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

Larissa Natany Almeida Martins

Geração e análise de dados sintéticos via Redes Bayesianas: Uma abordagem robusta para quantificação de incerteza via paradigma Bayesiano

Belo Horizonte
2024

Larissa Natany Almeida Martins

Geração e análise de dados sintéticos via Redes Bayesianas: Uma abordagem robusta para quantificação de incerteza via paradigma Bayesiano

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.

Orientador: Prof. Dr. Flávio Bambirra Gonçalves
Coorientadora: Profa. Dra. Thaís Paiva Galletti

Belo Horizonte
2024

Martins, Larissa Natany Almeida.

M386g Geração e análise de dados sintéticos via Redes Bayesianas: [recurso eletrônico] uma abordagem robusta para quantificação de incerteza via paradigma Bayesiano / Larissa Natany Almeida Martins. – 2024.
1 recurso online (76 f. il.) : pdf.

Orientador: Flávio Bambirra Gonçalves.
Coorientadora: Thaís Paiva Galletti.

Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 65-69.

1. Estatística - Teses. 2. Inferência Bayesiana – Teses. 3. Markov, Processos de – Teses. 4. Dados Sintéticos – Teses. I. Gonçalves, Flávio Bambirra. II. Galletti, Thaís Paiva. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

FOLHA DE APROVAÇÃO

"Geração e análise de dados sintéticos via Redes Bayesianas: Uma abordagem robusta para quantificação de incerteza via paradigma Bayesiano"

LARISSA NATANY ALMEIDA MARTINS

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Doutora em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada, em 29 de fevereiro de 2024, pela banca constituída pelos membros:

Prof. Flávio Bambirra Gonçalves - Orientador
DEST/UFMG

Prof. Vinícius Diniz Mayrink
DEST/UFMG

Prof. Guilherme Lopes de Oliveira
CEFET-MG

Profa. Lívia Maria Dutra
CEFET-MG

Profa. Kelly Cristina Mota Gonçalves
DME/UFRJ

Belo Horizonte, 29 de fevereiro de 2024.

Dedico este trabalho a todos aqueles que não largaram a minha mão durante este longo caminho, especialmente minha mãe.

Agradecimentos

Primeiramente, agradeço a Deus, que tornou tudo isso possível. Sou profundamente grata pela força, saúde e esperança que sempre me guiaram ao longo deste caminho.

À minha mãe, Ana Tércia, que sempre foi um exemplo de vida, perseverança e força. Você me ensinou, com seu próprio exemplo, a importância da dedicação, do trabalho árduo e da resiliência diante das adversidades. A sua capacidade de enfrentar desafios com serenidade e sabedoria sempre foi uma inspiração para mim. Agradeço por suas palavras de encorajamento, por estar ao meu lado em todos os momentos, oferecendo amor incondicional, apoio e orientação. Sem o seu carinho e sacrifícios, não seria possível chegar até aqui. Você é, sem dúvida, a base de tudo que conquistei.

Aos meus irmãos, Wallace e Mateus, por estarem sempre ao meu lado com carinho, incentivo e aquele apoio que, com momentos de descontração que tornam tudo mais leve. Vocês foram fundamentais em cada etapa desta jornada, seja nos momentos de dificuldade ou nas conquistas que compartilhamos juntos. A amizade, o companheirismo e, claro, as brincadeiras entre nós sempre me deram a energia para continuar. Sou imensamente grata por cada palavra de encorajamento, por acreditarem em mim e por me apoiarem. Vocês são uma parte essencial das minhas realizações e da minha felicidade.

À minha querida família, especialmente à minha avó, Dona Loya, por sua presença constante, carinho e sabedoria, que tanto me inspiram. Agradeço também à tia Dade, tia Cléria, tio Celso, tia Lena e Tia Maly, por todo o apoio, alegria e afeto que sempre trouxeram à minha vida. Sou profundamente grata por ter uma família tão amorosa e presente, que me acompanhou em todos os momentos.

Ao meu querido Alex, meu companheiro de vida, pelo amor incondicional, pela paciência infinita e pelos incontáveis momentos de incentivo e apoio ao longo desses anos de estudo. Você esteve ao meu lado em cada desafio, sempre com uma palavra de conforto, um gesto de carinho e, muitas vezes, com aquele humor que só nós compartilhamos. Sua parceria foi fundamental em cada passo desta jornada, e sou eternamente grata por tudo o que vivemos e construímos juntos. Obrigada por acreditar em mim, por me motivar nos dias difíceis e, acima de tudo, por ser meu porto seguro em todos os momentos. Não teria chegado até aqui sem você, meu amor.

Aos amigos e colegas de curso, que compartilharam comigo as alegrias e desafios dos últimos dois anos de mestrado.

Aos meus orientadores Flávio Bambirra e Thaís Paiva, sou eternamente grata pela oportunidade, pela orientação cuidadosa, pela paciência e pela dedicação incansável que

foram fundamentais para a realização deste trabalho. A sabedoria e o comprometimento de vocês me guiaram em cada etapa desta jornada, e o apoio que recebi foi essencial para que eu superasse os desafios e alcançasse este momento. Sou profundamente grata por terem acreditado em mim e por todo o conhecimento e apoio que me proporcionaram ao longo dessa caminhada.

Aos queridos amigos Vicente (*in memoriam*) e Vanderlei, pelo constante incentivo, apoio emocional e orações durante esses anos.

