UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

CRISTEL ECATERIN VERA TAPIA

ESTIMAÇÃO DO NÚMERO DE COMUNIDADES NO MODELO ESTOCÁSTICO DE BLOCOS COM CORREÇÃO DE GRAU

Tese apresentada ao Departamento de Estatística - DEs/UFSCar e ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Estatística - Programa Interinstitucional de Pós - Graduação em Estatística DEs-UFSCar/ICMC-USP.

Orientador: Prof. Dr. Alexsandro Giacomo Grimbert Gallo

Coorientadora: Profa. Dra. Florencia Graciela Leonardi

São Carlos

Fevereiro de 2023

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

CRISTEL ECATERIN VERA TAPIA

ESTIMATION OF THE NUMBER OF COMMUNITIES IN THE DEGREE CORRECTED STOCHASTIC BLOCK MODEL

Doctoral dissertation submitted to the Departamento de Estatística - DEs-UFSCar and to the Instituto de Ciências Matemáticas e de Computação - ICMC - USP, in partial fulfillment requirements for the degree of the Doctorate Interagency Program Graduate in Statistics DEs-UFSCar/ ICMC-USP

Advisor: Prof. Dr. Alexsandro Giacomo Grimbert Gallo

Coadvisor: Profa. Dra. Florencia Graciela Leonardi

São Carlos

February 2023

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Cristel Ecaterin Vera Tapia, realizada em 14/12/2022.

Comissão Julgadora:

Prof. Dr. Alexsandro Giacomo Grimbert Gallo (UFSCar)

Profa. Dra. Andressa Cerqueira (UFSCar)

Prof. Dr. Jefferson Antonio Galves (IME-USP)

Prof. Dr. Aline Duarte de Oliveira (IME-USP)

Profa. Dra. Nancy Lopes Garcia (UNICAMP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Agradecimentos

Em primeiro lugar, quero agradecer a Deus, por estar comigo sempre, sendo meu refúgio e fortaleza nos momentos mais difíceis durante esta caminhada.

Agradeço especialmente aos meus orientadores Sandro Gallo e Florencia Leonardi por todo o conhecimento transmitido, pela paciência, coompreensão e principalmente pelo apoio e preocupação nos momentos de crise.

À minha mãe, que sempre acreditou em mim e não permitiu que eu desistisse, seu carinho e apoio incondicional me deram muita força e coragem para seguir. Ao meu pai, por todo o carinho e por sempre estar presente em minha vida.

A todos os professores do curso, que compartilharam os seus conhecimentos.

O presente trabalho foi realizado com apoio da Coordinação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 e como parte das atividades do Centro de Pesquisa, Inovação e Difusão em Neuromatemática (CEPID NeuroMat), processo #2013/ 07699-0, Fundação de Amparo à Pesquisa do estado de São Paulo (FAPESP).

Também agradeço o apoio financeiro do projeto CNPq Universal 432310/2018-5.

RESUMO

VERA TAPIA, C. E. Estimação do número de comunidades no modelo estocástico de blocos com correção de grau. 2022. 75 p. Tese (Doutorado em Estatística – Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

O modelo estocástico de blocos (SBM, do inglês *Stochastic Block Model*) é um modelo de grafos aleatórios em que o conjunto de vértices é dividido em blocos, e a probabilidade de conexão entre cada par de vértices depende dos blocos aos quais os vértices pertencem. O SBM foi introduzido por Holland et al. (1983), é tipicamente aplicado em grafos simples, com cada entrada da matriz de adjacência seguindo uma distribuição de Bernoulli. Karrer e Newman (2011) estenderam o modelo em duas direções: definiram o modelo multigrafo ou também conhecido como modelo estocástico de blocos de Poisson (Poisson SBM), em que as entradas da matriz de adjacência seguem a distribuição de Poisson, e introduziram o modelo estocástico de blocos com correção de grau (DCSBM, do inglês *Degree Corrected Stochastic Block Model*), que permite que a distribuição dos graus dos vértices dependa também dos vértices, e não somente dos blocos aos quais pertencem.
A presente tese é dedicada ao problema de estimação do número de comunidades no Poisson SBM e no DCSBM. Consideramos o regime denso, no qual a probabilidade de conexão entre pares de vértices não depende do tamanho do grafo e também o regime semi-esparso, no qual a probabilidade de conexão entre pares de vértices pode decair para 0 (numa certa taxa) com o tamanho do grafo. Neste contexto geral, provamos que o estimador do número de comunidades introduzido por Cerqueira e Leonardi (2020) (com as devidas alterações) é fortemente consistente.

Palavras-chave: Modelo estocástico de blocos de Poisson, Modelo estocástico de blocos com correção de grau, Estimação do número de comunidades, regime semi-esparso.

ABSTRACT

VERA TAPIA, C. E. Estimation of the number of communities in degree corrected stochastic block model. 2022. 75 p. Tese (Doutorado em Estatística – Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

The stochastic block model (SBM) is a random graph model that splits the set of vertices into blocks, and the probability connection between each pair of vertices depends on the blocks to which the vertices belong. The SBM was introduced by Holland et al. (1983) and it is traditionally applied to simple graphs, with each entry in the adjacency matrix following the Bernoulli distribution. Karrer and Newman (2011) extended the model in two directions: they defined the multigraph model (Poisson SBM), in which the entries of the adjacency matrix follow the Poisson distribution, and introduced the degree corrected stochastic block model (DCSBM) that allows the degree distribution of vertices also depend on the vertices, and not just on the blocks they belong to.
This thesis is devoted to the problem of estimating the number of communities in the Poisson SBM and DCSBM. We consider the dense regime, in which the probability of connection between pairs of vertices does not depend on the size of the graph, or even the semi-sparse regime, in which the probability of connection between pairs of vertices can decay to 0 (at a certain rate) with the size of the graph. In this general context, we prove that the estimator of the number of communities introduced by Cerqueira and Leonardi (2020) (with the necessary changes) is still strongly consistent.

Keywords: Poisson stochastic block model, Degree corrected stochastic block model, Estimation of the number of communities, semi-sparse regime.

Contents

Chapter

1

Introdução

Grafo é uma estrutura matemática muito utilizada para representar redes. Consiste de um conjunto de objetos chamados vértices e um conjunto de relações chamadas arestas. Existem inúmeros exemplos de aplicações e utilizações em fenômenos da vida real (citamos por exemplo, redes de comunicação, sistemas de navegação, segurança de redes de computadores, redes de transporte, etc). Para fixar ideias, iremos utilizar o exemplo de uma rede de transporte aéreo. Neste contexto, os vértices representam os aeroportos e as arestas representam o fato de que pelo menos um voo ocorreu entre um par de aeroportos.

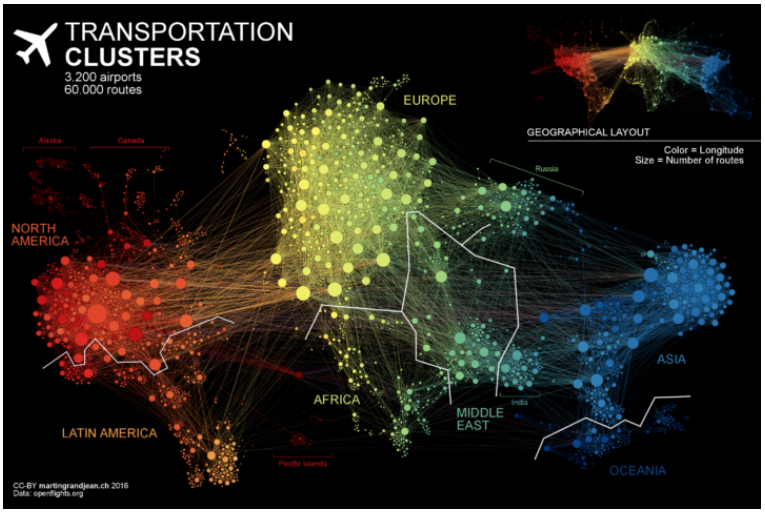

Figure 1.1: Rede de transporte aéreo.
Fonte: Grandjean (2016).

A Figura 1.1 mostra a análise de uma rede de transporte aéreo (malha de tráfego aéreo) feita por Grandjean (2016), em pelo menos 3200 aeroportos ao redor do mundo, as cores indicam comunidades e o tamanho dos pontos indica o número de rotas de aeroportos.

Formalmente, um grafo é um par $G = (V, E)$, onde $V = \{1, \ldots, n\}$ é um conjunto finito de vértices e $E = \{(i, j) : \text{existe uma aresta conectando i e j}\}$ é um conjunto de pares de vértices chamados arestas.

Um grafo G, no caso mais simples pode ser representado através de uma matriz de adjacência $X = (X_{ij})_{i,j\in[n]} \in \{0, 1\}^{n\times n}$, com entradas $X_{ij} = 1$ se $(i, j) \in E$ ou $X_{ij} = 0$ caso contrário. Se G é não orientado, isto é, se $(i, j) \in E \Rightarrow (j, i) \in E$, então X é simétrica.

Graficamente, um grafo não orientado pode ser representado pela Figura 1.2.

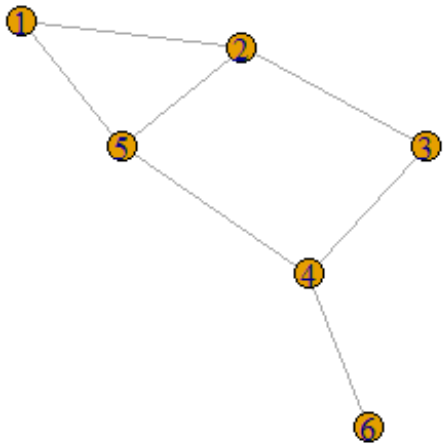

Figure 1.2: Grafo simples não orientado.

O grafo da Figura 1.2 tem conjunto de vértices $V = \{1, 2, 3, 4, 5, 6\}$ e conjunto de arestas $E = \{(1, 2), (1, 5), (2, 3), (2, 5), (3, 4), (4, 5), (4, 6)\}$.

Uma aresta que conecta um vértice para ele mesmo é chamada de laço e duas ou mais arestas conectando o mesmo par de vértices são chamadas de arestas múltiplas.

Um grafo G é dito de ser multigrafo quando este possui laços e arestas múltiplas. Graficamente um multigrafo pode ser representado pela Figura 1.3.

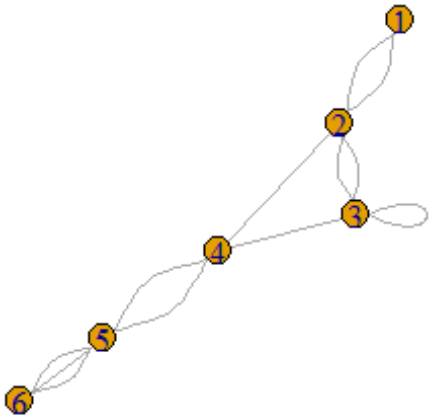

Figure 1.3: Multigrafo.

O multigrafo mostrado na Figura 1.3 tem conjunto de vértices $V = \{1, 2, 3, 4, 5, 6\}$ e conjunto de arestas $E = \{(1, 2), (1, 2), (2, 3), (2, 3), (3, 3), (3, 4), (2, 4), (4, 5), (4, 5), (5, 6), (5, 6), (5, 6)\}$.

Voltando ao exemplo da rede de transporte aéreo, um grafo simples (sem laços nem arestas múltiplas) poderia modelar a existência ou não de conexões entre um par de aeroportos, enquanto que um multigrafo incluiria a intensidade destas conexões (por exemplo, computando para cada par de aeroportos, a quantidade de voos diários operados).

Quando a construção do grafo envolve aleatoridade, este é chamado de grafo aleatório. Existem muitos modelos de grafos aleatórios, porém o mais básico e conhecido é o modelo Erdös Rényi, denotado por $G(n, p)$ e foi introduzido por (Erdős, 1959; Erdos et al., 1960), no caso o n é o número de vértices e cada par de vértices é conetado de forma independente por uma aresta somente, com probabilidade fixa p. O número de arestas no grafo $G(n, p)$ é uma variável aleatória com valor esperado dado por $\binom{n}{2}p$.

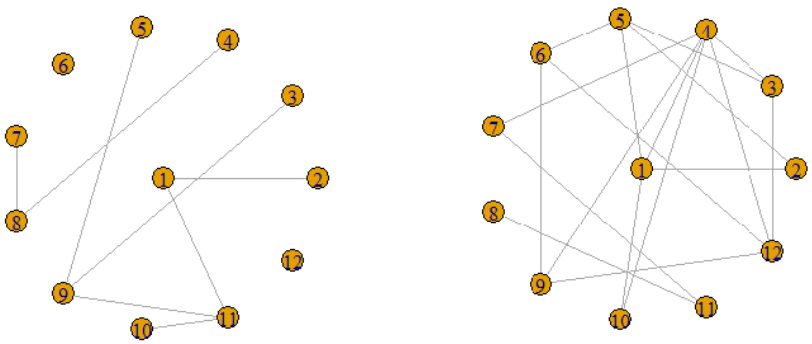

Figure 1.4: Erdös Rényi, $n = 12$.

Na Figura 1.4, podemos ver do lado esquerdo um grafo com $n = 12$ e $p = 0.10$ e no lado direito temos um grafo com $n = 12$ e $p = 0.30$.

Como dito anteriormente, o número médio de arestas no sistema cresce muito rápido com n e de fato, a imagem é de um grafo cheio de elos, muito denso. A razão para isso é que o número médio de conexões a cada vértice, np, vai para o infinito. Para evitar isso pode se tomar p dependendo de n, de forma que vai para zero quando n diverge, por exemplo $O(1/n)$. Desta forma temos um grafo esparso, no qual o número médio de conexões de cada vértice é limitado. A simplicidade do modelo Erdos Renyi permite que se obtenham muitos resultados matemáticos, mas é também uma limitação quando se trata de modelar fenômenos reais.

O modelo estocástico de blocos (SBM, do inglês *Stochastic Block Model*), originalmente introduzido por Holland et al. (1983) em ciências sociais. É uma generalização do modelo Erdös Renyi, caracterizado pela estrutura de blocos: o conjunto de vértices é particionado em subgrupos chamados blocos ou comunidades e condicionalmente na alocação destas comunidades, os vértices são conectados de forma independente, com uma probabilidade que depende apenas das suas comunidades. O SBM é tipicamente aplicado em grafos simples, que além de não ter laços nem arestas múltiplas, cada entrada da matriz de adjacência segue uma distribuição de Bernoulli.

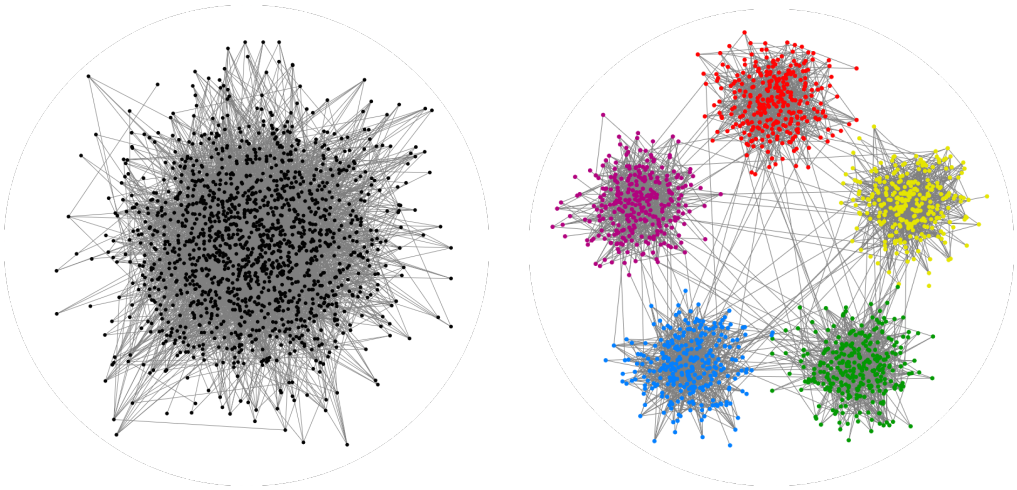

Figure 1.5: Estrututa de comunidades
Fonte: Abbe (2018).

Na figura 1.5, podemos observar no lado esquerdo, um grafo com $n = 1000$ e no lado direito temos o mesmo grafo, que foi particionado em 5 comunidades, a partir de um modelo estocástico de blocos.

Neste trabalho, vamos nos concentrar em duas extensões do SBM: O modelo estocástico de blocos de Poisson que é uma versão multigrafo do SBM, introduzida por

Karrer e Newman (2011) em que as entradas da matriz de adjacência seguem a distribuição de Poisson (ver Seção 2.1). E o modelo estocástico de blocos com correção de grau (DCSBM, do inglês *Degree Corrected Stochastic Block model*), também introduzido por Karrer e Newman (2011) que relaxa a limitação do SBM (homogeneidade das distribuições dos graus de pares de vértices dentro da mesma comunidade), permitindo distribuição arbitrária dos graus dentro dos blocos. O DCSM se comporta de forma similar ao modelo estocástico de blocos de Poisson só que agora é introduzido um novo parâmetro w_i que controla o grau esperado de cada vértice i (ver Seção 3.1.1).

No contexto do SBM o problema mais estudado na literatura é a detecção de comunidades, isto é, dado um grafo G, consiste na recuperação ou estimação do vetor aleatório $Z = (Z_1, Z_2, \ldots, Z_n) \in \{1, \ldots, k\}^n$ que é a classificação dos n vértices em k comunidades, em que Z_i representa a comunidade do vértice i. Para isso, foram propostos uma variedade de procedimentos com diferentes abordagens: Newman e Girvan (2004) propuseram um algoritmo baseado em uma medida de modularidade que mede a qualidade de uma determinada partição de uma rede, Bickel e Chen (2009) aplicaram sua abordagem à modularidade de Newman Girvan, (Celisse et al., 2012; Vu et al., 2013) usaram métodos variacionais, van der Pas e van der Vaart (2018) trabalharam com uma abordagem bayesiana, (Amini et al., 2013; Lei e Rinaldo, 2015) usaram métodos baseados em *espectral clustering*, (Rohe et al., 2011; Bickel et al., 2013) estudaram propriedades assintóticas de estimadores baseados nos métodos *espectral clustering* e variacional, respectivamente sob o SBM, Zhao et al. (2012) estabeleceram uma teoria geral para verificar a consistência da detecção de comunidades sob o DCSBM, Yan et al. (2014) propuseram uma abordagem baseada num teste da razão de verossimilhança para selecionar entre o SBM e o DCSBM, qual dos dois é melhor para estimar as comunidades, Chen et al. (2018) propuseram uma abordagem de maximização de modularidade convexa, baseada em um relaxamento da maximização de modularidade clássica, para detecção de comunidades sob o DCSBM.

Em alguns trabalhos (Bickel e Chen, 2009; van der Pas e van der Vaart, 2018; Celisse et al., 2012), etc., citados acima, o número de comunidades é assumido como conhecido, porém na prática este conhecimento a priori não é disponível. Em vista disso, a presente tese é dedicada ao problema de estimação do número de comunidades no modelo estocástico de blocos de Poisson e no DCSBM. Uma forma de abordar esse problema é o que se conhece como a seleção de modelos. Existem alguns critérios utilizados em trabalhos anteriores, tal como o *Bayesian Information Criterion* (BIC) ou o *Akaike Information Criterion* (AIC) que são baseados na verossimilhança dos dados observados da rede, o que faz esses critérios não tratáveis já que correm o risco de subestimar ou superestimar o verdadeiro número de comunidades, respectivamente. Para solucionar este problema, foi

proposto o critério *Integrated Classification Likelihood* (ICL) originalmente desenvolvido por Biernacki et al. (2000) para modelos de mistura gaussiana e depois adaptado por Daudin et al. (2008) para SBM. Este critério é baseado em uma aproximação assintótica da verosimilhança completa integrada dos dados. Entretanto, Mariadassou et al. (2010) obteve resultados mostrando que o ICL tende a subestimar o número de comunidades do SBM ao lidar com redes pequenas. Latouche et al. (2012) propuseram um novo critério chamado *Integrated Likelihood Variational Bayes* (ILvb), baseado em uma aproximação não assintótica da verossimilhança marginal. Ao contrário do ICL, Côme e Latouche (2015) apresentaram o critério *Exact ICL* (ICLex), que não é baseado em aproximações assintóticas, Yan (2016) propôs uma estrutura bayesiana para seleção de modelos no SBM e depois generalizaram esses resultados para o DCSBM.

Wang et al. (2017) também estudaram o problema de seleção de modelos sob o SBM e o DCSBM. Mostraram que a estatística do teste da razão de verossimilhança tem distribuição normal assintótica para o caso de subestimação e obtiveram a taxa de convergência para a estatística no caso de superestimação. Combinando esses resultados propuseram um critério de seleção de modelos baseado em verossimilhança penalizada, a função penalidade que eles derivaram foi da ordem $k^2 n \log n$, onde n é o número de vértices no grafo. Além disso mostraram que este critério é fracamente consistente (consistência em probabilidade). Hu et al. (2019) argumentaram que a função penalidade usada em Wang et al. (2017) tende a subestimar o número de comunidades e para esse fim sugeriram um novo critério chamado *Corrected Bayesian Information Criterion* (CBIC), a função penalidade usada foi da ordem n.

Cerqueira e Leonardi (2020) introduziram o estimador Krichevsky-Trofimov (KT) para determinar o número de comunidades do SBM. A abordagem proposta é um estimador penalizado baseado em uma distribuição de mistura do modelo, conhecido como distribuição de mistura KT. A principal contribuição é o resultado de consitência forte do estimador (consistência quase certa), tanto no regime denso (a probabilidade de conexão entre comunidades é constante) quanto no regime semi-esparso (a probabilidade de conexão entre comunidades pode decrescer para zero com n, numa determinada taxa) . A função penalidade proposta é dominada por um termo da ordem $k^3 \log n$, que é de menor ordem comparada com aquela usada em Wang et al. (2017). Usando um método de *espectral clustering*, Ma et al. (2021) propuseram um estimador do número de comunidades no DCSBM, baseado em uma estatística do teste da razão de verossimilhança.

O objetivo principal deste trabalho é provar consistência forte para o estimador do número de comunidades do modelo estocástico de blocos de Poisson, com correção de grau

e no regime semi-esparso, estendendo dessa forma o resultado de Cerqueira e Leonardi (2020).

