UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Inferência em redes aleatórias com pesos discretos

Laila Letícia da Silva Costa

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Laila Letícia da Silva Costa

Inference in random networks with discrete weights

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Andressa Cerqueira

USP – São Carlos
June 2023

Laila Letícia da Silva Costa

Inferência em redes aleatórias com pesos discretos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Andressa Cerqueira

USP – São Carlos
Junho de 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 Dissertação de Mestrado da candidata Laila Leticia da Silva Costa, realizada em 04/04/2023.

Comissão Julgadora:

Profa. Dra. Andressa Cerqueira (UFSCar)

Profa. Dra. Denise Duarte Scarpa Magalhães Alves (UFMG)

Profa. Dra. Florencia Graciela Leonardi (USP)

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

Dedico esse trabalho a minha família.

AGRADECIMENTOS

Gostaria de expressar meus agradecimentos a todas as pessoas e instituições que tornaram possível a conclusão desta etapa importante em minha vida.

Em primeiro lugar, agradeço a Deus por me proporcionar a oportunidade de alcançar este objetivo, pela proteção e força que me concedeu ao longo dos anos e por colocar pessoas incríveis em meu caminho durante esse processo.

À minha família, agradeço a meus pais: Roberto e Átila, meus avós: Valmir e Maria, meus irmãos: Lara e Pedro, e meus tios: Val, Beniti, Crisanto, Iza, Jane, Fernando e Elton, pelo incondicional suporte, por acreditarem em mim, por nunca me deixarem desistir e por serem uma fonte de inspiração e força. Vocês são o melhor de mim e sem vocês eu não teria conseguido. Agradeço também aos meus primos, Izaellen, Mateus, Sofia, Eduardo, Lucas, Davi, Kaleb e Ágatha, por me fazerem feliz simplesmente por existirem. Amo todos vocês.

À minha orientadora, Andressa Cerqueira, sou imensamente grata pela oportunidade, orientação, paciência, dedicação, incentivo e sugestões ao longo deste trabalho. Você é um exemplo de pessoa e profissional e não poderia ter escolhido melhor.

Aos amigos que fiz durante o mestrado, Edér, Marina, Loriz, Nayara, Patrícia, Alex, Asrat e Isaac, agradeço por tornarem este processo mais leve e as idas ao laboratório mais divertidas. Um agradecimento especial a Adriane, por me ouvir, pelos conselhos, apoio e por me fazer rir até nos momentos difíceis. Vocês foram fundamentais para a conclusão deste mestrado e sou muito grata por todos os momentos compartilhados, troca de informações e força nos momentos difíceis.

Aos secretários do programa de pós-graduação, Monique e Julio, agradeço pela atenção, paciência e disposição para ajudar. Aos professores do PIPGES e da UFPI, obrigada por todo o aprendizado e crescimento que obtive com vocês.

Aos meus amigos de Teresina e aos amigos que fiz em São Carlos, agradeço pelo apoio e carinho. Agradeço também a todos que de alguma forma contribuíram ou estiveram na torcida pela minha conclusão do mestrado.

À CAPES, como esse trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

À todas as pessoas que incentivam e acreditam na ciência. Sem o apoio de todos vocês, esta conquista não teria sido possível.

“Não sou nada.
Nunca serei nada.
Não posso querer ser nada.
À parte isso, tenho em mim todos os sonhos do mundo.”
(Fernando Pessoa)

RESUMO

COSTA, L. L. S. **Inferência em redes aleatórias com pesos discretos**. 2023. 82 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

As redes aleatórias têm sido amplamente utilizadas para descrever interações entre objetos, incluindo as relações interpessoais entre indivíduos. Uma das características mais importantes das redes é a presença de comunidades, que são grupos de nós com padrões de conexão semelhantes. Neste sentido, propomos um modelo em que as arestas entre pares de vértices são atribuídas de maneira aleatória, dadas as comunidades desses vértices, seguindo a distribuição de Poisson inflada de zeros (ZIP). Essa proposta nos permite modelar redes com estrutura de comunidades que sejam esparsas e que apresentem pesos nas arestas. A estimação dos parâmetros da distribuição ZIP é realizada por meio do algoritmo EM, enquanto a estimação das comunidades é feita usando o algoritmo EM-Variacional. O desempenho dos estimadores é avaliado por meio de estudos de simulação, utilizando a medida de comparação Informação Mútua Nomalizada (NMI), para comparar as comunidades verdadeiras e a estimadas pelo método. Para comparar os parâmetros estimados da distribuição ZIP, utilizamos o Erro Quadrático Médio (EQM). Por fim, aplicamos o modelo proposto em redes aeroportuárias do Brasil e detectamos a estrutura de comunidades nos anos de 2018 a 2021, a fim de avaliar as mudanças ocorridas nessas redes antes e durante o período de pandemia do COVID-19.

Palavras-chave: Redes aleatórias, Detecção de comunidades, Modelo estocástico em blocos , Distribuição de Poisson inflada de zeros, EM-Variacional.

ABSTRACT

COSTA, L. L. S. **Inference in random networks with discrete weights**. 2023. 82 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Random networks have been widely used to describe interactions between objects, including interpersonal relationships between individuals. One of the most important features of networks is the presence of communities, which are groups of nodes with similar patterns of connection. In this regard, we propose a model in which edges between pairs of vertices are randomly assigned, given the communities of those vertices, following the zero-inflated Poisson (ZIP) distribution. This proposal allows us to model networks with community structure that are sparse and have edge weights. The estimation of the parameters of the ZIP distribution is performed using the EM algorithm, while the estimation of communities is done using the EM-Variational algorithm. The performance of the estimators is evaluated through simulation studies, using the Normalized Mutual Information (NMI) comparison measure to compare the true and estimated communities. To compare the estimated parameters of the ZIP distribution, we use the Mean Squared Error (MSE). Finally, we apply the proposed model to airport networks in Brazil and detect the community structure from 2018 to 2021, in order to evaluate the changes that occurred in these networks before and during the COVID-19 pandemic period.

Keywords: Random network, Community detection, Stochastic block model, Zero-inflated Poisson Distribution, Variational EM.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

As redes estão presentes em diversas áreas, que vão desde nosso cérebro, onde temos conexões entre neurônios definidas pelas sinapses, até as regiões corticais que nos permitem estudar doenças mentais, processos cognitivos, dentre outros. No mercado financeiro, por exemplo, as ações estão conectadas em redes complexas, assim como nas redes sociais, onde há interação entre indivíduos, e nas cadeias alimentares, onde há relações entre presas e predadores. São sistemas distintos que podem ser analisados de forma similar, sendo possível entender como um conjunto de setas representando as interações que, isoladamente, não trazem informações, mas que, ao examinar o todo, proporcionam indicações sobre como essas interações podem afetar outros âmbitos e identificar padrões que ajudam a avaliar as interações existentes e as prováveis de existir (FURTADO; TóLVOLLI, 2015).

A teoria dos grafos foi introduzida por Euler em 1735, quando ele propôs uma solução para o problema das pontes de Konisgsberg, modelando-o em forma de grafo. Euler conseguiu provar matematicamente que seria possível passar por todas as pontes sem repeti-las somente se o número de conexões que saíssem de cada vértice fosse par. Com isso, ele propôs o primeiro grafo da história. A partir do século XX, mais precisamente na década de 1930, as redes sociais começaram a ser representadas por meio de aspectos reais, como demonstrado por Moreno em seu trabalho de sociometria (MORENO, 1934). Desde então, a representação de sistemas por meio de grafos tem sido amplamente utilizada para avaliar relações interpessoais e outros aspectos em diversas áreas do conhecimento.

O modelo de redes aleatórias introduzido por (ERDŐS; RÉNYI *et al.*, 1960) é caracterizado pela conexão entre cada par de vértices com a mesma probabilidade, gerando uma rede homogênea em que todos os vértices têm, em média, o mesmo número de conexões. Entretanto, as redes do mundo real são altamente heterogêneas, o que significa que alguns vértices têm muitas conexões e outros têm poucas. Embora o modelo de Erdős-Rényi seja útil para entender as propriedades das redes aleatórias, pode não ser adequado para descrever completamente as

redes do mundo real, que geralmente possuem características mais complexas e estruturas mais heterogêneas.

As redes aleatórias podem ser representadas por grafos, que consistem em um conjunto de vértices, que representam os objetos de estudo, e esses vértices são conectados por arestas se existir uma interação entre eles ou uma certa similaridade de interesses. As redes nos permitem observar a organização de um determinado sistema, uma vez que, ao observar um único vértice (indivíduo), não conseguimos explicar as interações existentes com o objeto de estudo. É necessário observar as interações entre o conjunto de vértices, como ocorre nos sistemas complexos, conforme explicado por (LUKOSEVICIUS; MARCHISOTTI; SOARES, 2016).

As redes aleatórias procuram descrever, por meio de probabilidades, a aleatoriedade que existe em redes reais representadas por meio de grafos. Existem dois tipos de grafos: os direcionados e os não direcionados. No caso dos grafos direcionados, cada aresta tem um sentido conectando um vértice de origem a um vértice de destino, e essa relação é representada por uma matriz de adjacência assimétrica (FORTUNATO, 2010). Já nos grafos não direcionados, a existência da aresta significa que existe uma conexão entre os vértices, e essa relação é representada por uma matriz de adjacência simétrica. Além disso, um grafo pode ser ponderado, o que significa que conseguimos definir o peso da ligação entre dois vértices, representando o tráfego de informação entre eles.

Uma das características das redes que serão abordadas nesta dissertação é a estrutura de comunidade, cuja análise foi pioneiramente realizada por (WEISS; JACOBSON, 1955). Essa análise nos permite agrupar os vértices que possuem funções semelhantes em um grafo, facilitando a compreensão da organização da rede. Espera-se que haja mais arestas dentro das comunidades do que entre elas, o que é semelhante ao voo de pássaros em formação, onde cada pássaro ajusta sua posição com os pássaros vizinhos, sem que haja um controle centralizado da direção e posição de todas as aves no voo (FURTADO; TóLVOLLI, 2015).

Embora seja de grande importância detectar comunidades em sistemas representados como grafos, esse é um problema que ainda é muito difícil de resolver, uma vez que se trata de uma variável latente, especialmente quando se trata de dados reais, onde as redes tendem a ser esparsas, ou seja, apresentam uma baixa densidade de conexões entre seus vértices e, consequentemente, menos informações disponíveis sobre a rede.

Nesse sentido, o objetivo deste trabalho é propor um modelo de rede que melhor represente a estrutura de comunidades em redes esparsas com arestas ponderadas, em um grafo não direcionado e sem laços, ou seja, onde um vértice não pode se conectar a si mesmo. Dessa forma, o grafo resultante terá uma matriz de adjacência simétrica e uma diagonal principal igual a zero.

Existem vários métodos na literatura para detecção de comunidades em redes, como a inferência baseada em verossimilhança ou no espectro da rede. Entre eles, podemos destacar o particionamento do grafo (SUARIS; KEDEM, 1988), agrupamento parcial (LLOYD, 1982),

detecção espectral (LUXBURG, 2007), e hierárquico (NEWMAN, 2004). Um modelo estatístico frequentemente utilizado é o modelo estocástico em blocos (SBM), que assume uma estrutura de comunidades onde as arestas são geradas aleatoriamente por meio da distribuição Bernoulli condicional aos grupos (LEE; WILKINSON, 2019).

No artigo de (DONG; CHEN; WANG, 2020), é proposto o uso da distribuição de Poisson multivariada inflada de zeros (MZIP) para modelar redes com múltiplas camadas. Nesse contexto, cada camada representa um tipo de interação entre dois vértices conectados por uma aresta. Para gerar as arestas ponderadas, os autores utilizaram o modelo estocástico em blocos com MZIP (SBM-MZIP) para capturar as interações entre redes multicamadas esparsas, ponderadas e direcionadas, visando preservar a estrutura de comunidade. As comunidades foram estimadas por meio do método variacional.

O objetivo desta dissertação é apresentar um modelo de rede baseado na distribuição Poisson inflada de zeros (ZIP), que é uma mistura de dois processos geradores de valores zero. O primeiro processo gera apenas valores zero, permitindo identificar onde não existem arestas na rede. Já o segundo processo utiliza a distribuição de Poisson para gerar contagens, que correspondem aos pesos das arestas. Dessa forma, essa distribuição determina quais arestas estão presentes no grafo e atribui peso às que existem, permitindo controlar a esparsidade da rede.

No artigo de (DONG; CHEN; WANG, 2020), os autores estimaram os parâmetros da MZIP por meio de métodos numéricos para maximizar a verossimilhança, enquanto as comunidades foram estimadas por um método variacional. Já em outra proposta de modelo para redes esparsas usando a distribuição ZIP, (MOTALEBI; STEVENS; STEINER, 2021) também empregaram métodos numéricos para estimar os parâmetros. Diferentemente dessas abordagens, nesta dissertação, não consideramos redes multicamadas e utilizamos o algoritmo EM para estimar os parâmetros da ZIP e o algoritmo EM-Variacional para estimar as comunidades de redes ponderadas e não direcionadas. Portanto, por meio do SBM-ZIP, é possível modelar redes com estrutura de comunidades, em que as arestas são geradas pela distribuição ZIP. Dessa forma, espera-se que esse modelo possa descrever redes mais esparsas e com pesos nas arestas.

Nos últimos anos, as análises de redes têm se mostrado bastante úteis em estudos de transportes, como ferrovias (SEN *et al.*, 2003) e aeroportos (GUIMERA; AMARAL, 2004; BAGLER, 2008; WU *et al.*, 2006). Em particular, as redes aéreas têm sido objeto de estudo para avaliar a proliferação de vírus que causam epidemias, como foi evidenciado no trabalho de Colizza et al. (COLIZZA *et al.*, 2006). A importância desse tipo de rede ficou ainda mais evidente durante a pandemia de COVID-19, em que os primeiros casos foram relatados na China no final de dezembro de 2019 e a doença se espalhou globalmente, levando à declaração de uma pandemia pela Organização Mundial da Saúde em 11 de março de 2020, fazendo com que diversas nações adotassem medidas protetivas.

Durante o período de pandemia, o setor aéreo sofreu significativas mudanças e adaptações. Embora tenha ocorrido uma diminuição no número de passageiros, a Agência Nacional de

Aviação Civil (ANAC, 2021), em conjunto com o Governo Federal, implementou diversas medidas emergenciais. Entre elas, destaca-se o ajuste na malha aérea, visando manter pelo menos uma ligação aérea em todos os estados brasileiros. Além disso, foi permitido o transporte de cargas por empresas de táxi-aéreo, visando transportar materiais para ações de combate à pandemia, como a vacinação.

Para destacar a importância da análise de redes no setor aéreo durante a pandemia, propomos a aplicação do modelo SBM-ZIP na rede aérea brasileira entre 2018 e 2021, a fim de investigar a estrutura de comunidades da rede antes e durante esse período crítico. A aplicação do modelo é viável, uma vez que a rede aérea é semelhante aos pressupostos do modelo, com vértices representando aeroportos, arestas indicando conexões e pesos associados ao número de voos. Adicionalmente, a existência de diversas conexões ausentes é representativa do setor. Essa análise pode fornecer informações importantes sobre a dinâmica da rede aérea e sua relação com a pandemia, auxiliando em futuras decisões de gestão. Os resultados dessa aplicação serão discutidos com mais detalhes no Capítulo 5.

Esta dissertação está estruturada em cinco capítulos e dois apêndices. No Capítulo 2, é feita uma breve introdução sobre grafos e apresentado o modelo proposto. No Capítulo 3, é descrito o método de estimação dos parâmetros e o algoritmo de detecção de comunidades. O Capítulo 4 apresenta os resultados da simulação do modelo, onde são avaliados o comportamento do modelo, o funcionamento do algoritmo, as estimativas dos parâmetros e a comparação com outros métodos. Já o Capítulo 5 examina a rede aeroportuária entre 2018 e 2021, com o objetivo de identificar a estrutura de comunidades e verificar se houve mudanças decorrentes da pandemia de COVID-19. No Capítulo 6, são discutidos os principais resultados obtidos e feita uma conclusão sobre o modelo proposto. Os apêndices A e B apresentam, respectivamente, uma forma de obter os estimadores de máxima verossimilhança para o modelo ZIP e uma explicação sobre o método variacional.

CAPÍTULO

2

MODELO

2.1 Representação de grafos

Para representar a organização de uma rede, usamos uma estrutura matemática chamada grafo. Um grafo simples $G(V,E)$ é descrito por um conjunto de vértices/nós (V) e um conjunto de arestas (E) que conectam pares de vértices. Seja n o número de nós em uma rede $(n = |V|)$. Podemos representar um grafo pela matriz de adjacência $\mathbf{A}$, dada por:

$$A_{ij} = \begin{cases} 1, & \text{se existe uma aresta entre i e j} \\ 0, & \text{caso contrário} \end{cases}$$

para $1 \leq i < j \leq n$, consideramos $\mathbf{A}$ como uma matriz $n \times n$ em que suas entradas indicam a existência de aresta no grafo. Nesse tipo de grafo não há laços, ou seja, quando um vértice não se conecta a si mesmo $(A_{ii} = 0)$. O grafo é não-direcionado, ou seja, a matriz de adjacência é simétrica $(A_{ij} = A_{ji})$ e não possui pesos $(A_{ij} \in 0,1)$. Portanto, a diagonal da matriz de adjacência é composta por zeros.

Em um grafo direcionado, a orientação das arestas é fundamental para o seu correto entendimento. Dessa forma, a matriz de adjacência, representada por A_{ij}, indica a existência de uma conexão direcionada saindo do vértice i e chegando ao vértice j. Já a entrada A_{ji} indica a existência de uma conexão saindo do vértice j e indo em direção ao vértice i.

Para redes com pesos, os grafos são representados por $G(V,E,C)$, onde C é um conjunto de pesos associados a cada aresta. Nesse caso, o peso entre dois vértices i e j indica a intensidade da conexão entre eles. Assim, a matriz de adjacência $\mathbf{A}$ pode assumir valores em $\mathbb{N}$ ou $\mathbb{R}$. Redes ponderadas são frequentemente utilizadas, pois permitem modelar o fluxo de informações de maneira mais completa e realista, já que a intensidade das conexões entre os vértices é heterogênea.

2.2 Modelo Estocásticos em Blocos

O modelo estocástico em blocos (SBM) (HOLLAND; LASKEY; LEINHARDT, 1983; SNIJDERS; NOWICKI, 1997) é amplamente utilizado para entender ou modelar a estrutura latente de uma rede, ou seja, a estrutura de comunidades. Para detectar as comunidades, é necessário observar o padrão de conexões da rede, uma vez que inicialmente só conhecemos a existência ou ausência de conexões entre os vértices.

Inicialmente, pressupõe-se a existência de comunidades em um grafo, em que os vértices que pertencem à mesma comunidade estão mais interconectados entre si do que com vértices de outras comunidades (STANLEY *et al.*, 2019). A detecção de comunidades permite compreender a estrutura de redes reais em diversas aplicações, seja em casos em que essa estrutura é nítida, como em redes sociais, ou em problemas em que essa estrutura é implícita, mas pode ser construída para solucionar problemas específicos.

Para aplicar o SBM, é necessário atribuir os vértices em grupos (comunidades) e inferir os parâmetros do modelo que melhor se ajustam aos dados observados. Geralmente, as arestas são geradas aleatoriamente e possuem valores binários, seguindo uma distribuição de Bernoulli cuja probabilidade depende da associação dos vértices aos grupos (LEE; WILKINSON, 2019).

Algumas especificações do SBM: pressupõe-se uma rede com n vértices, uma rede não direcionada, sem laço com matriz de adjacência $\mathbf{A}_{nxn}$, K é o número de comunidades, $\mathbf{Z}_n = (Z_1, Z_2, ..., Z_n)$ é um vetor aleatório latente que representa as comunidades sendo variáveis independentes. Assim, é possível informar em qual comunidade cada vértice pertence, em que $\mathbb{P}(Z_{ia} = 1) = \pi_a$ é a probabilidade do vértice i pertencer à comunidade a.

Atribuímos as conexões da rede através de uma matriz com elementos A_{ij} que são variáveis aleatórias condicionalmente independentes dado $\mathbf{Z}_n$ e que indicam a existência de uma conexão entre os vértices i e j, em que $A_{ij} \in \{0, 1\}$. A matriz de adjacência é gerada de tal forma que dependa apenas das comunidades. Então, a existência de uma conexão entre os vértices i e j, dados que o vértice i pertence à comunidade a e o vértice j pertence à comunidade b, segue uma distribuição de Bernoulli com probabilidade de sucesso que depende apenas das comunidades a e b, ou seja,

$$A_{ij}|Z_{ia} = 1, Z_{jb} = 1 \sim Ber(p_{ab}), \qquad 1 \leq i < j \leq n.$$

Portanto, temos que A_{ij} é condicionalmente independente das outras variáveis dado o vetor aleatório $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$.

