UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Rafael Honório Pereira Alves

Testes para Erro de Especificação em Modelos para Grafos Aleatórios

Belo Horizonte
2022

Rafael Honório Pereira Alves

Testes para Erro de Especificação em Modelos para Grafos Aleatórios

Versão Final

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Estatística.

Orientadora: Denise Duarte

Belo Horizonte
2022

Alves, Rafael Honório Pereira.

A474t Testes para erro de especificação em modelos para grafos aleatórios [manuscrito] / Rafael Honório Pereira Alves. – 2022.
1 recurso online (48 f. il, color.) : pdf.

Orientadora: Denise Duarte Scarpa Magalhães Alves.
Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 46-48.

1. Estatística – Teses. 2. Redes complexas – Modelagem – Teses. 3. Sistemas estocasticos – Teses. 4. Modelos de grafos exponenciais – Teses. 5. Erro de especificação – Teses. I. Alves, Denise Duarte Scarpa Magalhães. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III.Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

UFMG

ATA DA DEFESA DE TESE DE DOUTORADO DO ALUNO RAFAEL HONÓRIO PEREIRA ALVES, MATRICULADO, SOB O Nº 2016.656.357, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 23 DE FEVEREIRO DE 2022.

Aos 23 dias do mês de FEVEREIRO de 2022, às 14h00, em reunião pública virtual 72 (conforme orientações para a atividade de defesa de tese durante a vigência da Portaria PRPG nº 1819) OU na sala do Instituto de Ciências Exatas da UFMG, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de tese do aluno RAFAEL HONÓRIO PEREIRA ALVES, nº matrícula 2016.656.357, intitulada: "Testes para *Erro de Especificação em Modelos de Grafos Aleatórios*", requisito final para obtenção do Grau de doutor em Estatística. Abrindo a sessão, a Senhora Presidente da Comissão, Profa. DENISE DUARTE SCARPA MAGALHAES ALVES, passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

Denise Duarte
Profa. Denise Duarte Scarpa Magalhaes Alves- Orientadora (EST/UFMG)

Rodrigo Lambert
Prof. Rodrigo Lambert (EST/UFU)

Prof. Guilherme Ost de Aguiar (IM-UFRJ)

Prof. Miguel Natalio Abadi (IME/USP)

Prof. Luiz Henrique Duczmal (EST/UFMG)

XXXXXXX

O resultado final foi comunicado publicamente ao(à) aluno(a) pelo(a) Senhor(a) Presidente da Comissão. Nada mais havendo a tratar, o(a) Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 23 de FEVEREIRO de 2022.

Observações:
1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

Resumo

Nos últimos anos, houve um grande interesse em modelos de grafos aleatórios para modelar redes complexas nas mais diversas áreas como Ciências Sociais, Física, Biologia, Economia, Ecologia e Ciência da Computação. Uma classe de modelos que vêm sendo muito utilizados são os modelos de grafos aleatórios exponenciais (ERG), que formam uma família abrangente de modelos que inclui modelos de arestas independentes e diádicos, os grafos aleatórios de Markov e muitas outras distribuições de grafos, além de permitir a inclusão de covariáveis que podem levar a um ajuste melhor do modelo. Outra classe de modelos cada vez mais popular na análise estatística de redes são os modelos de blocos estocásticos (SBMs) . Eles podem ser usados para fins de agrupamento dos vertices em comunidades ou descobrir e analisar uma estrutura latente de uma rede. O modelo de bloco estocástico é um modelo generativo para grafos aleatórios que tende a produzir grafos contendo subconjuntos de vértices caracterizados por serem conectados uns aos outros, chamados comunidades. Muitos pesquisadores de várias áreas vêm usando ferramentas computacionais para o ajuste desses modelos sem, no entanto, fazer uma análise da adequação deles aos dados de redes que estão estudando. A complexidade envolvida no processo de estimação e nas metodologias de verificação de qualidade de ajuste pra esses modelos podem ser fatores que dificultam a análise de adequação e um possível descarte de um modelo em favor de outro. E é claro que os resultados obtidos através de um modelo não adequado podem levar o pesquisador a conclusões bastante equivocadas sobre o fenômeno estudado. A proposta deste trabalho é apresentar uma metodologia simples, baseada em Testes de Hipóteses, para verificar se há erro de especicificação de modelo para esses dois casos bastante utilizados na literatura para representar redes complexas: o ERG e o SBM. Acreditamos que essa ferramenta pode ser bastante útil para aqueles que querem utilizar esses modelos de uma forma mais cuidadosa, verificando antes se os modelos são adequados aos dados em estudo.

Palavras-chave: Modelagem de redes complexas. Modelos de Blocos Estocásticos. Modelos de Grafos Exponenciais. Erro de especificação de modelo. Testes de Hipóteses.

Abstract

In recent years, there has been a great interest in random graph models to model complex networks in the most diverse areas such as Social Sciences, Physics, Biology, Economics, Ecology and Computer Science. A class of models that have been widely used are the exponential random graph (ERG) models, which form a comprehensive family of models that include independent and dyadic edge models, Markov random graphs, and many other graph distributions, in addition to allow the inclusion of covariates that can lead to a better fit of the model. Another increasingly popular class of models in statistical network analysis are stochastic block models (SBMs). They can be used for the purpose of grouping nodes into communities or discovering and analyzing a latent structure of a network. The stochastic block model is a generative model for random graphs that tends to produce graphs containing subsets of nodes characterized by being connected to each other, called communities. Many researchers from various areas have been using computational tools to adjust these models without, however, analyzing their suitability for the data of the networks they are studying. The complexity involved in the estimation process and in the goodness-of-fit verification methodologies for these models can be factors that make the analysis of adequacy difficult and a possible discard of one model in favor of another. And it is clear that the results obtained through an inappropriate model can lead the researcher to very wrong conclusions about the phenomenon studied. The purpose of this work is to present a simple methodology, based on Hypothesis Tests, to verify if there is a model specification error for these two cases widely used in the literature to represent complex networks: the ERGM and the SBM. We believe that this tool can be very useful for those who want to use these models in a more careful way, verifying beforehand if the models are suitable for the data under study.

Keywords: Complex Network Modeling. Stochastic Block Models. Exponential Graph Models. Miss-Specification Error. Hypothesis Testing.

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A compreensão dos mecanismos de interação em redes complexas do mundo real através de modelos de grafos aleatórios é um tema que ganhou muita atenção na literatura nos útimos anos [20].

Aplicações práticas de grafos aleatórios são encontradas em todas as áreas em que redes complexas precisam ser modeladas, alguns exemplos são Ciências Sociais, Física, Biologia, Economia, Ecologia e Ciência da Computação. As interações econômicas ou sociais geralmente também estão organizadas em estruturas de redes complexas. Fenômenos semelhantes são observados em redes de comunicação como a internet ou no fluxo de tráfego. Nos problemas atuais em Biociências, as redes de proteínas na célula são exemplos importantes, e também as redes moleculares no genoma. Em escalas maiores encontram-se redes de células como em redes neurais, até a escala de organismos em teias alimentares ecológicas. Muitos modelos de grafos aleatórios tentam espelhar os diversos tipos de redes complexas encontradas em diferentes áreas. Para uma descrição mais abrangente sobre o tema e aplicações, indicamos a leitura de Newman (2010)[21] e Reuven e Shlomo (2010)[24].

O grafo aleatório de Erdös e Rényi[10] é um dos modelos de rede mais bem estudados, no entanto, para redes do mundo real, como redes sociais, a Internet ou redes biológicas, não é um bom modelo, pois algumas propriedades básicas não se encaixam bem. Redes complexas costumam apresentar características topológicas não triviais que diferem dos grafos aleatórios de Erdoős Renyi, como cauda pesada na distribuição de graus, alto coeficiente de agrupamento, estruturas hierárquicas e comprimento médio de caminhos curtos. Dois modelos de grafos aleatórios bastante usados para modelar fenômenos naturais que tentam capturar essas características são as redes livres de escala, proposto por Barabási e Albert (1999)[5] e os modelos de Small-World introduzidos por Watts e Strogats (1998)[31].

Nos últimos anos, houve um grande interesse em modelos de grafos aleatórios exponenciais para modelar redes, principalmente redes sociais(Frank e Strauss, 1986[11], Frank, 1991[14], Wasserman e Pattison, 1996[33]; ver também Pattison e Wasserman, 1999[22], Robins et al., 1999[25]). A classe de modelos de grafos aleatórios exponenciais é uma família abrangente de modelos que inclui modelos de arestas independentes e diádicos,

os grafos aleatórios de Markov de Frank e Strauss (1986)[11] e muitas outras distribuições de grafos, além de permitir a inclusão de covariáveis que podem levar a um ajuste melhor do modelo. A estimação dos parâmetros desse modelo é feita de através de métodos computacionais e está implementada em várias linguagens de programação, no software R, por exemplo, tem o pacote chamado ERGM implementado [16].

Outra classe de modelos cada vez mais popular na análise estatística de redes são os modelos de blocos estocásticos (SBMs) introduzidos por Holland et al (1983)[15]. Eles podem ser usados para fins de agrupamento dos vértices em comunidades ou descobrir e analisar uma estrutura latente de uma rede. Houve um desenvolvimento rápido no tema de agrupamento baseado em modelos de grafos aleatórios nos últimos dez anos. Citando apenas alguns, temos os dois trabalhos de Abbe e Sandom (2015)[1][2] e o de Karrer e Newman (2011)[18]. O modelo de bloco estocástico é um modelo generativo para grafos aleatórios que tende a produzir grafos contendo subconjuntos de vértices caracterizados por serem conectados uns aos outros, chamados comunidades. Esse modelo é hierárquico no sentido que primeiro sorteamos os vértices que pertecem a cada grupo, depois sorteamos as arestas entre os vértices. Cada um desses sorteios é governado por uma lei que depende de parâmetros específicos. Devido à complexidade da verossimilhança desse processo em dois estágios, a estimação dos parâmetros de um SBM também é baseada em métodos computacionais que estão implementados em várias linguagens. Por exemplo, no software R está implementado o pacote Stochastic Block Model (SBM) implementado por Leger (2016)[13]. Uma revisão detalhada sobre os SBM's pode ser encontrada em Lee e Wilkinson (2019)[17] e uma extensão muito interessante aplicada para redes grandes é apresentada em Peixoto (2014)[23].

Muitos pesquisadores de várias áreas vêm usando ferramentas computacionais para o ajuste de modelos sem, no entanto, fazer uma análise crítica a respeito da adequação do mesmo aos dados que estão sendo analisados, principalmente nos casos em que o modelo escolhido é um ERG ou um SBM. Para o ERG citamos, por exemplo, os procedimentos de verificação baseados em reamostragem apresentados em Kolaczyk e Csárdi (2014)[19]. No caso do SBM, uma das metodologias de verificação de ajuste de modelo são baseadas no Akaike Information Criterion, AIC ([3] e [4]) e estão implementadas no pacote SBM do R. A complexidade envolvida no processo de estimação e nas metodologias de verificação de qualidade de ajuste pra esse modelos podem ser fatores que dificultam a análise de adequação e um possível descarte de um modelo em favor de outro. E é claro que os resultados obtidos através de um modelo não adequado podem levar o pesquisador a conclusões bastante equivocadas sobre o fenômeno estudado.

A proposta deste trabalho é apresentar uma metodologia simples, baseada em Testes de Hipóteses, para verificar se há erro de especicificação de modelo para esses dois casos bastante utilizados na literatura para representar redes complexas: o ERG e o SBM. Acreditamos que essa ferramenta pode ser bastante útil para aqueles que querem utilizar

esses modelos de uma forma mais cuidadosa, verificando antes se os modelos são adequados aos dados em estudo. Destacamos a vantagem desse tipo de metodologia em relação aos processos de seleção de modelos do tipo AIC ou BIC ([29]) e outros com a mesma proposta, pois nessas ferramentas não temos uma indicação quanto à adequação dos modelos, mas uma seleção entre os candidatos propostos, que podem ser todos inadequados.

Derivaremos os testes levando em conta as verossimilhanças desses dois modelos, utilizando estimadores de máxima verossimilhança já propostos na literatura, para o ERG citamos Schmid e Desmarais (2017)[30] e para o SBM, citamos Celisse, Daudin e Pierre (2012)[7]. Seguiremos de perto o trabalho apresentado por White (1982)[32], onde é proposto um teste de verificação de ajuste de modelos para variáveis aleatórias cujas densidades satisfazem um conjunto geral de condições de regularidade. Mostraremos que os modelos ERG e SBM satisfazem essas condições e explicitaremos as estatísticas que devem ser calculadas para realizar os testes em cada um desses modelos. Desenvolveremos um pacote no R com a implementação dos testes para os dois modelos e apresentaremos um estudo de simulação mostrando que os testes propostos atigem os propósitos desejados.

Capítulo 2

Principais Definições

2.1 Alguns modelos para grafos aleatórios

2.1.1 Grafos Aleatórios de Erdos Rényi

Um grafo aleatório G é uma variável aleatória que assume valores em uma família de grafos $\mathcal{G}$. O estudo de tais grafos remonta à década de 1950, quando Paul Erdos e Alfréd Rényi, derivaram uma série de resultados sobre grafos aleatórios. Os grafos a que nos referimos aqui são todos rotulados, i.e., os vértices são distintos (por exemplo, há $\binom{n}{2}$ grafos com n vértices e exatamente uma aresta).

Existem duas formas de definir o modelo de grafos aleatórios Erdős-Rényi:

Definição 1 *O modelo de Erdos-Rényi denotado por $G(n, M)$, representa um grafo escolhido de forma aleatória da coleção de todos os grafos com n vértices e M arestas.*

Por exemplo, no modelo $G(3, 2)$ cada uma das três possibilidades de grafos de três vértices e duas arestas são incluídos com uma probabilidade de $\frac{1}{3}$.

