega}) = \arg\max_{\mu, \Omega \succ 0} \tilde{\ell}(\mu, \Omega) \quad (1.35)$$

which we call **robust regularized estimators** of Euclidean parameters. See beginning of Section 1.2 for the notation.

Penalized ML Estimation of μ and Ω: High-dimensional Case

Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values s indicates *sparsity*. In the high-dimensional case, we have the dimension of the Euclidean parameters, p, growing with number of samples, n. We consider the penalized approximated log-likelihood function under assumption of sparsity to be

$$\tilde{\ell}(\mu|\boldsymbol{X}) = -\sum_{i=1}^{n} a_i (X_i - \mu)^T (X_i - \mu) \tag{1.36}$$

$$\tilde{\ell}(\Omega|\boldsymbol{X}, \tilde{\mu}) = \frac{n}{2} \log|\Omega| - \text{tr}\,(S^*\Omega) + \nu|\Omega^-|_1 \tag{1.37}$$

where, $S^* \equiv \sum_{i=1}^{n} a_i (X_i - \tilde{\mu})^T (X_i - \tilde{\mu})$ and

$$\tilde{\mu} = \arg\max_{\mu} \tilde{\ell}(\mu) \tag{1.38}$$

$$\tilde{\Omega} = \arg\max_{\Omega \succ 0} \tilde{\ell}(\Omega) \tag{1.39}$$

are the **robust regularized estimators** of Euclidean parameters. See Section 1.2 for the notation.

Note that, if we had known Ω or if the proper form of elliptic density was known, we could have used penalization in the mean parameter too. So, if Ω is known, then, the penalized likelihood for μ becomes -

$$\tilde{\ell}(\mu|\boldsymbol{X}) = \sum_{i=1}^{n} -a_i (X_i - \mu)^T (X_i - \mu) + \nu_1 ||\mu||_1 \tag{1.40}$$

$$\tag{1.41}$$

and the penalized likelihood estimate, $\tilde{\mu}$ is

$$\tilde{\mu} = \arg\max_{\mu} \tilde{\ell}(\mu) \tag{1.42}$$

1.5 Inference III: Combined Approach and Theory

Now, we can summarize the estimation procedure based on the steps suggested in Section 1.3 and 1.4. We shall provide the estimation procedure in this section and we shall provide the theoretical justification for the method.

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. Then the density function $f(\cdot; \mu_0, \Omega_0)$ is of the form

$$f(x; \mu_0, \Omega_0) = |\Omega_0|^{1/2} g_p\left((x - \mu_0)^T \Omega (x - \mu_0)\right) \tag{1.43}$$

where $\theta_0 = (\mu_0, \Omega_0) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and inverse covariance parameters (Σ_0 is the covariance parameter) respectively with $\mu_0 \in \mathbb{R}^p$ and $\Omega_0 \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x)dx = 1$$

$\mathbb{R}^+ = [0, \infty)$.

Fixed Dimension Case

We first consider the fixed dimensional case, that is when we do not have the dimension of the Euclidean parameters, p, not growing with number of samples, n. The estimation steps are as follows -

(1) Assume that we have an initial consistent estimators $\hat{\mu}$ and $\hat{\Omega}$, such that, $||\hat{\mu} - \mu_0||_2$ and $||\hat{\Omega} - \Omega_0||_F$ concentrates to zero with tail bounds given by functions $J_1(t, n, p)$ and $J_2(t, n, p)$ such that,

$$\begin{aligned} \mathbb{P}[||\mu_0 - \hat{\mu}||_2 > t] &\leq J_1(t, n, p) \\ \mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] &\leq J_2(t, n, p). \end{aligned}$$

So, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(n))$, where, $\omega(n) \to 0$ as $n \to \infty$ with given tail bounds.

There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [73] is a good source

(2) Define $\hat{Y}_i = (X_i - \hat{\mu})^T \hat{\Omega}(X_i - \hat{\mu})$ and based on $(\hat{Y}_1, \ldots, \hat{Y}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (1.14). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (1.23).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ, given by equation (1.34)

$$\tilde{\ell}(\theta|\boldsymbol{X}) = \frac{n}{2} \log|\Omega| - \mathrm{tr}\,(S^*\Omega)\,.$$

where, $S^* = \sum_{i=1}^n a_i (X_i - \mu)(X_i - \mu)^T$. We maximize $\tilde{\ell}(\theta)$ with respect to θ, to get the robust estimates of θ_0.

(4) *(Optional)* We can use the estimates to obtained in Step 4 and repeat Steps 1-3, to get an estimate of g_p. But, both in theory and practice, that does not improve the error rates of the new estimate of g_p.

High-dimensional Case

Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small value of s indicates *sparsity*. We first consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n. The estimation steps are as follows -

(1) Assume that we have an initial consistent estimators $\hat{\mu}$ and $\hat{\Omega}$, such that, $||\hat{\mu} - \mu_0||_2$ and $||\hat{\Omega} - \Omega_0||_F$ concentrates around zero with tail bounds given by functions $J_1(t, n, p)$ and $J_2(t, n, p)$ such that,

$$\begin{aligned} \mathbb{P}[||\mu_0 - \hat{\mu}||_2 > t] &\leq J_1(t, n, p) \\ \mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] &\leq J_2(t, n, p). \end{aligned}$$

So, we have, $||\hat{\mu} - \mu_0||_F = O_P((\omega_1(p, n))$ and $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega_2(p, n))$, where, $\omega(p, n) \to 0$ as $p.n \to \infty$ with given tail bounds.

There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [73] is a good source

(2) Define $\hat{Y}_i = (X_i - \hat{\mu})^T \hat{\Omega} (X_i - \hat{\mu})$ and based on $(\hat{Y}_1, \ldots, \hat{Y}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (1.14). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (1.23).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated penalized log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ under sparsity assumptions, given by equation (1.36)

$$\begin{aligned} \tilde{\ell}(\mu | \boldsymbol{X}) &= \sum_{i=1}^{n} a_i (X_i - \mu)^T (X_i - \mu) \\ \tilde{\ell}(\Omega | \boldsymbol{X}, \tilde{\mu}) &= \frac{n}{2} \log|\Omega| - \text{tr}\,(S^* \Omega) + \nu |\Omega^-|_1 \end{aligned}$$

where, $S^* \equiv \sum_{i=1}^{n} a_i (X_i - \tilde{\mu})^T (X_i - \tilde{\mu})$ and

$$\begin{aligned} \tilde{\mu} &= \arg\max_{\mu} \tilde{\ell}(\mu) \\ \tilde{\Omega} &= \arg\max_{\Omega} \tilde{\ell}(\Omega) \end{aligned}$$

are the *robust regularized estimators* of Euclidean parameters of θ_0.

Note that, if we had known Ω or if the proper form of elliptic density was known, we could have used penalization in the mean parameter too. So, if Ω is known, then, the penalized likelihood for μ becomes -

$$\tilde{\ell}(\mu|\boldsymbol{X}) = -\sum_{i=1}^{n} a_i (X_i - \mu)^T (X_i - \mu) + \nu_1 ||\mu||_1$$

and the penalized likelihood estimate, $\tilde{\mu}$ is

$$\tilde{\mu} \;=\; \arg\max_{\mu} \tilde{\ell}(\mu)$$

The likelihood optimization problems are convex optimization problems and have been the focus of much study in statistical and optimization literature. One way of solving the optimization problem for μ is by using LARS algorithm of [55]. The optimization problem for Ω can be solved by using the graphical LASSO algorithms provided in [67], [164] and [83].

(4) *(Optional)* We can use the estimates to obtained in Step 4 and repeat Steps 1-3, to get an estimate of g_p. But, both in theory and practice, that does not improve the error rates of the new estimate of g_p.

In Section 4.4, we give proof of Theorem 1.2.6, by which we show that regularized estimators of the Euclidean parameters θ in the high-dimensional case is also robust to tail behavior of the underlying elliptical distribution.

Theory

We have described the estimation procedure of the Euclidean parameters and the non-parametric component of the elliptical density in Section 1.5. We now try to show that the estimators have nice behavior in the case of fixed and high dimension. The main theorem in this section is Therem 1.2.5 and Theorem 1.2.6given in Section 1.2. However, to prove the Theorems we first need to discuss the setup and the conditions.

We have $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mu_0, \Omega_0)$, where, $\Omega_0 \equiv \Sigma_0$. Then the density function $f(\cdot; \mu_0, \Omega_0)$ is of the form

$$f(x; \mu_0, \Omega_0) = |\Omega_0|^{1/2} g_p\left((x - \mu_0)^T \Omega (x - \mu_0)\right) \tag{1.44}$$

where $\theta_0 = (\mu_0, \Omega_0) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and inverse covariance parameters (Σ_0 is the covariance parameter) respectively with $\mu_0 \in \mathbb{R}^p$ and $\Omega_0 \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter. We shall also consider that g_p possess **consistency** property.

Now, according to the *consistency* condition of Elliptical distribution mentioned in Lemma 1.2.4, for independent random variables $(\xi_1, \ldots, \xi_n)$ with $\xi_i \stackrel{d}{=} \xi$, $\forall i$,

$$(\sqrt{\xi_1}\Omega_0^{1/2}(X_1 - \mu_0), \ldots, \sqrt{\xi_n}\Omega_0^{1/2}(X_n - \mu_0)) \stackrel{d}{=} (Z_1, \ldots, Z_n)$$

where, $(Z_1, \ldots, Z_n)$ are independent p-variate standard Gaussian random variables. If we define, $W_i \equiv \Omega_0^{1/2}(X_i - \mu_0)$, then,

$$(\sqrt{\xi_1}W_1, \ldots, \sqrt{\xi_n}W_n) \stackrel{d}{=} (Z_1, \ldots, Z_n) \tag{1.45}$$

Now, according to the estimation procedure we have proposed, after the estimation of the density generator by a log-lear spline, according to log-likelihood equation Eq (1.33), the resulting log-likelihood for θ becomes of the form

$$\ell(\theta|\boldsymbol{X}) = \frac{n}{2}\log|\Omega| - \sum_{i=1}^{n} a_i \left((X_i - \mu)^T\Omega(X_i - \mu)\right) + Constant$$

which is like the log-likelihood if estimated density $\hat{f}$ with parameters (μ_0, Ω_0) which has the form -

$$\hat{f}(X_1, \ldots, X_n|\theta_0) = C|\Omega|^{1/2}\exp\left(-\sum_{i=1}^{n} a_i(X_i - \mu_0)^T\Omega_0(X_i - \mu_0)\right).$$

that means that as if the data $W_i \equiv \Omega_0^{1/2}(X_i - \mu_0)$ has the following distributional form -

$$(\sqrt{a_1}W_1, \ldots, \sqrt{a_n}W_n) \stackrel{d}{=} (Z_1, \ldots, Z_n)$$

So, we can see that by our estimation of the non-parametric component, we have got an estimate of the latent scale variable ξ inherent to the consistent elliptical distribution. Our results on rate will thus depend on the tail behavior of ξ.

We wish to prove the Theorem 1.2.6 and Theorem 1.2.5 now. Recall the assumptions (A1)-(A5) given in Section 1.2, which preceded Theorem 1.2.6 and Theorem 1.2.5.

Before proving Theorem 1.2.6, we shall state and prove two lemma on concentration inequalities which are vital for the proof of Theorem 1.2.6. Lemma 1.5.1 is variant of the Lemma B.1 of [26]. Lemma 1.5.2 is variant of the Lemma 3 of [25] or Lemma 1 of [141].

Concentration inequality around the mean parameter μ_0 will be goal of our first Lemma.

Lemma 1.5.1. *Let X_i be i.i.d elliptically distributed random variables having elliptic distribution with parameters (μ_0, Ω_0) and a_i's are as stated in Eq. (1.45) and Ω be an estimate of Ω_0 satisfying Assumption (A2). Then,*

$$\mathbb{P}\left[\max_j \sum_{i=1}^{n} \sqrt{a_i}|(\Omega_0^{1/2}(X_i - (\mu_0)))_j| > 2nt\right] \leq p(c\exp(-dt)J_1(t, n, p)J_2(t, n, p))^n \tag{1.46}$$

where, c, d are some constants.

So, the rate of convergence is controlled by $\sigma_1(p, n)$, which is the function such that the above inequality satisfies if we replace $t = O(\sigma_1(p, n))$

Proof. We have $(\sqrt{\xi_i}\Omega_0^{1/2}(X_i-\mu_0) \stackrel{d}{=} Z_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$. So, we have, by Gaussian tail inequality

$$\mathbb{P}\left[\xi_i(X_i-(\mu_0))^T\Omega_0(X_i-\mu_0) > t^2\right] \;\leq\; c_1\exp(-c_2t^2)$$

From Lemma 1.3.3 and Lemma 1.4.1, we have that,

$$\mathbb{P}\left[(a_i-\xi_i)(X_i-(\mu_0))^T\Omega_0(X_i-\mu_0) > t^2\right] \;\leq\; (k_1\exp(-k_2t^2)J_1(t^2,n,p)J_2(t^2,n,p))$$

Combining the above two equations, we have for some constants $c_3, c_4 > 0$,

$$\begin{aligned}\mathbb{P}\left[a_i(X_i-(\mu_0))^T\Omega_0(X_i-\mu_0) > t^2\right] &\leq c_3\exp(-c_4t^2)J_1(t^2,n,p)J_2(t^2,n,p)\\ \Rightarrow \mathbb{P}\left[\sqrt{a_i}|(\Omega_0^{1/2}(X_i-(\mu_0)))_j|> t\right] &\leq c_3\exp(-c_4t)J_1(t,n,p)J_2(t,n,p)\end{aligned}$$

So, we get that,

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^n\sqrt{a_i}|(\Omega_0^{1/2}(X_i-(\mu_0)))_j|> nt\right] &\leq (c_3\exp(-c_4t)J_1(t,n,p)J_2(t,n,p))^n\\ \mathbb{P}\left[\max_j\sum_{i=1}^n\sqrt{a_i}|(\Omega_0^{1/2}(X_i-(\mu_0)))_j|> 2nt\right] &\leq p(c_3\exp(-c_4t)J_1(t,n,p)J_2(t,n,p))^n\end{aligned}$$

Let us consider that $\sigma_1(p,n)$ to be the function such that the above inequality satisfies if we replace $t = O(\sigma_1(p,n))$. □

So, we can see that depending on the behavior of $H_1(t)$, either $H_1(t)$ or $\exp(-c_2t^2)$ controls the rate in the above Lemma.

Concentration inequality around the covariance matrix parameter Σ_0 will be goal of our next Lemma.

Lemma 1.5.2. *Let X_i be i.i.d elliptically distributed random variables having elliptic distribution with parameters (μ_0, Σ_p) and a_i's are as stated in Eq. (1.45) and $\Omega_p \equiv \Sigma_p^{-1}$. We also have, $\lambda_{\max}(\sigma_p) \leq \overline{k} < \infty$. Then, if $(\Sigma_p)_{ab} = \sigma_{ab}$,*

$$\begin{aligned}&\mathbb{P}\left[\max_{j\neq k}\sum_{i=1}^n|a_i\hat{W}_{ij}\hat{W}_{ik} - (\Sigma_0)_{jk}|> nt\right]\\ &\qquad\leq d_1p^2(H(t)J_1(t)J_2(t))^n(\exp(-nd_2t))(\exp(-nd_3t^2)) \text{ for } |t|\leq\delta \quad (1.47)\end{aligned}$$

where, $\hat{\mu}$ is an of μ_0, d_1, d_2, d_3 and δ depend on $\overline{k}$ only.

So, the rate of convergence is controlled by $\sigma_2(p,n)$, which is the function such that the above inequality satisfies if we replace $t = O(\sigma_2(p,n))$.

Proof. Consider $W_i = X_i - \mu_0$ and $\hat{W}_i = X_i - \hat{\mu}$ for $i = 1, \ldots, n$. We have $\Omega_0^{1/2}\sqrt{\xi_i}W_i \stackrel{d}{=} Z_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$. So, we have, by Lemma 3 of [25], for $|t| \leq \delta$, for some constants $c_1, c_2 > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n}|\xi_i W_{ij}W_{ik} - (\Sigma_0)_{jk}| > nt\right] \leq c_1 \exp(-nc_2t^2) \tag{1.48}$$

We have $a_i > 0$ and $\xi_i > 0$. Now,

$$|a_i W_{ij}W_{ik} - \xi_i W_{ij}W_{ik}| \leq |a_i||W_i||_2^2 - \xi_i||W_i||_2^2|.\frac{|W_{ij}W_{ik}|}{\sum_j W_{ij}^2} \leq |a_i||W_i||_2^2 - \xi_i||W_i||_2^2|$$

Since, a_i from $\hat{\phi}$ is the slope estimate of the log density generator ϕ, whose slope is ξ, conditional on ξ. So, we have from Lemma 1.3.3 and Lemma 1.4.1 for $c_3, c_4 > 0$, that,

$$\mathbb{P}\left[|a_i||W_i||_2^2 - \xi_i||W_i||_2^2| > t\right] \leq c_3 \exp(c_4 t) J_1(t, n, p) J_2(t, n, p)$$

which implies,

$$\mathbb{P}\left[|a_i W_{ij}W_{ik} - \xi_i W_{ij}W_{ik}| > t\right] \leq c_3 \exp(c_4 t) J_1(t) J_2(t) \tag{1.49}$$

Now,

$$|a_i \hat{W}_{ij}\hat{W}_{ik} - a_i W_{ij}W_{ik}| \leq a_i\,|-X_{ij}(\hat{\mu}_k - (\mu_0)_k) - X_{ik}(\hat{\mu}_j - (\mu_0)_j) + (\hat{\mu}_j\hat{\mu}_k - (\mu_0)_j(\mu_0)_k)|$$

So, we have, for some constants, $c_5, c_6 > 0$,

$$\mathbb{P}\left[|a_i \hat{W}_{ij}\hat{W}_{ik} - a_i W_{ij}W_{ik}| > t\right] \leq H(t) c_5 \exp(c_6 t) \tag{1.50}$$

So, combining the above equations (1.48) (1.49) and (1.50), for constants $d_1, d_2, d_3 > 0$, we get that,

$$\mathbb{P}\left[\sum_{i=1}^{n}|a_i \hat{W}_{ij}\hat{W}_{ik} - (\Sigma_0)_{jk}| > 3nt\right] \leq d_1(H(t)J_1(t)J_2(t))^n(\exp(-nd_2t))(d_3\exp(-nd_3t^2))$$

$$\mathbb{P}\left[\max_{j\neq k}\sum_{i=1}^{n}|a_i \hat{W}_{ij}\hat{W}_{ik} - (\Sigma_0)_{jk}| > 3nt\right] \leq p(p-1)d_1(H(t)J_1(t)J_2(t))^n(\exp(-nd_2t))(\exp(-nd_3t^2))$$

Let us consider that $\sigma_2(p, n)$ to be the function such that the above inequality satisfies if we replace $t = O(\sigma_2(p, n))$. □

So, we can see that depending on the behavior of $H(t), J_1(t), J_2(t)$ and these rates along with $\exp(-c_2t^2)$ controls the rate in the above Lemma.

Proof of Theorem 1.2.6

(a) We consider

$$Q(\mu) = \sum_{i=1}^{n} a_i (X_i - \mu)^T \Omega_0 (X_i - \mu) - \sum_{i=1}^{n} a_i (X_i - \mu_0)^T \Omega_0 (X_i - \mu_0)$$

Our estimate $\tilde{\mu}$ given in Eq. (1.38) minimizes $Q(\mu)$ or equivalently $\hat{\delta} = \tilde{\mu} - \mu_0$ minimizes $G(\delta) \equiv Q(\mu_0 + \delta)$ given an estimate Ω of Ω_0. Consider the set

$$\Theta_n(M) = \{\delta : ||\delta||_2 = Mr_n\}$$

where,

$$r_n = \sqrt{p} O(\sigma_1(p, n)) \to 0$$

where, $\sigma_1(p, n)$ is taken from statement of Lemma 1.5.1. Note that $G(\delta)$ is a convex function and

$$G(\hat{\delta}) \leq G(0) = 0$$

Then, if we can show that

$$\inf\{G(\delta) : \delta \in \Theta(M)\} > 0$$

then the minimizer $\hat{\delta}$ must be inside the sphere $\Theta_n(M)$ and hence

$$||\hat{\delta}||_2 \leq Mr_n$$

Now,

$$\begin{aligned}
&\sum_{i=1}^{n} a_i (X_i - \mu)^T \Omega_0 (X_i - \mu) \\
&\qquad = \sum_{i=1}^{n} a_i (X_i - \mu_0)^T \Omega_0 (X_i - \mu_0) + 2 \sum_{i=1}^{n} a_i (X_i - \mu_0)^T \Omega_0 (\mu_0 - \mu) \\
&\qquad\qquad + \sum_{i=1}^{n} a_i (\mu_0 - \mu)^T \Omega_0 (\mu_0 - \mu)
\end{aligned}$$

So,

$$G(\delta) = 2 \sum_{i=1}^{n} a_i (X_i - \mu_0)^T \Omega_0 \delta + \sum_{i=1}^{n} a_i \delta^T \Omega_0 \delta$$

Now, by applying Cauchy-Scwartz inequality and Lemma 1.5.1, and $\bar{a}_1 = \sum_{i=1}^{n} \sqrt{a_i}$ and $\bar{a}_2 = \sum_{i=1}^{n} a_i$, we get that,

$$\begin{aligned} G(\delta) &\geq -2\sum_{i=1}^{n} \sqrt{a_i}\sigma_1(p,n)||\Omega_0^{1/2}\delta||_2 + \sum_{i=1}^{n} a_i \delta^T \Omega_0 \delta \\ &\geq -2\bar{a}_1 M\sqrt{s}(\sigma_1(p,n))^2 + \bar{a}_2 \underline{k} M^2 s(\sigma_1(p,n))^2 \\ &\geq \bar{a}_2(1/(\overline{k} + o_P(1)))M^2 s(\sigma_1(p,n))^2 - 2\bar{a}_1 M\sqrt{s}(\sigma_1(p,n))^2 \\ &> 0 \end{aligned}$$

for large enough $M > 0$. So, for large enough $M > 0$,

$$G(\delta) > 0$$

So, our proof follows.

(b) The proof closely follows proof of Theorem 1 in [141]. We do not repeat the proof as essentially the same proof follows. The only difference are

(i) Take S^* in stead of $\hat{\Sigma}$ in the whole proof.

(ii) Take $r_n = \sqrt{p+s}O(\sigma_2(p,n)) \to 0$, where,$\sigma_2(p,n)$ is taken from statement of Lemma 1.5.2

(iii) Use Lemma 1.5.2 instead of Lemma 1 after the equations (12) and (13) of the proof.

(iv) Use regularization parameter $\nu_2 = \frac{C_1}{\epsilon}O(\sigma_2(p,n))$, where, $\sigma_2(p,n)$ is taken from statement of Lemma 1.5.2

(c) Follows fro Lemma 1.3.1.

Proof of Theorem 1.2.5

Proof of Theorem 1.2.5 becomes a special case of proof of Theorem 1.2.6 as we do not have dependence on p in rate anymore.

1.6 Application to Special Problems

Application to Covariance and Precision Matrix Estimation

The general method of estimation given in Section 1.5 can be used in *robust regularized* estimation of covariance and inverse covariance matrices. Class of Elliptical densities contain densities having tails both thicker and thinner than sub-Gaussian random variables. So, adaptive estimation of covariance matrix from a class of elliptical densities lead to covariance

matrix estimators, which are robust to tail-behavior of the under distribution of the random variable.

Let $X_1, \ldots, X_n$ where $X_i \in \mathbb{R}^p$ are independent elliptically distributed random variables with density $f(\cdot; \mathbf{0}, \Omega)$. Then the density function $f(\cdot; \mu, \Omega)$ is of the form

$$f(x; \mathbf{0}, \Omega) = |\Omega|^{1/2} g_p\left(x^T \Omega x\right) \tag{1.51}$$

where $\theta = (\mathbf{0}, \Omega) \in \mathbb{R}^{p(p+3)/2}$ are the Euclidean mean and covariance parameters respectively with $\Omega \in \mathbb{R}^{p(p+1)/2}$ and $g_p : \mathbb{R}^+ \to \mathbb{R}^+$ is the infinite-dimensional parameter with the property

$$\int_{\mathbb{R}^p} g_p(x^T x) dx = 1$$

$\mathbb{R}^+ = [0, \infty)$. $\Omega \equiv \Sigma^{-1}$ is the inverse covariance parameter and Σ is the covariance parameter.

Now, we consider the high-dimensional situation under the additional structure of sparsity imposed on the Euclidean parameters θ_0. We consider that $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*. We consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n.

Method

If we follow the estimation procedure suggested in Section 1.5, we shell get the **robust regularized** estimate, $\tilde{\Omega}$ of Ω with nice theoretical properties given in Theorem 1.2.6. This procedure can be performed for any other additional structure on the parameters and for any other form of penalization on parameters. As a special case, assume *sparsity condition*: $||\Omega^-||_0 = s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form and ℓ_1 penalty on off-diagonal elements of Ω. The steps of estimation procedure can be stated as -

(1) Assume that we have an initial consistent estimators $\hat{\Omega}$, such that, $||\hat{\Omega} - \Omega_0||_F$ concentrates around zero with tail bounds given by functions $J(t, n, p)$ such that,

$$\mathbb{P}[||\Omega - \hat{\Omega}||_F > t] \quad \leq \quad J_2(t, n, p).$$

So, we have, $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega(p, n))$, where, $\omega(p, n) \to 0$ as $p.n \to \infty$ with given tail bounds.

There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [73] is a good source.

(2) Define $\hat{Y}_i = X_i^T \hat{\Omega} X_i$ and based on $(\hat{Y}_1, \ldots, \hat{Y}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (1.14). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (1.23).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated penalized log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ under sparsity assumptions, given by equation (1.36)

$$\tilde{\ell}(\Omega|\boldsymbol{X}, \tilde{\mu}) \quad = \quad \frac{n}{2}\log|\Omega| - \text{tr}\,(S^*\Omega) + \nu|\Omega^-|_1$$

where, $S^* \equiv \sum_{i=1}^n a_i X_i^T X_i$ and

$$\tilde{\Omega} \quad = \quad \arg\max_{\Omega} \tilde{\ell}(\Omega)$$

are the *robust regularized estimators* of Euclidean parameter Ω.

The likelihood optimization problems are convex optimization problems and have been the focus of much study in statistical and optimization literature. The optimization problem for Ω can be solved by using the graphical LASSO algorithms provided in [67], [164] and [83].

We can get *robust regularized* estimate of covariance matrix Σ by this method.

Theoretical Performance

We start with the following assumption -

(B1) Assume that we have an initial consistent estimators $\hat{\Omega}$, such that, $||\hat{\Omega} - \Omega_0||_F$ concentrates to zero with tail bounds given by functions $J(t, n, p)$ such that,

$$\mathbb{P}[||\Omega_0 - \hat{\Omega}||_F > t] \quad \leq \quad J(t, n, p). \tag{1.52}$$

We have, $||\hat{\Omega} - \Omega_0||_2 = O_P((\omega(p, n))$, where, $\omega(p, n) \to 0$ as $p, n \to \infty$ with given tail bounds.

In the high-dimensional situation we assume the additional structure of sparsity imposed on Ω_0. Density generator g comes from a consistent family elliptical distributions and so by Lemma 1.2.4, $\Omega_0^{1/2}(X) \stackrel{d}{=} Z_p/\sqrt{\xi}$. Let us consider the probability density function of X_j to be h $(j = 1, \ldots, p)$ and

$$H(u) \quad \equiv \quad \int_u^\infty h(v)dv \tag{1.53}$$

We consider the following assumptions -

(B2) Suppose that $||\Omega^-||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*.

(B3) $\lambda_{\min}(\Sigma_0) \geq \underline{k} > 0$, or equivalently $\lambda_{\max}(\Omega_0) \leq 1/\underline{k}$. $\lambda_{\max}(\Sigma_0) \leq \overline{k}$.

(B4) The function $H(t)$ defined in Eq. (1.53) and $J(t,n,p)$ defined in Eq. (1.52), satisfies the following conditions -
there exists a function $\sigma(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $d_1, d_2, d_3 > 0$, for $t = O(\sigma(p,n))$,

$$d_1 p^2 (J(t,n,p))^n (\exp(-nd_2 t))(d_3 \exp(-nd_3 t^2)) \to 0 \text{ as } p, n \to \infty \tag{1.54}$$

Let us consider that we have obtained estimators of the Euclidean parameters $\tilde{\Omega}$ and the nonparametric component $\hat{g}$ by following the estimation procedure of the elliptical density in Section 1.5. We consider the high-dimensional situation now, that means the number of dimensions p and the number of samples n both grow.

Theorem 1.6.1. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int (\phi'')^2 < \infty$. Then, under assumption (B1)-(B4),*

$$||\Omega_0 - \tilde{\Omega}||_F = O_P\left(\sqrt{(p+s)}\sigma(p,n)\right) \text{ for } \nu = O\left(\sigma(p,n)\right).$$

Proof. The proof follows from Theorem 1.2.6 with $\mu_0 = 0$. □

So, we get a consistent estimator $\tilde{\Omega}$ with computable rates of convergence, which is robust against tail behavior. Similarly, we can also get estimates of the covariance matrix Σ and correlation matrix by extending the methods suggested in [141] and [101] in our setup.

Application to Regression with Elliptical Errors

The general method of estimation given in Section 1.5 can be used in *robust regularized* regression. Class of Elliptical densities contain densities having tails both thicker and thinner than sub-Gaussian random variables. So, if we have linear regression with error variables having elliptical distributions, then, adaptive estimation from a class of elliptical densities lead to regression estimators, which are robust to tail-behavior of the error variables. Thus, we shall be able to get **robust regularized** regression estimators.

Let $((Y_1, X_1), \ldots, (Y_n, X_n))$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ are predictor and response variable such that

$$Y_i = \beta_0^T X_i + \epsilon_i$$

where, ϵ_i are independent elliptically distributed random variables with density $f(\cdot; \mathbf{0}, \mathbf{I})$. So, the density function of Y_i is $f(\cdot; \beta_0^T X_i, \mathbf{I})$, where, f is the elliptical density. So, the problem of estimation of β boils down to the problem of estimation of mean parameter of the elliptical density.

Method

For high-dimensional regression, we consider the most vanilla situation and method. We consider the high-dimensional situation under the additional structure of sparsity imposed on the regression coefficients β_0. We consider that $||\beta||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*. We consider the high-dimensional case, that is when we have the dimension of the Euclidean parameters, p, growing with number of samples, n.

(1) Assume that we have an initial consistent estimators $\hat{\beta}$, such that, $||\hat{\beta} - \beta_0||_2$ concentrates around zero with tail bounds given by functions $J(t, n, p)$ such that,

$$\mathbb{P}[||\beta_0 - \hat{\beta}||_2 > t] \quad \leq \quad J(t, n, p)$$

So, we have, $||\hat{\beta} - \beta_0||_F = O_P((\omega(p, n))$, where, $\omega(p, n) \to 0$ as $p.n \to \infty$ with given tail bounds.

There is a rich literature on robust estimates of multivariate location and scale parameters. The book by Hampel et.al. [73] is a good source.

(2) Define $\hat{E}_i = (Y_i - \hat{\beta}^T X_i)^2$ and based on $(\hat{E}_1, \ldots, \hat{E}_n)$, we construct the Grenander type estimator $\hat{g}_n(y)$ of the density generator g_p from the equation (1.14). If g_p is monotone, we get an isotonic linear spline estimate of $\phi(y)$, where, $\exp(\phi(y)) \equiv g_p(y)$, in the form of $\hat{\psi}(y)$, defined by the equation (1.23).

(3) Use the slope estimates of the linear spline estimators $\hat{\phi}_n$ or $\hat{\psi}_n$, to get an approximated penalized log-likelihood loss function, $\tilde{\ell}(\theta)$ for the Euclidean parameters θ under sparsity assumptions, given by equation (1.36)

$$\tilde{\ell}(\mu|\boldsymbol{X}) \quad = \quad \sum_{i=1}^{n} a_i (Y_i - \beta^T X_i)^T (Y_i - \beta^T X_i) + \nu ||\beta||_1$$

and

$$\tilde{\beta} \quad = \quad \arg\max_{\beta} \tilde{\ell}(\beta)$$

are the *robust regularized estimators* of Euclidean parameters of β_0.

The likelihood optimization problems are convex optimization problems and have been the focus of much study in statistical and optimization literature. One way of solving the optimization problem for β is by using LARS algorithm of [55].

Theoretical Performance

Thus following the estimation procedure suggested above, we shell get the **robust regularized** estimate, $\tilde{\beta}$ of β_0 with nice theoretical properties under restrictions on design matrix and coefficient parameters such as given in [26], [166] and [119]. Let us just give one such example of conditions on design matrix and coefficient parameters called *Restricted Eigenvalues* conditions given in [26]. The condition is stated as -

(C1) Assume

$$\kappa(s,p,c_0) \equiv \min_{J_0 \subseteq [p], |J_0| \leq s} \min_{\delta \neq 0, |\delta_{J_0^c}|_1 \leq c_0 |\delta_{J_{01}}|_1} \frac{||X\delta_2||_2}{\sqrt{n}||\delta_{J_{01}}||_2} > 0 \tag{1.55}$$

where, for integers s, m such that $1 \leq s \leq p/2$ and $m \geq s, s+m \leq p$, a vector $\delta \in \mathbb{R}^p$ and a set of indices $J_0 \subseteq \{1,\ldots,p\}$ with $|J_0| \leq s$; denote by J_1 the subset of $\{1,\ldots,p\}$ corresponding to the m largest in absolute value coordinates of δ outside of J_0, and define $J_{01} \equiv J_0 \cup J_1$.

Also,

(C2) Assume that we have an initial consistent estimators $\hat{\beta}$, such that, $||\hat{\beta} - \beta_0||_2$ concentrates to zero with tail bounds given by functions $J(t,n,p)$ such that,

$$\mathbb{P}[||\beta_0 - \hat{\beta}||_2 > t] \quad \leq \quad J(t,n,p) \tag{1.56}$$

We have, $||\hat{\beta} - \beta_0||_F = O_P((\omega(p,n))$, where, $\omega(p,n) \to 0$ as $p, n \to \infty$ with given tail bounds.

We consider the high-dimensional situation. So, in this case, the dimension of Euclidean parameters grows with n. In the high-dimensional situation we assume the additional structure of sparsity imposed on the coefficient parameters β_0. Density generator g comes from a consistent family elliptical distributions and so by Lemma 1.2.4, $\epsilon_i \stackrel{d}{=} Z/\sqrt{\xi}$. Let us consider the probability density function of ϵ_i to be h $(j = 1,\ldots,p)$ and

$$H(u) \quad \equiv \quad \int_u^\infty h(v)dv \tag{1.57}$$

We consider the following assumptions -

(C3) Suppose that $||\beta||_0 \leq s$, where, $||\cdot||_0$ calculates the number of non-zero entries in the vector or the matrix in vectorized form. Small values of s indicates *sparsity*.

(C4) The function $H(t)$ defined in Eq. (1.57) and $J(t,n,p)$ defined in Eq. (1.52), satisfies the following conditions -
there exists a function $\sigma(p,n) : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ is defined such that for constants, $c, d > 0$, for $t = O(\sigma(p,n))$,

$$p(c\exp(-dt)J(t,n,p))^n \to 0 \text{ as } p, n \to \infty \tag{1.58}$$

Let us consider that we have obtained estimators of the coefficient parameters $\tilde{\beta}$ and nonparametric component $\hat{g}$ by following the estimation procedure of the elliptical density in Section 1.5. We consider the high-dimensional situation now, that means the number of dimensions p and the number of samples n both grow.

Theorem 1.6.2. *Define $\phi_p(y) \equiv \log(g_p)$ and assume the following regularity conditions on density generator g_p and $\phi_p(y)$: g_p is twice continuously differentiable with bounded second derivative and derivative ϕ' and g' bounded away from 0 (from above) and $-\infty$ and $\int(\phi'')^2 < \infty$. Then, under assumption (C1)-(C4), with $c_0 = 3$ in (C1),*

$$||\beta_0 - \tilde{\beta}||_2 = O_P\left(\sqrt{s}\sigma(p,n)\right) \tag{1.59}$$

Proof. The proof follows the same steps as proof of Theorem 7.2 of [26] with only the concentration inequality portion replaced by the concentration inequality of Theorem 1.2.6 and Lemma 1.5.1. □

This procedure can be performed for any other additional structure on the regression parameters and for any other form of penalization of loss function. These gives a lot of scope for future work.

Mixture of Elliptic Distributions

Let $X_1, \ldots, X_n \sim \sum_{k=1}^{K} p_k f_k(x; \mu_k, \Omega_k)$, where, $X_i \in \mathbb{R}^p$ and $f_k(\cdot; \mu_k, \Omega_k)$ is the density of elliptic distribution of the form

$$f(x; \mu_k, \Omega_k) = |\Omega_k| g_k\left((x-\mu_k)^T \Omega_k (x-\mu_k)\right) \tag{1.60}$$

where, $g_k(\cdot)$ is a density generator, that is, a non-negative function on $[0, \infty)$ such that the spherically symmetric (around zero) function $g_k(x^T x)$, $x \in \mathbb{R}^p$ integrates to 1 and g_k is non-increasing in $[0, \infty)$ so that the density is unimodal.