No nosso estudo, vamos utilizar a distribuição de Poisson inflada de zeros (ZIP) para atribuir as arestas da rede. Essa distribuição é um modelo de mistura de dois processos que geram zeros, permitindo trabalhar com redes esparsas. Por meio desse modelo, podemos avaliar não apenas a existência das arestas, mas também a força da ligação entre elas, o que será abordado nas próximas seções.

2.3 Distribuição de Poisson inflada de zeros (ZIP)

Ao analisar dados de redes aleatórias, muitas vezes nos deparamos com uma variabilidade maior do que o esperado. Isso ocorre devido à alta proporção de ausência de arestas, o que resulta em redes mais esparsas e pode prejudicar a estimação, devido ao excesso de zeros na matriz de adjacência. Para contornar esse problema, recorremos a modelos inflados em zero, como a distribuição de Poisson inflada em zeros. Esse modelo é uma mistura de dois processos que geram zeros e permite observações frequentes de valor zero. Uma das principais vantagens dos modelos com inflação em zero é a capacidade de indicar heterogeneidade não observada nos dados de forma simples e eficiente (BÖHNING, 1998).

A distribuição ZIP (p, λ) é definida como segue:

$$\begin{aligned} \mathbb{P}(Y=0) &= p + (1-p)e^{-\lambda} \\ \mathbb{P}(Y=y_i) &= (1-p)\frac{\lambda^{y_i}e^{-\lambda}}{y_i!}, \quad y_i = 1,2,3,... \end{aligned} \tag{2.1}$$

onde p representa a probabilidade de que um nó na rede não tenha nenhuma conexão com outros vértices na rede, y_i é qualquer valor inteiro não negativo, e λ é o parâmetro de contagem da Poisson, que é interpretado como a taxa média de ocorrência de eventos em uma unidade de tempo e como um parâmetro de intensidade.

O modelo ZIP também pode ser representado de forma alternativa:

$$\mathbb{P}(Y=y_i) = p\mathbb{1}\{y_i=0\} + (1-p)\frac{\lambda^{y_i}e^{-\lambda}}{yi!}, \quad y_i = 0,1,2,3,... \tag{2.2}$$

onde $\mathbb{1}_{\{y_i=0\}}$ é uma função indicadora que assume o valor 1 se $y_i = 0$, e o valor 0 caso contrário. Essa representação é equivalente à anterior e é mais conveniente para cálculos e simulações, além de facilitar a interpretação dos parâmetros. Vale ressaltar que a média da variável Y é $(1-p)\lambda$ e a variância é $\lambda(1-p)(1+p\lambda)$, o que permite obter informações importantes sobre a distribuição de probabilidade representada pelo modelo ZIP.

2.4 Modelo Proposto

Consideramos uma rede não-direcionada com n vértices. Buscamos um modelo que apresente uma estrutura de comunidade, e para isso, começamos definindo essa estrutura latente. Os nós da rede são divididos em k grupos, ou comunidades, conhecidos previamente. Essa divisão é descrita por um vetor de variáveis latentes $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ associado ao vértice i, de modo que $Z_{ia} = 1$ se o vértice i está na comunidade a e $Z_{ia} = 0$ caso contrário.

Podemos definir uma variável aleatória Poisson para descrever o peso das conexões existentes na rede. No entanto, em muitos casos, há um grande número de pares de vértices que não estão conectados, o que excede a quantidade de zeros permitida pela distribuição de Poisson.

Para contabilizarmos essas duas estruturas, consideramos a distribuição de Poisson inflada de zeros, como discutido anteriormente.

A distribuição ZIP pode ser vista como uma combinação de duas distribuições: a distribuição Poisson discreta definida no intervalo $[0,\infty)$, que modela a contagem de arestas presentes na rede, e a distribuição Bernoulli, que atribui probabilidades não negativas aos valores 0 e 1, permitindo modelar a ausência de arestas na rede.

Cada comunidade associada ao vértice i de uma rede é gerada a partir de uma distribuição Multinomial, de modo que $Z_{ia} = 1$ se o nó i pertence à comunidade a, e $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ são independentes. Portanto, para $i = 1, \ldots, n$, temos

$$\mathbf{Z}_i \sim M(1, \pi = (\pi_1, \cdots, \pi_K)),$$

em que $\pi = (\pi_1, \ldots, \pi_K)$ é a distribuição que define as comunidades. Cada entrada da matriz de adjacência A_{ij} é condicionalmente independente dado $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$, que nos fornece o peso das arestas do grafo. O número de vértices é denotado por n, e a variável que modela o peso das arestas segue uma distribuição ZIP, a qual é um processo de mistura que indica a existência ou não de arestas e, caso existam, fornece informações sobre o seu peso.

Serão utilizados dois modelos: no modelo global, é possível identificar a existência de arestas no modelo sem considerar as comunidades do grafo, o que significa que a falta de arestas não depende das comunidades. Já no modelo local, a probabilidade de existência de arestas depende das comunidades. Em cada caso, temos:

$$\begin{aligned} &\textbf{Modelo Global:} && A_{ij}|Z_{ia}Z_{jb} = 1 \sim ZIP(p;\lambda_{ab}), && 1 \leq i < j \leq n. \\ &\textbf{Modelo Local:} && A_{ij}|Z_{ia}Z_{jb} = 1 \sim ZIP(p_{ab};\lambda_{ab}), && 1 \leq i < j \leq n. \end{aligned} \tag{2.3}$$

onde Λ é uma matriz $K \times K$ contendo os valores dos parâmetros λ_{ab} para $1 \leq a, b \leq K$. Quando $Z_{ia}Z_{jb} = 1$, isso indica que o nó i pertence à comunidade a e o nó j pertence à comunidade b. Considerando um grafo não-direcionado e sem laço, a distribuição de $\mathbf{Z}$ é dada por:

$$\mathbb{P}(\mathbf{Z}|\pi) = \prod_{i=1}^{n} \mathbb{P}(\mathbf{Z}_i|\pi) = \prod_{i=1}^{n} \prod_{a=1}^{K} \pi_a^{Z_{ia}}.$$

A distribuição condicional de $\mathbf{A}$ dado $\mathbf{Z}$ é dada por:

$$\mathbb{P}(\mathbf{A}|\mathbf{Z}, \Lambda, p) = \prod_{1<i<j<n} \prod_{a,b=1}^{K} f_{ab}(A_{ij})^{Z_{ia}Z_{jb}}. \tag{2.4}$$

em que definimos $f_{ab}(A_{ij})$ para o modelo global como:

$$f_{ab}(A_{ij}) = \left(p + (1-p)e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0\}}} \left((1-p)\frac{\lambda_{ab}^{a_{ij}} e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0\}}} \tag{2.5}$$

e para o modelo local temos que:

$$f_{ab}(A_{ij}) = \left(p_{ab} + (1-p_{ab})e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0\}}} \left((1-p_{ab})\frac{\lambda_{ab}^{a_{ij}} e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0\}}}. \tag{2.6}$$

Assim, escrevemos a distribuição conjunta como:

$$\mathbb{P}(\mathbf{A}, \mathbf{Z} \mid \pi, p, \Lambda) = \mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \Lambda, p)\mathbb{P}(\mathbf{Z} \mid \pi). \tag{2.7}$$

Conforme a independência condicional, podemos decompor a log-verossimilhança completa como:

$$\log(\mathbb{P}(\mathbf{A}, \mathbf{Z}|\pi, p, \Lambda)) = \log(\mathbb{P}(\mathbf{A}|\mathbf{Z}, \Lambda, p)) + \log(\mathbb{P}(\mathbf{Z}|\pi)). \tag{2.8}$$

No caso do modelo local, em que o vetor de parâmetros p_{ab}, $1 \leq a, b \leq k$, está presente, a log-verossimilhança desse modelo pode ser obtida a partir da expressão (2.8), substituindo na função $\mathbb{P}(\mathbf{A}|\mathbf{Z}, \Lambda, p)$ a correspondente função $f_{ab}(A_{ij})$ descrita em (2.6).

2.4.1 Força de um vértice

O grau do vértice é uma medida fundamental para descrever um nó e entender o comportamento do modelo, já que ele nos informa quantas arestas estão conectadas ao vértice i. No entanto, em nosso trabalho, usamos redes ponderadas, o que nos levou a definir uma nova medida para avaliar a força de conexão das arestas. Assim, definimos a força do nó i como a soma de todos os pesos das arestas conectadas a ele, conforme expresso em (2.9).

$$s_i = \sum_{j=1}^{n} A_{ij}. \tag{2.9}$$

A força média dos vértices, por sua vez, é dada pela média das forças de todos os vértices da rede, conforme a equação (2.10).

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{2.10}$$

A medida (2.9) permite observar a interação entre os vértices de uma rede ponderadas, levando em conta não só o número de arestas conectadas ao vértice i, mas também a força de cada uma dessas conexões. De fato, a medida s_i é obtida pela soma dos pesos de todas as arestas que incidem no vértice i. Dessa forma, a medida s_i nos fornece uma informação mais completa sobre a importância de cada vértice na rede. A medida $\bar{s}$, definida em (2.10), nos dá uma visão geral da força média de conexão dos vértices na rede. Essas medidas são importantes para avaliar diferentes padrões de conexões em redes ponderadas e entender o comportamento do modelo (NEWMAN, 2004).

Uma maneira de obter propriedades do modelo proposto de rede é calcular a esperança da força do vértice. Essa medida é baseada no número de arestas em um vértice, ponderada pelo peso de cada aresta. Basicamente, estamos somando o peso das arestas para cada vértice.

$$\mathrm{E}(s_i) = \mathrm{E}\left(\sum_{j=1}^{n} A_{ij}\right). \tag{2.11}$$

A esperança da força do vértice é calculada pela equação (2.11), que nos permite quantificar a força média do grau i. No entanto, para calcular a esperança da força do vértice, precisamos primeiro encontrar a esperança condicional.

$$\mathrm{E}(A_{ij}) = \mathrm{E}[\mathrm{E}(A_{ij}|\mathbf{Z})].$$

em que para o modelo global, temos:

$$\mathrm{E}(A_{ij}|\mathbf{Z}) = \sum_{a,b=1}^{K} (1-p)\lambda_{ab}\mathbb{1}_{\{Z_{ia}Z_{jb}=1\}}.$$

Logo,

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left(A_{ij}|\mathbf{Z}\right)\right] &= \sum_{a,b=1}^{k} (1-p)\lambda_{ab}\mathrm{E}\left(\mathbb{1}_{\{Z_{ia}Z_{jb}=1\}}\right) \\
&= \sum_{a,b=1}^{k} (1-p)\lambda_{ab}\mathbb{P}(Z_{ia}Z_{jb}=1) \\
&= \sum_{a,b=1}^{k} (1-p)\lambda_{ab}\mathbb{P}(Z_{ia}=1, Z_{jb}=1) \\
&= \sum_{a,b=1}^{k} (1-p)\lambda_{ab}\pi_a\pi_b.
\end{aligned} \tag{2.12}$$

Na última igualdade da equação anterior, utilizamos a independência das variáveis Z_i e Z_j, uma vez que para cada nó i temos uma comunidade de forma independente, e a distribuição multinomial.

De forma semelhante, podemos encontrar a esperança da força do vértice para o modelo local. Essa esperança é dada por:

$$\mathrm{E}\left[\mathrm{E}\left(A_{ij}|\mathbf{Z}\right)\right] = \sum_{a,b=1}^{k} (1-p_{ab})\lambda_{ab}\pi_a\pi_b. \tag{2.13}$$

Podemos observar que esperança da força do vértice não depende de i, o que sugere que a rede é homogênea em relação à força média dos vértices. Quando fixamos π e λ, o parâmetro p- que representa a probabilidade de que um nó na rede não tenha nenhuma conexão com outros

vértices na rede- se torna crucial para caracterizar a força da rede, já que ele controla a força média dos vértices. Se p é próximo de 1, a força média será baixa, enquanto que se p é próximo de 0, a rede se torna uma Poisson e a força média depende apenas de λ e π.

Para o modelo em questão, em que as comunidades seguem uma distribuição Multinomial e as arestas seguem uma Poisson inflada de zeros, é esperado que a rede apresente uma densidade de arestas relativamente baixa em relação ao tamanho da rede. Isso significa que muitos pares de vértices não estarão conectados por uma aresta. Portanto, é esperado que a força média dos vértices seja baixa, já que cada vértice terá poucas conexões em comparação com a quantidade total de vértices da rede.

2.5 Problemas Inferenciais

Para o modelo descrito, enfrentamos alguns desafios na inferência dos parâmetros (Λ, p, π). Para encontrar os estimadores dos parâmetros do modelo, é necessário maximizar a conjunta $\mathbb{P}(\mathbf{A}, \mathbf{Z} \mid \Lambda, p, \pi)$. Portanto, buscamos encontrar $\hat{\pi}, \hat{\lambda}_{ab}, \hat{p}$ que maximizam a conjunta para um $\mathbf{Z}$ fixo.

O segundo problema é a detecção das comunidades, que requer a estimativa de qual grupo cada vértice pertence, assumindo que K é conhecido, ou seja, queremos estimar $\mathbf{Z}$. Iremos abordar este problema usando métodos variacionais, que serão explicados detalhadamente no próximo capítulo.

O terceiro problema é a estimação do número de comunidades K, que não será abordado nesta dissertação.

CAPÍTULO

3

ESTIMAÇÃO

Nesse capítulo, desejamos estimar os parâmetros do modelo $\theta = (\pi, p, \Lambda)$ e identificar a comunidade à qual cada vértice pertence. Como a variável latente $\mathbf{Z}$ é desconhecida, a estimação será realizada por meio do método EM-Variacional. Esse método é uma técnica iterativa que combina o algoritmo EM com métodos de inferência variacional para encontrar o valor máximo da função de verossimilhança, permitindo a estimação dos parâmetros do modelo e a inferência dos valores de $\mathbf{Z}$ de forma eficiente.

3.1 Método EM - Variacional

Para estimar as comunidades do modelo, é necessário encontrar a probabilidade marginal $\mathbb{P}(\mathbf{Z} \mid \mathbf{A}, \pi, \Lambda, p)$. No entanto, não há solução analítica para esse problema, já que $\mathbf{Z}$ é uma variável latente. Portanto, aproximaremos $\mathbb{P}(\mathbf{Z} \mid \mathbf{A}, \pi, \Lambda, p)$ por outra distribuição $Q(\mathbf{Z})$ para encontrar a distribuição conjunta aproximada das variáveis não observadas. Esse método foi proposto por (HATHAWAY, 1986) e (NEAL; HINTON, 1998) e possui dois estágios. Dada uma distribuição $Q(\mathbf{Z})$, o logaritmo da probabilidade dos dados observados é decomposto em dois termos:

$$\log \mathbb{P}(\mathbf{A} \mid \pi, \Lambda, p) = D_{\mathrm{KL}}(Q(\cdot) \,\|\, \mathbb{P}(\cdot \mid \mathbf{A}, \pi, \Lambda, p)) + \mathfrak{L}(Q, \pi, \Lambda, p)\,,$$

em que

$$D_{\mathrm{KL}}(Q(\cdot) \,\|\, \mathbb{P}(\cdot \mid \mathbf{A}, \pi, \Lambda, p)) = -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{\mathbb{P}(\mathbf{Z} \mid \mathbf{A}, \pi, \Lambda, p)}{Q(\mathbf{Z})}\,, \tag{3.1}$$

onde $\log \mathbb{P}(\mathbf{A} \mid \pi, \Lambda, p)$ é conhecida como evidência, $D_{\mathrm{KL}}(Q(\cdot) \,\|\, \mathbb{P}(\cdot \mid \mathbf{A}, \pi, \Lambda, p))$ representa a divergência reversa de Kullback-Leibler e $\mathfrak{L}(Q, \pi, \Lambda, p)$ é o limite inferior da evidência (ELBO), uma vez que $\log \mathbb{P}(\mathbf{A} \mid \Lambda, p, \pi) \geq \mathfrak{L}(Q, \pi, \Lambda, p)$. Para obter mais informações sobre o cálculo realizado para obter a equação (3.1), por favor, consulte o Apêndice B na equação (B.6).

O ELBO pode ser escrito como

$$\begin{aligned}\mathfrak{L}(Q,\pi,\Lambda,p) &= -\sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{\mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p)}{Q(\mathbf{Z})} \\ &= \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p) - \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log Q(\mathbf{Z}) \\ &= \mathbb{E}_Q[\log \mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p)] - \mathbb{E}_Q[\log Q(\mathbf{Z})].\end{aligned} \tag{3.2}$$

É importante lembrar que a equação (2.7) nos fornece a expressão de $\mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p)$ para o modelo global. Para o modelo local, a função de probabilidade adequada deve ser utilizada. A divergência reversa de Kullback-Leibler nos ajuda a compreender a quantidade de informação que está sendo perdida ao utilizarmos uma distribuição aproximada em vez da verdadeira. Como a evidência é uma constante, para diminuir a divergência é necessário maximizar $\mathfrak{L}(Q,\pi,\Lambda,p)$, garantindo assim a menor divergência possível e encontrando a distribuição $Q(\mathbf{Z})$ que melhor se aproxima da distribuição verdadeira $\mathbb{P}(\mathbf{Z} \mid \mathbf{A},\pi,\Lambda,p)$.

Portanto, estamos diante de um problema de otimização variacional e para encontrar o $Q(\mathbf{Z})$ que maximize o ELBO, $\mathfrak{L}(Q,\pi,\Lambda,p)$, assumimos que as variáveis são independentes e que a distribuição de cada variável segue uma distribuição Multinomial.

$$Q(\mathbf{Z}) = \prod_{i=1}^{n} Q(\mathbf{Z}_i) = \prod_{i=1}^{n} \mathscr{M}(\mathbf{Z}_i, 1, \tau_i), \tag{3.3}$$

onde $\tau_i = (\tau_{i1},\ldots,\tau_{iK})$ é um vetor que representa as probabilidades do nó i pertencer a cada uma das K comunidades possíveis, onde τ_{ia} é a probabilidade do nó i pertencer à comunidade a.