A tese está organizada da seguinte forma: O Capítulo 2 é dedicado ao estimador consistente no modelo estocástico de blocos de Poisson. Começamos com a Seção 2.1 descrevendo formalmente o modelo, assim como a verossimilhança e estimadores de máxima verossimilhança do modelo. Na Seção 2.2, definimos o estimador do número de comunidades e enunciamos o resultado de consistência forte para este estimador (Teorema 2.2.1). A Seção 2.3 é dedicada à demonstração do Teorema 2.2.1, mas antes, na Subseção 2.3.1 enunciamos os resultados chave para a prova do teorema da consistência (Proposição 2.3.1 e Lema 2.3.2), em seguida nas Subseções 2.3.2 e 2.3.3 temos a prova do Teorema 2.2.1, por meio da não superestimação e não subestimação, respectivamente. O Capítulo 3 é uma extensão ao caso com correção de grau e contém as mudanças a serem feitas no Capítulo 2, para enunciar o teorema da consistência (Teorema 3.2.1).

Chapter

2

Estimação consistente no modelo estocástico de blocos de Poisson

Na Seção 2.1, definimos a versão multigrafo do Modelo Estocástico de Blocos, conhecido como modelo estocástico de blocos de Poisson, que foi introduzido por Karrer e Newman (2011). Na Seção 2.2, definimos o estimador do número de comunidades do modelo e enunciamos o teorema de consistência do estimador e por último na Seção 2.3 temos a demonstração do teorema da consistência.

2.1 Modelo estocástico de blocos de Poisson

2.1.1 Definição do modelo

O modelo estocástico de blocos de Poisson (*Poisson Stochastic Block Model*) é um modelo de grafo aleatório não orientado no qual os vértices estão separados em classes (comunidades, blocos) e o número de arestas entre pares de vértices depende da classe dos dois vértices envolvidos. Vamos formalizar este modelo abaixo.

Consideramos um espaço mensurável $(\Omega, \mathcal{F})$ no qual todas as v.a.'s introduzidas neste trabalho poderão ser definidas.

Seja $[n] := \{1, \ldots, n\}$, um conjunto de $n \geq 1$ vértices. Cada par de vértices $i, j \in [n]$ tem um número aleatório $N_{i,j}$ (possivelmente 0) de arestas ligando eles. A matriz de

adjacência $X = (X_{ij})_{i,j\in[n]}$ é simétrica (grafo não orientado), em que $X_{ij} = N_{ij}$ para $i < j$, enquanto $X_{ii} = 2N_{ii}$, ou seja, duas vezes o número de laços no vértice i.

Os n vértices são separados, de forma aleatória, entre $k_0 \geq 1$ possíveis grupos. O vetor aleatório $Z = (Z_i)_{i\in[n]}$, com $Z_i \in [k_0] := \{1, \ldots, k_0\}$, representa a associação de cada vértice a um grupo. Assumimos que este vetor é i.i.d. com distribuição marginal π sobre $[k_0]$.

No modelo Poisson, assumimos que existe uma matriz simétrica com entradas positivas $\lambda = (\lambda_{a,b})_{a,b\in[k_0]}$ tal que, condicionados no vetor Z, os $N_{ij}, i \leq j$ são independentes com distribuição

$$\begin{aligned} N_{ij}|Z_i = a, Z_j = b &\sim \text{Poisson}(\lambda_{a,b}), \ i < j \\ N_{ii}|Z_i = a &\sim \text{Poisson}(\lambda_{a,a}/2). \end{aligned}$$

Além disso, como $X_{ij} = N_{ij}$ para $i < j$, segue que

$$X_{ij}|Z_i = a, Z_j = b \sim \text{Poisson}(\lambda_{a,b}),$$

isto é, condicionalmente a Z, as entradas $X_{ij}, i \leq j$ da matriz de adjacência também são independentes e têm $\mathbb{E}(X_{ij}|Z_i = a, Z_j = b) = \lambda_{a,b}$, para $i < j$. Por outro lado, $X_{ii} = 2N_{ii}$ não tem distribuição Poisson, mas sabendo que $Z_i = a$, X_{ii} tem média $\lambda_{a,a}$.

No regime *denso*, a matriz de parâmetros $\lambda = (\lambda_{a,b})_{a,b\in[k_0]}$ é constante e não depende de n, neste caso cada vértice tem um número médio de conexões que cresce linearmente com n, o que faz a rede ficar superconectada (grafo com muitas arestas). Por esta razão, outro cenário a considerar é o regime *semi-esparso*, onde a matriz λ depende de n e cujas entradas podem decrescer para 0 quando $n \to \infty$. No presente trabalho, assumimos que existe uma sequência ρ_n tal que $\lambda = \rho_n\tilde{\lambda}$ com $\rho_n \to 0$ quando $n \to \infty$, onde $\tilde{\lambda}$ é uma matriz simétrica com entradas positivas, como no modelo *denso*. Daremos condições sobre ρ_n nos resultados principais da tese.

Suposição 2.1.1. *Suponhamos que o vetor de distribuição de probabilidades π tem todas as entradas positivas e a matriz simétrica $\tilde{\lambda}$ não tem duas colunas idênticas.*

Esta suposição quer dizer que o modelo de blocos é identificável, no sentido que este não pode ser reduzido para um modelo menor (com um menor número de comunidades).

2.1.2 A verossimilhança do modelo

Como já foi visto na seção anterior, o modelo é completamente definido pela matriz simétrica λ, de números reais positivos e pela distribuição de probabilidade π. Denotamos

por $(\Omega, \mathcal{F}, \mathbb{P}_{(\lambda,\pi)})$ o espaço de probabilidade associado a este par e temos a verossimilhança conjunta para todo $z \in [k_0]^n$ e $x \in \mathbb{N}^{n\times n}$

$$\mathbb{P}_{(\lambda,\pi)}(X = x, Z = z) = \mathbb{P}_{(\lambda,\pi)}(Z = z)\mathbb{P}_{(\lambda,\pi)}(X = x|Z = z),$$

onde

$$\mathbb{P}_{(\lambda,\pi)}(Z = z) = \prod_{a=1}^{k} \pi_a^{\sum_{i=1}^{n} \mathbb{1}\{z_i=a\}},$$

não depende de λ e

$$\mathbb{P}_{(\lambda,\pi)}(X = x|Z = z) = \left[\prod_{1\le i<j\le n} \frac{(\lambda_{z_i,z_j})^{x_{ij}} \exp\{-\lambda_{z_i,z_j}\}}{x_{ij}!}\right] \left[\prod_{i=1}^{n} \frac{(\frac{\lambda_{z_i,z_i}}{2})^{\frac{x_{ii}}{2}} \exp\{-\frac{\lambda_{z_i,z_i}}{2}\}}{\frac{x_{ii}}{2}!}\right], \tag{2.1}$$

não depende de π. Para escrever a verossimilhança conjunta como função de k, vamos definir os seguintes contadores, que permitirão fórmulas menos carregadas:

$$n_a(z) = \sum_{1\le i\le n} \mathbb{1}\{z_i = a\}, \quad 1 \le a \le k \tag{2.2}$$

e

$$o_{a,b}(x, z) = \sum_{1\le i,j\le n} x_{ij}\mathbb{1}\{z_i = a, z_j = b\}, \quad 1 \le a, b \le k \tag{2.3}$$

que contam o número de vértices na comunidade a e o número de arestas conectando vértices na comunidade a com vértices na comunidade b, respectivamente. Neste último caso, lembrando que x_{ii} é o dobro do número de arestas no vértice i, temos que $o_{a,a}(x, z)$ corresponde ao dobro do número de arestas entre vértices da comunidade a (já que a soma é feita para todos os pares i, j). Além disso, para contarmos o número de pares possíveis nas comunidades a, b podemos fazer o produto $n_a(z)n_b(z)$, mas no caso $a = b$ este valor também é o dobro do valor desejado. Por este motivo vamos definir os seguintes contadores que levam em consideração esta característica:

$$n_{a,b}(z) = \begin{cases} n_a(z)n_b(z) & a \neq b, \\ \frac{1}{2}n_a(z)n_b(z) & a = b \end{cases}$$

e

$$O_{a,b}(x, z) = \begin{cases} o_{a,b}(x, z) & a \neq b, \\ \frac{1}{2}o_{a,b}(x, z) & a = b. \end{cases}$$

Com estas definições, a verossimilhança conjunta pode ser reescrita como

$$\mathbb{P}_{(\lambda,\pi)}(X = x, Z = z) \;=\; \frac{1}{c(x)} \left[\prod_{1\le a\le b\le k} \lambda_{a,b}^{O_{ab}(x,z)} \exp\{-n_{a,b}(x)\lambda_{a,b}\} \right] \left[\prod_{a=1}^{k} {\pi_a}^{n_a(z)} \right], \quad (2.4)$$

com

$$c(x) \;:=\; \left[\prod_{i<j} x_{ij}! \right] \left[\prod_{i} 2^{x_{ii}/2}(x_{ii}/2)! \right].$$

2.1.3 Estimadores de máxima verossimilhança

No Lema 2.1.1 apresentamos os estimadores dos parâmetros do modelo, $\theta_0 = (\lambda^0, \pi^0)$ por meio do método de máxima verossimilhança.

Lema 2.1.1. *Seja uma realização (x, z) de um modelo estocástico de blocos de Poisson, com função de verossimilhança conjunta $p(x, z|\theta)$, dada em (2.4). Os estimadores de máxima verossimilhança dos parâmetros $\theta_0 = (\lambda^0, \pi^0)$ são*

$$\widehat{\lambda}_{a,b}(x, z) \;=\; \frac{O_{a,b}(x, z)}{n_{a,b}(z)}, \quad 1 \le a, b \le k$$

e

$$\widehat{\pi}_a(z) \;=\; \frac{n_a(z)}{n}, \quad 1 \le a \le k,$$

respectivamente.

Proof. Seja o logaritmo da função verossimilhança $p(x, z|\theta)$,

$$\log p(x, z|\theta) = \sum_{1\le a\le b\le k} O_{a,b}(x, z) \log \lambda_{a,b} - \sum_{1\le a\le b\le k} n_{a,b}(z)\lambda_{a,b} + \sum_{a=1}^{k} n_a(z) \log \pi_a. \quad (2.5)$$

Agora, usando o método de multiplicadores de Lagrange, definamos a função $\mathcal{L}(\lambda, \pi, \gamma|\theta)$ chamada a função de Lagrange, de tal forma que

$$\mathcal{L}(\lambda, \pi, \gamma|\theta) = \sum_{1\le a\le b\le k} O_{a,b}(x, z) \log \lambda_{a,b} - \sum_{1\le a\le b\le k} n_{a,b}(z)\lambda_{a,b} + \sum_{a=1}^{k} n_a(z) \log \pi_a + \gamma \left(1 - \sum_{a=1}^{k} \pi_a \right),$$

onde γ é chamado o multiplicador de Lagrange. Daqui, derivando primeiro $\mathcal{L}(\lambda, \pi, \gamma|\theta)$ com respeito a $\lambda_{a,b}$, $1 \le a, b \le k$ e igualando a zero, temos

$$\frac{\partial \mathcal{L}(\lambda, \pi, \gamma|\theta)}{\partial \lambda_{a,b}} = \frac{O_{a,b}(x, z)}{\lambda_{a,b}} - n_{a,b} = 0 \quad \Rightarrow \quad \lambda_{a,b}(x, z) = \frac{O_{a,b}(x, z)}{n_{a,b}(z)}. \quad (2.6)$$

Usando um critério da segunda derivada, de (2.6) resulta que $\widehat{\lambda}_{a,b}(x,z) = \frac{O_{a,b}(x,z)}{n_{a,b}(z)}$ é um ponto de máximo e portanto o estimador de máxima verossimilhança para $\lambda_{a,b}$, $1 \leq a, b \leq k$.

Por outro lado, derivando $\mathcal{L}(\lambda, \pi, \gamma|\theta)$ com respeito a π_a, $1 \leq a \leq k$ e igualando a zero temos

$$\frac{\partial \mathcal{L}(\lambda, \pi, \gamma|\theta)}{\partial \pi_a} = \frac{n_a(z)}{\pi_a} - \gamma = 0 \quad \Rightarrow \quad \pi_a = \frac{n_a(z)}{\gamma}. \tag{2.7}$$

Agora, usando a restrição $\sum_{a=1}^{k} \pi_a = 1$ em (2.7) resulta que $\gamma = n$. Portanto, temos que

$$\pi_a = \frac{n_a(z)}{n}.$$

Para verificar que é um ponto de máximo, usamos novamente o critério da segunda derivada com o qual podemos concluir que $\widehat{\pi}_a(z) = \frac{n_a(z)}{n}$ é o estimador de máxima verossimilhança para π_a, $1 \leq a \leq k$. □

O valor máximo de (2.4) em relação a π e λ é dado por

$$\sup_{\theta \in \Theta_k} p(x, z|\theta) \;=\; \frac{1}{c(x)} \left[\prod_{1 \leq a \leq b \leq k} \widehat{\lambda}_{a,b}^{O_{ab}(x,z)} \exp\{-n_{a,b}(z)\widehat{\lambda}_{a,b}\} \right] \left[\prod_{a=1}^{k} \widehat{\pi}_a^{n_a(z)} \right]. \tag{2.8}$$

Substituindo os valores de $\widehat{\pi}_a$ e $\widehat{\lambda}_{a,b}$ pelas expressões dadas no Lema 2.1.1 e tomando logaritmo temos que

$$\begin{aligned} \log \sup_{\theta \in \Theta_k} p(x, z|\theta) \;&= \bar{c}(x) + \sum_{1 \leq a \leq b \leq k} O_{a,b}(x,z) \log \frac{O_{a,b}(x,z)}{n_{a,b}(z)} \\ &\quad - M(x) + \sum_{1 \leq a \leq k} n_a(z) \log \frac{n_a(z)}{n}, \end{aligned} \tag{2.9}$$

em que

$$\bar{c}(x) \;=\; -\log c(x)$$

é um valor que não depende de z e portanto não depende de k. E

$$M(x) \;=\; \sum_{1 \leq a \leq b \leq k} O_{a,b}(x,z) \tag{2.10}$$

representa o número de arestas no grafo, também é um valor que não depende de z e portanto não depende de k.

O Lema 2.1.2 que apresentamos abaixo mostra a consistência dos estimadores de máxima verossimilhança para os parâmetros do modelo.

Lema 2.1.2. *Dada uma amostra (x, z) de um modelo estocástico de blocos de Poisson, com parâmetros $\theta_0 = (\lambda^0, \pi^0)$ e estimadores de máxima verossimilhança $(\widehat{\lambda}, \widehat{\pi})$ dados como no Lema 2.1.1, temos*

$$\widehat{\pi}_a \to \pi_a^0, \qquad a \in \{1, \dots, k_0\},$$
$$\widehat{\lambda}_{a,b} \to \lambda_{a,b}^0, \qquad a, b \in \{1, \dots, k_0\},$$

quase certamente quando $n \to \infty$.

Proof. Uma vez que as variáveis aleatórias $\mathbb{1}\{z_i = a\}$, $i \geq 1$ são independentes e identicamente distribuídas e além disso $\mathbb{E}(\mathbb{1}\{z_1 = a\}) = p(z_1 = a|\theta_0) = \pi_a^0 < \infty$, então pela lei forte dos grandes números, temos que

$$\widehat{\pi}_a = \frac{\sum_{i=1}^n \mathbb{1}\{z_i = a\}}{n} \to \pi_a^0,$$

quase certamente, quando $n \to \infty$. De forma similar, as variáveis aleatórias $x_{ij}\mathbb{1}\{z_i = a, z_j = b\}$, $i, j \geq 1$ são independentes e identicamente distribuídas e $\mathbb{E}(x_{12}\mathbb{1}\{z_1 = a, z_2 = b\}) = \mathbb{E}(x_{12}|z_1 = a, z_2 = b)p(z_1 = a|\theta_0)(z_2 = b|\theta_0) = \lambda_{a,b}^0\pi_a^0\pi_b^0$, então pela lei forte dos grandes números podemos concluir que

$$\widehat{\lambda}_{a,b} = \frac{\sum_{1 \leq i,j \leq n} x_{ij}\mathbb{1}\{z_i = a, z_j = b\}}{n^2} \frac{n^2}{n_a(z)n_b(z)} \to \lambda_{a,b}^0,$$

quase certamente, quando $n \to \infty$. □

2.2 Definição do estimador e teorema da consistência

Fixe $k \geq 1$. Denotamos por $\theta := (\lambda, \pi) \in \Theta_k^{(1)} \times \Theta_k^{(2)} =: \Theta_k$ o par de parâmetros do modelo, em que

$$\Theta_k^{(1)} := \{\lambda \in (\mathbb{R}^+)^{k \times k} : \lambda_{ab} = \lambda_{ba}, a, b \in [k]\}$$
$$\Theta_k^{(2)} := \{\pi \in (0, 1]^k : \sum_{a=1}^k \pi_a = 1\}.$$

Definição 2.2.1. *Dada uma realização do multigrafo $X = (X_{ij})_{i,j \in [n]}$, definimos a ordem do modelo como o menor inteiro $k_0 \geq 1$, tal que existe um par de parâmetros $\theta_0 = (\lambda^0, \pi^0) \in \Theta_{k_0}$ e $(X_{ij})_{i,j \in [n]}$ é distribuida de acordo a $\mathbb{P}_{(\lambda^0, \pi^0)}$.*

Nesse sentido, pela Suposição 2.1.1 podemos dizer que se o modelo estocástico de blocos de Poisson tem ordem k_0, então este não pode ser reduzido para um modelo com

um número de comunidades menor do que k_0. A seguir, vamos definir o estimador de Krichevsky-Trofimov (KT), que será utilizado neste trabalho.

A abordagem de KT é considerar uma distribuição *a priori* para $\theta = (\lambda, \pi)$ e integrar a verossimilhança marginal $\mathbb{P}_{(\lambda,\pi)}(X = x)$ sob esta *a priori*. Apesar de não ser uma abordagem Bayesiana, porque não estudamos a *a posteriori*, vamos usar a notação desta área para simplificar. Neste sentido, escrevemos agora λ, π, θ na condicional e omitimos a referência às variáveis aleatórias, mencionando somente os valores assumidos por elas, e finalmente, usamos p no lugar de $\mathbb{P}$. De forma que, agora, a verossimilhança conjunta $\mathbb{P}_{(\lambda,\pi)}(X = x, Z = z)$ passará a ser escrita $p(x, z|\theta)$, uma notação mais compacta.

Para todo $k \geq 1$, a partir de uma distribuição *a priori* ν_k sobre Θ_k, definimos a distribuição de mistura

$$p_k(x) := \sum_{z \in [k]^n} \int_{\Theta_k} p(x, z|\theta)\nu_k(\theta)d\theta, \tag{2.11}$$

onde $p(x, z|\theta)$ é a função de verossimilhança conjunta de (X, Z) e foi dada em (2.4). Por outro lado, definimos uma medida produto como *a priori*, $\nu_k = \nu_k^{(1)} \otimes \nu_k^{(2)}$ sobre $\Theta_k^{(1)} \times \Theta_k^{(2)}$, em que sob $\nu_k^{(1)}$ e $\nu_k^{(2)}$,

$$\begin{aligned} \lambda_{a,b} &\sim \text{Gama}(1/2, 1/2)\,,\ a, b \in [k], a \leq b \\ \pi &\sim \text{Dirichlet}\underbrace{(1/2, \ldots, 1/2)}_{k} \end{aligned}$$

respectivamente. Em outras palavras, para $\theta = (\lambda, \pi) \in \Theta_k^{(1)} \times \Theta_k^{(2)}$

$$\nu_k(\theta) = \nu_k^{(1)}(\lambda)\nu_k^{(2)}(\pi) := \left[\prod_{1 \leq a \leq b \leq k} \frac{1}{\Gamma(1/2)}(1/2)^{1/2}\lambda_{ab}^{-1/2}e^{-\lambda_{ab}/2}\right]\left[\frac{\Gamma(k/2)}{\Gamma(1/2)^k}\prod_{a=1}^{k}{\pi_a}^{-1/2}\right]. \tag{2.12}$$

A partir da distribuição de mistura em (2.11) baseada na *a priori* em (2.12), podemos definir agora o nosso estimador do número de comunidades do modelo: dada uma amostra x:

$$\widehat{k}_n(x) := \arg\max_k \{\log p_k(x) - k^3 \log n\}. \tag{2.13}$$

O motivo desta escolha será melhor entendido na demonstração do resultado de consistência forte do estimador do número de comunidades, que enunciamos a continuação.

Teorema 2.2.1 (Teorema da consistência)**.** *Seja um modelo estocástico de blocos de Poisson de ordem k_0 e $\rho_n \geq C\frac{\log n}{n}$, onde C é uma constante suficientemente grande.*

Então, o estimador $\widehat{k}_n$ *definido em* (2.13) *satisfaz*

$$\widehat{k}_n = k_0, \tag{2.14}$$

quase certamente, quando $n \to \infty$.

Wang et al. (2017) mostraram consistência fraca para um estimador do número de comunidades no SBM (distribuição de Bernoulli) com uma função de penalidade da ordem $k^2 n \log n$. Cerqueira e Leonardi (2020) mostraram consistência forte para o estimador do número de comunidades também no SBM, nesse trabalho usaram uma função de penalidade dominada por um termo da ordem $k^3 \log n$, que comparada com aquela usada em Wang et al. (2017), é de menor ordem. Nesta tese, provamos consistência forte para o estimador do número de comunidades no modelo estocástico de blocos de Poisson (distribuição de Poisson), introduzido por Karrer e Newman (2011). Até onde sabemos este é o primeiro resultado na literatura provando convergência forte para esse modelo, observamos que a função de penalidade dada em (2.13) é de menor ordem comparada com os trabalhos mencionados anteriormente.

2.3 Demonstração do teorema da consistência

Na demonstração do Teorema 2.2.1, mostraremos que o estimador da ordem não superestima (Proposição 2.3.3) e nem subestima (Proposição 2.3.4) a verdadeira ordem k_0 do modelo estocástico de blocos de Poisson, quando $n \to \infty$. Mas antes, vamos enunciar dois resultados chave.