Definição 2 *O modelo de Erdos-Rényi-Gilbert denotado por $G(n, p)$, representa um grafo com n vértices cujas arestas existem ou não com probabilidade p.*

Dessa forma para cada par u, v de vértices de $G = G(n, p)$, a aresta uv é aleatória(existe ou não existe), independentemente de qualquer outra aresta. Usa-se uma variável aleatória independente de Bernoulli de parâmetro p para decidir quanto à presença da aresta uv; a aresta uv está no grafo se, e só se, a variável resulta em sucesso. Pela independência entre as arestas, a probabilidade de um grafo G, com n vértices e M arestas, é

$$P(G) = p^M (1-p)^{\binom{n}{2}-M}. \tag{2.1}$$

Várias propriedades dos grafos de Erdos-Rényi são hoje bem conhecidas. Esse modelo é por vezes usado como referência para avaliação de outros, representando o papel de rede com comportamento aleatório uniforme.(Veja Frank e Strauss (1986)[11])

Vários outros modelos de grafos aleatórios são estudados na literatura. Eles atendem à necessidade de modelos mais complicados e realistas (do que o modelo ER) que descrevam as redes reais, observadas na prática. Algumas características tidas como típicas de redes reais são a presença de um grande número de vértices n, poucas arestas(e(G) = O(n)), diâmetro pequeno ($diam(G) = O(\log n)$, i.e., dois vértices tomados ao acaso estão ligados por um caminho curto), graus de vértices distribuídos segundo uma lei de potências (o número de vértices com grau k é proporcional a $k^{-\beta}$, para alguma constante β), e efeito de agrupamento (clustering, ou transitividade das ligações: vértices com vizinhança comum têm maior probabilidade de estarem ligados). O número de publicações e estudos de redes complexas e grafos aleatórios é grande. Para uma introdução à área, citamos Chatterjee e Diaconis (2011)[9] e Robins, Pattison, Kalish e Lusher(2007)[27].

2.1.2 Grafos Aleatórios Exponenciais (ERG)

O modelo ERG, também conhecido como ($p*$)modelo, foi primeiro proposto por Holland e Leinhardt (1983)[15], e foi construído com base nas fundamentações estatísticas estabelecidas por Besag (1974)[6]. Esses modelos constituem uma família de modelos estatísticos que vêm sendo bastante utilizados para modelar redes sociais. A importância deste modelo está em sua capacidade de representar os efeitos estruturais sociais comummente observados em muitas redes sociais humanas, incluindo efeitos gerais baseados no grau de cada vértice, como reciprocidade e transitividade, ou ainda, atividade baseada em atributos e efeitos de popularidade.

Desenvolvimentos substanciais foram feitos por Frank e Strauss (1986)[11], e continuaram a ser feitos por outros autores em toda a década de 1990 . E ainda hoje são objeto de estudo de vários pesquisadores, sempre na tentativa de tornar o ERG um modelo ainda melhor, mais aplicável a dados de rede reais. Uma revisão detalhada do assunto é apresentada em Wasserman and Pattison (1996)[33]. O modelo proposto por Besag, no contexto de Estatística Espacial, por sua vez, é centrado no Teorema de Hammerserley-Clifford (1971)[12], nascido na Física Estatística, que mostra que um modelo probabilístico para grafos, com certa estrutura de dependência tem, necessariamente, que pertencer a uma família exponencial. Mais que isso, Esse importante Teorema também mostra quais são as informações da rede que devem ser usadas para o cálculo das probabilidades de uma determinada configuração.

O problema que surge ao considerarmos estruturas de dependências entre os vértices, ainda que sejam estruturas consideravelmente simples como as consideradas no artigo referido acima, é que o número de parâmetros a serem estimados no modelo é muito grande, como detalharemos mais a frente neste texto. As inferências propostas para os parâmetros desse modelo são baseadas em pseudo máxima verossimilhança, usando uma analogia com o modelo de regressão logística, o que torna as inferências pouco confiáveis. Na tentativa de superar esses e outros problemas, na década de 2000 novas especificações sobre a estrutura de dependência dos vértices foram prospostas por pesquisadores, como as encontradas na série de artigos publicados por Snijders(1997)[28]. A ideia principal nesses trabalhos é diminuir a dimensão do vetor de parâmetros dos modelos e preservar as características da rede. Surgem também, os modelos ERG que permitem que características exógenas da rede possam ser usadas para modelar a probabilidade de ocorrência de uma configuração, melhorando assim as inferências para a rede. Um modelo ERG alternativo aos propostos pelos artigos de Snijders e colaboradores, mas com objetivo semelhante em relação à diminuição no número de parâmetros, foi proposto por Hunter e Handcock(2009) [16].

Definição 3 *Um modelo de grafo aleatório é um ERG se, para todo grafo $G \in \mathcal{G}$, podemos expressar a probabilidade de sua ocorrência por*

$$P(G;\theta) = \exp\left(\sum_{i=1}^{n} \theta_i T_i(G) - \varphi(\theta)\right) = \exp\left(\sum_{i=1}^{n} \theta_i T_i(G)\right) / z(\theta).$$

onde $\theta = (\theta_1, \ldots, \theta_n)$ é um vetor de parâmetros reais conhecido e $T_i(G)$ são funções de G (como o número de arestas, triângulos, estrelas, circuitos, etc.); i.e., se a distribuição sobre o espaço dos grafos é membro da família exponencial de distribuições.

Como no modelo de Erdos-Rényi, consideramos fixado o número n de vértices de G.

O fator $e^{-\varphi(\theta)} = z^{-1}(\theta)$ é por vezes chamado constante de normalização.

Duas dificuldades no uso de grafos aleatórios exponenciais são a estimação de $z(\theta)$ e o fato de que valores muito distintos de θ dão origem a distribuições essencialmente iguais no espaço dos grafos (Veja Chatterjee e Diaconis, 2011).

- Caso de arestas independentes
  Suponhamos que as ligações entre os vértices ocorram independentemente umas das outras, ou seja, que não haja dependência dentro da rede. Nesse caso, a função $T_i(G)$ se torna apenas a indicadora da aresta Y_{ij} da matriz de adjacências Y de G, dessa forma, modelo ERG geral se simplifica bastante, pois os parâmetros do modelo se reduzem aos coeficientes de ligação ij e o ERG se reduz ao modelo de Erdös-Rènyi.

- Modelo de Grafos de Markov
  Seguindo o trabalho de Besag (1974)[6] na área de Estatística Espacial, Frank e Strauss(1986)[11] propuseram uma dependência de Markov em um Grafo, postulando que uma possível ligação de i para j é assumida como dependente de qualquer outro vínculo possível envolvendo i ou j, mesmo que todos as outras ligações na rede sejam fixos. A dependência de Markov implica que duas possíveis arestas de uma rede são condicionalmente independentes, a menos que compartilhem um vértice comum. Eles mostraram que essa suposição resultou em modelos para grafos não direcionados que envolvem parâmetros associados a estatísticas simples da rede como número de arestas, de estruturas em forma de estrelas e de triângulos. Nesse modelo, dois vértices são considerados vizinhos se compartilham uma aresta. Um subconjunto do conjunto de vértices, v, onde todos os elementos são vizinhos é chamado de clique.

  Observamos que todas as especificações do modelo envolvem estatísticas que são apenas funções da própria rede Y , apenas efeitos endógenos são considerados. Ainda assim é natural esperar que a probabilidade de ocorrer uma ligação entre dois vértices pode depender também de características, atributos dos próprios vértices. Então, permitir a incorporação de efeitos exógenos pode levar a infências mais acuradas sobre a rede. Podemos incorporar atributos que foram medidos nos vérties, sob a forma de estatísticas adicionais na função dentro da exponencial.

2.1.3 Modelo de Blocos Estocásticos (SBM)

Ao analisar redes complexas, uma tarefa básica na área de detecção de comunidades (ou clusterização) consiste em particionar os vértices de um grafo em clusters que são mais densamente conectados. Mais geralmente, estruturas comunitárias também podem se referir a grupos de vértices que se conectam de forma semelhante ao resto do grafo, sem ter, necessariamente, uma maior densidade interna. No contexto mais geral, a detecção da comunidade refere-se ao problema de inferir relações de similaridade entre os itens de uma rede, observando suas interações locais.

A detecção da comunidade é um dos problemas centrais em redes e das ciências de dados. O Modelo de Blocos Estocásticos (SBM) tem sido amplamente utilizado como modelo canônico para estudar estas questões.

Definição 4 *O modelo de blocos estocásticos é uma família de distribuição de probabilidade para um grafo em blocos G com conjunto de vértices $\{1, \ldots, n\}$ e conjunto de blocos*

$\{1, \ldots, m\}$, *tal que:*

1. *Os parâmetros são o vetor* $\theta = (\theta_1, \ldots, \theta_m)$, *das probabilidades de blocos e a matriz* $\eta = (\eta_{kl})_{1 \leq k \leq l \leq m}$, *das probabilidades das arestas bloco-dependentes.*

2. *O vetor do conjunto de blocos consiste das* $(X_i)_{i=1}^n$ *variáveis aleatórias independentes e indenticamente distribuidas, onde* $P(X_i = k) = \theta_k$, *para* $k = 1 \ \ldots, m$.

3. *Condicional ao bloco do vértice* X_i, *as arestas* Y_{ij} *são independentes com* $Y_{ij} \sim$ *Bernoulli*(η_{X_i, X_j}).

Se (X, Y) *representa um grafo em blocos* G, *a função de probabilidade é dada por:*

$$P(\theta, \eta; X, Y) = \theta_1^{n_1} \cdots \theta_m^{n_m} \prod_{1 \leq k \leq l \leq m} \eta_{kl}^{e_{kl}} (1 - \eta_{kl})^{n_{kl} - e_{kl}},$$

onde $n_k = \sum_{i=1}^n I(X_i = k)$ *denota o número de vértices de* G *que pertencem ao bloco* k,

$$e_{kl} = \sum_{1 \leq i \neq j \leq n} Y_{ij} I(x_i = k) I(x_j = l).$$

denota o número de arestas de G *que tem um vértice no bloco* k *e um vértice no bloco* j, *e*

$$n_{kl} = \begin{cases} n_k n_l & \text{se } k \neq l \\ \binom{n_k}{2} & \text{se } k = l, \end{cases}$$

Um grafo G no conjunto de vértices $v(G)$ pode ser representado pela sua matriz de adjacência $Y = \{Y_{ij}\}_{1 \leq i \neq j \leq n}$, onde

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

onde $Y_{ii} = 0$ para todo i, i.e., não há ligação do vértice consigo mesmo.

Consideramos um grafo cujos vértices pertencem a m categorias diferentes. Essas categorias chamaremos de blocos. Seja $X = (X_i)_{i=1}^n$, em que $X_i = k$, se o vértice i pertence ao bloco k, para todo $i \in \{1, \ldots, n\}$ e $k \in \{1, \ldots, m\}$. Então o grafo em blocos pode ser representado por (Y, X), X é chamada a estrutura de bloco do grafo G.

Para um grafo aleatório em blocos, o número de vértices n é fixo, mas a matriz de adjacência Y e a estrutura de bloco X são aleatórias.

Seja o conjunto de vértices $\{1, \ldots, n\}$ e as seguintes condições:

1. $Y_{ij} = Y_{ji}$ e $Y_{ii} = 0$.

2. Existe uma partição dos n vértices em m blocos tal que para todos i, j, h com $i \neq j \neq h$, se i e h pertencem ao mesmo bloco, então Y_{ij} e Y_{hj} são identicamente distribuidos.

A distribuição condicional do grafo em blocos dado o vetor de blocos $(X_i)_{i=1}^n$ é um modelo de blocos estocásticos com arestas independentes em que os blocos são função dos parâmetros. No geral o número de parâmetros do modelo tende a infinito juntamente com n, o que dificulta a estimação dos mesmos. Várias propriedades estocásticas dos modelos de blocos estocásticos são estudadas na literatura.(Veja Snijders (1997)[28] e Celisse, Daudin e Pierre (2012)[7])

2.2 Estimação de Máxima-Verossimilhança

2.2.1 Estimação de Máxima-Verossimilhança para ERG

Suponha que $X_1, \ldots, X_n$ são variáveis aleatórias independentes e identicamente distribuídas seguindo distribuição $f(\cdot|\theta)$.

Definição 5 *Dados os valores observados $x_1, \ldots, x_n$ a função de verossimilhança é:*

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta).$$

Esta função é a densidade conjunta de $x_1, \ldots, x_n$, mas em função de θ.

Definição 6 *Seja $\hat{\theta}$ o valor de θ correspondente ao máximo global da função, $\hat{\theta}$ é chamado o estimador de maximaverossimilhança do modelo.*

$$\hat{\theta} = \max_{\theta \in \Theta} \mathcal{L}(\theta; x_1, \ldots, x_n)$$

Uma maneira mais fácil de descobrir $\hat{\theta}$ é fazendo uso da função log-verossimilhança. Esta função tem o mesmo estimador de máxima verossimilhança $\hat{\theta}$ como na função de verossimilhança.

Definição 7 *A função log-verossimilhança é dada por:*

$$l(\theta; x_1, \ldots, x_n) = \log \left\{ \prod_{i=1}^{n} f(x_i|\theta) \right\} = \sum_{i=1}^{n} \log \left[f(x_i|\theta) \right].$$

No caso do ERG temos que

$$\mathcal{L}(\theta; T) = \prod_{i=1}^{n} \frac{\exp(\theta_i T_i(G))}{z(\theta)},$$

O que implica em

$$l(\theta; T_1, \ldots, T_n) = \sum_{i=1}^{n} \theta_i T_i(G) - \log \left\{ \sum_{y \in \Omega} \exp \left[\theta_i T_i(y) \right] \right\} \tag{2.2}$$

em que Ω é o conjunto de todos possíveis grafos de n vértices.

Quando estamos lidando com redes muito grandes, é muito difícil diferenciar o segundo termo da equação 2.2 e a complexidade computacional aumenta conforme o número de estatísticas suficientes ultilizadas. Em ambos os casos o estimador obtido é consistente.

Alguns métodos são usados na literatura, como por exemplo o método de estimação por pseudo-verossimilhança ou método Monte Carlo Markov Chain's(MCMC) que é mais usado e implementado atualmente.(Veja Corander e Dahmstrom (1998)[8])

Nos casos mais simples, como por exemplo no modelo de apenas um parâmetro, θ, e com a função $T(G)$ igual ao número de arestas do grafo, a maximização direta é fácil de ser obtida e vai ser discutida com detalhes nesta tese.

2.2.2 Estimação de Máxima-Verossimilhança para SBM

No caso do SBM, temos que a função de verossimilhança é dada por

$$\mathcal{L}(\theta, \eta; x, y) = \prod_{i=1}^{n} \prod_{j=1}^{n} \left\{ \prod_{l=1}^{m} \theta_1^{\left(I_{x_i=l}\right)} \prod_{1 \leq k \leq l \leq m} \eta_{kl}^{e_{kl}} (1-\eta_{kl})^{n_{kl}-e_{kl}} \right\},$$

O que implica em

$$l(\theta, \eta; x, y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \log \left(\left\{ e^{\sum_{k=1}^{l} \sum_{l=1}^{m} \{e_{kl} \log(\eta kl) + (1-e_{kl}) \log(1-\eta_{kl})\}} \prod_{l=1}^{m} \theta_1^{\left(I_{x_i=l}\right)} \right\} \right).$$

Essa função não é fácil de ser maximizada, primeiro pelo fato de que Y_{ij} são independentes apenas condicionalmente aos X_i e X_j, segundo que o número de váriaveis aleatórias na expressão

$$\sum_{k=1}^{l} \sum_{l=1}^{m} \{e_{kl} \log(\eta kl) + (1-e_{kl}) \log(1-\eta_{kl})\} \tag{2.3}$$

é $\binom{n}{2}$, bem maior que n o número de vértices.