Dessa maneira, o algoritmo EM Variacional é dividido em dois passos: Passo-E e Passo-M. No Passo-E, encontramos a distribuição aproximada $Q(\mathbf{Z})$ por meio da classe de distribuições dada pela fatoração (3.3) que minimiza a divergência D_{KL}, isto é, maximizamos $\mathfrak{L}(Q,\pi,\Lambda,p)$ em relação a $Q(\mathbf{Z})$, considerando que os parâmetros π,Λ,p são fixos. No Passo-M, estimamos os parâmetros π,Λ e p maximizando o ELBO em relação a esses parâmetros, considerando que $Q(\mathbf{Z})$ é fixo. Para $Q(\mathbf{Z})$ fixo, maximizar o ELBO é equivalente a maximizar o primeiro termo em (3.2). Portanto, na etapa de iteração t, o método EM-Variacional é dado por:

$$\begin{aligned}&\text{E-step:} && Q^{(t)}(\mathbf{Z}) = \arg\max_{Q'} \mathfrak{L}(Q'(\mathbf{Z}), \pi^{(t-1)}, \Lambda^{(t-1)}, p^{(t-1)}) \\ &\text{M-step:} && (\pi^{(t)},\Lambda^{(t)},p^{(t)}) = \arg\max_{(\pi,\Lambda,p)} \mathfrak{L}(Q^{(t)}(\mathbf{Z}),\pi,\Lambda,p) \\ & && \quad = \arg\max_{(\pi,\Lambda,p)} \mathbb{E}_{Q^{(t)}}[\log \mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p)]\,.\end{aligned} \tag{3.4}$$

3.1.1 Passo-E

Nessa etapa, o objetivo é maximizar o ELBO no Passo-E, utilizando o resultado variacional para que, ao substituí-lo em (3.1), possamos minimizar a divergência de KL. Para encontrar a distribuição aproximada ótima, ou seja, aquela que faz a diferença entre a distribuição

aproximada e a verdadeira ser a menor possível, usaremos a fatoração da distribuição $Q(\mathbf{Z})$ usando a aproximação de campo médio em (3.3) em (3.2). A solução ótima, obtida conforme descrito no Apêndice B, é dada por:

$$\begin{aligned}\log Q(\mathbf{Z}_i) &= \mathbb{E}_{\mathbf{Z}_m|m\neq i}[\log \mathbb{P}(\mathbf{A},\mathbf{Z}\mid \pi,\Lambda,p)] + \mathrm{c} \\ &= \mathbb{E}_{\mathbf{Z}_m|m\neq i}[\log(\mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p)) + \log(\mathbb{P}(\mathbf{Z}|\pi)] + \mathrm{c} \\ &= \mathbb{E}_{\mathbf{Z}_m|m\neq i}\left[\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{a,b=1}^{K} Z_{ia}Z_{jb}\log f_{ab}(A_{ij})\right] \\ &\quad + \mathbb{E}_{\mathbf{Z}_m|m\neq i}\left[\sum_{a=1}^{K} Z_{ia}\log(\pi_a)\right] + \mathrm{c}\,,\end{aligned} \tag{3.5}$$

onde c é uma constante normalizadora e $\mathbb{E}_{\mathbf{Z}_m|m\neq i}$ denota a esperança sob todas as variáveis $\mathbf{Z}_m$, para $m \neq i$. A $f_{ab}(A_{ij})$ é a distribuição ZIP, dada por:

$$f_{ab}(A_{ij}) = \left(p_{ab} + (1-p_{ab})e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0\}}} \left((1-p_{ab})\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0\}}}. \tag{3.6}$$

Considere $\tau_i = (\tau_{i1},\cdots,\tau_{iK})$ tal que $\tau_{ia} = \mathbb{P}(Z_{ia}=1)$, assim $\mathbb{E}(Z_{ia}) = \tau_{ia}$ e $\mathbb{E}(Z_{ia}Z_{jb}) = \tau_{ia}\tau_{jb}$, visto que a distribuição $Q(\mathbf{Z})$ está na classe de distribuições Multinomiais. Logo,

$$\log Q(\mathbf{Z}_i) = \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{a,b=1}^{K} Z_{ia}\tau_{jb}\log f_{ab}(A_{ij}) + \sum_{a=1}^{K} Z_{ia}\log(\pi_a) + \mathrm{c}. \tag{3.7}$$

Organizando os resultados da Equação (3.7) para tentar encontrar a distribuição de $Q(\mathbf{Z}_i)$, temos:

$$\begin{aligned}\log Q(\mathbf{Z}_i) &= \sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{a=1}^{K}\sum_{b=1}^{K} Z_{ia}\tau_{jb}\log f_{ab}(A_{ij}) + \sum_{i=1}^{n}\sum_{a=1}^{K} Z_{ia}\log(\pi_a) \\ &= \sum_{a=1}^{K} Z_{ia}\left(\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{b=1}^{k}\tau_{jb}\log f_{ab}(A_{ij}) + \log(\pi_a)\right) \\ &= \sum_{a=1}^{K} Z_{ia}(\log\tau_{ia}).\end{aligned} \tag{3.8}$$

Assim, podemos observar que $Q(\mathbf{Z}_i) \sim \mathscr{M}(1,\tau_i = (\tau_{i1},\cdots,\tau_{iK}))$. Note que, devido à interdependência das variáveis, ao minimizá-las, precisamos considerar todas as funções $Q(\mathbf{Z}_j)$, pois para calcular o τ do vértice i é preciso conhecer o τ do vértice j. Portanto, temos que:

$$\log \tau_{ia} \propto \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{b=1}^{K} \tau_{jb} \log f_{ab}(A_{ij}) + \log(\pi_a). \tag{3.9}$$

Para encontrar $Q(\mathbf{Z}_i)$, temos que τ_{ia} é

$$\begin{aligned} \hat{\tau}_{ia} &\propto \exp\left\{ \sum_{\substack{j=1 \\ j\neq i}}^{n} \sum_{b=1}^{K} \hat{\tau}_{jb} \log f_{ab}(A_{ij}) + \log(\pi_a) \right\} \\ &\propto \prod_{\substack{j=1 \\ j\neq i}}^{n} \prod_{b=1}^{K} \exp\left\{ \hat{\tau}_{jb} \log f_{ab}(A_{ij}) \right\} \exp\left\{ \log(\pi_a) \right\} \\ &\propto \pi_a \prod_{\substack{j=1 \\ j\neq i}}^{n} \prod_{b=1}^{K} f_{ab}(A_{ij})^{\hat{\tau}_{jb}}. \end{aligned} \tag{3.10}$$

Desse modo, obtemos $\hat{\tau}$ usando um método iterativo, por meio do algoritmo do ponto fixo, garantindo assim a existência e unicidade do ponto fixo, o que auxilia na convergência das iterações do método.

Ao concluir as duas etapas, o estimador para as comunidades é obtido por:

$$\hat{\mathbf{Z}}_i = \max_{K=1,\cdots,k} \tau_{iK}^{(t)} \tag{3.11}$$

onde $\tau_{iK}^{(t)}$ representa a probabilidade do vértice i pertencer à comunidade K no t-ésimo instante. Assim, cada vértice é alocado à comunidade que possui a maior probabilidade de pertencer, conforme os valores de $\tau_{iK}^{(t)}$ calculados.

Após a etapa de atualização das comunidades, fixamos $Q(\mathbf{Z}_i)$ e maximizamos o limite inferior em relação aos parâmetros do modelo global π, Λ e p, sendo que o parâmetro p não depende das comunidades.

É importante destacar que o processo realizado no Passo-E é aplicável tanto para o modelo local quanto para o modelo global.

3.1.2 Passo-M

Primeiramente, calculamos

$$\begin{aligned} \mathbb{E}(\log Q(Z_i)) &= \mathbb{E}\left(\log\left(\prod_{a=1}^{K} \tau_{ia}^{Z_{ia}} \right) \right) = \mathbb{E}\left(\sum_{a=1}^{K} Z_{ia} \log \tau_{ia} \right) \\ &= \sum_{a=1}^{K} \mathbb{E}(Z_{ia} \log \tau_{ia}) = \sum_{a=1}^{K} \tau_{ia} \log \tau_{ia}. \end{aligned}$$

Para maximizar o ELBO, precisamos encontrar os estimadores dos parâmetros Λ, p e π. Primeiramente, usando a decomposição da log-verossimilhança (2.8), podemos escrever:

$$\begin{aligned}\mathfrak{L}(Q,\pi,\Lambda,p) &= \mathbb{E}_Q[\log \mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p)] - \mathbb{E}[\log Q(\mathbf{Z})] \\ &= \sum_{1<i,j<n} \sum_{a,b=1}^{K} \tau_{ia}\tau_{jb} \log f_{ab}(A_{ij}) + \sum_{i=1}^{n}\sum_{a=1}^{K} \tau_{ia}\log(\pi_a) - \sum_{i=1}^{n}\sum_{a=1}^{K} \tau_{ia}\log\tau_{ia}.\end{aligned} \tag{3.12}$$

Note que a entropia de $\mathbb{E}[\log Q(\mathbf{Z})]$ é uma constante, pois não depende dos parâmetros.

Para maximizar o ELBO, começamos por encontrar o valor dos parâmetros π, que possui uma solução analítica. Derivando (3.12) em relação a π_a, obtemos:

$$\begin{aligned}\frac{\partial}{\partial \pi_a}\mathfrak{L}(Q,\pi,\Lambda,p) &= \frac{\partial}{\partial \pi_a}\mathbb{E}_Q[\log \mathbb{P}(\mathbf{A},\mathbf{Z} \mid \pi,\Lambda,p)] \\ &= \frac{\partial}{\partial \pi_a}\left[\sum_{i=1}^{n}\sum_{a=1}^{K} \tau_{ia}\log(\pi_a) + C\right].\end{aligned} \tag{3.13}$$

Para encontrar $\hat{\pi}$, podemos utilizar o multiplicador de Lagrange para levar em conta a restrição de que $\sum_{a=1}^{K}\pi_a = 1$, já que π é um vetor de probabilidades. Para isso, podemos definir a função Lagrangiana:

$$f(\pi_a) = \sum_{i=1}^{n} \tau_{ia}\log(\pi_a) - \alpha\left(\sum_{a=1}^{K}\pi_a - 1\right),$$

derivando e igualando a zero, temos

$$\begin{aligned} f(\pi_a)' &= \frac{\sum_{i=1}^{n}\tau_{ia}}{\pi_a} - \alpha = 0 \\ \sum_{i=1}^{n}\tau_{ia} &= \alpha\pi_a. \end{aligned} \tag{3.14}$$

Sabe-se

$$\begin{aligned} 1 = \sum_{a=1}^{K}\pi_a &= \sum_{a=1}^{K}\sum_{i=1}^{n}\tau_{ia}\frac{1}{\alpha} \\ \alpha &= \sum_{a=1}^{K}\sum_{i=1}^{n}\tau_{ia}. \end{aligned} \tag{3.15}$$

Usando (3.14) em (3.15)

$$\hat{\pi}_a = \frac{\sum_{i=1}^{n}\tau_{ia}}{\sum_{a=1}^{K}\sum_{i=1}^{n}\tau_{ia}} = \frac{\sum_{i=1}^{n}\tau_{ia}}{n}. \tag{3.16}$$

Podemos observar que $\hat{\pi}_a$ só depende das atribuições dada a τ.

Para maximizar (3.12) para o modelo global em relação a Λ e p, é necessário maximizar a seguinte função

$$\mathscr{E}(\Lambda, p; \tau, A_{ij}) = \sum_{1<i<j<n} \sum_{a,b=1}^{K} \tau_{ia}\tau_{jb} \log f_{ab}(A_{ij}) \,. \tag{3.17}$$

que depende da log-verossimilhança da distribuição ZIP, $\log f_{ab}(A_{ij})$, não tem forma fechada, como descrito no Apêndice A e B. Para contornar esse problema, propomos o uso do Algoritmo EM para encontrar o estimador por meio de uma aproximação. Esse método consiste em imputar valores para os dados faltantes e usar métodos aproximados até que se obtenha a forma do estimador desejado.

O conjunto de dados observáveis A_{ij} é condicionado a $\mathbf{Z}$ e segue a distribuição ZIP (2.3). Para cada zero em A_{ij}, não conseguimos identificar se ele segue distribuição de Poisson ou distribuição de Bernoulli. Portanto, introduzimos uma variável latente $W_{ij}^{ab}, 1 \leq i < j \leq n$ e $1 < a, b < K$, no modelo para indicar se o valor que observamos em A_{ij} vêm da parte da distribuição de Poisson ou da distribuição de Bernoulli, condicionados às comunidades $\mathbf{Z}$.

$$W_{ij}^{ab} = \begin{cases} 1, & \text{se } A_{ij} \text{ foi gerada da Bernoulli} \\ 0, & \text{se } A_{ij} \text{ foi gerada da Poisson} \end{cases}$$

Então, reescrevemos (2.6) usando as variáveis latentes W_{ij}^{ab} como

$$f_{ab}(a_{ij}) = [p\mathbb{1}_{\{a_{ij}=0\}}]^{W_{ij}^{ab}} \left[(1-p) \frac{\lambda_{ab}^{a_{ij}} e^{-\lambda_{ab}}}{a_{ij}!} \right]^{1-W_{ij}^{ab}} \tag{3.18}$$

$a_{ij} \in \{0, 1, \dots\}$ e aplicando o logaritmo, temos

$$\log f_{ab}(a_{ij}) = W_{ij}^{ab} \log \left(\frac{p}{1-p} \right) + (1 - W_{ij}^{ab}) \log \left(\frac{\lambda_{ab}^{a_{ij}} e^{-\lambda_{ab}}}{a_{ij}!} \right) + \log(1-p) \,. \tag{3.19}$$

O método funciona em duas etapas:

- Passo-E: calcular a esperança de (3.17) em relação à distribuição condicional da variável latente $\mathbf{W}$ dada a rede $\mathbf{A}$, onde obtemos as s-ésimas estimativas dos parâmetros $\beta = (\Lambda, p)$.

$$Q(\beta, \beta^{(s)}) = \mathrm{E}_{\mathbf{W}|\mathbf{A},\beta,\pi}[\mathscr{E}(\beta; \tau, \mathbf{A})].$$

- Passo-M: encontrar $\beta^{(s+1)}$ que maximiza o valor esperado computado no Passo-E:

$$\beta^{(s+1)} = \mathrm{argmax}_{\beta} Q(\beta \mid \beta^{(s)}).$$

Demonstramos a seguir como o algoritmo EM funciona para o modelo ZIP. Para o modelo global, inicialmente temos que:

$$\begin{aligned}\mathscr{E}(\beta;\tau,\mathbf{A}) &= \sum_{1<i,j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left[(W_{ij}^{ab}\log\tfrac{p}{1-p})+(1-W_{ij}^{ab})\log\tfrac{e^{-\lambda_{ab}}\lambda_{ab}^{a_{ij}}}{a_{ij}!}+\log(1-p)\right]\\ &= \sum_{1<i,j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left[W_{ij}^{ab}\left(\log(p)-\log(1-p)\right)+\log(1-p)+\log\left(\frac{e^{-\lambda_{ab}}\lambda_{ab}^{a_{ij}}}{a_{ij}!}\right)(1-W_{ij}^{ab})\right]\\ &= \sum_{1<i,j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left[W_{ij}^{ab}\log\left(\frac{p}{1-p}\right)+\log(1-p)+(1-W_{ij}^{ab})\left(\frac{e^{-\lambda_{ab}}\lambda_{ab}^{a_{ij}}}{a_{ij}!}\right)\right],\end{aligned} \tag{3.20}$$

onde $(\Lambda, p) \in \mathbb{R}_+ \times [0,1]$ são os parâmetros subjacentes.

Ao aplicar a esperança, temos:

$$\begin{aligned}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta,\pi}[\mathscr{E}(\beta;\tau,\mathbf{A})] &= \sum_{1<i,j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left[\mathrm{E}_{\mathbf{W}|A,\beta}(W_{ij}^{ab})\log\left(\frac{p}{1-p}\right)\right]+\\ &+\sum_{1<i,j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left[\left(1-\mathrm{E}_{\mathbf{W}|A,\beta}(W_{ij}^{ab})\right)\log\left(\frac{e^{-\lambda_{ab}}\lambda_{ab}^{a_{ij}}}{a_{ij}!}\right)+\log(1-p)\right].\end{aligned} \tag{3.21}$$

Logo, no Passo-E, inicialmente calcula-se os valores esperados de W_{ij}^{ab}, assim temos que:

$$\mathrm{E}_{\mathbf{W}|A,\beta}(W_{ij}^{ab}) = \mathbb{P}(W_{ij}^{ab}=1 \mid A_{ij}=a_{ij},\beta) = \frac{p\mathbb{1}_{\{a_{ij}=0\}}}{p\mathbb{1}_{\{a_{ij}=0\}}+(1-p)e^{-\lambda_{ab}}}. \tag{3.22}$$

Para encontrar o estimador de p, é necessário substituir (3.22) na equação (3.21) e, em seguida, derivar a equação resultante em relação a p:

$$\begin{aligned}\frac{\partial}{\partial p}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}[\mathscr{E}(\beta;\tau,\mathbf{A})] = \sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\mathrm{E}_{\mathbf{W}|A,\beta}(W_{ij}^{ab})\frac{1}{p(1-p)}-\\ -\frac{\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}}{(1-p)}.\end{aligned} \tag{3.23}$$

igualando a zero o resultado obtido em (3.23), temos:

$$\begin{aligned}&\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\frac{1}{p(1-p)}-\frac{\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}}{(1-p)}=0\\ &\frac{\sum_{1<i,j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})}{p} = \sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}.\end{aligned} \tag{3.24}$$

Ao solucionar a equação anterior, podemos obter o estimador para p dado por:

$$\hat{p} = \frac{\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})}{\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}}.$$

Para maximizar Λ, é necessário derivar a equação (3.21) em relação a λ_{ab}. Dessa forma, obtém-se:

$$\begin{aligned}
\frac{\partial}{\partial\lambda_{ab}}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}[\mathscr{E}(\beta;\tau,\mathbf{A})] &= \frac{\partial}{\partial\lambda_{ab}}\left[\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)\log\left(\frac{e^{-\lambda_{ab}}\lambda_{ab}^{a_{ij}}}{a_{ij}!}\right)\right] \\
&= \frac{\partial}{\partial\lambda_{ab}}\left[\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)\left[-\lambda_{ab}+a_{ij}\log(\lambda_{ab})-\log(a_{ij}!)\right]\right] \\
&= \sum_{1<i<j<n}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij})\right)\left[-1+\frac{a_{ij}}{\lambda_{ab}}\right].
\end{aligned} \tag{3.25}$$

Igualando a zero temos:

$$\begin{aligned}
&\sum_{1<i<j<n}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)\left[-1+\frac{a_{ij}}{\lambda_{ab}}\right]=0 \\
&\frac{\sum_{1<i<j<n}\tau_{ia}\tau_{jb}a_{ij}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)}{\lambda_{ab}}-\sum_{1<i<j<n}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)=0.
\end{aligned} \tag{3.26}$$

Logo, o estimador para λ_{ab} resulta em:

$$\hat{\lambda}_{ab}=\frac{\sum_{1<i<j<n}a_{ij}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)}{\sum_{1<i<j<n}\tau_{ia}\tau_{jb}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta}(W_{ij}^{ab})\right)}. \tag{3.27}$$

Ao obter essas expressões podemos começar a entender o funcionamento do nosso algoritmo para o modelo global.

- Passo-E: calculamos o valor esperado de W_{ij}^{ab} no passo s da seguinte forma:

$$W_{ij}^{ab(s)}=\frac{p^{(s)}\mathbb{1}_{\{a_{ij}=0\}}}{p^{(s)}\mathbb{1}_{\{a_{ij}=0\}}+(1-p^{(s)})e^{-\lambda_{ab}^{(s)}}}. \tag{3.28}$$

na expressão:

$$\begin{aligned}
Q(\beta,\beta^{(s)})=\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})&=\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}W_{ij}^{ab(s)}\left[\log\left(\frac{p^{(s)}}{1-p^{(s)}}\right)-\log\left(\frac{e^{-\lambda_{ab}^{(s)}}\lambda_{ab}^{a_{ij}(s)}}{a_{ij}!}\right)\right] \\
&+\sum_{1<i<j<n}\sum_{a,b=1}^{K}\tau_{ia}\tau_{jb}\log(1-p^{(s)}).
\end{aligned}$$

em que $W_{ij}^{ab(s)}$ é uma função de $\beta^{(s)}$.

- No Passo-M, definimos $p^{(s+1)}$ e $\lambda_{ab}^{(s+1)}$ sendo a função que maximiza cada um deles, que encontramos acima, então temos que:

$$\hat{p}^{(s+1)} = \frac{\sum_{1<i<j<n} \sum_{a,b=1}^{K} \tau_{ia}^{(s)} \tau_{jb}^{(s)} \mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})}{\sum_{1<i<j<n} \sum_{a,b=1}^{K} \tau_{ia}^{(s)} \tau_{jb}^{(s)}}. \tag{3.29}$$

e para Λ temos:

$$\hat{\lambda}_{ab}^{(s+1)} = \frac{\sum_{1<i<j<n} a_{ij} \tau_{ia}^{(s)} \tau_{jb}^{(s)} \left(1 - \mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})\right)}{\sum_{1<i<j<n} \tau_{ia}^{(s)} \tau_{jb}^{(s)} \left(1 - \mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})\right)}. \tag{3.30}$$

o algoritmo para a estimação do parâmetro β, para o modelo global, funcionam da seguinte maneira:

Algoritmo 1- EM para maximizar $\beta = (\lambda_{ab}, p)$: modelo global	
Entrada:	$A_{ij}, 1 \leq i < j \leq n$. Número de comunidades: K.
Saída:	Os parâmetros do modelo $\beta = (\lambda_{ab}, p)$.

1) Inicialização:	definir s=1;
	definir $\Lambda^{(1)} = \Lambda^{(0)}$ (valor inicial);
	$p^{(t)} = p^0$ (valor inicial);
	ε pequeno, como critério de parada.