2.3.1 Resultados chave

A Proposição 2.3.1 fornece um limitante superior não assintótico para o logaritmo da razão entre a função de máxima verosimilhança de X e a distribuição de mistura.

Defina o conjunto

$$\Omega_n := \{x = \{x_{ij}\} \in \mathbb{N}^{n\times n} : \mathrm{x}_{ij} \leq \log n, \ \ \forall i, j = 1, \ldots, n\}. \tag{2.15}$$

Proposição 2.3.1. *Para todo* $k \geq 1$, $n \geq \max(4, k)$ *e para todo* $x \in \Omega_n$, *temos*

$$\log\left(\frac{\sup_{\theta\in\Theta_k} p(x|\theta)}{p_k(x)}\right) \leq d(k)\log n + c(k, n),$$

onde,

$$d(k) := \frac{k(2k+3)-1}{2}$$

$$c(k,n) := \frac{k(k-1)}{4n} + \frac{k(k+1)}{8n^2} + \frac{1}{12n} + \log\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{k}{2})}.$$

Proof. Analisaremos separadamente as razões $p(x|z,\theta)/p_k(x|z)$ e $p(z|\theta)/p_k(z)$ mostrando que são limitadas superiormente, uniformemente em x, z, por funções $c_1(n,k)$ e $c_2(n,k)$ respectivamente. Com isso, observamos que

$$\begin{aligned}\log\frac{p(x|\theta)}{p_k(x)} &= \log\frac{\sum_z p(x|\theta,z)p(z|\theta)}{\sum_z p_k(x|z)p_k(z)} \\ &\leq \log\frac{\sum_z p_k(x|z)c_1(n,k)p_k(z)c_2(n,k)}{\sum_z p_k(x|z)p_k(z)} \\ &= \log c_1(n,k) + \log c_2(n,k). \end{aligned} \tag{2.16}$$

Começamos procurando um limitante $c_1(n,k)$ para $p(x|z,\theta)/p_k(x|z)$. Da definição da distribuição de mistura, temos que

$$\begin{aligned} p_k(x) &= \sum_{z\in[k]^n}\int_{\Theta_k^{(1)}} p(x|z,\lambda)\nu_k^{(1)}(\lambda)d\lambda \int_{\Theta_k^{(2)}} p(z|\pi)\nu_k^{(2)}(\pi)d\pi \\ &= \sum_{z\in[k]^n} p_k(x|z)p_k(z). \end{aligned} \tag{2.17}$$

Nas contas que seguem, escrevemos $O_{a,b}, n_a, n_b$ diretamente, omitindo a referência a x e z, para aliviar a notação e as equações que já estão carregadas.

Calculamos:

$$\begin{aligned} p_k(x|z) &= \frac{1}{c(x)}\int_{\Theta_k^{(1)}} \prod_{1\leq a\leq b\leq k} \lambda_{a,b}^{O_{a,b}} e^{-n_{a,b}\lambda_{a,b}}\nu_k^{(1)}(\lambda)d\lambda \\ &= \frac{1}{c(x)2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}}\int_{\Theta_k^{(1)}} \prod_{1\leq a\leq b\leq k} \lambda_{a,b}^{\left(O_{a,b}-\frac{1}{2}\right)} e^{-\left(n_{a,b}+\frac{1}{2}\right)\lambda_{a,b}}d\lambda \\ &= \frac{1}{c(x)2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}} \prod_{1\leq a\leq b\leq k} \frac{\Gamma\left(O_{a,b}+\frac{1}{2}\right)}{\left(n_{a,b}+\frac{1}{2}\right)^{\left(O_{a,b}+\frac{1}{2}\right)}}, \end{aligned} \tag{2.18}$$

onde, na última igualdade foi usada a representação integral $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$ da função gamma para $r > 0$.

Por outro lado, considerando que o estimador de máxima verossimilhança para $\lambda_{a,b}$ é $\frac{O_{a,b}}{n_{a,b}}$, temos que

$$p(x|z,\lambda) \leq \sup_{\theta\in\Theta^k} p(x|z,\lambda) = \frac{1}{c(x)} \prod_{1\leq a\leq b\leq k} \left(\frac{O_{a,b}}{n_{a,b}}\right)^{O_{a,b}} e^{-O_{a,b}}. \tag{2.19}$$

Então, combinando (2.18) e (2.19), temos

$$\frac{p(x|z,\lambda)}{p_k(x|z)} \leq \frac{\prod_{1\leq a\leq b\leq k} \left(\frac{O_{a,b}}{n_{a,b}}\right)^{O_{a,b}} e^{-O_{a,b}}}{\frac{1}{2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}} \prod_{1\leq a\leq b\leq k} \frac{\Gamma\left(O_{a,b}+\frac{1}{2}\right)}{\left(n_{a,b}+\frac{1}{2}\right)^{\left(O_{a,b}+\frac{1}{2}\right)}}}.$$

Podemos então escrever

$$\frac{p(x|z,\lambda)}{p_k(x|z)} \leq D(k) \prod_{1\leq a\leq b\leq k} A_{a,b}$$

em que

$$D(k) = 2^{\frac{k(k+1)}{4}}\Gamma\left(\frac{1}{2}\right)^{\frac{k(k+1)}{2}}$$

e

$$A_{a,b} = \frac{\left(\frac{O_{a,b}}{n_{a,b}}\right)^{O_{a,b}}}{\Gamma\left(O_{a,b}+\frac{1}{2}\right)} \left(n_{a,b}+\frac{1}{2}\right)^{\left(O_{a,b}+\frac{1}{2}\right)} e^{-O_{a,b}}. \tag{2.20}$$

Agora, lembramos (2.10)

$$M := \sum_{1\leq a\leq b\leq k} O_{a,b}$$

para usar Lema A.1.1, que nós garante que

$$\prod_{1\leq a\leq b\leq k} \frac{\left(\frac{O_{a,b}}{M}\right)^{O_{a,b}}}{\Gamma\left(O_{a,b}+\frac{1}{2}\right)} \leq \frac{1}{\Gamma\left(M+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{\frac{k(k+1)}{2}-1}}. \tag{2.21}$$

Com isso,

$$\begin{aligned}\prod_{1\le a\le b\le k} A_{a,b} &\le \frac{1}{\Gamma\left(M+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{\frac{k(k+1)}{2}-1}} \prod_{1\le a\le b\le k}\left(\frac{M}{n_{a,b}}\right)^{O_{a,b}}\left(n_{a,b}+\frac{1}{2}\right)^{O_{a,b}+\frac{1}{2}} e^{-O_{a,b}} \\ &= \frac{1}{\Gamma\left(M+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{\frac{k(k+1)}{2}-1}} \prod_{1\le a\le b\le k} M^{O_{a,b}}\left(1+\frac{1}{2n_{a,b}}\right)^{O_{a,b}}\left(n_{a,b}+\frac{1}{2}\right)^{\frac{1}{2}} e^{-O_{a,b}} \\ &= \frac{\left(\frac{M}{e}\right)^M}{\Gamma\left(M+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{\frac{k(k+1)}{2}-1}} \prod_{1\le a\le b\le k}\left(1+\frac{1}{2n_{a,b}}\right)^{O_{a,b}}\left(n_{a,b}+\frac{1}{2}\right)^{\frac{1}{2}}.\end{aligned}$$

Podemos agora simplificar

$$\begin{aligned}\frac{p(x|z,\lambda)}{p_k(x|z)} &\le D(k)\prod_{1\le a\le b\le k} A_{a,b} \\ &\le \frac{2^{\frac{k(k+1)}{4}}\Gamma\left(\frac{1}{2}\right)\left(\frac{M}{e}\right)^M}{\Gamma\left(M+\frac{1}{2}\right)} \prod_{1\le a\le b\le k}\left(1+\frac{1}{2n_{a,b}}\right)^{O_{a,b}}\left(n_{a,b}+\frac{1}{2}\right)^{\frac{1}{2}}.\end{aligned}$$

Por outro lado, usando o fato que para $x>0$, $\Gamma(x)\ge x^{x-1/2}e^{-x}\sqrt{2\pi}$ temos que

$$\frac{1}{\Gamma\left(M+\frac{1}{2}\right)} \le \left(M+\frac{1}{2}\right)^{-M} e^{(M+\frac{1}{2})}\frac{1}{\sqrt{2\pi}}, \tag{2.22}$$

e portanto

$$\frac{p(x|z,\lambda)}{p_k(x|z)} \le \frac{1}{\sqrt{2\pi}}2^{\frac{k(k+1)}{4}}\left(\frac{M}{M+\frac{1}{2}}\right)^M \Gamma\left(\frac{1}{2}\right)e^{\frac{1}{2}} \prod_{1\le a\le b\le k}\left(1+\frac{1}{2n_{a,b}}\right)^{O_{a,b}}\left(n_{a,b}+\frac{1}{2}\right)^{\frac{1}{2}}.$$

Agora, usaremos

$$e^{\frac{1}{2}}\left(\frac{M}{M+\frac{1}{2}}\right)^M \le 1\ , \quad \left(1+\frac{1}{2n_{a,b}}\right)^{O_{a,b}} \le e^{\frac{O_{a,b}}{2n_{a,b}}}\ , \quad n_{a,b}\le n^2,$$

de forma que

$$\frac{p(x|z,\lambda)}{p_k(x|z)} \le \frac{1}{\sqrt{2\pi}}\left(2n^2+1\right)^{\frac{k(k+1)}{4}}\Gamma\left(\frac{1}{2}\right)e^{\sum_{a\le b}\frac{O_{a,b}}{2n_{a,b}}}.$$

Para simplificar ainda mais vamos majorar

$$\left(2n^2+1\right)^{\frac{k(k+1)}{4}} = \left(2n^2\left(1+\frac{1}{2n^2}\right)\right)^{\frac{k(k+1)}{4}} \le 2^{\frac{k(k+1)}{4}} n^{\frac{k(k+1)}{2}} e^{\frac{k(k+1)}{8n^2}}.$$

Por outro lado, para $x \in \Omega_n$ temos que $O_{a,b} \leq n_{a,b} \log n$ e portanto $\sum_{a \leq b} \frac{O_{a,b}}{2n_{a,b}} \leq \frac{k(k+1)}{4} \log n$. Juntando tudo isso, temos

$$\begin{aligned} \log c_1(n,k) &= \frac{k(k+1)}{2} \log n + \frac{k(k+1)}{4} \log n + \frac{k(k+1)}{8n^2} \\ &\leq k(k+1) \log n + \frac{k(k+1)}{8n^2}. \end{aligned} \tag{2.23}$$

Agora procuramos um limitante $c_2(n,k)$ para $p(z|\theta)/p_k(z)$. Esta parte é completamente similar ao caso de Cerqueira e Leonardi (2020) (ver a desigualdade (41) deste artigo). Mesmo assim, vamos incluir as contas aqui para tornar o texto completo. Iniciamos calculando $p_k(z)$

$$\begin{aligned} p_k(z) = \int_{\Theta_1^k} \left(\prod_{a=1}^k {\pi_a}^{n_a(z)} \right) \nu_k^1(\pi) d\pi &= \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k} \int_{\Theta_1^k} \prod_{a=1}^k {\pi_a}^{n_a(z)-\frac{1}{2}} d\pi \\ &= \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k} \frac{\prod_{a=1}^k \Gamma\left(n_a(z)+\frac{1}{2}\right)}{\Gamma\left(n+\frac{k}{2}\right)} \end{aligned} \tag{2.24}$$

na última igualdade de (2.24) integramos com respeito da distribuição de π usando para isso o fato de que a integral de uma função de densidade de probabilidade (distribuição Dirichlet) é igual a 1. Por outro lado, usando o fato que o estimador de máxima verosssimilhança para π_a é $\dfrac{n_a(z)}{n}$, segue que

$$p(z|\pi) \leq \sup_{\theta \in \Theta^k} p(z|\pi) = \prod_{a=1}^k \left(\frac{n_a(z)}{n} \right)^{n_a(z)}. \tag{2.25}$$

Assim, de (2.24) e (2.25) obtemos

$$\frac{p(z|\pi)}{p_k(z)} \leq \frac{\Gamma\left(\frac{1}{2}\right)^k \Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \prod_{a=1}^k \frac{\left(\frac{n_a(z)}{n}\right)^{n_a(z)}}{\Gamma\left(n_a(z)+\frac{1}{2}\right)}. \tag{2.26}$$

Por outro lado,

$$\prod_{a=1}^k \frac{\left(\frac{n_a(z)}{n}\right)^{n_a(z)}}{\Gamma\left(n_a(z)+\frac{1}{2}\right)} \leq \frac{1}{\Gamma\left(n+\frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right)^{k-1}}, \tag{2.27}$$

onde, usamos a igualdade $n = \sum_{a=1}^{k} n_a(z)$ e o Lema A.1.1. Substituindo (2.27) em (2.26), temos que

$$\frac{p(z|\pi)}{p_k(z)} \leq \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}. \tag{2.28}$$

Pegamos portanto

$$\log c_2(n,k) = \log \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(n+\frac{1}{2}\right)} \tag{2.29}$$

que majoramos usando o Lema A.1.2

$$\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(n+\frac{k}{2})}{\Gamma(\frac{k}{2})\Gamma(n+\frac{1}{2})}\right) \leq \frac{k-1}{2}\log n + \frac{k(k-1)}{4n} + \frac{1}{12n} - \log\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})}. \tag{2.30}$$

Voltamos agora a (2.16), usando (2.23) e (2.29) para concluir a prova da proposição: uniformemente em x, z temos

$$\begin{aligned}\log\frac{p(x|\theta)}{p_k(x)} &\leq \log c_1(n,k) + \log c_2(n,k)\\ &\leq \frac{k(2k+3)-1}{2}\log n + c(k,n)\end{aligned}$$

em que

$$c(k,n) := \frac{k(k-1)}{4n} + \frac{k(k+1)}{8n^2} + \frac{1}{12n} - \log\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})}.$$

□

Observamos que a Proposição 2.3.1 vale para um certo conjunto Ω_n. Intuitivamente, é importante garantir que Ω_n^c seja, de alguma forma, negligenciável. É o objeto do Lema 2.3.2 , que será utilizado em conjunto com a Proposição 2.3.1, na prova do Teorema 2.2.1.

Lema 2.3.2. *Sob as hipóteses da Proposição 2.3.1, temos que*

$$\sum_n p(\Omega_n^c|\theta_0) < \infty.$$

Proof. Queremos limitar $p(\Omega_n^c|\theta_0)$ superiormente. Temos

$$\begin{aligned} p(\Omega_n^c|\theta_0) &= \sum_{z\in[k_0]^n} p(\Omega_n^c, z|\theta_0) \\ &= \sum_{z\in[k_0]^n} p(z|\theta_0)p(\Omega_n^c|\theta_0, z) \\ &\leq \sum_{z\in[k_0]^n} p(z|\theta_0) \sum_{i,j=1}^{n} p(X_{ij} > \log n|\theta_0, z). \end{aligned} \tag{2.31}$$

Precisamos limitar superiormente $p(X_{ij} > \log n|\theta_0, z)$. As desigualdades de Chernoff permitem limitantes precisos que vamos relembrar rapidamente abaixo (baseamo-nos na Seção 2.2 de (Boucheron et al., 2013)). Consideramos uma v.a. Z a valores reais, e

$$\psi_Z(\alpha) = \log E(e^{\alpha Z}),$$

para todo $\alpha \geq 0$, o logaritmo da sua função geradora de momentos. A desigualdade de Chernoff estabelece que

$$P(Z \geq r) \leq e^{-\psi_Z^\star(r)},$$

em que

$$\psi^*(r) = \sup_\alpha(\alpha r - \psi_Z(\alpha))$$

é a transformada de Fenchel-Legendre de ψ_Z. Fazendo $Z = Y - \lambda$, onde Y tem distribuição Poisson(λ), temos

$$\psi^\star(r) = \lambda\left(1 + \frac{r}{\lambda}\right)\log\left(1 + \frac{r}{\lambda}\right) - r$$

e portanto

$$\begin{aligned} P(Z \geq r) &\leq e^{-\lambda\left(1+\frac{r}{\lambda}\right)\log\left(1+\frac{r}{\lambda}\right)+r} \\ &\leq e^{-r\left(\log\left(1+\frac{r}{\lambda}\right)-1\right)}. \end{aligned} \tag{2.32}$$

Voltando às X_{ij}, condicionalmente a z, elas têm distribuição Poisson(λ_{z_i,z_j}) para $z_i, z_j \in [k_0]$. Seja $\lambda^\star = \sup_{a,b}\lambda_{a,b}$, temos por (2.32)

$$p(X_{ij} \geq r + \lambda|\theta_0, z) \leq e^{-r\left(\log\left(1+\frac{r}{\lambda_{z_i,z_j}}\right)-1\right)}.$$

Dessa forma, temos que

$$\begin{aligned} p(X_{ij} > \log n|\theta_0, z) &= p(X_{ij} - \lambda_{z_i,z_j} > \log n - \lambda_{z_i,z_j}) \\ &\leq e^{(\log n - \lambda_{z_i,z_j})\left[\log\left(\frac{\log n}{\lambda_{z_i,z_j}}\right)-1\right]} \\ &= n^{-(\log(\log n)-\log\lambda_{z_i,z_j}-1)} e^{\lambda_{z_i,z_j}(\log(\log n)-\log\lambda_{z_i,z_j}-1)} \\ &\leq n^{-\log(\log n)} n^{(\log\lambda_{z_i,z_j}+1)} (\log n)^{\lambda_{z_i,z_j}} \\ &\leq n^{\log(\log n)} n^{(\log\lambda^*+1)} (\log n)^{\lambda^*}. \end{aligned} \tag{2.33}$$

Portanto, substituindo (2.33) em (2.31) segue que

$$\begin{aligned} p(\Omega_n^c|\theta_0) &\leq \sum_{z\in[k_0]^n} p(z|\theta_0) \sum_{i,j=1}^{n} p(X_{ij} > \log n|\theta_0, z) \\ &\leq n^2 n^{\log(\log n)} n^{(\log\lambda^*+1)} (\log n)^{\lambda^*} \sum_{z\in[k_0]^n} p(z|\theta_0) \\ &= n^{(\log(\log n)+\log\lambda^*+3)} (\log n)^{\lambda^*}. \end{aligned}$$

que é somável em n. □

2.3.2 Não superestimação

O objetivo desta Seção é provar que o estimador do número de comunidades $\widehat{k}_n$, não superestima a verdadeira ordem k_0 do modelo estocástico de blocos de Poisson. O resultado principal é a Proposição 2.3.3, enunciada a seguir.

Proposição 2.3.3. *Seja um modelo estocástico de blocos de Poisson de ordem k_0, com parâmetros $\theta_0 = (\lambda^0, \pi^0)$. Então, para o estimador da ordem, $\widehat{k}_n$ definido em (2.13), temos que*

$$p(\{\widehat{k}_n > k_0\} \textit{ infinitas vezes}\,|\theta_0) = 0.$$

Proof. Mostraremos que $\sum_n p(\widehat{k}_n > k_0|\theta_0) < \infty$, o que mostra, pelo Lema de Borel-Cantelli, a proposição.

Temos

$$\sum_{k=k_0+1}^{\infty} p(\widehat{k}_n = k|\theta_0) = \sum_{k=k_0+1}^{\infty} \sum_{x\in\Omega_n} p(x|\theta_0) \mathbb{1}_{\{\widehat{k}_n(x)=k\}} + \sum_{k=k_0+1}^{\infty} \sum_{x\in\Omega_n^c} p(x|\theta_0) \mathbb{1}_{\{\widehat{k}_n(x)=k\}}. \tag{2.34}$$

Iniciamos pelo segundo termo da soma em (2.34), que é mais fácil: temos que

$$\sum_{k=k_0+1}^{\infty}\sum_{x\in\Omega_n^c} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}} = \sum_{x\in\Omega_n^c} p(x|\theta_0)\sum_{k=k_0+1}^{\infty}\mathbb{1}_{\{\widehat{k}_n(x)=k\}} \leq \sum_{x\in\Omega_n^c} p(x|\theta_0)$$

que, por Lema 2.3.2 é somável em n.

No resto da prova, consideramos o primeiro termo em (2.34). Temos

$$\sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}} = \sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\arg\max_{k'}(\log p_{k'}(x)-\mathrm{pen}(k',n))=k\}}. \tag{2.35}$$

Da definição do estimador da ordem, temos que

$$\{\arg\max_{k'}(\log p_{k'}(x) - \mathrm{pen}(k',n)) = k\} \subset \{\log p_k(x) - k^3\log n \geq \log p_{k_0}(x) - k_0^3\log n\},$$

para todo $k > k_0$. Então, em (2.35) segue que

$$\begin{aligned}\sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}} &\leq \sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\log p_k(x)-k^3\log n\geq \log p_{k_0}(x)-k_0^3\log n\}} \\ &= \sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{p_{k_0}(x)\leq p_k(x)\exp\{k_0^3\log n - k^3\log n\}\}}. \end{aligned}\tag{2.36}$$

Por outro lado, para todo $x\in\Omega_n$, pela Proposição 2.3.1 temos que

$$\begin{aligned}\log p(x|\theta_0) &\leq \log\sup_{\theta\in\Theta_{k_0}} p(x|\theta). \\ &\leq \log p_{k_0}(x) + d(k_0)\log n + c(k_0,n). \end{aligned}\tag{2.37}$$

Daqui, tomando função exponencial em (2.37) resulta que

$$p(x|\theta_0) \leq p_{k_0}(x)\exp\left\{d(k_0)\log n + c(k_0,n)\right\}. \tag{2.38}$$

Defina

$$G(k,k_0,n) := d(k_0)\log n + c(k_0,n) + k_0^3\log n - k^3\log n.$$

Agora, substituindo (2.38) em (2.36), temos que

$$\sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}} \leq \exp\{G(k,k_0,n)\}\sum_{x\in\Omega_n} p_k(x) \leq \exp\{G(k,k_0,n)\}.$$

Portanto, o primeiro termo de (2.34) é majorado por

$$\sum_{k=k_0+1}^{\infty} \exp\{G(k,k_0,n)\} = \exp\left\{d(k_0)\log n + c(k_0,n) + k_0^3 \log n\right\} \sum_{k=k_0+1}^{\infty} \exp\left\{-k^3 \log n\right\}. \tag{2.39}$$

Por outro lado, temos que

$$\sum_{k=k_0+1}^{\infty} \exp\left\{-k^3 \log n\right\} = \sum_{k=k_0+1}^{\infty} \left(\frac{1}{n}\right)^{k^3} \tag{2.40}$$

em que,

$$\begin{aligned}\sum_{k=k_0+1}^{\infty} \left(\frac{1}{n}\right)^{k^3} = \sum_{k\geq 0} \left(\frac{1}{n}\right)^{[k+(k_0+1)]^3} \leq \sum_{k\geq 0} \left(\frac{1}{n}\right)^{k^3+(k_0+1)^3} &= \left(\frac{1}{n}\right)^{(k_0+1)^3} \sum_{k\geq 0} \left(\frac{1}{n}\right)^{k^3} \\ &= C_1(n) \left(\frac{1}{n}\right)^{(k_0+1)^3}\end{aligned}$$

onde, $C_1(n) = \sum_{k\geq 0} \left(\frac{1}{n}\right)^{k^3} = O((\log n)^{-1/3})$. Substituindo (2.40) em (2.39), segue que

$$\sum_{k=k_0+1}^{\infty} \exp\{G(k,k_0,n)\} \leq e^{c(k_0,n)} n^{\{d(k_0)+k_0^3-(k_0+1)^3\}}, \tag{2.41}$$

que é somável em n pois para todo k_0,

$$d(k_0) + k_0^3 - (k_0+1)^3 = \frac{k_0(2k_0+3)-1}{2} + k_0^3 - k_0^3 - 3k_0^2 - 3k_0 - 1 < -1.$$

□

2.3.3 Não subestimação

O objetivo desta Seção é mostrar que o estimador do número de comunidades $\widehat{k}_n$, não subestima a verdadeira ordem k_0 do modelo estocástico de blocos de Poisson. O resultado principal é a Proposição 2.3.4, enunciada a seguir.