Our goal is the estimation of Euclidean parameters $\boldsymbol{\theta} = (\mu_k, \Omega_k)_{k=1}^K$ or $\boldsymbol{\theta} = (\mu_k, \Omega_k)_{k=1}^K$ in the high-dimensional setting as well as the infinite-dimensional parameters $\boldsymbol{G} = (g_1, \ldots, g_K)$

EM Algorithm

We have data $\boldsymbol{X} = (X_1, \ldots, X_n)$, where, $X_i \overset{i.i.d}{\sim} p(x; \boldsymbol{\theta}, \boldsymbol{G}, \boldsymbol{\pi})$. The complete data vector is given by $\boldsymbol{X_c} = (\boldsymbol{Z}, \boldsymbol{X})$, where, $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ and each $Z_i \in \{0,1\}^K$ indicates component label. The complete data log likelihood is given by

$$\begin{aligned} \ell_c(\boldsymbol{\phi}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik} \log \pi_k + \sum_{k=1}^{K} \lambda_{1k}||\mu_k||_1 + \sum_{k=1}^{K} \lambda_{2k}||\Omega^-||_1 \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left(\frac{1}{2}\log|\Omega_k| + \log g\left(tr\left((x-\mu_k)^T(x-\mu_k)\right)\Omega_k\right)\right) \end{aligned}$$

The conditional log likelihood is given by

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\phi}) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\log\pi_k + \sum_{k=1}^{K}\lambda_{1k}||\mu_k||_1 + \sum_{k=1}^{K}\lambda_{2k}||\Omega^-||_1 \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}\left(\frac{1}{2}\log|\Omega_k| + \log g\left(tr\left((x-\mu_k)^T(x-\mu_k)\right)\Omega_k\right)\right)\end{aligned}$$

where, $\tau_{ik} = \mathbb{E}[Z_{ik}|\boldsymbol{X},\boldsymbol{\phi}]$

For m^{th} iteration of the algorithm, we start with $\boldsymbol{\phi}^{(m)}$ and
E step We estimate $\tau_{ik}^{(m+1)} = \mathbb{E}[Z_{ik}|\boldsymbol{X},\boldsymbol{\phi}^{(m)}]$.
M step We maximize $\mathcal{Q}(\boldsymbol{\phi})$ with $\boldsymbol{\tau}^{(m+1)}$ and the m^{th} iterate estimates are

$$\begin{aligned}\pi_k^{(m+1)} &= \frac{\sum_{i=1}^{n}\tau_{ik}^{(m+1)}}{\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(m+1)}} \\ \mu_k^{(m+1)} &= \arg\min_{\mu}\sum_{i=1}^{n}\tau_{ik}\left(\log g_k^{(m)}\left(tr\left((X_i-\mu)(X_i-\mu)^T\right)\Omega_k^{(m)}\right)\right) + \lambda_{1k}||\mu||_1 \\ \Omega_k^{(m+1)} &= \arg\min_{\Omega\succ 0}\sum_{i=1}^{n}\tau_{ik}\left(\log g_k^{(m)}\left(tr\left((X_i-\mu)(X_i-\mu)^T\right)\Omega\right) - \frac{1}{2}\log|\Omega|\right) \\ &+ \lambda_{2k}||\Omega^-||_1 \\ g_k^{(m+1)} &= \text{log-linear spline estimate of } g_k \text{ based on} \\ &\quad (X_i-\mu_k^{(m+1)})^T\Omega_k^{(m+1)}(X_i-\mu_k^{(m+1)}) \text{ for } i=1,\ldots,n\end{aligned}$$

Theoretical Results

The family of distributions $\boldsymbol{P} = \{P_{(\boldsymbol{\theta},\boldsymbol{G})}\}$ becomes identifiable under the conditions given in [81]. The condition states that for elliptical distributions with consistency property, that is, elliptical distributions of the form $X_p \overset{d}{=} \frac{Z_p}{\xi}$ given in Lemma 1.2.4, we have identifiability of Euclidean parameters for mixtures of such elliptic distributions if density of ξ, h exists and satisfies

$$\lim_{r\to 0}\frac{h(r)}{h(ar)} = 0 \text{ for } a > 1 \tag{1.61}$$

The result is given in Theorem 4 of [81].

Theorem 1.6.3. *Assume that we have a mixture of elliptical distributions $\boldsymbol{P} = \{P_{(\boldsymbol{\theta},\boldsymbol{G})}\}$ with consistency property. Also, the scale parameter, ξ as defined in Lemma 1.2.4 of each elliptical distribution component satisfies 1.61. Also, the Euclidean parameters of each elliptical distribution component satisfies the conditions mentioned in Theorem 1.2.6 and they lie within a compact set. Then, we have -*

The EM algorithm converges to a stationary point or local maxima of the penalized likelihood function.

Proof. We are maximizing the penalized likelihood at each **M Step** of the EM algorithm by following the steps of penalized maximum likelihood inference in Section 1.5. Thus, we get a hill-climbing algorithm. Also, the penalized likelihood function is bounded from above. So, the sequence of estimators obtained by EM iterations converges to a stationary point of the penalized likelihood function, since we are within a compact set. □

1.7 Simulation Examples

We simulate from high-dimensional Gaussian and t-distribution and try to estimate the inverse covariance matrix.

Estimation of High-dimensional Covariance Matrix and Density Generator

We have number of samples $n = 400$ and dimension of the data vectors as $p = 400$. We generate the Gaussian distribution with mean $\mathbf{0}$ and banded covariance matrix. The estimated covariance matrices are given in Figure 1.2. The top row of Figure 1.2 are the original covariance matrix, empirical covariance matrix and Graphical LASSO estimate from left to right. The bottom row of Figure 1.2 are the banded estimated covariance matrix, robust estimated covariance matrix and robust regularized estimated covariance matrix (our method) from left to right.

We have number of samples $n = 400$ and dimension of the data vectors as $p = 400$. We generate the t distribution having 2 degree of freedom and a banded covariance matrix. The estimated covariance matrices are given in Figure 1.6. The top row of Figure 1.6 are the original covariance matrix, empirical covariance matrix and Graphical LASSO estimate from left to right. The bottom row of Figure 1.6 are the banded estimated covariance matrix, robust estimated covariance matrix and robust regularized estimated covariance matrix (our method) from left to right. We also give the estimated density generators for the Gaussian distribution and t distribution cases.

1.8 Real Data Examples

We use our method in two different applications. One application is from biology and another is from astronomy.

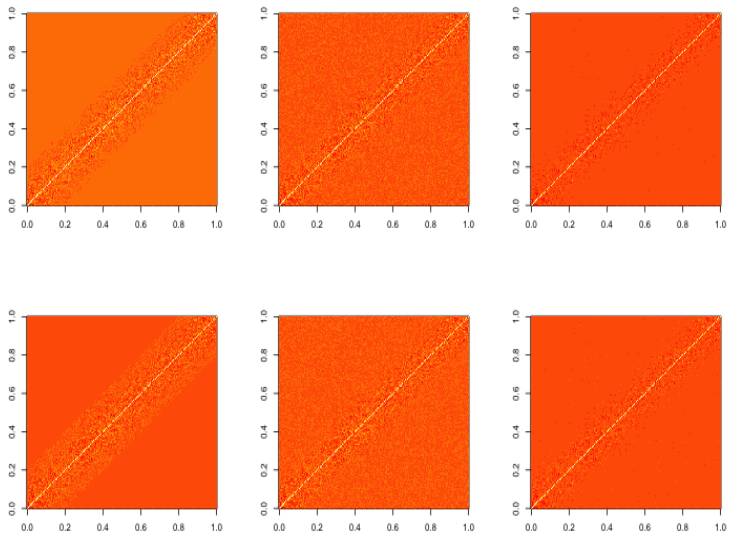


Figure 1.2: For $n = 400$ and $p = 400$, we get different covariance estimators for a banded covariance matrix of normal distribution.

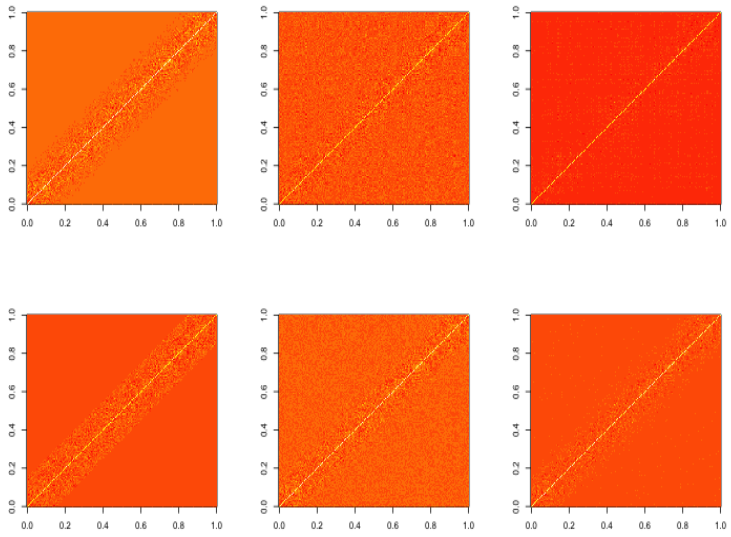
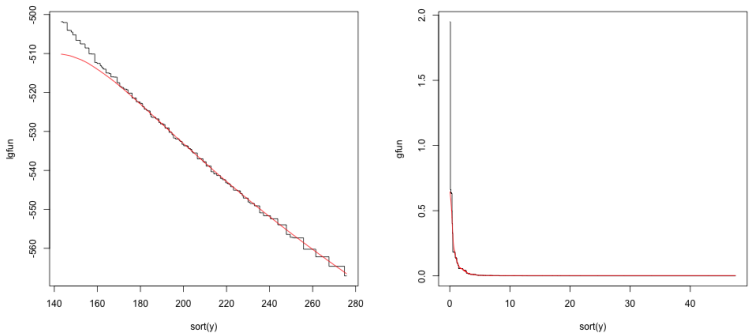


Figure 1.3: For $n = 400$ and $p = 400$, we get different covariance estimators for a banded covariance matrix. of t-dist



(a) (b)

Figure 1.4: For $n = 400$, we get estimator of density generator g_p for normal and t.

High-dimensional Covariance Estimation in Breast Cancer Data

In the biological data set, we focus on selecting gene expression profiling as a potential tool to predict them breast cancer patients who may achieve pathologic Complete Response (pCR), which is defined as no evidence of viable, invasive tumor cells left in surgical specimen. pCR after neoadjuvant chemotherapy has been described as a strong indicator of survival, justifying its use as a surrogate marker of chemosensitivity. Consequently, considerable interest has been developed in finding methods to predict which patients will have a pCR to preoperative therapy. In this study, we use the normalized gene expression data of 130 patients with stage I-III breast cancers analyzed by Hess et al. (2006) [77]. Among the 130 patients, 33 of them are from class 1 (achieved pCR), while the other 97 belong to class 2 (did not achieve pCR).

To evaluate the performance of the penalized precision matrix estimation using three different penalties, we randomly divide the data into training and testing sets of sizes 109 and 21, respectively, and repeat the whole process 100 times. To maintain similar class proportion for the training and testing datasets, we use a stratified sampling: each time we randomly select 5 subjects from class 1 and 16 subjects from class 2 (both are roughly 1/6 of their corresponding total class subjects) and these 21 subjects make up the testing set; the remaining will be used as the training set. From each training data, we first perform a two-sample t-test between the two groups and select the most significant 120 genes that have the smallest p-values. In this case, the dimensionality p = 120 is slightly larger than the sample size n = 109 for training datasets in our classification study. Due to the noise accumulation demonstrated in Fan and Fan (2008), p = 120 may be larger than needed for optimal classification, but allows us to examine the performance when $p > n$. Second, we perform a gene-wise standardization by dividing the data with the corresponding standard deviation, estimated from the training dataset. Finally, we estimate the precision matrix and covariance matrix for both the classes for the training data using our method and standard

graphical LASSO estimates. We find that our method gives sparser estimates of both inverse covariance and covariance matrices. The mean number of non zeros are given in the following table -

	Graphical LASSO	Our Method
Covariance Matrix	5312	4616
Inverse Covariance Matrix	486	412

Table 1.1: The number of non-zero elements in estimators of covariance and inverse covariance matrix in Breast Cancer Data.

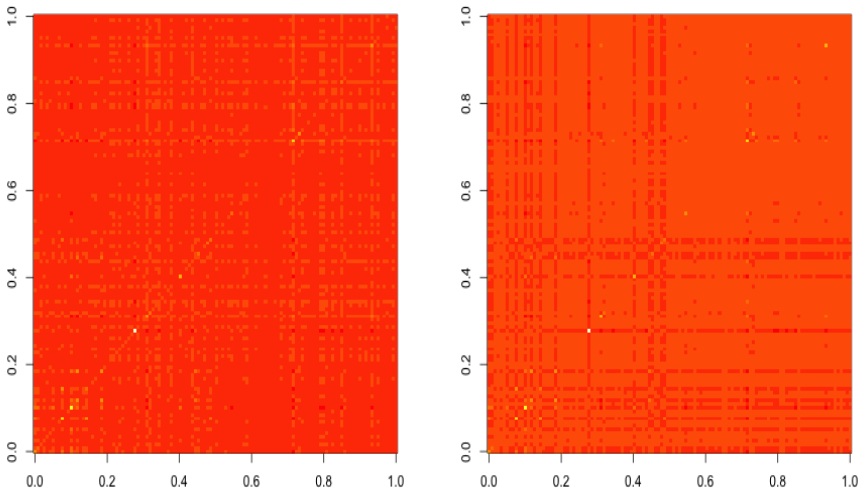


Figure 1.5: Breast Cancer Data covariance matrix estimators using Graphical Lasso (Left) and our method (Right).

1.9 Conclusion

We have developed adaptive estimation procedure for estimation of elliptical distribution for both low and high-dimensional cases. The method of estimation is novel and it gives us a way to move from fixed-dimensional to high-dimensional case quite naturally. We have developed estimation procedure for the density generator function of the elliptical distribution in a log-linear spline form in Section 1.3 and derive respective error bounds. We use

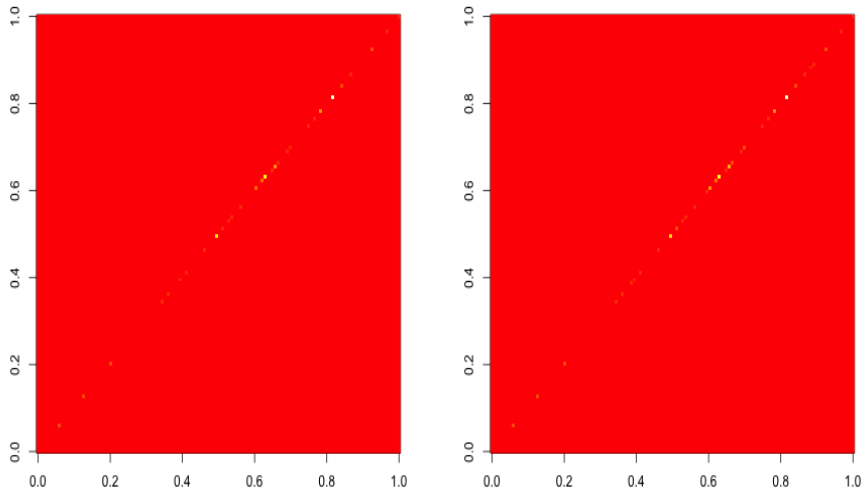


Figure 1.6: Breast Cancer Data inverse covariance matrix estimators using Graphical Lasso (Left) and our method (Right).

the estimate of density generator function of elliptical distribution to adaptively estimate Euclidean parameters of elliptical distribution.

For the estimation of Euclidean parameters, we devise a weighted loss function, where, the weights come from the slopes of estimated density generator function. As a result we have a very natural extension of squared error loss function and with the help of this weighted squared error loss function, we are able to estimate mean and covariance matrix parameters coming from distributions with widely varying tail behavior. So, we get robust estimates of the mean and covariance matrix.

Now, for the high-dimensions case too, weighted least squares loss function is a natural generalization of the least squares loss function, but with this simple generalized loss function, we are able to handle random variables coming from widely varying tail behaviors. As a result we can obtain estimators which are both regularized and robust in high-dimensions. Our approach is not the only approach in statistics literature which can produce estimators that have this dual property of being both robust to changing tail conditions and regularized to constrained parameter spaces in high-dimensions, but it is a quite natural one.

We have indicated three special cases, for which our method can be independently developed -(a) Estimation of Covariance and Precision matrix (b) Regression with Elliptical errors and (c) Clustering via mixtures of elliptical distribution in Section 1.6. For all of these cases, our method can give robust estimates, which can be regularized in high-dimensions.

Feasible algorithms are quite easily obtainable for our method, as most algorithms that

work on least squares loss function also work for weighted least square loss functions too. So, we give an easy approximation to a hard optimization problem and try to solve an optimization problem, which is much easier to handle. As a result our method can borrow strength from existing optimization literature.

So, we have provided an estimation procedure of mean and covariance matrix parameters of elliptic distributions, which is adaptive to the tail behavior and given some theoretical justification for the estimators. The procedure can be extended to use in several classical statistical problems of regression, classification and clustering, thus making our method a very important stepping stone for developing future natural robust regularized estimators.

Chapter 2

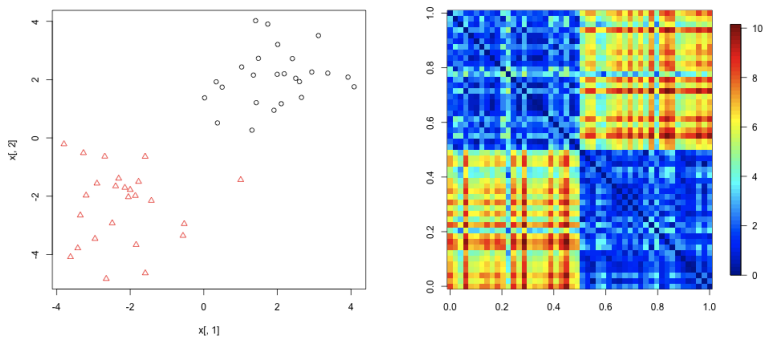
A Naive approach to detecting number of clusters

2.1 Introduction

Cluster analysis is an important unsupervised classification technique. In clustering, a set of unlabeled patterns, usually vectors in a multidimensional space, are grouped into clusters in such a way that patterns in same cluster are similar in some sense and patterns in different clusters are dissimilar in the same sense. One of the main approaches of clustering is optimal partitioning algorithms. The first step of optimal partitioning algorithm is choosing the number of groups or clusters.

The method of finding number of clusters that we have proposed depends upon exploiting the structure of the similarity (or distance) matrix after clustering. One way of viewing clustering is getting hold of the most block-diagonal form of the similarity matrix, by simultaneously permuting the rows and columns of the similarity matrix. In figure 1(a) we have a data set, whose distance matrix is given in figure 1(b) after permuting the row and columns according to the assignments obtained from output of k-means algorithm with 2 clusters on the data set. We see that the matrix in figure 1(b) is block-diagonal in nature. So, our method is based on the assumption that if the partitioning method is applied with correct number of clusters, then the resulting similarity (or distance) matrix will have a better block-diagonal structure.

Now, we test the 'block-diagonal-ness' of a matrix by hypothesis test of location shift. We test if there is a location shift between the distances in a diagonal block with the distances in an off-diagonal block. If there is evidence of location shift, that means that cluster is well-separated from other clusters. So, it is also a cluster validation technique, which determines, whether the current cluster under consideration is actually a well-separated cluster from the remaining clusters. If there is evidence of location shift for all blocks/clusters, then, that means that number of blocks/clusters is one possible choice for number of clusters. So, we have several possible choices for the number of clusters and this is expected if we consider



(a) (b)

Figure 2.1: (a) Sample data set (b)Distance matrix after 2-means clustering

Hartigan's (1985) [76] definition of high-density clusters, where, depending on the level, the number of disjoint components of the level set of the density (that means, number of clusters) vary. However, if we have to specify one number as the number of clusters, we shall prefer the one with most deviation from the null distribution. Also, note that our method works for selecting number of clusters for any clustering/partitioning algorithms.

Several Methods have been proposed for choosing the number of clusters in the literature. Milligan and Cooper (1985) [121] performed a simulation study comparing different statistical heuristics for choosing number of clusters, among which the best were by Calinski and Harabasz (1974) [35]:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{2.1}$$

where, $B(k)$ and $W(k)$ are the between and within cluster sums of squares with k clusters. Rousseeuw (1987) [**rousseeuw1987sihouettes**] proposed the cluster silhouette coefficient

$$SC(k) = \frac{1}{n}\sum_{i=1}^{n} s(i); \quad s(i) = \frac{b(i)-a(i)}{\max a(i), b(i)} \tag{2.2}$$

where, For observation i, let $a(i)$ be the average distance to other points in its cluster, and $b(i)$ the average distance to points in the nearest cluster besides its own and nearest is defined by the cluster minimizing this average distance. The $\hat{k}$ for which $SC(k)$ is maximized is considered the best number of clusters by this method. Tibshirani et. al. (2001) [154] proposed gap statistic for estimating the number of clusters. Another popular tool for selecting number of clusters is using cluster stability, as proposed by Ben-hur et. al. (2001) [15] and Lange et. al. (2004) [103]. There are also methods for selecting number of clusters through BIC in model-based clustering as proposed in Fraley and Raftery (1998) [64]. There are many other methods of finding number of clusters in the literature, but for brevity, we are not mentioning them here.

The chapter is arranged as follows. In section 2.2, we have introduced our method. In section 2.3, we have carried out simulation study, showing the efficacy of our method compared to other methods. In section 2.4, we have applied our method to two real-life data sets - an astronomical data set and a microarray data on leukemia study. In section 2.5, we have provided the discussion of the results and the method.

2.2 Our Method

Let us consider, for data $\mathbf{X} = (X_1, \ldots, X_n)$, where, $X_i \in \mathbb{R}^p$, we start with a distance matrix $D = ((d_{ij}))_{i,j=1}^n$, where, d_{ij} = distance between the observations X_i and X_j. We also have a clustering/partitioning method, which partitions the data into clusters, after the number of clusters have been specified. Let us consider, that for number of clusters, k, the partitioning method partitions the data into clusters $(C_1, \ldots, C_k)$, where, $C_j \subset \mathbf{X}$ for $j = 1, \ldots, k$ and $\cup_j C_j = \mathbf{X}$, $C_i \cap C_j = \phi$, for all $i \neq j$.

Now, if we consider the distance matrix with the row-column entries of the matrix being ordered according to the clusters, that is, consider the permutation of the data entries according to the clusters, $\mathbf{X}_\pi = (X_{\pi(1)}, \ldots, X_{\pi(n)} = (C_1, \ldots, C_k)$. We form the distance matrix $D^\pi = ((d_{ij}^\pi))$ from $\mathbf{X}_\pi$ by d_{ij}^π = the distance between $X_{\pi(1)}$ and $X_{\pi(j)}$.

One of the necessary conditions for the matrix D^π is - it should be 'block-diagonal'. That means the entries in the diagonal blocks of the matrix D^π should have a lower values than the value of the entries in the off-diagonal blocks. We can denote $D^\pi = ((D_{ii'}^\pi))$ as the $k \times k$ block matrix, where, $D_{ii'}^\pi$ is $|C_i| \times |C_{i'}|$ matrix containing the distances between the observations in clusters C_i and $C_{i'}$, where, $i, i' = 1, \ldots, n$. So, in ideal case, the entries in D_{ii}^π should have lower values than entries in D_{ij}^π, where, $i \neq j$. We judge whether a clustering is valid by testing this statement. Also, this is a cluster-wise validation, as for each cluster (or block), we are testing whether the corresponding diagonal block in D^π has smaller values than the corresponding off-diagonal blocks of D^π.

Block Diagonal Hypothesis Testing

We want to test the hypothesis that D^π has block diagonal structure. We proceed as follows - for each block C_i $(i = 1, \ldots, k)$, consider the $\frac{|C_i|(|C_i|-1)}{2}$ upper diagonal entries of D_{ii}^π in the vectorized form $(Y_1, \ldots, Y_m)$, where, $m = \frac{|C_i|(|C_i|-1)}{2}$. Now, for each $i' = (i+1, \ldots, n)$, consider the $|C_i||C_{i'}|$ entries of $D_{ii'}^\pi$ in a vectorized form $(Z_1, \ldots, Z_{m'})$, where, $m' = |C_i||C_{i'}|$.

So, we have two data sets $\mathbf{Y} = (Y_1, \ldots, Y_m)$ and $\mathbf{Z} = (Z_1, \ldots, Z_{m'})$ and we want to find whether $\mathbf{Y}$ has smaller values than $\mathbf{Z}$ in average. So, we perform a one-sided location shift test between $\mathbf{Y}$ and $\mathbf{Z}$, with the null hypothesis of no location shift. Now, note that, this is a non-standard location shift test, since, the data $\mathbf{Y}$ and $\mathbf{Z}$ are not independent. So, the null distribution of the standard location shift tests (t-test, Wilcoxon Rank-Sum test etc.) does not hold in this case. We have to go for a different route to get hold of a null distribution for the test statistic we use.

Using Permutation Test

Getting hold of a null or reference distribution for testing cluster structure is always a challenge, as indicated in Tibshirani et. al. [154]. In this case, let us consider $T_{ii'}$ is the test statistic we are using to test for a location shift between $\mathbf{Y}$ and $\mathbf{Z}$. Now, among among all $i' = i+1, \ldots, n$, consider T_i the test statistic $T_{ii'}$ that is least favorable towards the alternative, for example, for t-test statistic, $T_i = \max_{i'} T_{ii'}$. Now, the null distribution of this statistic is difficult to find theoretically. So, we perform permutation test instead. We permute the row-column entries of the distance matrix D^{π} to generate a new distance matrix $D^{\pi'}$ and generate the corresponding test-statistic T_i for the matrix $D^{\pi'}$. We, repeat this procedure, to get a null-distribution of the test-statistic T_i. Using the null distribution, we test for the location shift for i^{th} cluster.

Now, there is a problem with this approach. The objective of any clustering algorithm is to find the permutation of the row-column entries of the distance matrix D, such that, the permuted matrix D^{π} has most block-diagonal structure. So, through permutation, we are not actually finding the null distribution of T_i for each cluster i. So, we use the p-values generated from the permutation test to get a coefficient called permutation coefficient as follows - for fixed level α (usually 0.01) we find the α^{th} quantile, Q_i of the 'null' distribution of T_i generated through permutations. Now, we define, for each cluster, the *excess value* (ev_i) as

$$ev_i = Q_i - T_i \tag{2.3}$$

Then, we define the block-diagonal coefficient for the number of clusters k as

$$BDQ(k) = \min_i ev_i \tag{2.4}$$

The estimated number of clusters $\hat{k}$ is defined as

$$\hat{k} = \max_k BDQ(k) \tag{2.5}$$

Now, $\hat{k}$ is the most prominent number of clusters. But, there might be other possible number of clusters for which the partition of the data set makes sense. So, we also output a list of potential number of clusters. That is done as follows - let us consider i^* achieves the minimizer in (4). If the value of T_{i^*} is less than $Q_{i^*} - 1.5IQR$, where, IQR is the inter-quartile range of the distribution of permuted T_{i^*}, we call, the corresponding k as a potential cluster number.

Also, if the set of potential clusters is a null set. Then, it implies the lack of cluster structure, which can mean either that there is only one cluster in the data or the clustering algorithm is failing to properly cluster the data. So, by our method, we can also detect one cluster, which many of methods for detecting cluster number (like silhouette, C-H) fail to find.

An Example

We apply our method on a very well-known data set - Fisher's Iris data set [61]. The data set contains 4 measurements for a sample of 150 flowers. There are 3 types of flowers in the data set. The scatter plot based on first 2 types of measurements in given in figure 2.

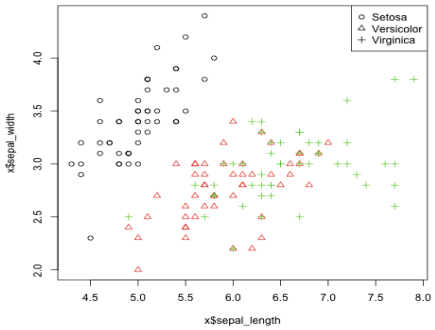


Figure 2.2: Iris Data with 2 dimensions sepal length and width.

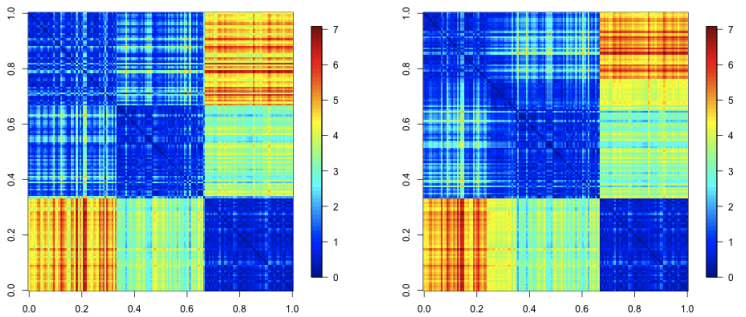
We use partitioning around medoid (PAM) as the clustering method. We use t-test statistic as the statistic for hypothesis testing here. Then, if we use, our method to choose the number of clusters, then the value of $BDQ(k)$ is maximized for $k = 2$. Actually, the only positive values of BDQ come to be $BDQ(2) = 96.5$ and $BDQ(3) = 7.03$. So, we see that by our analysis, $\hat{k} = 2$. However, if we try to find the set of potential number of clusters, then, $k = 3$ also becomes a potential cluster number, as $Q_{i^*} = -6.27$, $T_{i^*} = -13.3$ and $IQR = 2.6$ for $k = 3$. So, we see that we can identify clusters for different hierarchies according to our method and we know, where to stop. The distance matrices for $k = 2$ and $k = 3$ also gives the proper intuitions.

2.3 Simulation Study

For simulation study, we generated data for four different scenarios -

(a) *No Cluster*: We have generated data uniformly over a unit square in 10 dimensions.

(b) *2, 3, 4, 5 Random Clusters in 7, 8, 9 and 10 dimensions respectively*: We generated clusters centers randomly from $N(0, 2I)$ distribution, such that, two clusters centers are at least one unit apart. Then, we generated clusters of size 25 or 50 (randomly chosen) from normal distributions with mean as cluster centers and identity covariance matrix.



(a) (b)

Figure 2.3: (a) Distance matrix after 2-means clustering (b) Distance matrix after 3-means clustering

(c) *2 elongated clusters in three dimensions* We generated each cluster as follows: For cluster 1, set $x_1 = x_2 = x_3 = t$ with t taking on 100 equally spaced values from -.5 to .5 and then Gaussian noise with standard deviation .1 is added to each feature. Cluster 2 is generated in the same way, except that the value 2 is then added to each feature. The result is two elongated clusters, stretching out along the main diagonal of a three-dimensional cube.

(d) *2 close and elongated clusters in three dimensions* As in the previous scenario, with cluster 2 being generated in the same way as cluster 1, except that the value 1 is then added to the first feature only.

The scenarios are motivated from Tibshirani and Walther (2005) [153].

We have repeated each experiment 50 times. For scenario (a), we compare our method with gap statistic. For scenario(b)-(d), we compare our method with CH, silhouette and stability criterion. The stability method has been adopted from Brock et. al. (2008) [31]. We use PAM as the clustering method and t-test statistic as the statistic for location shift testing. We represent our methods by BDQ and BDQ.potential. The BDQ.potential lists the number of times a cluster number becomes a potential candidate for the data set. So, the sum of the elements in rows of BDQ.potential will not be 50, as each data set can have more than one potential clusters. The results are provided in table 1 - 4.

We can see that for scenario (a) BDQ performs better. For scenario (b), for number of clusters 3 and 4, BDQ.potential performs best. For the comparatively hard scenarios (c) and (d), BDQ and BDQ.potential performs quite well. Especially, BDQ.potential almost always include the correct number of clusters within its potential choices.

Table 2.1: Number of Clusters for Scenario (a)

	$k=1$	$k=2$	$k=3$
gap	38	11	1
BDQ	50	0	0

Table 2.2: Number of Clusters for Scenario (b)

$\mathbf{k}=2$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	48	0	0	0	0	0	2
CH	0	48	1	1	0	0	0	0
Stability	0	47	2	0	0	0	0	1
BDQ	11	39	0	0	0	0	0	0
BDQ.potential	11	39	0	0	0	0	0	0
$\mathbf{k}=3$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	25	25	0	0	0	0	0
CH	0	27	23	0	0	0	0	0
Stability	0	27	23	0	0	0	0	0
BDQ	4	30	16	0	0	0	0	0
BDQ.potential	4	40	33	0	0	0	0	0
$\mathbf{k}=4$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	11	16	23	0	0	0	0
CH	0	12	18	20	0	0	0	0
Stability	0	17	11	22	0	0	0	0
BDQ	4	24	7	15	0	0	0	0
BDQ.potential	4	35	30	28	0	0	0	0
$\mathbf{k}=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	5	7	13	25	0	0	0
CH	0	10	10	18	12	0	0	0
Stability	0	10	7	10	23	0	0	0
BDQ	5	26	9	3	7	0	0	0
BDQ.potential	5	31	18	12	17	0	0	0

Table 2.3: Number of Clusters for Scenario (c)

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	50	0	0	0	0	0	0
CH	0	0	0	7	0	32	3	8
Stability	0	50	0	0	0	0	0	0
BDQ	0	50	0	0	0	0	0	0

Table 2.4: Number of Clusters for Scenario (d)

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k \geq 9$
Silhouette	0	4	0	10	2	30	3	1	0
CH	0	0	0	0	0	10	2	26	12
Stability	0	44	1	5	0	0	0	0	0
BDQ	0	11	1	34	4	0	0	0	0
BDQ.potential	0	50	14	48	43	0	0	0	0

2.4 Study on Two Real Data Sets

We apply our method to two real data sets. We compare our method in these cases with CH, silhouette and stability criterion.

Leukemia Data

The Leukemia data is obtained from Monti et. al. (2003) [123]. The data is composed by instances representing diagnosed samples of bone marrow from pediatric acute leukemia patients, corresponding to six prognostically important leukemia subtypes - 43 T-lineage ALL; 27 E2A-PBX1; 15 BCR-ABL; 79 TEL-AML1 and 20 MLL rearrangements; and 64 hyperdiploid > 50? chromosomes. There are 248 total patients and for each patient the number of attributes is 985.

We use hierarchical clustering method as the clustering algorithm for our method in this case. The performance of our method on this data set compared to other methods is given in table 5 -

We see here that BDQ identifies the correct number of clusters. Also, we see that when we consider the BDQ.potential method, it gives the most information about the clustering picture of the data set, since if we see the cluster membership, after the clustering, one of the classes is spuriously broken and two classes remain merged to form the 6 clusters for the hierarchical clustering method considered. So, when, we have seven clusters, we are actually having all the 6 classes plus a broken part of one class. So, when, BDQ.potential says that the data potentially also has 7 clusters, it gives insight into the data.

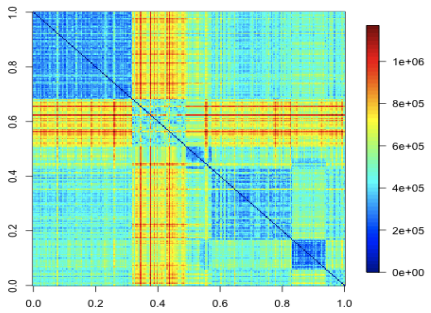


Figure 2.4: The distance matrix of Leukemia data with the classes arranged TEL-AML1, T-Lineage ALL, MLL, hyperdiploid, E2A-PBX1, BCR-ABL.

Table 2.5: Number of Clusters for Leukemia Data

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	0	0	0	0	1	0	0
CH	0	1	0	0	0	0	0	0
BDQ	0	0	0	0	0	1	0	0
BDQ.potential	0	0	0	0	0	1	1	0

Astronomy Data

The astronomy data is obtained from Richards et. al. (2011) [137]. The data is composed by instances representing light sources from sky surveys. The light sources are composed of 5 types of stars - 191 Classical Capheid, 145 Beta Lyrae, 114 Delta Scuti, 144 Mira, 58 W Ursae Majoris. There are 652 total light sources and for each light source the number of features is 64.

The performance of our method on this data set compared to other methods is given in table 6 -

Table 2.6: Number of Clusters for Astronomy Data

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	0	0	1	0	0	0	0
CH	0	0	1	0	0	0	0	0
BDQ	0	0	0	1	0	0	0	0
BDQ.potential	0	0	0	1	1	0	0	0

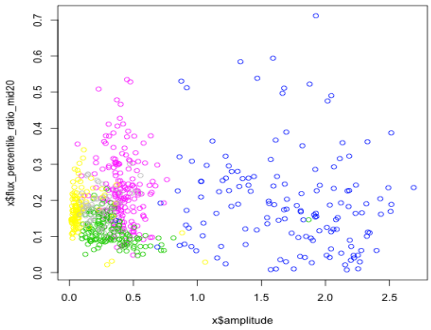


Figure 2.5: The astronomy data in two of its features.

So, BDQ.potential also performs good in this case and provides a nice insight to the data. Though number of clusters selected by BDQ is 4 in this case, we see that, $k = 5$ is one of the potential cluster numbers.

2.5 Discussion

So, we can see that methods of selecting cluster number by testing for block-diagonality of a matrix works nicely in practice. This method is highly general and can be applied in conjunction with any clustering method and any similarity (or distance) matrix. Also, the method can also provide a list of potential number of clusters, which is quite suggestive, since, a data set usually can be considered to have different number of clusters depending on the level of inspection, we are going to perform on the data set. Also, note that this a completely non-parametric approach, so it can be applied to quite general class of models

However, note that this method of selecting number of clusters is dependent on the performance of the clustering method itself. If the clustering method does not perform well, then, this method might produce unstable results.

Another issue is selecting the number of permutations. We have generally considered 1000 permutations to construct the 'null' distribution. However, it might be better to first sequentially test for the p-value 0.01, to see how many permutations are needed for the sequential rule to stop. Then, we can use a number of permutations slightly greater than the stopping number, to form the null distribution. Note that here, p-value is just an indicator, they do not have well-defined meaning.

Lastly, we have not derived any theoretical results for this method. However, assuming some underlying model space, we can try to prove the consistency and variance bounds of this method. Considering, gaussian model, we can easily see that, our procedure of finding location shift is a correct one. However, theoretically deriving the 'null' distribution is a challenge and we wish to address this issue later on.

Chapter 3

Stochastic Modeling of Networks

3.1 Introduction

Network is one of the most prominent way of representing relationships between different entities. These entities can be human or animals, living or non-living, real or imaginary and so on. And, due to the complex nature of the universe we live in, there are always relationships between such entities. A network is one way of representing that relationship and this way of representation has gained enormous acceptance among scientific community over years. As a result, we now see networks arising in all fields of physical and social sciences to represent relationships and interactions among entities.

Statistical study of networks is not a new field. Representing networks in form of graph structures with the fundamental components of *vertex* and *edge* can be traced back to Euler. However, from the later half of twentieth century, scientists have been increasingly interested in the empirical behavior of networks. One of the most well-known such empirical observation on networks is the *six-degree of separation* observation by Milgram [120] among human networks. It suggests that any person on earth is separated from another person by at most six people. This is also known as the '*small world phenomenon*'. Thus, study of networks give us a window to view the complex relationships between different entities in the real and imaginary world.

In this dissertation, we are concerned about statistical study of network data. When, relational data between entities are represented in form of network graphs, we shall try to infer about the different aspects of relational behavior by studying the empirical nature of the network data. Let us first see some examples of networks to understand the diverse situations of relational data, from which a network data can arise.

Examples of Networks

Network data arise in all fields of physical and social sciences to depict relationships between entities. With the information boom, a huge number of network data sets have come into prominence. In biology - gene transcription networks, protein-protein interaction network,

in social media - Facebook, Twitter, Linkedin networks, information networks arising in connection with text mining, technological networks such as the Internet, ecological and epidemiological networks and many others have come to the forefront. Here, we will give examples of a few well-known such networks.

Technological Networks

With the improvement of technology humans have created machines in a concerted and relational way. Thus creating a number of networks which connect entities which are human creations. Examples of such networks include Internet, Cell-phone tower and telephone exchange networks, Airport and Transport Networks. Among all these networks, Internet is probably the most prolific and interesting network. An example of Internet network data is given in Figure 3.1.