2)	definir:$s = t + 1$
3) Passo-E:	Calcular $\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})$ de acordo com (3.28)
4) Passo-M	Calcular $\beta^{(s)}$, onde: 4.1) $p^{(s)}$ é calculado de acordo com (3.29) 4.2) $\lambda_{ab}^{(s)}$ é calculado de acordo com (3.30), :
5)	Iterar os passos 3 e 4 ate que: $\left\|p^{(s)} - p^{(s+1)}\right\| < \varepsilon$ e $\left\|\lambda_{ab}^{(s)} - \lambda_{ab}^{(s+1)}\right\| < \varepsilon$, onde ocorre a convergência caso contrário, volte ao passo 2.
7)	Definir o máximo de p e λ_{ab}
8)	retornar $\hat{p}$ e $\hat{\lambda}_{ab}$

Assim, o algoritmo para estimar as comunidades do modelo global funciona da seguinte maneira:

Algoritmo 2- EM Variacional: modelo global	
Entrada:	$A_{ij}, 1 \leq i < j \leq n$. Número de comunidades: K Valor inicial: τ_0
Saída:	Vetor de comunidades estimadas $\hat{\mathbf{Z}}_i = (Z_{i1}, \ldots, Z_{iK})$; Os parâmetros estimados do modelo $\hat{\theta} = (\hat{\pi}, \hat{\Lambda}, \hat{p})$.
1) Inicialização:	definir t=0; definir $\tau^{(0)} = \tau_0$; $\mathfrak{L}(Q) = 0$; ε pequeno, como critério de parada.
2)	definir:$t = t + 1$
3) Passo-E:	Calcular $\tau^{(t)}$ de acordo com (3.10)
4) Passo-M	Calcular $\theta^{(t)}$, onde: 4.1) $\pi^{(t)}$ é calculado de acordo com (3.16) 4.2) Usar Algoritmo 1 para obter os parâmetros $\beta = (\lambda_{ab}, p)$, :
5)	Iterar os passos 3 e 4 ate que: $\left\lvert \mathfrak{L}(Q)^{(t)} - \mathfrak{L}(Q)^{(t+1)} \right\rvert < \varepsilon$, onde ocorre a convergência caso contrário, volte ao passo 2.
7)	Define $\hat{\mathbf{Z}}_i = \max_{K=1,\cdots,k} \tau_{iK}^{(t)}$
8)	retorna os valores de $\hat{\mathbf{Z}}, \theta^t$

Para estimar o parâmetro β no modelo local, temos que $\mathbf{W}$ é um conjunto de variáveis latentes que depende das comunidades, W_{ij}^{ab}, $1 < i < j < n$ e $1 < a, b < K$, que indica se os zeros observados em $\mathbf{A}$ segue distribuição de Poisson ou Bernoulli, condicionado as comunidades $\mathbf{Z}$. Então, reescrevemos (2.6) usando as variáveis latentes W_{ij}^{ab} como

$$f_{ab}(a_{ij}) = [p_{ab}\mathbb{1}_{\{a_{ij}=0\}}]^{W_{ij}^{ab}} \left[(1-p_{ab}) \frac{\lambda_{ab}^{a_{ij}} e^{-\lambda_{ab}}}{a_{ij}!} \right]^{1-W_{ij}^{ab}} \tag{3.31}$$

$a_{ij} \in \{0, 1, \ldots\}$ e aplicando o logaritmo, temos

$$\log f_{ab}(a_{ij}) = W_{ij}^{ab} \log \left(\frac{p_{ab}}{1-p_{ab}} \right) + (1 - W_{ij}^{ab}) \log \left(\frac{\lambda_{ab}^{a_{ij}} e^{-\lambda_{ab}}}{a_{ij}!} \right) + \log(1-p_{ab}). \tag{3.32}$$

E para estimar os parâmetros $\beta = (\lambda_{ab}, p_{ab})$ seguem de maneira análoga ao modelo global, o algoritmo EM para estimar o parâmetro β, do modelo local, funciona da seguinte maneira:

- Passo-E: calculamos o valor esperado de W_{ij}^{ab} no passo s da seguinte maneira:

$$W_{ij}^{ab(s)} = \frac{p_{ab}^{(s)} \mathbb{1}_{\{a_{ij}=0\}}}{p_{ab}^{(s)} \mathbb{1}_{\{a_{ij}=0\}} + (1 - p_{ab}^{(s)}) e^{-\lambda_{ab}^{(s)}}}. \tag{3.33}$$

na expressão:

$$Q(\beta,\beta^{(s)}) = \mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab}) = \sum_{1<i<j<n}\sum_{a,b=1}^{K} \tau_{ia}\tau_{jb}W_{ij}^{ab(s)}\left[\log\left(\frac{p_{ab}^{(s)}}{1-p_{ab}^{(s)}}\right) - \log\left(\frac{e^{-\lambda_{ab}^{(s)}}\lambda_{ab}^{a_{ij}(s)}}{a_{ij}!}\right)\right]$$
$$+ \sum_{1<i<j<n}\sum_{a,b=1}^{K} \tau_{ia}\tau_{jb}\log(1-p_{ab}^{(s)}).$$

em que $W_{ij}^{ab(s)}$ é uma função de $\beta^{(s)}$

- No Passo-M, definimos $p_{ab}^{(s+1)}$ e $\lambda_{ab}^{(s+1)}$ sendo a função que maximiza cada um deles, que encontramos acima, então temos que:

$$\hat{p}_{ab}^{(s+1)} = \frac{\sum_{1<i<j<n}\tau_{ia}^{(s)}\tau_{jb}^{(s)}\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})}{\sum_{1<i<j<n}\tau_{ia}^{(s)}\tau_{jb}^{(s)}}. \tag{3.34}$$

e para Λ temos:

$$\hat{\lambda}_{ab}^{(s+1)} = \frac{\sum_{1<i<j<n}a_{ij}\tau_{ia}^{(s)}\tau_{jb}^{(s)}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})\right)}{\sum_{1<i<j<n}\tau_{ia}^{(s)}\tau_{jb}^{(s)}\left(1-\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})\right)}. \tag{3.35}$$

O algoritmo para a estimação do parâmetro β, para o modelo local, funciona da seguinte maneira:

Algoritmo 3- EM para maximizar $\beta = (\lambda_{ab}, p_{ab})$: modelo local	
Entrada:	$A_{ij}, 1 \leq i < j \leq n$. Número de comunidades: K.
Saída:	Os parâmetros do modelo $\beta = (\Lambda, p_{ab})$.

1) Inicialização:	definir s=1;
	definir $\Lambda^{(1)} = \Lambda^0$ (valor inicial);
	$\mathbf{p}^{(t)} = \mathbf{p}^0$ (valor inicial);
	ε pequeno, como critério de parada.

2)	definir:$s = t+1$
3) Passo-E:	Calcular $\mathrm{E}_{\mathbf{W}|\mathbf{A},\beta^{(s)}}(W_{ij}^{ab})$ de acordo com (3.33)
4) Passo-M	Calcular $\beta^{(s)}$, onde: 4.1) $p^{(s)}$ é calculado de acordo com (3.34) 4.2) $\lambda_{ab}^{(s)}$ é calculado de acordo com (3.35), :
5)	Iterar os passos 3 e 4 ate que: $\left\|p_{ab}^{(s)} - p_{ab}^{(s+1)}\right\| < \varepsilon$ e $\left\|\lambda_{ab}^{(s)} - \lambda_{ab}^{(s+1)}\right\| < \varepsilon$, onde ocorre a convergência caso contrário, volte ao passo 2.
7)	Definir o máximo de p_{ab} e λ_{ab}
8)	retornar $\hat{p}_{ab}$ e $\hat{\lambda}_{ab}$

O algoritmo para estimar as comunidades do modelo local funciona da seguinte maneira:

	Algoritmo 4- EM Variacional: modelo local
Entrada:	$A_{ij}, 1 \leq i < j \leq n$. Número de comunidades: K. Valor inicial:τ_0
Saída:	Vetor de comunidades estimadas: $\hat{\mathbf{Z}}_i = (Z_{i1}, \ldots, Z_{iK})$; Os parâmetros estimados do modelo $\hat{\theta} = (\hat{\pi}, \hat{\Lambda}, \hat{p}_{ab})$.
1) Inicialização:	definir t=0; definir $\tau = \tau_0$; $\mathfrak{L}(Q)^{(t)} = 0$; ε pequeno, como critério de parada.
2)	definir:$t = t + 1$
3) Passo-E:	Calcular $\tau^{(t)}$ de acordo com (3.10)
4) Passo-M	Calcular $\theta^{(t)}$, onde: 4.1) $\pi^{(t)}$ é calculado de acordo com (3.16) 4.2) Usar Algoritmo 3 para obter os parâmetros $\beta = (\lambda_{ab}, p)$, :
5)	Iterar os passos 3 e 4 até que: $\left\lvert \mathfrak{L}(Q)^{(t)} - \mathfrak{L}(Q)^{(t+1)} \right\rvert < \varepsilon$, onde ocorre a convergência caso contrário, volte ao passo 2.
7)	Define $\hat{\mathbf{Z}}_i = \max_{K=1,\cdots,k} \tau_{iK}^{(t)}$
8)	retorna os valores de $\hat{\mathbf{Z}}, \theta^t$

CAPÍTULO

4

SIMULAÇÃO

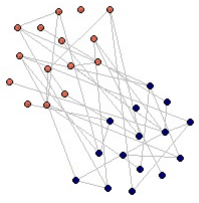
Nesse capítulo, realizamos um estudo de simulação para avaliar o comportamento do grafo em relação ao modelo proposto. Verificamos o desempenho do algoritmo em relação aos parâmetros do modelo e por fim, avaliamos a precisão das estimativas com auxílio do Software R.

4.1 Comportamento do modelo

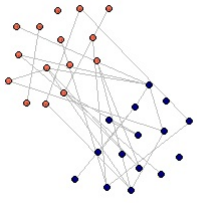
Para avaliar o comportamento do grafo em relação ao modelo proposto, iniciamos com uma simulação do modelo SBM-ZIP local e global para redes não direcionadas e ponderadas. Nas simulações, o número de vértices n foi fixado em 30 para facilitar a visualização da rede. Para todas as simulações, o grafo foi gerado com duas comunidades, e a probabilidade de um vértice pertencer a uma determinada comunidade foi definida como $\pi_a = \pi_b = 0.5$.

Na Figura 1a e na Figura 1b, apresentamos as redes geradas pelo modelo global, em que o parâmetro p é fixado em 0.9. É importante notar que os pesos das conexões dentro dos grupos são iguais entre si, ou seja, $\lambda_{11} = \lambda_{22}$, enquanto a diferença entre os grupos é dada por $\lambda_{12} = \lambda_{21}$.

Figura 1 – Avaliando o comportamento do modelo global.



(a) Comportamento do modelo global ao fixarmos p=0.9, onde o peso das conexões dentro dos grupos é $\lambda_{11} = \lambda_{22} = 2$ e a diferença entre os grupos de $\lambda_{12} = \lambda_{21} = 4$.



(b) Comportamento do modelo global ao fixarmos p=0.9, onde o peso das conexões dentro dos grupos é $\lambda_{11} = \lambda_{22} = 4$ e a diferença entre os grupos é dada por $\lambda_{12} = \lambda_{21} = 2$.

Na Figura 1a, a diferença entre os grupos é maior, com $\lambda_{12} = \lambda_{21} = 4$, enquanto que na Figura 1b, a diferença entre os grupos é menor, com $\lambda_{12} = \lambda_{21} = 2$. Assim, podemos observar que na Figura 1a há uma conexão maior entre vértices de comunidades diferentes em relação às conexões entre vértices da mesma comunidade. Já na Figura 1b, ocorre o oposto, com um maior peso nas arestas entre vértices que pertencem ao mesmo grupo.

Na Figura 2, podemos observar o comportamento da rede gerada pelo modelo global com pesos das arestas maiores entre grupos, representado por $\lambda_{12} = \lambda_{21} = 12$, e com pesos das arestas dentro das comunidades iguais a $\lambda_{11} = \lambda_{22} = 2$. O parâmetro p foi variado em 0.2, 0.5, e 0.9, respectivamente. Podemos notar que, à medida que p aumenta, o número de arestas diminui, resultando em redes mais esparsas com maior conexão entre as comunidades.

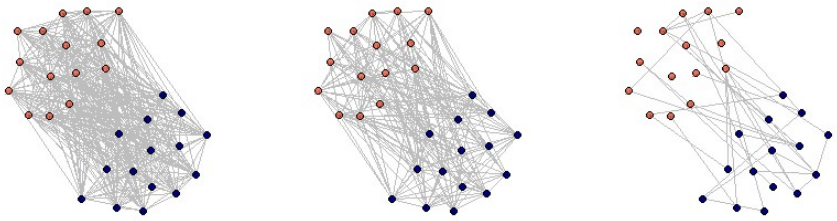


Figura 2 – Comportamento do modelo global ao fixarmos o parâmetro Λ, com valores de $\lambda_{11} = \lambda_{22} = 2$ e $\lambda_{12} = \lambda_{21} = 12$, e variarmos o parâmetro p entre 0.2, 0.5 e 0.9, respectivamente.

Na Figura 3, também estamos avaliando um modelo global com variação no parâmetro p, mas com uma configuração diferente de pesos nas arestas em comparação com a Figura 2. Neste caso, o peso das arestas é maior dentro dos grupos ($\lambda_{11} = \lambda_{22} = 12$) e menor entre os grupos ($\lambda_{12} = \lambda_{21} = 2$). Assim como na Figura 2, à medida que p aumenta, o número de arestas diminui, gerando redes mais esparsas. Além disso, observamos que os vértices que pertencem à mesma comunidade apresentam maior conexão do que aqueles que pertencem a comunidades diferentes.

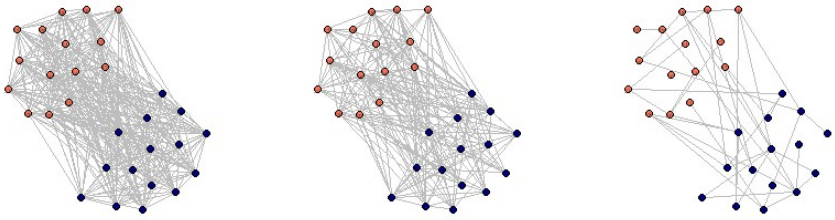


Figura 3 – Comportamento do modelo global ao fixarmos o parâmetro Λ, com valores de $\lambda_{11} = \lambda_{22} = 12$ e $\lambda_{12} = \lambda_{21} = 2$, enquanto variamos os valores de p para 0.2, 0.5 e 0.9, respectivamente.

Na Figura 4, podemos observar um cenário em que aplicamos o modelo local, com fixação de p_{ab} e com $p_{aa} = p_{bb} = 0.75$, gerando mais conexões entre vértices pertencentes à mesma comunidade. Por outro lado, $p_{ab} = p_{ba} = 0.95$, resultando em uma menor presença de

arestas entre vértices de diferentes comunidades. Neste caso, o peso das arestas dentro dos grupos é $\lambda_{11} = \lambda_{22} = 14$, enquanto o peso entre os grupos é $\lambda_{12} = \lambda_{21} = 2$.

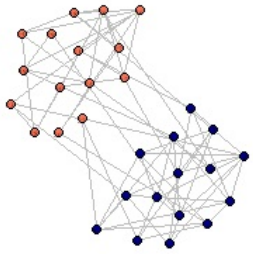


Figura 4 – Configuração dos parâmetros fixados para o modelo local: $p_{aa} = p_{bb} = 0.75$ e $p_{ab} = p_{ba} = 0.95$, $\lambda_{11} = \lambda_{22} = 14$ (peso das arestas dentro dos grupos) e $\lambda_{12} = \lambda_{21} = 2$ (peso das arestas entre os grupos)

A partir desta análise é possível avaliar o comportamento do parâmetro p, responsável por parte da proporção de zeros do modelo. Conseguimos observar que quanto maior o valor de p, mais esparsa se torna a rede. Além disso, é possível notar que as conexões entre vértices que pertencem à mesma comunidade são mais frequentes do que entre vértices de grupos diferentes. Vale ressaltar que o número esperado de arestas na rede é influenciado pelo parâmetro p, que por sua vez, tem impacto na força dos vértices. Ou seja, quanto maior o valor de p, menor a probabilidade de existir uma aresta entre dois vértices. Por fim, é importante destacar que a existência de pesos maiores em arestas entre vértices que pertencem à mesma comunidade é determinada pelo parâmetro Λ. No entanto, nas figuras apresentadas, os pesos das arestas não foram representados visualmente para tornar a imagem mais clara.

4.2 Detecção de Comunidades

Para avaliar o desempenho do algoritmo em relação ao procedimento de estimação dos parâmetros do modelo, foram conduzidas simulações utilizando 100 réplicas de redes em várias configurações de parâmetros e números de vértices, fixando o número de comunidades em $k = 2$, e considerando o cenário em que temos pouca informação sobre as comunidades verdadeiras, representado pelo parâmetro τ_0. As análises foram realizadas para o modelo local.

Para avaliar a precisão da estimação em cada cenário do algoritmo proposto, utilizou-se a medida de comparação conhecida como Informação Mútua Normalizada (NMI), baseada em informação mútua. Nessa medida, avaliamos o particionamento da rede e, assim, conseguimos descrever a semelhança entre uma comunidade estimada e a real. Quanto mais os particionamentos combinarem, melhor será o desempenho do algoritmo. Seguindo as notações de (YANG;

ALGESHEIMER; TESSONE, 2016) para definir o NMI, temos:

$$\begin{aligned} NMI(P,\tilde{P}) &= \frac{I(P,\tilde{P})}{\frac{1}{2}\left[H(P)+H(\tilde{P})\right]} \\ &= \frac{-\sum_{i=1}^{C}\sum_{j=1}^{\tilde{C}} N_{ij}\log\left(\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}}\right)}{\frac{1}{2}\left[\sum_{i=1}^{C} N_{i\cdot}\log\left(\frac{N_{i\cdot}}{N}\right)+\sum_{i=1}^{\tilde{C}} N_{\cdot j}\log\left(\frac{N_{\cdot j}}{N}\right)\right]}. \end{aligned} \tag{4.1}$$

onde P é o particionamento real com C grupos, $\tilde{P}$ é o particionamento estimado com $\tilde{C}$ grupos. Definido N como a matriz de confusão, onde N_{ij} é o número de vértices que estão na comunidade real i e na estimada j. Note que, as linhas (i) de N são as comunidades verdadeiras e as colunas (j) de N são as comunidades encontradas, $N_{i\cdot}$ é a soma da linha i e $N_{\cdot j}$ é a soma da coluna j.

Ao analisar a Equação (4.1), podemos observar que o NMI é uma medida baseada em informação mútua, $I(P,\tilde{P})$, que quantifica a quantidade de informação que temos sobre a variável aleatória em questão, que no nosso caso são as comunidades. Quando essa medida se aproxima de zero, significa que o particionamento real e o estimado não possuem similaridades. Por outro lado, quanto maior a informação mútua, maior é a relação entre os grupos verdadeiros e os estimados, o que indica um melhor desempenho do algoritmo de estimação dos parâmetros.

Note que o NMI também é baseado na entropia de Shannon, $H(P)$, sendo uma medida de incerteza de uma distribuição discreta P. (FORTUNATO, 2010) define essa medida como

$$H(P) = -\sum_{x\in X} \mathbb{P}(x)\log\mathbb{P}(x).$$

onde, e X é o conjunto de valores possíveis da distribuição. Usamos essa medida para normalizar a informação mútua e facilitar a interpretação dos resultados, visto que garante que $NMI \in [0,1]$, além de permitir a comparação mesmo se o número de comunidades estimadas e comunidades reais diferirem.

Desse modo, temos que se duas comunidades são independentes, quando $NMI = 0$ indica que não existe relação entre os agrupamentos. Quando mais próximo de 1 o NMI for, mais similares são as comunidades verdadeiras e simuladas. Portanto, usaremos essa medida nas simulações a seguir para avaliar a precisão do algoritmo ao comparar as comunidades reais com as estimadas.

Seguimos com o estudo de simulação para o modelo local, visando entender o comportamento do modelo ao variar o número de nós e quando aumentamos a diferença na força das conexões entre vértices que pertencem à mesma comunidade. Vale ressaltar que a diferença entre os pesos das arestas dentro e entre as comunidades é controlada pelo parâmetro Λ, o qual pode ser ajustado para aumentar ou diminuir a diferença de pesos. Assim, os parâmetros utilizados para o estudo de simulação são dados pela Tabela 1.

Tabela 1 – Parâmetros usados no estudo de simulação

Informações sobre estudo de simulação	
Número de comunidade:	K=2
Comunidades verdadeiras:	Metade na comunidade 1 e metade na comunidade 2
τ inicial do vetor aleatório da variável latente:	50% de trocas do τ verdadeiro.
Força das conexões: dentro e entre das comunidades:	$\Lambda = \begin{bmatrix} \lambda_{aa} & \lambda_{ab} \\ \lambda_{ba} & \lambda_{bb} \end{bmatrix} = \begin{bmatrix} a & 2 \\ 2 & a \end{bmatrix}$ $a = 2 + diff$, em que $diff = 2,4,6,8,10,12$
Definindo a existência de arestas: dentro e entre as comunidades:	$p_{ab} = \begin{bmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{bmatrix}$
Probabilidade de um vértice pertencer a uma determinada comunidade:	$\pi = (\pi_a, \pi_b)$, definido por: $\pi_a = \pi_b = 0.5$

Simulamos 100 réplicas e utilizamos a medida de Informação Mútua Normalizada (NMI) para avaliar a semelhança entre as comunidades verdadeiras e as estimadas pelo algoritmo proposto. Para isso, variamos o número de conexões entre as comunidades e o número de vértices da rede entre 30, 100, 150 e 250.