Proposição 2.3.4. *Seja um modelo estocástico de blocos de Poisson de ordem k_0 e $\rho_n \geq C\frac{\log n}{n}$, onde C é uma constante suficientemente grande. Então, para o estimador $\widehat{k}_n$ definido em* (2.13)*, temos que*

$$p(\{\widehat{k}_n < k_0\} \textit{ infinitas vezes} \,|\theta_0) = 0.$$

Para a prova deste resultado, definimos o estimador das comunidades com base no grafo observado sob o modelo de ordem k como sendo

$$z^* = \underset{z\in\{1,2,\ldots,k\}^n}{\arg\max} \sup_{\theta\in\Theta_k} p(x, z|\theta). \tag{2.42}$$

Logo temos que $p(x, z^*|\theta)$ é o máximo da função de verossimilhança conjunta em relação a θ e z.

Demonstração da Proposição 2.3.4. Em vista do Lema 2.3.2, onde foi mostrado que a soma em n da probabilidade do grafo ter mais do que $\log n$ elos é somável, no que segue desta prova assumiremos que $x \in \Omega_n$, para n suficientemente grande. Lembramos a definição do estimador da ordem, dado por

$$\widehat{k}_n(x) = \arg\max_k \{\log p_k(x) - k^3 \log n\}\,. \tag{2.43}$$

Para mostrar que $\widehat{k}_n(x) \geq k_0$, quase certamente quando $n \to \infty$, é suficiente mostrar que para todo $k < k_0$,

$$\log p_{k_0}(x) - k_0^3 \log n > \log p_k(x) - k^3 \log n\,, \tag{2.44}$$

quase certamente, quando $n \to \infty$. Mas, se mostrarmos que

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{p_{k_0}(x)}{p_k(x)} > 0 \tag{2.45}$$

pelo fato de

$$\lim_{n\to\infty} \frac{1}{\rho_n n^2} \left(k_0^3 \log n - k^3 \log n\right) = 0$$

então (2.45) implica (2.44). Observe que para todo $x \in \Omega_n$, pela Proposição 2.3.1 e o fato de $p_k(x) \leq \sup_{\theta\in\Theta_k} p(x|\theta)$, segue que

$$\begin{aligned}
\frac{1}{\rho_n n^2} \log \frac{p_{k_0}(x)}{p_k(x)} &= \frac{1}{\rho_n n^2} \log \frac{p_{k_0}(x)}{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)} + \frac{1}{\rho_n n^2} \log \frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} \\
&\quad + \frac{1}{\rho_n n^2} \log \frac{\sup_{\theta\in\Theta_k} p(x|\theta)}{p_k(x)} \\
&\geq -d(k_0)\frac{\log n}{\rho_n n^2} - \frac{c(k_0, n)}{\rho_n n^2} + \frac{1}{\rho_n n^2} \log \frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)}.
\end{aligned}$$

Assim, para mostrar (2.45) basta mostrar que para $k < k_0$

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} > 0 \tag{2.46}$$

já que

$$-d(k_0)\frac{\log n}{\rho_n n^2} - \frac{c(k_0,n)}{\rho_n n^2} \to 0$$

quando $n \to \infty$. Primeiro observamos que para todo z temos que

$$\log \sup_{\theta\in\Theta_{k_0}} p(x|\theta) \geq \log \sup_{\theta\in\Theta_{k_0}} p(x,z|\theta)\,. \tag{2.47}$$

A partir de (2.9) temos que a expressão no lado direito está dada por

$$\begin{aligned} \log \sup_{\theta\in\Theta_{k_0}} p(x,z|\theta) &= L(x) + \sum_{1\leq a\leq b\leq k_0} O_{a,b}(x,z)\log\widehat{\lambda}_{a,b}(x,z) + \sum_{a=1}^{k_0} n_a(z)\log\widehat{\pi}_a(z) \\ &= L(x) + \sum_{1\leq a\leq b\leq k_0} n_{a,b}(z)\varphi(\widehat{\lambda}_{a,b}(x,z)) + n\sum_{a=1}^{k_0}\widehat{\pi}_a(z)\log\widehat{\pi}_a(z), \end{aligned} \tag{2.48}$$

com

$$\begin{aligned} L(x) &= \bar{c}(x) - M(x)\,; \\ \widehat{\pi}_a(z) &= \frac{n_a(z)}{n}, \qquad 1 \leq a \leq k_0; \\ \widehat{\lambda}_{a,b}(x,z) &= \frac{O_{a,b}(x,z)}{n_{a,b}(z)}, \qquad 1 \leq a \leq b \leq k_0; \\ \varphi(u) &= u\log u, \qquad u > 0 \end{aligned}$$

Para o denominador em (2.46) usamos que

$$\log \sup_{\theta\in\Theta_k} p(x|\theta) \leq \log[k^n \sup_{\theta\in\Theta_k} p(x,z^*|\theta)] = n\log k + \log \sup_{\theta\in\Theta_k} p(x,z^*|\theta)\,, \tag{2.49}$$

com z^* definido por (2.42). Analogamente a (2.48) temos que

$$\begin{aligned} \log \sup_{\theta\in\Theta_k} p(x,z^*|\theta) &= L(x) + \sum_{1\leq a\leq b\leq k} O_{a,b}(x,z^*)\log\widehat{\lambda}_{a,b}(x,z^*) + \sum_{a=1}^{k} n_a(z^*)\log\widehat{\pi}_a(z^*) \\ &= L(x) + \sum_{1\leq a\leq b\leq k} n_{a,b}(z^*)\varphi(\widehat{\lambda}_{a,b}(x,z^*)) + n\sum_{a=1}^{k}\widehat{\pi}_a(z^*)\log\widehat{\pi}_a(z^*). \end{aligned} \tag{2.50}$$

Logo, o logaritmo em (2.46) pode ser limitado inferiormente pela diferença de (2.47) e (2.49), e usando as expressões em (2.48) e (2.50) obtemos que

$$\log \frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} \;\geq\; \sum_{1\leq a\leq b\leq k_0} n_{a,b}(z)\varphi(\widehat{\lambda}_{a,b}(x,z)) + n\sum_{a=1}^{k_0}\widehat{\pi}_a(z)\log\widehat{\pi}_a(z) - n\log k$$
$$- \sum_{1\leq a\leq b\leq k} n_{a,b}(z^*)\varphi(\widehat{\lambda}_{a,b}(x,z^*)) - n\sum_{a=1}^{k}\widehat{\pi}_a(z^*)\log\widehat{\pi}_a(z^*). \tag{2.51}$$

Observamos que dividindo os dois lados por $\rho_n n^2$ temos, primeiramente que

$$\frac{1}{\rho_n n^2}\Big[n\sum_{a=1}^{k_0}\widehat{\pi}_a(z)\log\widehat{\pi}_a(z) - n\log k - n\sum_{a=1}^{k}\widehat{\pi}_a(z^*)\log\widehat{\pi}_a(z^*)\Big] \;\to\; 0$$

já que $\rho_n n \to \infty$ e os termos nas somatórias ficam limitados, quando $n \to \infty$. Então, para provar (2.46) basta mostrar que

$$\liminf_{n\to\infty}\; \frac{1}{\rho_n n^2}\left[\sum_{1\leq a,b\leq k_0} n_{a,b}(z)\varphi(\widehat{\lambda}_{a,b}(x,z)) - \sum_{1\leq a,b\leq k} n_{a,b}(z^*)\varphi(\widehat{\lambda}_{a,b}(x,z^*))\right] \;>\; 0\,. \tag{2.52}$$

Observe que no caso esparso, $\widehat{\lambda}_{a,b}(x,z) \to 0$ quando $n \to \infty$, e nesse caso cada termo da diferença acima está indefinido. Mas somando e subtraindo a expressão

$$\sum_{1\leq a\leq b\leq k_0} O_{a,b}(x,z)\log\rho_n \;=\; \sum_{1\leq a\leq b\leq k} O_{a,b}(x,z)\log\rho_n \;=\; M(x)\log\rho_n$$

com $M(x)$ o número de arestas no grafo temos que o limite em (2.52) é equivalente a

$$\liminf_{n\to\infty}\; \frac{1}{n^2}\left[\sum_{1\leq a\leq b\leq k_0} n_{a,b}(z)\varphi\Big(\frac{\widehat{\lambda}_{a,b}(x,z)}{\rho_n}\Big) - \sum_{1\leq a\leq b\leq k} n_{a,b}(z^*)\varphi\Big(\frac{\widehat{\lambda}_{a,b}(x,z^*)}{\rho_n}\Big)\right] \;>\; 0\,. \tag{2.53}$$

Pela Lei dos Grandes Números temos que $\widehat{\lambda}_{a,b}(x,z)/\rho_n \to \tilde{\lambda}^0_{a,b}$, $\tilde{\lambda}^0_{a,b} > 0$ para quase todo par (x,z), quando $n \to \infty$. Lembramos que $\tilde{\lambda}^0$ é a matriz, tal que $\lambda^0 = \rho_n\tilde{\lambda}^0$. Logo, pelo fato de φ ser uma função convexa e portanto contínua segue que

$$\lim_{n\to\infty}\; \frac{1}{n^2}\sum_{1\leq a\leq b\leq k_0} n_{a,b}(z)\varphi\Big(\frac{\widehat{\lambda}_{a,b}(x,z)}{\rho_n}\Big) \;=\; \frac{1}{2}\sum_{1\leq a,b\leq k_0}\pi^0_a\pi^0_b\varphi(\tilde{\lambda}^0_{ab})\,. \tag{2.54}$$

Por outro lado, pelo Lemma 2.3.5 temos que

$$\limsup_{n\to\infty} \frac{1}{n^2} \sum_{1\le a\le b\le k} n_{a,b}(z^*)\varphi\Big(\frac{\widehat{\lambda}_{a,b}(x,z^*)}{\rho_n}\Big) \le \frac{1}{2}\sum_{1\le a,b\le k} \pi_a^*\pi_b^*\varphi(\lambda_{ab}^*)\,. \tag{2.55}$$

Finalmente, pelo Lemma 2.3.6 temos que a diferença de (2.54) e (2.55) é igual a

$$\frac{1}{2}\left(\sum_{1\le a,b\le k_0} \pi_a^0\pi_b^0\varphi(\tilde{\lambda}_{ab}^0) - \sum_{1\le a,b\le k} \pi_a^*\pi_b^*\varphi(\lambda_{ab}^*)\right) > 0$$

e isto conclui a demonstração da Proposição 2.3.4. □

A seguir, apresentamos os Lemas 2.3.5 e 2.3.6, que foram utilizados na prova da Proposição 2.3.4.

Lema 2.3.5. *Para $k < k_0$, temos que existe uma matriz positiva λ^*, $k\times k$ e um vetor π^*, k-dimensional tal que*

$$\limsup_{n\to\infty} \frac{1}{n^2} \sum_{1\le a\le b\le k} n_{a,b}(z^*)\varphi\Big(\frac{\widehat{\lambda}_{a,b}(x,z^*)}{\rho_n}\Big) \le \frac{1}{2}\sum_{1\le a,b\le k} \pi_a^*\pi_b^*\varphi(\lambda_{ab}^*), \tag{2.56}$$

em que (λ^*,π^*), são dados por:

$$\begin{aligned}\lambda_{a,b}^* &= \frac{\left[R^*\tilde{\lambda}^0 R^{*^T}\right]_{a,b}}{\left[R^*\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T R^{*^T}\right]_{a,b}}, \qquad a,b\in\{1,\dots,k\}\\ \pi_a^* &= [R^*\mathbf{1}_{k_0}]_a, \qquad a\in\{1,\dots,k\}\,.\end{aligned}$$

para uma matriz R^*, $k\times k_0$ que satisfaz $\|R^*\|_1 = 1$ e tem uma entrada diferente de zero em cada coluna.

Proof. Definamos para todo $z\in\{1,\dots,k\}^n$ e $z^0\in\{1,\dots,k_0\}^n$, $Q_n(z,z^0)$ como a matriz $k\times k_0$ com entradas dadas por

$$[Q_n(z,z^0)]_{a,a'} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{z_i = a, z_i^0 = a'\}. \tag{2.57}$$

Observamos que, os contadores $n_a(z^0)$, para $a\in[k_0]$, podem ser escritos como

$$n_{a'}(z^0) = \sum_{i=1}^n\sum_{a=1}^k \mathbb{1}\{z_i = a, z_i^0 = a'\}\,, \tag{2.58}$$

assim que, $n_a(z^0) = n(Q_n^T(z, z^0)\mathbf{1}_k)_a$, em que $\mathbf{1}_k$ é um vetor coluna de dimensão k com todas as entradas iguais a 1. E os contadores $n_a(z)$, para $a \in [k]$,

$$n_a(z) = \sum_{i=1}^{n} \sum_{a'=1}^{k_0} \mathbb{1}\{z_i = a, z_i^0 = a'\}\,, \tag{2.59}$$

assim que, $n_a(z) = n(Q_n(z, z^0)\mathbf{1}_{k_0})_a$. Além disso, a matriz $Q_n(z, z^0)$ satisfaz

$$\|Q_n(z, z^0)\|_1 \;=\; \sum_{a=1}^{k} \sum_{a'=1}^{k_0} [Q_n(z, z^0)]_{a,a'} = 1 \tag{2.60}$$

para todo (z, z^0). Por outro lado, temos para todo $a, b \in \{1, \ldots, k\}$ que

$$\begin{aligned} O_{a,b}(x, z) \;&=\; \sum_{1 \le i,j \le n} x_{ij} \mathbb{1}\{z_i = a, z_j = b\} \\ &= \sum_{1 \le a', b' \le k_0} \sum_{1 \le i,j \le n} x_{ij} \mathbb{1}\{z_i = a, z_i^0 = a', z_j = b, z_j^0 = b'\}\,. \end{aligned}$$

Por outro lado, observamos que

$$\begin{aligned} &\sum_{1 \le a,b \le k} \frac{n_a(z^*) n_b(z^*)}{n^2} \varphi\left(\frac{\widehat{\lambda}_{a,b}(x, z^*)}{\rho_n}\right) \;=\; \sum_{1 \le a,b \le k} \frac{n_a(z^*) n_b(z^*)}{n^2} \varphi\left(\frac{O_{a,b}(x, z^*)}{\rho_n n_a(z^*) n_b(z^*)}\right) \\ &= \sum_{1 \le a,b \le k} [Q_n(z^*, z^0)\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T Q_n^T(z^*, z^0)]_{a,b} \varphi\left(\frac{O_{a,b}(x, z^*)/\rho_n n^2}{[Q_n(z^*, z^0)\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T Q_n^T(z^*, z^0)]_{a,b}}\right), \end{aligned}$$

com $\varphi(x) = x \log x$. Então pelo Lema A.2.2, para alguma sequência $\epsilon_n \to 0$ temos que

$$[Q_n(z^*, z^0)\tilde{\lambda}^0 Q_n^T(z^*, z^0)]_{a,b} - \epsilon_n \le \frac{O_{a,b}(x, z^*)}{\rho_n n^2} \le [Q_n(z^*, z^0)\tilde{\lambda}^0 Q_n^T(z^*, z^0)]_{a,b} + \epsilon_n,$$

quase certamente quando $n \to \infty$. Então, como φ é contínua podemos ver que

$$\begin{aligned} &\sum_{1 \le a,b \le k} \frac{n_a(z^*) n_b(z^*)}{n^2} \varphi\left(\frac{\widehat{\lambda}_{a,b}(x, z^*)}{\rho_n}\right) \\ &\le \sum_{1 \le a,b \le k} [Q_n(z^*, z^0)\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T Q_n^T(z^*, z^0)]_{a,b} \varphi\left(\frac{[Q_n(z^*, z^0)\tilde{\lambda}^0 Q_n^T(z^*, z^0)]_{a,b} \pm \epsilon_n}{[Q_n(z^*, z^0)\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T Q_n^T(z^*, z^0)]_{a,b}}\right). \end{aligned} \tag{2.61}$$

Assumimos que $Q_n(z, z^0)$ tem uma subsequência convergente cujo limite iremos chamar de R, tal que $R(z, z^0)$ é alguma matriz de ordem $k \times k_0$ satisfazendo que $\|R(z, z^0)\| = 1$

e $R(z, z^0)^T \mathbf{1}_k = \pi^0$. Então, tomando limite a ambos lados de (2.61) segue que

$$\begin{aligned} \limsup_{n\to\infty} & \sum_{1\le a\le b\le k} \frac{n_a(z^*)n_b(z^*)}{n^2} \varphi\left(\frac{\widehat{\lambda}_{a,b}(x,z^*)}{\rho_n}\right) \\ & \le \sup_{\substack{R:\,\|R\|_1=1 \\ R^T\mathbf{1}_k=\pi^0}} \frac{1}{2} \sum_{1\le a,b\le k} \left[R\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T R^T\right]_{a,b} \varphi\left(\frac{\left[R\tilde{\lambda}^0 R^T\right]_{a,b}}{\left[R\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T R^T\right]_{a,b}}\right). \end{aligned} \tag{2.62}$$

Logo, o supremo em (2.62) corresponde a um problema de maximizar uma função convexa sobre um poliedro convexo definido por $\{R\colon \|R\|_1 = 1, R^T\mathbf{1}_k = \pi^0\}$. Observamos que o problema de maximizar uma função convexa é equivalente ao problema de minimizar uma função côncava. Logo, pelo Proposição A.2.1 temos que o máximo deve ser atingido em algum dos vértices do poliedro; isto é, naquelas matrizes R tais que uma e somente uma entrada por coluna é maior do que zero, considerando que $\pi_a^0 > 0$ para todo $a \in \{1,\dots,k_0\}$. Denotamos por R^* a um destes máximos e sejam

$$\begin{aligned} \lambda^*_{a,b} &= \frac{\left[R^*\tilde{\lambda}^0 {R^*}^T\right]_{a,b}}{\left[R^*\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T {R^*}^T\right]_{a,b}}, \qquad a,b \in \{1,\dots,k\} \\ \pi^*_a &= [R^*\mathbf{1}_{k_0}]_a \qquad a \in \{1,\dots,k\}. \end{aligned} \tag{2.63}$$

Portanto, temos que

$$\sup_{\substack{R:\,\|R\|_1=1 \\ R^T\mathbf{1}_k=\pi^0}} \frac{1}{2} \sum_{1\le a,b\le k} \left[R\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T R^T\right]_{a,b} \varphi\left(\frac{\left[R\tilde{\lambda}^0 R^T\right]_{a,b}}{\left[R\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T R^T\right]_{a,b}}\right) = \frac{1}{2}\sum_{1\le a,b\le k} \pi^*_a\pi^*_b\varphi(\lambda^*_{ab}).$$

Isso conclui a prova do Lema 2.3.5. □

Lema 2.3.6. *Para todo $k < k_0$ e (λ^*, π^*) como no Lema 2.3.5 temos que*

$$\sum_{1\le a,b\le k_0} \pi^0_a\pi^0_b\varphi(\tilde{\lambda}^0_{ab}) - \sum_{1\le a,b\le k} \pi^*_a\pi^*_b\varphi(\lambda^*_{ab}) \;>\; 0\,. \tag{2.64}$$

Proof. Consideramos primeiro o caso $k = k_0 - 1$. Então como h é uma função sobrejetora, há $k-1$ comunidades em $\{1,\dots,k_0\}$ que são mapeadas em $k-1$ comunidades em $\{1,\dots,k\}$ e duas comunidades em $\{1,\dots,k_0\}$ que são mapeadas numa única comunidade em $\{1,\dots,k\}$. Sem perda de generalidade, assumamos que as comunidades $k_0 - 1$ e k_0

satisfazem $h(k_0-1)=h(k_0)=k=k_0-1$. Logo, temos que

$$\begin{aligned}\sum_{1\le a,b\le k_0}\pi_a^0\pi_b^0\varphi(\tilde\lambda_{ab}^0)-\sum_{1\le a,b\le k}\pi_a^*\pi_b^*\varphi(\lambda_{ab}^*) \;=\; & 2\sum_{1\le a\le k_0-2}\pi_a^0\pi_{k_0-1}^0\varphi(\tilde\lambda_{a,k_0-1}^0)\\ &+2\sum_{1\le a\le k_0-2}\pi_a^0\pi_{k_0}^0\varphi(\tilde\lambda_{a,k_0}^0)+2\pi_{k_0-1}^0\pi_{k_0}^0\varphi(\tilde\lambda_{k_0-1,k_0}^0)+\pi_{k_0-1}^0\pi_{k_0-1}^0\varphi(\tilde\lambda_{k_0-1,k_0-1}^0)\\ &+\pi_{k_0}^0\pi_{k_0}^0\varphi(\tilde\lambda_{k_0,k_0}^0)-2\sum_{1\le a\le k_0-2}\pi_a^*\pi_{k_0-1}^*\varphi(\lambda_{a,k_0-1}^*)-\pi_{k_0-1}^*\pi_{k_0-1}^*\varphi(\lambda_{k_0-1,k_0-1}^*)\,.\end{aligned}\tag{2.65}$$