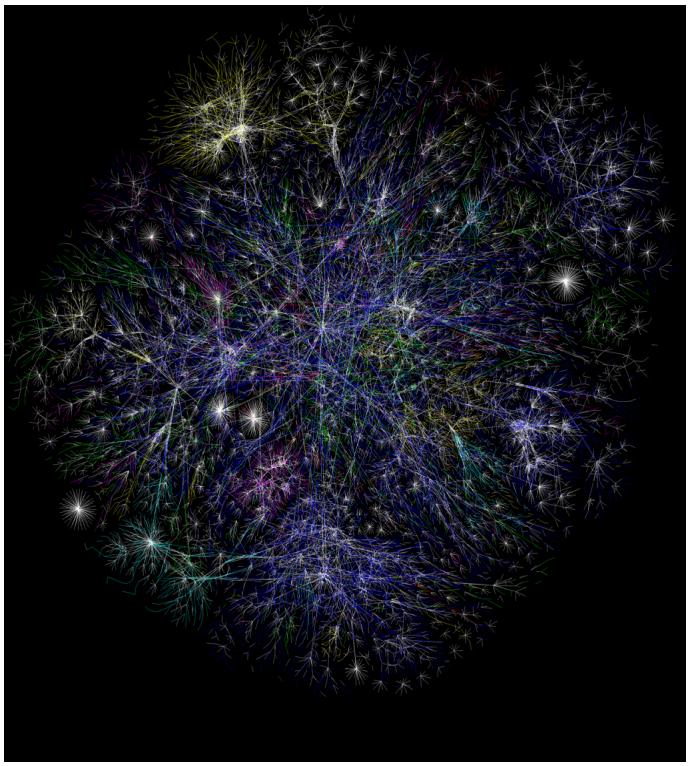


Figure 3.1: Internet Network from *www.opte.org*.

Social Networks

Social networks arise from interactions among human or any other social animals. This is one of highly studied form of network. Since, human relationships have always been an

enigma for scholars, study of human social networks have been highly popular among social scientists. With the advent of internet age, the number and complexity of human social networks have increased manifold. Examples of social network include social media networks such as Facebook, Twitter, LinkedIn and online gaming networks, academic networks such as collaboration and citation networks, networks arising from text-mining, networks arising from interaction of human or some other biological species. Figure 3.2 is a very famous social network which depicts friendship between members of a Karate club [165]. Figure 3.3 is Facebook friendship network in a college, Figure 3.4 is a romantic and sexual network among students in a high school, Figure 3.5 is a network of academic collaboration among High Energy Physics scientists.

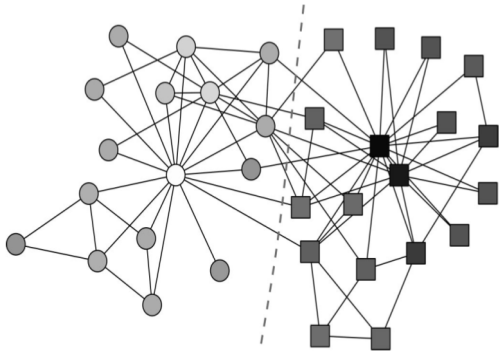


Figure 3.2: Karate Club (Newman, PNAS 2006)

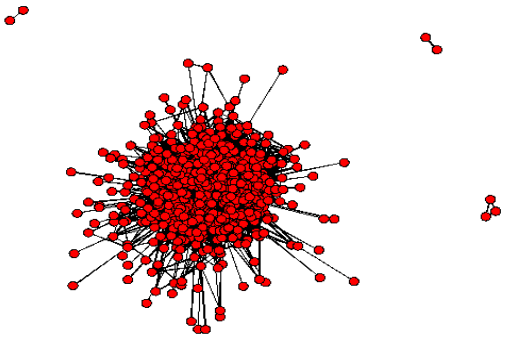


Figure 3.3: Facebook Network for Caltech with 769 nodes and average degree 43.

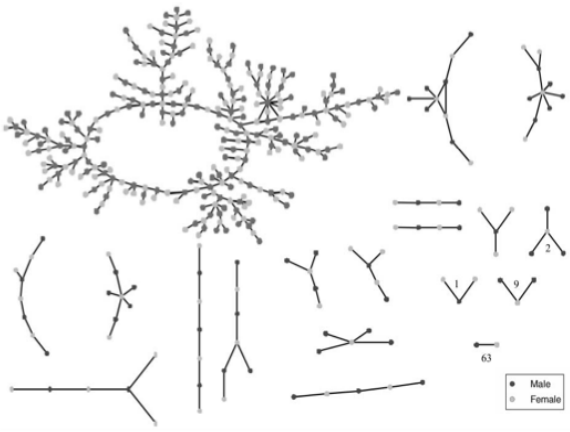


FIG. 2.—The direct relationship structure at Jefferson High

Figure 3.4: Network of romantic relationship between students of Jefferson High.

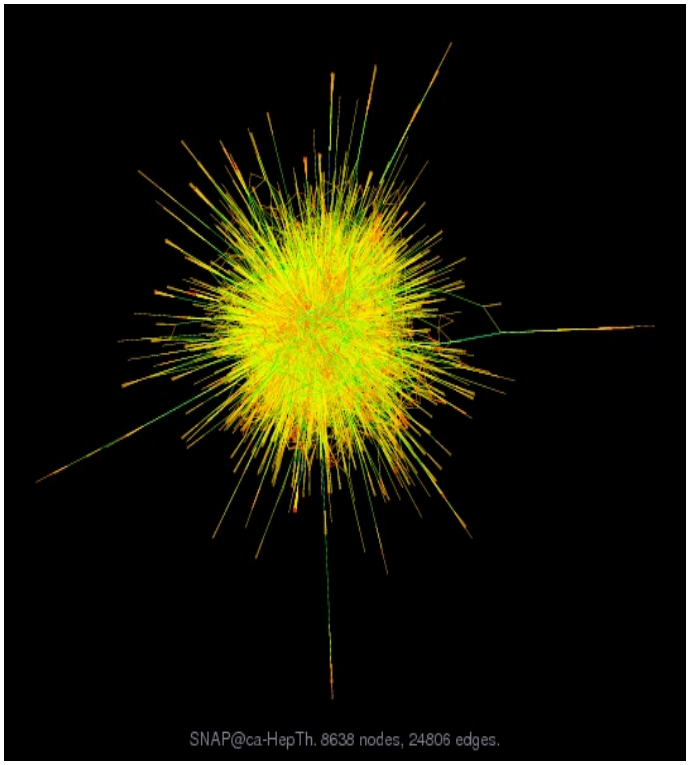


Figure 3.5: Collaboration Network in Arxiv for High Energy Physics with 8638 nodes and average degree 5.743.

Biological Network

Biological systems are one of the most complex systems known to mankind. Thus, study of relationships between biological entities through networks is highly important and relevant. Examples of such networks include Biochemical pathway networks, Protein-protein interaction networks, Gene transcription networks, Epidemiological Networks and so on. Each of these networks are important in their own right and gleaning knowledge about biological systems through analysis of these networks is an extremely significant endeavor. Figure 3.6 depicts a network of relationship between gene transcription factor of *E.Coli* bacteria and Figure 3.7 depicts a physical gene regulatory network.

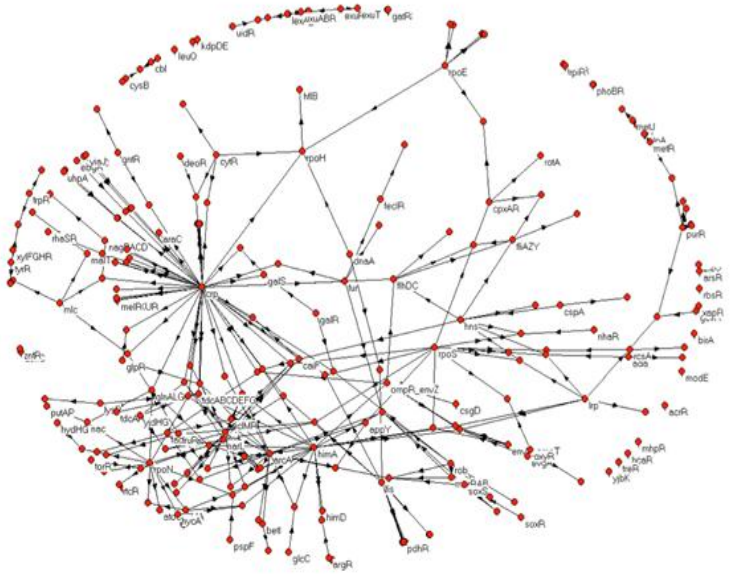


Figure 3.6: Transcription network of E. Coli with 423 nodes and 519 edges.

3.2 Research Questions on Networks

Statistical study of networks involve several important and interesting questions. The questions may be divided into two broad classes -

I. Given vectors of measurements $\boldsymbol{X_i}$ for each vertex, for example, given gene expression sequence nearby binding site information, physical (epigenetic information), protein assays etc., how do we decide to form edges that means causal or dependency relation between vertices (in the example, between genes).

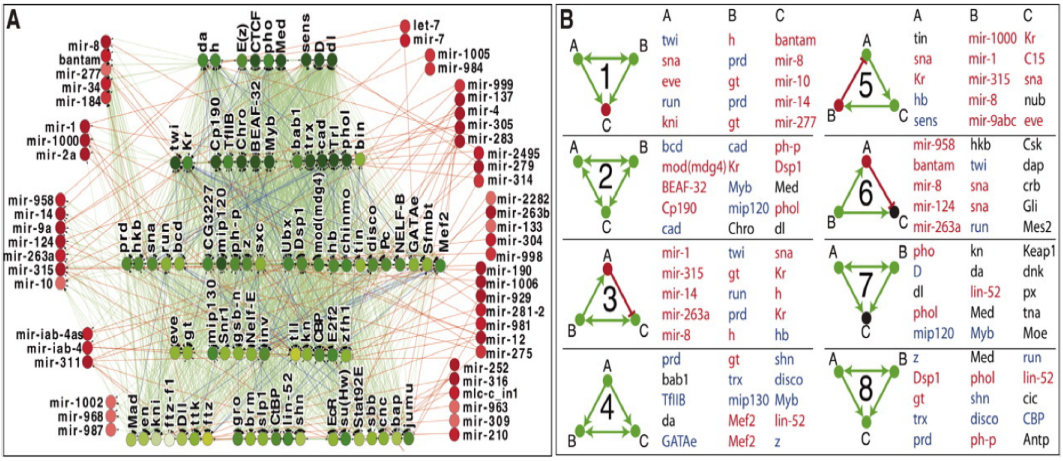


Figure 3.7: Physical Regulatory network (Science 2010; 330:1787-1797).

II. Given a network of relations (edges) identify higher level structures, clusters, like pathways in genomics.

The first broad class of problems is usually handled Gaussian or Markov Graphical Models and clustering. In Chapter 2, we addressed this issue for the general Elliptical Graphical Model case. But henceforth we shall mostly be interested in the second class of problems.

In the second class of problems, we already have a given network consisting of vertices and edges. Now, what are some of the important questions scientists wish to infer from these network data sets? Here, we list a subset of such questions.

(i) **Network Modeling:** Understanding the underlying general mechanism or model that is generating the network [132] [11]. Research in this area often focus on global network properties such as degree distribution, without addressing the semantics of individual edges. **We shall focus on this problem in Chapter 3.4**.

(ii) **Community Detection:** The task of finding hidden groups or communities in network based on the network topology is another common endeavor. Examples include protein complex finding in protein interaction network [2], detecting possible latent terrorist cells [13]. These applications often take the machine learning approach of graph partitioning. **We shall focus on this problem in Chapter 5**.

(v) **Sampling of nodes and subgraphs and descriptive statistics:** Descriptive statistics and their corresponding distribution help in summarizing and hypothesis testing on networks [122]. Motif finding, or more generally the search for subgraph patterns, also has many applications [14]. **We shall focus on this problem in Chapter 4**.

(iii) **Link Prediction:** In the machine learning community, network analysis often involves prediction [143], which can be edge related, e.g., predicting missing links in the network

[134], or attribute related, e.g., predicting how likely a movie is to be a box office hit [127]. Other applications include locating the crucial missing link in a business or a terrorist network, or calculating the probability that a customer will purchase a new product, given the pattern of purchases of his friends [78].

(iv) **Covariate or Latent Variable Estimation:** The related task of discovering the "roles" of individual nodes is useful for identity disambiguation and for business organization analysis [115].

(vi) **Information exchange:** The concept of information propagation also finds many applications in the network domain, such as virus propagation in computer networks [159], HIV infection networks [124] [87] and viral marketing [52]. Here the network structures are assumed to be known and the challenge is to find suitable models for disease spread. Theoretical works also give nice challenge here [126].

(vii) **Dynamic Network Behavior:** Studying dynamic behavior of network in terms of both modeling [10] and inference [144].

3.3 Stochastic Models of Networks

Let us consider that we have a random graph G_n as the data. Let $V(G_n) = \{v_i, \ldots, v_n\}$ denote the vertices of G_n and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of G_n. So, the number of vertices in G_n is $|V(G_n)| = n$ and number of edges of G_n is $|E(G_n)| = m$. Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. Nodes in the network which may represent individuals, organizations, or some other kind of unit of study. Edges correspond to types of links, relationships, or interactions between the units.

Parametric Models

We consider that the network graph is generated from an underlying probability model. When, the parameters of the probability model belong to finite-dimensional Euclidean space, we call them *parametric models*. We shall introduce here some of the well-known parametric models network data.

Erdös-Rényi Model

The mathematical biology literature of the 1950s contains a number of papers using what we now know as the network model $G(n, p)$, which for a network of n nodes sets the probability of an edge between each pair of nodes equal to p, independently of the other edges. But the formal properties of simple random graph network models are usually traced back to Gilbert [69], who examined $G(n, p)$, and to Erdös and Rényi [58]. The Erdös-Rényi random graph model, $G(n, N)$, describes an undirected graph involving n nodes and a fixed number of edges,

N, chosen randomly from the $\binom{n}{2}$ possible edges in the graph; an equivalent interpretation is that all $2^{\binom{n}{N}}$ graphs are equally likely. While the $G(n,p)$ model has a binomial likelihood where the probability of N edges is

$$p^N(1-p)^{\binom{n}{2}-N},$$

the likelihood of the $G(n,N)$ model is a hypergeometric distribution. The G(n,p) model is the more common one found in the modern literature on random graph theory, in part because the independence of edges simplifies analysis. Erdös and Rényi [59] went on to describe in detail the behavior of $G(n,N)$ as $p = N/\binom{n}{2}$ increased from 0 to 1. In the binomial version the key to asymptotic behavior is the value of $\lambda = pn$. One of the important Erdös-Rényi results is that there is a phase change associated with the value of $\lambda = 1$ and $\lambda = \log n$, with the emergence of a single giant connected component, while all the remaining components are relatively small and most of them take the form of trees [see 59; 70] and for the second phase the graph becomes asymptotically connected [16]. More formally,

- If $\lambda < 1$, then a graph in $G(n,p)$ will almost surely have no connected components of size larger than $O(\log n)$.
- If $\lambda = 1$, then a graph in $G(n,p)$ will almost surely have a largest component whose size is of $O(n^{2/3})$.
- If λ tends to a constant $c > 1$, then a graph in $G(n,p)$ will almost surely have a unique "giant" component containing a positive fraction of the nodes. No other component will contain more than $O(\log n)$ nodes.
- If $\liminf_{n\to\infty} \frac{\lambda}{\log n} = a$, where, $a > 1$, then a graph in $G(n,p)$ will be connected with high probability.

Stochastic Block Model

The stochastic block model is perhaps the most commonly used and best studied model for community detection. An SBM with K blocks states that each node belongs to a community $\boldsymbol{c} = (c_1, \ldots, c_n) \in \{1, \ldots, K\}$ which are drawn independently from the multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, where $\pi_i > 0$ for all i, and K is the number of communities, assumed known. Conditional on the labels, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with

$$\mathbb{E}[A_{ij}|\boldsymbol{c}] = \max\{\frac{P_{c_i c_j}}{n}, 1\}, \tag{3.1}$$

where $P = [P_{ab}]$ is a $K \times K$ symmetric matrix. The network is undirected, so $A_{ji} = A_{ij}$, and $A_{ii} = 0$ (no self-loops).

Preferential Attachment Model

Barabási and Albert [10] described a dynamic preferential attachment model specifically designed to generate scale-free networks. At time 0, the model starts out with n_0 unconnected nodes. At each sub- sequent time step, a new node is added with $m \leq n_0$ edges. The probability that the new node is connected to an existing node is proportional to the degree of the latter. In other words, the new node picks m nodes out of the existing network according to the multinomial distribution $p_i = \frac{\delta_i}{\sum_j \delta_j}$, where δ_i denotes the (undirected) degree of node i. This model is intended to describe networks that grow from a small nucleus of nodes and follow a "rich-get-richer" scheme. A new web page, for instance, will more likely link via a URL to a well-known web page as opposed to a little-known one.

The preferential attachment model of Barabási and Albert results in a network with a power law degree distribution empirically determined to have as its exponent ($\gamma_{BA} = 2.9 \pm 0.1$), whereas the Erdös-Rényi model has a Poisson degree distribution. Many extensions of the model have been proposed that allow for flexible power-law exponents, edge modifications, non-uniform dependence on the node degree distributions, etc. See Barabási et al. [11] and Durrett [54] for details. The generative process here could give an insight into the dynamics that led to the observed network. But data for the state of the network are typically gather at a small number of points in time (sometimes only once) and thus the network is only examined statically.

Exponential Random Graph Model

Under the assumption that two possible edges are dependent only if they share a common node, Frank and Strauss [66] devised the following characterization for the probability distribution of undirected Markov graphs:

$$\mathbb{P}_\theta(Y = y) = \exp\left(\sum_{k=1}^{n-1} \theta_k S_k(y) + \tau T(y) + \psi(\theta, \tau)\right) \quad y \in \mathcal{Y}, \tag{3.2}$$

where the statistics S_k and T count specific structures, such as edges, triangles, and k-stars: number of edges: number of k-stars ($k \geq 2$). $\theta \equiv \{\theta_k\}$ and τ are the parameters, and $\psi(\theta, \tau)$ is the normalizing constant. Note that there is a hierarchical structure to the parameters of this model, with edges being contained in 2-stars, and 2-stars being contained in both triangles and three-stars. There are also variations of the model involving directed edges.

The statistics $S_i(y)$ count graph structures. Although they are not independent, i.e., they count overlapping sets of edges, they are assumed independent in the pseudo-likelihood. Ignoring the correlations is a bad idea, and causes extreme sensitivity of the predicted number of edges to small changes in the value of certain parameters. Snijders et al. [145] recently proposed a variant of these models where the major problem of double-counting is mitigated, but not overcome. Hunter and Handcock [85] proposed an alternative estimation scheme that corrects parameter estimates for double-counting. This estimation procedure can be used for models based on distributions in the curved exponential family.

Robins et al. [139] describe problems associated with the estimation of parameters in many ERGMs, involving near degeneracies of the likelihood function and thus of methods used to estimate parameters using maximum likelihood. Bhamidi et. al. [17] point out to similar degeneracies. Chatterjee and Diaconis [39] provide some remedial measures.

There are two carefully constructed packages of routines that are available for analyzing network data using ERGMs: Statnet6 and SIENA7.

Nonparametric Model

Latent Space Model

The intuition at the core of latent space models is that each node $i \in V(G)$ can be represented as a point z_i in a "low dimensional" space, say $\mathbb{R}^k$. The existence of an edge in the adjacency matrix, $A_{ij} = 1$, is determined by the distance among the corresponding pair of nodes in the low dimensional space, $d(z_i, z_j)$, and by the values of a number of covariates measured on each node individually. The latent space model was first introduced by Hoff and Raftery [79] with applications to social network analysis, and has been recently extended in a number of directions to include treatment of transitivity, homophily on node-specific attributes, clustering, and heterogeneity of nodes [119; 109; 146].

Let A be an $n \times n$ adjacency matrix with binary entries A_{ij} denoting a relationship between nodes i and j. The probability model for A given in [79] is

$$\log \frac{\mathbb{P}(A_{ij} = 1)}{1 - \mathbb{P}(A_{ij} = 1)} = \alpha + \beta^T X_{ij} + |Z_i - Z_j| \equiv \eta_{ij}, \tag{3.3}$$

where $\boldsymbol{X}$ are covariates, Θ are parameters, and Z are the positions of the nodes in the latent space. Inference in latent space models has been carried out via Monte Carlo Markov chain.

Bickel-Chen Model

Consider any probability distribution $\mathbb{P}$ on an infinite undirected graph, or equivalently a probability distribution on the set of all matrices $||A_{ij} : i, j \geq 1||$ where $A_{ij} = 1$ or 0, $A_{ij} = A_{ji}$ for all i, j pairs and $A_{ii} = 0$ for all i, thus excluding self relation. If the graph is unlabeled, it is natural to restrict attention to $\mathbb{P}$ such that $||A_{\sigma_i \sigma_j}|| \sim \mathbb{P}$ for any permutation σ of $\{1, 2, 3, \ldots\}$. Hoover (see ref. 9) has shown that all such probability distributions can be represented as,

$$A_{ij} = g(\alpha, \xi_i, \xi_j, \lambda_{ij})$$

where $\sigma, \{\xi_i\}$ and $\{\lambda_{ij}\}$ are i.i.d. $U(0, 1)$ variables and $g(u, v, w, z) = g(u, w, v, z)$ for all u, v, w, z. The variables ξ correspond to latent variables, λ being completely individual specific, ξ generating relations between individuals and α a mixture variable which is unidentifiable even for an infinite graph. Note that g is unidentifiable and the ξ and λ could be put on another scale, e.g. Gaussian. It is clear that the distributions representable as,

$A = g(\xi_i, \xi_j, \lambda)$ where $\lambda_{ij} = \lambda_{ji}$, are the extreme points of this set and play the same role as sequences of i.i.d. variables play in de FinettiÕs theorem. Since given ξ_i and ξ_j, the λ_{ij} are i.i.d., these distributions are naturally parametrized by the function

$$h(u, v) = \mathbb{P}[A_{ij} = 1 | \xi_i = u, \xi_j = v]. \tag{3.4}$$

As Diaconis and Janson (13) point out $h(\cdot, \cdot)$ does not uniquely determine $\mathbb{P}$ but if h_1 and h_2 define the same $\mathbb{P}$, then there exists $\phi : [0, 1] \to [0, 1]$ which is measure preserving, i.e. such that $\phi(\xi_1)$ has a $U(0, 1)$ distribution and $h_1(u, v) = h_2(\phi(u), \phi(v))$.

Given any h corresponding to $\mathbb{P}$, let

$$\mathbb{P}[X_{ij} = 1 | \xi_i = u] = g(u) = \int_0^1 h(u, v)dv.$$

It is well known (see section 10 of ref. 14) that there exists a measure preserving ϕ such that, $g(\phi_g(v))$ is monotone non decreasing. Define

$$h_{CAN}(u, v) = h(\phi_g(u), \phi_g(v)) \tag{3.5}$$

$$g_{CAN}(u) = \int_0^1 h_{CAN}(u, v)dv = F^{-1}(u) \tag{3.6}$$

where F is the cdf of $g_{CAN}(\xi_i)$, and h_{CAN} is unique up to sets of measure 0. To see this note that if h corresponds to $\mathbb{P}$ and $g(u) = \int_0^1 h(u, v)dv$ is non decreasing, then since F is determined by $\mathbb{P}$ only, $g(u) = F^{-1}(u)$. But $g(\phi_g(u)) = g_{CAN}(u)$ and $\phi_g(u) = g^{-1}g_{CAN}(u) = u$. There is a reparametrization of h_{CAN} (we drop the CAN or *canonical* subscript in the future) which enables us to think of our model in terms more familiar to statisticians.

Let

$$\rho = \mathbb{P}(\text{Edge}) = \int_0^1 \int_0^1 h(u, v)dudv$$

Then the conditional density of (ξ_i, ξ_j) given that there is an edge between i and j is

$$w(u, v) = \rho^{-1}h(u, v). \tag{3.7}$$

This parametrization also permits us to decouple $\rho = \mathbb{E}(\text{Degree})$ of the graph from the inhomogeneity structure. It is natural finally to let ρ depend on n but $w(\cdot, \cdot)$ to be fixed. If $\lambda_n = \mathbb{E}(\text{Degree}) \to \infty$, we have what we may call the "dense graph" limit. If $\lambda_n = \Omega(1)$, we are in the case most studied in probability theory where, for instance, $\lambda_n = 1$ is the threshold at which the so called "giant component" appears.

So, the general non-parametric model can be described by the following equation -

$$\mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v) = h_n(u, v) = \rho_n w(u, v)\mathbf{1}(w \leq \rho_n^{-1}), \tag{3.8}$$

where, $w(u, v) \geq 0$, symmetric, $0 \leq u, v \leq 1$, $\rho_n \to 0$.

Inhomogeneous Random Graph Model

Let $\mathcal{S}$ be a separable metric space equipped with a Borel probability measure μ. For most cases $\mathcal{S} = (0, 1]$ with μ Lebesgue measure, that means a $U(0, 1)$ distribution. The "kernel" κ will be a symmetric non-negative function on $\mathcal{S} \times \mathcal{S}$. For each n we have a deterministic or random sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ of points in $\mathcal{S}$. Writing δ_x for the measure consisting of a point mass of weight 1 at x, and

$$\nu_n \equiv \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

for the empirical distribution of $\boldsymbol{x}$, it is assumed that ν_n converges in probability to μ as $n \to \infty$, with convergence in the usual space of probability measures on $\mathcal{S}$. One example where the convergence holds is the random case, where the x_i are independent and uniformly distributed on $\mathcal{S}$ with distribution μ convergence in probability holds by the law of large numbers.

For formal statements, the following definitions are made.

Definition 3.3.1. *A ground space is a pair $(\mathcal{S}, \mu)$, where $\mathcal{S}$ is a separable metric space and μ is a Borel probability measure on $\mathcal{S}$.*

Definition 3.3.2. *A vertex space $\mathcal{V}$ is a triple $(\mathcal{S}, \mu, (x_n)_{n\geq 1})$, where $(\mathcal{S}, \mu)$ is a ground space and, for each $n \geq 1$, $\boldsymbol{x}$ is a random sequence $(x_1, x_2, \ldots, x_n)$ of n points of $\mathcal{S}$, such that $\nu_n \xrightarrow{P} \mu$ holds.*

Of course, we do not need $(x_n)_{n\geq 1}$ to be defined for every n, but only for an infinite set of integers n.

Definition 3.3.3. *A kernel κ on a ground space $(\mathcal{S}, \mu)$ is a symmetric non-negative (Borel) measurable function on $\mathcal{S} \times \mathcal{S}$. By a kernel on a vertex space $(\mathcal{S}, \mu, (x_n)_{n\geq 1})$ we mean a kernel on $(\mathcal{S}, \mu)$.*

Let κ be a kernel on the vertex space $\mathcal{V}$. Given the (random) sequence $(x_1, \ldots, x_n)$, we let $G^{\mathcal{V}}(n, \kappa)$ be the random graph $G^{\mathcal{V}}(n, (p_{ij}))$ with

$$p_{ij} \equiv \min\{\kappa(x_i, x_j)/n, 1\}. \tag{3.9}$$

In other words, $G^{\mathcal{V}}(n, \kappa)$ has n vertices $\{1, \ldots, n\}$ and, given $x_1, \ldots, x_n$, an edge ij (with $i \neq j$) exists with probability p_{ij}, independently of all other (unordered) pairs ij.

The random graph $G(n, \kappa) = G^{\mathcal{V}}(n, \kappa)$ depends not only on κ but also on the choice of $x_1, \ldots, x_n$. The freedom of choice of x_i in this model is more than Bickel-Chen model. The asymptotic behavior of $G^{\mathcal{V}}(n, \kappa)$ depend very much on $\mathcal{S}$ and μ. Many of these key results such as existence of giant component, typical distance, phase transition properties are proved in [29]. In [29], we can see that many known parametric models of network are special cases of the inhomogeneous random graph model.

3.4 Inference on Network Models

Simple parametric models of networks are difficult to fit. We see that even for simple parametric models such as block models, the efficient estimation of the parameters is not easy [23]. But still many of the parametric models are not good enough representation of the naturally occurring graphs. The empirical and theoretical vulnerability of Exponential Random Graph Models have been pointed out by Chatterjee and Diaconis (2010) and Bhamidi et. al. (2008). Stochastic block models also have certain disadvantages like exponential degree distribution.

Nonparametric Inference on Bickel-Chen Model

We shall give a very simple non-parametric estimate of the kernel function $w(u,v)$ responsible for the link probability in Bickel-Chen model given in Section 3.3 for the dense case of networks when $\lambda_n \to \infty$.

By Theorem 1 of Bickel, Chen and Levina (2011) [23], as $\lambda \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tau(z_i) - \frac{D_i}{\bar{D}}\right)^2 = O\left(\frac{1}{\lambda}\right) \to 0 \tag{3.10}$$

here, $\tau(z) = T(\mathbf{1})(z)$.

Let

$$\hat{W}_n(u,v) \;=\; \int_0^u \int_0^v \frac{1}{n\bar{D}} \sum_{i,j} A_{ij}\mathbf{1}(\hat{\xi}_i \le s, \hat{\xi}_j \le t)dsdt$$

where $\hat{\xi}_i \equiv \hat{F}(\frac{D_i}{\bar{D}})$ and $\hat{F}$ is the empirical df of $\{\frac{D_i}{\bar{D}} : 1 \le i \le n\}$. Let

$$W_n(u,v) \;=\; \int_0^u \int_0^v \frac{1}{n\bar{D}} \sum_{i,j} A_{ij}\mathbf{1}(\xi_i \le s, \xi_j \le t)dsdt.$$

Degree-based Approach for Estimating w

a) Find smoothed empirical distribution function of $\frac{D_i}{\bar{D}}$,

$$\hat{F}(x) \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(\frac{D_i}{\bar{D}} \le x\right)$$

b) Divide $[0,1]$ into intervals $I_1, \ldots, I_M$, such that, $I_j = [\frac{j-1}{M}, \frac{j}{M})$,

$$\begin{aligned}\hat{w}(u,v) \;\equiv\; & \frac{1}{\bar{D}}\sum_{a,b=1}^{M}\frac{1}{n^*}\mathbf{1}(u \in I_a)\mathbf{1}(v \in I_b) \\ & \times\left[\sum_{i,j=1}^{n}\mathbf{1}\left\{A_{ij} \;:\; \hat{F}\left(\frac{D_i}{\bar{D}}\right) \in I_a,\; \hat{F}\left(\frac{D_j}{\bar{D}}\right) \in I_b\right\}\right]\end{aligned}$$

Figure 3.8: The LHS is estimate of h_{CAN} function for network of students of year 2008 and RHS is network of students of year 2008 residing in only 2 dorms. The proportions of classes in 2 distant modes are (0.3, 0.7) and (0.84, 0.16).

where, $n^* = |I_a||I_b|$, if, $a \neq b$ and $n^* = (|I_a|(|I_a|-1))/2$, if, $a = b$.

Application to Facebook Data

In this application, we try to quantitatively analyze the behavior of slink formation for Facebook collegiate networks. The networks were presented in the paper by Traud et.al. (2011) [155]. The network is formed by Facebook users acting as nodes and if two Facebook users are "friends" there is an edge between the corresponding nodes. Along with the network structure, we also have the data on covariates of the nodes. Each node has covariates: gender, class year, and data fields that represent (using anonymous numerical identifiers) high school, major, and dormitory residence. We try to find the probability of link formation given latent variables, that means estimating $w(\cdot,\cdot)$ function, defined in Eq. (3.7), for a part of the network.

However, as we have seen in discussions following Eq. (3.4), that either $w(u,v)$ or $h(u,v)$ are identifiable only unto a measure preserving transformation of the variables. So, we try to estimate $h_{CAN}(u,v)$ defined in Eq. (3.5) instead, which is identifiable and unique. For Facebook network, we try to estimate the h_{CAN} function for a part of the network. The subnetwork consists of distinct communities based on the dormitory affiliation of the students (vertices). However, since we are measuring h_{CAN}, the canonical h function, not the one associated with block model, we can not see the block structure properly for the 3-block case in Figure 3.9.

Figure 3.9: The LHS is estimate of h_{CAN} function for network of students of year 2008 residing in 3 dorms and RHS is sum of projections $\hat{h}_{CAN}(i,,i,)$ with two latent variables. The proportions of classes in 4 modes are (0.5, 0.13, 0.37), (0.67, 0.11, 0.22), (0.26, 0.66, 0.08), (0.32, 0.18, 0.5)

Chapter 4

Subsampling Bootstrap of Count Features of Networks

4.1 Introduction

The study of networks has received increased attention recently not only from the social sciences and statistics but also from physicists, computer scientists and mathematicians. With the information boom, a huge number of network data sets have come into prominence. In biology - gene transcription networks, protein-protein interaction network, in social media - Facebook, Twitter, Linkedin networks, information networks arising in connection with text mining, technological networks such as the Internet, ecological and epidemiological networks and many others have come to the forefront. Although the study of networks has a long history in physics and mathematics literature and informal methods of analysis have arisen in many fields of application, statistical inference on network models as opposed to descriptive statistics, empirical modeling and some Bayesian approaches [128] [98] [79] has not been addressed extensively in literature. A mathematical and systematic study of statistical inference on network models has only started in recent years.

Frequentist statistical inference involves proposing random models, fitting the proposed model to the data, checking goodness of fit in nonparametric context and given a good fit constructing tests and confidence statements about features of the model. Systematic analysis of complex models can only be done asymptotically and validated by simulation and network models are no exception. Much recent analysis has focussed on block models [80] and exponential random graph models (ERGM) [66]. The block models in their first incarnation did not fit large graphs well, for instance, their exponential degree distribution did not fit empirically observed degree distributions which often seemed to be of power law type [10] [132]. But they serve as shown by Bickel and Chen (2009) [22] the role of histograms. They also have until recently (Amini et. al. (2012) [5], Daudin et. al. (2008) [47]) proved hard to fit well. Nevertheless their analysis proceeds apace [125] [23] [140] [40]. For ERGM also there has been some work done on likelihood inference [145] [85] for these models. But

recently Bhamidi et al [17] has shown some issues with these models, as most of the time these models fail to represent real-world network properties. We will not dwell on these issues.

It follows from the work of Lovász [113] [30] and Aldous [4] and Hoover [82] that there is a representation of all probability models on n vertices which can be embedded in an infinite vertex model with natural invariance properties. This leads to the Bickel and Chen (2009) [22] characterization of "nonparametric" unlabeled graph models which is closely related to Lovász's notion of "graphons". It also follows from the work of Lovász [113], Diaconis and Janson [51] and in part from Bickel and Chen [22] that there is a unique set of statistics whose joint distribution characterize the probability distribution on graphs. These statistics, called "empirical moments" by Bickel, Chen and Levina [23], have appeared in various literatures earlier under the names of "motif" counts in biology [92], "subgraph" counts in probability [113]. Examples are the number of edges, the number of 'V's, the number of triangles contained in the observed graph.

The expectation and variances of the quantities can, in principle be computed (Picard et.al. [136]) and more usefully be asymptotically approximated [23] and under appropriate conditions these have limiting Gaussian distribution. They have many uses [161] [155] [14], particularly in distinguishing between the mechanisms generating different graphs as well as providing characterization, but in an outward form of the probability distribution.

A major stumbling block in their use has been the calculation of motifs that have more than 4 or 5 members. They have been used in testing equality of two distributions of count statistics and finding confidence intervals. Another problem that arises in dealing with the count statistics for large numbers is actually computing the count statistics. Finding the correct count statistics is a computationally hard problem for large networks as the complexity of finding the count of a subgraph is polynomial in terms of number of vertices and when number of vertices is in millions, the computation becomes infeasible.

In the statistical literature on networks, some work has been done on devising sampling designs to select network samples. Various sampling designs has been proposed at different points in the statistical literature to derive *meaningful* samples of a given network. Kolaczyk (2009) and [98] contains a nice summary of network sampling designs. Examples of such sampling designs include *random node selection, induced and incident sampling, star and snowball sampling, link-tracing sampling, random walks, forest fire* and several modifications of the stated methods [98] [108] [151]. Many of these sampling designs have been analyzed from design-based sampling point of view [152] [65]. There has also been work done on analyzing some of these methods from model-based sampling point of view, where, mostly the *exponential random graph model (ERGM)* was considered as the model generating the network and a likelihood-based approach was taken for inference [74]. As a result only parametric inference was possible in those approaches. On the other hand, our approach is not restricted to parametric models as we try to estimate the certain functionals of the underlying generating model, using the samples obtained from the random population network. So, in our work, we consider a "nonparametric" model as the underlying model in our analysis and try to see both theoretically and by examples how some of these sampling

schemes perform in estimating count features and their asymptotic variances.

Contribution and Structure of the Chapter

We use subsampling based bootstrap approaches to estimate the count statistics as well as find approximate distribution for such count statistics under the general model of Bickel and Chen [22]. We also state certain properties under which a network sampling design becomes *adaptive* to the network model for count statistics. By *adaptive*, we mean here that the network sampling design produces subsamples of network, such that the count statistics obtained from the subsampled network, becomes consistent to the original count statistics for the whole network.

We also apply our bootstrap method in simulated networks as well as two real-life networks. In simulation, we use two different models, stochastic block models [80] and preferential attachment models [10]. We try to compare the performance of different bootstrap methods for each of the simulated networks as well as to compare the networks generated by these two models using some well known descriptive statistics of networks [98]. One of the real-life networks is the Jefferson high-school network given in Bearman et.al. (2004) [14] and the others are the Facebook collegiate networks provided in Traud et.al. (2010) [155]. For the high-school network, we try to answer the question whether the number of small-cycles in the network is small. For the Facebook collegiate networks, we try to decide whether the node covariates given for the network have any potential clustering power. We also try to distinguish two different networks based on the partitioning properties. The test-statistics that we use in these comparisons is network transitivity, which has been argued to indicate network clustering capability [98].

In section 4.2 we outline our main results. In section 4.3 we describe the bootstrap subsampling methods and the theoretical properties of each bootstrap estimators. We also indicate a method for estimating asymptotic variances of these estimators using bootstrap. We also give a theoretical comparison of the methods. In section 4.5 we perform simulation under two special cases of the general "nonparametric" model: stochastic block model and preferential attachment model respectively. Under each of these cases we try to estimate count statistics and their variances with the the help of the bootstrap subsampling schemes and we compare the empirical performance of the three proposed bootstrap subsampling schemes as well as perform tests for model mis-specification. In 4.6 we apply our method to test hypotheses about the count statistics of the real network.

4.2 Main Results

Let us consider that a random graph G_n as the data. Let $V(G_n) = \{v_i, \ldots, v_n\}$ denote the vertices of G_n and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of G_n. So, the number of vertices in G_n is $|V(G_n)| = n$ and number of edges of G_n is $|E(G_n)| = m$. Let the adjacency matrix

of G_n be denoted by $A_{n\times n}$. For sake of notational simplicity, from here onwards we shall denote G_n by G having n vertices unless specifically mentioned.