À FUMP, CAPES, FAPEMIG e CNPq pelo suporte financeiro ao longo de toda a minha jornada acadêmica, especialmente à CAPES pelo financiamento desta pequisa.

Aos professores do Departamento de Estatística, com quem tive o privilégio de aprender, seja em sala de aula ou através dos projetos que enriqueceram minha formação. Cada um de vocês contribuiu de maneira única para o meu crescimento acadêmico e pessoal, e sou profundamente grata por todo o conhecimento compartilhado e pela dedicação com que conduzem seu trabalho. Vocês fazem parte dessa história, e tenho imenso orgulho de ter trilhado minha jornada acadêmica neste departamento, que é sinônimo de excelência e inspiração.

Agradeço à Universidade Federal de Minas Gerais, seu corpo docente, funcionários, direção e administração, por me proporcionarem a oportunidade de realizar este curso.

Por fim, meu muito obrigado a todas e todos que, de alguma forma, contribuíram para mais essa etapa em minha vida.

‘*In God we trust. All others must bring data.*”
(W. Edwards Deming)

Resumo

A divulgação segura de dados confidenciais representa uma área de grande interesse, e dentre as diversas metodologias existentes, a abordagem de dados sintéticos destaca-se por sua capacidade de gerar informações de forma sigilosa. Essa metodologia é altamente flexível, visando a divulgação de dados com distribuições muito semelhantes às dos dados originais e assim preservando também a segurança de informações sensíveis.
O modelo de rede Bayesiana, por sua vez, tem como propósito estimar de forma eficiente a distribuição conjunta de dados de interesse. Este método é uma escolha interessante para a geração de dados sintéticos, pois é um método flexível e robusto para a descrição das relações entre variáveis presentes no banco de dados original. Ao adotarmos o paradigma Bayesiano, conseguimos criar um modelo robusto não apenas para estimar a rede e os dados simulados, mas também para quantificar a incerteza intrínseca ao processo de geração desses novo dados.
Esta tese propõe um estudo que utiliza um modelo estado da arte *Markov chain Monte Carlo* (MCMC) para geração de dados sintéticos. Além disso, apresentamos uma abordagem inovadora para a divulgação de informações relevantes ao usuário final, com o intuito de reduzir a incerteza associada ao processo de estimação.
As principais contribuições deste trabalho incluem uma análise abrangente utilizando o paradigma Bayesiano para gerar dados sintéticos por meio de redes Bayesianas, incorporando um estudo robusto sobre a quantificação da incerteza no processo de geração desses novos dados. Introduzimos também uma classe geral de prioris penalizadoras para a rede. A tese compreende três estudos de simulação, bem como uma aplicação a dados reais que ilustra a análise do modelo proposto.

Palavras-chave: dados sintéticos; redes Bayesianas; inferência Bayesiana; quantificação de incerteza.

Abstract

The disclosure of confidential data represents an area of great interest, and among the various existing methodologies, the synthetic data approach stands out for its ability to generate information discreetly. This methodology is highly flexible, aiming to disclose data with distributions very similar to those of the original data, thus also preserving the security of sensitive information.
The Bayesian network model, in turn, is designed to efficiently estimate the joint distribution of relevant data. This method is an intriguing choice for generating synthetic data as it provides a flexible and robust approach to describing relationships between variables present in the original database. By adopting the Bayesian paradigm, we can create a robust model not only to estimate the network and simulated data but also to quantify the intrinsic uncertainty in the process of generating this new data.
This thesis proposes a study that utilizes a state-of-the-art Markov chain Monte Carlo (MCMC) model for generating synthetic data. Additionally, we introduce an innovative approach to disseminating relevant information to the end user, aiming to reduce the uncertainty associated with the estimation process.
The main contributions of this work include a comprehensive analysis using the Bayesian paradigm to generate synthetic data through Bayesian networks, incorporating a robust study on quantifying uncertainty in the process of generating this new data. We also introduce a general class of penalizing priors for the network. The thesis comprises three simulation studies as well as an application to real data that illustrates the analysis of the proposed model.

Keywords: synthetic data; Bayesian networks; Bayesian inference; uncertainty quantification.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Um dos objetivos de agências e instituições que lidam com divulgação de informações é a publicação de dados de forma segura (Duncan et al. 2001, Surendra & Mohan 2017, Raghunathan 2021). Dessa forma, essas instituições vêm investindo em pesquisas para melhorar a forma de liberação desses dados, ampliando o nível de informação disseminada e preservando o sigilo e a privacidade (Karr & Reiter 2014*a*, Reiter 2023).

Entre as alternativas mais simples de alteração de dados para serem divulgados, estão: a agregação, em que várias observações são condensadas de acordo com variáveis categóricas; a troca de dados, em que informações sensíveis são trocadas entre pares de registros com características parecidas; e a supressão, em que valores de alto risco são excluídos do banco de dados (Reiter 2003).

Embora esses métodos de alteração de dados possam resolver o problema da divulgação segura, muitas vezes as análises estatísticas possíveis de serem feitas com os dados alterados são limitadas e podem levar a resultados muito diferentes das análises realizadas no banco de dados original (Kennickell & Lane 2006). Devido a essas limitações, faz-se necessário o aprimoramento das técnicas de alteração de dados. Sendo assim, um método em que podemos conjuntamente proteger os dados e também fazer análises estatísticas válidas seria o ideal para que agências possam fazer o compartilhamento de bancos de dados (Caiola & Reiter 2010).