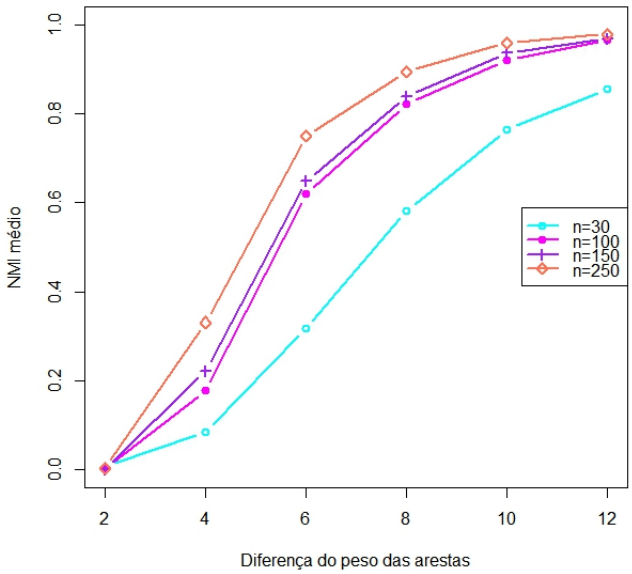


Figura 5 – Média do NMI em função da diferença de pesos das arestas dentro e entre comunidades, para 100 réplicas de redes com número de nós variando entre 30, 100, 150 e 250

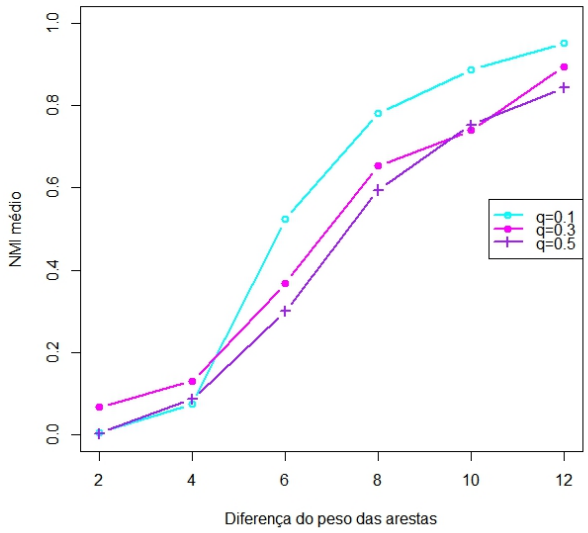
Na Figura 5, é apresentado o valor médio de NMI para diferentes números de vértices. É possível observar que o desempenho do algoritmo melhora à medida que o número de vértices

aumenta, pois a precisão das estimativas aumenta. Além disso, verificou-se que a diferença na força das conexões entre vértices que pertencem à mesma comunidade, controlada pelo parâmetro Λ, tem um impacto significativo nas estimativas. Uma diferença maior na força das conexões resulta em estimativas mais precisas.

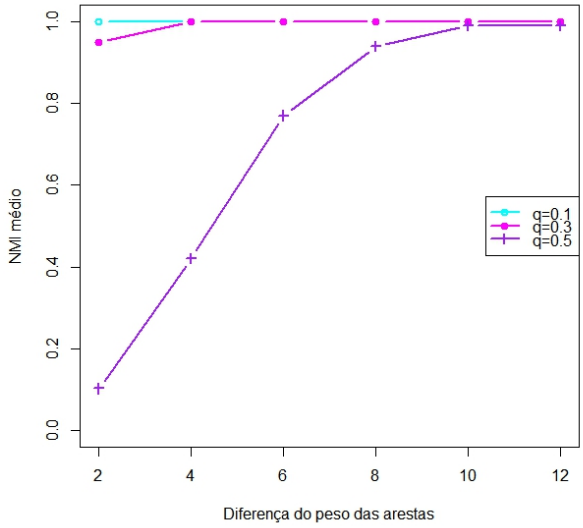
Para avaliar o efeito da escolha do valor inicial τ_0 no algoritmo, criamos uma função q, definida como a proporção de trocas em relação ao τ verdadeiro. Verificou-se o que ocorre quando mudamos 10%, 30% e 50% do τ verdadeiro para usar como τ_0. Isso permitiu avaliar o quão preciso deve ser o valor inicial do algoritmo. A simulação foi realizada para o número de vértices igual a 30 e 200.

Dessa forma, é possível observar que quanto maior o número de vértices, melhor é a estimativa das comunidades, mesmo que a diferença entre o número de arestas entre e dentro das comunidades seja pequena. Especialmente quando o valor inicial τ_0 está próximo do valor verdadeiro τ, obtêm-se boas estimativas, como pode ser observado quando mudamos apenas 10% do valor verdadeiro em um cenário com 200 nós.

Figura 6 – Avaliando o efeito da escolha do τ_0



(a) Resultado do *NMI* em função do aumento da diferença do peso das arestas entre e dentro das comunidades e da proporção de trocas do τ verdadeiro (q), para $n = 30$. Média calculada com 100 réplicas.



(b) Resultado do *NMI* em função do aumento da diferença do peso das arestas entre e dentro das comunidades e da proporção de trocas do τ verdadeiro (q), para $n = 200$. A média foi calculada com 100 réplicas.

Observou-se o comportamento do algoritmo em relação à variação do parâmetro p, para o modelo global, para diferentes tamanhos de rede (30, 50, 100 e 200 nós). O parâmetro p variou nos valores 0, 0.2, 0.4, 0.5, 0.6, 0.8 e 0.9, que influencia na existência das arestas da rede, sendo que quanto mais próximo de 1, mais esparsa é a rede e, consequentemente, temos menos informações sobre o modelo. Para avaliar esse comportamento, foi fixado o parâmetro Λ onde

dentro das comunidades $\lambda_{11} = \lambda_{22} = 12$ e entre as comunidades $\lambda_{12} = \lambda_{21} = 2$.

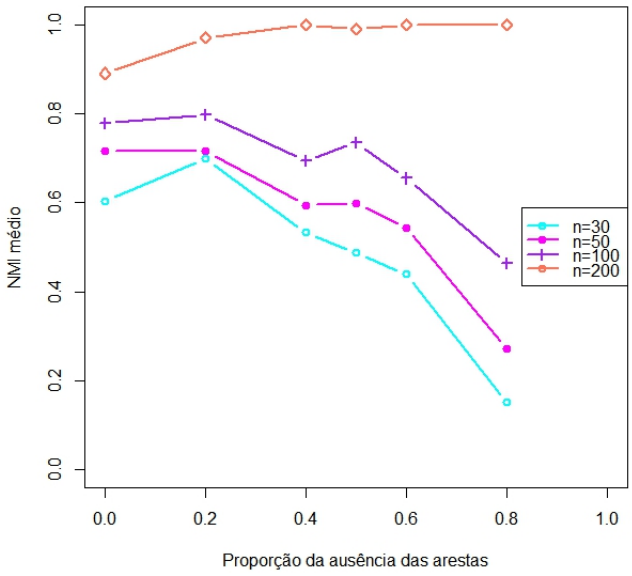


Figura 7 – Resultado do NMI em função da variação do parâmetro p para o modelo global, calculado a média sobre 100 réplicas, com o número de nós variando entre 30, 50 e 100. O parâmetro p influencia na existência de arestas no modelo e varia entre 0, 0.2, 0.4, 0.5, 0.6, 0.8 e 0.9, enquanto o parâmetro Λ é fixo, sendo $\lambda_{11} = \lambda_{22} = 12$ e $\lambda_{12} = \lambda_{21} = 2$.

Na Figura 7, é possível perceber que quanto menos esparsa é a rede, melhor é a estimação das comunidades, já que há mais informações disponíveis. Além disso, um comportamento interessante é observado: quando o número de vértices é igual a 200, a esparsidade da rede não tem tanta influência na estimação de comunidades.

Verificou-se, por meio do estudo de simulação, a importância de se ter um número de nós relativamente grande, bem como de se ter diferenças significativas nos pesos das arestas dentro e entre as comunidades. Também foi observada a influência da esparsidade da rede, assim como a importância da escolha de um bom valor inicial para o parâmetro τ, que deve ser o mais próximo possível da comunidade real para obtermos melhores estimativas. No entanto, ao analisar dados reais, não temos acesso a essa informação. Para lidar com esse problema, no Capítulo 5, o τ_0 foi definido como parte de um método de estimação já existente, o Agrupamento Espectral Esférico.

4.3 Estimativas dos parâmetros

Nesta seção, avaliamos a eficiência dos estimadores dos parâmetros do modelo (Λ, p_{ab}, π) por meio do Erro Quadrático Médio (EQM). Essa medida nos permite obter a diferença quadrática entre o valor estimado e o valor real dos parâmetros, permitindo avaliar a qualidade das estimativas.

Avaliamos essa medida para $n = 100$ e $n = 250$, utilizando os mesmos parâmetros da Tabela 1 e τ_0 correspondendo a 50% das trocas do τ verdadeiro, representando o cenário em que sabemos muito pouco sobre as comunidades.

Tabela 2 – Avaliando a eficiência dos estimadores dos parâmetros do modelo Λ, p_{ab}, π por meio do EQM para n= 100 e 200.

		EQM					
	i	2	4	6	8	10	21
n=100	λ_{11}	0.554	0.858	0.914	0.302	0.335	0.220
	$\lambda_{12} = \lambda_{21}$	1.900	5.262	5.652	2.116	2.209	1.707
	λ_{22}	0.998	0.959	0.729	0.390	0.310	0.381
	p_{11}	0.043	0.030	0.013	0.003	0.002	0.001
	$p_{12} = p_{21}$	0.028	0.018	0.008	0.001	0.001	<0.001
	p_{22}	0.046	0.028	0.013	0.002	0.002	<0.001
	π	0.060	0.025	0.006	<0.001	0.001	0.001
n=250	λ_{11}	0.418	0.669	0.696	0.053	0.290	0.002
	$\lambda_{12} = \lambda_{21}$	1.841	5.376	4.574	0.692	1.610	<0.001
	λ_{22}	0.298	0.665	0.491	0.143	0.166	0.002
	p_{11}	0.037	0.028	0.010	0.001	0.001	<0.001
	$p_{12} = p_{21}$	0.038	0.029	0.011	<0.001	0.001	<0.001
	p_{22}	0.038	0.029	0.011	<0.001	0.001	<0.001
	π	0.049	0.018	0.003	<0.001	<0.001	<0.001

A Tabela 2 apresenta os resultados do Erro Quadrático Médio (EQM) para os parâmetros do modelo (Λ, p_{ab}, π). Observamos que, à medida que o número de nós e a diferença entre os pesos das arestas dentro dos grupos aumentam, as estimativas melhoram, principalmente para o parâmetro Λ. No entanto, quando a diferença dos pesos entre as comunidades é pequena, a qualidade do estimador diminui.

Tabela 3 – Avaliando o custo computacional em relação ao tempo médio e média dos passos, para n= 30, 100 e 250

	i	2	4	6	8	10	12
n=30	Tempo médio (segundos)	12.10	11.49	33.54	10.97	8.60	28.96
	Nº médio de passos	83	82	78	33	26	42
n=100	Tempo médio (segundos)	131.73	101.86	51.75	117.40	25.09	21.16
	Nº médio de passos	327	116	57	47	19	16
n=250	Tempo médio (segundos)	644.18	363.78	203.97	81.23	84.79	87.43
	Nº médio de passos	431	177	57	22	21	19

Analisamos o custo computacional para obter as estimativas dos parâmetros do modelo por meio do algoritmo EM. A partir da Tabela 3, é possível observar que, à medida que a diferença dos pesos das arestas entre as comunidades aumenta, ocorre uma convergência mais rápida e o número de passos diminui. Quando o tamanho da rede n aumenta, para diferenças pequenas, o tempo médio e o número de passos para a convergência aumentam. No entanto, à medida que essa diferença cresce, o número médio de passos diminui. O τ_0 utilizado foi de 50% de trocas em relação ao τ verdadeiro.

4.4 Comparação com outros métodos

Realizamos um estudo de simulação para comparar o modelo SBM-ZIP com dois métodos utilizados na detecção de comunidades, o Cluster Louvain e o Agrupamento Espectral Esférico.

O método Cluster Louvain é baseado na medida de modularidade com uma abordagem hierárquica (BLONDEL *et al.*, 2008). Inicialmente, cada vértice é atribuído a uma comunidade própria e, posteriormente, os vértices são movidos para a comunidade em que atingem a maior contribuição para modularidade. O processo acaba quando resta apenas um vértice ou quando a modularidade não pode mais ser aumentada.

O agrupamento Espectral Esférico, proposto por (QIN; ROHE, 2013), é um método utilizado para detecção de comunidades que considera a heterogeneidade dos graus da rede e é bastante utilizado em redes esparsas. Ele é baseado em teoria dos grafos e utiliza o Laplaciano Normalizado de um grafo para detectar as comunidades. No entanto, uma desvantagem tanto do Cluster Louvain quanto do agrupamento Espectral Esférico é que essas metodologias não se baseiam nos critérios da função de verossimilhança.

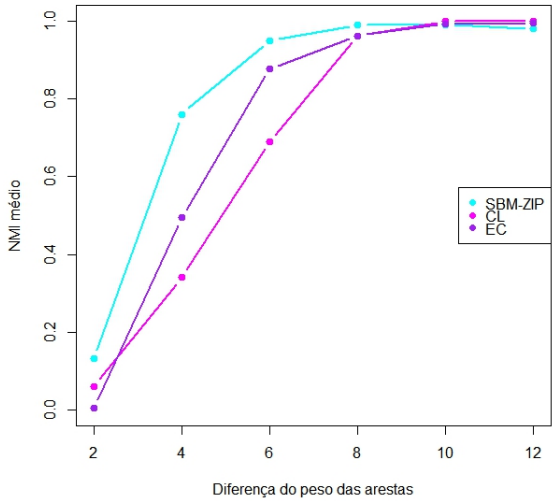
Para avaliar os métodos, utilizamos a medida de comparação Informação Mútua Normalizada (NMI), e avaliamos de duas maneiras. A primeira, utilizando os seguintes parâmetros para n=150.

Tabela 4 – Informações iniciais sobre os parâmetros usados para comparação de modelos

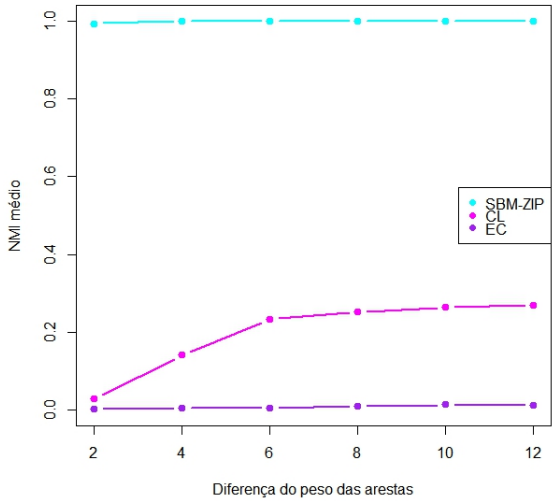
Informações sobre estudo de simulação para comparação de modelos	
Número de comunidade:	K=2
Comunidades verdadeiras:	Metade na comunidade 1 e metade na comunidade 2
τ inicial do vetor aleatório da variável latente:	50% de trocas do τ verdadeiro.
Força das conexões: dentro e entre das comunidades:	$\Lambda = \begin{bmatrix} \lambda_{aa} & \lambda_{ab} \\ \lambda_{ba} & \lambda_{bb} \end{bmatrix} = \begin{bmatrix} a & 2 \\ 2 & a \end{bmatrix}$ $a = 2 + diff$, em que $diff = 2,4,6,8,10,12$.
Definindo a existência de arestas: dentro e entre as comunidades:	$p_{ab} = \begin{bmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{bmatrix}$
Probabilidade de um vértice pertencer a uma determinada comunidade:	$\pi = (\pi_a, \pi_b)$, definido por: $\pi_a = \pi_b = 0.5$

No segundo cenário, estamos lidando com uma rede desbalanceada em que a probabilidade de um vértice pertencer a um determinado grupo é diferente. Isso significa que 80% dos vértices estão na comunidade 1 e os outros 20% estão na comunidade 2.

Figura 8 – Visualização gráfica das comparações dos métodos



(a) Para n=150. Resultado do NMI em função do aumento da diferença do peso das arestas entre e dentro das comunidades comparando os métodos SBM-ZIP, Cluster Louvain (CL) e agrupamento Espectral Esférico (EC), para o caso em que a rede é balanceada



(b) Para n=150. Resultado do NMI em função do aumento da diferença do peso das arestas entre e dentro das comunidades comparando os métodos SBM-ZIP, Cluster Louvain (CL) e agrupamento Espectral Esférico (EC), para o caso em que a rede é desbalanceada

Na Figura 8a, é possível notar que o método SBM-ZIP, o agrupamento espectral e o Cluster Louvain possuem comportamentos semelhantes, mas o SBM-ZIP apresenta um desempenho superior, pois não requer uma grande diferença na força das arestas para alcançar um NMI maior que 0.5 em comparação aos outros dois métodos. Já na Figura 8b, o método SBM-ZIP demonstra um desempenho ainda melhor, mesmo em casos em que a diferença na força das arestas é pequena. Em contrapartida, os outros dois métodos não conseguem alcançar um NMI acima de 0.4, mesmo em situações com grandes diferenças na força das arestas.

Tabela 5 – Avaliando a eficiência dos estimadores dos parâmetros do modelo Λ, p_{ab}, π por meio do EQM para rede balanceada e desbalanceada com n= 150.

Tipo de rede		EQM					
	i	2	4	6	8	10	21
Balanceada	λ_{11}	1.663	1.438	0.561	0.102	0.311	1.580
	$\lambda_{12} = \lambda_{21}$	0.884	0.819	0.501	0.171	0.275	0.781
	λ_{22}	1.683	1.545	0.722	0.299	0.640	0.740
	p_{11}	0.009	<0.001	<0.001	<0.001	<0.001	<0.001
	$p_{12} = p_{21}$	0.005	<0.001	<0.001	<0.001	<0.001	<0.001
	p_{22}	0.015	<0.001	<0.001	<0.001	<0.001	<0.001
	π	0.096	0.009	<0.001	<0.001	<0.001	<0.001
Desbalanceada	λ_{11}	0.031	0.052	0.066	0.083	0.086	0.126
	$\lambda_{12} = \lambda_{21}$	0.003	0.002	0.002	0.003	0.002	0.002
	λ_{22}	0.022	0.019	0.046	0.035	0.036	0.030
	p_{11}	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	$p_{12} = p_{21}$	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
	p_{22}	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
	π	<0.001	<0.001	<0.001	0.082	0.079	0.061

Ao avaliar a eficiência dos estimadores para redes balanceadas e desbalanceadas com 150 vértices, foi possível observar que, para redes balanceadas, o EQM apresenta um aumento moderado em relação ao parâmetro Λ, mas diminui à medida que a diferença do peso das arestas dentro das comunidades aumenta em relação ao peso das arestas entre as comunidades. No caso dos parâmetros p_{ab} e π, o mesmo padrão é observado, embora, mesmo para diferenças pequenas, o EQM já seja bastante próximo de zero.

Já para redes desbalanceadas, o EQM para os parâmetros Λ, p_{ab} e π é próximo de zero. Esses resultados demonstram que o método proposto é eficiente para a detecção de comunidades em redes tanto balanceadas quanto desbalanceadas.

Portanto, esses resultados sugerem que os métodos de estimação são eficazes na identificação dos parâmetros de redes esparsas. Além disso, a eficiência dos estimadores variou de acordo com a configuração dos parâmetros, mas em geral, apresentaram um EQM próximo a zero. Notamos que, por exemplo, a variação do parâmetro Λ teve um impacto maior no EQM em comparação com os parâmetros p_{ab} e π. Também verificamos que o desempenho dos estimadores foi consistente em diferentes configurações de parâmetros.

CAPÍTULO

5

APLICAÇÃO

Nesse capítulo, aplicamos a metodologia apresentada no Capítulo 3 a dados reais. O modelo proposto é usado para estudar o conjunto de dados de transporte aéreo brasileiro dos anos de 2018, 2019, 2020 e 2021 que foram obtidos através do site da Agência Nacional de Aviação Civil (ANAC, 2022). Os resultados numéricos necessários para análise de dados foram realizadas usando o software R.

5.1 Redes de aeroportos

A rede aérea brasileira estudada é composta por voos domésticos, sendo que cada vértice representa um aeroporto do Brasil e as arestas representam as conexões (voos) entre esses aeroportos. Uma aresta é adicionada se houver um voo direto entre dois aeroportos, seguindo a mesma abordagem utilizada em outros estudos (HOSSAIN; ALAM, 2017; DU *et al.*, 2016; REN; LI, 2018).

Certos aspectos da rede aérea, como a existência e a frequência de voos entre aeroportos, são influenciados por fatores complexos, tais como a demanda de passageiros e as estratégias das companhias aéreas. A demanda de passageiros depende de vários fatores socioeconômicos, geográficos e culturais, além de custos de viagem, disponibilidade de voos, entre outros (GEGOV *et al.*, 2013). As companhias aéreas, por sua vez, decidem em quais aeroportos operar com base em fatores como a demanda de passageiros, o custo de operação e a competição com outras companhias aéreas. Portanto, a frequência e a existência de voos entre aeroportos são influenciadas por uma complexa interação de fatores, tornando difícil prever exatamente quais conexões existirão e com que frequência.

Nesse contexto, optamos por considerar os pesos das arestas, que representam o número de voos entre os aeroportos, como aleatórios. Essa escolha reflete a dificuldade em prever exatamente a frequência de voos entre aeroportos, dada a complexidade dos fatores envolvidos.

Além disso, a aleatoriedade dos pesos permite levar em conta a variabilidade natural dos dados, uma vez que o número de voos entre aeroportos pode variar ao longo do tempo devido a fatores como flutuações sazonais na demanda ou mudanças nas estratégias das companhias aéreas.

A rede de transporte aéreo apresenta uma heterogeneidade no grau dos vértices, que pode ser observada através da quantidade de voos partindo de cada aeroporto, resultando em vértices com diferentes probabilidades de se conectar a outros vértices. Essa heterogeneidade pode ser representada pelo conceito de *hubs*, que correspondem a vértices com alta conectividade e alto peso, geralmente associados aos aeroportos com maior fluxo de passageiros. Vale ressaltar que a conectividade é uma medida de topologia da rede, enquanto o peso é uma medida de interação entre as arestas.