Várias tecnicas são usadas nesse caso. Uma maximização direta pode ser feita computacionalmente aplicando uma transformação à log-verossimilhança que passa a expressão 2.3 de um problema com $\binom{n}{2}$ variáveis para um problema com n variáveis. O

algoritmo Maximização da Esperança(EM) pode ser aplicado, para isso, é preciso tomar o vetor X como o vetor de dados faltantes e maximizar a esperança log-verossimilhança restrita à $P(Y, X, \theta, \eta|Y, \theta, \eta)$. Pode-se ainda usar métodos MCMC. Todos os métodos geram estimadores consistentes, no entanto na prática, a estimação só pode ser feita para grafos com n e m razoavelmente pequenos e geram estimativas de η geralmente instáveis.(Veja Snidjers (1997)[28])

Um método de estimação utilizando EM e técnicas variacionais, chamado EM Variacional, foi proposto e vem sendo o mais usado no estudo dos SBM. O método ainda pode gerar estimativas instáveis para η, mas é bastante útil na prática, pois consegue estimar os parâmetros para grafos com n e/ou m grandes e os estimadores obtidos são consistentes.(Veja Celisse, Daudin e Pierre(2012)[7]). O pacote *blockmodels* do software R utiliza o EM Variacional para estimar os parâmetros do SBM (Leger, 2016[13]).

2.3 Estimação de Quase-Verossimilhança para Erro de Especificação

Desde que Fisher postulou o método de máxima verossimilihança na década de 20, o método tornou-se uma das ferramentas mais importantes para estimativa e inferência disponível para estatísticos.

Uma hipotese fundamental subjacente aos resultados clássicos sobre as propriedades do estimador de máxima verossimilhança é que a lei estocástica que determina o comportamento dos fenômenos investigados (a estrutura verdadeira) é conhecida dentro de uma família paramétrica especificada de distribuições de probabilidade (o modelo). Em outras palavras, o modelo de probabilidade é considerado especificado corretamente. Em muitas (senão na maioria) circunstâncias, pode-se não ter total confiança de que é assim. Se não se assume que o modelo de probabilidade está corretamente especificado, é natural perguntar o que acontece com as propriedades do estimador de máxima verossimilhança. Ainda converge para algum limite assintoticamente, e esse limite tem algum significado? Se o estimador é de alguma forma consistente, ele também é assintoticamente normal? O estimador tem propriedades que podem ser usadas para decidir se a família especificada de distribuições de probabilidade contém ou não a estrutura verdadeira? White (1982)[32] deu resposta a essas perguntas.

Sob algumas condições, o estimador de quase-máxima verossimilhança (QMLE) é um estimador natural para os parâmetros que minimizam o critério de informação de Kullback-Leibler, assim o estimador de máxima verossimilhança converge para um limite

bem definido, mesmo quando o modelo de probabilidade não é especificado corretamente. Uma característica interessante desse resultado é que, com a especificação errada, a matriz de covariância assintótica do QMLE não é mais igual ao inverso da matriz de informação de Fisher. No entanto, a matriz de covariância pode ser estimada de forma consistente e, como esperado, simplifica a forma familiar na ausência de erros de especificação. Esta propriedade é explorada para produzir um novo teste para erros de especificação, aplicável a uma ampla gama de problemas, nesse trabalho estenderemos para ERG e SBM. (Veja White (1982)[32])

Definição 8 *A função de quase-log-verossimilhança da amostra é a função:*

$$L_n(U,\theta) \equiv n^{-1}\sum_{t=1}^{n}\log f(U_t,\theta)$$

e o estimador de quase-máxima verossimilhança(QMLE) é o vetor de parâmetros $\hat{\theta}_n$ que é solução da equação:

$$\hat{\theta}_n = \max_{\theta\in\Theta} L_n(U,\theta)$$

A seguir, apresentamos as suposições que serão necessárias para mostrar os resultados de existência e convergência de QMLE, essas suposições forma inicialmente apresentados em White (1982) [32]. As provas dos Teoremas enunciados nesta seção também podem ser encontradas no artigo de White (1982) [32].

Suposição 1 *Os vetores aleatórios independentes $1\times M$, U_t com $t = 1,\ldots,n$, têm função de distribuição conjunta comum $H \in \Omega$.*

Como H é desconhecido a priori, escolhemos uma família de funções de distribuição que pode ou não conter a estrutura verdadeira, H. Geralmente, é fácil escolher essa família para satisfazer a próxima suposição.[32]

Suposição 2 *A família de funções de distribuição $F(u,\theta)$ tem densidades $f(u,\theta) = dF(u,\theta)/d\nu$.*

Teorema 1 *Dados os pressupostos 1 e 2, para todo n existe um QMLE $\hat{\theta}_n$.*

Uma vez garantida a existência de um QMLE, passamos a examinar suas propriedades. Quando F contém a verdadeira estrutura H (isto é, $H(u) = F(u,\theta_0)$ para algum $\theta_0 \in \Theta$) a teoria geral dos estimadores de máxima verossimilhança garante que o MLE é consistente para θ_0 sob condições de regularidade adequadas. No entanto, sem essa restrição, observou que, como $L_n(U,\theta)$ é um estimador natural para $E(\log f(U_t,\theta))$, $\hat{\theta}_n$ é um estimador natural para θ_*, o vetor de parâmetro que minimiza o Critério de Informação de Kullback-Leibler (KLIC),

$$I(h:f,\theta) \equiv E\left(\log\left[\frac{h(U_t)}{f(U_t,\theta)}\right]\right).$$

Aqui as esperanças são tomadas com relação à distribuição verdadeira. Portanto,

$$I(h:f,\theta) \equiv \int \log(h(u))dH(u) - \int \log(f(u,\theta))dH(u).$$

O oposto de $I(h:f,\theta)$ é chamado de entropia da distribuição $H(u)$ em relação a $F(u,\theta)$. Intuitivamente, $I(h:f,\theta)$ mede nossa ignorância sobre a verdadeira estrutura[32]. Para $\hat{\theta}_n$ ser um estimador natural de θ_*, impomos a seguinte condição

Suposição 3 *a)* $E(\log(h(U_t))$ *existe e* $|\log f(U_t,\theta)| \leq m(u)$ *para todo* $\theta \in \Theta$,

b) $I(h:f,\theta)$ *tem mínimo único em* $\theta_* \in \Theta$.

A suposição 3 garante que o KLIC está bem definido[32].

Teorema 2 *Dados os pressupostos 1 até 3,* $\hat{\theta}_n \to \theta_*$ *quando* $n \to \infty$ *para quase toda sequência* (U_t).
$(\hat{\theta}_n \overset{q.c.}{\to} \theta_*)$.

Em outras palavras, o QMLE é geralmente um estimador fortemente consistente para o vetor de parâmetros que minimiza o KLIC[32].

O próximo passo é mostrar a normalidade assintótica do QMLE e, para isso, precisamos definir algumas matrizes auxiliares, quando as derivadas parciais existem:

$$A_n(\theta) = n^{-1}\sum_{t=1}^{n}\left[\frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_i\partial\theta_j}\right],$$

$$B_n(\theta) = n^{-1}\sum_{t=1}^{n}\left[\frac{\partial \log f(U_t,\theta)}{\partial\theta_i}.\frac{\partial \log f(U_t,\theta)}{\partial\theta_j}\right].$$

E consideramos as esperanças,

$$A(\theta) = E\left(\frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_i\partial\theta_j}\right),$$

$$B(\theta) = E\left(\frac{\partial \log f(U_t,\theta)}{\partial\theta_i}.\frac{\partial \log f(U_t,\theta)}{\partial\theta_j}\right).$$

Quando as inversas apropriadas existem, definimos:

$$C_n(\theta) = A_n(\theta)^{-1}B_n(\theta)A_n(\theta)^{-1},$$

$$C(\theta) = A(\theta)^{-1}B(\theta)A(\theta)^{-1}.$$

Suposição 4 $\partial \log f(u,\theta)/\partial\theta_i$, $i = 1,\ldots,p$, *são funções continuamente diferenciáveis de* θ *para cada* u *em* Ω.

Suposição 5 $|\partial^2 \log f(u,\theta)/\partial\theta_i\partial\theta_j|$ *e* $|\partial \log f(u,\theta)/\partial\theta_i.\partial \log f(u,\theta)/\partial\theta_j|$, $i,j = 1,\ldots,p$ *são dominadas por funções integráveis com respeito a* H *para todos os* u *em* Ω *e* θ *em* Θ.

Suposição 6 *a)* θ_* *é ponto interior de* Θ,

b) $B(\theta_*)$ *é não-singular,*

c) θ_* *é ponto regular de* $A(\theta)$.

A suposição 4 assegura que as duas primeiras derivadas com relação a θ existam. Estas condições nos permitem aplicar um teorema do valor médio para funções aleatórias. A premissa 5 garante que as derivadas sejam dominadas por funções integráveis com relação a H, o que garante que $A(\theta)$ e $B(\theta)$ sejam contínuas em θ e que possamos aplicar uma lei dos grandes números para $A_n(\theta)$ e $B_n(\theta)$. Na suposição 6, definimos um ponto regular da matriz $A(\theta)$ como um valor para θ tal que $A(\theta)$ tem *rank* constante em alguma vizinhança aberta de θ. Com essas suposições adicionais, o Teorema a seguir garante que o QMLE tem distribuição assintóticamente normal[32].

Teorema 3 ***(Normalidade Assintótica)*** *Dadas as suposições de 1 até 6:*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_*\right) \overset{A}{\sim} N(0, C(\theta_*)).$$

Além disso $C(\hat{\theta}_n) \overset{q.c.}{\sim} C(\theta_*)$*, elemento por elemento.*

Temos normalidade assintótica desde que

$$\int \frac{\partial^2 \log f(u,\theta)}{\partial\theta_i \partial\theta_j}.f(u,\theta)d\nu = -\int \frac{\partial \log f(u,\theta)}{\partial\theta_i}.\frac{\partial \log f(u,\theta)}{\partial\theta_j}.f(u,\theta)d\nu \tag{2.4}$$

A equação 2.4 é a igualdade familiar na teoria da máxima verossimilhança que assegura a equivalência da Hessiana(lado esquerdo) e inverso da Matriz de Informação de Fisher(lado direito). No presente caso, essa equivalência geralmente não será válida. No entanto, quando o modelo é especificado corretamente e a próxima suposição é válida, obtemos um resultado de equivalência da matriz de informação[32].

As suposições a seguir são necessárias para enunciar o Teorema 4 a respeito da Informação de Fisher associada a θ.

Suposição 7 $\partial[\partial \log f(u,\theta)/\partial\theta_i.f(u,\theta)]/\partial\theta_j$*,* $i,j = 1,\ldots,p$*, são dominadas por funções integráveis com respeito a* ν *para todo* $\theta \in \Theta$.

Juntas, as condições dadas de 1 até 7 e $h(u) = f(u,\theta_0)$ para algum θ_0 em Θ, podem ser consideradas como as *condições usuais de regularidade de máxima verossimilhança*, já que asseguram que todos os resultados familiares se mantêm[32].

Teorema 4 ***(Matriz de Informação)*** *Dadas as suposições de 1 até 7, se* $g(u) = f(u,\theta_0)$ *para algum* θ_0 *em* Θ*, então* $\theta_* = \theta_0$ *e* $A(\theta_0) = -B(\theta_0)$ *, daí* $C(\theta_0) = -A(\theta_0)^{-1} = B(\theta_0)^{-1}$*, em que* $-A(\theta_0)$ *é a matriz de informação de Fisher.*

O Teorema 4 diz essencialmente que, quando o modelo é especificado corretamente, a matriz de informação pode ser expressa na forma de Hessiana, $-A(\theta_0)$ ou na forma de produto, $B(\theta_0)$. Equivalentemente, $A(\theta_0) + B(\theta_0) = 0$. Quando essa igualdade falha, segue-se que o modelo é mal-especificado e essa especificação errada pode ter sérias consequências quando técnicas inferenciais padrão são aplicadas. Desse modo $A(\theta_*) + B(\theta_*)$ é um indicador útil para erros de especificação!

A matriz $A(\theta_*) + B(\theta_*)$ não é observável, mas pode ser consistentemente estimada por $A_n(\hat{\theta_n}) + B(\hat{\theta_n})$. Para obter uma estatística de teste, consideramos a distribuição assintótica dos elementos de $\sqrt{n}(A_n(\hat{\theta_n}) + B(\hat{\theta_n}))$, antecipando que, sob condições apropriadas, esses elementos tem assintoticamente uma distribuição normal, com média zero, na ausência de erros de especificação. Dado um estimador consistente para a matriz de covariância assintótica, podemos obter uma estatística de teste assintóticamente χ^2_q, para um q especificado[32].

Definimos agora outras matrizes auxiliares neccessárias para a construção da estatística do teste de especificação. Consideremos $l = 1, \ldots, p(p+1)/2; i = 1, \ldots, p; j = 1, \ldots, p$, onde p é o número de coordenadas do vetor θ (número de parâmetros do modelo). E seja

$$d_l(U_t, \theta) = \frac{\partial \log f(U_t, \theta)}{\partial \theta_i} . \frac{\partial \log f(U_t, \theta)}{\partial \theta_j} + \frac{\partial^2 \log f(U_t, \theta)}{\partial \theta_i \partial \theta_j},$$

Definimos também

$$D_{ln}(\hat{\theta}_n) = n^{-1} \sum_{t=1}^{n} d_l(U_t, \hat{\theta}_n),$$

que são os elementos de $A_n(\hat{\theta_n}) + B_n(\hat{\theta_n})$.

$$D_n(\hat{\theta_n}) = n^{-1} \sum_{t=1}^{n} d(U_t, \theta)$$

Quando as derivadas parciais e esperanças existem, definimos:

$$\nabla D_n(\theta) = n^{-1} \sum_{t=1}^{n} \left[\frac{\partial d_l(U_t, \theta)}{\partial \theta_k} \right]$$

$$\nabla D(\theta) = E\left(\frac{\partial d_l(U_t, \theta)}{\partial \theta_k} \right)$$

As suposições a seguir são necessárias para construir uma quantidade com uma distribuição assintótica que será usada na construção de um teste de hipóteses para verificar se existe erro de especificação em relação à família H.