Uma técnica que permite divulgação segura e também se preocupa com análises estatísticas para os dados compartilhados é a metodologia de dados sintéticos proposta por Rubin (1993). A geração de dados sintéticos pode ser vista como um método de imputação, em que os valores "ausentes" são na verdade os valores sensíveis que não podem ser divulgados (Reiter 2005).

Portanto, na abordagem para a geração de dados sintéticos, um ou mais bancos de dados com valores imputados são criados a partir das distribuições de probabilidade estimadas no banco de dados original. Logo, os dados imputados terão distribuições de probabilidades muito parecidas com as dos dados reais. Como os dados divulgados são valores simulados, mantém-se o sigilo dos dados originais. Consequentemente, é possível ampliar o nível de informação disseminada e preservar o sigilo e a privacidade dos dados. Além disso, com uso da metodologia de dados sintéticos, é possível realizar inferências

muito próximas àquelas que seriam feitas nos dados originais (Rubin 1993).

Algumas técnicas podem ser utilizadas para descrever a relação entre as variáveis sensíveis, e assim gerar os dados sintéticos. Diferentes metodologias estatísticas podem ser utilizadas para esse fim, como as propostas por Raghunathan et al. (2003), Drechsler & Reiter (2011*a*), Drechsler (2011). Nesses trabalhos, os autores usam metodologias como imputação múltipla, e métodos não-paramétricos, como árvores de classificação e florestas aleatórias. Embora o objetivo seja gerar dados sintéticos, essas abordagens são predominantemente baseadas em modelos não-paramétricos e seus resultados são frequentemente limitados pelo número de variáveis presentes no banco de dados. No entanto, essas limitações podem ser superadas por métodos que utilizam abordagens paramétricas e Bayesianas, que oferecem uma forma mais robusta e flexível de modelar a complexidade das relações entre variáveis e gerar dados sintéticos com uma maior variedade de características e complexidade.

Uma metodologia usada para estimar as relações entre variáveis é o modelo de redes Bayesianas (Koller & Friedman 2009). O objetivo da rede Bayesiana é descrever a distribuição conjunta dos dados de forma robusta e parcimoniosa (Friedman & Koller 2003). Ela consiste em duas partes: a parte gráfica, que apresenta a rede na forma de um grafo, definindo as dependências condicionais entre as variáveis; e a parte estatística, que apresenta o conjunto de distribuições ou tabelas de probabilidades de cada ligação do grafo (Friedman et al. n.d.). Utilizando a rede Bayesiana, podemos estimar a distribuição conjunta dos dados a partir das relações de independências condicionais entre as variáveis. Visto que essa abordagem é bem flexível à natureza das variáveis, podemos utilizar essas informações para imputar novos dados com base nas distribuições de probabilidade estimadas na rede (Friedman & Koller 2003).

Sun & Erath (2015*a*) aplicam a metodologia de redes para dados sintéticos. Os autores estimam uma rede Bayesiana para os dados. Nesse estudo, os autores usam algoritmo *Expectation Maximization*(EM) para estimar as probabilidades condicionais do modelo, baseando-se nos dados originais. Após esta etapa, os dados sintéticos são gerados a partir das probabilidades condicionais estimadas pela rede. Embora os resultados mostrem que a metodologia é adequada, o trabalho se limita a um número pequeno de variáveis no banco original.

Young et al. (2009) também utilizam redes Bayesianas no contexto de dados sintéticos. Os autores usam abordagem Bayesiana para estimar a rede, e geram os dados sintéticos através da maximização dos parâmetros da rede. Um algoritmo (*greedy*) de otimização é usado para gerar e selecionar estruturas de rede com maior probabilidade *a posteriori*. Este algoritmo realiza a escolha da rede de forma descentralizada, optando por escolhas locais em vez de uma abordagem global. Essa estratégia confere maior agilidade ao algoritmo, uma vez que elimina a necessidade de coordenação global, mas também pode levar a redes que não representam bem os dados. Uma vez que a rede mais provável é

estimada, os dados sintéticos são gerados a partir das distribuições condicionais ajustadas. Todavia, os resultados apresentados são limitados a bancos de dados com conhecimento *a priori*, visto que algumas relações entre variáveis devem ser estabelecidas antes de se estimar a rede.

A metodologia de Redes Bayesianas não se limita à geração de dados sintéticos. Embora poucos estudos usem redes Bayesianas exclusivamente para criar dados sintéticos, é essencial considerar outras metodologias quando o objetivo é apenas a estimação da rede (Korb & Nicholson 2010). A partir dessas abordagens, é possível gerar dados sintéticos mais detalhados e relevantes. Além disso, explorar diferentes metodologias permite identificar as vantagens e limitações de cada uma, ajudando na escolha da técnica mais apropriada para o contexto específico. Esse processo não apenas aprimora a precisão na estimação da rede, mas também enriquece a interpretação dos dados gerados, oferecendo uma visão mais robusta e completa.

Em um contexto estatístico, a estimação de redes Bayesianas pode ser realizada por meio de inferência utilizando abordagens clássicas ou Bayesianas. Na abordagem clássica, também conhecida como frequentista, a estimação é feita maximizando a verossimilhança dos dados observados. Essa maximização é realizada por meio de métodos como a Máxima Verossimilhança (MLE), onde o objetivo é encontrar os parâmetros que tornam os dados mais prováveis. Além disso, técnicas de seleção de modelos, como o Critério de Informação de Akaike (AIC) ou o Critério de Informação Bayesiano (BIC), são usadas para comparar diferentes estruturas de redes e escolher aquela que melhor equilibra ajuste e complexidade.