Além disso, é comum encontrar redes esparsas, ou seja, com muitas conexões inexistentes, resultando em pares de vértices sem conexões diretas. No caso da rede aérea brasileira, essa esparsidade é evidenciada pelo grande número de aeroportos no país e pela possibilidade de conexões indiretas entre os aeroportos, através de voos com escalas em outros aeroportos.

5.2 Estrutura dos Dados

Para avaliar se houve diferença entre as redes de aeroportos após a pandemia causada pelo COVID-19, foram avaliadas as evoluções das redes aeroportuárias entre os anos de 2018 à 2021, utilizando os dados disponibilizados pela ANAC. As variáveis utilizadas neste trabalho foram:

- Natureza do voo: foram considerados apenas voos domésticos em que a decolagem ocorreu dentro do território brasileiro.
- Ano e mês do voo: foram considerados os dados de cada ano entre 2018 e 2021, divididos por mês.
- Aeroporto de origem e destino: foram coletados os dados de sigla do aeroporto, cidade, estado e região tanto do aeroporto de origem quanto do aeroporto de destino.
- Tipo de voo: foram considerados apenas os voos regulares, que são os voos regulamentados.
- Decolagens: foram coletados o número de decolagens entre os aeroportos de origem e destino.

Essas variáveis foram utilizadas em um estudo realizado por (AMANCIO, 2021).

Os dados iniciais possuem arestas direcionadas, já que a ida do aeroporto A para B difere da ida de B para A. Para o modelo em estudo, que é uma rede não direcionada, é necessário somar todos os voos que saíram do aeroporto i para j com os que saíram do aeroporto j para i.

Matematicamente, a soma da matriz de adjacência com sua transposta resulta em uma matriz simétrica, criando um trajeto (rota) onde a rede é não direcionada. Como estamos interessados em investigar redes ponderadas, consideramos o número de voos que circulam entre quaisquer dois aeroportos. Isso nos fornece informações sobre a conectividade entre aeroportos e a dinâmica do tráfego na rede.

Na Tabela 6, é possível observar que para cada ano, o número de aeroportos e a força média do vértice, que representa a quantidade média de voos entre os aeroportos, apresentam variações. Após o início da pandemia, houve uma queda na força média do vértice devido à diminuição no número de passageiros, o que sugere que o fluxo de voos entre os aeroportos foi reduzido. Além disso, nota-se um aumento no número de aeroportos nesse período devido ao ajuste na malha aérea para manter pelo menos uma ligação aérea em todos os estados, como medida emergencial.

Tabela 6 – Número de aeroportos por ano e sua respectiva força média do vértice.

Ano	Nº de aeroportos	Força média do vértice ($\bar{s}$)
2018	138	17427.08
2019	164	17858.05
2020	180	8800.06
2021	163	10922.08

Como um dos objetivos deste estudo é comparar as comunidades estimadas entre os anos analisados, decidimos fixar os aeroportos que aparecem em todos os anos estudados. Como resultado, o número de vértices em cada grafo é igual a 227. Essa escolha garante que a comparação das comunidades estimadas seja feita em uma base consistente e comparável, além de permitir que possamos identificar padrões consistentes na estrutura das redes aéreas ao longo dos anos.

5.3 Detecção de comunidades

Para determinar a estrutura de agrupamento da rede de aeroportos, utilizamos o método proposto no Capítulo 2 para o modelo local, com o objetivo de estimar as comunidades. Inicialmente, estimamos o valor inicial do parâmetro τ utilizando o método Espectral Esférico. Em seguida, fixamos o número de comunidades em $K = 5$, uma vez que o Brasil é dividido em 5 regiões geográficas.

Na Figura 10, são apresentadas as comunidades estimadas para a rede aeroportuária nos anos de 2018, 2019, 2020 e 2021, representadas por grafos, onde cada comunidade é representada por uma cor. É possível notar que a rede é bastante desbalanceada, com muitos aeroportos pertencentes a uma única comunidade, o que pode ser observado em todos os anos.

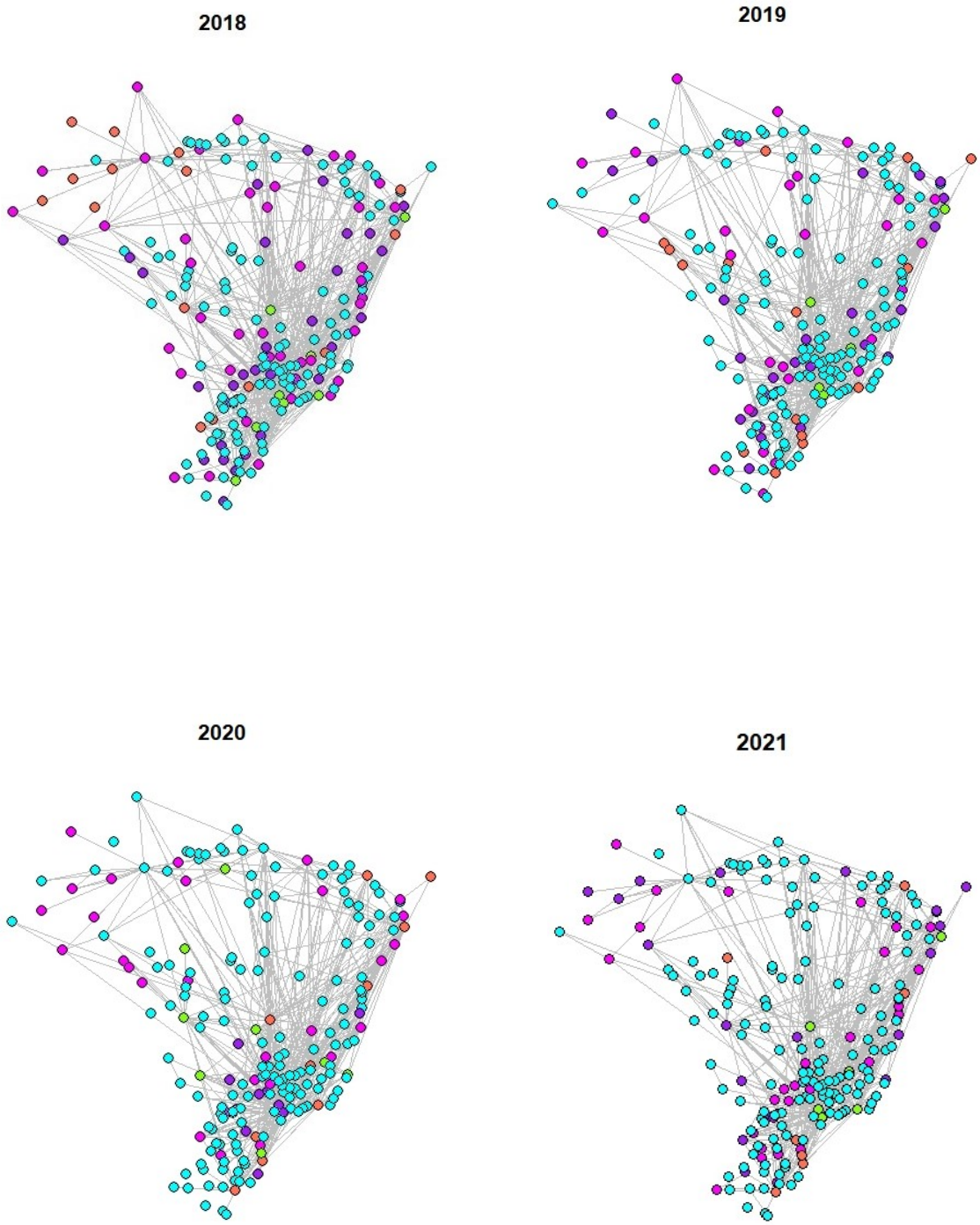


Figura 10 – Comunidades estimadas para redes aeroportuárias nos anos de 2018,2019,2020 e 2021.

Para uma melhor compreensão dos resultados obtidos, realizamos uma análise descritiva dos grupos para cada ano. Na Tabela 7, seguimos a mesma abordagem utilizada no trabalho (AMANCIO, 2021), na qual apresentamos os grupos e suas respectivas cores para cada ano, o número de aeroportos em cada grupo (**Nº**), a força média dos vértices dentro de cada grupo ($\bar{s}$) e o número de aeroportos por região, sendo elas: Norte (**N**), Nordeste (**NE**), Centro-Oeste (**CO**), Sudeste (**SE**) e Sul (**S**).

Tabela 7 – Medidas descritivas das comunidades estimadas.

Ano	Grupo	Cores	Nº	$\bar{s}$	Regiões				
					N	NE	CO	SE	S
2018	1	verde-ciano	124	77.70	18	20	17	38	31
	2	magenta	45	2.35	9	13	8	10	5
	3	roxo	30	83.4	5	8	2	8	7
	4	coral	18	352.88	10	2	1	2	3
	5	verde-lima	10	193771.4	0	1	1	6	2
2019	1	verde-ciano	148	0.621	23	28	20	47	30
	2	magenta	29	20.06	10	6	3	4	6
	3	roxo	28	411.92	5	6	3	7	7
	4	coral	15	5683.86	4	3	2	1	5
	5	verde-lima	7	144733.4	0	1	1	5	0
2020	1	verde-ciano	164	1.58	28	31	20	49	36
	2	magenta	30	20.93	11	7	2	5	5
	3	roxo	12	256.83	0	2	2	5	3
	4	coral	11	22398	0	4	1	3	3
	5	verde-lima	10	843.4	3	0	4	2	1
2021	1	verde-ciano	161	0.124	29	28	24	48	32
	2	magenta	30	1.86	6	7	0	9	8
	3	roxo	19	272.73	5	6	3	1	4
	4	coral	10	5926.2	2	2	1	1	4
	5	verde-lima	7	101553.4	0	1	1	5	0

Os resultados obtidos pela análise descritiva são bastante interessantes. Para o ano de 2018, por exemplo, o grupo com menor força do grau foi o grupo 2, representado pela cor magenta. Este grupo é composto principalmente por aeroportos localizados na região Nordeste do Brasil, e é composto por voos em aeroportos de pequeno porte, muitos dos quais estão localizados no interior da região. Um exemplo é o aeroporto de Parnaíba, no Piauí, que recebe voos de apenas uma companhia aérea e não necessariamente diariamente, não sendo uma rota muito frequente.

Por outro lado, o grupo com a maior força do grau foi o grupo 5, representado pela cor verde-lima, e quase todos os aeroportos deste grupo pertencem à região Sudeste do Brasil. Estes aeroportos possuem um grande fluxo de voos, como os aeroportos de Guarulhos, Congonhas e Campinas em São Paulo, o aeroporto do Rio de Janeiro, o aeroporto de Brasília e o aeroporto de Recife em Pernambuco, sendo em sua maioria aeroportos internacionais e aeroportos de grande porte.

Em 2019, o que muda é que a maior quantidade de aeroportos com fluxo de voos está presente no grupo 1, representado pela cor verde ciano, com 148 aeroportos. A maioria desses aeroportos está na região Sudeste, com 47 aeroportos. Nesse grupo, encontramos aeroportos de pequeno porte, como os do interior de São Paulo, com rotas como Araraquara e Jundiaí. Já o grupo com maior força nos vértices é o 5, representado pela cor verde-lima, com a maioria dos

aeroportos na região Sudeste. Esse grupo é bem parecido com o de 2018, mas é composto apenas por aeroportos que realizam voos internacionais, como os aeroportos de Confins-MG, Recife-PE e Guarulhos-SP.

Em 2020, ano do início da pandemia que trouxe várias restrições em relação a voos, foi observado um aumento no número de rotas com baixo fluxo de voos, como pode ser visto no grupo 1. No entanto, o mais interessante ocorreu no grupo 4, coral, que possui a maior força do grau, onde as rotas mais frequentes foram para cidades turísticas. No segundo semestre de 2020, houve o início da vacinação e uma flexibilização para voos nacionais, o que não ocorreu para destinos internacionais. Esse foi o único ano em que os aeroportos de São Paulo, como Campinas e Guarulhos, não estavam no grupo com o maior fluxo de voos. Em vez disso, as rotas mais frequentes foram para destinos como Florianópolis-SC, Fortaleza-CE, Salvador-BA e até mesmo Fernando de Noronha-PE, mesmo que a ilha tenha ficado fechada por 5 meses em decorrência do coronavírus, reabrindo somente em outubro. Segundo dados do governo de Pernambuco (NORONHA.PE.GOV, 2021), em 2020, 87% dos visitantes em Fernando de Noronha eram brasileiros. A maioria do grupo está concentrada na região Nordeste, diferentemente dos outros anos, onde os grupos com maior força média do vértice estão concentrados na região Sudeste.

Segundo o (BUTANTAN, 2021), de janeiro até o final de 2021, o país atingiu a marca de 80% da população vacinada e fechou o ano com a retomada das atividades e inclusive com o fim das restrições para voos internacionais. E pela nossa análise, podemos perceber essa retomada ao observarmos o grupo com maior força no vértice nesse ano, o grupo 5, verde-lima. A região Sudeste voltou a ser a maioria nesse grupo, com os mesmos aeroportos do ano de 2019, ou seja, aeroportos de grande porte e internacionais.

As estimativas dos parâmetros do modelo para as redes aeroportuárias de cada ano foram avaliadas e seus resultados são apresentados na Tabela 8. Observamos que, para todos os anos, o parâmetro π é desbalanceado, o que reforça a constatação anterior de que a proporção de aeroportos difere entre os grupos. É importante salientar que, quanto maior for o valor de π dentro de um grupo, maior será a proporção de vértices pertencentes a esse grupo. Esse comportamento é observado em todas as estimativas do grupo 1, verde-ciano.

No que diz respeito ao parâmetro $\mathbf{p}$, as estimativas mais próximas de 1 representam o grupo com as rotas feitas com baixa frequência, em sua maioria, pertencentes ao grupo 1, verde-ciano. Esse resultado é coerente com as observações anteriores, já que esse grupo é composto por aeroportos de pequeno porte, cujas rotas não são frequentes e estão normalmente localizadas em cidades menores. Além disso, podemos notar que os grupos com maiores força media do vértice apresentam estimativas de p mais próximas de zero. Mesmo que essa estimativa ainda seja próxima de 0.5 em alguns grupos, isso não ocorre com o grupo 1, verde-ciano, que possui menor força média do vértice. Nesse grupo, mesmo as conexões entre os aeroportos de maior força média do vértice apresentam uma estimativa próxima de 1, o que indica uma quase inexistência de conexão entre os aeroportos desse grupo.

Tabela 8 – Estimativas dos parâmetros do modelo.

Ano	Parâmetros	Estimativas					
		Grupos:	verde-ciano (1)	magenta (2)	roxo (3)	coral (4)	verde-lima (5)
2018	Λ	1	602.25	194.84	379.24	799.09	6626.75
		2	194.84	5.87	144.66	1671.64	544.57
		3	379.24	144.66	89.35	294.33	2007.61
		4	799.09	1671.64	294.33	211.73	3753.98
		5	6626.75	544.57	2007.61	3753.98	21530.15
	p	1	0.998	0.996	0.990	0.990	0.945
		2	0.996	0.992	0.986	0.984	0.805
		3	0.990	0.986	0.968	0.961	0.627
		4	0.990	0.984	0.961	0.907	0.666
		5	0.945	0.805	0.627	0.666	0.100
	π		0.522	0.222	0.132	0.079	0.044
2019	Λ	1	7.66	226.16	28.06	179.68	412.70
		2	226.16	48.50	1293.59	628.50	1258.02
		3	28.06	1293.59	134.11	675.01	3776.14
		4	179.68	628.50	675.01	2029.95	15446.94
		5	412.70	1258.02	3776.14	15446.94	25328.35
	p	1	0.999	0.998	0.996	0.991	0.941
		2	0.998	0.986	0.973	0.964	0.654
		3	0.996	0.973	0.890	0.830	0.418
		4	0.991	0.964	0.830	0.813	0.485
		5	0.941	0.654	0.418	0.485	0.183
	π		0.648	0.131	0.123	0.066	0.030
2020	Λ	1	14.44	38.63	486.75	302.43	787.70
		2	38.63	19.62	2204.69	685.20	100.78
		3	486.75	2204.69	102.73	17266.63	5483.94
		4	302.43	685.20	17266.63	3329.43	997.60
		5	787.70	100.78	5483.94	997.60	527.12
	p	1	0.999	0.997	0.971	0.963	0.989
		2	0.997	0.964	0.880	0.788	0.906
		3	0.971	0.880	0.791	0.772	0.841
		4	0.963	0.788	0.772	0.388	0.627
		5	0.989	0.906	0.841	0.627	0.840
	π		0.722	0.132	0.052	0.048	0.044
2021	Λ	1	2.23	146.75	269.83	919.41	397.89
		2	146.75	6.99	42.08	146.26	1565.04
		3	269.83	42.08	368.88	470.70	3293.20
		4	919.41	146.26	470.70	1411.00	8567.15
		5	397.89	1565.04	3293.20	8567.15	17771.85
	p	1	0.999	0.999	0.996	0.989	0.934
		2	0.999	0.991	0.757	0.901	0.699
		3	0.996	0.959	0.960	0.757	0.483
		4	0.989	0.901	0.757	0.580	0.257
		5	0.934	0.699	0.483	0.257	0.183
	π		0.705	0.133	0.085	0.044	0.030

Na Tabela 8, também é possível observar que nos parâmetros Λ, que indicam a força das conexões entre os vértices, as maiores estimativas pertencem aos grupos que possuem maior

força média do vértice. Esses grupos são compostos por aeroportos que possuem um peso maior nas conexões, representando as rotas mais frequentes e importantes. Por outro lado, os grupos de menor força média do vértice, como o grupo 1, verde-ciano, que contém aeroportos de menor porte e com rotas menos frequentes, apresentam estimativas menores de Λ, indicando uma conexão menos expressiva entre esses vértices.

É importante destacar que o grupo 5, verde-lima, apresenta as maiores estimativas de lambda em relação aos outros grupos nos anos de 2018, 2019 e 2021. Esse grupo é composto por aeroportos de grande porte e internacionais, o que justifica a sua maior conexão com os vértices de outros grupos. O mesmo ocorre com o grupo coral no ano de 2020. Em resumo, as estimativas de Λ corroboram com as observações realizadas anteriormente em relação a força das conexões de cada grupo na rede aeroportuária.

Comparamos por meio do NMI se as comunidades estimadas entre cada ano apresentavam alguma similaridade. A Tabela 9 apresenta os resultados obtidos.

Tabela 9 – Comparação entre as comunidades por anos, por meio do NMI.

	NMI			
	2018	2019	2020	2021
2018	1	0.231	0.202	0.177
2019	0.231	1	0.265	0.263
2020	0.202	0.265	1	0.271
2021	0.177	0.263	0.271	1

Observamos que as estimativas são bastante diferentes, com a maior similaridade ocorrendo entre os anos de 2020 e 2021, que não chegam a 30% de semelhança. Embora tenhamos encontrado algumas similaridades entre as comunidades que agrupam os aeroportos com grande fluxo de voo e maior força média dos vértices, como vimos nas análises anteriores, as diferenças nas comunidades estimadas entre os anos indicam uma grande variação na estrutura da rede aeroportuária ao longo do tempo.

CAPÍTULO

6

DISCUSSÃO

Nesta dissertação, estudamos um modelo estatístico para redes esparsas e ponderadas com estrutura de comunidades. Uma vez que a estrutura de comunidade é uma variável não observada e, ao lidar com dados reais, frequentemente nos deparamos com redes bastante esparsas, propusemos o modelo SBM-ZIP para detectar comunidades em redes ponderadas, não direcionadas e esparsas. Com este método, agrupamos os vértices que apresentam fortes interações em um grupo, e conseguimos lidar com a esparsidade devido às características da distribuição que permite observações frequentes de zeros.

Nesse modelo, as comunidades são estimadas por meio do algoritmo EM-Variacional, enquanto os parâmetros da distribuição ZIP são estimados por meio do algoritmo EM. Para validar as estimações, realizamos um estudo de simulação no qual avaliamos o efeito na detecção das comunidades ao testar diferentes cenários com mudanças nos parâmetros.

Observamos que o valor inicial atribuído às comunidades é crucial, principalmente em casos de um número pequeno de vértices ou se a diferença nas relações entre vértices do mesmo grupo e vértices de outros grupos for pequena. À medida que o número de vértices aumenta, mais informações são obtidas sobre a rede, o que leva a uma estimativa mais precisa dos parâmetros.

Em nosso estudo de simulação, observamos um resultado interessante: à medida que aumentamos o valor do parâmetro p, responsável por parte da proporção de zeros na rede, torna-se mais difícil recuperar as comunidades. No entanto, notamos que, a partir de um número suficientemente grande de arestas - como vimos em nosso estudo com $n = 200$ - o valor de p já não tem tanta influência na precisão da estimação das comunidades bem como a escolha do τ_0. Para avaliar a eficiência dos estimadores, utilizamos o EQM como medida de comparação e notamos que, à medida que o número de vértices aumenta, essa medida se aproxima de zero, indicando que obtivemos bons resultados na estimação dos parâmetros.