Suposição 8 *$\partial d_l(u, \theta)/\partial \theta_k$, $l = 1, \ldots, q$, $k = 1, \ldots, p$, existem e são funções contínuas de θ para cada u.*

Suposição 9 *$|d_l(u,\theta)d_m(u,\theta)|$, $|\partial d_l(U_t,\theta)/\partial\theta_k|$, e $|d_l(u,\theta).\partial\log(u,\theta)/\partial\theta_k|$, $l,m=1,\ldots,q$, $k=1,\ldots,p$ são dominadas por funções integráveis com respeito a H para todo u e θ em Θ.*

Essas suposições desempenham papéis análogos as suposições 4 e 5. A hipótese 8 requer terceiras derivadas contínuas para a função de quase-log-verossimilhança. Entre outras coisas, a hipótese 9 garante que $\nabla D(\theta)$ seja finito para todos os θ em Θ[32]. Definimos:

$$V(\theta) = E\left(\left[d(U_t,\theta) - \nabla D(\theta)A(\theta)^{-1}\nabla\log f(U_t,\theta)\right].\left[d(U_t,\theta) - \nabla D(\theta)A(\theta)^{-1}\nabla\log f(U_t,\theta)\right]'\right)$$

$V(\theta_*)$ é a matrix de covariância assintótica de $\sqrt{n}D_n(\hat{\theta_n})$ e temos:

Suposição 10 *$V(\theta_*)$ é não-singular.*

Um estimador consistente para $V(\theta_*)$ é

$$\begin{aligned} V_n(\hat{\theta_n}) &= n^{-1}\sum_{t=1}^{n}\left\{\left[d(U_t,\hat{\theta}_n) - \nabla D_n(\hat{\theta}_n)A_n(\hat{\theta}_n)^{-1}\nabla\log f(U_t,\hat{\theta}_n)\right].\right. \\ &\quad \left.\left[d(U_t,\hat{\theta}_n) - \nabla D_n(\hat{\theta}_n)A_n(\hat{\theta}_n)^{-1}\nabla\log f(U_t,\hat{\theta}_n)\right]'\right\} \end{aligned}$$

Temos então que

Teorema 5 ***(Teste para erro de especificação)*** *Satisfeitas as suposições de 1 à 10, se $g(U) = f(U,\theta_0)$, para $\theta_0 \in \Theta$, então*

i) $\sqrt{n}D_n(\hat{\theta_n}) \overset{A}{\sim} N(0, V(\theta_0))$;

ii) $V_n(\hat{\theta_n}) \overset{q.c.}{\to} V(\theta_0)$, *e* $V_n(\hat{\theta_n})$ *é não singular quase certamente para todo n suficientemente grande;*

iii) O teste para erro de especificação:

$$\mathcal{I}_n = nD_n(\hat{\theta_n})'(V_n(\hat{\theta_n}))^{-1}D_n(\hat{\theta_n}) \tag{2.5}$$

tem distribuição assintótica χ^2_q.

Para realizar o teste calcula-se $\mathcal{I}_n$ e compara-se com o valor crítico da distribuição χ^2_q para um dado tamanho de teste. Se 2.5 não exceder este valor, não se pode rejeitar a hipótese nula de que o modelo foi especificado corretamente[32].

Capítulo 3

Testes de Erro de Especificação para Modelos de Grafos Aleatórios

Neste trabalho construímos testes de erros de especificação para os modelos ERG e SBM a partir dos testes desenvolvidos por White (1982)[32] apresentados na seção anterior. Esses testes são importantes pois ambos os modelos tem sido bastante utilizados na prática para modelar redes sociais e esta ferramenta que propomos poderá ser usada para verificar se o modelo é realmente adequado para o banco de dados

Mostraremos primeiro que os modelos ERG e SBM satisfazem as condições de regularidade descritas nas suposições de 1 à 10. Desta forma, provaremos que as estatísticas $\mathcal{I}_n$ podem ser usadas para construir testes de hipóteses para verificar a adequação desses modelos a bancos de dados de redes. Em seguida, encontraremos as matrizes auxiliares necessárias para a construção das estatísticas dos teses em cada um dos casos e apresentaremos os testes para cada modelo.

A construção dos testes de especificação é possível devido a dois motivos: *P*rimeiro Observamos que as funçôes de verossimilhanmça desses modelos podem ser escritas em função das distribuições de probabilidades de seus vértices e arestas, dessa forma uma única amostra de uma rede pode ser tomada como a amostra de n vértices ou m arestas dessa rede. *S*egundo É possível obter estimadores de máxima verossimilhança assintóticamente consistentes tanto para o ERG quanto para o SBM, como descrito na Seção 2.2.

3.1 Testes de Erro de Especificação para modelos de Grafos Aleatórios Exponenciais

Nesta seção vamos construir um teste de erro de especificação para o ERG. Vimos que foi possível escrever a distribuição de probabilidade de um ERG em função das

distribuições de probabilidades de suas arestas, o que possibilitou a obtenção de uma função de quase-verossimilhança para o modelo. Verificamos que todas as condições de regularidade da função do ERG são válidas, assim foi possível obter um estimador de quase-verossimilhança, estimadores assintoticamente consistentes para as matrizes auxiliares e um teste de erro de especificação para o modelo.

Consideramos o modelo com apenas um parâmetro, θ, e com a função $T(G) = e(G) = k$ igual ao número de arestas do grafo.

Proposição 1 *Para o ERG a distribuição de probabilidade de um grafo G é igual ao produtório das distribuições de probabilidade de suas arestas. Ou seja,*

$$P(k;\theta) = e^{k\theta}(1+e^{\theta})^{-\binom{n}{2}} = \prod_{t=1}^{\binom{n}{2}} e^{U_t\theta}(1+e^{\theta})^{-1}$$

Demonstração 1

Resolvendo a equação $1 = \sum_{G\in\mathcal{G}} P(\theta;G)$, encontramos a constante de normalização $\varphi(\theta)$:

$$1 = \sum_{i=0}^{\binom{n}{2}} \sum_{G\in\mathcal{G}(e(G)=i)} \exp(i\theta - \varphi(\theta)) = e^{-\varphi(\theta)} \sum_{i=0}^{\binom{n}{2}} \binom{\binom{n}{2}}{i} e^{\theta i} = e^{-\varphi(\theta)} \left(1+e^{\theta}\right)^{\binom{n}{2}}.$$

Segue que $e^{-\varphi(\theta)} = z^{-1}(\theta) = \left(1+e^{\theta}\right)^{-\binom{n}{2}}$. Então, a função de probabilidade de um grafo particular G, com k arestas é dada por:

$$P(G;\theta) = e^{k\theta}\left(1+e^{\theta}\right)^{-\binom{n}{2}} = \left(\frac{e^{\theta}}{1+e^{\theta}}\right)^{k} \left(1-\frac{e^{\theta}}{1+e^{\theta}}\right)^{\binom{n}{2}-k}$$

que é a expressão 2.1 do modelo Erdos-Rényi com parâmetro $p = \left(\frac{e^{\theta}}{1+e^{\theta}}\right)$. Assim, os modelos ER com $p \neq 0, 1$ são grafos aleatórios exponenciais.

Nesse caso, a função de probabilidade de um grafo particular G, com k arestas é dada por:

$$P(k;\theta) = e^{k\theta}(1+e^{\theta})^{-\binom{n}{2}},$$

Vamos usar como amostra U_t, $t = 1, \ldots, \binom{n}{2}$, os $\binom{n}{2}$ elementos abaixo da diagonal principal da matriz Y em que Y_{ij} é a variável aleatória de Bernoulli $\left(\frac{e^{\theta}}{1+e^{\theta}}\right)$, variável indicadora da existência da aresta entre os vértices i e j, para $i = 1, \ldots, n$ e $j = 1, \ldots, n$. Tomaremos os elementos abaixo da diagonal principal de Y na seguinte ordem: coluna por coluna, da coluna 1 até a coluna $n-1$, no sentido da linha de menor índice para a linha n. Dessa forma:

$$
\begin{array}{cccc}
U_1 = y_{2,1} & \cdots & & \\
U_2 = y_{3,1} & U_{(n-1)+1} = y_{3,2} & \cdots & \\
U_3 = y_{4,1} & U_{(n-1)+2} = y_{4,2} & U_{(n-1)+(n-2)+1} = y_{4,3} & \cdots \\
U_4 = y_{5,1} & U_{(n-1)+3} = y_{5,2} & U_{(n-1)+(n-2)+2} = y_{5,3} & U_{(n-1)+(n-2)+(n-3)+1} = y_{5,4} \\
U_5 = y_{6,1} & U_{(n-1)+4} = y_{6,2} & U_{(n-1)+(n-2)+3} = y_{6,3} & U_{(n-1)+(n-2)+(n-3)+2} = y_{6,4} \\
\vdots & \vdots & \vdots & \vdots \\
U_{n-1} = y_{n,1} & U_{(n-1)+(n-2)} = y_{n,2} & U_{(n-1)+(n-2)+(n-3)} = y_{n,3} & U_{(n-1)+(n-2)+(n-3)+(n-4)} = y_{n,4}
\end{array}
$$

Assim, temos que $k = \sum_{t=1}^{\binom{n}{2}} U_t$ *e*

$$P(k;\theta) = e^{k\theta}(1+e^{\theta})^{-\binom{n}{2}} = \prod_{t=1}^{\binom{n}{2}} e^{U_t\theta}(1+e^{\theta})^{-1}$$

ou seja, $P(k;\theta) = \prod_{t=1}^{\binom{n}{2}} f(U_t;\theta)$ *em que* $f(U_t;\theta) = e^{U_t\theta}(1+e^{\theta})^{-1}$.

Proposição 2 *As condições de regularidade dadas nas suposições 1 a 10 são válidas para* $f(U_t;\theta)$ *no ERG.*

Demonstração 2

1. *Como* H *é desconhecida, a priori, escolhemos uma família de funções de distribuição que pode ou não conter a estrutura verdadeira,* H*. No caso do ERG,* H *tem distribuição Bernoulli* $\left(\frac{e^{\theta}}{1+e^{\theta}}\right)$ *para todo* U_t*, satisfazendo a suposição 1.*

2. *Para o ERG,* $f(U_t,\theta) = e^{U_t\theta}(1+e^{\theta})^{-1}$ *é contínua em* θ *para cada* $U_t \in \{0,1\}$*, satisfazendo a suposição 2.*

3. $E(\log(h(U_t))$ *existe e* $|\log f(U_t,\theta)| = |\theta U_t - \log(1+e^{\theta})|$ *para todo* $\theta \in \mathbb{R}$ *é integravel com respeito a* H*.*

   Desse modo, a suposição 3 é satisfeita.

4. $$\frac{\partial \log f(U_t,\theta)}{\partial\theta} = \left(U_t - \frac{e^{\theta}}{1+e^{\theta}}\right)$$

   é continuamente diferenciável em θ *para cada* $U_t \in \{0,1\}$*, satisfazendo a suposição 4.*

5. *Temos que* $\left|\frac{\partial^2 \log f(u,\theta)}{\partial\theta^2}\right| = \frac{e^{\theta}}{(1+e^{\theta})^2}$ *e* $\left|\frac{\partial \log f(u,\theta)}{\partial\theta}.\frac{\partial \log f(u,\theta)}{\partial\theta}\right| = \left(U_t - \frac{e^{\theta}}{1+e^{\theta}}\right)^2$ *são dominadas por funções integráveis com respeito a* H *para todos os* U_t *em* $\{0,1\}$ *e* θ *em* $\mathbb{R}$*. Assim suposição 5 satisfeita.*

6. *No caso do ERG* $\theta_* \in \mathbb{R}$, $B(\theta_*)$ *é não-singular e* θ_* *é ponto regular de* $A(\theta)$ *o que satisfaz a suposição 6.*

7. *Notemos que* $\dfrac{\partial[\frac{\partial \log f(u,\theta)}{\partial\theta}.f(u,\theta)]}{\partial\theta} = \dfrac{e^{U_t\theta}(2U_t^2 - 2U_t - 1)e^{\theta} + (U_t-1)^2e^{2\theta}}{(1+e^{\theta})^3}$
é integrável com respeito a ν *para todo* $\theta \in \mathbb{R}$ *sendo verificada a suposição 7.*

8. *No ERG* $\dfrac{\partial d_1(u,\theta)}{\partial\theta_k} = \left(-2U_t\dfrac{e^{\theta}}{(1+e^{\theta})^2} + e^{\theta}\left(\dfrac{3e^{\theta}-1}{(1+e^{\theta})^3}\right)\right)$ *é uma função contínua de* θ *para cada* U_t*, sendo satisfeita a suposição 8.*

9. *As funções:*

$$|d_1(u,\theta)d_1(u,\theta)| = \left[U_t^2 - 2U_t\frac{e^{\theta}}{1+e^{\theta}} + e^{\theta}\left(\frac{e^{\theta}-1}{(1+e^{\theta})^2}\right)\right]^2,$$

$$\left|\frac{\partial d_1(U_t,\theta)}{\partial\theta}\right| = \left|\left(-2U_t\frac{e^{\hat{\theta}}}{(1+e^{\hat{\theta}})^2} + e^{\hat{\theta}}\left(\frac{3e^{\hat{\theta}}-1}{(1+e^{\hat{\theta}})^3}\right)\right)\right| \quad e$$

$$\left|d_1(u,\theta).\frac{\partial \log(u,\theta)}{\partial\theta}\right| = \left|\left[U_t^2 - 2U_t\frac{e^{\theta}}{1+e^{\theta}} + e^{\theta}\left(\frac{e^{\theta}-1}{(1+e^{\theta})^2}\right)\right].\left[U_t - \frac{e^{\theta}}{1+e^{\theta}}\right]\right|$$

são integráveis com respeito a H *para todo* U_t *e* θ *em* $\mathbb{R}$*, sendo satisfeita a suposição 9.*

10. *No ERG,* $V(\theta_*)$ *é não-singular, sendo verificada a suposição 10.*

Como são válidas as suposições de 1 à 10, obtemos um QMLE, estimadores assintoticamente consistentes para as matrizes auxiliares e um teste de erro de especificação para o ERG.