Cooper & Herskovits (1992), Cano et al. (2004), Di Zio et al. (2004) e Hruschka et al. (2004) utilizam métodos de otimização para estimação da rede. Esses trabalhos utilizam algoritmos como o *hill climbing* (Russell & Norvig 2010). Esses algoritmos geralmente começam seu processo criando a estrutura inicial da rede, que sofre pequenas modificações de acordo com um critério de avaliação das possíveis estruturas modificadas. O intuito com esse processo de avaliação é tentar encontrar uma configuração ótima para a estrutura da rede Bayesiana.

Outra metodologia usada por Cooper & Herskovits (1992), Cano et al. (2004), Di Zio et al. (2004) e Hruschka et al. (2004) é o algoritmo de *simulated annealing* (Kirkpatrick et al. 1983). Semelhante ao *hill climbing*, ele faz pequenas modificações na rede e permite retroceder a estados anteriores da estrutura durante a busca.

Nos métodos de inferência clássica para a estimação de redes, a escolha da melhor estrutura é frequentemente baseada no cálculo de um *score*, que avalia a adequação da rede aos dados observados. Esse *score* desempenha um papel crucial na seleção da estrutura de rede que melhor descreve as propriedades estatísticas dos dados, oferecendo uma abordagem sistemática para a construção e estimação da rede Bayesiana (Mihaljević et al. 2021).

A inferência Bayesiana para a estimação de redes Bayesianas adota uma abordagem diferenciada em relação aos métodos clássicos. Em vez de buscar uma única estrutura de rede, a inferência Bayesiana considera distribuições de probabilidade sobre todas as possíveis estruturas, incorporando conhecimentos *a priori*. Métodos Bayesianos, como a amostragem de Gibbs ou o Metropolis-Hastings, exploram o espaço de configurações de rede, proporcionando estimativas a *posteriori* tanto para os parâmetros quanto para a estrutura da rede (Heckerman, Geiger & Chickering 1995). Essa metodologia oferece uma perspectiva probabilística mais robusta e abrangente para a estimação de redes Bayesianas.

Koller & Friedman (2009), Grzegorczyk & Husmeier (2008) e Goudie & Mukherjee (2016) utilizam inferência Bayesiana para a estimação da rede por meio de algoritmos MCMC (*Monte Carlo Markov Chain*) para fazer a estimação da rede.

Em particular, Goudie & Mukherjee (2016) propõem uma abordagem avançada para a estimação Bayesiana de redes Bayesianas por meio do método de um amostrador Gibbs *sampling*, que é considerado o estado da arte na área. O artigo apresenta um método robusto de Monte Carlo via Cadeias de Markov (MCMC) que utiliza uma variação do Gibbs *sampling* conhecida como *random scan* Gibbs *sampling*. Essa abordagem é notável por sua capacidade de lidar eficientemente com modelos complexos.

O método de *random scan* Gibbs *sampling* descrito pelos autores melhora o desempenho computacional ao atualizar variáveis do modelo de forma estocástica. Em vez de atualizar todas as variáveis simultaneamente, o algoritmo seleciona aleatoriamente uma variável para atualização a cada iteração. Essa técnica reduz significativamente o custo computacional, especialmente em modelos com um grande número de variáveis, onde a atualização simultânea poderia ser excessivamente dispendiosa.

Além disso, o trabalho introduz a definição de blocos de variáveis no algoritmo. Estes blocos são formados com base nas variáveis do modelo e permitem que o *sampling* de Gibbs trate grupos de variáveis de maneira conjunta. Essa abordagem não só melhora a eficiência do processo de amostragem, mas também resulta em estimativas mais precisas da estrutura da rede, aproveitando as interações entre variáveis capturadas pelos blocos definidos.

Em resumo, a metodologia proposta por Goudie & Mukherjee (2016) fornece uma solução inovadora para a estimação Bayesiana de redes Bayesianas, combinando a eficácia do *random scan* Gibbs *sampling* com a estratégia de blocos de variáveis para otimizar o desempenho computacional. Isso reduz o tempo de processamento e melhora a precisão das estimativas, tornando-a uma ferramenta valiosa para a análise e modelagem de redes complexas. No contexto de dados sintéticos, a quantificação de incerteza é uma tarefa fundamental para estudar a qualidade dos dados gerados. Consequentemente, é de grande importância considerar abordagens robustas e eficientes em relação à quantificação de incerteza nesse cenário. Embora medidas pontuais possam ser relevantes, elas não cumprem

essa tarefa quando a abordagem é completamente Bayesiana. Quando usamos modelagem Bayesiana em todas as etapas da imputação de dados sintéticos, podemos considerar e analisar toda a incerteza da modelagem. Nos trabalhos presentes na literatura, embora façam uso de metodologia Bayesiana nesse contexto, não existe um estudo completo da utilização dessas ferramentas com intuito de considerar toda a incerteza do modelo.

Redes Bayesianas são modelos versáteis que podem ser aplicados a diferentes tipos de dados, incluindo variáveis contínuas, binárias, e categóricas. Essa flexibilidade torna as Redes Bayesianas uma ferramenta poderosa em diversas áreas do conhecimento, permitindo modelar de maneira eficiente as relações de dependência entre variáveis de naturezas distintas.