We consider a general non-parametric model, as described in Bickel, Chen and Levina (2011) [23], generates the random data network G. The general non-parametric model can be described by the following equation -

$$\mathbb{P}(A_{ij} = 1|\xi_i = u, \xi_j = v) = h_n(u, v) = \rho_n w(u, v)\mathbf{1}(w \leq \rho_n^{-1}), \tag{4.1}$$

where, $w(u, v) \geq 0$, symmetric, $0 \leq u, v \leq 1$, $\rho_n \to 0$. This model assumes exchangeability

The graph statistics that we are concerned with, are count statistics of subgraphs. Let R be a subgraph of G, with $V(R) \subseteq V(G)$ and $E(R) \subseteq E(G)$. We have $|V(R)|= p$ and $E(R)|= e$. For notation, if two graphs R and S are equivalent, we denote them by $R \cong S$ and if R is a subgraph of S, we denote them by $R \subseteq S$. Now, the empirical statistic of our concern is

$$T_G(R) = \frac{1}{\binom{n}{p}|Iso(R)|} \sum_{S\subseteq K_n, S\cong R} \mathbf{1}(S \subseteq G) \tag{4.2}$$

where, $Iso(R)$ is the group of Isomorphisms of R and K_n is the complete graph on n vertices.

The population version of the sample statistic $T_G(R)$ can be defined as $P(R)$,

$$P(R) = \mathbb{E}\left\{ \prod_{(i,j)\in R} h(\xi_i, \xi_j) \prod_{(i,j)\in \bar{R}} h(\xi_i, \xi_j) \right\}$$

where, $\bar{R} = \{(i, j) \notin R, i \in V(G), j \in V(G)\}$. Evidently, we have

$$\mathbb{E}(T_G(R)) = P(R)$$

If we define normalized versions of parameter $P(R)$ as

$$\tilde{P}(R) = \rho^{-e} P(R)$$

where, $e \equiv |E(R)|$, then, we can define the corresponding normalized statistic to be

$$\hat{T}_G(R) = \hat{\rho}^{-e} T_G(R)$$

where,

$$\hat{\rho} = \frac{\bar{D}}{n-1}. \tag{4.3}$$

where, $D_i =$ degree of v_i, $v_i \in V(G_n)$ for $i = 1, \ldots, n$ and $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$.

We wish to approximate the functionals $\mathbb{E}(T_G(R))$ and $\mathrm{Var}(T_G(R))$ by nonparametric bootstrap. Let us consider the bootstrap estimate of $\hat{T}_G(R)$ to be $\hat{T}_b(R)$ and bootstrap estimate of $\mathrm{Var}(T_G(R))$ to be $\hat{\sigma}_b^2(R)$. We consider b as the bootstrap repetition or resampling parameter. How, we get the bootstrap estimates will be discussed in next section. But, for such a bootstrap estimates, we can state the general theorem that we proved -

Theorem 4.2.1. *Suppose R is fixed, acyclic with $|V(R)|= p$ and $\int_0^\infty \int_0^\infty w^{2|R|}(u,v)dudv < \infty$. Then if $\lambda_n \to \infty$ and $b \to \infty$*

$$\hat{T}_b(R) \xrightarrow{P} \tilde{P}(R) \tag{4.4}$$

$$\sqrt{n}\left(\frac{\hat{T}_b(R) - \tilde{P}(R)}{\hat{\sigma}_b(R)}\right) \xrightarrow{w} N(0,1) \tag{4.5}$$

If for fixed, acyclic subgraphs $(R_1, \ldots, R_k)$, we define, $\mathbf{T}_b(\mathbf{R}) = \left(\hat{T}_b(R_1), \ldots, \hat{T}_b(R_k)\right)$ and $\mathbf{P}(\mathbf{R}) = \left(\tilde{P}(R_1), \ldots, \tilde{P}(R_k)\right)$

$$\sqrt{n}\left((\mathbf{T}_b(\mathbf{R}) - \mathbf{P}(\mathbf{R}))^T \hat{\Sigma}_b^{-1/2}(\mathbf{R}) (\mathbf{T}_b(\mathbf{R}) - \mathbf{P}(\mathbf{R}))\right) \xrightarrow{w} N(\mathbf{0}, \mathbf{I}) \tag{4.6}$$

where, $[\Sigma_b]_{st} = \hat{\sigma}_b(R_s, R_t)$, $s,t = 1, \ldots, k$ and if $R_s = R_t = R$, $\sigma_b(R_s, R_t) = \hat{\sigma}_b^2(R)$. These results also hold for subgraphs R, which are k-cycles.

Note that the above theorem is the *master theorem*. The proof of this theorem depends upon how we obtain the bootstrap estimates $\hat{T}_b(R)$ and $\hat{\sigma}_b^2(R)$. We consider two bootstrap procedures.

(I) *uniform subsampling* bootstrap procedure

(II) *sampling-based* bootstrap procedure.

Both bootstrap procedures have different small-sample behavior. That means depending on λ_n and size of R for a given network G, the efficiency of the bootstrap estimates differ.

We also considered another bootstrap procedure, which was a variant of the common *snowball sampling*. However, we do not discuss that method in the main discourse, instead, we relegate discussion on that method to the Appendix, since the method performs poorly for all types of graphs compared to the other two methods, both theoretically and empirically.

For each of the bootstrap methods, we prove a theorem of following type -

Theorem 4.2.2. *Suppose R is fixed, acyclic with $|V(R)|= p$, then, if $b \to \infty$,*

$$\sqrt{n}\left(\hat{T}_b(R) - \hat{T}_G(R)\right) \xrightarrow{P} 0 \tag{4.7}$$

Also, if $n \to \infty$ and $\lambda_n \to \infty$ and under certain conditions depending on the bootstrap method

$$\frac{\hat{\sigma}_b^2(R)}{\sigma^2(R)} \xrightarrow{P} 1 \tag{4.8}$$

where, $\sigma^2(R)$ is the asymptotic variance of $\hat{T}_G(R)$ as defined in Theorem 1 of [23]. These results also hold for subgraphs R, which are k-cycles.

We shall prove Theorem 4.2.2 for each of the two bootstrap cases in Section 4.3. Then, we shall use the Theorem 4.2.2 to prove the general theorem 4.2.1 in Section 4.4.

Bootstrap and Model-based Sampling

Our work can be viewed from two different perspectives. The first perspective is that of *bootstrap*. In *non-parametric bootstrap*, we use resamples or subsamples of the data, where the data comes from an unknown distribution, to find the functionals of the unknown distribution. In our situation also, we have a network that has been generated from an underlying probability model. We want to *subsample* networks from our given network and use those subsampled networks to find estimates of functionals of the underlying population model generating the given network. Note that here we are interested in *subsampling* not *resampling* a network. This is precisely because one of our goals is to reduce the burden of performing computation on the *large* original network.

The second perspective is that of *model-based sampling*. In model-based sampling, we consider that the *population*, from which the sample is selected according to some sampling design, is a realization of a probabilistic event. So, in our case, we consider the given network as the population and it is generated from an underlying probability model.

4.3 Bootstrap Methods

We propose three different bootstrap methods. Each of them are subsampling approaches of bootstrap. In the following subsections, we shall define each of these subsampling bootstrap methods. We shall also compare the theoretical performance between the three bootstrap schemes.

Let us consider that we have a random graph G_n as the data with $|V(G_n)|= n$ and number of edges of G_n is $|E(G_n)|= m$. Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. For sake of notational simplicity, from here onwards we shall denote G_n by G having n vertices unless specifically mentioned. Let R be a subgraph of G, with $V(R) \subseteq V(G)$ and $E(R) \subseteq E(G)$. We have $|V(R)|= p$ and $|E(R)|= e$.

Uniform Subsampling Bootstrap

In the *uniform subsampling* bootstrap scheme at each bootstrap iteration a subset of vertices of the full network G is selected without replacement and the graph induced by the selected subset is the subsample we consider. This is a vertex subsampling or induced network sampling scheme. The full bootstrap procedure given the subsample size, m and number of bootstrap iterates, B, is as follows –

1. For b^{th} iterate of the bootstrap, $b = 1, \ldots, B$,
2. Choose m vertices without replacement from $V(G)$ and form the induced subgraph of G based on the selected vertices. Denote the graph formed by H.

3. Calculate $T_{b1}(R)$, given by formula

$$T_{b1}(R) = \frac{1}{\binom{m}{p}|Iso(R)|} \sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H) \tag{4.9}$$

The bootstrap estimate of $T_G(R)$ is given by

$$\tilde{T}_{b1}(R) = \frac{1}{B} \sum_{b=1}^{B} T_{b1}(R) \tag{4.10}$$

The uniform subsampling bootstrap scheme is the network version of the common subsampling bootstrap scheme seen in Bickel et. al. [20]. Note that, there are other ways of forming uniformly subsampled bootstrap estimates as mentioned in [20], however, we just mention one of them in this discourse. The properties of the bootstrap estimator $\tilde{T}_{b1}(R)$ is given is Lemma 4.3.1

Lemma 4.3.1. *The estimator $\tilde{T}_{b1}(R)$ has the following properties*

(i) Given G, $\tilde{T}_{b1}(R)$ is an unbiased estimate of $T_G(R)$.

(ii) As $B \to \infty$, $n \to \infty$, $m \to \infty$ and $m/n \to 0$,

$$\sqrt{n}(\rho^{-e}\tilde{T}_{b1}(R) - \rho^{-e}T_G(R)) \xrightarrow{P} 0$$

Proof. (i) Now, let us try to try to find the expectation of $T_{b1}(R)$ under the sampling distribution conditional on the given data G.

$$\begin{aligned}
&\mathbb{E}_b\left[\frac{1}{\binom{m}{p}|Iso(R)|} \sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H)\Big|G\right] \\
&= \frac{1}{\binom{m}{p}|Iso(R)|}\mathbb{E}\left[\sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H)\Big|G\right] \\
&= \frac{1}{\binom{m}{p}|Iso(R)|} \sum_{H \subseteq G, |H|=m} \frac{1}{\binom{n}{m}} \sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H) \\
&= \frac{1}{\binom{m}{p}|Iso(R)|} \sum_{\substack{S \subseteq K_n \\ S \cong R}} \sum_{\substack{H \supseteq S, H \subseteq G \\ |H|=m}} \frac{1}{\binom{n}{m}} \mathbf{1}(S \subseteq G) \\
&= \frac{1}{\binom{m}{p}|Iso(R)|} \sum_{S \subseteq K_n, S \cong R} \frac{\binom{n-p}{m-p}}{\binom{n}{m}} \mathbf{1}(S \subseteq G) \\
&= \frac{1}{\binom{n}{p}|Iso(R)|} \sum_{S \subseteq K_n, S \cong R} \mathbf{1}(S \subseteq G)
\end{aligned}$$

So, we have,

$$\mathbb{E}_b[\tilde{T}_{b1}(R)|G] = \mathbb{E}_b[T_{b1}(R)|G] = T_G(R)$$

(ii) Here, we use properties of the underlying model. Let us condition on $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_n\}$ and the whole graph G separately. Now, conditioning on $\boldsymbol{\xi}$, we get the main term of $T_G(R)$ to be,

$$\mathbb{E}(\hat{P}(R)|\boldsymbol{\xi}) = \frac{1}{\binom{n}{p}|Iso(R)|} \sum_{S \subseteq K_n, S \cong R} \left(\prod_{(i,j) \in E(S)} w(\xi_i, \xi_j) \right) + O(n^{-1}\lambda_n). \tag{4.11}$$

We shall use the same decomposition as used in [23] of $(\rho_n^{-e}\tilde{T}_{b1}(R) - \tilde{P}(R))$ into

$$\begin{aligned}(\rho_n^{-e}\tilde{T}_{b1}(R) - \tilde{P}(R)) &= \rho_n^{-e}\left(\tilde{T}_{b1} - \mathbb{E}_b[T_{b1}(R)|G]\right) \\ &\quad + \rho_n^{-e}\left(T_G(R) - \mathbb{E}(T_G(R)|\boldsymbol{\xi})\right) \\ &\quad + \mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-e} - \tilde{P}(R)\end{aligned}$$

Let us define,

$$\begin{aligned}U_3 &= \mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-e} - \tilde{P}(R) \\ U_2 &= \rho_n^{-e}\left(T_G(R) - \mathbb{E}(T_G(R)|\boldsymbol{\xi})\right) \\ U_1 &= \rho_n^{-e}\left(\tilde{T}_{b1} - \mathbb{E}_b[T_{b1}(R)|G]\right)\end{aligned}$$

Now, it is easy to see that

$$\begin{aligned}\mathrm{Var}(\rho^{-e}\tilde{T}_{b1}(R)) &= \mathbb{E}(\mathrm{Var}(\rho^{-e}\tilde{T}_{b1}(R)|G) + \mathrm{Var}(\mathbb{E}(\rho^{-e}\tilde{T}_{b1}(R)|G)) \\ &= \mathbb{E}(\mathrm{Var}(\rho^{-e}\tilde{T}_{b1}(R) - T_G(R)|G) + \mathrm{Var}(T_G(R)) \\ &= \mathbb{E}(\mathrm{Var}(U_1|G)) + \mathbb{E}(\mathrm{Var}(T_G(R)|\boldsymbol{\xi})) + \mathrm{Var}(\mathbb{E}(T_G(R)|\boldsymbol{\xi})) \\ &= \mathbb{E}(\mathrm{Var}(U_1|G)) + \mathbb{E}(\mathrm{Var}(U_2|\boldsymbol{\xi})) + \mathrm{Var}(U_3)\end{aligned}$$

We shall try to see the behavior of $\mathrm{Var}(U_1|G) = \mathrm{Var}_b[\rho^{-e}\tilde{T}_{b1}(R)|G]$. Now,

$$\mathrm{Var}_b[\rho^{-e}\tilde{T}_{b1}(R)|G] = \rho^{-2e}\frac{1}{B^2}\left(\sum_{b=1}^{B}\mathrm{Var}_b[T_{b1}(R)] + \sum_{b,b'=1,b\neq b'}^{B}\mathrm{Cov}_b(T_{b1}(R), T_{b'1}(R))\right)$$

Now, the formula for $\mathrm{Var}_b[T_{b1}(R)] = O(\frac{1}{m})$ and $\mathrm{Cov}_b[T_{b1}(R), T_{b'1}(R)] = O(\frac{1}{m})$ for acyclic and k-cycle R is given in Appendix A1. If we consider the uniform probability for bootstrap to be γ, then, $B = O(\gamma n^p)$. Note that, if $E(H_b) \cap E(H_{b'}) = \phi$, then, $\mathrm{Cov}_b(T_{b1}(R), T_{b'1}(R)) = 0$. The number of pairs such that $E(H_b) \cap E(H_{b'}) \neq \phi$

is $O(m^2\gamma^2 n^{2m-2})$. Also, the number of edges for the leading term in the covariance is equal to or more than $2e$. So,

$$\mathbb{E}(\text{Var}(U_1|G)) = O\left(\frac{m^2\gamma^2 n^{2m-2}}{m\gamma^2 n^{2m}}\right) = O\left(\frac{m}{n^2}\right) = o(n^{-1})$$

The last equality follows since we have $m/n \to 0$ as $n \to \infty$.

Now, by proof of Theorem 1 in [23], we have,

$$\begin{aligned} \text{Var}(U_2) &= o(n^{-1}) \\ \text{Var}(U_3) &= o(n^{-1}) \end{aligned}$$

So, we get, $\text{Var}(\rho^{-e}\tilde{T}_{b1}(R)) = o(n^{-1})$. Since, we already know $\sqrt{n}$-consistency of $\left(\rho_n^{-e}T_G(R) - \tilde{P}(R))\right)$, this proves the $\sqrt{n}$-consistency of $\rho_n^{-e}\tilde{T}_{b1}(R)$ to $\rho_n^{-e}T_G(R)$.

□

The variance of $\tilde{T}_{b1}(R)$ given G can also be calculated and is given in the Appendix A1.

Sampling based bootstrap

In this bootstrap scheme we use an enumeration scheme of finding all possible subgraph R in the graph G and convert the enumeration scheme into a sampling scheme. The enumeration scheme was proposed by Wernicke et. al. (2006) [162]. A random version of the enumeration scheme was also proposed in the paper [162]. We use the random version of the enumeration scheme to form our sampling scheme.

Let us first discuss the enumeration scheme of Wernicke et al [162], which we shall henceforth call EnumerateSubgraph or `ESU`. The enumeration algorithm is a *breadth-first search* algorithm. The algorithm strives to create a forest of tree structures such that each leaf of each tree is a size-p subgraph (we have, $|R| = p$). Since, the counting scheme follows a breadth-first search route, before performing the `ESU` algorithm, we need an ordering of the vertices based on breadth-first search of the graph starting from any particular vertex (say v). So, we get a particular fixed ordering of the vertices of the network with v getting lowest order value and subsequently searched vertices getting higher order values. The ordering is described in the algorithm AssignOrder or `AO`. Based on that ordering of G we perform `ESU`.

When counting, `ESU` algorithm creates a forest of tree structures such that each tree represents one vertex of the network and each leaf of each tree is a size-p subgraph (we have, $|R| = p$). We start with an available vertex of lowest possible order, say u. We construct a tree with the vertex u as the root node. We consider u as he "parent" node and neighbors of u, which have a higher order than u as its "children". In the next step, the "children" node become the "parent" node in the tree and has its own neighbors as their "children". The tree is allowed to grow unto a height p, if we are counting size-p subgraphs. So, we can see

that each leaf of the tree represents a collection of p nodes coming from the path connecting the leaf to the root. For each vertex, we have such a tree and over counting is averted as we maintain the order while forming the trees. So, with the help of the particular ordering of vertices, each of the size-p subgraphs ($|R|= p$) is counted only once.

Algorithm 4.3.1 AssignOrder(G, p)

Require: A graph $G = (V, E)$, where, $|V(G)|= n$.
Ensure: A vector $\sigma = (\sigma(1), \ldots, \sigma(n))$, where, σ is some permutation of $\{1, \ldots, n\}$ and $\sigma(i)$ is associated with vertex $v_{\sigma(i)} \in V(G)$ for all $i = 1, \ldots, n$.

$\sigma_1 \leftarrow 1$
$\mathcal{V} \leftarrow \{v_1\}$
$i \leftarrow 1$
while $|\mathcal{V}|< n$ **do**
    Denote $k \leftarrow |N(\mathcal{V})|$ and $\{v_{h_1}, \ldots, v_{h_k}\} = N(\mathcal{V})$
    Define $\sigma(i + j) \leftarrow h_j$ for $j = 1, \ldots, k$.
    $i \leftarrow i + k$
    $\mathcal{V} \leftarrow \mathcal{V} \cup N(\mathcal{V})$
end while

Once, we have the ordering σ for G, **we define,** $\mathbf{v_i \succ v_j}$ if $\boldsymbol{\sigma^{-1}(i) > \sigma^{-1}(j)}$., where, $v_i, v_j \in V(G)$ and $i, j = 1, \ldots, n$ with $i \neq j$. This ordering is needed for success of the `ESU` algorithm and its randomized counterpart 4.3.4. We shall only formally state the randomized version of the algorithm, `RAND-ESU` 4.3.4 in this paper. The enumeration version can be found in [162].

Algorithm 4.3.2 EnumerateSubgraph(G, p)

Require: A graph $G = (V, E)$ and an integer p, where, $1 \leq p \leq |V|$.
Ensure: $\mathcal{S}_p$ = All subgraphs, R of G, such that $|R|= p$.

for each vertex $v \in V$ **do**
    $V_{Extension} \leftarrow \{u \in N(\{v\}) : u \succ v\}$
    Call ExtendSubgraph$(\{v\}, V_{Extension}, v\})$
end for

In Theorem 2 of [162] it was proved that the output of `ESU` algorithm, $\mathcal{S}_p$ contains all subgraphs R of G, such that $|R|= p$, exactly once. So, we can write the statistic (4.2) for a specific subgraph R with $|R|= p$ in the following way

$$T_G(R) = \frac{1}{\binom{n}{p}|Iso(R)|} \sum_{S \in \mathcal{S}_p} \mathbf{1}(S \cong R) \tag{4.12}$$

Essentially we have a normalized *population total* in terms of sampling theory. Our goal is to form a sampling design and devise a corresponding sampling estimator of $T_G(R)$ given G. To meet that goal we use a sampling version of the enumeration scheme `ESU`.

Algorithm 4.3.3 ExtendSubgraph($V_{Subgraph}, V_{Extension}, v$)

Require: Graphs $V_{Subgraph}, V_{Extension}$ and vertex v.
Ensure: All subgraphs, R of G, such that $|R|= p$ and v is a vertex of R.
1: **if** $|V_{Subgraph}|= p$ **then**
2: **return** Subgraph of G induced by $V_{Subgraph}$
3: **else**
4: **while** $V_{Extension} \neq \phi$ **do**
5: Remove an arbitrarily chosen vertex w from $V_{Extension}$
6: $V'_{Extension} \leftarrow V_{Extension} \cup \{u \in N_{excl}(w, V_{Subgraph}) : u \succ v\}$
7: **Call** ExtendSubgraph($V_{Subgraph} \cup \{w\}, V'_{Extension}, v$)
8: **end while**
9: **end ifreturn**

Now, the sampling version of the `ESU` algorithm has an extra set of parameters $(q_1, \ldots, q_p)$. We shall call the new algorithm as RandomizedEnumerateSubgraph or `RAND-ESU`.

Algorithm 4.3.4 RandomizedEnumerateSubgraph(G, p)

Require: A graph $G = (V, E)$, an integer p and an vector $(q_1, \ldots, q_p)$, where, $1 \leq p \leq |V|$ and $q_d \leq 1$ for all $d = 1, \ldots, p$.
Ensure: $\mathcal{S}_p^R$ = A sample of subgraphs, R of G, such that $|R|= p$.
1: **for** each vertex $v \in V$ **do**
2: $V_{Extension} \leftarrow \{u \in N(\{v\}) : u \succ v\}$
3: **With probability** q_1 **Call** RandExtendSubgraph($\{v\}, V_{Extension}, v\}$)
4: **end for**

From the sampling scheme `RAND-ESU` we have a sample $\mathcal{S}_p^R$ of $size - p$ subgraphs of G. Now, if we consider each item to be one $size - p$ subgraph of G, that is, an element of $\mathcal{S}_p$, then, we can try to calculate the *inclusion probability* of each item in the sample $\mathcal{S}_p^R$.

An item, $S \in \mathcal{S}_p$ is a subgraph of G induced by the set of vertices $\{w_1, \ldots, w_p\}$, where, we take that $w_{i+1} \succ w_i$, $i = 1, \ldots, p-1$. So,

$$\begin{aligned}
\pi \equiv \text{Inclusion Probability of } S &= \mathbb{P}[(w_1, \ldots, w_p) \text{ is selected}] \\
&= \mathbb{P}[w_p | (w_1, \ldots, w_{p-1}) \text{ is selected}]. \\
&\quad \mathbb{P}[(w_1, \ldots, w_{p-1}) \text{ is selected}] \\
&= q_p \cdot \mathbb{P}[(w_1, \ldots, w_{p-1}) \text{ is selected}] \\
&= q_p \cdot q_{p-1} \cdot \mathbb{P}[(w_1, \ldots, w_{p-2}) \text{ is selected}] \\
&= \cdots = q_p \cdot q_{p-1} \cdots q_1 = \prod_{d=1}^{p} q_d
\end{aligned}$$

Algorithm 4.3.5 RandExtendSubgraph($V_{Subgraph}, V_{Extension}, v$)

Require: Graphs $V_{Subgraph}, V_{Extension}$ and vertex v.
Ensure: A sample of subgraphs, R of G, such that $|R|= p$ and v is a vertex of R.

```
if |V_Subgraph|= p then
    return Subgraph of G induced by V_Subgraph
else
    while V_Extension ≠ φ do
        Remove an arbitrarily chosen vertex w from V_Extension
        V'_Extension ← V_Extension ∪ {u ∈ N_excl(w, V_Subgraph) : u ≻ v}
        d ← |V_Subgraph|+1
        With probability q_d Call RandExtendSubgraph(V_Subgraph ∪ {w}, V'_Extension, v)
    end while
end if
return
```

So, each item $S \in \mathcal{S}_p$ has an inclusion probability π to be in the sample $\mathcal{S}_p^R$. So, we can define a *Horvitz-Thompson* estimator of $T_G(R)$ based on $\mathcal{S}_p^R$ as

$$\tilde{T}_{b2}(R) = \frac{1}{(\prod_{d=1}^{p} q_d)\binom{n}{p}|Iso(R)|} \sum_{S \in \mathcal{S}_p^R} \mathbf{1}(S \cong R) \tag{4.13}$$

For variance calculation, we also need the joint inclusion probability of two items, $S, S' \in \mathcal{S}_p$, which are subgraphs of G induced by the set of vertices $\{w_1, \ldots, w_p\}$ and $\{w'_1, \ldots, w'_p\}$ respectively, where, we take that $w_{i+1} \succ w_i$ and $w'_{i+1} \succ w'_i$, $i = 1, \ldots, p-1$. So,

$$\begin{aligned}\pi_{SS'} \equiv \text{Inclusion Probability of } S \text{ and } S' \\ = \mathbb{P}[(w_1, \ldots, w_p) \text{ is selected} \& (w'_1, \ldots, w'_p) \text{ is selected}] \\ = \prod_{d=1}^{p} (q_d)^{z_{1d}} \prod_{d=1}^{p} \left(q_d^2\right)^{z_{2d}}\end{aligned}$$

where,

$$z_{1d} = \begin{cases} \mathbf{1}(w_d = w'_d), & \text{for } d = 1 \\ \mathbf{1}((w_d, w_{d-1}) = (w'_d, w'_{d-1})), & \text{for } d = 2, \ldots, p \end{cases}$$

$$z_{2d} = \begin{cases} \mathbf{1}(w_d \neq w'_d), & \text{for } d = 1 \\ \mathbf{1}((w_d, w_{d-1}) \neq (w'_d, w'_{d-1})), & \text{for } d = 2, \ldots, p \end{cases}$$

Now, let us try to see the properties of the bootstrap estimator $\tilde{T}_{b2}(R)$ -

Lemma 4.3.2. *The estimator $\tilde{T}_{b2}(R)$ has the following properties*

(i) Given G, $\tilde{T}_{b2}(R)$ is an unbiased estimate of $T_G(R)$.

(ii) As $B \to \infty$, $q_1 = 1$ or $q_1 \to 1$ and $q_d \to 0$ and $nq_d \to \infty$ for $d = 2, \ldots, p$ and $n \to \infty$,

$$\sqrt{n}(\rho^{-e}\tilde{T}_{b2}(R) - \rho^{-e}T_G(R)) \xrightarrow{P} 0$$

Proof. (i) according to the sampling theory [151], we have that $\tilde{T}_{b2}(R)$ is an unbiased estimator of $T_G(R)$ given the network G.

(ii) The variance of $\tilde{T}_{b2}(R)$ coming from the bootstrap sampling only is given by

$$\mathrm{Var}_b\left[\tilde{T}_{b2}(R)\right] = \frac{1}{N^2}\left[\frac{1-\pi}{\pi}\sum_{S\in\mathcal{S}_p}\mathbf{1}(S \cong R) + \sum_{S,S'\in\mathcal{S}_p, S\neq S'}\frac{\pi_{SS'}-\pi^2}{\pi^2}\mathbf{1}(S \cong R,\ S' \cong R)\right] \tag{4.14}$$

where,

$$N = \binom{n}{p}|Iso(R)|$$

We shall use the same decomposition as used in [23] of $(\rho_n^{-e}\tilde{T}_{b2}(R) - \tilde{P}(R))$ into

$$\begin{aligned}(\rho_n^{-e}\tilde{T}_{b2}(R) - \tilde{P}(R)) &= \rho_n^{-e}\left(\tilde{T}_{b2} - \mathbb{E}_b[T_{b1}(R)|G]\right) \\ &\quad +\rho_n^{-e}\left(T_G(R) - \mathbb{E}(T_G(R)|\boldsymbol{\xi})\right) \\ &\quad +\mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-e} - \tilde{P}(R)\end{aligned}$$

Let us define,

$$\begin{aligned}U_3 &= \mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-e} - \tilde{P}(R) \\ U_2 &= \rho_n^{-e}\left(T_G(R) - \mathbb{E}(T_G(R)|\boldsymbol{\xi})\right) \\ U_1 &= \rho_n^{-e}\left(\tilde{T}_{b2} - \mathbb{E}_b[T_{b2}(R)|G]\right)\end{aligned}$$

Now, it is easy to see that

$$\begin{aligned}\mathrm{Var}(\hat{T}^b(R)) &= \mathbb{E}(\mathrm{Var}(T_{b2}(R)|G) + \mathrm{Var}(\mathbb{E}(T_{b2}(R)|G)) \\ &= \mathbb{E}(\mathrm{Var}(\hat{T}_{b2}(R) - T_R(G)|G) + \mathrm{Var}(T_G(R)) \\ &= \mathbb{E}(\mathrm{Var}(U_1|G)) + \mathbb{E}(\mathrm{Var}(T_G(R)|\boldsymbol{\xi})) + \mathrm{Var}(\mathbb{E}(T_G(R)|\boldsymbol{\xi})) \\ &= \mathbb{E}(\mathrm{Var}(U_1|G)) + \mathbb{E}(\mathrm{Var}(U_2|\boldsymbol{\xi})) + \mathrm{Var}(U_3)\end{aligned}$$

We shall try to see the behavior of $\mathrm{Var}(U_1|G) = \mathrm{Var}_b[T_{b3}(R)|G]$. From the formula of $\mathrm{Var}_b[T_{b2}(R)|G]$, we see that, the covariance terms vanishes when $\pi_{SS'} = \pi^2$. Now,

if $q_1 = 1$, then, $\pi_{SS'} = \pi^2$ if $E(S) \cap E(S') = \phi$. The number of pairs such that $E(S) \cap E(S') \neq \phi$ is $O(p^2 n^{2p-2})$. So,

$$\mathbb{E}(\text{Var}(U_1|G)) = O\left(\frac{p^2 n^{2p-2}}{N}\right) = O(p^2/n^2) = o(n^{-1})$$

Now, the condition of $q_1 = 1$ is a bit restrictive. In stead, if we have $q_1 \to 1$ as $n \to \infty$, then, the highest order term of covariance term comes from the case when $E(S) \cap E(S') \neq \phi$ but the root nodes are same that is $w_1 = w_1'$. So, for some constant $C > 0$,

$$\begin{aligned}
&\frac{1}{N^2} \sum_{S,S' \in \mathcal{S}_p, S \neq S'} \frac{\pi_{SS'} - \pi^2}{\pi^2} \mathbf{1}(S \cong R,\ S' \cong R) \\
&\leq \frac{C}{N^2} \sum_{S,S' \in \mathcal{S}_p, S \neq S'} \frac{q_1 - q_1^2}{q_1^2} \mathbf{1}(S \cong R,\ S' \cong R) \\
&= O\left(\left(\frac{1}{q_1} - 1\right) \frac{n^{2p-1}}{n^{2p}}\right) \\
&= O\left(\left(\frac{1}{q_1} - 1\right) \frac{1}{n}\right) \\
&= o(n^{-1})
\end{aligned}$$

Now, for the variance term to vanish we need the conditions $q_1 = 1$ or $q_1 \to 1$ and $q_d \to 0$ and $nq_d \to \infty$ for $d = 2, \ldots, p$ as $n \to \infty$. Under these conditions, we have

$$\begin{aligned}
\frac{1}{N^2} \frac{1-\pi}{\pi} \sum_{S \in \mathcal{S}_p} \mathbf{1}(S \cong R) &= \left(\frac{1}{\pi} - 1\right) O\left(\frac{n^p}{n^{2p}}\right) \\
&= O\left(\frac{1}{n^p \pi}\right) \\
&= O\left(\frac{1}{n} \cdot \prod_{d=2}^{p} \frac{1}{nq_d}\right) \\
&= o(n^{-1})
\end{aligned}$$

So, we have,

$$\text{Var}(U_1) = o(n^{-1})$$

Now, by proof of Theorem 1 in [23], we have,

$$\begin{aligned}
\text{Var}(U_2) &= o(n^{-1}) \\
\text{Var}(U_3) &= o(n^{-1})
\end{aligned}$$

So, we get, $\text{Var}(\rho^{-e}\tilde{T}_{b2}(R)) = o(n^{-1})$. Since, we already know $\sqrt{n}$-consistency of $\left(\rho_n^{-e}T_G(R) - \tilde{P}(R))\right)$, this proves the $\sqrt{n}$-consistency of $\rho_n^{-e}\tilde{T}_{b2}(R)$ to $\rho_n^{-e}T_G(R)$.

□

Comparison of the Bootstrap Methods

Among the two bootstrap methods, the *uniform subsampling* scheme works for dense graphs only. If the subgraph pattern becomes large or if the original network is sparse, then the subgraph size also need to be large leading to slowing of the bootstrap scheme. However for dense graphs the *uniform subsampling* bootstrap scheme is fast and accurate.

The *sampling-based* bootstrap scheme is more accurate and stable, that means it has less bootstrap variance than the *uniform subsampling* bootstrap scheme. But, for dense graphs it becomes slower than *uniform subsampling* bootstrap to maintain its low bootstrap variance.

So, if speed is your concern, *uniform subsampling* bootstrap might be better choice, however, if you want a more reliable estimate then, *sampling based* bootstrap would be a better choice.

4.4 Theoretical Results

In this section, we try to give an estimate of asymptotic variance of $\rho^{-e}T_G(R)$, $\sigma^2(R)$, which is defined in Theorem 1 of [23]. By obtaining an estimate of the asymptotic variance of $\rho^{-e}T_G(R)$, we can estimate its asymptotic distribution and thus construct hypothesis tests based on the asymptotic distribution. We combine the results obtained in Section 4.2 to prove Theorem 4.2.1

Estimation of Variance and Covariance

We shall try to find the variance of the statistic $\rho^{-e}T_G(R)$ and then, using it give an estimate of the variance of the statistic $\hat{T}_G(R)$. The source of variation here is the randomness coming from sampling from the underlying model (5.2).

$$\begin{aligned}
\text{Var}\left[\rho^{-e}T_G(R)\right] &= \text{Var}\left[\sum_{S\subseteq K_n, S\cong R} \frac{\mathbf{1}(S\subseteq H)}{\rho^e\binom{n}{p}|Iso(R)|}\right] \\
&= \frac{1}{\left(\rho^e\binom{n}{p}|Iso(R)|\right)^2}\mathbb{E}\left[\sum_{S\subseteq K_n, S\cong R}\mathbf{1}(S\subseteq H)\right]^2 - \left(\tilde{P}(R)\right)^2 \\
&= \frac{1}{\left(\rho^e\binom{n}{p}|Iso(R)|\right)^2}\mathbb{E}\left[\sum_{\substack{S,T\subseteq K_n \\ S,T\cong R, S\cap T\neq\phi}}\mathbf{1}(S,T\subseteq H)\right]
\end{aligned}$$

If R is a connected subgraph and $W = S \cup T$ and $k = |W|$ and $e_W \equiv |E(W)|$, then, $k = p, \ldots, 2p-1$ and each term of sum in RHS

$$\frac{1}{\left(\rho^e \binom{n}{p} |Iso(R)|\right)^2} \mathbb{E}\left[\sum_{W \subseteq K_n} \mathbf{1}(W \subseteq H)\right] = \frac{\rho^{e_W}\binom{n}{k}|Hom(W)|}{\left(\rho^e \binom{n}{p}|Iso(R)|\right)^2} \tilde{P}(W) = O(n^{k-2p}\rho^{e_W-2e}) \tag{4.15}$$

If $|W| = |S \cup T| = 2p-1$, then, W is a connected graph, with $e_W = 2e$. So, we have the main leading term equals $\mathbf{O}(\frac{\mathbf{1}}{\mathbf{n}})$ for acyclic R. For, any other W, such that, $|W| = k < 2p-1$ then the Eq (4.15) becomes

$$\frac{1}{\left(\rho^e \binom{n}{p} |Iso(R)|\right)^2} \mathbb{E}\left[\sum_{W \subseteq K_n} \mathbf{1}(W \subseteq H)\right]$$
$$= \frac{\rho^{e_W}\binom{n}{k}|Hom(W)|}{\left(\rho^e \binom{n}{p}|Iso(R)|\right)^2} \tilde{P}(W) = O(n^{k-2p}\rho^{e_W-2e}) = o(n^{-1})$$

under the condition that ρ decreases at a rate slower than $O(n^{-1})$, that means except when we are in the constant degree case. The condition is equivalent to stating that $\lambda_n \to \infty$.

So, for calculation of variance, we only estimate the count of the features which are $W = S \cup T$ and $|W| = 2p-1$, that means S and T have only one node in common. So, the estimator of variance becomes -

$$\tilde{\sigma}^2(R) = \frac{1}{\left(\rho^e \binom{n}{p}|Iso(R)|\right)^2} \sum_{W \subseteq K_n, W = S \cup T, S, T \cong R, |S \cap T| = 1} \mathbf{1}(W \subseteq G) \tag{4.16}$$

and we have

$$\mathbb{E}\tilde{\sigma}^2(R) = \text{Var}\left[\rho^{-e} T_G(R)\right] + o(n^{-1})$$

Similarly, for calculation of covariance between two count statistics, $T_G(R_1)$ and $T_G(R_2)$, we only estimate the count of the features which are $W = S \cup T$ and $|S \cap T| = 1$, $S \cong R_1$, $T \cong R_2$, that means S and T have only one node in common. So, the estimator of covariance becomes -

$$\tilde{\sigma}(R_1, R_2)$$
$$= \frac{1}{\left(\rho^{e_{R1}} \binom{n}{p}|Hom(R_1)|\right)} \frac{1}{\left(\rho^{e_{R2}} \binom{n}{p}|Hom(R_2)|\right)} \sum_{W \subseteq K_n, W = S \cup T, S \cong R_1, T \cong R_2, |S \cap T| = 1} \mathbf{1}(W \subseteq G) \tag{4.17}$$

where $e_{R_1} = |E(R_1)|$ and $e_{R_2} = |E(R_2)|$. So, we have

$$\mathbb{E}\tilde{\sigma}(R_1, R_2) = \text{Cov}\left[\rho^{-e_{R_1}} T_G(R), \rho^{-e_{R_2}} T_G(R_2)\right] + o(n^{-1})$$

Now, from the Theorem 1(a) in [23], we know that as $\lambda_n \to \infty$, if $\hat{\rho}_n = \frac{\bar{D}}{n-1}$ as defined in (4.3),

$$\frac{\hat{\rho}_n}{\rho_n} \xrightarrow{P} 1$$

So, using the estimate $\hat{\rho}_n$, we define the estimate of variance -

$$\hat{\sigma}^2(R) = \frac{1}{\left(\hat{\rho}_n^e \binom{n}{p} |Iso(R)|\right)^2} \sum_{W \subseteq K_n, W = S \cup T, S, T \cong R, |S \cap T| = 1} \mathbf{1}(W \subseteq G) \tag{4.18}$$

and the estimate of covariance is -

$$\begin{aligned} &\hat{\sigma}(R_1, R_2) \\ &= \frac{1}{\left(\rho^{e_{R1}} \binom{n}{p} |Hom(R_1)|\right)} \frac{1}{\left(\rho^{e_{R2}} \binom{n}{p} |Hom(R_2)|\right)} \sum_{W \subseteq K_n, W = S \cup T, S \cong R_1, T \cong R_2, |S \cap T| = 1} \mathbf{1}(W \subseteq G) \end{aligned} \tag{4.19}$$

So, $\hat{\sigma}^2(R)$ and $\hat{\sigma}(R_1, R_2)$ become consistent estimates of $\text{Var}\left[\rho^{-e} T_G(R)\right]$ and $\text{Cov}\left[\rho^{-e_{R_1}} T_G(R_1), \rho^{-e_{R2}} T_G(R_2)\right]$ respectively and consequently a consistent estimate of $\text{Var}\left[\hat{T}_G(R)\right]$ and $\text{Cov}\left[\hat{T}_G(R_1), \hat{T}_G(R_2)\right]$ respectively.