Teorema 6 *Dados os pressupostos 1 e 2, para todo n existe um QMLE* $\hat{\theta}_n$*.*

Demonstração 3

De fato para o ERG,

$$L_n(U,\theta) \equiv \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\log f(U_t,\theta) = \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\theta U_t - \log(1+e^{\theta})\right) =$$

$$= \left(\theta\frac{\sum_{t=1}^{\binom{n}{2}}U_t}{\binom{n}{2}} - log(1+e^{\theta})\right)$$

possui máximo para $\theta \in \mathbb{R}$*.*

Nesse caso o máximo é dado por $\hat{\theta}_n = \log\left(\dfrac{\frac{\sum_{t=1}^{\binom{n}{2}}U_t}{\binom{n}{2}}}{1 - \frac{\sum_{t=1}^{\binom{n}{2}}U_t}{\binom{n}{2}}}\right)$.

Teorema 7 ***(Teste para erro de especificação em Grafos Aleatórios Exponenciais)*** *Satisfeitas as suposições de 1 à 10, se $h(U) = f(U, \theta_0)$, para $\theta_0 \in \Theta$, então*

$$\mathcal{I}_n = \frac{1}{V_n(\hat{\theta})} \left(\binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(U_t^2 - 2U_t \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} + e^{\hat{\theta}} \left(\frac{e^{\hat{\theta}}-1}{(1+e^{\hat{\theta}})^2} \right) \right) \right)^2 \tag{3.1}$$

tem distribuição assintótica χ^2_1.

Para um teste de hipóteses com α de significância, calcula-la se 3.1 e usa-se o critério: se $\mathcal{I}_n \leq \chi^2_{(\alpha,1)}$, em que $\chi^2_{(\alpha,1)}$ é o valor da distribuição acumulada da χ^2_1 em α, o modelo foi bem especificado, caso contrário o modelo foi mal especificado.

Demonstração 4

Seja $\hat{\theta} = \hat{\theta}_n$, vamos definir as seguintes matrizes auxiliares:

$$A_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[\frac{\partial^2 \log f(U_t, \hat{\theta})}{\partial \theta^2} \right] = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(-\frac{e^{\hat{\theta}}}{(1+e^{\hat{\theta}})^2} \right) = -\frac{e^{\hat{\theta}}}{(1+e^{\hat{\theta}})^2},$$

$$B_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[\frac{\partial \log f(U_t, \hat{\theta})}{\partial \theta} . \frac{\partial \log f(U_t, \hat{\theta})}{\partial \theta} \right] = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\left(U_t - \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} \right)^2 \right).$$

$$C_n(\hat{\theta}) = A_n(\hat{\theta})^{-1} B_n(\hat{\theta}) A_n(\hat{\theta})^{-1} = \frac{\left(1+e^{\hat{\theta}}\right)^4}{e^{2\hat{\theta}}} \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\left(U_t - \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} \right)^2 \right),$$

Vamos definir

$$d_l(U, \theta) = \frac{\partial \log f(U_t, \theta)}{\partial \theta_i} . \frac{\partial \log f(U_t, \theta)}{\partial \theta_j} + \frac{\partial^2 \log f(U_t, \theta)}{\partial \theta_i \partial \theta_j},$$

$l = 1, \ldots, p(p+1)/2; i = 1, \ldots, p; j = 1, \ldots, p$. Em que p é o número de coordenadas do vetor θ.

No caso do ERG, temos apenas um parâmetro, dessa forma calculamos o d_1, dado por

$$d_1(U_t, \theta) = \frac{\partial \log f(U_t, \theta)}{\partial \theta} . \frac{\partial \log f(U_t, \theta)}{\partial \theta} + \frac{\partial^2 \log f(U_t, \theta)}{\partial \theta^2} = U_t^2 - 2U_t \frac{e^{\theta}}{1+e^{\theta}} + e^{\theta} \left(\frac{e^{\theta}-1}{(1+e^{\theta})^2} \right)$$

O teste será baseado em $D_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} d_l(U_t, \hat{\theta})$, que são os elementos de $A_n(\hat{\theta}) + B(\hat{\theta})$.

Seja $l = 1, \dots\ q = p(p+1)/2$, *defina o vetor* $d(U_t, \theta)$, *de dimensão* $q \times 1$, *assim*

$$d(U_t, \theta) = U_t^2 - 2U_t \frac{e^\theta}{1+e^\theta} + e^\theta \left(\frac{e^\theta - 1}{(1+e^\theta)^2} \right)$$

Então $D_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} d(U_t, \hat{\theta})$. *Para o ERG,*

$$D_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(U_t^2 - 2U_t \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} + e^{\hat{\theta}} \left(\frac{e^{\hat{\theta}} - 1}{(1+e^{\hat{\theta}})^2} \right) \right)$$

A partir de $D_n(\hat{\theta})$, *definimos:*

$$\nabla D_n(\theta) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[\frac{\partial d_1(U_t, \hat{\theta})}{\partial \theta} \right] = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(-2U_t \frac{e^{\hat{\theta}}}{(1+e^{\hat{\theta}})^2} + e^{\hat{\theta}} \left(\frac{3e^{\hat{\theta}} - 1}{(1+e^{\hat{\theta}})^3} \right) \right)$$

$V(\theta_*)$ *é a matriz de covariância assintótica de* $\sqrt{n} D_n(\hat{\theta})$ *e temos que um estimador consistente para* $V(\theta_*)$ *é*

$$V_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[d(U_t, \hat{\theta}) - \nabla D_n(\hat{\theta}) A_n(\hat{\theta})^{-1} \nabla \log f(U_t, \hat{\theta}) \right] . \left[d(U_t, \hat{\theta}) - \nabla D_n(\hat{\theta}) A_n(\hat{\theta})^{-1} \nabla \log f(U_t, \hat{\theta}) \right]' .$$

No ERG:

$$V_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[\left(U_t^2 - 2U_t \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} + e^{\hat{\theta}} \left(\frac{e^{\hat{\theta}} - 1}{(1+e^{\hat{\theta}})^2} \right) \right) + \right.$$

$$\left. + \left(\binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(U_t^2 - 2U_t \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} + e^{\hat{\theta}} \left(\frac{e^{\hat{\theta}} - 1}{(1+e^{\hat{\theta}})^2} \right) \right) \right) . \frac{(1+e^{\hat{\theta}})^2}{e^{\hat{\theta}}} . \left(U_t - \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} \right) \right]^2 .$$

Daí,

$$\mathcal{I}_n = \frac{1}{V_n(\hat{\theta})} \left(\binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(U_t^2 - 2U_t \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} + e^{\hat{\theta}} \left(\frac{e^{\hat{\theta}} - 1}{(1+e^{\hat{\theta}})^2} \right) \right) \right)^2$$

tem distribuição assintótica χ_1^2.

3.2 Testes de Erro de Especificação para modelos de Blocos Estocásticos

Nesta seção vamos construir um teste de erro de especificação para o SBM. Vimos que foi possível escrever a distribuição de probabilidade de um SBM em função das

distribuições de probabilidades de suas arestas, quando são dadas as classes e o número de ligações de cada um dos seus vértices, o que possibilitou a obtenção de uma função de quase-verossimilhança para o modelo. Verificamos que todas as condições de regularidade da função SBM são válidas, assim foi possível obter um estimador de quase-verossimilhança, estimadores assintoticamente consistentes para as matrizes auxiliares e um teste de erro de especificação para o SBM.

A função de distribuição de probabilidade para um grafo no SBM é dada por:

$$P(\theta, \eta; X, Y) = \theta_1^{n_1} \cdots \theta_m^{n_m} \prod_{1 \leq k \leq l \leq m} \eta_{kl}^{e_{kl}} (1 - \eta_{kl})^{n_{kl} - e_{kl}},$$

onde $n_k = \sum_{i=1}^{n} I(X_i = k)$ denota o número de vértices de G que pertencem ao bloco k,

$$e_{kl} = \sum_{1 \leq i \neq j \leq n} Y_{ij} I(x_i = k) I(x_j = l).$$

denota o número de arestas de G que tem um vértice no bloco k e um vértice no bloco j, e

$$n_{kl} = \begin{cases} n_k n_l & \text{se } k \neq l \\ \binom{n_k}{2} & \text{se } k = l, \end{cases}$$

Para construirmos o Teste de Erro de Especificação no caso do SBM precisamos reescrever
$P(\theta, \eta; X, Y)$ em função da variável aleatória U_t. A Proposição a seguir mostra como fazer isso.

Proposição 3 *Para o SBM, a distribuição de probabilidade de um grafo G é igual ao produtório das distribuições de probabilidade de suas arestas, quando são dadas os blocos e o número de ligações de cada um dos seus vértices. Ou seja,*

$$P(\theta, \eta; X, Y) = \prod_{t=1}^{\binom{n}{2}} f(U_t, \theta, \eta) = \prod_{t=1}^{\binom{n}{2}} \left(\theta_k^{\frac{I_{x_i = k}}{n_i}} \theta_l^{\frac{I_{x_j = l}}{n_j}} \right) \eta_{kl}^{y_{ij}} (1 - \eta_{kl})^{(1 - y_{ij})}.$$

Demonstração 5

Vamos construir o vetor amostral U_t, $t = 1, \ldots, \binom{n}{2}$, para cada aresta do grafo G, vamos associar a cada aresta t dois vértices i^t (vértice i da aresta t) e j^t (vértice j da aresta t) , U_t será obtido a partir das matrizes amostrais X e Y.

O vetor U_t, de 5 dimensões, para cada aresta t, deve conter as seguintes informações:

- *o bloco k do vértice i^t, denotaremos este bloco por k^t;*
- *o bloco l do vértice j^t, denotaremos este bloco por l^t;*
- *o número de ligações do vértice i^t: $n_i^t = \sum_{p=1}^{n} y_{ip}$;*

- *o número de ligações do vértice* j^t*:* $n_j^t = \sum_{p=1}^{n} y_{jp}$*;*
- *a variável aleatória* y_{ij}^t *de Bernoulli*(η_{kl})*(variável indicadora da aresta* t *entre os vértices* i^t *e* j^t*.)*

Ou seja, $U_t = \left(k^t, l^t, n_i^t, n_j^t, y_{ij}^t\right)$, $t = 1, \ldots, \binom{n}{2}$.

Essa construção de U_t *é plausível pois é possível associar a cada aresta* t *dois vértices,* i^t*(vértice* i *da aresta* t*) e* j^t*(vértice* j *da aresta* t*), basta notar que para os* $\binom{n}{2}$ *elementos* y_{ij} *que estão abaixo da diagonal principal da matriz* Y*, o seu vértice* i^t *está no bloco* x_i*(onde* i *é o índice da linha do elemento* y_{ij}*) e o vértice* j^t *está no bloco* x_j *(onde* j *é o índice da coluna do elemento* y_{ij}*), para cada* t*. Dessa forma mais uma vez tomaremos os elementos abaixo da diagonal principal de* Y *na seguinte ordem: coluna por coluna, da coluna* 1 *até a coluna* $n-1$*, no sentido da linha de menor índice para a linha* n*. Assim:*

$$
\begin{array}{lll}
U_1 = (x_2, x_1, n_2, n_1, y_{2,1}) & & \\
U_2 = (x_3, x_1, n_3, n_1, y_{3,1}) & U_{(n-1)+1} = (x_3, x_2, n_3, n_2, y_{3,2}) & \\
U_3 = (x_4, x_1, n_4, n_1, y_{4,1}) & U_{(n-1)+2} = (x_4, x_2, n_4, n_2, y_{4,2}) & U_{(n-1)+(n-2)+1} = (x_4, x_3, n_4, n_3, y_{4,3}) \\
U_4 = (x_5, x_1, n_5, n_1, y_{5,1}) & U_{(n-1)+3} = (x_5, x_2, n_5, n_2, y_{5,2}) & U_{(n-1)+(n-2)+2} = (x_5, x_3, n_5, n_3, y_{5,3}) \\
U_5 = (x_6, x_1, n_6, n_1, y_{6,1}) & U_{(n-1)+4} = (x_6, x_2, n_6, n_2, y_{6,2}) & U_{(n-1)+(n-2)+3} = (x_6, x_3, n_6, n_3, y_{6,3}) \\
\vdots & \vdots & \vdots \\
U_{n-1} = (x_n, x_1, n_n, n_1, y_{n,1}) & U_{(n-1)+(n-2)} = (x_n, x_2, n_n, n_2, y_{n,2}) & U_{(n-1)+(n-2)+(n-3)} = (x_n, x_3, n_n, n_3, y_{n,3})
\end{array}
$$

$$
\begin{array}{l}
U_{(n-1)+(n-2)+(n-3)+1} = (x_5, x_4, n_5, n_4, y_{5,4}) \\
U_{(n-1)+(n-2)+(n-3)+2} = (x_6, x_4, n_6, n_4, y_{6,4}) \\
\vdots \\
U_{(n-1)+(n-2)+(n-3)+(n-4)} = (x_n, x_4, n_n, n_4, y_{n,4}) \quad \ldots \quad U_{\binom{n}{2}} = (x_n, x_{(n-1)}, n_n, n_{(n-1)}, y_{n,n-1})
\end{array}
$$

Daí, temos que

$$
P(\theta, \eta; X, Y) = \theta_1^{n_1} \cdots \theta_m^{n_m} \prod_{1 \le k \le l \le m} \eta_{kl}^{e_{kl}} (1 - \eta_{kl})^{n_{kl} - e_{kl}} = \prod_{t=1}^{\binom{n}{2}} \theta_k^{\frac{I_{x_{it}=k}}{n_i^t}} \theta_l^{\frac{I_{x_{jt}=l}}{n_j^t}} \eta_{kl}^{y_{ij}^t} (1 - \eta_{kl})^{\left(1 - y_{ij}^t\right)}
$$

ou seja, $P(\theta, \eta; X, Y) = \prod_{t=1}^{\binom{n}{2}} f(U_t; \theta, \eta)$

em que $f(U_t, \theta, \eta) = \left(\theta_k^{\frac{I_{x_{it}=k}}{n_i^t}} \theta_l^{\frac{I_{x_{jt}=l}}{n_j^t}} \right) \eta_{kl}^{y_{ij}^t} (1 - \eta_{kl})^{\left(1 - y_{ij}^t\right)}$

Para simplificar um pouco a notação vamos usar $U_t = \left(k^t, l^t, n_i^t, n_j^t, y_{ij}^t\right) = (k, l, n_i, n_j, y_{ij})$*, ou seja, vamos omitir o subscrito* t *em todos os elementos de* U_t*, sabendo que cada um deles depende de* t*.*