No caso de dados contínuos, as redes Bayesianas podem ser usadas para modelar variáveis como medições, resultados de testes ou outras quantidades numéricas. Já para dados binários, que são o foco de muitas aplicações em campos como medicina, biologia e ciências sociais, as redes Bayesianas são particularmente eficazes em lidar com variáveis categóricas e estruturas de dependência complexas (Korb & Nicholson 2010).

A aplicação de Redes Bayesianas com abordagem Bayesiana em dados binários tem se mostrado eficaz em diversas áreas, onde é comum lidar com variáveis como presença ou ausência de sintomas, ocorrência de eventos, ou classificações binárias em um conjunto de dados. Essa metodologia permite não apenas a modelagem precisa das interações entre variáveis, mas também a previsão e a tomada de decisão, considerando a incerteza inerente aos dados e ao conhecimento prévio disponível (Bishop 2006).

Alguns trabalhos que exploram o uso de redes Bayesianas com abordagem Bayesiana para dados binários, destacando-se em áreas como diagnóstico médico, biologia computacional e análise de risco. Por exemplo, Hikosaka et al. (1999) exploraram redes Bayesianas para prever padrões binários em diagnósticos médicos, utilizando dados binários para modelar a presença ou ausência de sintomas e inferir a probabilidade de diferentes doenças. Esse estudo demonstrou a eficácia da abordagem Bayesiana na incorporação de conhecimento prévio e na obtenção de estimativas mais robustas para situações de incerteza, especialmente quando os dados disponíveis são limitados.

Outro trabalho relevante é o de Neal (2012), que aplicou redes Bayesianas para a classificação de dados binários em problemas de aprendizado de máquina. Eles usaram uma abordagem Bayesiana para explorar o espaço de estruturas de rede, mostrando que a consideração de distribuições a posteriori pode levar a modelos que generalizam melhor em dados desconhecidos. Esses estudos reforçam a utilidade das redes Bayesianas em contextos onde as variáveis binárias desempenham um papel crucial, e onde a robustez na estimação dos parâmetros é essencial para a tomada de decisão. No entanto, essas abordagens apresentam algumas limitações. Muitos desses estudos enfrentam desafios relacionados à complexidade computacional e à necessidade de grandes quantidades de dados para alcançar estimativas precisas. Além disso, a maioria desses trabalhos não

aborda diretamente a geração de dados sintéticos, limitando a aplicação das redes Bayesianas a dados reais e, frequentemente, escassos.

Portanto, o objetivo desta tese é propor uma metodologia completamente Bayesiana para a geração e análise de dados sintéticos via redes Bayesianas no contexto de dados binários. Este estudo terá também o foco de estudar e quantificar toda incerteza intrínseca ao modelo. Será utilizado o algoritmo de Goudie & Mukherjee (2016) para estimação da rede. Uma classe de distribuições *a priori* para a rede será proposta. Além disso, a quantificação da incerteza associada aos dados sintéticos será feita através da distribuição *a posteriori* preditiva dos mesmos.

Esta tese está organizada da seguinte forma. O Capítulo 2 apresenta uma revisão sobre redes Bayesianas e sobre geração de dados sintéticos, além de uma ampla revisão bibliográfica dessas áreas.

O Capítulo 3 descreve detalhadamente a metodologia estatística proposta para a geração e análise de dados sintéticos via redes Bayesianas.

Utilizando a metodologia proposta, duas análises serão apresentadas no Capítulo 4. A primeira considera dados simulados para investigar a eficiência da metodologia e a segunda aplica a mesma a um conjunto de dados reais para ilustrar sua aplicabilidade.

Por fim, conclusões e direcionamentos de trabalhos futuros serão apresentados no Capítulo 5.

Capítulo 2

Dados Sintéticos para Redes Bayesianas

2.1 Dados Sintéticos

Métodos para divulgação de dados de forma segura são utilizados por agências a fim de evitar a quebra de sigilo de informações. Sendo por questões éticas ou até mesmo jurídicas, essas entidades têm buscado maneiras de disponibilizar bancos para uso público, preservando a confidencialidade dos dados originais. Karr & Reiter (2014*b*) apresentam algumas técnicas para alteração de dados com intuito de miniminizar o risco de divulgação de informações confidenciais. Uma das técnicas citadas por Karr & Reiter (2014*b*) é a agregação. A agregação condensa valores de variáveis em algumas categorias, ou seja, variáveis são re-codificadas e resumidas. Essa mudança, embora dificulte a identificação de valores mais incomuns, afeta a inferência a ser feita com esses dados (Kennickell & Lane 2006).

Outra técnica mencionada pelos autores é a supressão, em que valores considerados sensíveis são simplesmente excluídos da amostra, o que também afeta as distribuições de determinadas variáveis (Cox 1980). A troca de dados também é utilizada como técnica de divulgação de dados e apenas permuta os valores entre indivíduos, o que pode também impactar nas inferências da amostra (Drechsler & Reiter 2010).

Adicionar ruído aleatório em variáveis contínuas também é uma técnica de alteração. Contudo, isso altera a distribuição verdadeira dos dados. Embora a segurança aumente, a precisão das análises dos dados pode ser seriamente afetada (Yancey et al. 2002).

Embora as técnicas mencionadas anteriormente ajudem a preservar a confidencialidade dos dados e, consequentemente, diminuir o risco da divulgação, a utilidade dos dados fica seriamente afetada. A utilidade é vista como o que podemos fazer com os dados divulgados, ou seja, se esses dados podem ser analisados e fornecer inferências parecidas com aquelas feitas com os dados originais. Sendo assim, o interesse é tentar controlar o

trade-off entre o risco e a utilidade, ou seja, proteger os dados e divulgar informações que possam ser analisadas (Woo et al. 2009).