Lemma 4.4.1. *As $\lambda_n \to \infty$ and $n \to \infty$,*

$$\frac{\hat{\sigma}^2(R)}{Var\left[\rho^{-e} T_G(R)\right]} \xrightarrow{P} 1 \tag{4.20}$$

$$\frac{\hat{\sigma}(R_1, R_2)}{Cov\left[\hat{T}_G(R_1), \hat{T}_G(R_2)\right]} \xrightarrow{P} 1 \tag{4.21}$$

Proof. The proof follows from previous discussion. □

Now, we can see that $\hat{\sigma}^2(R)$ and $\hat{\sigma}(R_1, R_2)$ are nothing but count statistics on the statistic $W = S \cup T$, where, $S, T \cong R$ and $|S \cap T| = 1$. So, using bootstrap methods, we can get an estimate of $\hat{\sigma}^2(R)$ -

$$\hat{\sigma}^2_{bi}(R) = \frac{\left(\hat{\rho}_n^{eW} \binom{n}{2p-1} |Iso(R)|\right)}{\left(\hat{\rho}_n^e \binom{n}{p} |Iso(R)|\right)^2} \tilde{T}_{bi}(W) \text{ for } i = 1, 2. \tag{4.22}$$

and an estimate of $\hat{\sigma}(R_1, R_2)$ -

$$\hat{\sigma}_{bi}(R_1, R_2) = \frac{\left(\hat{\rho}_n^{eW} \binom{n}{2p-1} |Iso(R)|\right)}{\left(\hat{\rho}_n^{e_{R_1}} \binom{n}{p} |Hom(R_1)|\right) \left(\hat{\rho}_n^{e_{R_2}} \binom{n}{p} |Hom(R_2)|\right)} \tilde{T}_{bi}(W) \text{ for } i = 1, 2. \tag{4.23}$$

where, $W = S \cup T$ with $S, T \cong R$ and $|S \cap T| = 1$ and $|V(W)| = 2p - 1$ and $e_W = |E(W)|$. $\tilde{T}_{bi}(W)$ $(i = 1, 2)$ are bootstrap count statistics estimates, defined in Eq (4.10) and (4.13). Also from Lemma 4.3.1 and Lemma 4.3.2, we get the $\sqrt{n}$-consistency of the bootstrap count estimates $\tilde{T}_{bi}(W)$ $(i = 1, 2, 3)$. So, we can now combine the Lemma 4.3.1 and 4.3.2 and Lemma 4.4.1, to see that $\hat{\sigma}^2_{bi}(R)$ and $\hat{\sigma}_{bi}(R_1, R_2)$ $(i = 1, 2)$ are consistent estimators of $\sigma^2(R)$ and $\sigma(R_1, R_2)$ respectively by Slutsky's Theorem and Convergence of Types Theorem. So, we get an estimate of variance, $\hat{\sigma}^2_{bi}(R)$ $(i = 1, 2)$ with the property

Lemma 4.4.2. *as $\lambda_n \to \infty$, $n \to \infty$ and under conditions of Lemma 4.3.1 and Lemma 4.3.2,*

$$\frac{\hat{\sigma}^2_{bi}(R)}{\sigma^2(R)} \xrightarrow{P} 1 \text{ for } i = 1, 2 \tag{4.24}$$

$$\frac{\hat{\sigma}_{bi}(R_1, R_2)}{\sigma(R_1, R_2)} \xrightarrow{P} 1 \text{ for } i = 1, 2 \tag{4.25}$$

Proof. The proof follows from previous discussion. □

Proof of Theorem 4.2.2

The proof of the internal theorem follows from the lemmas of previous section. Since, we have $\sqrt{n}$-consistent bootstrap estimators of $\rho^{-e}\tilde{T}_b(R)$. Now, from the Theorem 1(a) in [23], we know that as $\lambda_n \to \infty$, if $\hat{\rho}_n = \frac{\bar{D}}{n-1}$ as defined in (4.3),

$$\begin{aligned} \frac{\hat{\rho}_n}{\rho_n} &\xrightarrow{P} 1 \\ \sqrt{n}\left(\frac{\hat{\rho}_n}{\rho_n} - 1\right) &\xrightarrow{w} N(0, \sigma^2) \end{aligned}$$

Now, we can define the bootstrap estimates as -

$$\hat{T}_{bi}(R) = \hat{\rho}^{-e}\tilde{T}_{bi}(R) \text{ for } i = 1, 2, 3. \tag{4.26}$$

So, we get by applying Slutsky's Theorem that

$$\sqrt{n}\left(\hat{T}_{bi}(R) - \hat{T}_G(R)\right) \xrightarrow{P} 0 \text{ for } i = 1, 2, 3.$$

The statement about bootstrap estimate of variance follows from Lemma 4.4.2 and the definitions of bootstrap variance in the form of equation (4.22).

Proof of Theorem 4.2.1

The proof of the main theorem follows from the lemma of Section 4.4 and Theorem 4.2.2. We have $\sqrt{n}$-consistent bootstrap estimators, $\hat{T}_{bi}(R)$ (for $i = 1, 2, 3$) of $\hat{T}_G(R)$ and consistent

estimators, $\hat{\sigma}^2_{bi}(R)$ (for $i = 1, 2$) of $\sigma^2(R)$. Also from Theorem 1 of [23], we have, for subgraphs $R_1, \ldots, R_k$ of G_n,

$$\sqrt{n}\left(\left(\hat{T}_G(R_1), \ldots, \hat{T}_G(R_k)\right) - \left(\tilde{P}(R_1), \ldots, \tilde{P}(R_k)\right)\right) \xrightarrow{w} N(\mathbf{0}, \Sigma(\mathbf{R}))$$

So, we can combine the result from Theorem 4.2.2 with the above theorem, using Slutsky and convergence of types theorem, to get the symptomatic normality behavior of $\hat{T}_{bi}(R)$. As $n \to \infty$, $\lambda_n \to \infty$ and under conditions of Lemma 4.3.1, Lemma 4.7.1 and Lemma 4.3.2, if we define, $\mathbf{T}_{bi}(\mathbf{R}) = \left(\hat{T}_{bi}(R_1), \ldots, \hat{T}_{bi}(R_k)\right)$ and $\mathbf{P}(\mathbf{R}) = \left(\tilde{P}(R_1), \ldots, \tilde{P}(R_k)\right)$

$$\sqrt{n}\left(\left(\mathbf{T}_{bi}(\mathbf{R}) - \mathbf{P}(\mathbf{R})\right)\hat{\Sigma}^{-1/2}_{bi}(\mathbf{R})\left(\mathbf{T}_{bi}(\mathbf{R}) - \mathbf{P}(\mathbf{R})\right)\right) \xrightarrow{w} N(\mathbf{0}, \mathbf{I}) \text{ for } i = 1, 2, 3$$

where, $[\Sigma_{bi}]_{st} = \hat{\sigma}_{bi}(R_s, R_t)$, $s, t = 1, \ldots, k$ and if $R_s = R_t = R$, $\sigma_{bi}(R_s, R_t) = \hat{\sigma}^2_{bi}(R)$ for $i = 1, 2$.

4.5 Simulation Results

We apply three (the two described and the snowball sampling variant given in Appendix) representative bootstrap subsampling schemes for simulated datasets to find out their performances. We generate data from two different simulation models. Both models are special cases of the nonparametric model described in [22]. The two models that we consider are -

- **Stochastic block model**
- **Preferential attachment model**

For each of the models, we try to find estimate of the count statistics features and their confidence intervals through bootstrap subsampling. The features that we consider are generalized (k, l)-wheels, triangles and a smooth function of them, transitivity.

Count Statistics

The main class of acyclic features we consider are generalized (k, l)-wheels.

Definition 4.5.1 (Wheels)**.** *A (k, l)-wheel is an acyclic graph with $kl+1$ vertices and kl edges isomorphic to the graph with edges $\{(1, 2), \ldots, (k, k+1)\ (1, k+2), \ldots, (2k, 2k+1)\ \ldots, (1, (l-1)k + 2), \ldots, (lk, lk + 1)\}$.*

In other words a (k, l)-wheel is a subgraph R, such that it contains

i) A "hub" vertex

ii) l spokes from hub

iii) Each spoke has k connected vertices.

Edges, 'V', 'W' are all examples of (k, l)-wheels. An edge is a $(1, 1)$-wheel, a 'V' is a $(1, 2)$-wheel and a 'W' is a $(2, 2)$-wheel.

Definition 4.5.2 (Generalized Wheels)**.** *A generalized $(\boldsymbol{k}, \boldsymbol{l})$-wheel, where $\boldsymbol{k} = (k_1, \ldots, k_t)$, $\boldsymbol{l} = (l_1, \ldots, l_t)$ are vectors and the k_j's are distinct integers, is the union $R_1 \cup \ldots \cup R_t$, where R_j is a (k_j, l_j)-wheel, $j = 1, \ldots, t$ and the wheels $R_1, \ldots, R_t$ share a common hub but all their spokes are disjoint.*

A $(\boldsymbol{k}, \boldsymbol{l})$-wheel has a total of $p = \sum_j k_j l_j + 1$ vertices and $\sum_j k_j l_j$ edges. The following picture is an example of $((2, 1, 1), (1, 1, 1))$-wheel and it is an union of two 'V's, where the common vertex is the hub of one 'V' and leaf of the other 'V'.

In these simulations, we consider counts of simple (k, l)-wheels such as $(1, 2)$ and $(1, 3)$. We also consider the count of the cyclic pattern such as triangle and *quadrilaterals* or 4-cycles. We consider a smooth function of counts of triangle and 'V's, known as **transitivity**, T_{Tr}, defined as

$$T_{Tr} = \frac{\hat{P}(R_1)}{\hat{P}(R_2) + \hat{P}(R_2)}$$

where, R_1 is a triangle or a 3-cycle and R_2 is a 'V' or a $(1, 2)$-wheel.

Stochastic Block Model

Let w correspond to a K-block model defined by parameters $\theta = (\boldsymbol{\pi}, \rho_n, S)$, where π_a is the probability of a node being assigned to block a as before, and

$$\mathbf{F}_{ab} = \mathbb{P}(A_{ij} = 1 | i \in a, j \in b) = \rho_n S_{ab}, \quad 1 \leq a, b \leq K.$$

and the probability of node i to be assigned to block a to be π_a $(a = 1, \ldots, K)$.

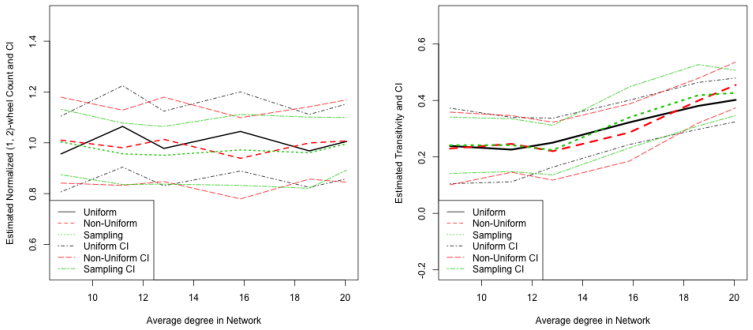
We consider a stochastic block model with $K = 2$. We consider the parameter matrix $\mathbf{F} = \tilde{\lambda} F^{(1)} + (1 - \tilde{\lambda}) F^{(2)}$, where, $F^{(1)}_{2\times 2} = \text{Diag}(0.0525, 0.0975)$ and $F^{(2)}_{2\times 2} = 0.015\mathbf{J}_2$, where, $\mathbf{J}_2$ is a 2×2 matrix of all 1's. So, we get $\rho_n = \boldsymbol{\pi}^T \mathbf{F} \boldsymbol{\pi}$. We now, vary $\tilde{\lambda}$ to get different combinations of $\mathbf{F}$ as well as ρ_n.

In the following figures, we try to see the behavior of mean and variances of the count statistics, as we vary λ_n and n for the model. In Figure 4.1(a), we compare the mean and variance of the normalized count statistic, $\hat{T}_G(R)$, where, R is $(1, 2)$-wheel. We find the bootstrap estimator $\tilde{T}_{bi}(R)$ for all three bootstrap schemes - $i = 1, 2, 3$ and we also find the corresponding estimates of variance by bootstrap, $\hat{\sigma}^2_{bi}(R)$, for all three bootstrap schemes - $i = 1, 2, 3$. We then plot the plot the estimator $\tilde{T}_{bi}(R)$ along with the asymptotic 95% confidence interval using the asymptotic normality result of Theorem 4.2.1 and the

bootstrap estimates of variance $\hat{\sigma}^2_{bi}(R)$. In Figure 4.1(b) we have a similar plot, but instead of $T_G(R)$, we use the statistic T_{Tr}. We find the bootstrap estimate of T_{Tr} in the form,

$$\hat{T}^b_{Tr} = \frac{\hat{T}_b(R_1)}{\hat{T}_b(R_2) + \hat{T}_b(R_2)}$$

where, $\hat{T}_b$ is the bootstrap estimate of count statistic $T_G(R)$ and R_1 is a triangle or a 3-cycle and R_2 is a 'V' or a $(1,2)$-wheel. The bootstrap estimate of asymptotic variance of $\hat{T}^b_{Tr}$ is obtained from the bootstrap estimates of $\hat{\sigma}^2_b(R_1)$, $\hat{\sigma}^2_b(R_2)$ and $\hat{\sigma}_b(R_1, R_2)$ by using Delta method and using the Theorem 4.2.1.



(a) (b)

Figure 4.1: For $n = 500$, we vary average degree (λ_n) and (a) Plot estimated normalized (1, 2)-wheel count (b) Plot estimated Transitivity and their 95% Confidence Interval (CI), where, CI is estimated using bootstrap estimates of variance of the estimators. We use different colors to indicate different bootstrap subsampling schemes.

In Figure 4.2, we see that the variance of the bootstrap estimates, arising solely from bootstrap iterations and not from from model-based iterations, decrease, as we increase the number of subsamples in the *Unifrom Subsampling Scheme*. In Figure 4.3, we compare the variance of the bootstrap estimates, based on bootstrap iterations for the three different bootstrap schemes. We see that bootstrap variance is universally low for the *Sampling-based Scheme* as we vary average degree, λ_n of the graph. However, the bootstrap variance of the *Non-uniform Snowball Sampling Scheme*, decrease, as average degree λ_n increase. We expect such a behavior, as, λ_n increase, the bootstrap subsample in *Non-uniform Snowball Sampling Scheme* becomes large and stable. So, based on simulations, we recommend *Sampling-based Scheme* for bootstrap.

In Figure 4.4, we try to see the behavior of mean and variances of the count statistics, $(1,3)$-wheels and 4-cycles. We use only *Sampling-based Scheme* for bootstrap in this case. Like in Figure 4.1, we plot the plot the estimator $\tilde{T}_{b2}(R)$ along with the asymptotic 95%

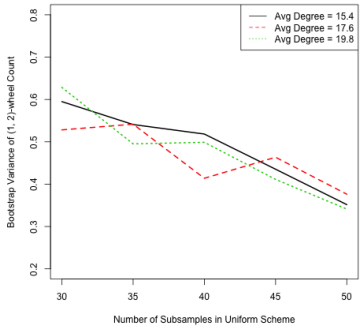
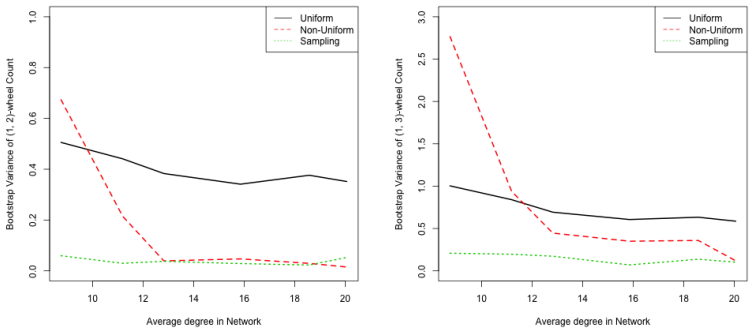


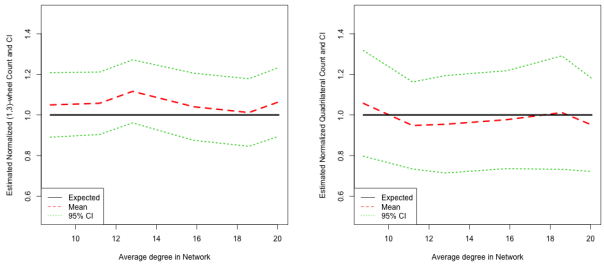
Figure 4.2: For $n = 500$ and $\lambda_n = 19.875$, we vary the subsample size of the Uniform subsampling scheme and plot the bootstrap variance of bootstrap estimators of Uniform subsampling scheme.



(a) (b)

Figure 4.3: For $n = 500$, we vary average degree (λ_n) and plot (a) bootstrap variance of estimated normalized (1, 2)-wheel count (b) bootstrap variance of normalized (1,3)-wheel count.We use different colors to indicate different bootstrap subsampling schemes.

confidence interval using the asymptotic normality result of Theorem 4.2.1 and the bootstrap estimates of variance $\hat{\sigma}^2_{b2}(R)$, for R as $(1,3)$-wheel in Figure 4.4(a) and or R as 4-cycle in Figure 4.4(b).



(a) (b)

Figure 4.4: For $n = 500$, we vary average degree (λ_n) and (a) Plot estimated normalized (1, 3)-wheel count (b) Plot estimated normalized 4-cycle count and their 95% Confidence Interval (CI), where, CI is estimated using bootstrap estimates of variance of the estimators. We use only Sampling-based bootstrap scheme.

Preferential Attachement Model

In Preferential Attachement Model, given k initial vertices, $k+1$th vertex attach to one of the preceding k vertices with probability proportional to degree. Now, we have degree of vertex v defined as D_v and $\bar{D} = \frac{1}{n}\sum_{v=1} D_v$. Also, we have,

$$\tau(v) \simeq \frac{D_v}{\bar{D}}$$

So, following Eq. (5.2), we have the probability of edge formation as

$$w(u,v) = \frac{\tau(u)}{T(u)}\mathbf{1}(u \le v) + \frac{\tau(v)}{T'(u)}\mathbf{1}(v \le u)$$

where, $T(u) = \int_u^1 \tau(s)ds$ and $T'(v) = 1 - T(v)$ and

$$\tau(u) = \int_0^1 w(u,v)dv$$

So, the preferential attachment model can be defined by the following formula on w

$$w(u,v) = \frac{\tau(u)}{\int_u^1 \tau(s)ds}\mathbf{1}(u \le v) + \frac{\tau(v)}{\int_v^1 \tau(s)ds}\mathbf{1}(v \le u)$$

Thus, for

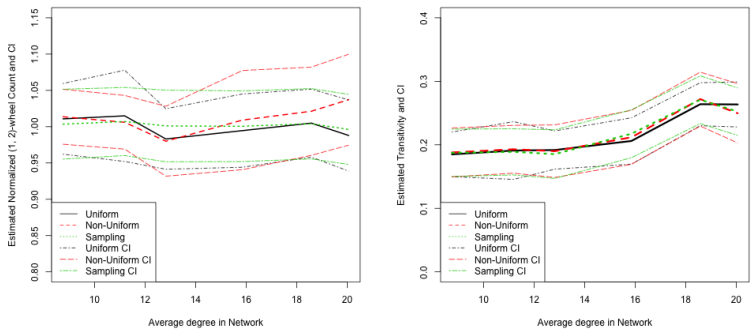
$$w(u,v) \;=\; (1-u)^{-1/2}(1-v)^{-1/2}$$

we have,

$$\tau(v) = c(1-v)^{-1/2}$$

which is equivalent to power law of *degree distribution* $F \equiv \tau^{-1}$.

In the following figure, we try to see the behavior of mean and variances of the count statistics, as we vary λ_n for the model. In Figure 4.5(a), we compare the mean and variance of the normalized count statistic, $\hat{T}_G(R)$, where, R is $(1,2)$-wheel. We find the bootstrap estimator $\tilde{T}_{bi}(R)$ for all three bootstrap schemes - $i = 1,2,3$ and we also find the corresponding estimates of variance by bootstrap, $\hat{\sigma}^2_{bi}(R)$, for all three bootstrap schemes - $i = 1,2,3$. We then plot the plot the estimator $\tilde{T}_{bi}(R)$ along with the asymptotic 95% confidence interval using the asymptotic normality result of Theorem 4.2.1 and the bootstrap estimates of variance $\hat{\sigma}^2_{bi}(R)$. In Figure 4.1(b) we have a similar plot, but instead of $T_G(R)$, we use the statistic T_{Tr}. We find the bootstrap estimate of T_{Tr} and the bootstrap estimate of asymptotic variance of $\hat{T}^b_{Tr}$ is obtained from the bootstrap estimates of $\hat{\sigma}^2_b(R_1)$, $\hat{\sigma}^2_b(R_2)$ and $\hat{\sigma}_b(R_1,R_2)$ by using Delta method and using the Theorem 4.2.1.



(a) (b)

Figure 4.5: For $n = 500$, we vary average degree (λ_n) and (a) Plot estimated normalized (1, 2)-wheel count (b) Plot estimated Transitivity and their 95% Confidence Interval (CI), where, CI is estimated using bootstrap estimates of variance of the estimators. We use different colors to indicate different bootstrap subsampling schemes.

Comparison Between Stochastic Block Model and Preferential Attachment Model

We simulate networks from both stochastic block model and preferential attachment model and then, we try to compare the distribution of a statistic of the graph for two different

networks. As a statistic, we use transitivity here. In Figure 4.6, compare the bootstrap estimated mean and variance of transitivity of the networks simulated from the two different models. We keep the average degree, λ_n, of the two simulated networks same and then, we

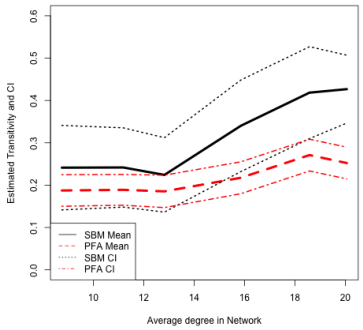


Figure 4.6: For $n = 500$, we vary λ_n and we plot estimated Transitivity and their 95% Confidence Interval (CI), where, CI is estimated using bootstrap estimates of variance of the estimators. We use different colors to indicate different models from which networks are generated.

try to get the asymptotic distribution of the transitivity statistic for the two cases for each λ_n. We see here that, for low average degree, we can not statistically distinguish between the transitivity of networks generated from two different models, but, they become statistically distinguishable as average degree, λ_n, increases.

4.6 Real Data Examples

Social networks recently has become quite large after the introduction of social networking sites. We consider two different social networks as a platform for our experiments. The first one, High School Romantic Relations data (HS) is a small social network, whereas the second one, Facebook College Social Network (FB) has greater number of nodes and links. The Facebook data was presented in [155]. In this dataset, the Facebook social network of different colleges are represented. In a network, the nodes are people of the colleges and a link represents online friendship between the nodes. The High School data was presented in [14]. In this network, each student is a node and a link between students indicate that they had romantic relations.

High School Network

In this application, we try to quantitatively verify some of the hypothesis mentioned by the authors in the paper [14] when presenting the data. The network here is formed by students

of Jefferson High school as nodes and if two students have romantic relations then there exists a link between those two nodes. In the paper, [14], where the data was presented, an observation was made about the dearth of short cycles in the network. Our application here is trying to answer the question whether the absence of short cycles in this graph is significant or not. We consider a very simple model for the data.

We consider the data has been generated from two different models -

(a) Stochastic block model with two blocks (Male and Female) and the connection probability matrix is given by

$$P = \begin{pmatrix} \hat{P}_{11} & \hat{P}_{12} \\ \hat{P}_{12} & \hat{P}_{22} \end{pmatrix}$$

where, $\hat{P}_{ab} =$ Average number of edges between blocks$a$ and b in the network, where, $a, b = 1, 2$ are the two blocks with Male $= 1$ and Female $= 2$. In this network, we have $\hat{P}_{11} = 0$, $\hat{P}_{12} = 0.0058$ and $\hat{P}_{22} = 0.000025$. The probability of belonging to the two blocks are $(0.497, 0.503)$.

(b) Preferential attachment model with $\rho = \frac{\lambda_n}{n}$, where, $\lambda_n =$ Average degree of the network $= 1.66$ and n is the number of nodes.

Now, for these two simple models, we can theoretically find the normalized count of small cycles. Then, we can perform a hypothesis test to find out whether the number of small cycles we see in this network is significantly small or not. We use the results of Theorem 1 to form the asymptotic test. The results are given in Table 4.1. We see in the results that, according to the two simple models, it is extremely unlikely for 3-cycles and 4-cycles to occur in the graph. In fact, the original network has *too many* 4-cycles short cycles not *too few*. This is an interesting observation coming out of our simple exploratory analysis. So, our simple models do not capture the probabilistic mechanism of the original network correctly and we need to analyze the short cycles in the network more closely to understand their formation.

Subgraph	Normalized Count	Standard Deviation	Count (SBM)	Count (PFA)
(1,2)-wheel	2.27	0.17	1.01	2.97
3-cycle	1.31	0.1	0.01	1.04
4-cycle	9.47	3.16	0.63	3.06

Table 4.1: The normalized subgraph counts, their standard deviation and the expected counts from the stochastic block model (SBM) and preferential attachment model (PFA) for the whole high school network.

Note that, this is a very small and sparse network. For this network, the use of Theorem 2 from [23] would have sufficed, but we give the example as an example of the use of count

statistics and their quantitative behavior. Here in the paper [14] permutation tests were used. We use asymptotic Gaussian tests and we can directly answer the questions without the possible awkwardness of permutation tests.

Facebook Network

In this application, we try to quantitatively analyze the behavior of some of the known descriptive statistics for Facebook collegiate networks. The networks were presented in the paper by Traud et.al. (2011) [155]. The network is formed by Facebook users acting as nodes and if two Facebook users are "friends" there is an edge between the corresponding nodes. Along with the network structure, we also have the data on covariates of the nodes. Each node has covariates: gender, class year, and data fields that represent (using anonymous numerical identifiers) high school, major, and dormitory residence. We try to answer two very basic questions quantitatively for these networks -

1. Can the node covariates act as cluster identifiers?

2. Can two college networks be distinguishable in terms of some basic descriptive statistics?

In order to address the first question, we consider the network of a specific college (Caltech). We consider the covariates class year, major and dormitory residence as our covariates of interest. Note that each of these covariates are district covariates. We take the induced network created by levels of each of these covariates and try to see if those networks have different clustering properties. For example, consider class year and major as the covariates of interest. We consider the nodes belonging two different class years and find their induced network from the whole collegiate network. Similarly, we consider the nodes belonging two different majors and find their induced network from the whole collegiate network. Now, we have two different networks, one of which has nodes coming exclusively from two different class years and another has nodes coming exclusively from two different majors. We now try to find which of two networks is more "clustered" by comparing *transitivity* of the two networks. We can repeat the same exercise for any two covariates and choosing a subset of their levels.

The second question can also be answered in the similar spirit as the first one. We consider the full collegiate network of two different colleges (Caltech and Princeton). Then, we try to compare the mean degree and transitivity of these two collegiate networks.

These comparisons could in principle be possible using the results given in Bickel et. al. (2011) [23], but computationally intractable. Using bootstrap estimators, we can estimate the variance of the estimators and thus perform hypothesis testing in reasonable time.

In Tables 4.6, 4.3 and 4.4, we present an excerpt of the result of our analysis and the answer the both the questions. The results indicate that some conclusions in [155] are questionable.

	Class Year(CY)	Dormitory(DM)	Major(MJ)
Estimated Transitivity	0.15	0.22	0.12

Table 4.2: Transitivity of induced networks formed by considering only two levels of a specific covariate of a specific collegiate network.

Difference	CY and DM	DM and MJ
Estimated	0.07	0.1
Estimated SD	0.05	0.035

Table 4.3: The Difference between Class Year and Dorm is not significant but difference between Dorm and Major is significant by asymptotic normal test at 5% level. The data was presented in Traud et. al. (2011) SIAM Review.

	Network 1	Network 2
Estimated Transitivity	0.29	0.16
Estimated Difference	0.13	
Estimated Difference SD	0.11	

Table 4.4: The Difference of transitivity between two networks is not significant by asymptotic normal test at 5% level. Therefore Network 1 can not be said to be more *'clusterable'*. The data was presented in Traud et. al. (2011) SIAM Review.

Now, without finding the bootstrap estimate of count statistics and its variance, finding the asymptotic distribution of these count statistics will not have been possible. So, now, with the help of the bootstrap based estimates we can perform hypothesis testing on the count statistics and provide their estimates of their asymptotic distribution.

4.7 Conclusion and Future Works

In this chapter, we have considered three known subsampling schemes of networks and tried to show situations, where, they are applicable to find the asymptotic distribution of certain *local statistics* of the network. We consider the count of fixed subgraphs of the network as *local statistics* and call them *count statistics*. These have also been referred to as motif counts. We showed that the bootstrap subsample estimates of the count statistics and their smooth functions have asymptotic normal distribution. We proposed bootstrap schemes by which we could efficiently compute the asymptotic mean and variance of these count statistics. We also showed that the *Sampling based* bootstrap subsampling scheme seemed most stable and we recommend that scheme for use as bootstrap subsampling scheme.

We also use the estimated asymptotic mean and variances of the count statistics to construct hypothesis tests. These hypothesis tests can serve several purposes, such as

(a) Distinguish between the count statistics of two different networks

(b) Distinguish between parts of same network

(b) Testing whether a network has been generated from a specified model, by comparing the empirical and population version of the count statistic.

(c) Testing how close parameters of two different network models can become.

All of these different qualitative tests can be made quantitative by using hypothesis tests using the count statistics. We showed during simulations, that transitivity of networks from stochastic block models becomes easier to differentiate from transitivity of preferential attachment model as average degree grows. Similarly, in real networks, such as Facebook collegiate network, we show that certain covariate based subnetworks have more 'cluster' structure than others. We were also able to show that even in large networks conclusions based on means only as opposed to confidence statements using variances could be unreliable.

Future Works

Here we used bootstrap subsampling scheme to estimate *local statistics* only. But, one natural generalization can be use of bootstrap scheme to get asymptotic distribution of global statistics - such as graph cut, conductance, functionals of graphon (non-integral fundtionals) and such parameters. Sample and bootstrap estimates of such parameters are sometimes obtainable, but their theoretical properties are still unknown. It would be a nice future endeavor to extend bootstrap subsampling scheme to estimate such global characteristics of the networks.

Appendix

A1. Variance of $\hat{T}_{b1}(R)$

The variance of $T^b(R)$ is

$$\mathrm{Var}_b\left[\frac{1}{\binom{m}{p}|Iso(R)|}\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right] = \left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2 \mathrm{Var}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right]$$

$$\begin{aligned}\mathrm{Var}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right] &= \mathbb{E}_b\left[\left(\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\right)^2\Big|G\right] \\ &\quad -\left(\mathbb{E}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right]\right)^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}_b\left[\left(\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\right)^2\Big|G\right] &= \mathbb{E}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right] \\ &\quad + \mathbb{E}_b\left[\sum_{\substack{S,T\subseteq K_m \\ S,T\cong R, S\neq T}}\mathbf{1}(S,T\subseteq H)\Big|G\right] \\ &= I + II \text{ (Suppose)}\end{aligned}$$

Thus,

$$I = \sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)$$

$$II = \mathbb{E}_b\left[\sum_{\substack{S,T\subseteq K_m \\ S,T\cong R, S\neq T}}\mathbf{1}(S,T\subseteq H)\Big|G\right]$$

Now, a host of subgraphs can be formed by the intersection of two copies of R. The number of intersected vertices can range from 0 to $p-1$. Let us consider, that for number of vertices in intersection as k ($k=1,\ldots,(p-1)$), the number of graph structures that can be formed is g_k and we represent that graph structure by W_{jk}, where, $j=1,\ldots,g_k$. Thus,

$$II = \sum_{k=0}^{p-1}\sum_{j=1}^{g_k}\sum_{S\subseteq K_n, S\cong W_{jk}}\frac{\binom{n-(2p-k)}{m-(2p-k)}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)$$

So,

$$\begin{aligned}\mathbb{E}_b\left[\left(\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\right)^2\Big|G\right] &= \sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G) \\ &\quad + \sum_{k=0}^{p-1}\sum_{j=1}^{g_k}\sum_{S\subseteq K_n, S\cong W_{jk}}\frac{\binom{n-(2p-k)}{m-(2p-k)}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\end{aligned}$$

$$\begin{aligned}\mathrm{Var}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right] &= \sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\\ &\quad+\sum_{k=0}^{p-1}\sum_{j=1}^{g_k}\sum_{S\subseteq K_n, S\cong W_{jk}}\frac{\binom{n-(2p-k)}{m-(2p-k)}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\\ &\quad-\left(\sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right)^2\end{aligned}$$

$$\begin{aligned}&\mathrm{Var}_b\left[\frac{1}{\binom{m}{p}|Iso(R)|}\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\Big|G\right]\\ &\quad=\left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2\left[\sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right]\\ &\quad-\left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2\left[\left(\sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right)^2\right]\\ &\quad+\left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2\left[\sum_{k=0}^{p-1}\sum_{j=1}^{g_k}\sum_{\substack{S\subseteq K_n\\ S\cong W_{jk}}}\frac{\binom{n-(2p-k)}{m-(2p-k)}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right]\end{aligned}$$

So,

$$\begin{aligned}\mathrm{Var}_b\left[T^b(R)\right] &= \left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2\left[\sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right]\\ &\quad-\left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2\left[\left(\sum_{S\subseteq K_n, S\cong R}\frac{\binom{n-p}{m-p}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right)^2\right]\\ &\quad+\left(\frac{1}{\binom{m}{p}|Iso(R)|}\right)^2\left[\sum_{k=0}^{p-1}\sum_{j=1}^{g_k}\sum_{\substack{S\subseteq K_n\\ S\cong W_{jk}}}\frac{\binom{n-(2p-k)}{m-(2p-k)}}{\binom{n}{m}}\mathbf{1}(S\subseteq G)\right]\end{aligned}$$

A2. Degree-based non-uniform sampling bootstrap

In the *non-uniform subsampling* bootstrap scheme at each bootstrap iteration a subset of vertices of the full network G is selected by looking at the neighborhood graph of a randomly selected vertex and the graph induced by the selected subset is the subsample we

consider. This is a vertex subsampling scheme and is a variant of common snowball sampling scheme. The full bootstrap procedure given the neighborhood size, d and number of bootstrap iterates, B, is as follows – -

1. For b^{th} iterate of the bootstrap, $b = 1, \ldots, B$,
2. Fix d = Depth of R.
3. Choose $i \in V(G)$, as the central vertex. Form the graph H as the neighborhood graph with i as root and depth d. Let us denote $|V(H)|$ by m.
4. Calculate $T_{b2}(R)$, given by formula

$$T_{b3}(R) = \frac{n}{m\binom{n}{p}|Iso(R)|} \sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H, i \in S) \tag{4.27}$$

The bootstrap estimate of $T_G(R)$ is given by

$$\tilde{T}_{b3}(R) = \frac{1}{B} \sum_{b=1}^{B} T_{b3}(R) \tag{4.28}$$

In this sampling, all subgraphs H do not have the same probability of being sampled. Actually some subgraphs may have zero probability of being selected. In uniform sampling bootstrap, we were considering all possible subgraphs with vertex count equal to m. In degree-based bootstrap we consider all subgraph H, which are d-neighborhood graph a $i \in V(G)$. But, unlike all possible subgraphs with vertex count equal to m, all d-neighborhood graph are not distinct and so we have considered that there are n_R of them, where, $n_R \leq n$.

Lemma 4.7.1. *The estimator $\tilde{T}_{b3}(R)$ has the following properties*

(i) Given G, $\tilde{T}_{b3}(R)$ is an unbiased estimate of $T_G(R)$.