Desse modo $f(U_t, \theta, \eta) = \left(\theta_k^{\frac{I_{x_i=k}}{n_i}} \theta_l^{\frac{I_{x_j=l}}{n_j}} \right) \eta_{kl}^{y_{ij}} (1 - \eta_{kl})^{(1 - y_{ij})}$

Proposição 4 *As condições de regularidade dadas nas suposições 1 a 10 são válidas para* $f(U_t, \theta, \eta)$ *no SBM.*

Demonstração 6

1. *Como* H *é desconhecido a priori, escolhemos uma família de funções de distribuição que pode ou não conter a estrutura verdadeira,* H*. Para o SBM,* H *tem distribuição conjunta* $P(\theta, \eta; X, Y)$ *com* $P(x_i = k) = \theta_k \in (0,1)$ *e* $\eta_{kl} \in (0,1)$*, para todo* $k, l = 1, \ldots, m$*, satisfazendo a suposição 1.*

   Para simplificar a notação, denotaremos o conjunto (θ, η) *apenas por* θ *desse ponto em diante.*

2. *Para o SBM,* $f(U_t, \theta) = \left(\theta_k^{\frac{\boldsymbol{I}_{x_i=k}}{n_i}} \theta_l^{\frac{\boldsymbol{I}_{x_j=l}}{n_j}} \right) \eta_{kl}^{y_{ij}} (1-\eta_{kl})^{(1-y_{ij})}$ *é contínua em* θ *para cada* $U_t \in \Omega$*, satisfazendo a suposição 2.*

3. $E(\log(h(U_t))$ *existe e*

$$|\log f(U_t, \theta)| = \left| \left(\frac{\boldsymbol{I}_{x_i=k}}{n_i} \right) \log(\theta_k) + \left(\frac{\boldsymbol{I}_{x_j=l}}{n_j} \right) \log(\theta_l) + y_{ij} \log(\eta_{kl}) + (1 - y_{ij}) \log(1 - \eta_{kl}) \right|$$

   para todo $\theta \in \Theta$ *é integravel com respeito a* H*. Desse modo, a suposição 3 é satisfeita.*

4. *As funções*

$$\frac{\partial \log f(U_t, \theta)}{\partial \theta_k} = \frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i} \right)}{\theta_k}$$

$$\frac{\partial \log f(U_t, \theta)}{\partial \theta_l} = \frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j} \right)}{\theta_l}$$

$$\frac{\partial \log f(U_t, \theta)}{\partial \eta_{kl}} = \frac{y_{ij}}{\eta_{kl}} - \frac{(1 - y_{ij})}{1 - \eta_{kl}}$$

   são continuamente diferenciável em θ *para cada* $U_t \in \Omega$*, satisfazendo a suposição 4.*

5. *Vejamos que o módulo dos produtos das derivadas de primeira ordem e o módulo das derivadas de segunda ordem de* $\log f(U_t, \theta)$ *em relação a cada um dos parâmetros, são dominadas por funções integráveis com respeito a* H *para todos os* U_t *em* Ω *e* θ *em* Θ*. Assim a suposição 5 será satisfeita. Seguem as funções:*

$$\left| \left(\frac{\partial \log f(U_t, \theta)}{\partial \theta_k} \right)^2 \right| = \left| \frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i} \right)^2}{\theta_k^2} \right|$$

$$\left| \frac{\partial \log f(U_t, \theta)}{\partial \theta_k} \frac{\partial \log f(U_t, \theta)}{\partial \theta_l} \right| = \left| \frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i} \right) \left(\frac{\boldsymbol{I}_{x_j=l}}{n_j} \right)}{\theta_k \theta_l} \right|$$

$$\left|\frac{\partial \log f(U_t,\theta)}{\partial \theta_k}\frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}}\right| = \left|\frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k}\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)\right|$$

$$\left|\frac{\partial \log f(U_t,\theta)}{\partial \theta_l}\frac{\partial \log f(U_t,\theta)}{\partial \theta_k}\right| = \left|\frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l\theta_k}\right|$$

$$\left|\left(\frac{\partial \log f(U_t,\theta)}{\partial \theta_l}\right)^2\right| = \left|\frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)^2}{\theta_l^2}\right|$$

$$\left|\frac{\partial \log f(U_t,\theta)}{\partial \theta_l}\frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}}\right| = \left|\frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l}\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)\right|$$

$$\left|\frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}}\frac{\partial \log f(U_t,\theta)}{\partial \theta_k}\right| = \left|\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)\frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k}\right|$$

$$\left|\frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}}\frac{\partial \log f(U_t,\theta)}{\partial \theta_l}\right| = \left|\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)\frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l}\right|$$

$$\left|\left(\frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}}\right)^2\right| = \left|\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)^2\right|$$

$$\left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \theta_k^2}\right| = \left|\frac{-\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k^2}\right|, \left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \theta_k \partial \theta_l}\right| = 0, \left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \theta_k \partial \eta_{kl}}\right| = 0$$

$$\left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \theta_l \partial \theta_k}\right| = 0, \left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \theta_l^2}\right| = \left|\frac{-\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l^2}\right|, \left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \theta_l \partial \eta_{kl}}\right| = 0$$

$$\left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \eta_{kl} \partial \theta_k}\right| = 0, \left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \eta_{kl} \partial \theta_l}\right| = 0, \left|\frac{\partial^2 \log f(U_t,\theta)}{\partial \eta_{kl}^2}\right| = \left|\frac{-y_{ij}}{\eta_{kl}^2} - \frac{(1-y_{ij})}{(1-\eta_{kl})^2}\right|$$

que são todas dominadas por funções integráveis com respeito a H.

6. *No caso do SBM $\theta_* \in \mathbb{R}$, $B(\theta_*)$ é não-singular e θ_* é ponto regular de $A(\theta)$ o que satisfaz a suposição 6.*

7. *Notemos que*

$$\frac{\partial[\frac{\partial \log f(U_t,\theta)}{\partial \theta_k}.f(U_t,\theta)]}{\partial \theta_k} = \left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i} - 1\right)\left(\theta_k^{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}-2\right)}\theta_l^{\frac{\boldsymbol{I}_{x_j=l}}{n_j}}\right)\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{(1-y_{ij})}$$

$$\frac{\partial[\frac{\partial \log f(U_t,\theta)}{\partial \theta_l}.f(U_t,\theta)]}{\partial \theta_l} = \left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}-1\right)\left(\theta_k^{\frac{I_{x_i=k}}{n_i}}\theta_l^{\left(\frac{I_{x_j=l}}{n_i}-2\right)}\right)\eta_{kl}^{y_{ij}}(1-\eta_{kl})^{(1-y_{ij})}$$

$$\frac{\partial[\frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}}.f(U_t,\theta)]}{\partial \eta_{kl}} = \left(\theta_k^{\frac{I_{x_i=k}}{n_i}}\theta_l^{\frac{I_{x_j=l}}{n_j}}\right)\left[\left(\frac{\eta_{kl}^{y_{ij}-1}(y_{ij}-\eta_{kl})}{1-\eta_{kl}}\right).\left(\frac{y_{ij}-\eta_{kl}}{\eta_{kl}-\eta_{kl}^2}\right)+\right.$$
$$\left.+\left(\frac{(1-\eta_{kl})\,\eta_{kl}^{y_{ij}}}{1-\eta_{kl}}\right)\left(\frac{y_{ij}\,(2\eta_{kl}-1)-\eta_{kl}^2}{(1-\eta_{kl})^2\,\eta_{kl}}\right)\right]$$

são integráveis com respeito a ν para todo $\theta \in \Theta$ sendo verificada a suposição 7.

8. *No SBM*

$$\frac{\partial d_1(u,\theta)}{\partial \theta_k} = \left(\frac{-2\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)^2}{\theta_k^3}+\frac{2\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k^3}\right),\ \frac{\partial d_1(u,\theta)}{\partial \theta_l}=0,\ \frac{\partial d_1(u,\theta)}{\partial \eta_{kl}}=0,$$

$$\frac{\partial d_2(u,\theta)}{\partial \theta_k} = \left(\frac{-\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_k^2\theta_l}\right),\ \frac{\partial d_2(u,\theta)}{\partial \theta_l} = \left(\frac{-\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_k\theta_l^2}\right),$$

$$\frac{\partial d_2(u,\theta)}{\partial \eta_{kl}}=0,\ \frac{\partial d_3(u,\theta)}{\partial \theta_k} = \left(\frac{-\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k^2}\left(\frac{y_{ij}}{\eta_{kl}}-\frac{(1-y_{ij})}{1-\eta_{kl}}\right)\right),\ \frac{\partial d_3(u,\theta)}{\partial \theta_l}=0,$$

$$\frac{\partial d_3(u,\theta)}{\partial \eta_{kl}} = \left(\frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k}\left(\frac{-y_{ij}}{\eta_{kl}^2}-\frac{(1-y_{ij})}{(1-\eta_{kl})^2}\right)\right),\ \frac{\partial d_4(u,\theta)}{\partial \theta_k}=0,$$

$$\frac{\partial d_4(u,\theta)}{\partial \theta_l} = \left(\frac{-2\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)^2}{\theta_l^3}+\frac{2\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l^3}\right),\ \frac{\partial d_4(u,\theta)}{\partial \eta_{kl}}=0,$$

$$\frac{\partial d_5(u,\theta)}{\partial \theta_k}=0,\ \frac{\partial d_5(u,\theta)}{\partial \theta_l} = \left(-\frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l^2}\left(\frac{y_{ij}}{\eta_{kl}}-\frac{(1-y_{ij})}{1-\eta_{kl}}\right)\right),$$

$$\frac{\partial d_5(u,\theta)}{\partial \eta_{kl}} = \left(\frac{\left(\frac{\boldsymbol{I}_{x_j=k}}{n_j}\right)}{\theta_l}\left(\frac{-y_{ij}}{\eta_{kl}^2}-\frac{(1-y_{ij})}{(1-\eta_{kl})^2}\right)\right),\ \frac{\partial d_6(u,\theta)}{\partial \theta_k}=0,\ \frac{\partial d_6(u,\theta)}{\partial \theta_l}=0,$$

$$\frac{\partial d_6(u,\theta)}{\partial \eta_{kl}} = \left(2\left(\frac{y_{ij}}{\eta_{kl}}-\frac{(1-y_{ij})}{1-\eta_{kl}}\right)\left(\frac{-y_{ij}}{\eta_{kl}^2}-\frac{(1-y_{ij})}{(1-\eta_{kl})^2}\right)+\frac{2y_{ij}}{\eta_{kl}^3}-\frac{2\,(1-y_{ij})}{(1-\eta_{kl})^3}\right),$$

são funções contínuas de $\theta \in \Theta$ para cada $U_t \in \Omega$, sendo satisfeita a suposição 8.

9. *Todas as funções:*

$d_i(u,\theta)d_j(u,\theta)$, $\left|\frac{\partial d_i(U_t,\theta)}{\partial \theta_k}\right|$, $\left|\frac{\partial d_i(U_t,\theta)}{\partial \theta_l}\right|$, $\left|\frac{\partial d_i(U_t,\theta)}{\partial \eta_{kl}}\right|$, $\left|d_i(u,\theta).\frac{\partial \log(u,\theta)}{\partial \theta_k}\right|$, $\left|d_i(u,\theta).\frac{\partial \log(u,\theta)}{\partial \theta_l}\right|$, $\left|d_i(u,\theta).\frac{\partial \log(u,\theta)}{\partial \eta_{kl}}\right|$,

são produtos de funções integráveis com respeito a H, logo essas funções são integráveis com respeito a H para todo $1 \leq i \leq 6$, U_t e $\theta \in \Theta$, sendo satisfeita a suposição 9.

10. *No SBM, $V(\theta_*)$ é não-singular, sendo verificada a suposição 10.*

Como são válidas as suposições de 1 à 10, obtemos um QMLE, estimadores assintoticamente consistentes para as matrizes auxiliares e um teste de erro de especificação para o SBM.

Teorema 8 *Dados os pressupostos 1 e 2, para todo n existe um QMLE $\hat{\theta}_n$.*

Demonstração 7

De fato para o SBM, $L_n(U,\theta) \equiv \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \log f(U_t,\theta) =$

$$\binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i} \right) \log(\theta_k) + \left(\frac{\boldsymbol{I}_{x_j=l}}{n_j} \right) \log(\theta_l) + y_{ij} \log(\eta_{kl}) + (1-y_{ij}) \log(1-\eta_{kl}) \right]$$

é mensurável e possui máximo para $\theta \in \Theta$.

Nesse caso o máximo pode ser obtido via EM Variacional e nesse trabalho usaremos as estimativas via EM Variacional, obtidas pelo pacote *blockmodels* do software R.

Teorema 9 ***(Teste para erro de especificação em SBM)*** *Satisfeitas as suposições de 1 à 10, se $h(U) = f(U,\theta_0)$, para $\theta_0 \in \Theta$, então*

$$\mathcal{I}_n = n D_n(\hat{\theta})' (V_n(\hat{\theta}))^{-1} D_n(\hat{\theta}) \tag{3.2}$$

tem distribuição assintótica χ^2_6.

Para um teste de hipóteses com α de significância, calcula-la se 3.2 e usa-se o critério: se $\mathcal{I}_n \leq \chi^2_{(\alpha,6)}$, em que $\chi^2_{(\alpha,6)}$ é o valor da distribuição acumulada da χ^2_6 em α, o modelo foi bem especificado, caso contrário o modelo foi mal especificado.