Rubin (1993) desenvolveu um método para gerar bases de dados capazes de conter características muito próximas à base de dados original. A metodologia de dados sintéticos busca gerar novos dados com base nas distribuições de probabilidade estimadas com os dados originais. Os dados são simulados com o foco de manter ao máximo as relações existentes no banco de dados original. Assim, são geradas várias versões dos dados simulados para que usuários possam medir a variabilidade do modelo de simulação de dados (Reiter & Raghunathan 2007).

É importante ressaltar que os dados sintéticos também podem ser completamente sintéticos. Dessa forma todo o banco de dados é simulado. Logo, nenhum dado original é divulgado. Podemos também divulgar uma parte do banco original, sendo aquelas variáveis que não precisam de sigilo, então divulgamos um banco de dados parcialmente sintético. Ainda assim podemos ter também dados parcialmente sintéticos, em que apenas algumas informações de uma determinada variável são divulgadas. A escolha entre os esses tipos de dados sintéticos pode ser feita dependendo do tipo de dado envolvido no estudo (Drechsler & Reiter 2011*a*).

Uma vez que os dados sintéticos são simulados a partir da distribuição dos dados originais, o mais importante é obter uma boa estimativa da distribuição conjunta dos dados originais. Após essa estimação, os dados sintéticos podem ser gerados. Sendo assim, utiliza-se várias técnicas, paramétricas ou não-paramétricas, para a estimação dessa distribuição de interesse.

Algumas metodologias já são utilizadas por agências para a imputação de dados sintéticos. Reiter (2005) apresenta um estudo em que utiliza-se uma metodologia não-paramétrica de árvores de classificação e regressão (CART) para estimar a distribuição conjunta dos dados. Caiola & Reiter (2010) fazem um extensão do trabalho de Reiter (2005), utilizando florestas aleatórias.

Alguns trabalhos na literatura na área de inferência em redes Bayesianas também consideram o problema de se utilizar a rede para a geração de dados sintéticos, assim como nesta tese. Nesta seção, formalizaremos o problema de geração dos dados sintéticos e apresentaremos uma revisão bibliográfica sobre o assunto.

A metodologia de dados sintéticos tem como objetivo fornecer ao usuário final bases de dados analisáveis e preservar as relações contidas no banco de dados original. O intuito é gerar novas bases de dados que permitam que diversas análises estatísticas possam ser feitas. Essas novas bases de dados são criadas a partir de métodos semelhantes à imputação de dados ausentes proposta por Rubin (1993).

Usualmente, mais de um banco de dados sintético é gerado a partir dos dados originais. Dessa forma, podemos medir a variabilidade do método de imputação (Rubin 1993). As imputações $i = 1, ..., m$ são independentes, produzindo assim m bancos de

dados sintéticos diferentes que serão divulgadas para os usuários.

Para fazer uma análise estatística frequentista a partir dos dados sintéticos, a inferência será feita pela combinação dos m bancos completos gerados, para estimadores como média, variância e outras medidas de interesse (Reiter 2003).

Seja $P(\mathbf{X})$ a distribuição de probabilidade conjunta das variáveis no conjunto de dados $\mathbf{X}$. Essa distribuição descreve a probabilidade de ocorrência de diferentes combinações de valores das variáveis, capturando as relações de dependência entre elas. A estimativa de $P(\mathbf{X})$ é crucial para a geração de dados sintéticos, pois permite a criação de novas amostras que preservam as propriedades estatísticas do conjunto de dados original, como a estrutura de correlação e as distribuições marginais.

No contexto da imputação de dados sintéticos, a natureza das variáveis em $\mathbf{X}$ — sejam elas contínuas ou discretas, determina a escolha das metodologias apropriadas para a estimação de $P(\mathbf{X})$. Por exemplo, Reiter (2003) desenvolveu uma metodologia para gerar dados parcialmente sintéticos, na qual nem todas as variáveis são consideradas sensíveis para divulgação. Nessa abordagem, apenas uma parte do banco de dados é substituída por dados sintéticos, permitindo assim a preservação de informações essenciais enquanto se protege a privacidade dos dados originais.

Metodologias não-paramétricas para a imputação de dados sintéticos foram introduzidas por Reiter (2005). Nesse estudo, são utilizadas árvores de regressão (CART) e outras metodologias não-paramétricas para fazer a estimação de $P(\mathbf{X})$ e gerar os dados sintéticos. Posteriormente, Caiola & Reiter (2010) e Drechsler & Reiter (2011*a*) fizeram uma comparação entre metodologias como CART, florestas aleatórias e outras metodologias de *machine learning*. Em todos esses trabalhos, após a estimação de $P(\mathbf{X})$, os novos dados gerados são imputados a partir do *bootstrap* Bayesiano.

Reiter et al. (2014) utilizam inferência Bayesiana para a geração de dados sintéticos. O objetivo é a geração de dados parcialmente sintéticos. O modelo de regressão é utilizado para estimar a distribuição *a posteriori* da variável sensível, e gera os dados sintéticos a partir da distribuição preditiva *a posteriori* da variável sensível.