(ii) As $b \to \infty$, $\sqrt{n}\left(\rho^{-e}\tilde{T}_{b3}(R) - \rho^{-e}\hat{T}_G(R)\right) \xrightarrow{P} 0$

Proof. (i) Now, let us try to try to find the expectation of $T^b(R)$ under the sampling distribution conditional on the given data G.

$$\begin{aligned}
&\mathbb{E}_b\left[\frac{n}{m\binom{n}{p}|Iso(R)|} \sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H, i \in S)\Big|G\right] \\
&= \frac{n}{m\binom{n}{p}|Iso(R)|}\mathbb{E}\left[\sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H, i \in S)\Big|G\right] \\
&= \frac{n}{m\binom{n}{p}|Iso(R)|} \sum_{H \subseteq G} \mathbb{P}(H) \sum_{S \subseteq K_m, S \cong R} \mathbf{1}(S \subseteq H, i \in S)
\end{aligned}$$

Then, RHS is

$$\begin{aligned}
&\mathbb{E}_b\left[\frac{n}{m\binom{n}{p}|Iso(R)|}\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i\in S)\Big|G\right]\\
&=\frac{n}{m\binom{n}{p}|Iso(R)|}\sum_{H\subseteq G}\mathbb{P}(H)\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i\in S)\\
&=\frac{n}{m\binom{n}{p}|Iso(R)|}\sum_{H\subseteq G_d}\frac{1}{n}\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i\in S)\\
&=\frac{n}{m\binom{n}{p}|Iso(R)|}\frac{1}{n}\sum_{S\subseteq K_n, S\cong R}\sum_{H\subseteq G_d, H\supseteq S}\mathbf{1}(S\subseteq G, i\in S)\\
&=\frac{n}{m\binom{n}{p}|Iso(R)|}\frac{m}{n}\sum_{S\subseteq K_n, S\cong R}\mathbf{1}(S\subseteq G)\\
&=\frac{1}{\binom{n}{p}|Iso(R)|}\sum_{S\subseteq K_n, S\cong R}\mathbf{1}(S\subseteq G)
\end{aligned}$$

So, we have,

$$\mathbb{E}_b[T_{b3}(R)|G]=T_G(R)$$

(ii) Here, we use properties of the underlying model. Let us condition on $\boldsymbol{\xi}=\{\xi_1,\ldots,\xi_n\}$ and the whole graph G separately. Now, conditioning on $\boldsymbol{\xi}$, we get the main term of $T_G(R)$ to be,

$$\mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})=\frac{1}{\binom{n}{p}|Iso(R)|}\sum_{S\subseteq K_n, S\cong R}\left(\prod_{(i,j)\in E(S)}w(\xi_i,\xi_j)\right)+O(n^{-1}\lambda_n). \tag{4.29}$$

We shall use the same decomposition as used in [23] of $(\rho_n^{-e}\tilde{T}_{b3}(R)-\tilde{P}(R))$ into

$$\begin{aligned}
(\rho_n^{-e}\tilde{T}_{b3}(R)-\tilde{P}(R)) &= \rho_n^{-e}\left(\tilde{T}_{b3}-\mathbb{E}_b[T_{b3}(R)|G]\right)\\
&\quad+\rho_n^{-e}\left(T_G(R)-\mathbb{E}(T_G(R)|\boldsymbol{\xi})\right)\\
&\quad+\mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-e}-\tilde{P}(R)
\end{aligned}$$

Let us define,

$$\begin{aligned}
U_3 &= \mathbb{E}(\hat{P}(R)|\boldsymbol{\xi})\rho_n^{-e}-\tilde{P}(R)\\
U_2 &= \rho_n^{-e}\left(T_G(R)-\mathbb{E}(T_G(R)|\boldsymbol{\xi})\right)\\
U_1 &= \rho_n^{-e}\left(\tilde{T}_{b3}-\mathbb{E}_b[T_{b3}(R)|G]\right)
\end{aligned}$$

Now, it is easy to see that

$$\begin{aligned}\mathrm{Var}(\rho^{-e}\tilde{T}_{b3}(R)) &= \mathbb{E}(\mathrm{Var}(\rho^{-e}\tilde{T}_{b3}(R)|G) + \mathrm{Var}(\mathbb{E}(\rho^{-e}\tilde{T}_{b3}(R)|G)) \\ &= \mathbb{E}(\mathrm{Var}(\rho^{-e}\tilde{T}_{b2}(R) - T_G(R)|G) + \mathrm{Var}(T_G(R)) \\ &= \mathbb{E}(\mathrm{Var}(U_1|G)) + \mathbb{E}(\mathrm{Var}(T_G(R)|\boldsymbol{\xi})) + \mathrm{Var}(\mathbb{E}(T_G(R)|\boldsymbol{\xi})) \\ &= \mathbb{E}(\mathrm{Var}(U_1|G)) + \mathbb{E}(\mathrm{Var}(U_2|\boldsymbol{\xi})) + \mathrm{Var}(U_3)\end{aligned}$$

We shall try to see the behavior of $\mathrm{Var}(U_1|G) = \mathrm{Var}_b[\rho^{-e}\tilde{T}_{b3}(R)|G]$. Now,

$$\mathrm{Var}_b[\rho^{-e}\tilde{T}_{b3}(R)|G] = \rho^{-2e}\frac{1}{B^2}\left(\sum_{b=1}^{B}\mathrm{Var}_b[T_{b3}(R)] + \sum_{b,b'=1,b\neq b'}^{B}\mathrm{Cov}_b(T_{b3}(R), T_{b'3}(R))\right)$$

Now, the formula for $\mathrm{Var}_b[T_{b3}(R)]$ is given in Appendix A3. Note that, if $E(H_b) \cap E(H_{b'}) = \phi$, then, $\mathrm{Cov}_b(T_{b3}(R), T_{b'3}(R)) = 0$. The number of pairs such that $E(H_b) \cap E(H_{b'}) \neq \phi$ is depends on the density of the nodes and size of R. If we have $\lambda^{|V(R)|}$ and if $\lambda^{|V(R)|} = o(n)$ and also, the number of edges for the leading term in the covariance is equal to or more than $2e$. So,

$$\mathbb{E}(\mathrm{Var}(U_1|G)) = o(n^{-1})$$

Now, by proof of Theorem 1 in [23], we have,

$$\begin{aligned}\mathrm{Var}(U_2) &= o(n^{-1}) \\ \mathrm{Var}(U_3) &= o(n^{-1})\end{aligned}$$

So, we get, $\mathrm{Var}(\rho^{-2e}\tilde{T}_{b3}(R)) = o(n^{-1})$. Since, we already know $\sqrt{n}$-consistency of $\left(\rho_n^{-e}T_G(R) - \tilde{P}(R))\right)$, this proves the $\sqrt{n}$-consistency of $\rho_n^{-e}\tilde{T}_{b3}(R)$ to $\rho_n^{-e}T_G(R)$.

□

The variance of $\tilde{T}_{b2}(R)$ given G can also be calculated and is given in the Appendix.

A3. Variance of $\hat{T}_{b2}(R)$

The variance of $T^b(R)$ is

$$\mathrm{Var}_b\left[\frac{n_R}{\binom{n}{p}|Iso(R)|}\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right]$$

$$= \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2 \mathrm{Var}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right]$$

$$\begin{aligned}
&\mathrm{Var}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right]\\
&=\mathbb{E}_b\left[\left(\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\right)^2\Big|G\right]\\
&-\left(\mathbb{E}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right]\right)^2
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_b\left[\left(\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\right)^2\Big|G\right]\\
&=\mathbb{E}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right]\\
&-\mathbb{E}_b\left[\sum_{\substack{S,T\subseteq K_m\\ S,T\cong R, S\neq T}}\mathbf{1}(S,T\subseteq H, i \text{ central vertex of } S,T)\Big|G\right]\\
&= I - II \text{ (Suppose)}
\end{aligned}$$

Thus,

$$I=\sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)$$

$$II = \mathbb{E}_b\left[\sum_{\substack{S,T\subseteq K_m\\ S,T\cong R, S\neq T}}\mathbf{1}(S,T\subseteq H, i \text{ central vertex of } S,T)\Big|G\right]$$

Now, a host of subgraphs can be formed by the intersection of two copies of R, having the same central vertex i. The number of intersected vertices can range from 1 to d, the depth of the graph R. Let us consider, that for number of vertices in intersection as k $(k = 1,\ldots,(p-1))$, the number of graph structures that can be formed is d_k and we represent that graph structure by R_{jk}, where, $j = 1,\ldots,d_k$. Thus,

$$II = \sum_{k=1}^{d}\sum_{j=1}^{d_k}\sum_{S\subseteq K_n, S\cong R_{jk}}\frac{1}{n_{R_{jk}}}\mathbf{1}(S\subseteq G)$$

So,

$$\mathbb{E}_b\left[\left(\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H)\right)^2\Big|G\right] = \sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)+\sum_{k=1}^{d}\sum_{j=1}^{d_k}\sum_{S\subseteq K_n, S\cong R_{jk}}\frac{1}{n_{R_{jk}}}\mathbf{1}(S\subseteq G)$$

$$\begin{aligned}\mathrm{Var}_b\left[\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right] &= \sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)\\ &+ \sum_{k=1}^{d}\sum_{j=1}^{d_k}\sum_{S\subseteq K_n, S\cong R_{jk}}\frac{1}{n_{R_{jk}}}\mathbf{1}(S\subseteq G)\\ &- \left(\sum_{S\subseteq K_n, S\cong R}\frac{1}{n_H}\mathbf{1}(S\subseteq G)\right)^2\end{aligned}$$

$$\begin{aligned}&\mathrm{Var}_b\left[\frac{n_R}{\binom{n}{p}|Iso(R)|}\sum_{S\subseteq K_m, S\cong R}\mathbf{1}(S\subseteq H, i \text{ central vertex of } S)\Big|G\right]\\ &= \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2\left[\sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)\right]\\ &- \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2\left[\left(\sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)\right)^2\right]\\ &+ \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2\left[\sum_{k=1}^{d}\sum_{j=1}^{d_k}\sum_{S\subseteq K_n, S\cong R_{jk}}\frac{1}{n_{R_{jk}}}\mathbf{1}(S\subseteq G)\right]\end{aligned}$$

So,

$$\begin{aligned}\mathrm{Var}_b\left[T^b(R)\right] &= \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2\left[\sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)\right]\\ &- \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2\left[\left(\sum_{S\subseteq K_n, S\cong R}\frac{1}{n_R}\mathbf{1}(S\subseteq G)\right)^2\right]\\ &+ \left(\frac{n_R}{\binom{n}{p}|Iso(R)|}\right)^2\left[\sum_{k=1}^{d}\sum_{j=1}^{d_k}\sum_{S\subseteq K_n, S\cong R_{jk}}\frac{1}{n_{R_{jk}}}\mathbf{1}(S\subseteq G)\right]\end{aligned}$$

Chapter 5

Community Detection in Networks using Graph Distance

5.1 Introduction

The study of networks has received increased attention recently not only from the social sciences and statistics but also from physicists, computer scientists and mathematicians. With the information boom, a huge number of network data sets have come into prominence. In biology - gene transcription networks, protein-protein interaction network, in social media - Facebook, Twitter, Linkedin networks, information networks arising in connection with text mining, technological networks such as the Internet, ecological and epidemiological networks and many others have appeared. Although the study of networks has a long history in physics, social sciences and mathematics literature and informal methods of analysis have arisen in many fields of application, statistical inference on network models as opposed to descriptive statistics, empirical modeling and some Bayesian approaches [128] [98] [79] has not been addressed extensively in the literature. A mathematical and systematic study of statistical inference on network models has only started in recent years.

One of the fundamental questions in analysis of such data is detecting and modeling community structure within the network. A lot of algorithmic approaches to community detection have been proposed, particularly in the physics and computer science literature [130] [109] [63]. In terms of community detection, there are two different goals that researchers have tried to pursue -

- **Algorithmic Goal:** Identify the community each vertex of the network belongs to.

- **Theoretical Goal:** If the network is generated by an underlying generative model, then, what is the probability of success for the algorithm.

Algorithms

Several popular algorithms for community detection have been proposed in physics, computer science and statistics literature. Most of these algorithms show decent performance in community detection for selected real-world and simulated networks [102] and have polynomial time complexity. We shall briefly mention some of these algorithms.

1. Modularity maximizing methods [131]. One of the most popular method of community detection. The problem is NP hard but spectral relaxations of polynomial complexity exist [129].

2. Hierarchical clustering techniques [43].

3. Spectral clustering based methods [116] [44], [140] [40]. These methods are also very popular. Most of the time these methods have linear or polynomial running times. Mostly shown to work for dense graphs only.

4. Profile likelihood maximization [22]. The problem is NP hard, but heuristic algorithms have been proposed, which have good performance for dense graphs.

5. Stochastic Model based methods:

   - MCMC based likelihood maximization by Gibbs Sampling, the cavity method and belief propagation based on stochastic block model. [50]
   - Variational Likelihood Maximization based on stochastic block model [37], [21]. Polynomial running time but appears to work only for dense graphs.
   - Pseudo-likelihood Maximization [41]. Fast method which works well for both dense and sparse graphs. But the method is not fully justified.
   - Model-based:
     (a) Mixed Membership Block Model [3]. Iterative method and works for dense graphs. The algorithm for this model is based on variational approximation of the maximum likelihood estimation.
     (b) Degree-corrected block model [91]: Incorporates degree inhomogeneity in the model. Algorithms based on maximum likelihood and profile likelihood estimation has been developed.
     (c) Overlapping stochastic block model [104]: Stochastic block model where each vertex can lie within more than one community. The algorithm for this model is based on variational approximation of the maximum likelihood estimation.
     (d) Mixed configurations model [9]: Another extension to degree-corrected stochastic block model, where, the model is a mixture of configurations model (degree-corrected block model with one block) and each vertex can lie in more than one community. The algorithm for this model is based on the EM algorithm and maximum likelihood estimation.

6. Model based clustering [75].

Theoretical Goal

The stochastic block model (SBM) is perhaps the most commonly used and best studied model for community detection. An SBM with Q blocks states that each node belongs to a community $\boldsymbol{c} = (c_1, \ldots, c_n) \in \{1, \ldots, Q\}$ which are drawn independently from the multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$, where $\pi_i > 0$ for all i, and Q is the number of communities, assumed known. Conditional on the labels, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with

$$\mathbb{E}[A_{ij}|\boldsymbol{c}] = P_{c_i c_j}, \tag{5.1}$$

where $P = [P_{ab}]$ and $K = [K_{ab}]$ are $Q \times Q$ symmetric matrix. P can be considered the *connection probability* matrix, where as K is the *kernel* matrix for the connection. So, we have $P_{ab} \leq 1$ for all $a, b = 1, \ldots, Q$, $P\mathbf{1} \leq \mathbf{1}$ and $\mathbf{1}^T P \leq \mathbf{1}$ element-wise. The network is undirected, so $A_{ji} = A_{ij}$, and $A_{ii} = 0$ (no self-loops). The problem of community detection is then to infer the node labels $\boldsymbol{c}$ from A. Thus we are not really interested in estimation or inference on parameters $\boldsymbol{\pi}$ and P, but, rather we are interested in estimating $\boldsymbol{c}$. But, it does not mean the two problems are mutually exclusive. In reality, the inferential problem and the community detection problem are quite interlinked.

The theoretical results of community detection for stochastic block models can be divided into 3 different regimes -

(a) $\frac{\mathbb{E}(\text{degree})}{\log n} \to \infty$, equivalent to, $\mathbb{P}[\text{there exists an isolated point}] \to 0$.

(b) $\mathbb{E}(\text{degree}) \to \infty$, which means existence of giant component, but also presence of isolated small components from Theorem 5.2.7.

(c) If $\mathbb{E}(\text{degree}) = O(1)$, phase boundaries exist, below which community identification is not possible.

Note:

(a) All of the above mentioned algorithms perform satisfactorily on regime (a).

(b) None of the above algorithms have been shown to have near perfect probability of success under either regime (b) or (c), for the full parameter space. Some algorithms like [44] [22] [40] [41] are shown to partially work in the sparse setting. Some very recent algorithms include [100] [133].

In this paper, we shall only concentrate on stochastic block models. In the future, we shall try to extend our method and results for more general models.

Contributions and Outline of the Chapter

In real life networks, most of the time we seem to see moderately sparse networks [107] [111] [110]. Most of the large or small complex networks we see seem to fall in the (b) regime of Section 5.1 we describe before, that is, $\mathbb{E}(\text{degree}) \to \infty$. We propose a simple algorithm, which performs well in practice in both regimes (b) and (c) and has some theoretical backing If degree distribution can identify block parameters then classification using our method should give reasonable result in practice.

Our algorithm is based on graph distance between vertices of the graph. We perform spectral clustering based on the graph distance matrix of the graph. By looking at the graph distance matrix instead of adjacency matrix for spectral clustering increases the performance of the community detection, as the normalized distance between cluster centers increases when we go from the adjacency matrix to the graph distance matrix. This helps in community detection even for sparse matrices. We only show theoretical results for stochastic block models. The theoretical proofs are quite intricate and involve careful coupling of the stochastic block model with multi-type branching process to find asymptotic distribution of the typical graph distances. Then, a careful analysis of the eigenvector of the asymptotic graph distance matrix reveals the existence of separation needed for spectral clustering to succeed. This method of analysis has been used for spectral clustering analysis using the adjacency matrix also [149], but the analysis is simpler.

The rest of the chapter is organized as follows. We give a summary of the preliminary results needed in Section 5.2. We present the algorithms in Section 5.3. We give an outline of proof of theoretical guarantee of performance of the method and then the details in Section 5.4. The numerical performance of the methods is demonstrated on a range of simulated networks and on some real world networks in Section 5.5. Section 5.6 concludes with discussion, and the Appendix contains some additional technical results.

5.2 Preliminaries

Let us suppose that we have a random graph G_n as the data. Let $V(G_n) = \{v_i, \ldots, v_n\}$ denote the vertices of G_n and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of G_n. So, the number of vertices in G_n is $|V(G_n)| = n$ and number of edges of G_n is $|E(G_n)| = m$. Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. For the sake of notational simplicity, from here onwards we shall denote G_n by G having n vertices unless specifically mentioned. We consider the n vertices of G are clustered into Q different communities with each community having size n_a, $a = 1, \ldots, Q$ and $\sum_a n_a = n$. In this paper, we are interested in the problem of *vertex community identification* or *graph partitioning*. That means that we are interested in finding which of the Q different community each vertex of G belongs to. However, the problem is an *unsupervised learning* problem. So, we assume that the data is coming from an underlying model and we try to verify how good 'our' *community detection* method works for that model.

Model for Community Detection

As a model for community detection, we consider the stochastic block model. We shall define the stochastic block model shortly, but, we first we shall introduce some more general models, of which stochastic block model is a special case.

Bickel-Chen Model

The general non-parametric model, as described in Bickel, Chen and Levina (2011) [23], that generates the random data network G can be defined by the following equation -

$$\mathbb{P}(A_{ij} = 1|\xi_i = u, \xi_j = v) = h_n(u, v) = \rho_n w(u, v)\mathbf{1}(w \leq \rho_n^{-1}), \tag{5.2}$$

where, $w(u, v) \geq 0$, symmetric, $0 \leq u, v \leq 1$, $\rho_n \to 0$. For block models, the latent variable for each vertex $(\xi_1, \ldots, \xi_n)$ can be considered to be coming from a discrete and finite set. Then, each element of that set can be considered to be inducing a partition in the vertex set $V(G_n)$. Thus, we get a model for vertex partitioning, where, the set of vertices can be partitioned into finite number of disjoint classes, but however the partition to which each vertex belongs to is the latent variable in the model and thus unknown. The main goal becomes estimating this latent variable.

Inhomogeneous Random Graph Model

The inhomogeneous random graph model (IRGM) was introduced in Bollobás et. al. (2007) [29]. Let $\mathcal{S}$ be a separable metric space equipped with a Borel probability measure μ. For most cases $\mathcal{S} = (0, 1]$ with μ Lebesgue measure, that means a $U(0, 1)$ distribution. The "kernel" κ will be a symmetric non-negative function on $\mathcal{S} \times \mathcal{S}$. For each n we have a deterministic or random sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ of points in $\mathcal{S}$. Writing δ_x for the measure consisting of a point mass of weight 1 at x, and

$$\nu_n \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

for the empirical distribution of $\boldsymbol{x}$, it is assumed that ν_n converges in probability to μ as $n \to \infty$, with convergence in the usual space of probability measures on $\mathcal{S}$. One example where the convergence holds is the random case, where the x_i are independent and identically distributed on $\mathcal{S}$ with distribution μ convergence in probability holds by the law of large numbers. Of course, we do not need $(x_n)_{n\geq 1}$ to be defined for every n, but only for an infinite set of integers n. From here onwards, we shall only focus on this special case, where, $(x_1, \ldots, x_n) \overset{iid}{\sim} \mu$.

Definition 5.2.1. *A **kernel** κ_n on a ground space $(\mathcal{S}, \mu)$ is a symmetric non-negative (Borel) measurable function on $\mathcal{S} \times \mathcal{S}$. κ is also continuous a.e. on $\mathcal{S} \times \mathcal{S}$. By a kernel on a vertex space $(\mathcal{S}, \mu, (x_n)_{n\geq 1})$ we mean a kernel on $(\mathcal{S}, \mu)$.*

Given the (random) sequence $(x_1, \ldots, x_n)$, we let $G(n, \kappa)$ be the random graph $G(n, (p_{ij}))$ with

$$p_{ij} \equiv \min\{\kappa(x_i, x_j)/n, 1\}. \tag{5.3}$$

In other words, $G^{\mathcal{V}}(n, \kappa)$ has n vertices $\{1, \ldots, n\}$ and, given $x_1, \ldots, x_n$, an edge ij (with $i \neq j$) exists with probability p_{ij}, independently of all other (unordered) pairs ij. Based on the graph kernel we can also define an integral operator T_κ in the following way

Definition 5.2.2. *The **integral operator** $T_\kappa : L^2(\mathcal{S}) \to L^2(\mathcal{S})$ corresponding to $G(n, \kappa)$, is defined as*

$$T_\kappa f(x)(\cdot) = \int_0^1 \kappa(x, y) f(y) d\mu(y),$$

where, $x \in \mathcal{S}$ and any measurable function $f \in L^1(\mathcal{S})$.

The random graph $G(n, \kappa)$ depends not only on κ but also on the choice of $x_1, \ldots, x_n$. The freedom of choice of x_i in this model gives some more flexibility than Bickel-Chen model. The asymptotic behavior of $G(n, \kappa)$ depend very much on $\mathcal{S}$ and μ. Many of these key results such as existence of giant component, typical distance, phase transition properties are proved in [29]. We shall use these results on inhomogeneous random graphs in order to prove results on graph distance for stochastic block models.

Here is further comparison of the Inhomogeneous random graph model (IRGM) with the Bickel-Chen model (BCM), to understand their similarities and dissimilarities -

(a) In BCM, $(\xi_1, \ldots, \xi_n) \overset{iid}{\sim} U(0, 1)$ are the latent variables associated with the vertices $(v_1, \ldots, v_n)$ of random graph G_n. Similarly, in IRGM, $(x_1, \ldots, x_n) \sim \mu$ are the latent variables associated with the vertices $(v_1, \ldots, v_n)$ of random graph G_n. Now, if in IRGM, $(x_1, \ldots, x_n) \overset{iid}{\sim} \mu$ then the latent variable structure of the two models become equivalent.

(b) In BCM, the conditional probability of connection between two vertices given the value of their latent variables is controlled by the kernel function $h_n(u, v)$. In IRGM, the conditional probability of connection between two vertices given the value of their latent variables is controlled by the kernel function $\frac{\kappa(u,v)}{n}$.

(c) So, if $h_n(u, v) = \kappa(u, v)/n$, $\mathcal{S}[(0, 1)$ and the underlying measure spaces are same and the measure μ is a uniform measure on interval $\mathcal{S} = (0, 1)$, then, BCM and IRGM generates graphs from the same distribution. In fact, as noted in [22], if $\mathcal{S} = \mathbb{R}$ and μ has a positive density with respect to Lebesgue measure, then the (limiting) IRGM is equivalent to Bickel-Chen model with suitable h_n.

(d) For IRGM, let us define

$$\lambda \equiv ||T_\kappa|| \equiv \sup_{f \in L^2(\mathcal{S}), ||f||_{L^2(\mathcal{S})}=1} \int_{\mathcal{S}} \int_{\mathcal{S}} \kappa(u,v) f(u) f(v) d\mu(u) d\mu(v),$$

where, T_κ is the operator define in Definition 5.2.2 and $||\cdot||$ is the operator norm. In BCM,

$$\rho_n \equiv \int_0^1 \int_0^1 h_n(u,v) du dv.$$

If BCM and IRGM have same underlying measure spaces ($\mathcal{S} = (0,1), \mu = U(0,1)$) and $h_n(u,v) = \kappa(u,v)/n$ and

Case 1: $\mathbf{1}$ is the principal eigenfunction of T_κ, then

$$n\rho_n \to \lambda$$

where, λ is as defined above.

Case 2: $\mathbf{1}$ is not the principal eigenfunction of T_κ, then

$$n\rho_n \leq \lambda$$

In case of BCM $n\rho_n$ is the natural scaling parameter for the random graph, since, $\mathbb{E}[\text{Number of Edges in } G_n] = \frac{1}{2} n\rho_n$. In case of IRGM, λ is fixed. However, we shall see that the limiting behavior of the graph distance between two vertices of the network becomes dependent on the parameter λ. So, the parameter λ still remains of importance. We shall henceforth focus on IRGM, with parameter of importance being λ

Stochastic Block Model

The stochastic block model is perhaps the most commonly used and best studied model for community detection. We continue with IRGM framework, so the graph is sparse.

Definition 5.2.3. *A graph $G^Q(,(P,\boldsymbol{\pi}))$ generated from **stochastic block model** (SBM) with Q blocks and parameters $P \in (0,1)^{Q\times Q}$ and $\boldsymbol{\pi} \in (0,1)^Q$ can be defined in following way - each vertex of graph G_n from an SBM belongs to a community $\boldsymbol{c} = (c_1, \ldots, c_n) \in \{1, \ldots, Q\}$ which are drawn independently from the multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$, where $\pi_i > 0$ for all i. Conditional on the labels, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with*

$$\mathbb{E}[A_{ij}|\boldsymbol{c}] = P_{c_i c_j} = \min\{\frac{K_{c_i c_j}}{n}, 1\}, \tag{5.4}$$

*where $P = [P_{ab}]$ and $K = [K_{ab}]$ are $Q \times Q$ symmetric matrices. P is known as the **connection probability** matrix and K as the **kernel** matrix for the connection. So, we have $P_{ab} \leq 1$ for all $a, b = 1, \ldots, Q$, $P\mathbf{1} \leq \mathbf{1}$ and $\mathbf{1}^T P \leq \mathbf{1}$ element-wise.*

The network is undirected, so $A_{ji} = A_{ij}$, and $A_{ii} = 0$ (no self-loops). The problem of community detection is then to infer the node labels $\boldsymbol{c}$ from A. Thus we are not really interested in estimation or inference on parameters $\boldsymbol{\pi}$ and P, but, rather we are interested in estimating $\boldsymbol{c}$. But, it does not mean the two problems are mutually exclusive, in reality, the inferential problem and the community detection problem are quite interlinked.

We can see that SBM is a special case of both Bickel-Chen model and IRGM. In IRGM, if we consider $\mathcal{S}$ to be a finite set, $(x_1, \ldots, x_n) \in [Q]^n$ ($[Q] = \{1, \ldots, Q\}$) with $x_i \overset{iid}{\sim} Mult(n, \boldsymbol{\pi})$ and kernel $\kappa : [Q] \to [Q]$ as $\kappa(a, b) = K_{ab}$ $(a, b = 1, \ldots, Q)$, then the resulting IRGM graph follows stochastic block model. So, for SBM we can define an *integral operator* on $[Q]$ with measure $\{\pi_1, \ldots, \pi_Q\}$.

Definition 5.2.4. *The **integral operator** $T_K : \ell^1(\mathcal{S}) \to \ell^1(\mathcal{S})$ corresponding to $G^Q(n, (P, \boldsymbol{\pi}))$, is defined as*

$$(T_K(x))_a = \sum_{b=1}^{Q} K_{ab}\pi_b x_b, \text{ for } a = 1, \ldots, Q$$

where, $x \in \mathbb{R}^Q$.

The stochastic block model has deep connections with Multi-type branching process, just as, Erodös-Rényi random graph model (ERRGM) has connections with the branching process. Let us introduce branching process first.

Multi-type Branching Process

We shall try to link network formed by SBM with the tree network generated by multi-type Galton-Watson branching process. In our case, the Multi-type branching process (MTBP) has type space $S = \{1, \ldots, Q\}$, where a particle of type $a \in S$ is replaced in the next generation by a set of particles distributed as a Poisson process on S with intensity $(K_{ab}\pi_b)_{b=1}^{Q}$. We denote this branching process, started with a single particle of type a, by $\mathcal{B}_{K,\pi}(a)$. We write $\mathcal{B}_{K,\pi}$ for the same process with the type of the initial particle random, distributed according to $\boldsymbol{\pi}$.

Definition 5.2.5. *(a) Define $\rho_k(K, \pi; a)$ as the probability that the branching process $\mathcal{B}_{K,\pi}(a)$ has a total population of exactly k particles.*

(b) Define $\rho_{\geq k}(K, \pi; a)$ as the probability that the total population is at least k.

(c) Define $\rho(K, \pi; a)$ as the probability that the branching process survives for eternity.

(d) Define,

$$\rho_k(K, \pi) \equiv \sum_{a=1}^{Q} \rho_k(K, \pi; a)\pi_a, \quad \rho \equiv \rho(K, \pi) \equiv \sum_{a=1}^{Q} \rho(K, \pi; a)\pi_a \tag{5.5}$$

and define $\rho_{\geq k}(K)$ analogously. Thus, $\rho(K,\pi)$ is the ***survival probability*** *of the branching process $\mathcal{B}_{K,\pi}$ given that its initial distribution is $\boldsymbol{\pi}$*

If the probability that a particle has infinitely many children is 0, then $\rho(K,\pi;a)$ is equal to $\rho_\infty(a)$, the probability that the total population is infinite. As we shall see later, the branching process $\mathcal{B}_{K,\pi}(a)$ arises naturally when exploring a component of G_n starting at a vertex of type a; this is directly analogous to the use of the single-type Poisson branching process in the analysis of the Erdös-Rényi graph $G(n,c/n)$.

Known Results for Stochastic Block Model

The performance of community detection algorithms depends on the parameters $\boldsymbol{\pi}$ and P. We refer to Definition 5.2.3 for definition of stochastic block models. An important condition that we usually put on parameter P is *irreducibility*.

Definition 5.2.6. *A connection matrix P on a $\mathcal{S}=\{1,\dots,Q\}$ is* ***reducible*** *if there exists $A\subset\mathcal{S}$ with $0<|A|<Q$ such that $P=0$ a.e. on $A\times(\mathcal{S}-A)$; otherwise P is* ***irreducible****. Thus P is* ***irreducible*** *if $A\subseteq\mathcal{S}$ and $P=0$ a.e. on $A\times(\mathcal{S}-A)$ implies $|A|=0$ or $|A|=Q$.*

So, the results on existence of giant components in [29] also apply for SBM. The following theorem describes the result on existence of giant components.

Theorem 5.2.7 ([29])**.** *Let us define operator T_K as in definition 5.2.4,*

(i) If $||T_K||\leq 1$ ($||\cdot||$ refer to operator norm), then the size of largest component is $o_P(n)$, while if $||T_K||>1$, then the size of largest component is $\Theta_P(n)$ whp.

(ii) If P is irreducible, then $\frac{1}{n}$(Size of largest component) $\to \boldsymbol{\pi}^T\rho$, where, $\rho\in[0,1]^Q$ is the survival probability as defined in (5.5).

The theoretical results on community detection depend on the 3 different regime on which the generative model is based on -

(a) $\frac{\mathbb{E}(\text{degree})}{\log n}\to\infty$, equivalent to, $\mathbb{P}[\text{there exists an isolated point}]\to 0$. In this setting, there are several algorithms, such as those described in Section 1, can identify correct community with high probability under quite relaxed conditions on parameters P and $\boldsymbol{\pi}$. See [40] (Theorem 2 and 3), [140] (Theorem 3.1), [44] (Theorem 1).

(b) $\mathbb{E}(\text{degree})\to\infty$, which means existence of giant component, but also presence of isolated small components from Theorem 5.2.7. In this setting, algorithms proposed in [44], [41] is proved to identify community labels that are highly correlated with original community labels with high probability.

(c) If $\mathbb{E}(\text{degree})=O(1)$, phase boundaries exist, below which community identification is not possible. These results and rigorous proof are given in [125]. The results can be summarized for 2-block model with parameters $P_{11}=a, P_{12}=b, P_{22}=a$ as

Theorem 5.2.8 ([125]). *(i) If $(a-b)^2 < 2(a+b)$ then probability model of SBM and ERRGM with $p = \frac{a+b}{2n}$ are mutually contiguous. Moreover, if $(a-b)^2 < 2(a+b)$, there exists no consistent estimators of a and b.*

(ii) If $(a-b)^2 > 2(a+b)$ then probability model of SBM and ERRGM with $p = \frac{a+b}{2n}$ are asymptotically orthogonal.

So, in the range $(a-b)^2 > 2(a+b)$, there should exists an algorithm which identifies highly correct clustering with high probability at least within the giant components.

5.3 Algorithm

The algorithm we propose depend on the graph distance or geodesic distance between vertices in a graph.

Definition 5.3.1. ***Graph distance*** *or* ***Geodesic distance*** *between two vertices i and j of graph G is given by the length of the shortest path between the vertices i and j, if they are connected. Otherwise, the distance is infinite.*

So, for any two vertices $u, v \in V(G)$, graph distance, d_g is defined by

$$d_g(u,v) = \begin{cases} |V(e)|, & \text{if } e \text{ is the shortest path connecting } u \text{ and } v \\ \infty, & u \text{ and } v \text{ are not connected} \end{cases}$$

For sake of numerical convenience, we shall replace ∞ by a large number for value of $d_g(u,v)$, when, u and v are not connected. The main steps of the algorithm can be described as follows

1. Find the graph distance matrix $D = [d_g(v_i, v_j)]_{i,j=1}^n$ for a given network but with distance upper bounded by $k \log n$. Assign non-connected vertices an arbitrary high value B.

2. Perform hierarchical clustering to identify the giant component G^C of graph G. Let $n_C = |V(G^C)|$.

3. Normalize the graph distance matrix on G^C, D^C by

$$\bar{D}^C = -\left(I - \frac{1}{n_C}\mathbf{1}\mathbf{1}^T\right)(D^C)^2\left(I - \frac{1}{n_C}\mathbf{1}\mathbf{1}^T\right)$$

4. Perform eigenvalue decomposition on $\bar{D}^C$.

5. Consider the top Q eigenvectors of normalized distance matrix $\bar{D}^C$ and $\tilde{\mathbf{W}}$ be the $n \times Q$ matrix formed by arranging the Q eigenvectors as columns in $\tilde{\mathbf{W}}$. Perform Q-means clustering on the rows $\tilde{\mathbf{W}}$, that means, find an $n \times Q$ matrix $\mathbf{C}$, which has Q distinct rows and minimizes $||\mathbf{C} - \tilde{\mathbf{W}}||_F$.

6. (Alternative to 5.) Perform Gaussian mixture model based clustering on the rows of $\tilde{\mathbf{W}}$, when there is an indication of highly-varying average degree between the communities.

7. Let $\hat{\xi} : V \mapsto [Q]$ be the block assignment function according to the clustering of the rows of $\tilde{\mathbf{W}}$ performed in either Step 5 or 6.

Here are some important observations about the algorithm -

(a) There are standard algorithms for graph distance finding in the algorithmic graph theory literature. In algorithmic graph theory literature the problem is known as the **all pairs shortest path** problem. The two most popular algorithms are Floyd-Warshall [62] [160] and Johnson's algorithm [86]. The time complexity of the Floyd-Warshall algorithm is $O(n^3)$, where as, the time complexity of Johnson's algorithm is $O(n^2 \log n + ne)$ [106] ($n = |V(G_n)|$ and $e = |E(G_n)|$). So, for sparse graphs, Johnson's algorithm is faster than Floyd-Warshall. Memory storage is also another issue for this algorithm, since the algorithm involves a matrix multiplication step of complexity $\Omega(n^2)$. Recently, there also has been some progress on parallel implementation of all-pairs shortest path problem [146] [33] [72], which addresses both memory and computation aspects of the algorithm and lets us scale the algorithm for large graphs, both dense and sparse.

(b) The Step 3 of the algorithm is nothing but the classical multi-dimensional scaling (MDS) of the graph distance matrix. In MDS, we try to find vectors $(x_1, \ldots, x_n)$, where, $x_i \in \mathbb{R}^Q$, such that,

$$\sum_{i,j=1}^{n} \left(||x_i - x_j||_2 - (D^C)_{ij} \right)^2$$

is minimized. The minimizer is attained by the rows of the matrix formed by the top Q eigenvectors of $\bar{D}^C$ as columns. So, performing spectral clustering on $\bar{D}^C$ is the same as performing Q-means clustering on the multi-dimensional scaled space.

Instead of $\bar{D}^C$, we could also use the matrix $(D^C)^2$, but then, the topmost eigenvector does not carry any information about the clustering. Similarly, we can also use the matrix D^C directly for spectral clustering, but, in that case, D^C is not a positive semi-definite matrix and as a result we have to consider the eigenvectors corresponding to largest absolute eigenvalues (since eigenvalues can be negative).

(c) In the Step 5 of the algorithm Q-means clustering if the expected degree of the blocks are equal. However, if the expected degree of the blocks are different, it leads to multi scale behavior in the eigenvectors of the normalized distance matrix. So, we perform Gaussian Mixture Model (GMM) based clustering instead of Q-means to take into account the multi scale behavior.

5.4 Theory

Let us consider that we have a random graph G_n as the data. Let $V(G_n) = \{v_i, \ldots, v_n\}$ denote the vertices of G_n and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of G_n. So, the number of vertices in G_n is $|V(G_n)| = n$ and number of edges of G_n is $|E(G_n)| = m$. Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. For sake of notational simplicity, from here onwards we shall denote G_n by G having n vertices unless specifically mentioned. There are Q communities for the vertices and each community has $(n_a)_{a=1}^{Q}$ number of vertices. In this paper, we are interested in the problem of *vertex community identification* or *graph partitioning*. However, the problem is an *unsupervised learning* problem. So, we assume that the data is coming from an underlying model and we try to verify how good 'our' *community detection* method works for that model.

The theoretical analysis of the algorithm has two main parts -

I. Finding the limiting distribution of graph distance between two typical vertices of type a and type b (where, $a, b = 1, \ldots, Q$). This part of the analysis is highly dependent on results from multi-type branching processes and their relation with stochastic block models. The proof techniques and results are borrowed from [29], [18] and [7].

II. Finding the behavior of top Q eigenvectors of the graph distance matrix D using the limiting distribution of the typical graph distances. This part of analysis is highly dependent on perturbation theory of linear operators. The proof techniques and results are borrowed from [93], [38] and [149].

Results of Part I

We shall give limiting results for *typical distance* between vertices in G_n. If u and $v \in V(G_n)$ are two vertices in G_n, which has been selected uniformly at random from type a and type b respectively, where, $a, b = 1, \ldots, Q$ are the different communities. Then, the graph distance between u and v is $d_G(u, v)$. Now, the operator that controls the process is T_K as defined in Definition 5.2.4. T_K is another representation of the matrix $\tilde{K}_{Q\times Q}$, which is defined as

$$\tilde{K}_{ab} \equiv \pi_a K_{ab} \pi_b, \text{ for } a, b = 1, \ldots, Q \tag{5.6}$$

The matrix $\tilde{K}$ defines the quadratic form for $T_K : \ell^1(\mathcal{S}, \pi) \to \ell^1(\mathcal{S}, \pi)$. So, we have that

$$\lambda \equiv ||T_K|| = \lambda_{max}(\tilde{K}). \tag{5.7}$$

The relation between λ and $\mathbb{E}$[number of Edges in G_n] is given Section 5.2. Here, we use λ as the scaling operator, not either average, minimum or maximum degree of vertices as used in [149] and [140]. But, we already know that, if the graph is *homogeneous*, then, $\mathbb{E}$[number of Edges in G_n] $= \frac{1}{2}\lambda$ and otherwise $\mathbb{E}$[number of Edges in G_n] $\leq \lambda$.

Let us also denote, $\nu \in \mathbb{R}^Q$ as the eigenvector of $\tilde{K}$ corresponding to λ. We at first, try to find an asymptotic bound on the graph distance $d_G(u, v)$ for vertices $u, v \in V(G)$.