Demonstração 8

Seja $\hat{\theta} = \hat{\theta}_n$, vamos definir as seguintes matrizes auxiliares:

$$\nabla \log f(U_t,\theta) = \begin{pmatrix} \frac{\partial \log f(U_t,\theta)}{\partial \theta_k} \\ \frac{\partial \log f(U_t,\theta)}{\partial \theta_l} \\ \frac{\partial \log f(U_t,\theta)}{\partial \eta_{kl}} \end{pmatrix} = \begin{pmatrix} \frac{\left(\frac{I_{x_i=k}}{n_i} \right)}{\theta_k} \\ \frac{\left(\frac{I_{x_j=l}}{n_j} \right)}{\theta_l} \\ \left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}} \right) \end{pmatrix}$$

$$A_n(\hat{\theta}) = \begin{pmatrix} \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\theta_k^2}\right] & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\theta_k\partial\theta_l}\right] & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\theta_k\partial\eta_{kl}}\right] \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\theta_l\partial\theta_k}\right] & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\theta_l^2}\right] & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\theta_l\partial\eta_{kl}}\right] \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\eta_{kl}\partial\theta_k}\right] & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\eta_{kl}\partial\theta_l}\right] & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[\frac{\partial^2 \log f(U_t,\hat{\theta})}{\partial\eta_{kl}^2}\right] \end{pmatrix}$$

$$A_n(\hat{\theta}) = \begin{pmatrix} \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{-\left(\frac{I_{x_i=k}}{n_i}\right)}{\hat{\theta_k}^2}\right) & 0 & 0 \\ 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{-\left(\frac{I_{x_j=l}}{n_j}\right)}{\hat{\theta_l}^2}\right) & 0 \\ 0 & 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{-y_{ij}}{\hat{\eta_{kl}}^2} - \frac{(1-y_{ij})}{(1-\hat{\eta_{kl}})^2}\right) \end{pmatrix}$$

Vamos definir

$$d_l(U,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\theta_i}.\frac{\partial \log f(U_t,\theta)}{\partial\theta_j} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_i\partial\theta_j},$$

$l = 1,\ldots,p(p+1)/2; i = 1,\ldots,p; j = 1,\ldots,p$. *Em que* p *é o número de coordenadas do vetor* θ.

No caso do SBM, temos os parâmetro θ_k, θ_l *e* η_{kl} *para cada* U_t, *ou seja* 3 *parâmetros, dessa forma calculamos os* d_l, *para* $l = 1,2,\ldots,6$ *dados por*

$$d_1(U_t,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\theta_k}.\frac{\partial \log f(U_t,\theta)}{\partial\theta_k} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_k^2} = \frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)^2}{\theta_k^2} + \frac{-\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k^2}$$

$$d_2(U_t,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\theta_k}.\frac{\partial \log f(U_t,\theta)}{\partial\theta_l} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_k\partial\theta_l} = \frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_k\theta_l}$$

$$d_3(U_t,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\theta_k}.\frac{\partial \log f(U_t,\theta)}{\partial\eta_{kl}} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_k\partial\eta_{kl}} = \frac{\left(\frac{\boldsymbol{I}_{x_i=k}}{n_i}\right)}{\theta_k}\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)$$

$$d_4(U_t,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\theta_l}.\frac{\partial \log f(U_t,\theta)}{\partial\theta_l} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_l^2} = \frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)^2}{\theta_l^2} + \frac{-\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l^2}$$

$$d_5(U_t,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\theta_l}.\frac{\partial \log f(U_t,\theta)}{\partial\eta_{kl}} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\theta_l\partial\eta_{kl}} = \frac{\left(\frac{\boldsymbol{I}_{x_j=l}}{n_j}\right)}{\theta_l}\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)$$

$$d_6(U_t,\theta) = \frac{\partial \log f(U_t,\theta)}{\partial\eta_{kl}}.\frac{\partial \log f(U_t,\theta)}{\partial\eta_{kl}} + \frac{\partial^2 \log f(U_t,\theta)}{\partial\eta_{kl}^2} = \left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)^2 + \frac{-y_{ij}}{\eta_{kl}^2} - \frac{(1-y_{ij})}{(1-\eta_{kl})^2}$$

O teste será baseado em $D_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} d_l(U_t, \hat{\theta})$, *que são os elementos de* $A_n(\hat{\theta}) + B_n(\hat{\theta})$.

Seja $l = 1, \dots\ q = p(p+1)/2 = 6$, *defina o vetor* $d(U_t, \theta)$, *de dimensão* $q \times 1 = 6 \times 1$, *assim*

$$d(U_t,\theta) = \begin{pmatrix} \frac{\left(\frac{I_{x_i=k}}{n_i}\right)^2}{\theta_k^2} + \frac{-\left(\frac{I_{x_i=k}}{n_i}\right)}{\theta_k^2} \\ \frac{\left(\frac{I_{x_i=k}}{n_i}\right)\left(\frac{I_{x_j=l}}{n_j}\right)}{\theta_k\theta_l} \\ \frac{\left(\frac{I_{x_i=k}}{n_i}\right)}{\theta_k}\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right) \\ \frac{\left(\frac{I_{x_j=l}}{n_j}\right)^2}{\theta_l^2} + \frac{-\left(\frac{I_{x_j=l}}{n_j}\right)}{\theta_l^2} \\ \frac{\left(\frac{I_{x_j=l}}{n_j}\right)}{\theta_l}\left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right) \\ \left(\frac{y_{ij}}{\eta_{kl}} - \frac{(1-y_{ij})}{1-\eta_{kl}}\right)^2 + \frac{-y_{ij}}{\eta_{kl}^2} - \frac{(1-y_{ij})}{(1-\eta_{kl})^2} \end{pmatrix}$$

Então $D_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} d(U_t, \hat{\theta})$. *Para o SBM,*

$$D_n(\hat{\theta_n}) = \begin{pmatrix} \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\frac{\left(\frac{I_{x_i=k}}{n_i}\right)^2}{\hat{\theta_k}^2} + \frac{-\left(\frac{I_{x_i=k}}{n_i}\right)}{\hat{\theta_k}^2} \right) \\ \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\frac{\left(\frac{I_{x_i=k}}{n_i}\right)\left(\frac{I_{x_j=l}}{n_j}\right)}{\hat{\theta_k}\hat{\theta_l}} \right) \\ \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\frac{\left(\frac{I_{x_i=k}}{n_i}\right)}{\hat{\theta_k}} \left(\frac{y_{ij}}{\hat{\eta_{kl}}} - \frac{(1-y_{ij})}{1-\hat{\eta_{kl}}}\right) \right) \\ \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\frac{\left(\frac{I_{x_j=l}}{n_j}\right)^2}{\hat{\theta_l}^2} + \frac{-\left(\frac{I_{x_j=l}}{n_j}\right)}{\hat{\theta_l}^2} \right) \\ \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\frac{\left(\frac{I_{x_j=l}}{n_j}\right)}{\hat{\theta_l}} \left(\frac{y_{ij}}{\hat{\eta_{kl}}} - \frac{(1-y_{ij})}{1-\hat{\eta_{kl}}}\right) \right) \\ \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left(\left(\frac{y_{ij}}{\hat{\eta_{kl}}} - \frac{(1-y_{ij})}{1-\hat{\eta_{kl}}}\right)^2 + \frac{-y_{ij}}{\hat{\eta_{kl}}^2} - \frac{(1-y_{ij})}{(1-\hat{\eta_{kl}})^2} \right) \end{pmatrix}$$

A partir de $D_n(\hat{\theta})$, *definimos* $\nabla D_n(\hat{\theta}) = \binom{n}{2}^{-1} \sum_{t=1}^{\binom{n}{2}} \left[\frac{\partial d_l(U_t, \hat{\theta})}{\partial \theta_k} \right]$.

$$\nabla D_n(\hat{\theta}) = \begin{pmatrix} \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_1(u,\hat{\theta})}{\partial \theta_k}\right) & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_1(u,\hat{\theta})}{\partial \theta_l}\right) = 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_1(u,\hat{\theta})}{\partial \eta_{kl}}\right) = 0 \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_2(u,\hat{\theta})}{\partial \theta_k}\right) & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_2(u,\hat{\theta})}{\partial \theta_l}\right) & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_2(u,\hat{\theta})}{\partial \eta_{kl}}\right) = 0 \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_3(u,\hat{\theta})}{\partial \theta_k}\right) & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_3(u,\hat{\theta})}{\partial \theta_l}\right) = 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_3(u,\hat{\theta})}{\partial \eta_{kl}}\right) \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_4(u,\hat{\theta})}{\partial \theta_k}\right) = 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_4(u,\hat{\theta})}{\partial \theta_l}\right) & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_4(u,\hat{\theta})}{\partial \eta_{kl}}\right) = 0 \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_5(u,\hat{\theta})}{\partial \theta_k}\right) = 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_5(u,\hat{\theta})}{\partial \theta_l}\right) & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_5(u,\hat{\theta})}{\partial \eta_{kl}}\right) \\ \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_6(u,\hat{\theta})}{\partial \theta_k}\right) = 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_6(u,\hat{\theta})}{\partial \theta_l}\right) = 0 & \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left(\frac{\partial d_6(u,\hat{\theta})}{\partial \eta_{kl}}\right) \end{pmatrix}$$

As funções que devem ser somadas nas componentes de $\nabla D_n(\theta)$ *são as funções que foram obtidas na suposição 8.*

$V(\theta_*)$ *é a matriz de covariância assintótica de* $\sqrt{n}D_n(\hat{\theta})$ *e temos que um estimador consistente para* $V(\theta_*)$ *é*

$$V_n(\hat{\theta}) = \binom{n}{2}^{-1}\sum_{t=1}^{\binom{n}{2}}\left[d(U_t,\hat{\theta}) - \nabla D_n(\hat{\theta})A_n(\hat{\theta})^{-1}\nabla \log f(U_t,\hat{\theta})\right].\left[d(U_t,\hat{\theta}) - \nabla D_n(\hat{\theta})A_n(\hat{\theta})^{-1}\nabla \log f(U_t,\hat{\theta})\right]'.$$

No SBM, para cada um dos vetores U_t *vamos obter*

$$V_t(\hat{\theta}) = \left[d(U_t,\hat{\theta}) - \nabla D_n(\hat{\theta})A_n(\hat{\theta})^{-1}\nabla \log f(U_t,\hat{\theta})\right].\left[d(U_t,\hat{\theta}) - \nabla D_n(\hat{\theta})A_n(\hat{\theta})^{-1}\nabla \log f(U_t,\hat{\theta})\right]',$$

$V_t(\hat{\theta})$ *é uma matriz de dimensão* 6×6*, e* $V_n(\hat{\theta})$ *será a matriz* 6×6*, cujas entradas são a média das entradas respectivas das* $V_t(\hat{\theta})$*.*

Então,

$$\mathcal{I}_n = nD_n(\hat{\theta})'(V_n(\hat{\theta}))^{-1}D_n(\hat{\theta})$$

tem distribuição assintótica χ^2_6*.*

Capítulo 4

Simulações

Nas seções anteriores construímos testes de erro de especificação para o ERG e o SBM, especificando as estatísticas de teste e suas respectivas distribuições. Nesta Seção iremos apresentar simulações para verificar o comportamento dos testes propostos em cenários conhecidos.

Os testes foram aplicados em amostras de grafos que foram gerados variando os valores de cada parâmetro dos modelos, do número de vértices e quantidade de classes, no caso do SBM. Para aplicar os testes, simulamos amostras aleatórias de grafos em dois cenários: (1) grafos gerados a partir de cada um dos modelos e (2) grafos gerados com outros modelos. A partir das amostras estimamos os parâmetros via máxima verossimilhança e construímos os vetores amostrais U_t, que serão utilizados em cada teste. Em seguida, encontramos as matrizes auxiliares $d(U_t, \hat{\theta})$, $D_n(\hat{\theta})$, $\nabla D_n(\hat{\theta})$, $A_n(\hat{\theta})$, $\nabla \log f(U_t, \hat{\theta})$ para calcular $V_n(\hat{\theta})$ e por fim, obter $\mathcal{I}_n$ e comparar com o valor da distribuição acumulada da χ^2_q, fixado um α.

Os códigos foram implementados em R, versão 4.1.2, e executados em um computador com processador AMD Ryzen 5 $3,6$ GHz, 16GB de memória ram e $512GB$ de SSDM2. O tempo das simulaçõees em cada cenário teve relação direta com o número de vértices e quantidade de classes, no caso do SBM. Alguns testes foram executados em segundos enquanto outros levaram alguns minutos.

Para cada um dos cenários, fizemos replicações de Monte Carlo(MC) e investigamos se os testes aceitavam ou rejeitavam a hipótese do modelo como bem especificado. Nas Tabelas 4.1, 4.2, 4.3 e 4.4 é possível observar a quantidade dos testes simulados que indicam, ou não, um modelo bem especificado em cada um dos cenários.

4.1 Comportamento dos Testes de Erro de Especificação para ERG

Para o ERG, consideramos testes em que os grafos estão divididos em dois cenários. No cenário 1, geramos os grafos com distribuição do ERG com apenas um parâmetro θ, já no cenário 2, vamos gerar grafos que, intencionalmente, não tem distribuição do ERG com apenas um parâmetro θ. Em ambos os cenários variamos o número de vértices dos grafos, fizemos 1000 ou 10000 replicações de Monte Carlo(MC) e investigamos se os testes simulados indicam o modelo como bem especificado.

1. **C**enário 1: Em cada replicação geramos um α aleatório no intervalo $(0, 1)$, os grafos tem $n = 50$, $n = 100$, $n = 200$, $n = 1000$, ou $n = 10000$ vértices. A probabilidade de um vértice i se ligar a outro vértice qualquer j é exatamente α para quaisquer vértices i e j, ou seja, teoricamente, estamos em um grafo com distribuição do ERG com apenas um parâmetro $\theta = \log\left(\frac{\alpha}{1-\alpha}\right)$.

2. **C**enário 2: Em cada replicação geramos um α aleatório no intervalo $(0, 1)$, os grafos tem $n = 50$, $n = 100$, $n = 200$, $n = 1000$ ou $n = 10000$ vértices. A probabilidade de um vértice i se ligar a outro vértice j é $p_k.\alpha$ para cada $\frac{n}{10}$ vértices, em que p_k é um valor aleatório no intervalo $(0, 1)$, para $k = 1, 2, \ldots, 10$, ou seja, geramos grafos cuja probabilidade para cada grupo de $\frac{n}{10}$ vértices é uma porcentagem aleatória de α, assim, teoricamente, estamos em um grafo cuja distribuição não é a do ERG com parâmetro $\theta = \log\left(\frac{\alpha}{1-\alpha}\right)$.

O teste para erro de especificação do modelo ERG tem as seguintes hipóteses:

H_0 : o grafo gerador tem distribuição ERG(θ),

contra

H_a : o grafo gerador não tem distribuição ERG(θ).