A metodologia de redes Bayesianas também pode ser utilizada para a geração de dados sintéticos. A vantagem dessa metodologia está ligada ao fato de que em um cenário em que precisamos gerar bancos de dados completamente sintéticos, podemos estimar a distribuição de $P(\mathbf{X})$ mais facilmente. Como a rede Bayesiana procura descrever a distribuição conjunta dos dados através da fatoração das variáveis por independências condicionais, podemos utilizar essas informações para a geração dos dados sintéticos. Outra vantagem da rede Bayesiana é que, após a estimação da rede, podemos amostrar valores diretamente da distribuição de $\mathbf{X}$ dada pela distribuição conjunta fatorada. Dessa forma, os dados sintéticos serão semelhantes ao dados originais, pois a rede preserva e estima as relações presentes no banco de dados.

Deeva et al. (2020), e Sun & Erath (2015*b*) utilizam a metodologia de redes Baye-

sianas para a geração de dados sintéticos. Nesses estudos, a rede é utilizada para estimar $P(\mathbf{X})$ com abordagem clássica. Após aprender a estrutura do modelo, os dados sintéticos são gerados a partir das distribuições condicionais estimadas pela rede. Com isso, além de gerar as amostras diretamente das distribuições estimadas, ainda existe a vantagem de entender as relações entre as variáveis, uma vez que a rede é construída sequencialmente.

Embora Deeva et al. (2020), e Sun & Erath (2015*b*) utilizem redes Bayesianas para gerar dados sintéticos, suas abordagens são exclusivamente clássicas, limitando-se a variáveis discretas na estimação da rede. Embora os novos dados possam ser gerados diretamente a partir das distribuições estimadas, esses estudos não investigam o impacto dessa metodologia na proteção dos dados sigilosos. Quando a imputação se assemelha aos dados reais, há um risco elevado de divulgação de informações confidenciais. Além disso, essa abordagem apresenta limitações em cenários, pois as relações entre as variáveis já são conhecidas, ou seja, em conjuntos de dados maiores com mais independências entre variáveis, podendo não ser a melhor escolha nesse contexto.

Drechsler & Reiter (2011*b*) apresentam métricas que podem ser empregadas para avaliar tanto o risco de divulgação quanto a utilidade dos dados sintéticos. A utilidade refere-se à precisão com que análises estatísticas podem ser realizadas utilizando os dados sintéticos. Por outro lado, é crucial considerar o risco de divulgação de informações confidenciais. Quando os dados sintéticos são muito semelhantes aos dados reais, a utilidade tende a ser alta, mas o risco de divulgação também aumenta. Assim, é necessário desenvolver um algoritmo flexível que permita, ao final da estimação da rede e da imputação dos dados, avaliar e balancear a utilidade e o risco dos dados sintéticos. Dessa forma, um algoritmo flexível para a estimação da rede possibilitará ajustes necessários para equilibrar esses dois aspectos.

2.1.1 Análise de dados sintéticos

O tipo de variável presente no banco de dados, e até mesmo sua importância para a divulgação de dados, pode estipular como a geração dos dados sintéticos é realizada. Em alguns casos, tem-se o interesse em divulgar dados completamente sintéticos para aumentar a proteção dos dados, mas em algumas bases nem todas as variáveis possuem algum risco de divulgação. Gerar bancos inteiros com dados sintéticos pode não ser uma tarefa simples e muitas vezes nem todas as variáveis são de risco.

Uma proposta apresentada por Reiter (2003) envolve o uso de dados parcialmente sintéticos. Nessa abordagem, apenas os valores mais sensíveis à divulgação são substituídos por dados sintéticos. Como resultado, as fórmulas utilizadas para a análise desses

dados sintéticos diferem daquelas descritas por Raghunathan et al. (2003).

Dados parcialmente sintéticos são atraentes porque podem manter muitos dos benefícios de dados totalmente sintéticos, protegendo a confidencialidade, e também sua implementação computacional pode se tornar mais simples.

A inferência para os bancos de dados sintéticos será feita pela combinação dos m bancos completos gerados, para estimadores como média, variância e outras medidas de interesse, visto que os m bancos de dados sintéticos são independentes.

Raghunathan et al. (2003) propõe algumas medidas de utilidade para dados sintéticos. Essas medidas tem como objetivo fornecer ao usuário final ferramentas para a análise conjunta de todos os bancos de dados divulgados. Sendo assim, podemos combinar esses m bancos de dados sintéticos e assim conseguir fazer inferências válidas.

Seja Q uma estatística de interesse, q seu estimador pontual e v a variância desse estimador pontual. Considere um conjunto de bancos de dados sintéticos m. Para cada banco de dados sintético i (com $i = 1, \ldots, m$), definem-se q_i e v_i como os respectivos valores do estimador pontual q e sua variância v.

Os valores de q_i e v_i são calculados como se estivessem no banco de dados original. Para estimar a estatística Q com base nos resultados obtidos dos m bancos de dados sintéticos, os seguintes cálculos são realizados:

1. **Média dos Estimadores Pontuais**: A média dos estimadores pontuais q_i é dada por:

$$\bar{q}_m = \frac{1}{m}\sum_{i=1}^{m} q_i, \tag{2.1}$$

   essa fórmula fornece uma estimativa central para a estatística Q, considerando os valores dos estimadores pontuais obtidos de todos os bancos de dados sintéticos.

2. **Variância dos Estimadores Pontuais**: A variância dos estimadores pontuais q_i é calculada por:

$$b_m = \frac{1}{m-1}\sum_{i=1}^{m}(q_i - \bar{q}_m)^2, \tag{2.2}$$

   aqui calculamos a variabilidade entre os estimadores pontuais q_i dos diferentes bancos de dados sintéticos. Logo, ela fornece uma ideia de como os estimadores variam em torno da média $\bar{q}_m$.