Esta igualdade segue do fato que $\pi_i^*=\pi_i^0$ para todo $1\le i\le k_0-2$ e $\lambda_{l,r}^*=\tilde\lambda_{l,r}^0$ para todo $1\le l<r\le k_0-2$. Além disso, observamos que

$$\pi_{k_0-1}^*=\pi_{k_0-1}^0+\pi_{k_0}^0$$

e

$$\begin{aligned}\lambda_{l,k_0-1}^*&=\frac{\pi_l^0\pi_{k_0-1}^0\tilde\lambda_{l,k_0-1}^0+\pi_l^0\pi_{k_0}^0\tilde\lambda_{l,k_0}^0}{\pi_l^0\pi_{k_0-1}^0+\pi_l^0\pi_{k_0}^0}, && 1\le l\le k_0-2,\\ \lambda_{k_0-1,k_0-1}^*&=\frac{\pi_{k_0-1}^0\pi_{k_0-1}^0\tilde\lambda_{k_0-1,k_0-1}^0+2\pi_{k_0-1}^0\pi_{k_0}^0\tilde\lambda_{k_0-1,k_0}^0+\pi_{k_0}^0\pi_{k_0}^0\tilde\lambda_{k_0,k_0}^0}{\pi_{k_0-1}^0\pi_{k_0-1}^0+2\pi_{k_0-1}^0\pi_{k_0}^0+\pi_{k_0}^0\pi_{k_0}^0}.\end{aligned}$$

Para $1\le a\le k_0-2$ segue que

$$\begin{aligned}\pi_a^*\pi_{k_0-1}^*\varphi(\lambda_{a,k_0-1}^*)&=(\pi_a^0\pi_{k_0-1}^0+\pi_a^0\pi_{k_0}^0)\varphi\left(\frac{\pi_a^0\pi_{k_0-1}^0\tilde\lambda_{a,k_0-1}^0+\pi_a^0\pi_{k_0}^0\tilde\lambda_{a,k_0}^0}{\pi_a^0\pi_{k_0-1}^0+\pi_a^0\pi_{k_0}^0}\right)\\ &\le\pi_a^0\pi_{k_0-1}^0\varphi(\tilde\lambda_{a,k_0-1}^0)+\pi_a^0\pi_{k_0}^0\varphi(\tilde\lambda_{a,k_0}^0),\end{aligned}\tag{2.66}$$

dado que φ é uma função convexa, e a desigualdade em (2.66) segue pela desigualdade de Jensen. Para $a=k_0-1$ e $b=k_0$, também pela convexidade de φ e a desigualdade de Jensen temos que

$$\begin{aligned}\pi_{k_0-1}^*\pi_{k_0-1}^*\varphi(\lambda_{k_0-1,k_0-1}^*) \;\le\; & \pi_{k_0-1}^0\pi_{k_0-1}^0\varphi(\tilde\lambda_{k_0-1,k_0-1}^0)\\ &+\;2\pi_{k_0-1}^0\pi_{k_0}^0\varphi(\tilde\lambda_{k_0-1,k_0}^0)+\pi_{k_0}^0\pi_{k_0}^0\varphi(\tilde\lambda_{k_0,k_0}^0).\end{aligned}\tag{2.67}$$

Logo, substituindo (2.66) e (2.67) em (2.65) resulta que

$$\sum_{1\le a,b\le k_0}\pi_a^0\pi_b^0\varphi(\tilde\lambda_{ab}^0)-\sum_{1\le a,b\le k}\pi_a^*\pi_b^*\varphi(\lambda_{ab}^*)\;\ge\;0\,.\tag{2.68}$$

Ainda, a desigualdade acima deve ser estrita, a não ser que $\tilde{\lambda}^0_{a,k_0} = \tilde{\lambda}^0_{a,k_0-1}$ para todo $a = 1, \ldots, k_0$. Isso significaria que a matriz $\tilde{\lambda}^0$ tem duas colunas iguais, o que contradiz a Suposição 2.1.1, que é a suposição de identificabilidade do modelo. □

Chapter

3

Estimação consistente no modelo estocástico de blocos com correção de grau

No capítulo 2 vimos que no modelo estocástico de blocos de Poisson, condicionalmente na alocação das comunidades, os vértices são conectados de forma independente com uma probabilidade que depende das comunidades. Porém, este fato impõe distribuições de grau homogêneas para todos os vértices dentro de cada comunidade, o que pode ser uma limitação quando se trata de modelar redes da vida real. Para isso, o modelo estocástico de blocos com correção de grau, também introduzido por Karrer e Newman (2011), permite a heterogeneidade dos graus dentro das comunidades.

3.1 Modelo estocástico de blocos com correção de grau

3.1.1 Definição do modelo

O modelo estocástico de blocos com correção de grau (Degree Corrected Stochastic Block Model) é um modelo de grafos aleatórios não orientado.

De forma similar ao modelo anterior, consideramos $[n] := \{1, \ldots, n\}$ um conjunto de $n \geq 1$ vértices, separados de forma aleatória entre $k_0 \geq 1$ possíveis grupos. O vetor aleatório $Z = (Z_i)_{i \in [n]}$, com $Z_i \in [k_0] := \{1, \ldots, k_0\}$, representa a associação de cada

vértice a um grupo. Assumimos que este vetor é i.i.d. com distribuição marginal π sobre $[k_0]$.

O modelo de blocos com correção de grau se comporta da mesma forma que o modelo de blocos de Poisson só que agora incluimos um peso w_i, tal que w_i representa o grau esperado do vértice i, para cada vértice $i \in [n]$. Dessa forma, as entradas $X_{ij}, i \geq j$ da matriz de adjacência são independentes com distribuição de Poisson, em que o número esperado de arestas entre vértices $i \neq j$ é dado por

$$\mathbb{E}(X_{ij}|Z_i = a, Z_j = b) = w_i w_j \lambda_{a,b}.$$

Note que para $w_i = 1$, para todo $i \in [n]$ o modelo de blocos com correção de grau se reduz ao modelo de blocos de Poisson (ver Seção 2.1). Além disso, dentro de cada comunidade, os pesos w_i satisfazem a seguinte condição:

$$\sum_{i:\ z_i=a} w_i \ = \ n_a(z),$$

para todo $a \in [k_0]$. Isso implica que o peso total no grafo é n, o número de vértices.

Suposição 3.1.1. *Suponhamos que nenhuma coluna da matriz simétrica $\tilde{\lambda}$ é proporcional a qualquer outra coluna.*

A Suposição 3.1.1 é uma condição de identificabilidade do modelo. Esta condição é usualmente assumida na literatura, por exemplo Ma et al. (2021).

3.1.2 A verossimilhança do modelo

Denotamos por $(\Omega, \mathcal{F}, \mathbb{P}_{(\lambda,w,\pi)})$ o espaço de probabilidade associado à matriz simétrica λ, ao vetor de pesos w e à distribuição de probabilidades π. Temos que a verossimilhança conjunta para todo $z \in [k_0]^n$ e $x \in \mathbb{N}^{n\times n}$ é dada por:

$$\mathbb{P}_{(\lambda,w,\pi)}(X = x, Z = z) = \mathbb{P}_{(\lambda,w,\pi)}(Z = z)\mathbb{P}_{(\lambda,w,\pi)}(X = x|Z = z),$$

onde

$$\mathbb{P}_{(\lambda,w,\pi)}(Z = z) = \prod_{a=1}^{k} \pi_a^{\sum_{i=1}^{n} \mathbb{1}\{z_i=a\}},$$

não depende de λ nem de w e

$$\begin{aligned}&\mathbb{P}_{(\lambda,w,\pi)}(X = x|Z = z)\\ &= \left[\prod_{1\le i<j\le n} \frac{(w_i w_j \lambda_{z_i z_j})^{x_{ij}} \exp\{-w_i w_j \lambda_{z_i z_j}\}}{x_{ij}!}\right]\left[\prod_{1\le i\le n} \frac{(w_i^2 \lambda_{z_i z_i})^{x_{ii}/2} \exp\{-w_i^2 \lambda_{z_i z_i}\}}{\frac{x_{ii}}{2}!}\right],\end{aligned}$$

não depende de π.

Para escrever a verossimilhança conjunta de (X, Z) como função de k, escrevemos novamente os contadores definidos na seção 2.1.2 e que são dados por:

$$\begin{aligned} o_{a,b}(x,z) &= \sum_{1\le i,j\le n} x_{ij}\mathbb{1}\{z_i = a, z_j = b\}, \quad 1 \le a, b \le k\\ n_a(z) &= \sum_{1\le i\le n} \mathbb{1}\{z_i = a\}, \quad 1 \le a \le k\\ O_{a,b}(x,z) &= \begin{cases} o_{a,b}(x,z) & a \neq b,\\ \frac{1}{2}o_{a,b}(x,z) & a = b \end{cases}\\ n_{a,b}(z) &= \begin{cases} n_a(z)n_b(z) & a \neq b,\\ \frac{1}{2}n_a(z)n_b(z) & a = b. \end{cases} \end{aligned} \tag{3.1}$$

Por outro lado, defininos o grau do vértice i por:

$$g_i(x) = \sum_{1\le j\le n} x_{ij}. \tag{3.2}$$

Logo, o grau total na comunidade a é dado por:

$$\bar{g}_a(x,z) = \sum_{i:\, z_i=a} g_i(x) = \sum_{1\le i\le n} g_i(x)\mathbb{1}\{z_i = a\} = \sum_{1\le i,j\le n} x_{ij}\mathbb{1}\{z_i = a\}.$$

Daqui, podemos ver que

$$\bar{g}_a(x,z) = \sum_{b\neq a} o_{ab}(x,z) + 2o_{aa}(x,z).$$

Com estas definições, a verossimilhança conjunta de (X, Z) para o modelo com k comunidades pode ser reescrita como:

$$\mathbb{P}_{(\lambda,\pi)}(X = x, Z = z) = \frac{1}{c(x)} \left[\prod_{1 \le i \le n} {w_i}^{g_i(x)} \right] \left[\prod_{1 \le a \le b \le k} \lambda_{ab}^{o_{ab}(x,z)} \exp\{-n_{ab}(z)\lambda_{ab}\} \right] \left[\prod_{a=1}^{k} {\pi_a}^{n_a(z)} \right], \tag{3.3}$$

com

$$c(x) := \left[\prod_{i<j} x_{ij}! \right] \left[\prod_i 2^{x_{ii}/2} (x_{ii}/2)! \right].$$

3.1.3 Estimadores de máxima verossimilhança

No Lema 3.1.1 apresentamos os estimadores dos parâmetros $\theta = (\lambda, w, \pi)$, do modelo, obtidos por meio do método de máxima verossimilhança.

Lema 3.1.1. *Dada uma realização (x, z) do modelo estocástico de blocos com correção de grau, com função de verossimilhança conjunta $p(x, z|\theta)$ dada em (3.3), então os estimadores de máxima verossimilhança dos parâmetros $\theta_0 = (\lambda^0, w^0, \pi^0)$ são dados por:*

$$\widehat{\lambda}_{a,b}(x, z) \;=\; \frac{O_{a,b}(x, z)}{n_{a,b}(z)}, \quad a, b \in [k]$$

$$\widehat{w}_i(x, z) = \sum_{1 \le a \le k} \mathbb{1}\{z_i = a\} \frac{n_a(z) g_i(x)}{\sum_{i:z_i=a} g_i(x)}, \quad i \in [n]$$

$$\widehat{\pi}_a(z) \;=\; \frac{n_a(z)}{n}, \quad a \in [k],$$

respectivamente.

3.2 Definição do estimador e teorema da consistência

Fixe $k \ge 1$. Para $z \in [k]^n$ denotamos por $\theta := (\lambda, w, \pi) \in \Theta_k^{(1)} \times \Theta^{(2)}(z) \times \Theta_k^{(3)} =: \Theta_k(z)$ os parâmetros do modelo, em que

$$\Theta_k^{(1)} := \{\lambda \in (\mathbb{R}^+)^{k \times k} : \lambda_{ab} = \lambda_{ba}, a, b \in [k]\}$$

é o conjunto de matrizes simétricas com entradas positivas,

$$\Theta^{(2)}(z) = \Theta_1^{(2)}(z) \times \Theta_2^{(2)}(z) \times \cdots \times \Theta_k^{(2)}(z),$$

com

$$\Theta_a^{(2)}(z) := \{w \in (\mathbb{R}^+)^{n_a(z)} : \sum_i w_i = n_a(z)\}$$

que é o conjunto de possíveis w's na comunidade a. E

$$\Theta_k^{(3)} := \{\pi \in (0,1]^k : \sum_{a=1}^{k} \pi_a = 1\}.$$

Para todo $k \geq 1$, a partir de uma distribuição *a priori* ν_k sobre Θ_k, definimos a distribuição de mistura da mesma forma que foi definida na Seção 2.2, que é dada por:

$$p_k(x) := \sum_{z \in [k]^n} \int_{\Theta_k} p(x, z|\theta)\nu_k(\theta)d\theta \tag{3.4}$$

em que $p(x, z|\theta)$ é a função de verossimilhança conjunta de (X, Z), dada em (3.3). Agora, definimos uma medida produto como *a priori*, $\nu_k = \nu_k^{(1)} \otimes \nu^{(2)} \otimes \nu_k^{(3)}$ sobre $\Theta_k^{(1)} \times \Theta_k^{(2)}(z) \times \Theta_k^{(3)}$, em que sob $\nu_k^{(1)}$, $\nu^{(2)}$ e $\nu_k^{(3)}$

$$\begin{aligned}
\lambda_{a,b} &\sim \text{Gama}(1/2, 1/2)\,, \ a, b \in [k],\ a \leq b \\
(w_i)_{i\in[a]} \,|\, z &\sim n_a(z)\ \text{Dirichlet}(\underbrace{\frac{1}{2}, \ldots, \frac{1}{2}}_{n_a(z)})\,, \quad a \in [k] \\
\pi &\sim \text{Dirichlet}\,\underbrace{(1/2, \ldots, 1/2)}_{k}
\end{aligned}$$

respectivamente. Em outras palavras, para $\theta = (\lambda, w, \pi) \in \Theta_k^{(1)} \times \Theta^{(2)}(z) \times \Theta_k^{(3)}$ temos que a distribuição a *priori* é dada por:

$$\nu_k(\theta) := \nu_k^{(1)}(\lambda)\nu^{(2)}(w|z)\nu_k^{(3)}(\pi) \tag{3.5}$$

onde

$$\nu_k^{(1)}(\lambda) = \left[\prod_{1\leq a\leq b\leq k} \frac{1}{\Gamma(1/2)}(1/2)^{1/2}\lambda_{ab}^{-1/2}e^{-\lambda_{ab}/2}\right],$$

$$\nu^{(2)}(w|z) = \left[\prod_{1\leq a\leq k} \frac{\Gamma\left(\frac{n_a(z)}{2}\right)}{n_a(z)\Gamma\left(\frac{1}{2}\right)^{n_a(z)}} \prod_{i:\, z_i=a} \left(\frac{w_i}{n_a(z)}\right)^{-1/2}\right],$$

$$\nu_k^{(3)}(\pi) = \left[\frac{\Gamma\left(k/2\right)}{\Gamma\left(1/2\right)^k} \prod_{a=1}^{k} {\pi_a}^{-1/2}\right].$$

A partir da distribuição de mistura em (3.4) baseada na *a priori* em (3.5), podemos definir o estimador do número de comunidades do modelo estocástico de blocos com correção de grau: dada uma amostra x

$$\widehat{k}_n(x) := \arg\max_k \{\log p_k(x) - (k^3 - 3kn)\log n\}. \tag{3.6}$$

Observamos que o estimador definido em (3.6) tem um termo de penalidade extra, da ordem $n \log n$, comparado com a penalidade do estimador no modelo estocástico de blocos de Poisson (ver (2.13)). Este termo extra é necessário pela adição de n parâmetros w_i, $i \in [n]$ no modelo. A continuação enunciamos o teorema de consistência, que é o resultado principal deste capítulo.

Teorema 3.2.1 (Teorema da consistência)**.** *Seja um modelo estocástico de blocos com correção de grau de ordem k_0 e $\rho_n \geq C\frac{\log n}{n}$, onde C é uma constante suficientemente grande. Então, o estimador $\widehat{k}_n$ definido em* (3.6) *satisfaz*

$$\widehat{k}_n(x) = k_0, \tag{3.7}$$

quase certamente, quando $n \to \infty$.

3.3 Demonstração do teorema da consistência

As provas do Teorema 3.2.1 são similares às provas do Teorema 2.2.1, porém um pouco mais complicadas, por conta dos graus que entram como parâmetros do modelo. De fato existem algumas diferenças na prova da não subestimação, que fizeram preferirmos apresentar os enunciados e provas separadamente. Abaixo enunciamos e provamos a Proposição 3.3.1 que é chave na demonstração do Teorema 3.2.1.

Lembramos a definição do conjunto de observações,

$$\Omega_n := \{x = \{x_{ij}\} \in \mathbb{N}^{n\times n} : \mathrm{x}_{ij} \leq \log n, \ \ \forall i, j \in 1, \ldots, n\}.$$

Proposição 3.3.1. *Para todo $k \geq 1$, $n \geq \max(4, k)$ e para todo $x \in \Omega_n$, temos que*

$$\log\left(\frac{\sup_{\theta\in\Theta^k} p(x|\theta)}{p_k(x)}\right) \leq d(k)\log n + 3n\log n + c(k,n)$$

onde,

$$d(k) = \frac{k(2k+3)-1}{2}$$

$$c(k,n) := \frac{k(k-1)}{4n} + \frac{k(k+1)}{4n^2} + \frac{1}{12n} + \log\frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{k}{2})}.$$

Proof. De forma similar ao caso do modelo estocástico de blocos de Poisson, analisaremos separadamente as razões $p(x|z,\theta)/p_k(x|z)$ e $p(z|\theta)/p_k(z)$ mostrando que são limitadas superiormente, uniformemente em x, z, por funções $c_1(n,k)c_2(n,k)$ e $c_3(n,k)$ respectivamente. Com isso, observamos que

$$\begin{aligned}\log\frac{p(x|\theta)}{p_k(x)} &= \log\frac{\sum_z p(x|\theta,z)p(z|\theta)}{\sum_z p_k(x|z)p_k(z)} \\ &\leq \log\frac{\sum_z p_k(x|z)c_1(n,k)c_2(n,k)p_k(z)c_3(n,k)}{\sum_z p_k(x|z)p_k(z)} \\ &= \log c_1(n,k) + \log c_2(n,k) + \log c_3(n,k). \end{aligned} \tag{3.8}$$

Com efeito, da definição da distribuição de mistura, temos que

$$p_k(x) = \sum_{z\in[k]^n}\int_{\Theta_k} p(x,z|\theta)\nu_k(\theta)d\theta \tag{3.9}$$

Por outro lado, descompondo temos

$$\begin{aligned}\int_{\Theta_k} p(x,z|\theta) &= \sum_{z[k]^n}\int_{\Theta^{(2)}}\int_{\Theta_k^{(1)}} p(x|z,\lambda)\nu_k^{(1)}(\lambda)\nu^{(2)}(w|z)d\lambda dw\int_{\Theta_k^{(3)}} p(z|\pi)\nu_k^{(3)}(\pi)d\pi \\ &= \sum_{z\in[k]^n} p_k(x|z)p_k(z).\end{aligned}$$

Nas contas que seguem, escrevemos os contadores diretamente, omitindo a referência a x e z, para aliviar a notação e as equações que já estão carregadas. Dessa forma, calculamos

$$\begin{aligned}&\int_{\Theta_k^{(1)}}\frac{1}{c(x)}\prod_{1\leq i\leq n} {w_i}^{g_i}\prod_{1\leq a\leq b\leq k}\lambda_{a,b}^{O_{a,b}}e^{-n_{a,b}\lambda_{a,b}}\nu_k^{(1)}(\lambda)d\lambda \\ &= \frac{1}{c(x)2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}}\prod_{1\leq i\leq n}{w_i}^{g_i}\int_{\Theta_k^{(1)}}\prod_{1\leq a\leq b\leq k}\lambda_{a,b}^{\left(O_{a,b}-\frac{1}{2}\right)}e^{-\left(n_{a,b}+\frac{1}{2}\right)\lambda_{a,b}}d\lambda \\ &= \frac{1}{c(x)2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}}\prod_{1\leq a\leq b\leq k}\frac{\Gamma\left(O_{a,b}+\frac{1}{2}\right)}{\left(n_{a,b}+\frac{1}{2}\right)^{\left(O_{a,b}+\frac{1}{2}\right)}}\prod_{1\leq i\leq n}{w_i}^{g_i} \\ &= C\prod_{1\leq i\leq n}{w_i}^{g_i}.\end{aligned} \tag{3.10}$$

De tal forma que,

$$\begin{aligned} p_k(x|z) = C\int_{\Theta^{(2)}} \prod_{1\le i\le n} {w_i}^{g_i}\nu^{(2)}(w|z)dw &= \prod_{1\le a\le k}\frac{\Gamma\left(\frac{n_a}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{n_a}}\int_{\Theta_a^{(2)}}\prod_{i:\ z_i=a} {w_i}^{g_i}\left(\frac{w_i}{n_a}\right)^{-\frac{1}{2}}dw_a \\ &= \prod_{1\le a\le k}\frac{\Gamma\left(\frac{n_a}{2}\right)n_a^{\bar{g}_a-1}}{\Gamma\left(\frac{1}{2}\right)^{n_a}}\int_{\Theta_a^{(2)}}\prod_{i:\ z_i=a}\left(\frac{w_i}{n_a}\right)^{\left(g_i-\frac{1}{2}\right)}dw_a \\ &= C\prod_{1\le a\le k}\frac{n_a^{\bar{g}_a}\Gamma\left(\frac{n_a}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{n_a}}\frac{\prod_{i:\ z_i=a}\Gamma(g_i+\frac{1}{2})}{\Gamma(\bar{g}_a+\frac{n_a}{2})}. \end{aligned} \tag{3.11}$$