Vamos tomar como estimativa de θ o estimador de máxima verrosimilhança dado por

$$\hat{\theta} = \log\left(\frac{\frac{\sum_{t=1}^{\binom{n}{2}} U_t}{\binom{n}{2}}}{1 - \frac{\sum_{t=1}^{\binom{n}{2}} U_t}{\binom{n}{2}}}\right).$$

Em ambos os cenários, seguindo o processo descrito na seção 3.1, encontramos os vetores amostrais U_t e as matrizes auxiliares $d(U_t, \hat{\theta})$, $D_n(\hat{\theta})$, $\nabla D_n(\hat{\theta})$, $A_n(\hat{\theta})$, $\nabla \log f(U_t, \hat{\theta})$ para calcular $V_n(\hat{\theta})$. Por fim, obtemos $\mathcal{I}_n$ e comparamos com o valor da distribuição acumulada da χ^2_1 em $\alpha = 0,05$. Se $\mathcal{I}_n$ for menor ou igual ao valor da distribuição acumulada

da χ^2_1 em $\alpha = 0,05$, não rejeitamos H_0, ou seja, o teste de hipóteses indica que o modelo foi bem especificado.

A Tabela 4.1 apresenta os testes simulados, no cenário 1, podemos inferir que para diversos valores de θ em grafos com 50, 100, 200, 1000 e 10000 vértices o teste de hipóteses com nível de 5% de significância, indicou boa adequação, como era esperado.

Número de vértices	Replicações	Aceitaram H_0	Rejeitaram H_0
50	1000	998	2
100	1000	999	1
200	1000	999	1
1000	1000	1000	0
10000	1000	1000	0
50	10000	9984	16
100	10000	9993	7
200	10000	9999	1
1000	10000	10000	0
10000	10000	10000	0

Tabela 4.1: Testes de hipóteses simulados no Cenário 1 para o ERG.

A Tabela 4.2 apresenta os testes simulados, no cenário 2, podemos inferir que para diversos valores de θ em grafos com 50, 100, 200, 1000 e 10000 vértices o teste de hipóteses com nível de 5% de significância, indicou má adequação, como também era esperado.

Número de vértices	Replicações	Aceitaram H_0	Rejeitaram H_0
50	1000	0	1000
100	1000	0	1000
200	1000	0	1000
1000	1000	0	1000
10000	1000	0	1000
50	10000	29	9971
100	10000	14	9986
200	10000	6	9994
1000	10000	0	10000
10000	10000	0	10000

Tabela 4.2: Testes de hipóteses simulados no Cenário 2 para o ERG.

4.2 Comportamento dos Testes de Erro de Especificação para SBM

Para o SBM também consideramos testes em que os grafos estão divididos em dois cenários: no **C**enário 1, geramos os grafos com distribuição do SBM com os parâmetros (θ, η). No **C**enário 2, geramos grafos aleatórios perturbando os parâmetros originais (θ, η). Em ambos cenários, variamos o número de vértices e o número de blocos dos grafos, fizemos 100 ou 1000 replicações de Monte Carlo(MC) e investigamos qual a porcentagem dos testes indicava quando o modelo estava bem especificado.

1. **C**enário 1: Em cada replicação geramos a matriz η em que η_{kl} aleatório no intervalo $(0, 1)$ e $\eta_{kl} = \eta_{lk}$, para $k = 1, 2, \ldots, m$ e $l = 1, 2, \ldots, m$, em que m é o número de blocos existêntes e geramos o vetor X tal que, para $i = 1, 2, \ldots, n$, x_i é aleatório entre $\{1, 2, \ldots, m\}$ e representa a classe do vértice i. A probabilidade de um vértice i se ligar a outro vértice qualquer j é exatamente η_{kl} se o vértice i é da classe k e o vértice j é da classe l, ou seja, um grafo com distribuição do SBM com parâmetro (θ, η) em que η é formado pelos η_{kl} e $\theta_k = \frac{\sum_{i=1}^{n} I_{x_i=k}}{m}$.

   Geramos grafos com $n = 90$ e $m = 3$ ou $m = 6$, $n = 120$ e $m = 4$ ou $m = 6$, $n = 200$ e $m = 4$ ou $m = 10$, $n = 300$ e $m = 3$ ou $m = 10$ ou $m = 15$, $n = 1000$ e $m = 4$ ou $m = 10$ ou $m = 20$ ou $m = 100$.

2. **C**enário 2: Em cada replicação geramos a matriz η em que η_{kl} aleatório no intervalo $(0, 1)$ e $\eta_{kl} = \eta_{lk}$, para $k = 1, 2, \ldots, m$ e $l = 1, 2, \ldots, m$, em que m é o número de blocos existêntes e geramos o vetor X tal que, para $i = 1, 2, \ldots, n$, x_i é aleatório entre $\{1, 2, \ldots, m\}$ e representa a classe do vértice i. A probabilidade de um vértice i se ligar a outro vértice j é $p_k.\eta$ para cada $\frac{n}{10}$ vértices, em que p_k é um valor aleatório no intervalo $(0, 1)$, para $k = 1, 2, \ldots, 10$, ou seja, geramos grafos cuja probabilidade para cada grupo de $\frac{n}{10}$ vértices é uma porcentagem aleatória de η, assim, um grafo cuja distribuição não é a do SBM com os parâmetro (θ, η) em que η é formado pelos η_{kl} e $\theta_k = \frac{\sum_{i=1}^{n} I_{x_i=k}}{m}$. Geramos grafos em que $n = 90$ e $m = 3$ ou $m = 6$, $n = 120$ e $m = 4$ ou $m = 6$, $n = 200$ e $m = 4$ ou $m = 10$ e $n = 300$ e $m = 3$ ou $m = 10$ ou $m = 30$.

O teste para erro de especificação do SBM tem as seguintes hipóteses:

H_0 : o grafo gerador tem distribuição SBM(θ, η),

contra

H_a : o grafo gerador não tem distribuição SBM(θ, η).

Vamos tomar como estimativa de (θ, η) o estimador de máxima verrosimilhança $\left(\hat{\theta}, \hat{\eta}\right)$ obtido via EM Variacional pelo pacote *blockmodels* do software R,

Em ambos os cenários, seguindo o processo descrito na seção 3.2, construímos os vetores amostrais U_t. Em seguida, construímos as matrizes auxiliares $d(U_t, \hat{\theta})$, $D_n(\hat{\theta})$, $\nabla D_n(\hat{\theta})$, $A_n(\hat{\theta})$, $\nabla \log f(U_t, \hat{\theta})$ para calcular $V_n(\hat{\theta})$ e por fim, obter $\mathcal{I}_n$ e comparar com o valor da distribuição acumulada da χ^2_6 em $\alpha = 0,05$. Se $\mathcal{I}_n$ for menor ou igual ao valor da distribuição acumulada da χ^2_6 em $\alpha = 0,05$, aceita-se H_0, ou seja, o teste de hipóteses indica que o modelo foi bem especificado como um SBM.

A Tabela 4.3 apresenta os testes simulados, no cenário 1, podemos inferir que para diversos valores de (θ, η) em grafos com variações do número de blocos e vértices o teste indicou boa adequação como era esperado.

Número de vértices	Número de blocos	Replicações	Aceitaram H_0	Rejeitaram H_0
90	3	100	100	0
90	6	100	99	1
120	4	100	100	0
120	6	100	98	2
200	4	100	99	1
200	10	100	97	3
300	3	100	100	0
300	10	100	100	0
300	30	100	99	1
90	3	1000	963	37
90	6	1000	972	28
120	4	1000	968	32
120	6	1000	981	19
200	4	1000	971	29
200	10	1000	982	18
300	3	1000	983	17
300	10	1000	989	11
300	30	1000	991	9

Tabela 4.3: Testes de hipóteses simulados no Cenário 1 para o SBM.

A Tabela 4.4 apresenta os testes simulados, no cenário 2, podemos inferir que para diversos valores de (θ, η) em grafos com variações do número de blocos e vértices o teste indicou má adequação como também era esperado.

Número de vértices	Número de blocos	Replicações	Aceitaram H_0	Rejeitaram H_0
90	3	100	0	100
90	6	100	2	98
120	4	100	0	100
120	6	100	1	99
200	4	100	0	100
200	10	100	2	98
300	3	100	1	99
300	10	100	2	98
300	30	100	2	98
90	3	1000	42	958
90	6	1000	54	946
120	4	1000	36	964
120	6	1000	44	956
200	4	1000	31	969
200	10	1000	38	962
300	3	1000	22	978
300	10	1000	24	976
300	30	1000	32	968

Tabela 4.4: Testes de hipóteses simulados no Cenário 2 para o SBM.

4.3 Conclusões

As simulações foram feitas com amostras aleatórias de grafos em dois cenários para cada um dos modelos analisados: ERG e SBM: (1) No primeiro cenário, os grafos foram gerados a partir de cada um dos modelos especificados. (2) No segundo cenário, os grafos foram gerados com outros modelos. Foram feitas as estimativas dos parâmetros via máxima verossimilhança e calculamos as estatísticas de teste para cada amostra. As Tabelas 4.1 e 4.3 são referentes ao Cenário 1, onde foram gerados grafos de acordo com os modelos ERG e SBM, mostram que a maioria dos testes simulados indicam que o modelo é adequado aos dados. Mas quando os grafos gerados não são gerados seguindo os modelos, os resultados mostrados nas Tabelas 4.2 e 4.4, referentes ao cenário 2, mostram que a maioria dos testes aplicados indicam que o modelo não é adequado aos dados.

Então, observamos que quando o teste é aplicado a uma amostra que, de fato, foi gerada de acordo com os modelos testados, ele acerta (aceita H_0) em praticamente todas as simulações. Tanto para o ERG, quanto para o SBM , Tabelas 4.1 e 4.3.

No caso do teste para erro de especificação do ERG, notamos pela Tabela 4.2 que, mesmo que o número de vértices seja consideravelmente grande, a proporção de erros é muito pequena, menos que 1% em todos as réplicas.

Observamos pela Tabela 4.4 que, mesmo quando temos um grande número de

vértices, e de blocos, a proporção de erro do teste proposto também é bem pequena, menor do que os 5% esperados do erro tipo I.

Com essas simulações ilustramos a eficácia dos testes propostos para verificar erro de especificação dos modelos ERG e SBM.

Os códigos em R utilizados para gerar os grafos e calcular os testes podem ser encontrados no endereço:

https://drive.google.com/drive/folders/1epDTdqS42853_Arrg7TzMsz3oNaLQfbG

Referências

[1] ABBE, E.; SANDOM, C. (2015) "Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms". arXiv:1503.00609.

[2] ABBE, E.; SANDOM, C. (2015) . "Recovering communities in the general stochastic block model without knowing the parameters". arXiv:1506.03729.

[3] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.),Second International Symposium on Information Theory, (pp. 267–281). Academiai Kiado: Budapest.

[4] AKAIKE, H. (1974). A new look at the statistical model identification.IEEE Transactions on Automatic Control, AC-19, 716–723.

[5] BARABÁSI A.L., ALBERT,R. (1999) Emergence of scaling in random networks. Science 286: 509–512.

[6] BESAG, J. (1974) Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. Ser. B 36(2), 192–236.

[7] CELISSE, A. , DAUDIN, J. J. AND PIERRE, L. (2012) Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics 6: 1847-1899.

[8] CORANDER, J. , DAHMSTRÖM, K. AND DAHMSTRÖM, P. (1998) Maximum likelihood estimation for Markov graphs. Research report. Stockholm: Department of Statistics, Stockholm University.

[9] CHATTERJEE, S. AND DIACONIS, P. (2011) Estimating and understanding exponential random graph models. ArXiv e-prints.

[10] ERDÖS, P. , RÉNYI A. (1959) On Random Graphs I. Budapest.

[11] FRANK, O. AND STRAUSS, D. (1986) Markov graphs. JASA.

[12] HAMMERSLEY, J.M.; CLIFFORD, P. (1971) , Markov fields on – (unpublished).

[13] LEGER, J. (2016). Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates, arXiv:1602.07587.

[14] FRANK,O. (1991), Statistical analysis of change in networks O. Frank, Statistica Neerlandica, Volume45, Issue3, 283-293.

[15] HOLLAND. P.W.; LASKEY,K.Laskey, K. B.; LEINHARDT, S., S, (1983). "Stochastic blockmodels: First steps". Social Networks. 5 (2): 109–137.

[16] HUNTER, D.R., HANDCOCK, M.S., BUTTS, C.T., Steven M. GOODREAU, S.M. and Martina MORRIS, M. (2009) ergm: A package to fit, simulate and diagnose exponential-family models for networks- Journal of statistical Software, 24(3), 1-29.

[17] LEE, C. and WILKINSON, D. J. (2019) A review of stochastic block models and extensions for graph clustering. Applied Network Science, 4:122.

[18] KARRER, B; NEWMAN, M. E. J. (2011). "Stochastic blockmodels and community structure in networks". Physical Review E. 83 (1).

[19] KOLACZYK, E.D. : CSÁARDI, G. (2014) Statistical Analysis of Network Data with R, Springer.

[20] NEWMAN, M.E.J., STROGATZ, S. H and WATTS, D. J (2001) Random graphs with arbitrary degree distributions and their applications, Physical review E, Vol 64,no 2: 026-118.

[21] NEWMAN, M. E. J. (2010). Networks: An Introduction. Oxford.

[22] PATTISON, P. ANDP WASSERMAN, S.(1999), Logit models and logistic regressions for social networks: II. Multivariate relations, British journal of mathematical and Statistical Psychology. Volume52, Issue2, 169-193.

[23] PEIXOTO, T. (2014). "Hierarchical block structures and high-resolution model selection in large networks". Physical Review X. 4 (1).

[24] REUVEN, C. and SHLOMO, H. (2010). Complex Networks: Structure, Robustness and Function. Cambridge University Press.

[25] ROBINS, G., PATTISON, S, WASSERMAN (1999) Logit models and logistic regressions for social networks: III. Valued relations, Psychometrika, 64, 371–394.

[26] ROBINS, G. AND MORRIS, M. (2007) Advances in exponential random graph (p*) models. Social Networks 29: 169-172.

[27] ROBINS, G. , PATTISON, P. , KALISH, Y. AND LUSHER D. (2007) An introduction to exponential random graph (p*) models for social networks. Social Networks 29: 173-191.

[28] SNIJDERS, T. (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of Classification 14: 75-100.

[29] SCHWARZ, G. E. (1978), "Estimating the dimension of a model", Annals of Statistics, 6 (2): 461–464.

[30] SCHMID, C. S. and DESMARAIS, B.A. Desmarais (2017) "Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap,"IEEE International Conference on Big Data (Big Data), 2017, pp. 116-121.

[31] WATTS D.J., STROGATZ, S.H. (1998) Collective dynamics of small-world networks. Nature 393: 440–442.

[32] WHITE, H. (1982) Maximum Likelihood estimation of misspecified models, Econometrica 50: 1-2.

[33] WASSERMAN, S. e PATTISON, P.(1996) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p, Psychometrika, 1996 - Springer.