Theorem 5.4.1. *Let $\lambda > 1$ (defined in Eq. (5.7)), then, the graph distance $d_G(u, v)$ between two uniformly chosen vertices of type a and b respectively, conditioned on being connected, satisfies the following asymptotic relation -*

(i)

$$\mathbb{P}\left[d_G(u,v) < (1-\varepsilon)\frac{\log n}{\log|\lambda|/\log(\nu_a\nu_b)}\right] = o(1) \tag{5.8}$$

(ii)

$$\mathbb{P}\left[d_G(u,v) > (1+\varepsilon)\frac{\log n}{\log|\lambda|/\log(\nu_a\nu_b)}\right] = o(1) \tag{5.9}$$

Now, let us consider the limiting operator $\mathbb{D}$ defined as

Definition 5.4.2. *The **normalized limiting matrix** is an $n \times n$ matrix, $\mathbb{D}$, which in limit as $n \to \infty$ becomes an operator on l_2 (space of convergent sequences), is defined as $\mathbb{D} = [\mathbb{D}_{ij}]_{i,j=1}^n$, where,*

$$\mathbb{D}_{ij} \;=\; \begin{cases} \frac{\log(\nu_a\nu_b)}{\log|\lambda|}, & \text{if type of } v_i = a \neq b = \text{ type of } v_j \\ \frac{2\log(\nu_a)}{\log|\lambda|}, & \text{if type of } v_i = \text{ type of } v_j = a \end{cases}$$

and $\mathbb{D}_{ii} = 0$ for all $i = 1, \ldots, n$.
*The **graph distance matrix** $\mathbf{D}$ can be defined as*

$$\mathbf{D} = [d(v_i, v_j)]_{i,j=1}^n.$$

In Theorem 5.4.1 we had a point-wise result, so, we combine these point-wise results to give a matrix result -

Theorem 5.4.3. *Let $\lambda = ||T_K|| > 1$, then, within the big connected component,*

$$\mathbb{P}\left[\left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F \leq O(n^{1-\varepsilon})\right] = 1 - o(1)$$

Thus, the above theorem gives us an idea about the limiting behavior of the normalized version of geodesic matrix $\mathbf{D}$.

Sketch of Analysis of Part I

A rough idea of the proof of part I is as follows. Fix two vertices, say 1 and 2, in the giant component. Think of a branching process starting from vertices of type 1 and 2, so that at time t, $\mathcal{B}_{P\pi}(a)(t)$ is the branching process tree from vertex of type a and includes the shortest paths to all vertices in the tree at or before time t from vertex a, $a = 1, 2$. When

these two trees meet via the formation of an edge (v_1, v_2) between two vertices $v_1 \in \mathcal{B}_{P\pi}(1)(\cdot)$ and $v_2 \in \mathcal{B}_{P\pi}(2)(\cdot)$, then the shortest-length path between the two vertices 1 and 2 has been found. If $D_n(v_a)$, $a = 1, 2$, denotes the number of edges between the source a and the vertex v_a along the tree $\mathcal{B}_{P\pi}(a)$, then the graph distance $d_n(1, 2)$ is given by

$$d_n(1, 2) = D_n(v_1) + D_n(v_2) + 1 \tag{5.10}$$

The above idea is indeed a very rough sketch of our proof and it follows from the graph distance finding idea developed in [29]. In the paper, we embed the SBM in a multi-type branching process (MTMBP) or a single-type marked branching process (MBP), depending on whether the types of two vertices are same or not. The offspring distribution is binomial with parameters $n - 1$ and kernel P (see Section 5.4). With high probability, the vertex exploration process in the SBM can be coupled with two multi-type branching processes, bounding the vertex exploration process on SBM on both sides. Now, using the property of the two multi-type branching processes, we can bound the number of vertices explored in the vertex exploration process of a SBM graph and infer about the asymptotic limit of the graph distance.

With the above sketch of proof can be organized as follows.

1. We analyze various properties of a Galton-watson process conditioned on non-extinction, including times to grow to a particular size. In this branching process, the offspring will have a Poisson distribution.

2. We introduce multi-type branching process trees with binomially distributed offspring and make the connection between these trees and the SBM. We bound the vertices explored for an SBM graph, starting from a fixed vertex, by considering a muti-type branching process coupled with it.

3. We bound the geodesic distance using the number of vertices explored in the coupled multi-type branching processes within a certain generation. The limiting behavior of the generation give us the limiting behavior of graph distance.

4. The whole analysis is true for IRGM. So, the results are true for SBM with increasing block numbers and degree-corrected block models also.

The idea of the argument is quite simple, but making these ideas rigorous takes some technical work, particularly because we need to condition on our vertices being in the giant component.

Results of Part II

So, from Part I of the analysis, we get an idea about the point-wise asymptotic convergence of the matrix $\mathbf{D} = [d(v_i, v_j)]_{i,j=1}^n$ to the normalized limiting operator $\mathbb{D}$, defined in Definition 5.4.2.

The limiting matrix $\mathbb{D}$ can also be written in terms of limiting low-dimensional matrix, $\mathcal{D}$, which is defined as follows -

Definition 5.4.4. *The limiting kernel matrix, $\mathcal{D}_{Q\times Q}$ is defined as*

$$\mathcal{D}_{ab} = \begin{cases} \frac{\log(\nu_a\nu_b)}{\log|\lambda|}, & \text{if } a \neq b \\ \frac{2\log(\nu_a)}{\log|\lambda|}, & \text{if } a = b \end{cases}$$

So, we can see that if $\mathbf{J}_{n\times n} = \mathbf{1}\mathbf{1}^T$ is an $n \times n$ matrix of all ones, then, there exists a permutation of rows of $\mathbb{D}$, which is obtained by multiplying $\mathbb{D}$ with permutation matrix $\mathbf{R}$, such that,

$$\mathbb{D}\mathbf{R} = \mathcal{D} \star \mathbf{J} - Diag(\tilde{d}) \equiv [\mathcal{D}_{ab}\mathbf{J}_{ab}]_{a,b=1}^{Q} - Diag(\tilde{d}) \tag{5.11}$$

where, $[\mathbf{J}_{ab}]_{a,b=1}^{Q}$ is a $Q \times Q$ partition of $\mathbf{J}$ in the following way - the rows and columns are partitioned in similar fashion according to $(n_1, \ldots, n_Q)$. Note that, $(n_a)_{a=1}^{Q}$ are the number of vertices of type a in the graph G_n. So, $\mathbf{J}_{ab}$ is an $n_a \times n_b$ matrix of all ones. $\tilde{d}$ is a vector of length containing n_a elements of value $\frac{2\log(\nu_a)}{\log|\lambda|}$, $a = 1, \ldots, Q$. Note that product $\star$ can also be seen as a Khatri-Rao product of two partitioned matrices [95].

Now, we assume some conditions on the limiting low-dimensional matrix $\mathcal{D}$.

(C1) The operator T_K or the matrix $\tilde{K}$ can not have $\mathbf{1}$ as the principal eigenvector. If the principal eigenvector $\nu = \mathbf{1}$, then, $\mathcal{D}$ becomes a matrix with no difference between diagonal and off-diagonal elements and thus have no discriminatory power to do community detection.

(C2) The eigenvalues of $\mathcal{D}$, $\lambda_1(\mathcal{D}) \geq \cdots \geq \lambda_Q(\mathcal{D})$, satisfy the condition that there exists an constant α, such that, $0 < \alpha \leq \lambda_Q(\mathcal{D})$.

(C3) The eigenvectors of $\mathcal{D}$, $(v_1(\mathcal{D}), \ldots, v_Q(\mathcal{D}))$ corresponding to $\lambda_1, \ldots, \lambda_Q$, satisfy the condition that there exists a constant β, such that, rows of the $Q \times Q$ matrix $\mathbf{V} = [v_1 \cdots v_Q]$, represented as $(u_1, \ldots, u_Q)$ $(u_a \in \mathbb{R}^Q)$, satisfies the condition $0 < \beta \leq ||u_a - u_b||_2$ for all pairs of rows of $\mathbf{V}$.

(C4) The number of vertices in each type $(n_1, \ldots, n_Q)$, satisfy the condition that there exists a constant θ such that $0 < \theta < \frac{n_a}{n}$ for all $a = 1, \ldots, Q$ and all n.

Theorem 5.4.5. *Under the conditions (C1)-(C4), suppose that the number of blocks Q is known. Let $\hat{\xi} : V \mapsto [Q]$ be the block assignment function according to a clustering of the rows of $\tilde{\mathbf{W}}^{(n)}$ satisfying algorithm in Section 5.3 and $\xi : V \mapsto [Q]$ be the actual assignment. Let $\mathcal{P}_Q$ be the set of permutations on $[Q]$. With high probability and for large n it holds that*

$$\min_{\pi\in\mathcal{P}_Q} |\{u \in V : \xi(u) \neq \pi(\hat{\xi}(u))\}| = O(n^{1/2-\varepsilon}) \tag{5.12}$$

Sketch of Proof of Part II

We can consider the limiting distribution of the graph distance matrix as $\mathbf{D}$ which was proposed in Theorem 5.4.3, with $(\mathbf{D}_{ij}) = d_G(v_i, v_j)$, where, $v_i, v_j \in V(G)$. Our goal is to show that the eigenvectors of $\mathbf{D}$ or normalized version of it, converge to eigenvectors of $\mathcal{D}$ or $\mathbb{D}$. For that reason, we use the perturbation theory of operators, as given in Kato [93] and Davis-Kahan [48]. The steps are as follows

- We use Davis-Kahan to show convergence of eigenspace $\tilde{\mathbf{W}}$, formed by top Q eigenvectors of $\mathbf{D}/\log n$ to $\mathbf{WR}$, where, $\mathbf{W}$ is the eigenspace formed by the top Q eigenvectors of $\mathbb{D}$ and $\mathbf{R}$ is some orthogonal permutation matrix, which permutes the rows of $\mathbf{W}$.
- We show by contradiction that if the clustering assignment makes too many mistakes then the rate of convergence of $\tilde{\mathbf{W}}$ to $\mathbf{WR}$ would be violated.

Branching Process Results

The branching process $\mathcal{B}_K(a)$ is a multi-type Galton-Watson branching processes with type space $\mathcal{S} \equiv \{1, \ldots, Q\}$, a particle of type $a \in \mathcal{S}$ is replaced in the next generation by its "children", a set of particles whose types are distributed as a Poisson process on $\mathcal{S}$ with intensity $\{K_{ab}\pi_b\}_{b=1}^{Q}$. Recall the parameters $K \in \mathbb{R}^{Q\times Q}$ and $\pi \in [0,1]^Q$ with $\sum_{a=1}^{Q} \pi_a = 1$, from the definition of Stochastic block model in equation (5.4). The zeroth generation of $\mathcal{B}_K(a)$ consists of a single particle of type a. Also, the branching process $\mathcal{B}_K$ is just the process $\mathcal{B}_K(a)$ started with a single particle whose (random) type is distributed according to the probability measure $(\pi_1, \ldots, \pi_Q)$.

Let us recall our notation for the survival probabilities of particles in $\mathcal{B}_K(a)$. We write $\rho_k(K; a)$ for the probability that the total population consists of exactly k particles, and $\rho_{\geq k}(K; a)$ for the probability that the total population contains at least k particles. Furthermore, $\rho(K; a)$ is the probability that the branching process survives for eternity. We write $\rho_k(K), \rho_{\geq k}(K)$ and $\rho(K)$ for the corresponding probabilities for $\mathcal{B}_K$, so that, e.g., $\rho_k(K) = \sum_{a=1}^{Q} \rho_k(K; a)\pi_a$.

Now, we try to find a coupling relation between *neighborhood exploration process* of a vertex of type a in stochastic block model and multi-type Galton-Watson process, $\mathcal{B}(a)$ starting from a vertex of type a.

We assume all vertices of graph G_n generated from a stochastic block model has been assigned a community or type ξ_i (say) for vertex $v_i \in V(G_n)$. By *neighborhood exploration process* of a vertex of type a in stochastic block model, we mean that we start from a random vertex v_i of type a in the random graph G_n generated from stochastic block model. Then, we count the number of vertices of the random graph G_n are neighbors of v_i, $N(v_i)$. We repeat the neighborhood exploration process by looking at the neighbors of the vertices in $N(v_i)$. We continue until we have covered all the vertices in G_n. Since, we either consider G_n connected or only the giant component of G_n, the neighborhood exploration process will end in finite steps but the number of steps may depend on n.

Lemma 5.4.6. *Within the giant component, the neighborhood exploration process for a stochastic block model graph with parameters $(P, \pi) = (K/n, \pi)$, can be bounded with high probability by two multi-type branching processes with kernels $(1-2\epsilon)K$ and $(1+\epsilon)K$ for some $\epsilon > 0$.*

Proof. We have n_a vertices of type a ,$a = 1, \ldots, Q$, and that $n_a/n \overset{a.s.}{\to} \pi_a$. From now on we condition on $n_1, \ldots, n_Q$; we may thus assume that $n_1, \ldots, n_Q$ are deterministic with $n_a/n \to \pi_a$. Let $\omega(n)$ be any function such that $\omega(n) \to \infty$ and $\omega(n)/n \to 0$. We call a component of $G_n \equiv G(n, P) = G(n, K/n)$ big if it has at least $\omega(n)$ vertices. Let B be the union of the big components, so $|B| = N_{\geq \omega(n)}(G_n)$. Fix $\epsilon > 0$.We may assume that n is so large that $\omega(n)/n < \epsilon\, \pi_i$ and $|n_a/n - \pi_a| < \epsilon\, \pi_a$ for every a; thus $(1-\epsilon)\pi_a n < n_a < (1+\epsilon)\pi_a n$. We may also assume that $n > \max K$, as K is a function on the finite set $\mathcal{S} \times \mathcal{S}$. Since, n_a/n is a $\sqrt{n}$-consistent estimator of π_a, we get that

$$\epsilon = O(n^{-1/2}). \tag{5.13}$$

Select a vertex and explore its component in the usual way, that means looking at its neighbors, one vertex at a time. We first reveal all edges from the initial vertex, and put all neighbors that we find in a list of unexplored vertices; we then choose one of these and reveal its entire neighborhood, and so on. Stop when we have found at least $\omega(n)$ vertices (so $x \in B$), or when there are no unexplored vertices left (so we have found the entire component and $x \notin B$).

Consider one step in this exploration, and assume that we are about to reveal the neighborhood of a vertex x of type a. Let us write n_b' for the number of unused vertices of type b remaining. Note that $n_b \geq n_b' \geq n_b - \omega(n)$, so

$$(1-2\epsilon)\pi_b < n_b'/n < (1+\epsilon)\pi_b \tag{5.14}$$

The number of new neighbors of x of type b has a binomial $Bin(n_b', K_{ab}/n)$ distribution, and the numbers for different b are independent. The total variation distance between a binomial $Bin(n, p)$ distribution and the Poisson distribution with the same mean is at most p. Hence the total variation distance between the binomial distribution above and the Poisson distribution $Poi(K_{ab}n_b'/n)$ is at most $K_{ab}/n = O(1/n)$. Also, by (5.14),

$$(1-2\epsilon)K_{ab}\pi_b \leq K_{ab}n_b'/n \leq (1+\epsilon)K_{ab}\pi_b. \tag{5.15}$$

Since we perform at most $\omega(n)$ steps in the exploration, we may, with an error probability of $O(\omega(n)/n) = o(1)$, couple the exploration with two multi-type branching processes $\mathcal{B}((1-2\epsilon)K)$ and $\mathcal{B}((1+\epsilon)K)$ such that the first process always finds at most as many new vertices of each type as the exploration, and the second process finds at least as many. Consequently, for a vertex x of type a,

$$\rho_{\geq\omega(n)}((1-2\epsilon)K; a) + o(1) \leq \mathbb{P}(x \in B) \leq \rho_{\geq\omega(n)}((1+\epsilon)K; a) + o(1). \tag{5.16}$$

Since $\omega(n) \to \infty$, by Lemma 9.5 of [29], we have $\rho_{\geq \omega(n)}(K; a) \to \rho(K; a)$ for every matrix or finitary kernel K, which parametrizes the offspring distribution of the branching process in the sense that the number of offsprings of type b coming from a parent of type a follows $Poi(K_{ab}\pi_b)$ distribution. So we can rewrite (5.16) as

$$\rho((1-2\epsilon)K; a) + o(1) \leq P(x \in B) \leq \rho((1+\epsilon)K; a) + o(1). \tag{5.17}$$

□

Now, we restrict ourselves to the giant component only. So, if we condition that the exploration process does not leave the giant component, it is same as conditioning that the branching process does not die out. Under this additional condition, the branching process can be coupled with another branching process with a different kernel. The kernel of that branching process is given in following lemma.

Lemma 5.4.7. *If we condition a branching process, $\mathcal{B}_{K\pi}$ on survival, the new branching process has kernel $\left(K_{ab}\left(\rho(K; a) + \rho(K; b) - \rho(K; a)\rho(K; b)\right)\right)_{a,b=1}^{Q}$.*

Proof. We need to consider certain branching process expectations $\sigma(K)$ and $\sigma_{\geq k}(K)$ in place of $\rho(K)$ and $\rho_{\geq k}(K)$. In preparation for the proof, we shall relate $\zeta(K)$ to the branching process $\mathcal{B}_K$ via $\sigma(K)$. As before, we assume that K is a kernel on $(\mathcal{S}, \pi)$ with $K \in L^1$.

Let A be a Poisson process on $\mathcal{S}$, with intensity given by a finite measure λ, so that A is a random multi-set on $\mathcal{S}$. If g is a bounded measurable function on multi-sets on $\mathcal{S}$, it is easy to see that

$$\mathbb{E}(|A|g(A)) = \sum_{i \in \mathcal{S}} \mathbb{E}g(A \cup \{i\})\lambda_i \tag{5.18}$$

For details see Proposition 10.4 of [29].

Let $B(x)$ denote the first generation of the branching process $\mathcal{B}_K(x)$. Thus $B(x)$ is given by a Poisson process on $\mathcal{S}$ with intensity $K(x, y)\pi_x$. Suppose that $\sum_b K_{ab}\pi_b < \infty$ for every $a = 1, \ldots, Q$, so $B(x)$ is finite. Let $\sigma(K; x)$ denote the expectation of $|B(x)|\mathbf{1}[|\mathcal{B}_K(x)| = \infty]$, recalling that under the assumption $\sum_b K_{ab}\pi_b < \infty$ for every a, the branching process $\mathcal{B}_K(x)$ dies out if and only if $|\mathcal{B}_K(x)| < \infty$. Then

$$\begin{aligned}
\sum_{b=1}^{Q} K_{xb}\pi_b - \sigma(K; x) &= \mathbb{E}\left[|B(x)|\mathbf{1}(\mathcal{B}_K(x) < \infty)\right] \\
&= \mathbb{E}\left(|B(x)| \prod_{z \in B(x)} \rho(K; z)\right) \\
&= \sum_{b=1}^{Q} K_{xb}(1 - \rho(K; b))\mathbb{E}\left(\prod_{z \in B(x)} \rho(K; z)\right)\pi_b \\
&= \sum_{b=1}^{Q} K_{xb}(1 - \rho(K; b))(1 - \rho(K; x))\pi_b
\end{aligned}$$

Here the penultimate step is from (5.18); the last step uses the fact that the branching process dies out if and only if none of the children of the initial particle survives. Writing B for the first generation of $\mathcal{B}_K$ conditioned on survival becomes

$$\sigma(K) \equiv \mathbb{E}|B|\mathbf{1}[|\mathcal{B}_K|=\infty] = \sum_{x=1}^{Q} \sigma(K;x)\pi_x$$

Then, integrating over x and subtracting from $\sum_{a,b} K_{ab}\pi_a\pi_b$, we get,

$$\sigma(K) = \sum_{a,b} K_{ab}\left(1-(1-\rho(K;a))(1-\rho(K;b))\right)\pi_a\pi_b \tag{5.19}$$

So, the kernel for the conditioned branching process becomes

$$K_{ab}\left(\rho(K;a)+\rho(K;b)-\rho(K;a)\rho(K;b)\right) \tag{5.20}$$

□

Now, we shall try to prove the limiting behavior of typical distance between vertices v and w of G_n, where, $v, w \in V(G_n)$.

Lemma 5.4.8. *Let us have $\lambda \equiv ||T_K|| > 1$ and let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ, then,*

$$\mathbb{E}\left|\{\{v,w\} : d_G(v,w) \le (1-\varepsilon)\log n/\log|\lambda|\}\right| = O(n^{2-\varepsilon})$$

and so

$$\left|\left\{\{v,w\} : d_G(v,w) \le (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| \le O(n^{2-\varepsilon/2}) \text{ with high probability}$$

Proof. We have $\mathcal{S}$ is finite, say $\mathcal{S} = \{1, 2, \ldots, Q\}$. Let $\Gamma_d(v) \equiv \Gamma_d(v, G_n)$ denote the d-distance set of v in G_n, i.e., the set of vertices of G_n at graph distance exactly d from v, and let $\Gamma_{\le d}(v) \equiv \Gamma_{\le d}(v, G_n)$ denote the d-neighborhood $\cup_{d' \le d}\Gamma_{d'}(v)$ of v.

Let $0 < \varepsilon < 1/10$ be arbitrary. The proof of (5.17) involved first showing that, for n large enough, the neighborhood exploration process starting at a given vertex v of G_n with type a (chosen without inspecting G_n) could be coupled with the branching process $\mathcal{B}_{(1+\varepsilon)K'}(i)$, where the K' is defined by equation (5.20), so that the branching process is conditioned to survive. However, henceforth we shall abuse notation and denote K' as K.

The neighborhood exploration process and multi-type branching process can be coupled so that for every d, $|\Gamma_d(v)|$ is at most the number N_d of particles in generation d of $\mathcal{B}_{(1+2\varepsilon)K}(i)$. The number of vertices at generation d of type c of branching process $\mathcal{B}_{(1+2\varepsilon)K}(a)$, denoted by $N^a_{d,c}$ and the number of vertices of type c at distance d from v for the neighborhood exploration process of G_n is denoted by $|\Gamma^a_{d,c}(v)|$f, where, $c = 1, \ldots, Q$.

Elementary properties of the branching process imply that $\mathbb{E}N_d = O\left(||T_{(1+2\varepsilon)K}||^d\right) = O(((1+2\varepsilon)\lambda)^d)$, where $\lambda = ||T_K|| > 1$.

Let $N_t^a(c)$ be the number of particles of type c in the t-th generation of $\mathcal{B}_K(a)$, then, N_t^a is the vector $(N_t^a(1), \ldots, N_t^a(Q))$. Also, let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ (unique, up to normalization, as P is irreducible). From standard branching process results, we have

$$N_t^a/\lambda^t \overset{a.s.}{\to} X\nu, \tag{5.21}$$

where $X \geq 0$ is a real-valued random variable, X is continuous except that it has some mass at 0, and $X = 0$ if and only if the branching process eventually dies out and lastly,

$$\mathbb{E}X = \nu_a.$$

under the conditions given in Theorem V.6.1 and Theorem V.6.2 of [7].

Set $D = (1-10\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda$. Then $D < (1-\varepsilon)\log(n/\nu_a\nu_b)/\log((1+2\varepsilon)\lambda)$ if ε is small enough, which we shall assume. Thus,

$$\mathbb{E}|\Gamma_{\leq D}(v)| \leq \mathbb{E}\sum_{d=0}^{D} N_d = O(((1+2\varepsilon)\lambda)^D) = O(n^{1-\varepsilon})$$

So, summing over v, we have

$$\sum_{v\in V(G_n)} |\Gamma_{\leq D}(v)| = |\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda\}|$$

and its expected value to be

$$\mathbb{E}\,|\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda\}| = \mathbb{E}\sum_{v\in V(G_n)} |\Gamma_{\leq D}(v)| = O(n^{2-\varepsilon})$$

The above statement is equivalent to

$$\mathbb{E}\left|\left\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| = \mathbb{E}\sum_{v\in V(G_n)} |\Gamma_{\leq D}(v)| = O(n^{2-\varepsilon})$$

So, by Markov's Theorem, we have,

$$\mathbb{P}\left[\left|\left\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| \leq O(n^{2-\varepsilon/2})\right] = o(1)$$

for any fixed $\varepsilon > 0$. □

Now, let us try to bound the typical distance between two vertices of the any type. We shall only give an upper bound for typical distance between two vertices of any type.

Lemma 5.4.9. *Let us have* $\lambda \equiv ||T_K|| = \lambda_{\max}(\tilde{K}) > 1$ *from Eq* (5.7) *and let* $\nu = (\nu_1, \ldots, \nu_Q)$ *be the eigenvector of* T_K *with eigenvalue* λ*. For uniformly selected vertices* $v, w \in V(G)$*,*

$$\mathbb{P}\left(d_G(v,w) < (1+\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right) = 1 - exp(-\Omega(n^{2\eta}))$$

conditioned on the event that the branching process $\mathcal{B}_K$ *survives.*

Proof. We consider the multi-type branching process with probability kernel $P_{ab} = \frac{K_{ab}}{n}$ $\forall a, b = 1, \ldots, Q$ and the corresponding random graph G_n generated from stochastic block model has in total n nodes. We condition that branching process $\mathcal{B}_K$ survives.

Note that an upper bound 1 is obvious, since we are bounding a probability, so it suffices to prove a corresponding lower bound. We may and shall assume that $K_{ab} > 0$ for some a, b.

Fix $0 < \eta < 1/10$. We shall assume that η is small enough that $(1-2\eta)\lambda > 1$. In the argument leading to (5.17) in proof of Lemma 5.4.6, we showed that, given $\omega(n)$ with $\omega(n) = o(n)$ and a vertex v of type a, the neighborhood exploration process of v in G_n could be coupled with the branching process $\mathcal{B}_{(1-2\eta)K}(a)$ so that whp the former dominates until it reaches size $\omega(n)$. More precisely, writing $N_{d,c}$ for the number of particles of type c in generation d of $\mathcal{B}_{(1-2\eta)K}(a)$, and $\Gamma_{d,c}(v)$ for the set of type c vertices at graph distance d from v, whp

$$|\Gamma_{d,c}(v)| \geq N_{d,c},\ c = 1, \ldots, Q, \quad \text{for all } d \text{ s.t. } |\Gamma_{\leq d}(v)| < \omega(n). \tag{5.22}$$

This relation between the number of vertices at generation d of type c of branching process $\mathcal{B}_{(1-2\eta)K}(a)$, denoted by $N_{d,c}$ and the number of vertices of type c at distance d from v for the neighborhood exploration process of G_n, denoted by $|\Gamma_{d,c}(v)|$ becomes highly important later on in this proof, where, $c = 1, \ldots, Q$. Note that the relation only holds when $|\Gamma_{\leq d}(v)| < \omega(n)$ for some $\omega(n)$ such that $\omega(n)/n \to 0$ as $n \to \infty$.

Let $N_t^a(c)$ be the number of particles of type c in the t-th generation of $\mathcal{B}_K(a)$, then, N_t^a is the vector $(N_t^a(1), \ldots, N_t^a(Q))$. Also, let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ (unique, up to normalization, as P is irreducible). From standard branching process results, we have

$$N_t^a/\lambda^t \xrightarrow{a.s.} X\nu, \tag{5.23}$$

where $X \geq 0$ is a real-valued random variable, X is continuous except that it has some mass at 0, and $X = 0$ if and only if the branching process eventually dies out and lastly,

$$\mathbb{E}X = \nu_a$$

under the conditions given in Theorem V.6.1 and Theorem V.6.2 of [7].

Let D be the integer part of $\log((n/\nu_a\nu_b)^{1/2+2\eta})/\log((1-2\eta)\lambda)$. From (5.23), conditioned on survival of branching process $\mathcal{B}_K(a)$, whp either $N_D^a = 0$, or $N_{D,c}^a \geq n^{1/2+\eta}$ for each c (note that $N_{D,c}^a$ comes from branching process $\mathcal{B}_{(1-2\eta)K}(a)$ not branching process $\mathcal{B}_K(a)$).

Furthermore, as $\lim_{d\to\infty}\mathbb{P}(N_d^a \neq 0) = \rho((1-2\eta)K)$ and $D \to \infty$, we have $\mathbb{P}(N_D^a \neq 0) \to \rho((1-2\eta)K)$. Thus, if n is large enough,

$$\mathbb{P}\left(\forall c : N_{D,c}^a \geq n^{1/2+\eta}\right) \geq \rho((1-2\eta)K) - \eta.$$

Now, we have conditioned that the branching process with kernel K is conditioned to survive. The right-hand side tends to $\rho(K) = 1$ as $\eta \to 0$. Hence, given any fixed $\gamma > 0$, if we choose $\eta > 0$ small enough we have

$$\mathbb{P}\left(\forall c : N_{D,c}^a \geq n^{1/2+\eta}\right) \geq 1 - \gamma$$

for n large enough.

Now, the neighborhood exploration process and branching process can be coupled so that for every d, $|\Gamma_d(v)|$ is at most the number M_d of particles in generation d of $\mathcal{B}_{(1+2\varepsilon)K}(a)$ from Lemma 5.4.6 and Eq (5.15). So, we have,

$$\mathbb{E}|\Gamma_{\leq D}(v)| \leq \mathbb{E}\sum_{d=0}^{D} M_d = O(((1+2\varepsilon)\lambda)^D) = o(n^{2/3})$$

if η is small enough, since D be the integer part of $\log(n^{1/2+2\eta})/\log((1-2\eta)\lambda)$. Note that the power $2/3$ here is arbitrary, we could have any power in the range $(1/2, 1)$. Hence,

$$|\Gamma_{\leq D}(v)| \leq n^{2/3} \quad whp,$$

and whp the coupling described in (5.22) extends at least to the D-neighborhood. So, now, we are in a position to apply Eq (5.22), as we have $|\Gamma_{\leq D}(v)| \leq n_a^{2/3} < \omega(n)$, with $\omega(n)/n \to 0$.

Now let v and w be two fixed vertices of $G(n, P)$, of types a and b respectively. We explore both their neighborhoods at the same time, stopping either when we reach distance D in both neighborhoods, or we find an edge from one to the other, in which case v and w are within graph distance $2D + 1$. We consider two independent branching processes $\mathcal{B}_{(1-2\eta)K}(a)$, $\mathcal{B}'_{(1-2\eta)K}(b)$, with $N_{d,c}^a$ and $N_{d,c}^b$ vertices of type c in generation d respectively. By previous equation, whp we encounter $o(n)$ vertices in the explorations so, by the argument leading to (5.22), whp either the explorations meet, or $|\Gamma_{D,c}^a(v)| \geq N_{D,c}^a$ and $|\Gamma_{D,c}^b(w)| \geq N_{D,c}^b$, $c = 1, \ldots, Q$, with the explorations not meeting. Using bound on $N_{d,c}^a$ and the independence of the branching processes, it follows that

$$\mathbb{P}\left(d(v,w) \leq 2D+1 \text{ or } \forall c : |\Gamma_{D,c}^a(v)|, |\Gamma_{D,c}^b(w)| \geq n^{1/2+\eta}\right) \geq (\rho(K)-\gamma)^2 - o(1).$$

Note that the two events in the above probability statement are not disjoint. We shall try to find the probability that the second event in the above equation holds but not the first. We have not examined any edges from $\Gamma_D(v)$ to $\Gamma_D(w)$, so these edges are present independently with their original unconditioned probabilities. For any c_1, c_2, the expected number of these edges is at least $|\Gamma_{D,c_1}^a(v)||\Gamma_{D,c_2}^b(w)|K_{c_1c_2}/n$. Choosing c_1, c_2 such that $K_{c_1c_2} > 0$, this

expectation is $\Omega((n^{1/2+\eta})^2/n) = \Omega(n^{2\eta})$. It follows that at least one edge is present with probability $1 - \exp(-\Omega(n^{2\eta})) = 1 - o(1)$. If such an edge is present, then $d(v, w) \leq 2D + 1$. So, the probability that the second event in the above equation holds but not the first is $o(1)$. Thus, the last equation implies that

$$\mathbb{P}(d(v, w) \leq 2D + 1) \geq (1 - \gamma)^2 - o(1) \geq 1 - 2\gamma - o(1).$$

Choosing η small enough, we have $2D + 1 \leq (1 + \varepsilon) \log(n/\nu_a\nu_b)/\log \lambda$. As γ is arbitrary, we have

$$\mathbb{P}(d(v, w) \leq (1 + \varepsilon) \log(n/\nu_a\nu_b)/\log \lambda) \geq 1 - \exp(-\Omega(n^{2\eta})).$$

The above statement is equivalent to

$$\mathbb{P}\left(d(v, w) \leq (1 + \varepsilon)\frac{\log n}{\log \lambda/\log(\nu_a\nu_b)}\right) \geq 1 - \exp(-\Omega(n^{2\eta})).$$

and the lemma follows.

□

Proof of Theorem 5.4.1 and Theorem 5.4.3

Proof of Theorem 5.4.1

We shall try to prove the limiting behavior of typical graph distance in the giant component as $n \to \infty$. The Theorem essentially follows from Lemma 5.4.8 and Lemma 5.4.9. Under the conditions mentioned in the Theorem, part (a) follows from Lemma 5.4.8 and part (b) follows from Lemma 5.4.9.

Proof of Theorem 5.4.3

From the definition 5.4.2, we have that $\mathbf{D}_{ij}$ = graph distance between vertices v_i and v_j, where, $v_i, v_j \in V(G_n)$.

From Lemma 5.4.8, we get for any vertices v and w with high probability,

$$\left|\left\{\{v, w\} : d_G(v, w) \leq (1 - \varepsilon)\frac{\log n}{\log \lambda/\log(\nu_a\nu_b)}\right\}\right| \leq O(n^{2-\varepsilon}).$$

Also, from Lemma 5.4.9, we get

$$\mathbb{P}\left(d_G(v, w) < (1 + \varepsilon)\frac{\log n}{\log \lambda/\log(\nu_a\nu_b)}\right) = 1 - exp(-\Omega(n^{2\eta}))$$

So, putting the two statements together, we get that with high probability,

$$\sum_{i,j=1:type(v_i)\neq type(v_j)}^{n} \left(\frac{\mathbf{D}_{ij}}{\log n} - \mathbb{D}_{ij}\right)^2 = O(n^{2-\varepsilon}) + O(n^2).\varepsilon^2 = O(n^{2-\varepsilon})$$

since, by Eq. (5.13) $\epsilon = O(1/\sqrt{n})$ and and $(1-exp(-\Omega(n^{2\eta})))^{n^2} \to 1$ as $n \to \infty$. So, putting the two cases together, we get that with high probability,

$$\sum_{i,j=1}^{n} \left(\frac{\mathbf{D}_{ij}}{\log n} - \mathbb{D}_{ij} \right)^2 = O(n^{2-\varepsilon}) + O(n^2).\varepsilon^2 = O(n^{2-\varepsilon}).$$

Hence,

$$\left|\left| \frac{\mathbf{D}}{\log n} - \mathbb{D} \right|\right|_F \leq O(n^{1-\varepsilon/2}).$$

Perturbation Theory of Linear Operators

Once, we have the limiting behavior of the matrix D established in Theorem 5.4.3, we shall now try to see the behavior of the eigenvectors of the matrix D. Now, matrix D can be considered as a perturbation of the operator $\mathbb{D}$.

The Davis-Kahan Theorem states a bound on perturbation of eigenspace instead of eigenvector, as discussed previously. The $\sin\theta$ Theorem of Davis-Kahan [48]

Theorem 5.4.10 (Davis-Kahan (1970)[48])**.** *Let $\mathbf{H}, \mathbf{H}' \in \mathbb{R}^{n\times n}$ be symmetric, suppose $\mathcal{V} \subset \mathbb{R}$ is an interval, and suppose for some positive integer d that $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{n\times d}$ are such that the columns of $\mathbf{W}$ form an orthonormal basis for the sum of the eigenspaces of $\mathbf{H}$ associated with the eigenvalues of $\mathbf{H}$ in $\mathcal{V}$ and that the columns of $\mathbf{W}'$ form an orthonormal basis for the sum of the eigenspaces of $\mathbf{H}'$ associated with the eigenvalues of $\mathbf{H}'$ in $\mathcal{V}$. Let δ be the minimum distance between any eigenvalue of $\mathbf{H}$ in $\mathcal{V}$ and any eigenvalue of $\mathbf{H}$ not in $\mathcal{V}$. Then there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ such that $||\mathbf{WR} - \mathbf{W}'||_F \leq \sqrt{2}\frac{||\mathbf{H}-\mathbf{H}'||_F}{\delta}$.*

Proof of Theorem 5.4.5

Now, we can try to approximate limiting operator by the graph distance matrix $\mathbf{D}$ in Frobenius norm based on Theorem 5.4.3 of Part I. The behavior of the eigenvalues of the limiting operator $\mathbb{D}$ can be stated as follows -

Lemma 5.4.11. *The eigenvalues of $\mathbb{D}$ - $|\lambda_1(\mathbb{D})| \geq |\lambda_2(\mathbb{D})| \geq \cdots \geq |\lambda_n(\mathbb{D})|$, can be bounded as follows -*

$$\lambda_1(\mathbb{D}) < n,\ |\lambda_K(\mathbb{D})| > Cn,\ \lambda_{Q+1}(\mathbb{D}) = -\min\{\tilde{d}_1, \ldots, \tilde{d}_Q\}, \ldots, \lambda_n = -\max\{\tilde{d}_1, \ldots, \tilde{d}_Q\} \tag{5.24}$$

where, $\tilde{d}$, a vector of length Q, is defined in Eq. (5.11) and the smallest $(n-Q)$ absolute eigenvalues of $\mathbb{D}$ are $-\tilde{d}$ where $-\tilde{d}_a$ has multiplicity $(n_a - 1)$ for $a = 1, \ldots, Q$.

Proof. The matrix $\mathbb{D}$ can be considered as a Khatri-Rao product of the matrices $\mathcal{D}$ and $\mathbf{J}$ according to equation (5.11). Now, there exists a constant τ such that $\log||T_K||> \tau > 0$, since $||T_K||> 1$. So, we have $\lambda_1(\mathcal{D}) < \tau$. So, we have $\lambda_1(\mathcal{D}) < 1$ and since $n_a \leq n$ for all a and $\sum_a n_a = n$. So, we have $\lambda_1(\mathbb{D}) \leq n$. Now, By Assumption (C2) and (C4), $\lambda_Q(\mathcal{D}) \geq \alpha$ and $n_a \geq \gamma n$, so, $\lambda_Q(\mathbb{D}) \geq \alpha\gamma n$. Now, it is easy to see that the remaining eigenvalues of $\mathbb{D}$ is -1, since, $\mathcal{B} \star \mathbf{J}$ is a rank Q matrix and its remaining eigenvalues are zero and the eigenvalues of diagonal matrix are $\tilde{d}$ with $\tilde{d}_a$ having multiplicity (n_a) for $a = 1, \ldots, Q$. □

Corollary 5.4.12. *With high probability it holds that* $|\lambda_Q(\mathbf{D}/\log n)|\geq O(n)$ *and* $\lambda_{Q+1}(\mathbf{D}/\log n) \leq O(n^{1-\varepsilon})$.