Por outro lado, considerando que os estimadores de máxima verossimilhança para $\lambda_{a,b}$ e w_i são $\frac{O_{a,b}}{n_{a,b}}$ e $\frac{n_a g_i}{\sum_{i:z_i=a} g_i}$ respectivamente, temos que

$$p(x|z,\lambda,w)\le \sup_{\theta\in\Theta^k} p(x|z,\lambda,w) = \frac{1}{c(x)}\prod_{1\le a\le b\le k}\left(\frac{O_{a,b}}{n_{a,b}}\right)^{O_{a,b}} e^{-O_{a,b}}\prod_{a\in[k],i\in[n]:z_i=a}\left(\frac{n_a g_i}{\bar{g}_a}\right)^{g_i}. \tag{3.12}$$

Então, combinando (3.11) e (3.12), temos

$$\frac{p(x|z,\lambda,w)}{p_k(x|z)} \le \frac{\prod_{1\le a\le b\le k}\left(\frac{O_{a,b}}{n_{a,b}}\right)^{O_{a,b}} e^{-O_{a,b}}}{\frac{1}{2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}}\prod_{1\le a\le b\le k}\frac{\Gamma\left(O_{a,b}+\frac{1}{2}\right)}{\left(n_{a,b}+\frac{1}{2}\right)^{\left(O_{a,b}+\frac{1}{2}\right)}}}\frac{\prod_{a\in[k],i\in[n]:z_i=a}\left(\frac{n_a g_i}{\bar{g}_a}\right)^{g_i}}{\prod_{1\le a\le k}\frac{n_a^{\bar{g}_a}\Gamma\left(\frac{n_a}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{n_a}}\frac{\prod_{i:\ z_i=a}\Gamma(g_i+\frac{1}{2})}{\Gamma(\bar{g}_a+\frac{n_a}{2})}}$$

Daqui, da mesma forma que o modelo de blocos de Poisson temos que

$$\frac{\prod_{1\le a\le b\le k}\left(\frac{O_{a,b}}{n_{a,b}}\right)^{O_{a,b}} e^{-O_{a,b}}}{\frac{1}{2^{\frac{k(k+1)}{4}}\Gamma(\frac{1}{2})^{\frac{k(k+1)}{2}}}\prod_{1\le a\le b\le k}\frac{\Gamma\left(O_{a,b}+\frac{1}{2}\right)}{\left(n_{a,b}+\frac{1}{2}\right)^{\left(O_{a,b}+\frac{1}{2}\right)}}} \le 2^{\frac{k(k+1)}{4}} n^{\frac{k(k+1)}{2}} e^{\frac{k(k+1)}{8n^2}} e^{\sum_{a\le b}\frac{O_{a,b}}{2n_{a,b}}}.$$

Por outro lado, para $x\in\Omega_n$ temos que $O_{a,b}\le n_{a,b}\log n$ e portanto $\sum_{a\le b}\frac{O_{a,b}}{2n_{a,b}} \le \frac{k(k+1)}{4}\log n$. Dessa forma, podemos pegar

$$\log c_1(n,k) = \frac{k(k+1)}{2}\log n + \frac{k(k+1)}{4}\log n + \frac{k(k+1)}{8n^2}. \tag{3.13}$$

Por outro lado, temos que

$$\frac{\prod_{a\in[k],i\in[n]:z_i=a}\left(\frac{n_a g_i}{\bar{g}_a}\right)^{g_i}}{\prod_{1\le a\le k}\frac{n_a^{\bar{g}_a}\Gamma\left(\frac{n_a}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{n_a}}\frac{\prod_{i:\ z_i=a}\Gamma(g_i+\frac{1}{2})}{\Gamma(\bar{g}_a+\frac{n_a}{2})}} = \prod_{1\le a\le k}\frac{\Gamma\left(\frac{1}{2}\right)^{n_a}\Gamma(\bar{g}_a+\frac{n_a}{2})}{\Gamma\left(\frac{n_a}{2}\right)}\prod_{i:z_i=a}\frac{\left(\frac{g_i}{\bar{g}_a}\right)^{g_i}}{\Gamma(g_i+\frac{1}{2})}. \tag{3.14}$$

Agora, para todo $a \in [k]$, pelo Lema A.1.1 temos que

$$\prod_{i:z_i=a} \frac{\left(\frac{g_i}{\bar{g}_a}\right)^{g_i}}{\Gamma(g_i+\frac{1}{2})} \;\leq\; \frac{1}{\Gamma\left(\bar{g}_a+\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{n_a-1}}. \tag{3.15}$$

Então, substituindo (3.15) em (3.14) segue que

$$\frac{\prod_{a\in[k],i\in[n]:z_i=a}\left(\frac{n_a g_i}{\bar{g}_a}\right)^{g_i}}{\prod_{1\leq a\leq k}\frac{n_a^{\bar{g}_a}\Gamma\left(\frac{n_a}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{n_a}}\frac{\prod_{i:\ z_i=a}\Gamma(g_i+\frac{1}{2})}{\Gamma(\bar{g}_a+\frac{n_a}{2})}} \leq \prod_{1\leq a\leq k}\frac{\Gamma\left(\frac{1}{2}\right)\Gamma(\bar{g}_a+\frac{n_a}{2})}{\Gamma\left(\frac{n_a}{2}\right)\Gamma\left(\bar{g}_a+\frac{1}{2}\right)}.$$

Agora, pelo A.1.2 e o fato de $g_a(x) \leq n^2 \log n$, para todo $a \in [k]$ e $\sum_a n_a(z) = n$, temos que

$$\log\left(\prod_{1\leq a\leq k}\frac{\Gamma\left(\frac{1}{2}\right)\Gamma(\bar{g}_a+\frac{n_a}{2})}{\Gamma\left(\frac{n_a}{2}\right)\Gamma\left(\bar{g}_a+\frac{1}{2}\right)}\right) \leq n\log(n^2\log n). \tag{3.16}$$

Dessa forma, podemos tomar

$$\log c_2(n,k) \leq 3n\log n. \tag{3.17}$$

Agora procuramos um limitante $c_3(n,k)$ para $p(z|\theta)/p_k(z)$. Esta parte é completamente similar ao modelo estocástico de blocos de Poisson (Proposição 2.3.1), isto é

$$\frac{p(z|\pi)}{p_k(z)} \leq \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}. \tag{3.18}$$

Portanto, tomamos

$$\log c_3(n,k) = \log\left(\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(n+\frac{k}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(n+\frac{1}{2}\right)}\right). \tag{3.19}$$

Daqui, majoramos (3.19) usando o Lema A.1.2 e segue que

$$\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(n+\frac{k}{2})}{\Gamma(\frac{k}{2})\Gamma(n+\frac{1}{2})}\right) \leq \frac{k-1}{2}\log n + \frac{k(k-1)}{4n} + \frac{1}{12n} - \log\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})}. \tag{3.20}$$

Voltamos agora a (3.8), usando (3.13), (3.17) e (3.19) para concluir a prova da proposição: uniformemente em x, z temos

$$\log \frac{p(x|\theta)}{p_k(x)} \leq \log c_1(n,k) + \log c_2(n,k) + \log c_3(n,k)$$
$$= \frac{k(k+2)-1}{2}\log n + \frac{k(k+1)}{4}\log n + 3n\log n + c(k,n)$$

em que

$$c(k,n) := \frac{k(k-1)}{4n} + \frac{k(k+1)}{8n^2} + \frac{1}{12n} - \log\frac{\Gamma(\frac{k}{2})}{\Gamma(\frac{1}{2})}.$$

□

3.3.1 Não superestimação

De forma similar que na Seção 2.3.2, o objetivo desta Seção é mostrar que o estimador $\widehat{k}_n$, não superestima a verdadeira ordem k_0, do modelo estocástico de blocos com correção de grau. O resultado principal é a Proposição 3.3.2, enunciada a seguir.

Proposição 3.3.2. *Seja um modelo estocástico de blocos com correção de grau de ordem k_0, com parâmetros $\theta_0 = (\lambda^0, w^0, \pi^0)$. Então, para o estimador da ordem, $\widehat{k}_n$ definido em* (3.6)*, temos que*

$$p(\{\widehat{k}_n(x) > k_0\} \text{ infinitas vezes} \,|\theta_0) = 0.$$

Proof. Provamos esta proposição seguindo o mesmo caminho que na prova da Proposição 2.3.3. Desta forma, precisamos provar a convergência da seguinte série

$$\sum_{k=k_0+1}^{\infty} p(\widehat{k}_n = k|\theta_0) = \sum_{k=k_0+1}^{\infty}\sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}} + \sum_{k=k_0+1}^{\infty}\sum_{x\in\Omega_n^c} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}}. \quad (3.21)$$

Temos que o segundo termo em (3.21), pelo Lema 2.3.2 é somável em n.

Agora, para o primeiro termo em (3.21), definindo

$$D(k,k_0,n) := d(k_0)\log n + 3n\log n + c(k_0,n) + (k_0^3 + 3k_0 n)\log n - (k^3 - 3kn)\log n,$$

temos que

$$\sum_{x\in\Omega_n} p(x|\theta_0)\mathbb{1}_{\{\widehat{k}_n(x)=k\}} \leq \exp\{D(k,k_0,n)\}\sum_{x\in\Omega_n} p_k(x) \leq \exp\{D(k,k_0,n)\}.$$

Portanto, o primeiro termo em (3.21) é majorado por

$$\sum_{k=k_0+1}^{\infty} \exp\{D(k,k_0,n)\}$$

$$= \exp\left\{d(k_0)\log n + c(k_0,n) + 3n\log n + (k_0^3+3k_0n)\log n\right\} \sum_{k=k_0+1}^{\infty} \exp\left\{-(k^3+3kn)\log n\right\}. \tag{3.22}$$

Por outro lado, temos que

$$\sum_{k=k_0+1}^{\infty} \exp\left\{-(k^3+3kn)\log n\right\} \leq \sum_{k=k_0+1}^{\infty} \exp\{-k^3\log n\} \sum_{k=k_0+1}^{\infty} \exp\{-3kn\log n\}. \tag{3.23}$$

Daqui, da mesma forma que em (2.40) temos que o primeiro somatório no lado direito de (3.23) é limitado superiormente por

$$\sum_{k=k_0+1}^{\infty} \exp\left\{-k^3\log n\right\} \leq C_1(n)\left(\frac{1}{n}\right)^{(k_0+1)^3}, \tag{3.24}$$

em que $C_1(n) = \sum_{k\geq 0}\left(\frac{1}{n}\right)^{k^3} = O((\log n)^{-1/3})$. E o segundo somatório no lado direito de (3.23) fica como

$$\begin{aligned}\sum_{k=k_0+1}^{\infty} \exp\{-3kn\log n\} &= \sum_{k=k_0+1}^{\infty} (e^{-3n\log n})^k \\ &= \frac{(e^{-3n\log n})^{(k_0+1)}}{1-e^{-3n\log n}}\end{aligned} \tag{3.25}$$

onde, na segunda igualdade foi usada a convergência de uma série geométrica. Logo, substituindo (3.24) e (3.25) em (3.23) temos que

$$\sum_{k=k_0+1}^{\infty} \exp\{D(k,k_0,n)\} \leq e^{c(k_0,n)} n^{\{d(k_0)+3n+k_0^3+3k_0n-(k_0+1)^3-3n(k_0+1)\}}. \tag{3.26}$$

Assim, (3.26) é somável em n pois para todo k_0, temos que

$$d(k_0)+3n+k_0^3+3k_0n-(k_0+1)^3-3n(k_0+1) = \frac{k_0(2k_0+3)-1}{2} - 3k_0^2 - 3k_0 - 1 < -1.$$

□

3.3.2 Não subestimação

O objetivo desta Seção é mostrar que o estimador $\widehat{k}_n$, não subestima a verdadeira ordem k_0, do modelo estocástico de blocos com correção de grau. O resultado principal é a Proposição 3.3.3, de fato existem algumas diferenças com respeito das provas da Proposição 2.3.4.

Proposição 3.3.3. *Seja um modelo estocástico de blocos com correção de grau de ordem k_0 e $\rho_n \geq C\frac{\log n}{n}$, onde C é uma constante suficientemente grande. Então, para o estimador $\widehat{k}_n$ definido em* (3.6)*, temos que*

$$p(\{\widehat{k}_n(x) < k_0\} \textit{ infinitas vezes} \,|\theta_0) = 0.$$

Proof. Lembramos a definição do estimador da ordem, dado por

$$\widehat{k}_n(x) = \arg\max_k \{\log p_k(x) - (k^3 + 3kn)\log n\}\,, \tag{3.27}$$

Para mostrar que $\widehat{k}_n(x) \geq k_0$, quase certamente quando $n \to \infty$, é suficiente mostrar que para todo $k < k_0$,

$$\log p_{k_0}(x) - (k_0^3 + 3k_0 n)\log n > \log p_k(x) - (k^3 + 3kn)\log n\,, \tag{3.28}$$

quase certamente, quando $n \to \infty$. Mas, se mostrarmos que

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{p_{k_0}(x)}{p_k(x)} > 0 \tag{3.29}$$

pelo fato de

$$\lim_{n\to\infty} \frac{1}{\rho_n n^2}\left(k_0^3 + 3k_0 n - k^3 - 3kn\right)\log n = 0$$

como $\log n/n\rho_n \to 0$, quando $n \to \infty$, então (3.29) implica (3.28). Observemos que para todo $x \in \Omega_n$, pela Proposição 3.3.1 e o fato de $p_k(x) \leq \sup_{\theta\in\Theta_k} p(x|\theta)$, segue que

$$\begin{aligned}
\frac{1}{\rho_n n^2}\log\frac{p_{k_0}(x)}{p_k(x)} &= \frac{1}{\rho_n n^2}\log\frac{p_{k_0}(x)}{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)} + \frac{1}{\rho_n n^2}\log\frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} \\
&\quad + \frac{1}{\rho_n n^2}\log\frac{\sup_{\theta\in\Theta_k} p(x|\theta)}{p_k(x)} \\
&\geq -d(k_0)\frac{\log n}{\rho_n n^2} - \frac{3\log n}{\rho_n n} - \frac{c(k_0,n)}{\rho_n n^2} + \frac{1}{\rho_n n^2}\log\frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)}
\end{aligned}$$

Assim, para mostrar (3.29) basta mostrar que para $k < k_0$

$$\liminf_{n\to\infty} \frac{1}{\rho_n n^2} \log \frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} > 0 \tag{3.30}$$

já que

$$-d(k_0)\frac{\log n}{\rho_n n^2} - \frac{3\log n}{\rho_n n} - \frac{c(k_0,n)}{\rho_n n^2} \to 0$$

quando $n \to \infty$. Primeiro observemos que para todo z temos que

$$\log \sup_{\theta\in\Theta_{k_0}} p(x|\theta) \geq \log \sup_{\theta\in\Theta_{k_0}} p(x,z|\theta)\,. \tag{3.31}$$

Pela expressão em (3.3) e pelo Lema 3.1.1, temos que

$$\begin{aligned}\log \sup_{\theta\in\Theta_{k_0}} p(x,z|\theta) &= U(x) + \frac{1}{2}\sum_{1\leq a\leq b\leq k_0} O_{a,b}(x,z)\log\widehat{\lambda}_{a,b}(x,z) + \sum_{1\leq a\leq k_0} \bar{g}_a(x,z)\log\frac{n_a(z)}{\bar{g}_a(x,z)} \\ &\quad + n\sum_{a=1}^{k_0}\widehat{\pi}_a(z)\log\widehat{\pi}_a(z),\end{aligned} \tag{3.32}$$

com

$$\begin{aligned} L(x) &= \bar{c}(x) - M(x) + \sum_i g_i(x)\log g_i(x)\,; \\ \widehat{\pi}_a(z) &= \frac{n_a(z)}{n}, \qquad 1\leq a\leq k_0; \\ \widehat{\lambda}_{a,b}(x,z) &= \frac{O_{a,b}(x,z)}{n_{a,b}(z)}, \qquad 1\leq a\leq b\leq k_0; \\ \bar{g}_a(x,z) &= \sum_i x_{ij}\mathbb{1}\{z_i = a\} = \sum_b O_{ab}(x,z), \qquad 1\leq a\leq k_0\,. \end{aligned}$$

Por outro lado, para o denominador em (3.30) usamos que

$$\begin{aligned}\log \sup_{\theta\in\Theta_k} p(x|\theta) &\leq \log k^n \sup_{\theta\in\Theta_k} p(x,z^*|\theta) \\ &\leq n\log k + \log\sup_{\theta\in\Theta_k} p(x,z^*|\theta)\,,\end{aligned} \tag{3.33}$$

com z^* definido por (2.42). Analogamente a (3.32) temos que

$$\begin{aligned}\log \sup_{\theta\in\Theta_k} p(x,z^*|\theta) \; &= U(x) + \frac{1}{2}\sum_{1\le a\le b\le k_0} O_{a,b}(x,z^*)\log\widehat{\lambda}_{a,b}(x,z^*) + \sum_{1\le a\le k_0}\bar{g}_a(x,z^*)\log\frac{n_a(z^*)}{\bar{g}_a(x,z^*)} \\ &\quad + n\sum_{a=1}^{k_0}\widehat{\pi}_a(z^*)\log\widehat{\pi}_a(z^*),\end{aligned} \tag{3.34}$$

Logo, o logaritmo em (3.30) pode ser limitado inferiormente pela diferença de (3.31) e (3.33) e usando as expressões (3.32) e (3.34) temos que

$$\begin{aligned}\log\frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} \;\ge\; &\frac{1}{2}\sum_{1\le a\le b\le k_0} O_{a,b}(x,z)\log\widehat{\lambda}_{a,b}(x,z) + \sum_{1\le a\le k_0}\bar{g}_a(x,z)\log\frac{n_a(z)}{\bar{g}_a(x,z)} \\ &+ n\sum_{a=1}^{k_0}\widehat{\pi}_a(z)\log\widehat{\pi}_a(z) - n\log k - \frac{1}{2}\sum_{1\le a\le b\le k_0} O_{a,b}(x,z^*)\log\widehat{\lambda}_{a,b}(x,z^*) \\ &- \sum_{1\le a\le k_0}\bar{g}_a(x,z^*)\log\frac{n_a(z^*)}{\bar{g}_a(x,z^*)} - n\sum_{a=1}^{k}\widehat{\pi}_a(z^*)\log\widehat{\pi}_a(z^*)\,.\end{aligned} \tag{3.35}$$

Observemos que dividindo ambos lados de (3.35) por $\rho_n n^2$, temos primeiramente que

$$\frac{1}{\rho_n n^2}\Big[n\sum_{a=1}^{k_0}\widehat{\pi}_a(z)\log\widehat{\pi}_a(z) - n\log k - n\sum_{a=1}^{k}\widehat{\pi}_a(z^*)\log\widehat{\pi}_a(z^*)\Big] \;\to\; 0$$

já que $\rho_n n\to\infty$ e os termos nas somatórias ficam limitados, quando $n\to\infty$. Então, para provar (3.30) basta mostrar que os termos restantes são limitados inferiormente por zero. Por outro lado, observemos que podemos escrever

$$\begin{aligned}\sum_{1\le a\le k_0}\bar{g}_a(x,z)\log\frac{n_a(z)}{\bar{g}_a(x,z)} \;&=\; \frac{1}{2}\sum_{1\le a\le k_0}\bar{g}_a(x,z)\log\frac{n_a(z)}{\bar{g}_a(x,z)} + \frac{1}{2}\sum_{1\le b\le k_0}\bar{g}_b(x,z)\log\frac{n_b(z)}{\bar{g}_b(x,z)} \\ &=\; \frac{1}{2}\sum_{1\le a,b\le k_0} O_{ab}(x,z)\log\frac{n_a(z)n_b(z)}{\bar{g}_a(x,z)\bar{g}_b(x,z)}\end{aligned} \tag{3.36}$$

e portanto

$$\begin{aligned}\frac{1}{2}\sum_{1\le a,b\le k_0} O_{ab}(x,z)\log\widehat{\lambda}_{ab}(x,z) + \sum_{1\le a\le k_0}\bar{g}_a(x,z)\log\frac{n_a(z)}{\bar{g}_a(x,z)} \\ = \sum_{1\le a,b\le k_0}\tilde{o}_{ab}(x,z)\log\frac{O_{ab}(x,z)}{\bar{g}_a(x,z)\bar{g}_b(x,z)}\end{aligned} \tag{3.37}$$

e similarmente para os termos envolvendo o modelo de ordem k. Então, somando e subtraindo a expressão

$$\sum_{1\le a\le b\le k_0} O_{a,b}(x,z)\log\rho_n \;=\; \sum_{1\le a\le b\le k} O_{a,b}(x,z)\log\rho_n$$

e substituindo (3.37) em (3.35) temos

$$\begin{aligned}\frac{1}{\rho_n n^2}\log\frac{\sup_{\theta\in\Theta_{k_0}} p(x|\theta)}{\sup_{\theta\in\Theta_k} p(x|\theta)} \;\ge\; \frac{1}{2}\Bigg(&\sum_{1\le a,b\le k_0}\frac{O_{ab}(x,z)}{\rho_n n^2}\log\frac{\rho_n n^2 O_{ab}(x,z)}{\bar{g}_a(x,z)\bar{g}_b(x,z)} \\ &- \sum_{1\le a,b\le k}\frac{O_{ab}(x,z^*)}{\rho_n n^2}\log\frac{\rho_n n^2 O_{ab}(x,z^*)}{\bar{g}_a(x,z^*)\bar{g}_b(x,z^*)}\Bigg) + O(\rho_n^{-1}n^{-1})\,.\end{aligned} \tag{3.38}$$