Durante o estudo comparativo entre o método SBM-ZIP e os métodos Cluster Louvain e Espectral Esférico, que não são baseados em critérios de verossimilhança, notamos que, em

uma rede balanceada, onde a probabilidade de um vértice pertencer aos grupos é a mesma, os modelos tiveram desempenhos semelhantes na estimação, embora o SBM-ZIP tenha se destacado. Entretanto, em uma rede desbalanceada, onde a probabilidade de um vértice pertencer a um determinado grupo difere, o modelo SBM-ZIP conseguiu capturar as comunidades com NMI igual a 1, ou seja, as comunidades reais e estimadas foram as mesmas. Em contrapartida, os outros métodos não conseguiram alcançar nem mesmo 50% de similaridade para todos os cenários estudados.

Ao analisar o modelo de redes de aeroportos, notamos que os grupos com menor força média dos vértices são compostos por aeroportos de cidades do interior, de pequeno porte e com poucos voos disponíveis. Por outro lado, os grupos de maior força média dos vértices correspondem aos aeroportos de grande porte, que em sua maioria possuem voos internacionais e estão localizados na região sudeste, ou seja, são considerados os *hubs* da rede em estudo.

Durante o primeiro ano da pandemia do COVID-19, foi observado um destaque dos aeroportos com grande volume de conexões de arestas, conhecidos como *hubs*, localizados em cidades turísticas do nordeste na rede de aeroportos. Isso difere dos anos anteriores à pandemia em que a maioria dos *hubs* estavam localizados no sudeste. No entanto, em 2021, com o avanço da vacinação e a flexibilização das medidas de prevenção à disseminação do vírus, os *hubs* voltou a ser em aeroportos internacionais, predominantemente localizados no sudeste, voltaram a ser os mais importantes na rede.

Por fim, é importante destacar que apesar dos resultados satisfatórios obtidos neste trabalho, é necessário considerar suas limitações. Para obter resultados mais precisos, seria necessário estudar com mais detalhes os procedimentos técnicos utilizados para escolha do número de comunidades. Além disso, em trabalhos futuros, seria interessante realizar uma correção do grau, pois a inclusão desse parâmetro no modelo pode ajudar a capturar a heterogeneidade presente em redes reais, uma variável importante para a detecção de comunidades.

REFERÊNCIAS

AMANCIO, D. G. **Um estudo estatístico sobre as redes de aeroportos no Brasil**. 79 p. Monografia (Trabalho de Conclusão de curso) — Universidade de São Carlos, UFSCAR, São Paulo, 2021. Citado nas páginas 58 e 60.

ANAC. **Principais Medidas do Setor Aéreo Após Início da Pandemia de Covid-19 – Linha do tempo**. 2021. Disponível em: <https://www.gov.br/anac/pt-br/assuntos/coronavirus/anac-covid-19-linha-do-tempo>. Acesso em: 15 jan. 2023. Citado na página 24.

____. **Agência Nacional de Aviação Civil**. 2022. Disponível em: <https://www.gov.br/anac/pt-br>. Acesso em: 22 set. 2022. Citado na página 57.

BAGLER, G. Analysis of the airport network of india as a complex weighted network. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 387, n. 12, p. 2972–2980, 2008. Citado na página 23.

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of statistical mechanics: theory and experiment**, IOP Publishing, v. 2008, n. 10, p. P10008, 2008. Citado na página 53.

BÖHNING, D. Zero-inflated poisson models and ca man: A tutorial collection of evidence. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, Wiley Online Library, v. 40, n. 7, p. 833–843, 1998. Citado na página 27.

BUTANTAN. **Segundo Ano da Pandemia É marcado Pelo Avanço da Vacinação contra covid-19 no Brasil**. 2021. Disponível em: <https://butantan.gov.br/noticias/retrospectiva-2021-segundo-ano-da-pandemia-e-marcado-pelo-avanco-da-vacinacao-contra-covid-19-no-br Acesso em: 30 jan. 2023. Citado na página 62.

COLIZZA, V.; BARRAT, A.; BARTHÉLEMY, M.; VESPIGNANI, A. The role of the airline transportation network in the prediction and predictability of global epidemics. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 7, p. 2015–2020, 2006. Citado na página 23.

DONG, H.; CHEN, N.; WANG, K. Modeling and change detection for count-weighted multilayer networks. **Technometrics**, v. 62, n. 2, p. 184–195, 2020. Citado na página 23.

DU, W.-B.; ZHOU, X.-L.; LORDAN, O.; WANG, Z.; ZHAO, C.; ZHU, Y.-B. Analysis of the chinese airline network as multi-layer networks. **Transportation Research Part E: Logistics and Transportation Review**, Elsevier, v. 89, p. 108–116, 2016. Citado na página 57.

ERDŐS, 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. Citado na página 21.

FORTUNATO, S. Community detection in graphs. **Physics reports**, Elsevier, v. 486, n. 3-5, p. 75–174, 2010. Citado nas páginas 22 e 48.

FURTADO, S.; TóLVOLLI. **Modelagem de sistemas complexos para políticas públicas**. Brasília: Instituto de Pesquisa Econômica Aplicada (Ipea), 2015. Citado nas páginas 21 e 22.

GEGOV, E.; POSTORINO, M. N.; ATHERTON, M.; GOBET, F. Community structure detection in the evolution of the united states airport network. **Advances in Complex Systems**, World Scientific, v. 16, n. 01, p. 1350003, 2013. Citado na página 57.

GUIMERA, R.; AMARAL, L. A. N. Modeling the world-wide airport network. **The European Physical Journal B**, Springer, v. 38, p. 381–385, 2004. Citado na página 23.

HATHAWAY, R. J. Another interpretation of the em algorithm for mixture distributions. **Statistics & probability letters**, Elsevier, v. 4, n. 2, p. 53–56, 1986. Citado na página 33.

HOLLAND, P. W.; LASKEY, K. B.; LEINHARDT, S. Stochastic blockmodels: First steps. **Social networks**, Elsevier, v. 5, n. 2, p. 109–137, 1983. Citado na página 26.

HOSSAIN, M. M.; ALAM, S. A complex network approach towards modeling and analysis of the australian airport network. **Journal of Air Transport Management**, Elsevier, v. 60, p. 1–9, 2017. Citado na página 57.

LEE, C.; WILKINSON, D. J. A review of stochastic block models and extensions for graph clustering. **Applied Network Science**, Springer, v. 4, n. 1, p. 1–50, 2019. Citado nas páginas 23 e 26.

LLOYD, S. Least squares quantization in pcm. **IEEE transactions on information theory**, IEEE, v. 28, n. 2, p. 129–137, 1982. Citado na página 22.

LUKOSEVICIUS, A. P.; MARCHISOTTI, G. G.; SOARES, C. A. P. Overview of complexity: main currents, definitions and constructs. **Systems & Management**, v. 11, n. 4, p. 455–465, 2016. Citado na página 22.

LUXBURG, U. V. A tutorial on spectral clustering. **Statistics and computing**, Springer, v. 17, n. 4, p. 395–416, 2007. Citado na página 23.

MORENO, J. L. Who shall survive?: A new approach to the problem of human interrelations. Nervous and mental disease publishing co, 1934. Citado na página 21.

MOTALEBI, N.; STEVENS, N. T.; STEINER, S. H. Hurdle blockmodels for sparse network modeling. **The American Statistician**, Taylor & Francis, v. 75, n. 4, p. 383–393, 2021. Citado na página 23.

NEAL, R. M.; HINTON, G. E. A view of the em algorithm that justifies incremental, sparse, and other variants. In: **Learning in graphical models**. [S.l.]: Springer, 1998. p. 355–368. Citado na página 33.

NEWMAN, M. E. Analysis of weighted networks. **Physical review E**, APS, v. 70, n. 5, p. 056131, 2004. Citado nas páginas 23 e 29.

NORONHA.PE.GOV. **Fernando de Noronha apresenta fluxo turístico de 2020**. 2021. Disponível em: <https://www.noronha.pe.gov.br/fernando-de-noronha-apresenta-fluxo-turistico-de-2020-2/>. Acesso em: 30 jan. 2023. Citado na página 62.

QIN, T.; ROHE, K. Regularized spectral clustering under the degree-corrected stochastic blockmodel. **Advances in neural information processing systems**, v. 26, 2013. Citado na página 53.

REN, P.; LI, L. Characterizing air traffic networks via large-scale aircraft tracking data: A comparison between china and the us networks. **Journal of Air Transport Management**, Elsevier, v. 67, p. 181–196, 2018. Citado na página 57.

SEN, P.; DASGUPTA, S.; CHATTERJEE, A.; SREERAM, P.; MUKHERJEE, G.; MANNA, S. Small-world properties of the indian railway network. **Physical Review E**, APS, v. 67, n. 3, p. 036106, 2003. Citado na página 23.

SNIJDERS, T. A.; NOWICKI, K. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. **Journal of classification**, Springer, v. 14, n. 1, p. 75–100, 1997. Citado na página 26.

STANLEY, N.; BONACCI, T.; KWITT, R.; NIETHAMMER, M.; MUCHA, P. J. Stochastic block models with multiple continuous attributes. **Applied Network Science**, SpringerOpen, v. 4, n. 1, p. 1–22, 2019. Citado na página 26.

SUARIS, P. R.; KEDEM, G. An algorithm for quadrisection and its application to standard cell placement. **IEEE Transactions on Circuits and Systems**, IEEE, v. 35, n. 3, p. 294–303, 1988. Citado na página 22.

WEISS, R. S.; JACOBSON, E. A method for the analysis of the structure of complex organizations. **American Sociological Review**, JSTOR, v. 20, n. 6, p. 661–668, 1955. Citado na página 22.

WU, Z.; BRAUNSTEIN, L. A.; COLIZZA, V.; COHEN, R.; HAVLIN, S.; STANLEY, H. E. Optimal paths in complex networks with correlated weights: The worldwide airport network. **Physical Review E**, APS, v. 74, n. 5, p. 056104, 2006. Citado na página 23.

YANG, Z.; ALGESHEIMER, R.; TESSONE, C. J. A comparative analysis of community detection algorithms on artificial networks. **Scientific reports**, Springer, v. 6, n. 1, p. 1–18, 2016. Citado na página 48.

APÊNDICE

A

ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA

Neste apêndice, apresentaremos a obtenção dos estimadores por meio do método de máxima verossimilhança para os parâmetros π, p, Λ em um grafo não-direcionado e sem laços, onde p e λ_{ab} são os parâmetros a serem estimados.

$$\mathbb{P}(\mathbf{Z}|\pi) = \prod_{i=1}^{n} \mathbb{P}(\mathbf{Z}_i|\pi) = \prod_{i=1}^{n}\prod_{a=1}^{k} \pi_a^{Z_{ia}}.$$

essa é a probabilidade de pertencer à comunidade de a elevado ao fato de pertencer ou não a comunidade a. A distribuição condicional de $\mathbf{A}$ dado $\mathbf{Z}$ é dada por:

$$\begin{aligned}
\mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p) &= \prod_{1<i,j<n}\prod_{a,b=1}^{k} f_{ab}(A_{ij})^{Z_{ia}Z_{jb}} \\
&= \prod_{1<i<j<n}\prod_{a,b=1}^{k}\left[\left(p+(1-p)e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0\}}}\left((1-p)\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0\}}}\right]^{Z_{ia}Z_{jb}} \\
&= \prod_{1<i<j<n}\prod_{a,b=1}^{k}\left[\left(p+(1-p)e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}}\left((1-p)\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}}\right] \\
&= \prod_{a,b=1}^{k}\left(p+(1-p)e^{-\lambda_{ab}}\right)^{\sum_{1<i,j<n}\mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}}\prod_{1<i,j<n}\prod_{a,b=1}^{k}\left((1-p)\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \qquad \text{(A.1)} \\
&= \prod_{a,b=1}^{k}\left(p+(1-p)e^{-\lambda_{ab}}\right)^{\sum_{1<i,j<n}\mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}}(1-p)^{\sum_{1<i,j<n}\sum_{a,b}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \\
&\left[\prod_{1<i<j<n}\prod_{a,b=1}^{k}\frac{1}{a_{ij}!}\right]^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}}\prod_{a,b=1}^{k}\lambda_{ab}^{\sum_{1<i<j<n}\left(a_{ij}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}\right)} \\
&\prod_{a,b=1}^{k} e^{-\lambda_{ab}\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}}.
\end{aligned}$$

Defina, $n_{ab}(A,Z) = \sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}$ como o número de arestas entre vértices da

comunidade a e b e $n^0_{ab}(A,Z) = \sum_{1<i<j<n} \mathbb{1}_{\{a_{ij}=0, Z_{ia}Z_{jb}=1\}}$ sendo o número de não arestas entre a e b. Além disso, $O_{ab}(A,Z) = \sum\limits_{1<i<j<n} \left(a_{ij} \mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{jb}=1\}} \right)$ como a soma dos pesos das arestas entre vértices das comunidades a e b. Temos que

$$\begin{aligned} \mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p) = &\prod_{a,b=1}^{k} \left(p + (1-p)e^{-\lambda_{ab}} \right)^{n^0_{ab}(A,Z)} (1-p)^{\sum\limits_{a,b} n_{ab}(A,Z)} \\ &\left[\prod_{1<i<j<n} \prod_{a,b=1}^{k} \frac{1}{a_{ij}!} \right]^{\mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{jb}=1\}}} \prod_{a,b=1}^{k} \lambda_{ab}^{O_{ab}(A,Z)} \prod_{a,b=1}^{k} e^{-\lambda_{ab} n_{ab}(A,Z)}. \end{aligned} \tag{A.2}$$

Portanto, escrevemos a distribuição conjunta como

$$\begin{aligned} \mathbb{P}(\mathbf{A},\mathbf{Z}|\pi,p,\Lambda) &= \mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p)\mathbb{P}(Z|\pi) \\ &= \left[\prod_{i=1}^{n} \prod_{a=1}^{k} \pi_a^{Z_{ia}} \right] \prod_{a,b=1}^{k} \left(p + (1-p)e^{-\lambda_{ab}} \right)^{n^0_{ab}(A,Z)} (1-p)^{\sum\limits_{a,b} n_{ab}(A,Z)} \\ &\left[\prod_{1<i<j<n} \prod_{a,b=1}^{k} \frac{1}{a_{ij}!} \right]^{\mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{ij}=1\}}} \prod_{a,b=1}^{k} \lambda_{ab}^{O_{ab}(A,Z)} \prod_{a,b=1}^{k} e^{-\lambda_{ab} n_{ab}(A,Z)}. \end{aligned} \tag{A.3}$$

Conforme a independência condicional, decompomos a log-verossimilhança, para poder estimar os parâmetros

$$\log(\mathbb{P}(\mathbf{A},\mathbf{Z}|\pi,p,\Lambda)) = \log(\mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p)) + \log(\mathbb{P}(Z|\pi)). \tag{A.4}$$

Como inicialmente temos o objetivo de encontrar $\hat{\pi}$, aplicaremos o log em $\mathbb{P}(Z|\pi)$, pois é a parte da equação anterior que depende de π

$$\log(\mathbb{P}(Z|\pi)) = \sum_{i-1}^{n} \sum_{a=1}^{k} Z_{ia} \log(\pi_a). \tag{A.5}$$

Utilizaremos o multiplicador de Lagrange em (A.5) por conta da restrição que $\sum_{a=1}^{k} \pi_a = 1$, já que π é um vetor de probabilidades. Para isso, definimos a função

$$f(\pi_a) = \sum_{i=1}^{n} Z_{ia} \log(\pi_a) - \alpha \left(\sum_{a=1}^{k} \pi_a - 1 \right).$$

derivando a função e igualando a zero, temos:

$$f(\pi_a)' = \frac{\sum_{i=1}^{n} Z_{ia}}{\pi_a} - \alpha = 0 \sum_{i=1}^{n} Z_{ia} = \alpha \pi_a. \tag{A.6}$$

sabe-se que

$$\begin{aligned} 1 = \sum_{a=1}^{k} \pi_a &= \sum_{a=1}^{k} \sum_{i=1}^{n} Z_{ia} \frac{1}{\alpha} \\ \alpha &= \sum_{a=1}^{k} \sum_{i=1}^{n} Z_{ia}. \end{aligned} \tag{A.7}$$

Usando (A.7) em (A.6)

$$\hat{\pi}_a = \frac{\sum_{i=1}^{n} Z_{ia}}{\sum_{a=1}^{k}\sum_{i=1}^{n} Z_{ia}} = \frac{\sum_{i=1}^{n} Z_{ia}}{n}. \tag{A.8}$$

Podemos observar que $\hat{\pi}_a$ só depende das atribuições dada a $\mathbf{Z}$, onde nesse resultado temos a soma de quantos vértices está na comunidade a dividido pelo total de vértices da rede, fazendo sentido, se quero somar qual a probabilidade de alguém está na comunidade a.

Para encontra os estimadores de p e λ_{ab} que maximiza a conjunta, aplicamos o log na Equação (A.4), no primeiro termo, pois apenas ele depende dos parâmetros

$$\begin{aligned}\log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p) = \sum_{a,b=1}^{k} n_{ab}^{0}(A,Z)\log(p+(1-p)e^{-\lambda_{ab}}) + \sum_{a,b=1}^{k} n_{ab}(A,Z)\log(1-p) + \\ \sum_{a,b=1}^{k} O_{ab}(A,Z)\log(\lambda_{ab}) - \sum_{a,b=1}^{k}\lambda_{ab}n_{ab}(A,Z) - \mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{jb}=1\}}\log\left[\prod_{1<i<j<n}\prod_{a,b=1}^{k} a_{ij}!.\right]\end{aligned} \tag{A.9}$$

Derivando (A.9) a respeito de p e λ_{ab},

$$\frac{\partial \log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p)}{\partial p} = \sum_{a,b=1}^{k}\left[n_{ab}^{0}(A,Z)\frac{\left(1-e^{-\lambda_{ab}}\right)}{p+(1-p)e^{-\lambda_{ab}}}\right] - \sum_{a,b=1}^{k}\frac{n_{ab}(A,Z)}{1-p}.$$

Temos que,

$$\begin{aligned}\sum_{a,b=1}^{k} n_{ab}(A,Z) &= \sum_{a.b}^{k}\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{ij}=1\}} = \sum_{a.b}^{k}\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0\}}\mathbb{1}_{\{Z_{ia}Z_{ij}=1\}} \\ &= \sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0\}}\sum_{a.b}^{k}\mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{ij}=1\}} = E(A).\end{aligned}$$

Como $\sum_{a,b=1}^{k}\mathbb{1}_{\{Z_{ia}Z_{ij}=1\}} = 1$ e fazendo $\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0\}} = E(A)$. Sendo $E(A)$ o número de arestas na rede. Então, a derivada em relação a p é

$$\frac{\partial \log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p)}{\partial p} = \sum_{a,b=1}^{k}\left[n_{ab}^{0}(A,Z)\frac{\left(1-e^{-\lambda_{ab}}\right)}{p+(1-p)e^{-\lambda_{ab}}}\right] - \frac{E(A)}{1-p}.$$

e em relação a λ_{ab} é

$$\frac{\partial \log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p)}{\partial \lambda_{ab}} = -\left[n_{ab}^{0}(A,Z)\frac{\left((1-p)e^{-\lambda_{ab}}\right)}{p+(1-p)e^{-\lambda_{ab}}}\right] + \frac{O_{ab}(A,Z)}{\lambda_{ab}} - n_{ab}(A,Z).$$

Igualando as derivadas a zero obtemos o seguinte sistema de equações, para o parametro p, temos

$$E(A) = \sum_{a,b=1}^{k}\left[n_{ab}^{0}(A,Z)\frac{\left(1-e^{-\lambda_{ab}}\right)}{p+(1-p)e^{-\lambda_{ab}}}\right](1-p). \tag{A.10}$$

para o parâmetro λ_{ab}, temos a seguinte equação

$$\frac{O_{ab}(A,Z)}{\lambda_{a,b}} = \left[n^0_{ab}(A,Z) \frac{\left((1-p)e^{-\lambda_{ab}}\right)}{p+(1-p)e^{-\lambda_{ab}}} \right] + n_{ab}(A,Z). \tag{A.11}$$

Assim, uma vez que não se tem uma solução fechada para o problema, é necessário resolver por meio de métodos iterativos.