Proof. By Weyl's Inequality, for all $i = 1, \ldots, n$,

$$\begin{aligned} ||\lambda_i(\mathbf{D}/\log n)|-|\lambda_i(\mathbb{D})|| &\leq \left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F \leq O(n^{1-\varepsilon/2}) \\ &\leq O(n^{1-\varepsilon}) \end{aligned}$$

So, $|\lambda_Q(\mathbf{D}/\log n)|\geq O(n)-O(n^{1-\varepsilon}) = O(n)$ for large n and $|\lambda_{Q+1}(\mathbf{D}/\log n)|\leq -1+O(n^{1-\varepsilon}) = O(n^{1-\varepsilon})$. □

Now, if we consider $\mathbf{W}$ is the eigenspace corresponding to top Q absolute eigenvalues of $\mathbb{D}$ and $\tilde{\mathbf{W}}$ is the eigenspace corresponding to top Q absolute eigenvalues of $\mathbf{D}$. Using Davis-Kahan

Lemma 5.4.13. *With high probability, there exists an orthogonal matrix* $\mathbf{R} \in \mathbb{R}^{Q\times Q}$ *such that* $||\mathbf{WR} - \tilde{\mathbf{W}}||_F \leq O(n^{-\varepsilon})$

Proof. The top Q eigenvalues of both $\mathbb{D}$ and $\mathbf{D}/\log n$ lies in (Cn, ∞) for some $C > 0$. Also, the gap $\delta = O(n)$ between top Q and $Q + 1$th eigenvalues of matrix $\mathbb{D}$. So, now, we can apply Davis-Kahan Theorem 5.4.10 and Theorem 5.4.3, to get that,

$$||\mathbf{WR} - \tilde{\mathbf{W}}||_F \leq \sqrt{2}\frac{\left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F}{\delta} \leq \frac{O(n^{1-\varepsilon})}{O(n)} = O(n^{-\varepsilon})$$

□

Now, the relationship between the rows of W can be specified based on Assumption (C3) as follows -

Lemma 5.4.14. *For any two rows* i, j *of* $\mathbf{W}_{n\times Q}$ *matrix,* $||u_i - u_j||_2 \geq O(1/\sqrt{n})$*, if type of* $v_i \neq$ *type of* v_j.

Proof. The matrix $\mathbb{D}$ can be considered as a Khatri-Rao product of the matrices $\mathcal{D}$ and $\mathbf{J}$ according to equation (5.11). Now, by Assumption (C3), we have a constant difference between the rows of matrix $\mathcal{D}$. So, rows of $\mathbb{D}$ as well as the projection of $\mathbb{D}$ into into its top Q eigenspace has difference of order $O(n^{-1/2})$ between rows of matrix. □

Now, if we consider Q-means criterion as the clustering criterion on $\tilde{\mathbf{W}}$, then, for the Q-means minimizer centroid matrix $\mathbf{C}$ is an $n \times Q$ matrix with Q distinct rows corresponding to the Q centroids of Q-means algorithm. By property of Q-means objective function and Lemma 5.4.13, with high probability,

$$\begin{aligned} ||\mathbf{C} - \tilde{\mathbf{W}}||_F &\leq ||\mathbf{WR} - \tilde{\mathbf{W}}||_F \\ ||\mathbf{C} - \mathbf{WR}||_F &\leq ||\mathbf{C} - \tilde{\mathbf{W}}||_F + ||\mathbf{WR} - \tilde{\mathbf{W}}||_F \\ &\leq 2||\mathbf{WR} - \tilde{\mathbf{W}}||_F \\ &\leq O(n^{-\varepsilon}) \end{aligned}$$

By Lemma 5.4.14, for large n, we can get constant C, such that, Q balls, $B_1, \ldots, B_Q$, of radius $r = Cn^{-1/2}$ around Q distinct rows of $\mathbf{W}$ are disjoint.

Now note that with high probability the number of rows i such that $||\mathbf{C}_i - (\mathbf{WR})_i|| > r$ is at most $O(n^{1/2-\varepsilon})$. If the statement does not hold then,

$$\begin{aligned} ||\mathbf{C} - \mathbf{WR}||_F &> r.O(n^{1/2-\varepsilon}) \\ &\geq Cn^{-1/2}.O(n^{1/2-\varepsilon}) = O(n^{-\varepsilon}) \end{aligned}$$

So, we get a contradiction, since $||\mathbf{C} - \mathbf{WR}||_F \leq O(n^{-\varepsilon})$. Thus, the number of mistakes should be at most of order $O(n^{1/2-\varepsilon})$.

So, for each $v_i \in V(G_n)$, if ξ_i is the type of v_i and $\hat{\xi}_i$ is the type of v_i as estimated from applying Q-means on top Q eigenspace of geodesic matrix $\mathbf{D}$, we get that with high probability, for some small $0 < \varepsilon$,

$$\min_{\pi \in \mathcal{P}_Q} |\{u \in V : \xi(u) \neq \pi(\hat{\xi}(u))\}| = O(n^{1/2-\varepsilon})$$

5.5 Application

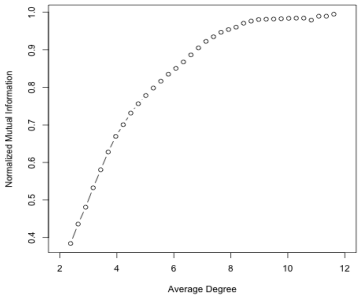
We investigate the empirical performance of the algorithm in several different setup. At first, we use simulated networks from stochastic block model to find the empirical performance of the algorithm. Then, we apply our method to find communities in several real world networks.

Simulation

We simulate networks from stochastic block models with $Q = 3$ blocks. Let w correspond to a Q-block model defined by parameters $\theta = (\boldsymbol{\pi}, \rho_n, S)$, where π_a is the probability of a node being assigned to block a as before, and

$$\mathbf{F}_{ab} = \mathbb{P}(A_{ij} = 1 | i \in a, j \in b) = \rho_n S_{ab}, \quad 1 \leq a, b \leq K.$$

and the probability of node i to be assigned to block a to be π_a $(a = 1, \ldots, K)$.



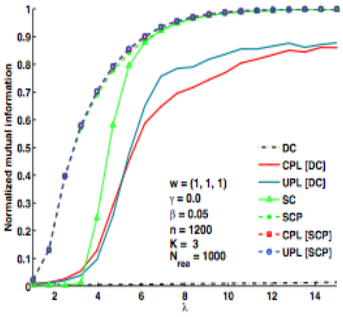


Figure 5.1: The LHS is the performance of graph distance based method and RHS is the performance of Pseudolikelihood method on same generative SBM.

Equal Density Clusters

We consider a stochastic block model with $Q = 3$. We consider the parameter matrix $\mathbf{F} = 0.012(1+0.1\nu)(\tilde{\lambda}F^{(1)} + (1-\tilde{\lambda})F^{(2)})$, where, $F^{(1)}_{3\times3} = \text{Diag}(0.9, 0.9, 0.9)$ and $F^{(2)}_{3\times3} = 0.1\mathbf{J}_2$, where, $\mathbf{J}_2$ is a 2×2 matrix of all 1's and ν varies from 1 to 15 to give networks of different density. So, we get $\rho_n = \boldsymbol{\pi}^T\mathbf{F}\boldsymbol{\pi}$. We now, vary $\tilde{\lambda}$ to get different combinations of $\mathbf{F}$ as well as ρ_n.

In the following figures, we try to see the behavior of mean and variances of the count statistics, as we vary λ_n as we vary ν.

Unequal Density Clusters

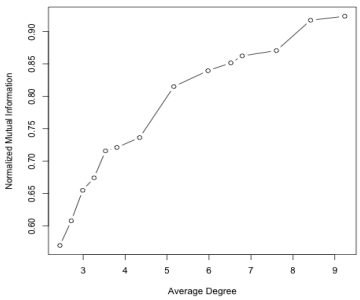
We consider a stochastic block model with $Q = 3$. We consider the parameter matrix $\mathbf{F} = 0.012(1+0.1\nu)(\tilde{\lambda}F^{(1)} + (1-\tilde{\lambda})F^{(2)})$, where, $F^{(1)}_{3\times3} = \text{Diag}(0.1, 0.5, 0.9)$ and $F^{(2)}_{3\times3} = 0.1\mathbf{J}_2$, where, $\mathbf{J}_2$ is a 2×2 matrix of all 1's and ν varies from 1 to 15 to give networks of different density. So, we get $\rho_n = \boldsymbol{\pi}^T\mathbf{F}\boldsymbol{\pi}$. We now, vary $\tilde{\lambda}$ to get different combinations of $\mathbf{F}$ as well as ρ_n.

In the following figures, we try to see the behavior of mean and variances of the count statistics, as we vary λ_n as we vary ν.

Application to Real Network Data

Facebook Collegiate Network

In this application, we try to find communities for Facebook collegiate networks. The networks were presented in the paper by Traud et.al. (2011) [155]. The network is formed by Facebook users acting as nodes and if two Facebook users are "friends" there is an edge between the corresponding nodes. Along with the network structure, we also have the data on covariates of the nodes. Each node has covariates: gender, class year, and data fields that



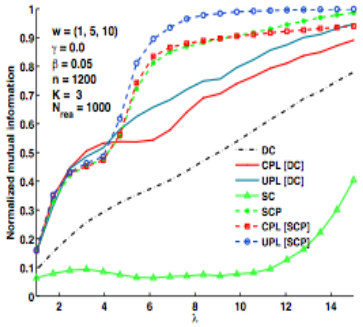


Figure 5.2: The LHS is the performance of graph distance based method and RHS is the performance of Pseudolikelihood method on same generative SBM.

Figure 5.3: The LHS is community allocation and RHS is the one estimated by graph distance for Facebook Caltech network with 3 dorms.

represent (using anonymous numerical identifiers) high school, major, and dormitory residence. We consider the network of a specific college (Caltech). We compare the communities found with the dormitory affiliation of the nodes.

Political Web Blogs Network

This dataset on political blogs was compiled by [1] soon after the 2004 U.S. presidential election. The nodes are blogs focused on US politics and the edges are hyperlinks between these blogs. Each blog was manually labeled as liberal or conservative by [1], and we treat these as true community labels. We ignore directions of the hyperlinks and analyze the largest connected component of this network, which has 1222 nodes and the average degree of 27. The distribution of degrees is highly skewed to the right (the median degree is 13, and the maximum is 351). This is a network where the degree distribution is heavy-tailed and the graph is inhomogeneous.

Figure 5.4: The LHS is community allocation and RHS is the one estimated by graph distance for Political Web blogs Network.

5.6 Conclusion

The proposed graph distance based community detection algorithm gives a very general way for community detection for graphs over a large range of densities - from very sparse graphs to very dense graphs. We theoretically prove the efficacy of the method under the model that the graph is generated from stochastic block model with fixed number of blocks. We prove that the proportion of mislabeled communities goes to zero as the number of vertices $n \to \infty$. This result is true for graphs coming from stochastic block model under certain conditions on the stochastic block model parameters. These conditions are satisfied above the threshold of block identification for two blocks as given in [125]. The condition (C1) of $\mathbf{1}$ not being the eigenvector of $\tilde{K}$ for our community identification result to hold, seems to be an artificial one, as simulation suggests that our method is able to identify communities, even when $\mathbf{1}$ is an eigenvector of $\tilde{K}$.

We demonstrate the empirical performance of the method by using both simulated and real world networks. We compare with the pseudo-likelihood method and show that they have similar empirical performances. We demonstrate the empirical performance by applying the method for community detection in several real world networks too.

The method also works when number of blocks in the stochastic block model brows with n (number of vertices) and for degree-corrected block model [91]. We conjecture that under these models too the method will have the theoretical guarantee of correct community detection. The proof can be obtained by using similar techniques that we have used in this chapter.

Bibliography

[1] Lada A Adamic and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog". In: *Proceedings of the 3rd international workshop on Link discovery*. ACM. 2005, pp. 36–43.

[2] Edoardo M Airoldi et al. "Mixed membership stochastic block models for relational data with application to protein-protein interactions". In: *Proceedings of the international biometrics society annual meeting*. 2006.

[3] Edoardo M Airoldi et al. "Mixed membership stochastic blockmodels". In: *The Journal of Machine Learning Research* 9 (2008), pp. 1981–2014.

[4] David J Aldous. "Representations for partially exchangeable arrays of random variables". In: *Journal of Multivariate Analysis* 11.4 (1981), pp. 581–598.

[5] Arash A Amini et al. "Pseudo-likelihood methods for community detection in large sparse networks". In: (2012).

[6] TW Anderson, Huang Hsu, and Kai-Tai Fang. "Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions". In: *Canadian Journal of Statistics* 14.1 (1986), pp. 55–59.

[7] Krishna B Athreya and Peter E Ney. *Branching processes*. Vol. 28. Springer-Verlag Berlin, 1972.

[8] Fadoua Balabdaoui and Jon A. Wellner. "Estimation of a k-monotone density: limit distribution theory and the spline connection". In: *Ann. Statist.* 35.6 (2007), pp. 2536–2564. ISSN: 0090-5364. DOI: 10.1214/009053607000000262. URL: http://dx.doi.org/10.1214/009053607000000262.

[9] Brian Ball, Brian Karrer, and MEJ Newman. "Efficient and principled method for detecting communities in networks". In: *Physical Review E* 84.3 (2011), p. 036103.

[10] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[11] Albert-Laszlo Barabâsi et al. "Evolution of the social network of scientific collaborations". In: *Physica A: Statistical Mechanics and its Applications* 311.3 (2002), pp. 590–614.

[12] Heather Battey and Oliver Linton. "Nonparametric estimation of multivariate elliptic densities via finite mixture sieves". In: (2012).

[13] Jeff Baumes et al. "Discovering hidden groups in communication networks". In: *Intelligence and Security Informatics*. Springer, 2004, pp. 378–389.

[14] Peter S Bearman, James Moody, and Katherine Stovel. "Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks1". In: *American Journal of Sociology* 110.1 (2004), pp. 44–91.

[15] A. Ben-Hur, A. Elisseeff, and I. Guyon. "A stability based method for discovering structure in clustered data". In: *Pacific Symposium on Biocomputing 2002: Kauai, Hawaii, 3-7 January 2002*. World Scientific Pub Co Inc. 2001, p. 6.

[16] Shankar Bhamidi. "First passage percolation on locally treelike networks. I. Dense random graphs". In: *Journal of Mathematical Physics* 49 (2008), p. 125218.

[17] Shankar Bhamidi, Guy Bresler, and Allan Sly. "Mixing time of exponential random graphs". In: *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*. IEEE. 2008, pp. 803–812.

[18] Shankar Bhamidi, Remco Van der Hofstad, and Gerard Hooghiemstra. "First Passage Percolation on the Erds-Renyi Random Graph". In: *Combinatorics, Probability & Computing* 20.5 (2011), pp. 683–707.

[19] P. J. Bickel. "On adaptive estimation". In: *Ann. Statist.* 10.3 (1982), pp. 647–671. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(198209)10:3<647:OAE>2.0.CO;2-1&origin=MSN.

[20] P. J. Bickel, F. Götze, and W. R. van Zwet. "Resampling fewer than n observations: gains, losses, and remedies for losses". In: *Statist. Sinica* 7.1 (1997). Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995), pp. 1–31. ISSN: 1017-0405.

[21] Peter Bickel et al. "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *arXiv preprint arXiv:1207.0865* (2012).

[22] Peter J Bickel and Aiyou Chen. "A nonparametric view of network models and Newman–Girvan and other modularities". In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21068–21073.

[23] Peter J Bickel, Aiyou Chen, and Elizaveta Levina. "The method of moments and degree distributions for network models". In: *The Annals of Statistics* 39.5 (2011), pp. 2280–2301.

[24] Peter J. Bickel and Elizaveta Levina. "Covariance regularization by thresholding". In: *Ann. Statist.* 36.6 (2008), pp. 2577–2604. ISSN: 0090-5364. DOI: 10.1214/08-AOS600. URL: http://dx.doi.org/10.1214/08-AOS600.

[25] Peter J. Bickel and Elizaveta Levina. "Regularized estimation of large covariance matrices". In: *Ann. Statist.* 36.1 (2008), pp. 199–227. ISSN: 0090-5364. DOI: 10.1214/009053607000000758. URL: http://dx.doi.org/10.1214/009053607000000758.

[26] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. "Simultaneous analysis of lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (2009), pp. 1705–1732. ISSN: 0090-5364. DOI: 10.1214/08-AOS620. URL: http://dx.doi.org/10.1214/08-AOS620.

[27] Peter J Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[28] Peter J. Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Johns Hopkins University Press, 1993, pp. xxii+560. ISBN: 0-8018-4541-6.

[29] Béla Bollobás, Svante Janson, and Oliver Riordan. "The phase transition in inhomogeneous random graphs". In: *Random Structures & Algorithms* 31.1 (2007), pp. 3–122.

[30] Christian Borgs et al. "Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing". In: *Advances in Mathematics* 219.6 (2008), pp. 1801–1851.

[31] G. Brock et al. "clValid: An R package for cluster validation". In: *Journal of Statistical Software* 25.4 (2008), pp. 1–22.

[32] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Methods, theory and applications. Heidelberg: Springer, 2011, pp. xviii+556. ISBN: 978-3-642-20191-2. DOI: 10.1007/978-3-642-20192-9. URL: http://dx.doi.org/10.1007/978-3-642-20192-9.

[33] Aydın Buluç, John R Gilbert, and Ceren Budak. "Solving path problems on the GPU". In: *Parallel Computing* 36.5 (2010), pp. 241–253.

[34] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. "Optimal rates of convergence for covariance matrix estimation". In: *Ann. Statist.* 38.4 (2010), pp. 2118–2144. ISSN: 0090-5364. DOI: 10.1214/09-AOS752. URL: http://dx.doi.org/10.1214/09-AOS752.

[35] T. Caliński and J. Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-Theory and Methods* 3.1 (1974), pp. 1–27.

[36] Emmanuel Candes and Terence Tao. "The Dantzig selector: statistical estimation when p is much larger than n". In: *Ann. Statist.* 35.6 (2007), pp. 2313–2351. ISSN: 0090-5364. DOI: 10.1214/009053606000001523. URL: http://dx.doi.org/10.1214/009053606000001523.

[37] Alain Celisse, J-J Daudin, and Laurent Pierre. "Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model". In: *arXiv:1105.3288* (2011).

[38] Françoise Chatelin. *Spectral Approximation of Linear Operators*. SIAM, 1983.

[39] Sourav Chatterjee and Persi Diaconis. "Estimating and understanding exponential random graph models". In: *arXiv preprint arXiv:1102.2650* (2011).

[40] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. "Spectral clustering of graphs with general degrees in the extended planted partition model". In: *Journal of Machine Learning Research* 2012 (2012), pp. 1–23.

[41] Aiyou Chen et al. "Fitting community models to large sparse networks". In: *arXiv preprint arXiv:1207.2340* (2012).

[42] Yilun Chen, Ami Wiesel, and Alfred O. Hero III. "Robust shrinkage estimation of high-dimensional covariance matrices". In: *IEEE Trans. Signal Process.* 59.9 (2011), pp. 4097–4107. ISSN: 1053-587X. DOI: 10.1109/TSP.2011.2138698. URL: http://dx.doi.org/10.1109/TSP.2011.2138698.

[43] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. "Finding community structure in very large networks". In: *Physical review E* 70.6 (2004), p. 066111.

[44] Amin Coja-Oghlan and André Lanka. "Finding planted partitions in random graphs with general degree distributions". In: *SIAM Journal on Discrete Mathematics* 23.4 (2009), pp. 1682–1714.

[45] H Cui and X He. "The consistence of semiparametric estimation of elliptic densities". In: *Acta Math. Sin. New Ser* 11 (1995), pp. 44–58.

[46] Madeleine Cule, Richard Samworth, and Michael Stewart. "Maximum likelihood estimation of a multi-dimensional log-concave density". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72.5 (2010), pp. 545–607. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2010.00753.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2010.00753.x.

[47] J-J Daudin, Franck Picard, and Stéphane Robin. "A mixture model for random graphs". In: *Statistics and computing* 18.2 (2008), pp. 173–183.

[48] Chandler Davis and William Morton Kahan. "The rotation of eigenvectors by a perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pp. 1–46.

[49] Richard Davis, Oliver Pfaffel, and Robert Stelzer. "Limit Theory for the largest eigenvalues of sample covariance matrices with heavy-tails". In: *arXiv: 1108.5464* (2011).

[50] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Physical Review E* 84.6 (2011), p. 066106.

[51] Persi Diaconis and Svante Janson. "Graph limits and exchangeable random graphs". In: *arXiv preprint arXiv:0712.2749* (2007).

[52] Pedro Domingos. "Mining social networks for viral marketing". In: *IEEE Intelligent Systems* 20.1 (2005), pp. 80–82.

[53] Lutz Dümbgen and Kaspar Rufibach. "Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency". In: *Bernoulli* 15.1 (2009), pp. 40–68. ISSN: 1350-7265. DOI: 10.3150/08-BEJ141. URL: http://dx.doi.org/10.3150/08-BEJ141.

[54] Richard Durrett. *Random graph dynamics*. Vol. 20. Cambridge university press, 2007.

[55] Bradley Efron et al. "Least angle regression". In: *Ann. Statist.* 32.2 (2004). With discussion, and a rejoinder by the authors, pp. 407–499. ISSN: 0090-5364. DOI: 10.1214/009053604000000067. URL: http://dx.doi.org/10.1214/009053604000000067.

[56] Noureddine El Karoui. "Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond". In: *Ann. Appl. Probab.* 19.6 (2009), pp. 2362–2405. ISSN: 1050-5164. DOI: 10.1214/08-AAP548. URL: http://dx.doi.org/10.1214/08-AAP548.

[57] Noureddine El Karoui. "Operator norm consistent estimation of large-dimensional sparse covariance matrices". In: *Ann. Statist.* 36.6 (2008), pp. 2717–2756. ISSN: 0090-5364. DOI: 10.1214/07-AOS559. URL: http://dx.doi.org/10.1214/07-AOS559.

[58] Paul Erdős and Alfréd Rényi. "On random graphs". In: *Publicationes Mathematicae Debrecen* 6 (1959), pp. 290–297.

[59] Paul Erdos and Alfréd Rényi. "On the evolution of random graphs". In: *Bull. Inst. Internat. Statist* 38.4 (1961), pp. 343–347.

[60] Kai Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric multivariate and related distributions*. Vol. 36. Monographs on Statistics and Applied Probability. London: Chapman and Hall Ltd., 1990, pp. x+220. ISBN: 0-412-31430-4.

[61] R.A. Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of Human Genetics* 7.2 (1936), pp. 179–188.

[62] Robert W Floyd. "Algorithm 97: shortest path". In: *Communications of the ACM* 5.6 (1962), p. 345.

[63] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3 (2010), pp. 75–174.

[64] C. Fraley and A.E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis". In: *The computer journal* 41.8 (1998), p. 578.

[65] Ove Frank. "Network sampling and model fitting". In: *Models and methods in social network analysis* (2005), pp. 31–56.

[66] Ove Frank and David Strauss. "Markov graphs". In: *Journal of the american Statistical association* 81.395 (1986), pp. 832–842.

[67] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[68] Jerome Friedman and Robert Tibshirani. "The monotone smoothing of scatterplots". In: *Technometrics* 26.3 (1984), pp. 243–250.

[69] Edgar N Gilbert. "Random graphs". In: *The Annals of Mathematical Statistics* 30.4 (1959), pp. 1141–1144.

[70] Ulf Grenander. "On the theory of mortality measurement. II". In: *Skand. Aktuarietidskr.* 39 (1956), 125–153 (1957).

[71] P. Groeneboom. "Estimating a monotone density". In: *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983).* Wadsworth Statist./Probab. Ser. Belmont, CA: Wadsworth, 1985, pp. 539–555.

[72] Mayiz B Habbal, Haris N Koutsopoulos, and Steven R Lerman. "A decomposition algorithm for the all-pairs shortest path problem on massively parallel computer architectures". In: *Transportation Science* 28.4 (1994), pp. 292–308.

[73] Frank R. Hampel et al. *Robust statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. The approach based on influence functions. New York: John Wiley & Sons Inc., 1986, pp. xxiv+502. ISBN: 0-471-82921-8.

[74] Mark S Handcock and Krista J Gile. "Modeling social networks from sampled data". In: *The Annals of Applied Statistics* 4.1 (2010), pp. 5–25.

[75] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. "Model-based clustering for social networks". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.2 (2007), pp. 301–354.

[76] J.A. Hartigan. "Statistical theory in clustering". In: *Journal of classification* 2.1 (1985), pp. 63–76.

[77] Kenneth R Hess et al. "Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer". In: *Journal of clinical oncology* 24.26 (2006), pp. 4236–4244.

[78] Shawndra Hill, Foster Provost, and Chris Volinsky. "Network-based marketing: Identifying likely adopters via consumer networks". In: *Statistical Science* (2006), pp. 256–276.

[79] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. "Latent space approaches to social network analysis". In: *Journal of the american Statistical association* 97.460 (2002), pp. 1090–1098.

[80] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social networks* 5.2 (1983), pp. 109–137.

[81] Hajo Holzmann, Axel Munk, and Tilmann Gneiting. "Identifiability of finite mixtures of elliptical distributions". In: *Scandinavian journal of statistics* 33.4 (2006), pp. 753–763.

[82] Douglas N Hoover. "Relations on probability spaces and arrays of random variables". In: *Institute for Advanced Study, Princeton, NJ* (1979).

[83] Cho-Jui Hsieh et al. "Sparse inverse covariance matrix estimation using quadratic approximation". In: *arXiv preprint arXiv:1306.3212* (2013).

[84] Peter J. Huber. *Robust statistics*. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons Inc., 1981, pp. ix+308. ISBN: 0-471-41805-6.

[85] David R Hunter and Mark S Handcock. "Inference in curved exponential family models for networks". In: *Journal of Computational and Graphical Statistics* 15.3 (2006), pp. 565–583.

[86] Donald B Johnson. "Efficient algorithms for shortest paths in sparse networks". In: *Journal of the ACM (JACM)* 24.1 (1977), pp. 1–13.

[87] James Holland Jones and Mark S Handcock. "Social networks (communication arising): Sexual contacts and epidemic thresholds". In: *Nature* 423.6940 (2003), pp. 605–606.

[88] Marianne A. Jonker and Aad W. van der Vaart. "A semi-parametric model for censored and passively registered data". In: *Bernoulli* 7.1 (2001), pp. 1–31. ISSN: 1350-7265. DOI: 10.2307/3318600. URL: http://dx.doi.org/10.2307/3318600.

[89] O. Kallenberg. *Probabilistic symmetries and invariance principles*. Springer Verlag, 2005. ISBN: 0387251154.

[90] Yutaka Kano. "Consistency property of elliptical probability density functions". In: *J. Multivariate Anal.* 51.1 (1994), pp. 139–147. ISSN: 0047-259X. DOI: 10.1006/jmva.1994.1054. URL: http://dx.doi.org/10.1006/jmva.1994.1054.

[91] Brian Karrer and Mark EJ Newman. "Stochastic blockmodels and community structure in networks". In: *Physical Review E* 83.1 (2011), p. 016107.

[92] Nadav Kashtan et al. "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs". In: *Bioinformatics* 20.11 (2004), pp. 1746–1758.

[93] Tosio Katō. *Perturbation theory for linear operators*. Vol. 132. springer, 1995.

[94] John T. Kent and David E. Tyler. "Redescending M-estimates of multivariate location and scatter". In: *Ann. Statist.* 19.4 (1991), pp. 2102–2119. ISSN: 0090-5364. DOI: 10.1214/aos/1176348388. URL: http://dx.doi.org/10.1214/aos/1176348388.

[95] CG Khatri and C Radhakrishna Rao. "Solutions to some functional equations and their applications to characterization of probability distributions". In: *Sankhyā: The Indian Journal of Statistics, Series A* (1968), pp. 167–180.

[96] J. Kiefer and J. Wolfowitz. "Asymptotically minimax estimation of concave and convex distribution functions". In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 34.1 (1976), pp. 73–85.

[97] Roger Koenker and Ivan Mizera. "Quasi-concave density estimation". In: *Ann. Statist.* 38.5 (2010), pp. 2998–3027. ISSN: 0090-5364. DOI: 10.1214/10-AOS814. URL: http://dx.doi.org/10.1214/10-AOS814.

[98] E.D. Kolaczyk. *Statistical analysis of network data: methods and models*. Springer Verlag, 2009. ISBN: 038788145X.

[99] János Komlós, Péter Major, and Gábor Tusnády. "An Approximation of Partial Sums of Independent rv'-s, and the Sample df. I". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 32.1-2 (1975), pp. 111–131.

[100] Florent Krzakala et al. "Spectral redemption: clustering sparse networks". In: *arXiv preprint arXiv:1306.5550* (2013).

[101] Clifford Lam and Jianqing Fan. "Sparsistency and rates of convergence in large covariance matrix estimation". In: *Ann. Statist.* 37.6B (2009), pp. 4254–4278. ISSN: 0090-5364. DOI: 10.1214/09-AOS720. URL: http://dx.doi.org/10.1214/09-AOS720.

[102] Andrea Lancichinetti and Santo Fortunato. "Community detection algorithms: a comparative analysis". In: *Physical review E* 80.5 (2009), p. 056117.

[103] T. Lange et al. "Stability-based validation of clustering solutions". In: *Neural computation* 16.6 (2004), pp. 1299–1323.

[104] Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. "Overlapping stochastic block models with application to the french political blogosphere". In: *The Annals of Applied Statistics* 5.1 (2011), pp. 309–336.

[105] Olivier Ledoit and Michael Wolf. "A well-conditioned estimator for large-dimensional covariance matrices". In: *J. Multivariate Anal.* 88.2 (2004), pp. 365–411. ISSN: 0047-259X. DOI: 10.1016/S0047-259X(03)00096-4. URL: http://dx.doi.org/10.1016/S0047-259X(03)00096-4.

[106] Charles E Leiserson et al. *Introduction to algorithms*. The MIT press, 2001.

[107] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graph evolution: Densification and shrinking diameters". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), p. 2.

[108] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. "Graphs over time: densification laws, shrinking diameters and possible explanations". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 177–187.

[109] Jure Leskovec, Kevin J Lang, and Michael Mahoney. "Empirical comparison of algorithms for network community detection". In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 631–640.

[110] Jure Leskovec et al. "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters". In: *Internet Mathematics* 6.1 (2009), pp. 29–123.

[111] Jure Leskovec et al. "Statistical properties of community structure in large social and information networks". In: *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, pp. 695–704.

[112] Eckhard Liebscher. "A semiparametric density estimator based on elliptical distributions". In: *J. Multivariate Anal.* 92.1 (2005), pp. 205–225. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2003.09.007. URL: http://dx.doi.org/10.1016/j.jmva.2003.09.007.

[113] László Lovász. *Large networks and graph limits*. Vol. 60. American Mathematical Soc., 2012.

[114] Ricardo A. Maronna and Víctor J. Yohai. "The behavior of the Stahel-Donoho robust multivariate estimator". In: *J. Amer. Statist. Assoc.* 90.429 (1995), pp. 330–341. ISSN: 0162-1459. URL: http://links.jstor.org/sici?sici=0162-1459(199503)90:429<330:TBOTSR>2.0.CO;2-Q&origin=MSN.

[115] Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. "Topic and role discovery in social networks with experiments on enron and academic email." In: *J. Artif. Intell. Res.(JAIR)* 30 (2007), pp. 249–272.

[116] Frank McSherry. "Spectral partitioning of random graphs". In: *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*. IEEE. 2001, pp. 529–537.

[117] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *Ann. Statist.* 34.3 (2006), pp. 1436–1462. ISSN: 0090-5364. DOI: 10.1214/009053606000000281. URL: http://dx.doi.org/10.1214/009053606000000281.

[118] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72.4 (2010), pp. 417–473. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2010.00740.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x.

[119] Nicolai Meinshausen and Bin Yu. "Lasso-type recovery of sparse representations for high-dimensional data". In: *Ann. Statist.* 37.1 (2009), pp. 246–270. ISSN: 0090-5364. DOI: 10.1214/07-AOS582. URL: http://dx.doi.org/10.1214/07-AOS582.

[120] Stanley Milgram. "The small world problem". In: *Psychology today* 2.1 (1967), pp. 60–67.

[121] G.W. Milligan and M.C. Cooper. "An examination of procedures for determining the number of clusters in a data set". In: *Psychometrika* 50.2 (1985), pp. 159–179.

[122] Ron Milo et al. "Network motifs: simple building blocks of complex networks". In: *Science* 298.5594 (2002), pp. 824–827.

[123] S. Monti et al. "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data". In: *Machine learning* 52.1 (2003), pp. 91–118.

[124] Martina Morris and Mirjam Kretzschmar. "Concurrent partnerships and the spread of HIV". In: *Aids* 11.5 (1997), pp. 641–648.

[125] Elchanan Mossel, Joe Neeman, and Allan Sly. "Stochastic block models and reconstruction". In: *arXiv preprint arXiv:1202.1499* (2012).

[126] Elchanan Mossel, Joe Neeman, and Omer Tamuz. "Majority dynamics and aggregation of information in social networks". In: *Autonomous Agents and Multi-Agent Systems* (2012), pp. 1–22.

[127] Jennifer Neville and David Jensen. "Collective classification with relational dependency networks". In: *Proceedings of the Second International Workshop on Multi-Relational Data Mining*. Citeseer. 2003, pp. 77–91.

[128] Mark Newman. *Networks: an introduction*. OUP Oxford, 2009.

[129] Mark EJ Newman. "Finding community structure in networks using the eigenvectors of matrices". In: *Physical review E* 74.3 (2006), p. 036104.

[130] Mark EJ Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23 (2006), pp. 8577–8582.

[131] Mark EJ Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical review E* 69.2 (2004), p. 026113.

[132] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. "Random graph models of social networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.Suppl 1 (2002), pp. 2566–2572.

[133] MEJ Newman. "Spectral community detection in sparse networks". In: *arXiv preprint arXiv:1308.6494* (2013).

[134] Joshua OÕMadadhain, Padhraic Smyth, and Lada Adamic. "Learning predictive models for link formation". In: *International Sunbelt Social Network Conference*. 2005.

[135] Jayanta Kumar Pal and Michael Woodroofe. "Large sample properties of shape restricted regression estimators with smoothness adjustments". In: *Statistica Sinica* 17.4 (2007), p. 1601.

[136] Franck Picard et al. "Assessing the exceptionality of network motifs". In: *Journal of Computational Biology* 15.1 (2008), pp. 1–20.

[137] J.W. Richards et al. "On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data". In: *The Astrophysical Journal* 733 (2011), p. 10.

[138] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd., 1988, pp. xx+521. ISBN: 0-471-91787-7.

[139] Garry Robins et al. "Recent developments in exponential random graph (¡ i¿ p¡/i¿*) models for social networks". In: *Social networks* 29.2 (2007), pp. 192–215.

[140] Karl Rohe, Sourav Chatterjee, and Bin Yu. "Spectral clustering and the high- dimensional stochastic blockmodel". In: *The Annals of Statistics* 39.4 (2011), pp. 1878–1915.

[141] Adam J. Rothman, Elizaveta Levina, and Ji Zhu. "Generalized thresholding of large covariance matrices". In: *J. Amer. Statist. Assoc.* 104.485 (2009), pp. 177–186. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.0101. URL: http://dx.doi.org/10.1198/jasa.2009.0101.

[142] Kaspar Rufibach. "Computing maximum likelihood estimators of a log-concave density function". In: *J. Stat. Comput. Simul.* 77.7-8 (2007), pp. 561–574. ISSN: 0094-9655. DOI: 10.1080/10629360600569097. URL: http://dx.doi.org/10.1080/10629360600569097.

[143] Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. "Theoretical justification of popular link prediction heuristics". In: *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*. AAAI Press. 2011, pp. 2722–2727.

[144] Purnamrita Sarkar and Andrew W Moore. "Dynamic social network analysis using latent space models". In: *ACM SIGKDD Explorations Newsletter* 7.2 (2005), pp. 31–40.

[145] Tom AB Snijders et al. "New specifications for exponential random graph models". In: *Sociological methodology* 36.1 (2006), pp. 99–153.

[146] Edgar Solomonik, Aydın Buluç, and James Demmel. "Minimizing communication in all-pairs shortest paths". In: *University of California at Berkeley, Berkeley, US* (2012).

[147] Nikhil Srivastava and Roman Vershynin. "Covariance Estimation for Distributions with 2+\ epsilon Moments". In: *arXiv preprint arXiv:1106.2775* (2011).

[148] Winfried Stute and U Werner. "Nonparametric estimation of elliptically contoured densities". In: *Nonparametric Functional Estimation and Related Topics*. Springer, 1991, pp. 173–190.

[149] Daniel L Sussman et al. "A consistent adjacency spectral embedding for stochastic blockmodel graphs". In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1119–1128.

[150] Chim Tantiyaswasdikul and Michael B Woodroofe. "Isotonic smoothing splines under sequential designs". In: *Journal of Statistical Planning and Inference* 38.1 (1994), pp. 75–87.

[151] SK Thompson. *Sampling – 3rd Ed.* Vol. 755. John Wiley, Sons: Wiley series in probability, and statistics, 2012.

[152] Steven K Thompson and Ove Frank. "Model-based estimation with link-tracing sampling designs". In: *Survey Methodology* 26.1 (2000), pp. 87–98.

[153] R. Tibshirani and G. Walther. "Cluster validation by prediction strength". In: *Journal of Computational and Graphical Statistics* 14.3 (2005), pp. 511–528.

[154] R. Tibshirani, G. Walther, and T. Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[155] Amanda L Traud et al. "Comparing community structure to characteristics in online collegiate social networks". In: *SIAM review* 53.3 (2011), pp. 526–543.

[156] David E. Tyler. "A distribution-free M-estimator of multivariate scatter". In: *Ann. Statist.* 15.1 (1987), pp. 234–251. ISSN: 0090-5364. DOI: 10.1214/aos/1176350263. URL: http://dx.doi.org/10.1214/aos/1176350263.

[157] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. With applications to statistics. New York: Springer-Verlag, 1996, pp. xvi+508. ISBN: 0-387-94640-3.

[158] Guenther Walther. "Inference and modeling with log-concave distributions". In: *Statist. Sci.* 24.3 (2009), pp. 319–327. ISSN: 0883-4237. DOI: 10.1214/09-STS303. URL: http://dx.doi.org/10.1214/09-STS303.

[159] Yang Wang et al. "Epidemic spreading in real networks: An eigenvalue viewpoint". In: *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on.* IEEE. 2003, pp. 25–34.

[160] Stephen Warshall. "A theorem on boolean matrices". In: *Journal of the ACM (JACM)* 9.1 (1962), pp. 11–12.

[161] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications.* Vol. 8. Cambridge university press, 1994.

[162] Sebastian Wernicke. "Efficient detection of network motifs". In: *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 3.4 (2006), pp. 347–359.

[163] Ami Wiesel. "Unified framework to regularized covariance estimation in scaled Gaussian models". In: *IEEE Trans. Signal Process.* 60.1 (2012), pp. 29–38. ISSN: 1053-587X. DOI: 10.1109/TSP.2011.2170685. URL: http://dx.doi.org/10.1109/TSP.2011.2170685.

[164] Daniela M Witten, Jerome H Friedman, and Noah Simon. "New insights and faster computations for the graphical lasso". In: *Journal of Computational and Graphical Statistics* 20.4 (2011), pp. 892–900.

[165] Wayne W Zachary. "An information flow model for conflict and fission in small groups". In: *Journal of anthropological research* (1977), pp. 452–473.

[166] Peng Zhao and Bin Yu. "On model selection consistency of Lasso". In: *J. Mach. Learn. Res.* 7 (2006), pp. 2541–2563. ISSN: 1532-4435.