Daqui, mostrar que (3.38) é limitado inferiormente por zero, quase certamente quando $n\to\infty$ é equivalente a mostrar que

$$\begin{aligned}\liminf_{n\to\infty}\;\frac{1}{2}\sum_{1\le a,b\le k_0}\frac{\bar{g}_a(x,z)\bar{g}_b(x,z)}{\rho_n^2 n^4}\varphi\Big(\frac{\rho_n n^2 O_{ab}(x,z)}{\bar{g}_a(x,z)\bar{g}_b(x,z)}\Big) \\ - \frac{1}{2}\sum_{1\le a,b\le k}\frac{\bar{g}_a(x,z^*)\bar{g}_b(x,z^*)}{\rho_n^2 n^4}\varphi\Big(\frac{\rho_n n^2 O_{ab}(x,z^*)}{\bar{g}_a(x,z^*)\bar{g}_b(x,z^*)}\Big) \;>\; 0\,,\end{aligned} \tag{3.39}$$

com $\varphi(u) = u\log u$, $u>0$. Agora pelo Lema A.2.3, temos que quase certamente que

$$\begin{aligned}\frac{\rho_n n^2 O_{ab}(x,z)}{\bar{g}_a(x,z)\bar{g}_b(x,z)} \;=\; \frac{O_{ab}(x,z)/\rho_n n^2}{\bar{g}_a(x,z)\bar{g}_b(x,z)/\rho_n^2 n^4} \;&\to\; \frac{[\mathrm{diag}(\pi)\tilde{\lambda}\mathrm{diag}(\pi)^T]_{ab}}{[\mathrm{diag}(\pi)\tilde{\lambda}\mathrm{diag}(\pi)^T\mathbf{1}_k]_a[\mathrm{diag}(\pi)\tilde{\lambda}\mathrm{diag}(\pi)^T\mathbf{1}_k]_b} \\ &= \frac{\pi_a^0\pi_b^0\tilde{\lambda}_{ab}^0}{\pi_a^0[\tilde{\lambda}^0\pi^0]_b\pi_b^0[\tilde{\lambda}^0\pi^0]_a} \\ &= \frac{\tilde{\lambda}_{ab}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b}\,.\end{aligned} \tag{3.40}$$

Lembramos que $\tilde{\lambda}^0$ é a matriz tal que $\lambda^0 = \rho_n \tilde{\lambda}^0$. Logo, resulta que

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{2} \sum_{1\le a,b\le k_0} \frac{\bar{g}_a(x,z)\bar{g}_b(x,z)}{\rho_n^2 n^4} \varphi\Big(\frac{\rho_n n^2 O_{ab}(x,z)}{\bar{g}_a(x,z)\bar{g}_b(x,z)}\Big) \\ = \frac{1}{2} \sum_{1\le a,b\le k_0} \pi_a^0\pi_b^0[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b \varphi\Big(\frac{\tilde{\lambda}^0_{ab}}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b}\Big). \end{aligned} \tag{3.41}$$

Por outro lado, pelo Lema 3.3.4 temos que

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{2} \sum_{1\le a,b\le k} \frac{\bar{g}_a(x,z_k^*)\bar{g}_b(x,z_k^*)}{\rho_n^2 n^4} \varphi\Big(\frac{\rho_n n^2 O_{ab}(x,z_k^*)}{\bar{g}_a(x,z_k^*)\bar{g}_b(x,z_k^*)}\Big) \\ \le \frac{1}{2} \sum_{1\le a,b\le k} \pi_a^*\pi_b^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b \varphi\Big(\frac{\lambda^*_{ab}}{[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b}\Big) \end{aligned} \tag{3.42}$$

para alguma matriz λ^* de ordem $k \times k$ e um vetor π^* k-dimensional, dado em (3.45). Finalmente, pelo Lema 3.3.5 temos que da diferença de (3.41) e (3.42) resulta que

$$\begin{aligned} \sum_{1\le a,b\le k_0} \pi_a^0\pi_b^0[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b \varphi\Big(\frac{\tilde{\lambda}^0_{ab}}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b}\Big) \\ - \sum_{1\le a,b\le k} \pi_a^*\pi_b^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b \varphi\Big(\frac{\lambda^*_{ab}}{[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b}\Big) > 0. \end{aligned} \tag{3.43}$$

□

Lema 3.3.4. *Para $k < k_0$, temos que existe uma matriz positiva λ^*, $k \times k$ e um vetor π^*, k-dimensional tal que*

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{2} \sum_{1\le a,b\le k} \frac{\bar{g}_a(x,z_k^*)\bar{g}_b(x,z_k^*)}{\rho_n^2 n^4} \varphi\Big(\frac{\rho_n n^2 O_{ab}(x,z_k^*)}{\bar{g}_a(x,z_k^*)\bar{g}_b(x,z_k^*)}\Big) \\ \le \frac{1}{2} \sum_{1\le a,b\le k} \pi_a^*\pi_b^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b \varphi\Big(\frac{\lambda^*_{ab}}{[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b}\Big). \end{aligned} \tag{3.44}$$

em que, (λ^*, π^*) são dados por

$$\begin{aligned} \lambda^*_{ab} &= \frac{[I^*\tilde{\lambda}^0 I^{*T}]_{ab}}{[I^*\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T I^{*T}]_{ab}}, \qquad a,b \in \{1,\dots,k\} \\ \pi^*_a &= [I^*\mathbf{1}_{k_0}]_a, \qquad a \in \{1,\dots,k\} \end{aligned} \tag{3.45}$$

para uma matriz I^*, $k \times k_0$ que satisfaz $\|I^*\|_1 = 1$ e tem uma entrada diferente de zero em cada coluna.

Proof. Definamos para todo $z \in \{1, \ldots, k\}^n$ e $z^0 \in \{1, \ldots, k_0\}^n$, $H_n(z, z^0)$ como a matriz $k \times k_0$ com entradas dadas por

$$[H_n(z, z^0)]_{a,a'} \;=\; \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1}\{z_i = a, z_i^0 = a'\}. \tag{3.46}$$

Observemos que, os contadores $n_a(z^0)$, para $a \in [k_0]$ podem ser escritos como

$$n_{a'}(z^0) = \sum_{i=1}^{n} \sum_{a=1}^{k} w_i \mathbb{1}\{z_i = a, z_i^0 = a'\}\,. \tag{3.47}$$

De tal forma que, $n'_a(z^0) = n(H_n^T(z, z^0)\mathbf{1}_k)_a$, em que $\mathbf{1}_k$ é um vetor coluna de dimensão k com todas as entradas iguais a 1. E os contadores $n_a(z)$, para $a \in [k]$

$$n_a(z) = \sum_{i=1}^{n} \sum_{a'=1}^{k_0} w_i \mathbb{1}\{z_i = a, z_i^0 = a'\}\,. \tag{3.48}$$

Tal que, $n_a(z) = n(H_n(z, z^0)\mathbf{1}_{k_0})_a$, de forma similar que antes, $\mathbf{1}_{k_0}$ é um vetor coluna de dimensão k_0 com todas as entradas iguais a 1. Além disso, a matriz $H_n(z, z^0)$ satisfaz

$$\|H_n(z, z^0)\|_1 \;=\; \sum_{a=1}^{k} \sum_{a'=1}^{k_0} [H_n(z, z^0)]_{a,a'} = 1, \tag{3.49}$$

para todo (z, z^0). Por outro lado, temos que

$$\begin{aligned} &\sum_{1 \le a,b \le k} \frac{\bar{g}_a(x, z_k^*)\bar{g}_b(x, z_k^*)}{\rho_n^2 n^4} \varphi\Big(\frac{\rho_n n^2 O_{ab}(x, z_k^*)}{\rho_n\, \bar{g}_a(x, z_k^*)\bar{g}_b(x, z_k^*)}\Big) \\ &\qquad = \sum_{1 \le a,b \le k} \frac{\bar{g}_a(x, z_k^*)\bar{g}_b(x, z_k^*)}{\rho_n^2 n^4} \varphi\left(\frac{O_{ab}(x, z_k^*)/\rho_n n^2}{\bar{g}_a(x, z_k^*)\bar{g}_b(x, z_k^*)/\rho_n^2 n^4}\right), \end{aligned} \tag{3.50}$$

em que $\varphi(u) = u \log u$. Então, pelo Lema A.2.3 tomando uma sequência $\epsilon_n = \rho_n = \frac{\log n}{n}$, temos quase certamente que

$$\begin{aligned} &\left|\frac{O_{a,b}(x, z^*)}{\rho_n n^2} - [H_n(z, z^0)\tilde{\lambda} H_n(z, z^0)^T]_{a,b}\right| \le \epsilon_n, \\ &\left|\frac{\bar{g}_a(x, z^*)}{\rho_n n^2} - [H_n(z, z^0)\tilde{\lambda} H_n(z, z^0)^T \mathbf{1}_k]_a\right| \le \epsilon_n \end{aligned} \tag{3.51}$$

Daqui, como φ é continua, substituindo (3.51) no lado direito de (3.50) e tomando limsup a ambos lados, segue que

$$\limsup_{n\to\infty} \sum_{1\le a,b\le k} \frac{\bar{g}_a(x,z_k^*)\bar{g}_b(x,z_k^*)}{\rho_n^2 n^4}\,\varphi\Big(\frac{\rho_n n^2 O_{ab}(x,z_k^*)}{\bar{g}_a(x,z_k^*)\bar{g}_b(x,z_k^*)}\Big) \le \sup_{\substack{I\colon \|I\|_1=1 \\ I^T\mathbf{1}_k=\pi}} \frac{1}{2}\sum_{1\le a,b\le k} [I\tilde{\lambda}^0 I^T\mathbf{1}_k]_a[I\tilde{\lambda}^0 I^T\mathbf{1}_k]_b\,\varphi\bigg(\frac{[I\tilde{\lambda}^0 I^T]_{ab}}{[I\tilde{\lambda}^0 I^T\mathbf{1}_k]_a[I\tilde{\lambda}^0 I^T\mathbf{1}_k]_b}\bigg), \tag{3.52}$$

quase certamente. Logo, o supremo em (3.52) corresponde a um problema de maximizar uma função convexa sobre um poliedro convexo definido por $\{I\colon \|I\|_1 = 1, I^T\mathbf{1}_k = \pi^0\}$. Portanto, pelo Proposição A.2.1 temos que o máximo deve ser atingido em algum dos vértices do poliedro; isto é, naquelas matrizes I tais que uma e somente uma entrada por coluna é maior do que zero, considerando que $\pi_a^0 > 0$ para todo $a \in \{1,\dots,k_0\}$. Denotemos por I^* a um destes máximos e sejam

$$\begin{aligned} \pi_a^* &= [I^*\mathbf{1}_{k_0}]_a, \qquad a\in\{1,\dots,k\} \\ \lambda_{ab}^* &= \frac{[I^*\tilde{\lambda}^0 I^{*^T}]_{ab}}{[I^*\mathbf{1}_{k_0}\mathbf{1}_{k_0}^T I^{*^T}]_{ab}}\,, \qquad a,b\in\{1,\dots,k\}\,. \end{aligned} \tag{3.53}$$

Logo, temos que

$$\begin{aligned} \sup_{\substack{I\colon \|I\|_1=1 \\ I^T\mathbf{1}_k=\pi}} \frac{1}{2}\sum_{1\le a,b\le k} [I\tilde{\lambda}^0 I^T\mathbf{1}_k]_a[I\tilde{\lambda}^0 I^T\mathbf{1}_k]_b\,\varphi\bigg(\frac{[I\tilde{\lambda}^0 I^T]_{ab}}{[I\tilde{\lambda}^0 I^T\mathbf{1}_k]_a[I\tilde{\lambda}^0 I^T\mathbf{1}_k]_b}\bigg) \\ = \frac{1}{2}\sum_{1\le a,b\le k} \pi_a^*\pi_b^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b\,\varphi\bigg(\frac{\lambda_{ab}^*}{[\lambda^*\pi^*]_a[\lambda^*\pi^{*]}]_b}\bigg). \end{aligned} \tag{3.54}$$

Isso conclui a prova do Lema 3.3.4. □

Lema 3.3.5. *Para $k < k_0$ e (π^*,λ^*) como no Lema 3.3.4, temos que*

$$\begin{aligned} \sum_{1\le a,b\le k_0} \pi_a^0\pi_b^0[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b\,\varphi\Big(\frac{\tilde{\lambda}_{ab}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b}\Big) \\ - \sum_{1\le a,b\le k} \pi_a^*\pi_b^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b\,\varphi\bigg(\frac{\lambda_{ab}^*}{[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b}\bigg) \;>\; 0\,. \end{aligned} \tag{3.55}$$

Proof. Observemos que de forma similar que na prova do Lema 2.3.6, os parâmetros π^* e λ^* são dados por:

$$\begin{aligned}\pi_i^* &= \pi_i^0\,, \qquad 1 \le i \le k_0 - 2\\ \pi_{k_0-1}^* &= \pi_{k_0-1}^0 + \pi_{k_0}^0\end{aligned}$$

e

$$\begin{aligned}\lambda_{l,r}^* &= \tilde{\lambda}_{l,r}^0\,, \qquad 1 \le l < r \le k_0 - 2,\\ \lambda_{l,k_0-1}^* &= \frac{\pi_l^0\pi_{k_0-1}^0\tilde{\lambda}_{l,k_0-1}^0 + \pi_l^0\pi_{k_0}^0\tilde{\lambda}_{l,k_0}^0}{\pi_l^0\pi_{k_0-1}^0 + \pi_l^0\pi_{k_0}^0}\,, \qquad 1 \le l \le k_0 - 2,\\ \lambda_{k_0-1,k_0-1}^* &= \frac{\pi_{k_0-1}^0\pi_{k_0-1}^0\tilde{\lambda}_{k_0-1,k_0-1}^0 + 2\pi_{k_0-1}^0\pi_{k_0}^0\tilde{\lambda}_{k_0-1,k_0}^0 + \pi_{k_0}^0\pi_{k_0}^0\tilde{\lambda}_{k_0,k_0}^0}{\pi_{k_0-1}^0\pi_{k_0-1}^0 + 2\pi_{k_0-1}^0\pi_{k_0}^0 + \pi_{k_0}^0\pi_{k_0}^0}.\end{aligned}$$

Agora, observemos que para $1 \le a \le k_0 - 2$ temos que $[\lambda^*\pi^*]_a = [\tilde{\lambda}^0\pi^0]_a$, então para $1 \le a, b \le k_0 - 2$ segue que

$$\pi_a^*\pi_b^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b\,\varphi\!\left(\frac{\lambda_{a,b}^*}{[\lambda^*\pi^*]_a[\lambda^*\pi^*]_b}\right) \;=\; \pi_a^0\pi_b^0[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b\,\varphi\!\left(\frac{\tilde{\lambda}_{a,b}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_b}\right).$$

Por outro lado, temos que

$$[\lambda^*\pi^*]_{k_0-1} = \frac{\pi_{k_0-1}^0[\tilde{\lambda}^0\pi^0]_{k_0-1} + \pi_{k_0}^0[\tilde{\lambda}^0\pi^0]_{k_0}}{\pi_{k_0-1}^0 + \pi_{k_0}^0}\,.$$

Dado que φ é uma função convexa, então para $1 \le a \le k_0 - 2$, pela desigualdade de Jensen segue que

$$\begin{aligned}&\pi_a^*\pi_{k_0-1}^*[\lambda^*\pi^*]_a[\lambda^*\pi^*]_{k_0-1}\varphi\!\left(\frac{\lambda_{a,(k_0-1)}^*}{[\lambda^*\pi^*]_a[\lambda^*\pi^*]_{k_0-1}}\right)\\ &\quad= \pi_a^0[\tilde{\lambda}^0\pi^0]_a(\pi_{k_0-1}^0[\tilde{\lambda}^0\pi^0]_{k_0-1} + \pi_{k_0}^0[\tilde{\lambda}^0\pi^0]_{k_0})\varphi\left(\frac{\pi_{k_0-1}^0\tilde{\lambda}_{a,(k_0-1)}^0 + \pi_{k_0}^0\tilde{\lambda}_{a,k_0}^0}{[\tilde{\lambda}^0\pi^0]_a(\pi_{k_0-1}^0[\tilde{\lambda}^0\pi^0]_{k_0-1} + \pi_{k_0}^0[\tilde{\lambda}^0\pi^0]_{k_0})}\right)\\ &\quad= \pi_a^0(\pi_{k_0-1}^0\tilde{\lambda}_{a,(k_0-1)}^0 + \pi_{k_0}^0\tilde{\lambda}_{a,k_0}^0)\log\left(\frac{\pi_a^0(\pi_{k_0-1}^0\tilde{\lambda}_{a,(k_0-1)}^0 + \pi_{k_0}^0\tilde{\lambda}_{a,k_0}^0)}{\pi_a^0[\tilde{\lambda}^0\pi^0]_a(\pi_{k_0-1}^0[\tilde{\lambda}^0\pi^0]_{k_0-1} + \pi_{k_0}^0[\tilde{\lambda}^0\pi^0]_{k_0})}\right)\\ &\quad\le \pi_a^0\pi_{k_0-1}^0\tilde{\lambda}_{a,(k_0-1)}^0\log\left(\frac{\tilde{\lambda}_{a,(k_0-1)}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0-1}}\right) + \pi_a^0\pi_{k_0}^0\tilde{\lambda}_{a,k_0}^0\log\left(\frac{\tilde{\lambda}_{a,k_0}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0}}\right)\\ &\quad= \pi_a^0\pi_{k_0-1}^0[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0-1}\varphi\!\left(\frac{\tilde{\lambda}_{a,(k_0-1)}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0-1}}\right) + \pi_a^0\pi_{k_0}^0[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0}\varphi\!\left(\frac{\tilde{\lambda}_{a,k_0}^0}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0}}\right).\end{aligned} \tag{3.56}$$

Ainda, a desigualdade acima deve ser estrita, a não ser que

$$\frac{\tilde{\lambda}^0_{a,(k_0-1)}}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0-1}} = \frac{\tilde{\lambda}^0_{a,k_0}}{[\tilde{\lambda}^0\pi^0]_a[\tilde{\lambda}^0\pi^0]_{k_0}}, \qquad a \leq k_0 - 2\,. \tag{3.57}$$

Agora, para $a = k_0 - 1$ e $b = k_0$, também pela convexidade de φ e a desigualdade de Jensen temos que

$$\begin{aligned}
&\pi^*_{k_0-1}\pi^*_{k_0-1}[\lambda^*\pi^*]_{k_0-1}[\lambda^*\pi^*]_{k_0-1}\varphi\left(\frac{\lambda^*_{(k_0-1),(k_0-1)}}{[\lambda^*\pi^*]_{k_0-1}[\lambda^*\pi^*]_{k_0-1}}\right)\\
&= (\pi^0_{k_0-1}\pi^0_{k_0-1}\tilde{\lambda}^0_{(k_0-1),(k_0-1)} + 2\pi^0_{k_0-1}\pi^0_{k_0}\tilde{\lambda}^0_{(k_0-1),k_0} + \pi^0_{k_0}\pi^0_{k_0}\tilde{\lambda}^0_{k_0,k_0})\\
&\qquad \times \log\left(\frac{\pi^0_{k_0-1}\pi^0_{k_0-1}\tilde{\lambda}^0_{(k_0-1),(k_0-1)} + 2\pi^0_{k_0-1}\pi^0_{k_0}\tilde{\lambda}^0_{(k_0-1),k_0} + \pi^0_{k_0}\pi^0_{k_0}\tilde{\lambda}^0_{k_0,k_0}}{\pi^0_{k_0-1}\pi^0_{k_0-1}[\tilde{\lambda}^0\pi^0]^2_{k_0-1} + 2\pi^0_{k_0-1}\pi^0_{k_0}[\tilde{\lambda}^0\pi^0]_{k_0-1}[\tilde{\lambda}^0\pi^0]_{k_0} + \pi^0_{k_0}\pi^0_{k_0}[\tilde{\lambda}^0\pi^0]^2_{k_0}}\right)\\
&\leq\ \pi^0_{k_0-1}\pi^0_{k_0-1}\tilde{\lambda}^0_{(k_0-1),(k_0-1)}\log\left(\frac{\tilde{\lambda}^0_{(k_0-1),(k_0-1)}}{[\tilde{\lambda}^0\pi^0]^2_{k_0-1}}\right) + 2\pi^0_{k_0-1}\pi^0_{k_0}\tilde{\lambda}^0_{(k_0-1),k_0}\log\left(\frac{\tilde{\lambda}^0_{(k_0-1),k_0}}{[\tilde{\lambda}^0\pi^0]_{k_0-1}[\tilde{\lambda}^0\pi^0]_{k_0}}\right)\\
&\qquad + \pi^0_{k_0}\pi^0_{k_0}\tilde{\lambda}^0_{k_0,k_0}\log\left(\frac{\tilde{\lambda}^0_{k_0,k_0}}{[\tilde{\lambda}^0\pi^0]^2_{k_0}}\right)\\
&= \pi^0_{k_0-1}\pi^0_{k_0-1}[\tilde{\lambda}^0\pi^0]^2_{k_0-1}\varphi\left(\frac{\tilde{\lambda}^0_{(k_0-1),(k_0-1)}}{[\tilde{\lambda}^0\pi^0]^2_{k_0-1}}\right) + 2\pi^0_{k_0-1}\pi^0_{k_0}[\tilde{\lambda}^0\pi^0]_{k_0-1}[\tilde{\lambda}^0\pi^0]_{k_0}\varphi\left(\frac{\tilde{\lambda}^0_{(k_0-1),k_0}}{[\tilde{\lambda}^0\pi^0]_{k_0-1}[\tilde{\lambda}^0\pi^0]_{k_0}}\right)\\
&\qquad + \pi^0_{k_0}\pi^0_{k_0}[\tilde{\lambda}^0\pi^0]^2_{k_0}\varphi\left(\frac{\tilde{\lambda}_{k_0,k_0}}{[\tilde{\lambda}^0\pi^0]^2_{k_0}}\right),
\end{aligned} \tag{3.58}$$

Daqui, a desigualdade acima deve ser estrita, a menos que

$$\frac{\tilde{\lambda}^0_{(k_0-1),(k_0-1)}}{[\tilde{\lambda}^0\pi^0]^2_{k_0-1}} = \frac{\tilde{\lambda}^0_{(k_0-1),k_0}}{[\tilde{\lambda}^0\pi^0]_{k_0-1}[\tilde{\lambda}^0\pi^0]_{k_0}} = \frac{\tilde{\lambda}^0_{k_0,k_0}}{[\tilde{\lambda}^0\pi^0]^2_{k_0}}\,. \tag{3.59}$$

Logo, das igualdades em (3.57) e (3.59) temos que a desigualdade em (3.55) deve ser estrita, a menos que

$$\tilde{\lambda}^0_{a,(k_0-1)} = \frac{[\tilde{\lambda}^0\pi^0]_{k_0-1}}{[\tilde{\lambda}^0\pi^0]_{k_0}}\tilde{\lambda}_{a,k_0}, \qquad a \leq k_0. \tag{3.60}$$

O que contradiz a Suposição 3.1.1. □

Chapter

4

Considerações Finais

Nesta tese, provamos consistência forte para o estimador do número de comunidades no modelo estocástico de blocos de Poisson, em que as entradas da matriz de adjacência seguem uma distribuição de Poisson e no modelo estocástico de blocos com correção de grau, que permite que a distribuição dos graus dos vértices dependa também dos vértices e não apenas das comunidades às quais pertencem.