Para realizar os cálculos considerando o modelo local com p_{ab} e λ_{ab}, temos a seguinte distribuição condicional de $\mathbf{A}$ dado $\mathbf{Z}$:

$$\begin{aligned}
\mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p_{ab}) &= \prod_{1<i,j<n} \prod_{a,b=1}^{k} f_{ab}(A_{ij})^{Z_{ia}Z_{jb}} \\
&= \prod_{1<i<j<n} \prod_{a,b=1}^{k} \left[\left(p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0\}}} \left((1-p_{ab})\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0\}}} \right]^{Z_{ia}Z_{jb}} \\
&= \prod_{1<i<j<n} \prod_{a,b=1}^{k} \left[\left(p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}\right)^{\mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}} \left((1-p_{ab})\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \right] \\
&= \prod_{a,b=1}^{k} \left(p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}\right)^{\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}} \prod_{1<i<j<n} \prod_{a,b=1}^{k} \left((1-p_{ab})\frac{\lambda_{ab}^{a_{ij}}e^{-\lambda_{ab}}}{a_{ij}!}\right)^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \\
&= \prod_{a,b=1}^{k} \left(p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}\right)^{\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}} \prod_{a,b=1}^{k} (1-p)^{\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \\
&\times \left[\prod_{1<i<j<n} \prod_{a,b=1}^{k} \frac{1}{a_{ij}!} \right]^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \prod_{a,b=1}^{k} \lambda_{ab}^{\sum_{1<i<j<n}\left(a_{ij}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}\right)} \\
&\times \prod_{a,b=1}^{k} e^{-\lambda_{ab}\sum_{1<i<j<n}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}}.
\end{aligned} \tag{A.12}$$

Fazendo, $n_{ab}(A,Z) = \sum_{1<i<j<n} \mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}$ como o número de arestas entre vértices da comunidade a e b e $n^0_{ab}(A,Z) = \sum_{1<i,j<n} \mathbb{1}_{\{a_{ij}=0,Z_{ia}Z_{jb}=1\}}$ sendo o número de não arestas entre a. Além disso, b, e $O_{ab}(A,Z) = \sum_{1<i<j<n} \left(a_{ij}\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}\right)$ sendo a soma dos pesos das arestas entre vértices das comunidades a e b . Temos que,

$$\begin{aligned}
\mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p_{ab}) &= \prod_{a,b=1}^{k} \left(p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}\right)^{n^0_{ab}(A,Z)} \prod_{a,b=1}^{k} (1-p_{ab})^{n_{ab}(A,Z)} \\
&\left[\prod_{1<i<j<n} \prod_{a,b=1}^{k} \frac{1}{a_{ij}!} \right]^{\mathbb{1}_{\{a_{ij}\neq 0,Z_{ia}Z_{jb}=1\}}} \prod_{a,b=1}^{k} \lambda_{ab}^{O_{ab}(A,Z)} \prod_{a,b=1}^{k} e^{-\lambda_{ab}n_{ab}(A,Z)}.
\end{aligned} \tag{A.13}$$

Ao aplicar o log na equação anterior, temos

$$\log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p_{ab}) = \sum_{a,b=1}^{k} n^0_{ab}(A,Z)\log(p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}) + \sum_{a,b=1}^{k} n_{ab}(A,Z)\log(1-p_{ab}) + \sum_{a,b=1}^{k} O_{ab}(A,Z)\log(\lambda_{ab}) - \sum_{a,b=1}^{k} \lambda_{ab} n_{ab}(A,Z) - \mathbb{1}_{\{a_{ij}\neq 0, Z_{ia}Z_{jb}=1\}} \log\left[\prod_{1<i<j<n}\prod_{a,b=1}^{k} a_{ij}!.\right] \tag{A.14}$$

Derivando (A.14) a respeito de p_{ab}, temos

$$\frac{\partial \log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p_{ab})}{\partial p_{ab}} = \left[n^0_{ab}(A,Z)\frac{\left(1-e^{-\lambda_{ab}}\right)}{p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}}\right] - \frac{n_{ab}(A,Z)}{1-p_{ab}}.$$

e derivando (A.14) a respeito de λ_{ab}, temos

$$\frac{\partial \log \mathbb{P}(\mathbf{A}|\mathbf{Z}=z,\Lambda,p_{ab})}{\partial \lambda_{ab}} = -\left[n^0_{ab}(A,Z)\frac{\left((1-p_{ab})e^{-\lambda_{ab}}\right)}{p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}}\right] + \frac{O_{ab}(A,Z)}{\lambda_{ab}} - n_{ab}(A,Z).$$

Igualando as derivadas de p_{ab} a zero, obtemos o seguinte sistema de equações:

$$n_{ab}(A,Z) = n^0_{ab}(A,Z)\frac{\left(1-e^{-\lambda_{ab}}\right)}{p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}}(1-p_{ab}). \tag{A.15}$$

e ao igualar as derivadas de λ_{ab} a zero, obtemos

$$\frac{O_{ab}(A,Z)}{\lambda_{a,b}} = \left[n^0_{ab}(A,Z)\frac{\left((1-p_{ab})e^{-\lambda_{ab}}\right)}{p_{ab}+(1-p_{ab})e^{-\lambda_{ab}}}\right] + n_{ab}(A,Z). \tag{A.16}$$

Como não foi possível encontrar uma fórmula fechada para a solução, é necessário recorrer a métodos iterativos para resolvê-la.

APÊNDICE

B

MÉTODO VARIACIONAL

O método variacional é uma técnica amplamente utilizada para encontrar distribuições aproximadas para variáveis não observadas. Essas distribuições são obtidas através do Teorema de Bayes, no entanto, como se trata de uma distribuição para variáveis não observadas, geralmente não conseguimos chegar a uma solução analítica devido à integração marginal. O método variacional é uma das alternativas para resolver esse problema, pois permite encontrar uma distribuição conjunta para cada variável desconhecida e, assim, determinar a distribuição adequada para cada variável usando apenas a probabilidade a priori.

No método variacional, obtemos uma aproximação analítica da distribuição de interesse usando a divergência de Kullback-Leibler (KL), que é uma medida não simétrica da diferença entre duas distribuições de probabilidade: a distribuição verdadeira, que neste trabalho está relacionada a $\mathbb{P}(\mathbf{A}|\mathbf{Z},\Lambda,p)$, e a distribuição aproximada, associada a $\mathbb{P}(\mathbf{A};\mathbf{Z},\Lambda,p)$. Através da KL, podemos medir o quão bem nossa distribuição aproximada se ajusta à distribuição verdadeira. Usaremos a Aproximação de Campo Médio para calcular a distribuição Q para uma determinada partição, assumindo que as variáveis latentes podem ser particionadas de forma que cada partição seja independente das outras. Este método é detalhado na seção B.2.

B.1 Divergência de Kullback-Leibler

Essa divergência é também conhecida como ganho de informação e, como dito anteriormente, é uma forma de comparar duas distribuições.

No caso em que P e Q são distribuições discretas, definimos

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{B.1}$$

No caso em que P e Q são distribuições contínuas, definimos

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx,$$

onde p e q denotam as densidades de P e Q.

É interessante observar que a equação (B.1) nos mostra que quando Q é igual a P, a divergência KL é igual a zero. Portanto, quanto mais próximo de zero a divergência estiver, mais próximas estarão as distribuições. É por isso que nosso objetivo é minimizar essa divergência, de forma a obter uma boa aproximação da distribuição verdadeira.

Podemos entender a divergência de KL como uma medida de entropia de informação, que indica o quão próximas ou distintas duas distribuições são entre si. A entropia é uma medida do grau de incerteza em uma distribuição de probabilidade e pode ser definida como:

$$H(P) := E_P[I_P(X)] = -\sum_{i=1}^{n} P(i) \log(P(i)) \tag{B.2}$$

$$H(P) := E_P[I_P(X)] = \int_{-\infty}^{\infty} p(x) \log(p(x)) dx$$

Na equação (B.2), temos o valor esperado da informação de $I_P(X)$, em que $I_P(X) = -\log(P(X))$. Nessa equação, relacionamos o quanto de informação a distribuição está trazendo. Observe que a informação aumenta conforme a probabilidade de um evento diminui. Ou seja, quando $P(\cdot)$ é próxima de 1, a informação desse evento é baixa e, quando $P(\cdot)$ é próxima de 0, a informação é alta.

A divergência KL também pode ser vista como o comprimento médio de uma codificação, onde podemos relacionar como chegamos na divergência por meio da perda de informação. Para isso, podemos associar o que chamamos de entropia cruzada com a divergência KL. Nesse caso, a entropia cruzada é dada por: $H(P,Q) := E_P[I_Q(X)] = E_P[-\log(Q(X))]$. Essa definição pode ser reformulada usando a divergência KL:

$$\begin{aligned} D_{KL}(P||Q) &= H(P,Q) - H(P) \\ &= -\sum_{i=1}^{n} P(i) \log(Q(i)) + \sum_{i=1}^{n} P(i) \log(P(i)) \\ &= \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)} \end{aligned} \tag{B.3}$$

Através da equação (B.3), busca-se encontrar uma distribuição Q que se aproxime o máximo possível da distribuição verdadeira P. Isso é feito visando minimizar a divergência KL, que mede a perda de informação entre as duas distribuições. Dessa forma, é possível avaliar o quanto de informação se perde ao usar a distribuição aproximada em vez da distribuição verdadeira. O objetivo é minimizar essa perda, fazendo com que a distribuição Q se aproxime o máximo possível de P.

Mas na prática, o que realmente nos interessa é saber quanto de informação perdemos ao usar a distribuição Q em vez da distribuição P. Nesse sentido, trabalharemos com a divergência KL reversa, em que usamos a distribuição estimada (Q) em vez da distribuição verdadeira (P). Isso é interessante, pois podemos relacionar o quanto de informação estamos perdendo se, em vez de usar a distribuição verdadeira, usarmos a distribuição aproximada. Como nosso objetivo é minimizar a perda de informação, precisamos minimizar essa divergência, que é dada por:

$$D_{KL}(Q||P) = \sum_{i=1}^{n} \log Q(i) \frac{Q(i)}{P(i)} \tag{B.4}$$

Observe que aqui temos uma situação oposta à vista anteriormente, onde queremos que se P for pequeno, Q também seja proporcionalmente pequeno. Além disso, se P for grande, a proporção se aproxima de zero. Isso indica que a divergência KL reversa penaliza mais quando Q atribui alta probabilidade a eventos com baixa probabilidade em P, o que pode levar a uma maior perda de informação ao usar Q como aproximação de P. Assim, a minimização da divergência KL reversa nos ajuda a encontrar uma distribuição Q que seja uma boa aproximação de P, sem perder muita informação.

Então, nosso objetivo é encontrar uma distribuição aproximada Q que minimize a divergência KL reversa. Suponha que θ seja a variável aleatória desconhecida, $p(X|\theta)$ a distribuição condicional conhecida, $p(\theta|x)$ a distribuição verdadeira que queremos aproximar e $q(\theta)$ a densidade aproximada de $p(\theta|x)$.

$$\begin{aligned}
D_{KL}(Q||P) &= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(\theta|X)} d\theta \\
&= \int_{-\infty}^{\infty} q(\theta) \log \left(\frac{q(\theta)}{p(X,\theta)} p(X) d\theta \right) \\
&= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(X,\theta)} + \int_{-\infty}^{\infty} q(\theta) \log(p(X)) d\theta \\
&= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(X,\theta)} + \log(p(X)) \int_{-\infty}^{\infty} q(\theta) d\theta \\
&= \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(X,\theta)} + \log(p(X))
\end{aligned} \tag{B.5}$$

Então, na equação (B.5), estamos interessados na distribuição $p(\theta|x)$, em que θ é uma variável latente. O objetivo é avaliar o quanto de informação é perdido ao aproximar $p(\theta|x)$ por uma distribuição aproximada $q(\theta)$. Podemos escrever $\log(X)$ em termos da divergência e do ELBO (evidence lower bound), $\mathfrak{L}$, como segue:

$$\begin{aligned}
\log p(X) &= D_{KL}(Q||P) - \int_{-\infty}^{\infty} q(\theta) \log \frac{q(\theta)}{p(X,\theta)} \\
&= D_{KL}(Q||P) + \mathfrak{L}(Q)
\end{aligned} \tag{B.6}$$

Onde $\log p(X)$ é a evidência e é fixa, pois depende apenas dos dados observados. Quando aumentamos o ELBO, $\mathfrak{L}(Q)$, diminuímos a divergência $D_{KL}(Q||P)$. Portanto, o objetivo é maximizar $\mathfrak{L}(Q)$ para minimizar a divergência e encontrar a distribuição $q(\theta)$ que melhor se aproxima da verdadeira distribuição $p(\theta|x)$. Isso nos leva a um problema de otimização.

E assim chegamos à aproximação de campo médio, onde buscamos aproximar a distribuição $p(\theta|X)$ por uma distribuição que pode ser fatorada, facilitando o seu manuseio. Esse é o ponto crucial do método variacional: não precisamos calcular explicitamente a probabilidade marginal, podemos resolver um problema de otimização para encontrar a distribuição correta Q que melhor se adapta à energia livre variacional. Isso nos permite encontrar uma aproximação mais simples e computacionalmente viável para a distribuição verdadeira, sem a necessidade de cálculos complexos.

B.2 Aproximação de campo médio

A aproximação de campo médio é uma suposição simplificadora da distribuição Q, onde se pode particionar as variáveis em partes independentes. Essa suposição é usada para aproximar a distribuição $p(\theta|X)$ por uma distribuição que pode ser fatorada, tornando o problema mais fácil de lidar. Com esse método, é possível "relaxar" um problema de otimização difícil para um mais simples, tornando a aproximação de campo médio um ponto crucial do método variacional.

Por exemplo, suponha que tenhamos N variáveis desconhecidas $\theta = (\theta_1, ..., \theta_N)$. Podemos particionar essas variáveis de diferentes maneiras. Por exemplo, podemos separar cada θ_i em um conjunto. Isso nos permite escrever as equações (B.5) e (B.6) com múltiplas integrais, e as variáveis serão particionadas em partes independentes que não dependem de x.

$$p(\theta|X) \approx q(\theta) = q(\theta_1, ..., \theta_N) = \prod_{i=1}^{N} q_i(\theta_i) \tag{B.7}$$

Podemos entender a equação (B.7) como um processo de aproximação da função densidade $p(\theta \mid x)$, que é desconhecida, por uma distribuição que fatora. Nesse contexto, precisamos encontrar entre as distribuições que fatoram qual é a melhor aproximação para a função de interesse. Em outras palavras, estamos buscando a distribuição $q(\theta)$ que melhor se aproxima de $p(\theta \mid x)$ e que, ao mesmo tempo, é mais fácil de ser manipulada.

Na equação (B.7), associamos a densidade de probabilidade $p(\theta \mid x)$ ao ELBO $\mathfrak{L}(Q)$ e usamos o cálculo variacional para derivar q_j em sua forma funcional, visando encontrar a função que maximiza o ELBO, não apenas um ponto de máximo. Para isso, partimos da expressão do ELBO $\mathfrak{L}$ e tentamos reescrevê-la de forma a isolar os termos $q_j(\theta_j)$, na esperança de obter uma derivada funcional que nos permita encontrar a função ótima, aquela que maximiza o funcional $\mathfrak{L}$. É importante observar que o ELBO $\mathfrak{L}$ é um funcional que depende das densidades aproximadas

$q_1, \ldots, q_N$.

$$\begin{aligned}
\mathfrak{L}[q_1,...,q_N] &= -\int_{\theta_1,...,\theta_N} \left[\prod_{i=1}^{N} q_i(\theta_i)\right] \log \frac{\prod_{k=1}^{N} q_k(\theta_k)}{p(\theta,X)} d\theta_1 ... d\theta_N \\
&= \int_{\theta_1,...,\theta_N} \left[\prod_{i=1}^{N} q_i(\theta_i)\right] \left[\log p(\theta,X) - \sum_{k=1}^{N} \log q_k(\theta_k)\right] d\theta_1 ... d\theta_N \\
&= \int_{\theta_j} q_j(\theta_j) \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \left[\log p(\theta,X) - \sum_{k=1}^{N} \log q_k(\theta_k)\right] d\theta_1 ... d\theta_N \\
&= \int_{\theta_j} q_j(\theta_j) \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \log p(\theta,X) d\theta_1 ... d\theta_N \\
&- \int_{\theta_j} q_j(\theta_j) \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \sum_{k=1}^{N} \log q_k(\theta_k) d\theta_1 ... d\theta_N
\end{aligned} \tag{B.8}$$

Separamos o q_j para chegamos a esperança e assim podemos defini-la como esperança para todas as variáveis, exceto j.

$$E_{m|m\neq j}[\log p(\theta,X)] = \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \log p(\theta,X) d\theta_1 ..., d\theta_{j-1}, d\theta_{j+1} d\theta_N \tag{B.9}$$

Assim chegamos na esperança para todas as variáveis, exceto j. Substituindo a equação (B.9) em (B.8), usando a esperança e expandindo o segundo termo:

$$\begin{aligned}
&\mathfrak{L}[q_1,...,q_N] = \int_{\theta_j} q_j(\theta_j) E_{m|m\neq j}[\log p(\theta,X)] d\theta_j \\
&- \int_{\theta_j} q_j(\theta_j) \log q_j(\theta_j) \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] d\theta_1 ... d\theta_N \\
&- \int_{\theta_j} q_j(\theta_j) d\theta_j \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \sum_{k\neq j} \log q_k(\theta_k) d\theta_1 ..., d\theta_{j-1}, d\theta_{j+1}, ... d\theta_N \\
&= \int_{\theta_j} q_j(\theta_j) E_{m|m\neq j}[\log p(\theta,X)] d\theta_j - \int_{\theta_j} q_j(\theta_j) \log q_j(\theta_j) d\theta_j \\
&- \int_{\theta_{m|m\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \sum_{k\neq j} \log q_k(\theta_k) d\theta_1 ..., d\theta_{j-1}, d\theta_{j+1}, ... d\theta_N \\
&= \int_{\theta_j} q_j(\theta_j) \left[E_{m|m\neq j}[\log p(\theta,X)] - \log q_j(\theta_j)\right] d\theta_j - G[q_1,...,q_{j-1},q_{j+1},...,q_N],
\end{aligned} \tag{B.10}$$

Então conseguimos fazer manipulações para um funcional onde temos um termo que depende de q_j e outro que não depende de q_j.

Usando o método Lagrangiano para equação (B.10), temos:

$$\mathfrak{L}[q_1,...,q_N] - \sum_{i=1}^{N} \lambda_i \int_{\theta_i} q_i(\theta_i) d\theta_i \tag{B.11}$$

onde $q_i(\theta_i)$ devem ser funções de densidade de probabilidade.

Ao realizar a derivada funcional na equação (B.11), buscamos encontrar a função de densidade $q_j(\theta_j)$ que maximiza o ELBO $\mathfrak{L}(Q)$. Nesse caso, a derivada funcional é uma derivada parcial em relação a $q_j(\theta_j)$, pois estamos considerando uma densidade por vez. Encontrando a solução dessa derivada parcial, obtemos a distribuição ótima para cada θ_j, mantendo as outras variáveis fixas. Esse processo é repetido para cada $q_j(\theta_j)$, de forma iterativa, até que a convergência seja alcançada e tenhamos encontrado a distribuição $q(\theta)$ que melhor se aproxima de $p(\theta \mid x)$.

$$\begin{aligned}\frac{\delta \mathfrak{L}[q_1,...,q_N]}{\delta q_j(\theta)} &= \frac{\partial}{\partial q_j}\left[q_j(\theta_j)\left[E_{m|m\neq j}[\log p(\theta,X)] - \log q_j(\theta_j)\right] - \lambda_j q_j(\theta_j)\right] \\ &= E_{m|m\neq j}[\log p(\theta,X)] - \log q_j(\theta_j) - 1 - \lambda_j\end{aligned} \tag{B.12}$$

Igualando a derivada a zero e isolando $q_j(\theta_j)$, obtemos

$$\begin{aligned}\log q_j(\theta_j) &= E_{m|m\neq j}[\log p(\theta,X)] - 1 - \lambda_j \\ &= E_{m|m\neq j}[\log p(\theta,X)] + const \\ q_j(\theta_j) &= \frac{e^{E_{m|m\neq j}[\log p(\theta,X)]}}{Z_j}\end{aligned} \tag{B.13}$$

onde Z_j é uma constante normalizadora. Este resultado nos permite maximizar o ELBO, que por sua vez, através da equação (B.6), ajuda a minimizar a divergência KL entre a distribuição aproximada e a distribuição verdadeira. A minimização da divergência KL garante que a distribuição aproximada esteja o mais próximo possível da distribuição verdadeira, o que é o objetivo da abordagem de inferência variacional.

O algoritmo de otimização iterativo funciona da seguinte forma:

1. Inicialização: gerar valores aleatórios para cada um dos parâmetros (funções) $q_j(\theta_j)$.
2. Iteração: usar a Equação (B.13) para minimizar a divergência KL em cada q_j, atualizando $q_j(\theta_j)$.
3. Convergência: repetir a iteração até atingir uma convergência satisfatória.

A cada iteração, estamos diminuindo a divergência entre as distribuições Q e P. O objetivo final é encontrar uma distribuição aproximada Q que esteja o mais próximo possível da distribuição verdadeira $P(\theta \mid X)$, de forma que possamos utilizar essa distribuição aproximada para fazer inferências e predições sobre a variável θ com base nos dados X observados.


ICMC
USP
SÃO CARLOS