Consideramos o regime denso, em que a probalidade de conexão entre pares de vértices não depende do tamanho do grafo e o regime semi-esparso, em que a probabilidade de conexão entre pares de vértices pode decrescer para zero (numa certa taxa), em função do tamanho do grafo.

O nosso resultado estende em particular o trabalho de Wang et al. (2017), que provaram consistência fraca para o modelo estocástico de blocos, com entradas da matriz de adjacência seguindo uma distribuição de Bernoulli e o trabalho de Cerqueira e Leonardi (2020), que provaram consistência forte e consideraram também um regime semi-esparso, porém no modelo estocástico de blocos com entradas da matriz de adjacência seguindo uma distribuição de Bernoulli.

Appendix

A

Demonstração de resultados auxiliares

A.1 Resultados básicos

Lema A.1.1. *Para $m \geq 0$ e inteiros m_j, $j = 1, \ldots, J$ tal que $m = \sum_{j=1}^{J} m_j$, temos que*

$$\frac{\prod_{j=1}^{J} \left(\frac{m_j}{m}\right)^{m_j}}{\prod_{j=1}^{J} \Gamma(m_j + \frac{1}{2})} \leq \frac{1}{\Gamma\left(m + \frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right)^{J-1}}.$$

Proof. A prova deste Lema segue por Davisson et al. (1981). Seja a igualdade

$$\frac{\prod_{j=1}^{J} \Gamma\left(m_j + \frac{1}{2}\right)}{\Gamma\left(m + \frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right)^{J-1}} = \frac{\prod_{j=1}^{J} (m_j + 1)(m_j + 2) \ldots (2m_j)}{(m+1)(m+2)\ldots(2m)}. \tag{A.1}$$

Com efeito, usando as propriedades da função Gama: $\Gamma(m+\frac{1}{2}) = \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2m-1)}{2^m}\sqrt{\pi}$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ e a igualdade

$$1.3.5. \cdots .(2m-1) = \frac{1.2.3.4.5.6. \cdots .(2m-1).2m}{2.4.6. \cdots .2m} = \frac{(2m)!}{2^m m!},$$

temos que

$$\frac{\prod_{j=1}^{J} \Gamma\left(m_j + \frac{1}{2}\right)}{\Gamma\left(m + \frac{1}{2}\right) \Gamma\left(\frac{1}{2}\right)^{J-1}} = \frac{\frac{(2m_1)!}{2^{m_1} m_1!} \frac{\sqrt{\pi}}{2^{m_1}} \frac{(2m_2)!}{2^{m_2} m_2!} \frac{\sqrt{\pi}}{2^{m_2}} \cdots \frac{(2m_J)!}{2^{m_J} m_J!} \frac{\sqrt{\pi}}{2^{m_J}}}{\frac{(2m)!}{2^m m!} \frac{\sqrt{\pi}}{2^m} (\sqrt{\pi})^{J-1}} = \frac{\frac{(2m_1)!}{m_1!} \frac{(2m_2)!}{m_2!} \cdots \frac{(2m_J)!}{m_J!}}{\frac{(2m)!}{m!}}. \tag{A.2}$$

Por outro lado,

$$\frac{\frac{(2m_1)!}{m_1!}\frac{(2m_2)!}{m_2!}\cdots\frac{(2m_J)!}{m_J!}}{\frac{(2m)!}{m!}} = \frac{\prod_{j=1}^{J}(2mj)(2m_j-1)\cdots(m_j+2)(m_j+1)}{(2m)(2m-1)\cdots(m+2)(m+1)}. \tag{A.3}$$

substituindo (A.3) em (A.2), obtemos a igualdade (A.1). Por outro lado, para $k \geq 0$ e $x \geq 0$ definimos

$$g_k(x) = \prod_{i=1}^{k}\left(x + \frac{i}{k}\right).$$

Agora iremos provar que, se $m = \sum_{j=1}^{J} m_j$ e $x \geq 0$

$$g_m(x) \leq \prod_{j=1}^{J} g_{m_j}(x). \tag{A.4}$$

Seguindo a prova feita em Davisson et al. (1981) mostraremos (A.4) para o caso $J = 2$, pois o caso geral segue direitamente usando indução. Então, dessa forma precisamos mostrar que para inteiros m_1, m_2 e $x \geq 0$,

$$\prod_{k=1}^{m_1+m_2}\left(x + \frac{k}{m_1+m_2}\right) \leq \prod_{l=1}^{m_1}\left(x + \frac{l}{m_1}\right)\prod_{r=1}^{m_2}\left(x + \frac{r}{m_2}\right). \tag{A.5}$$

Uma vez que em ambos os lados de (A.5) existem m_1+m_2 termos multiplicados, mostraremos que existe uma correspondência biunívoca entre esses termos, de tal forma que cada termo no lado esquerdo seja menor ou igual do que o termo que lhe corresponde no lado direito. Agora, denotemos por S_k o conjunto de termos do lado direito, que são maiores ou iguais do que $\left(x + \frac{k}{m_1+m_2}\right)$, que é o termo que aparece no lado esquerdo de (A.5). Com isso, daqui em frente o nosso objetivo será mostrar que existe uma ordem $(t_1, \ldots, t_{m_1+m_2})$ dos $m_1 + m_2$ termos do lado direito de (A.5), tal que $t_k \in S_k$, para $k = 1, \ldots, m_1 + m_2$, podemos ver que

$$\begin{aligned}
\left(x + \frac{k}{m_1+m_2}\right) \leq \left(x + \frac{l}{m_1}\right) &\iff l \geq \frac{m_1 k}{m_1+m_2} \\
\left(x + \frac{k}{m_1+m_2}\right) \leq \left(x + \frac{r}{m_2}\right) &\iff r \geq \frac{m_2 k}{m_1+m_2}.
\end{aligned}$$

Portanto, temos que o número total de termos em S_k é dado por

$$|S_k| = \left(m_1 - \left\lceil\frac{m_1 k}{m_1+m_2}\right\rceil + 1\right) + \left(m_2 - \left\lceil\frac{m_2 k}{m_1+m_2}\right\rceil + 1\right).$$

Usando a desigualdade $\lceil x \rceil + \lceil y \rceil - 1 \leq \lceil x + y \rceil$ obtemos que,

$$|S_k| \geq m_1 + m_2 + 1 - k. \tag{A.6}$$

Com a desigualdade (A.6) podemos obter a ordem desejada, pois tomando um $t_{m_1+m_2} \in S_{m_1+m_2}$, podemos tomar o $t_i \in S_i$ já que no máximo foram tomados $m_1 + m_2 - i$ termos em S_i. Com isso, mostramos (A.4), para $J = 2$. Dessa forma, usando (A.4) para $x = 1$ temos

$$\begin{aligned}\prod_{j=1}^{J} \left(\frac{m_j}{m}\right)^{m_j} &\leq \frac{\prod_{i=1}^{m_1}(m_1+i)\prod_{i=1}^{m_2}(m_2+i)\cdots\prod_{i=1}^{m_J}(m_J+i)}{\prod_{i=1}^{m}(m+i)} \\ &= \frac{\prod_{j=1}^{J}(m_j+1)(m_j+2)\cdots(2m_j)}{(m+1)(m+2)\cdots(2m)}.\end{aligned} \tag{A.7}$$

Logo, juntando (A.1) e (A.7) temos que o Lema A.1.1 fica provado. □

Lema A.1.2. *Para todo J e para todo $m \geq \max\{4, J\}$ temos,*

$$\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{J}{2})}{\Gamma(\frac{J}{2})\Gamma(m+\frac{1}{2})}\right) \leq \frac{J-1}{2}\log m + \frac{J(J-1)}{4m} + \frac{1}{12m} - \log\frac{\Gamma(\frac{J}{2})}{\Gamma(\frac{1}{2})}. \tag{A.8}$$

Proof. A prova deste Lema é dada em Gassiat e Boucheron (2003). Usando os limitantes sobre a função Γ, que estendem a aproximação de Robbins- Stirling,

$$x^{x-\frac{1}{2}}e^{-x}\sqrt{2\pi} \leq \Gamma(x) \leq x^{x-\frac{1}{2}}e^{-x}\sqrt{2\pi}e^{\frac{1}{12x}},$$

temos que,

$$\begin{aligned}\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{J}{2})}{\Gamma(\frac{J}{2})\Gamma(m+\frac{1}{2})}\right) \leq\ & m\log\left(m+\frac{J}{2}\right) + \frac{(J-1)}{2}\log\left(m+\frac{J}{2}\right) + \frac{1}{12(m+J/2)} - \frac{(J-1)}{2} \\ & - m\log\left(m+\frac{1}{2}\right) - \log\frac{\Gamma(\frac{J}{2})}{\Gamma(\frac{1}{2})}.\end{aligned}$$

Daqui, fazendo algumas contas temos

$$\begin{aligned}\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{J}{2})}{\Gamma(\frac{J}{2})\Gamma(m+\frac{1}{2})}\right) \leq\ & \frac{(J-1)}{2}\log m + m\log\left(\frac{2m+J}{2m+1}\right) + \frac{(J-1)}{2}\log\left(1+\frac{J}{2m}\right) + \frac{1}{12m} \\ & - \frac{(J-1)}{2} - \log\frac{\Gamma(\frac{J}{2})}{\Gamma(\frac{1}{2})}.\end{aligned}$$

Agora, usando a desigualdade $\log x \leq x - 1$, para $x > 0$, temos que

$$\begin{aligned}\log\left(\frac{\Gamma(\frac{1}{2})\Gamma(m+\frac{J}{2})}{\Gamma(\frac{J}{2})\Gamma(m+\frac{1}{2})}\right) &\leq \frac{(J-1)}{2}\log m + \frac{m(J-1)}{(2m+1)} + \frac{J(J-1)}{4m} + \frac{1}{12m} - \frac{(J-1)}{2} - \log\frac{\Gamma(\frac{J}{2})}{\Gamma(\frac{1}{2})}. \\ &= \frac{(J-1)}{2}\log m + \frac{J(J-1)}{4m} + \frac{1}{12m} - \frac{(J-1)}{4m+2} - \log\frac{\Gamma(\frac{J}{2})}{\Gamma(\frac{1}{2})}\end{aligned}$$

□

A.2 Outros resultados auxiliares

Definição A.2.1. *Dado um conjunto convexo não vazio C, diz-se que um vetor $x \in C$ é um ponto extremal de C se não existem vetores $y \in C$ e $z \in C$, com $y \neq x$ e $z \neq x$ e um escalar $\alpha \in (0,1)$ tais que $x = \alpha y + (1-\alpha)z$.*

Proposição A.2.1. *Seja C um subconjunto convexo e fechado de $\mathbb{R}^n$ que tem como mínimo um ponto extremo. Uma função côncava $f\colon C \to \mathbb{R}$ que atinge um mínimo sobre C atinge seu mínimo em algum ponto extremo de C.*

Proof. Ver Bertsekas (2009, Proposition 2.4.1, pg. 112). □

Lema A.2.2. *Para qualquer $\epsilon > 0$ e $a, b \in [k]$ temos que*

$$\mathbb{P}\left(\sup_{z\in[k]^n}\left|\frac{O_{a,b}(X,z)}{\rho_n n^2} - [Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}\right| > \epsilon\right) \leq e^{\left(-\frac{\rho_n n^2\epsilon^2}{2(\lambda_{max}+\epsilon)} + n\log k\right)}.$$

Proof. Para quaisquer $z \in [k]^n$ e $z^0 \in [k_0]^n$, temos que

$$\begin{aligned}&\left|O_{a,b}(X,z) - n^2[Q_n(z,z^0)\lambda Q_n(z,z^0)^T]_{a,b}\right| \\ &= \left|O_{a,b}(X,z) - \rho_n n^2[Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}\right| \\ &= \left|\sum_{1\leq i,j\leq n}\sum_{1\leq a',b'\leq k_0}(X_{ij} - \rho_n\tilde{\lambda}_{a',b'})\mathbb{1}\{z_i = a, z_i^0 = a'\}\mathbb{1}\{z_j = b, z_j^0 = b'\}\right|\end{aligned} \tag{A.9}$$

Aqui podemos ver que dado $Z = z$, as $O_{a,b}(X,z)$ correspondem à soma de $n_a(z)n_b(z)$ variáveis aleatórias independentes com distribuição de Poisson, dadas por $X_{ij}\mathbb{1}\{z_i = a, z_j = b\}$ cujo valor esperado é dado por $\rho_n\tilde{\lambda}_{z_i,z_j}$. Logo, temos que a soma também é Poisson distribuída com parâmetro sendo a correspondente soma de parâmetros. Dessa forma, podemos usar a desigualdade de concentração, tal que se $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$,

então

$$\mathbb{P}(|X-\lambda|>\alpha)\leq e^{-\frac{\alpha^2}{2(\lambda+\alpha)}}.$$

Assim, para todo $\epsilon>0$, temos que

$$\begin{aligned}&\mathbb{P}\left(\left|O_{a,b}(X,z)-\rho_n n^2[Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}\right|>\epsilon|Z=z^0\right)\\&\leq e^{-\frac{\epsilon^2}{2\left(\rho_n n^2[Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}+\epsilon\right)}}.\end{aligned}$$

Por outro lado, temos que para qualquer z e z^0

$$\rho_n n^2[Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}\leq\rho_n n^2\tilde{\lambda}_{max},$$

em que $\tilde{\lambda}_{max}=\max_{a',b'}\tilde{\lambda}_{a',b'}$. Logo, para quaisquer z,z^0 e $\epsilon>0$ segue

$$\mathbb{P}\left(\left|\frac{O_{a,b}(X,z)}{\rho_n n^2}-[Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}\right|>\epsilon|Z=z^0\right)\leq e^{-\frac{\rho_n n^2\epsilon^2}{2(\lambda_{max}+\epsilon)}}.$$

Agora, tomando união sobre todos os $z\in[k]^n$ e integrando em z^0, temos que

$$\mathbb{P}\left(\sup_{z\in[k]^n}\left|\frac{O_{a,b}(X,z)}{\rho_n n^2}-[Q_n(z,z^0)\tilde{\lambda}Q_n(z,z^0)^T]_{a,b}\right|>\epsilon\right)\leq e^{\left(-\frac{\rho_n n^2\epsilon^2}{2(\lambda_{max}+\epsilon)}+n\log k\right)}.$$

Isso conclui a prova do Lema A.2.2. □

Lema A.2.3. *Para qualquer $\epsilon>0$ and $a,b\in[k]$, temos que*

$$\mathbb{P}\Big(\sup_{z\in[k]^n}\left|\frac{O_{ab}(X,z)}{\rho_n n^2}-[H_n(z,z^0)\tilde{\lambda}H_n(z,z^0)^T]_{ab}\right|>\epsilon\Big)\ \leq\ \exp\Big(-\frac{\rho_n n^2\epsilon^2}{2(\tilde{\lambda}_{max}+\epsilon)}+n\log k\Big)$$

e

$$\mathbb{P}\Big(\sup_{z\in[k]^n}\left|\frac{\bar{g}_a(X,z)}{\rho_n n^2}-[H_n(z,z^0)\tilde{\lambda}H_n(z,z^0)^T\mathbf{1}_k]_a\right|\ >\ \epsilon\Big)\ \leq\ \exp\Big(-\frac{\rho_n n^2\epsilon^2}{2(\tilde{\lambda}_{max}+\epsilon)}+n\log k\Big).$$

Proof. A prova da primeira desigualdade segue de forma similar à prova do Lema A.2.2. Para a segunda desigualdade temos que, dado $Z=z^0$

$$\bar{g}_a(X,z)=\sum_{b\in[k]}o_{ab}(X,z),$$

também é soma de variáveis aleatórias independentes com distribuição de Poisson, tal que

$$\mathbb{E}(\bar{g}_a(X,z)\mid Z=z^0)=[\rho_n n^2 H_n(z,z^0)\tilde{\lambda}H_n(z,z^0)^T\mathbf{1}_k]_a.$$

Logo, podemos ter que

$$\mathbb{P}\Big(\sup_{z\in[k]^n} \Big| \frac{\bar{g}_a(X,z)}{\rho_n n^2} - [H_n(z,z^0)\tilde{\lambda}H_n(z,z^0)^T\mathbf{1}_k]_a \Big| \ > \ \epsilon \Big) \ \leq \ \exp\Big(-\frac{\rho_n n^2 \epsilon^2}{2(\tilde{\lambda}_{max}+\epsilon)} + n\log k \Big).$$

Com isso concluimos a prova do Lema A.2.3. □

Bibliography

Abbe, E. *Community detection and stochastic block models: Recent developments. Journal of Machine Learning Research, v. 18, n. 177, p. 1–86, 2018.* *Disponível em* `http://jmlr.org/papers/v18/16-480.html`

Amini, A. A.; Chen, A.; Bickel, P. J.; Levina, E. *Pseudo-likelihood methods for community detection in large sparse networks. The Annals of Statistics, v. 41, n. 4, p. 2097–2122, 2013.*

Bertsekas, D. P. Convex optimization theory. *Athena Scientific, 2009.*

Bickel, P.; Choi, D.; Chang, X.; Zhang, H. *Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. The Annals of Statistics, v. 41, n. 4, p. 1922–1943, 2013.*

Bickel, P. J.; Chen, A. *A nonparametric view of network models and newman–girvan and other modularities. Proceedings of the National Academy of Sciences, v. 106, n. 50, p. 21068–21073, 2009.*

Biernacki, C.; Celeux, G.; Govaert, G. *Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 22, n. 7, p. 719–725, 2000.*

Boucheron, S.; Lugosi, G.; Massart, P. Concentration inequalities: A nonasymptotic theory of independence. *Oxford university press, 2013.*

Celisse, A.; Daudin, J.-J.; Pierre, L. *Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics, v. 6, p. 1847–1899, 2012.*

Cerqueira, A.; Leonardi, F. *Estimation of the number of communities in the stochastic block model. IEEE Transactions on Information Theory, v. 66, n. 10, p. 6403–6412, 2020.*

Chen, Y.; Li, X.; Xu, J. *Convexified modularity maximization for degree-corrected stochastic block models. The Annals of Statistics, v. 46, n. 4, p. 1573–1602, 2018.*

Côme, E.; Latouche, P. *Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. Statistical Modelling, v. 15, n. 6, p. 564–589, 2015.*

Daudin, J.-J.; Picard, F.; Robin, S. *A mixture model for random graphs. Statistics and computing, v. 18, n. 2, p. 173–183, 2008.*

Davisson, L.; McEliece, R.; Pursley, M.; Wallace, M. *Efficient universal noiseless source codes. IEEE Transactions on Information Theory, v. 27, n. 3, p. 269–279, 1981.*

Erdős, P. *Rényi, a.:" on random graphs. I". Publicationes Mathematicae (Debre, 1959.*

Erdos, P.; Rényi, A.; et al. *On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, v. 5, n. 1, p. 17–60, 1960.*

Gassiat, E.; Boucheron, S. *Optimal error exponents in hidden markov models order estimation. IEEE Transactions on Information Theory, v. 49, n. 4, p. 964–980, 2003.*

Grandjean, M. *Connected world: Untangling the air traffic network. 2016.*

Holland, P. W.; Laskey, K. B.; Leinhardt, S. *Stochastic blockmodels: First steps. Social networks, v. 5, n. 2, p. 109–137, 1983.*

Hu, J.; Qin, H.; Yan, T.; Zhao, Y. *Corrected bayesian information criterion for stochastic block models. Journal of the American Statistical Association, p. 1–13, 2019.*

Karrer, B.; Newman, M. E. *Stochastic blockmodels and community structure in networks. Physical review E, v. 83, n. 1, p. 016107, 2011.*

Latouche, P.; Birmele, E.; Ambroise, C. *Variational bayesian inference and complexity control for stochastic block models. Statistical Modelling, v. 12, n. 1, p. 93–115, 2012.*

Lei, J.; Rinaldo, A. *Consistency of spectral clustering in stochastic block models. The Annals of Statistics, v. 43, n. 1, p. 215–237, 2015.*

Ma, S.; Su, L.; Zhang, Y. *Determining the number of communities in degree-corrected stochastic block models. Journal of machine learning research, v. 22, n. 69, 2021.*

Mariadassou, M.; Robin, S.; Vacher, C. *Uncovering latent structure in valued graphs: a variational approach. The Annals of Applied Statistics, v. 4, n. 2, p. 715–742, 2010.*

Newman, M. E.; Girvan, M. *Finding and evaluating community structure in networks. Physical review E, v. 69, n. 2, p. 026113, 2004.*

Pas, S. L.; Vaart, A. W. *Bayesian community detection. Bayesian Analysis, v. 13, n. 3, p. 767–796, 2018.*

Rohe, K.; Chatterjee, S.; Yu, B. *Spectral clustering and the high-dimensional stochastic blockmodel. The Annals of Statistics, v. 39, n. 4, p. 1878–1915, 2011.*

Vu, D. Q.; Hunter, D. R.; Schweinberger, M. *Model-based clustering of large networks. The annals of applied statistics, v. 7, n. 2, p. 1010, 2013.*

Wang, Y. R.; Bickel, P. J.; et al. *Likelihood-based model selection for stochastic block models. The Annals of Statistics, v. 45, n. 2, p. 500–528, 2017.*

Yan, X. *Bayesian model selection of stochastic block models. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, p. 323–328.*

Yan, X.; Shalizi, C.; Jensen, J. E.; Krzakala, F.; Moore, C.; Zdeborová, L.; Zhang, P.; Zhu, Y. *Model selection for degree-corrected block models. Journal of Statistical Mechanics: Theory and Experiment, v. 2014, n. 5, p. P05007, 2014.*

Zhao, Y.; Levina, E.; Zhu, J. *Consistency of community detection in networks under degree-corrected stochastic block models. The Annals of Statistics, v. 40, n. 4, p. 2266–2292, 2012.*