3. **Média das Variâncias dos Estimadores Pontuais**: A média das variâncias v_i é dada por:

$$\bar{v}_m = \frac{1}{m}\sum_{i=1}^{m} v_i, \tag{2.3}$$

o que fornece uma estimativa da variância média dos estimadores pontuais em todos os bancos de dados sintéticos.

A estimativa final para a estatística Q é dada pela média dos estimadores pontuais $\bar{q}_m$. Para calcular a variância associada a essa estimativa, utiliza-se a fórmula:

$$T_p = \frac{b_m}{m} + \bar{v}_m. \tag{2.4}$$

Nesta fórmula, $\frac{b_m}{m}$ representa a variância média entre os estimadores pontuais ajustada pelo número de bancos de dados sintéticos, e $\bar{v}_m$ é a variância média dos estimadores pontuais. Juntas, essas componentes fornecem uma estimativa combinada da variância da estatística Q.

Considere um exemplo simples com $m = 2$, em que Q é a média de uma variável binária. Suponha que temos dois bancos de dados sintéticos e queremos estimar a média da variável binária a partir desses dados. Defina os valores dos estimadores pontuais q_i e suas variâncias v_i para $i = 1$ e $i = 2$. Dessa forma, suponha que os estimadores pontuais e as variâncias para os dois bancos de dados sintéticos sejam:

$$\begin{aligned} q_1 &= 0.60, \quad v_1 = 0.02 \\ q_2 &= 0.55, \quad v_2 = 0.03 \end{aligned}$$

A média dos estimadores pontuais q_i é dada por:

$$\begin{aligned} \bar{q}_2 &= \frac{1}{2}(q_1 + q_2) \\ &= \frac{1}{2}(0.60 + 0.55) \\ &= 0.575 \end{aligned}$$

A variância dos estimadores pontuais q_i é calculada por:

$$\begin{aligned} b_2 &= \frac{1}{2-1}\left((q_1 - \bar{q}_2)^2 + (q_2 - \bar{q}_2)^2\right) \\ &= \frac{1}{1}\left((0.60 - 0.575)^2 + (0.55 - 0.575)^2\right) \\ &= (0.025)^2 + (-0.025)^2 \\ &= 0.000625 + 0.000625 \\ &= 0.00125 \end{aligned}$$

A média das variâncias v_i é dada por:

$$\begin{aligned}\bar{v}_2 &= \frac{1}{2}(v_1 + v_2) \\ &= \frac{1}{2}(0.02 + 0.03) \\ &= \frac{0.05}{2} \\ &= 0.025\end{aligned}$$

A variância estimada para a estatística Q é dada por:

$$\begin{aligned}T_p &= \frac{b_2}{2} + \bar{v}_2 \\ &= \frac{0.00125}{2} + 0.025 \\ &= 0.000625 + 0.025 \\ &= 0.025625\end{aligned}$$

Este exemplo ilustra como calcular a média e a variância de uma variável binária utilizando dados de dois bancos de dados sintéticos. Com a utilização de dados sintéticos, a realização desses cálculos se torna mais prática e flexível, pois permite uma análise e verificação de métodos estatísticos sem depender de dados reais, que podem ser limitados ou confidenciais. Além disso, a repetição de cálculos em diferentes conjuntos de dados sintéticos possibilita a avaliação da variabilidade das estimativas e a comparação de diferentes abordagens estatísticas. Isso facilita a validação de modelos e a implementação de técnicas de estimativa, oferecendo uma maneira eficiente de explorar e ajustar os métodos de análise antes de aplicá-los a dados reais (Rubin 1993).

A escolha de utilizar entre 5 e 10 bancos de dados sintéticos é uma prática recomendada em metodologias de imputação múltipla, conforme discutido por Rubin (1993) e Raghunathan et al. (2003). Esse intervalo é ideal para equilibrar a precisão das estimativas e o custo computacional: um número inferior a 5 conjuntos pode resultar em variabilidade insuficiente, prejudicando a precisão dos intervalos de confiança, enquanto mais de 10 conjuntos pode aumentar significativamente o custo computacional e o tempo de processamento (Rubin 1993).

Apesar dessa recomendação tradicional, a crescente capacidade computacional permite agora a utilização de um número maior de bancos de dados sintéticos. Essa expansão pode melhorar a precisão das estimativas e a robustez dos resultados, reduzindo o risco de viés e aumentando a confiança nas inferências. Assim, com a tecnologia atual, é viável e benéfico considerar um número maior de conjuntos de dados sintéticos para obter análises mais precisas e confiáveis (Rubin 1993).

Ao trabalhar com dados sintéticos, é essencial considerar não apenas a utilidade dos dados gerados, mas também o risco de divulgação de informações sensíveis. O risco

de divulgação avalia o potencial de exposição de informações privadas ou confidenciais. Métodos para medir o risco de divulgação incluem técnicas como a análise de privacidade diferencial, que quantifica a probabilidade de que a inclusão de um indivíduo em um conjunto de dados possa ser detectada (Dwork et al. 2006). Outras abordagens envolvem a verificação de re-identificação, em que se avalia a possibilidade de que indivíduos possam ser re-identificados a partir dos dados sintéticos (Cui 2019).

No caso de dados binários, o risco de divulgação pode ser especialmente crítico. Os dados binários, por serem discretos e muitas vezes agregados em categorias simples, podem facilitar a identificação se não forem adequadamente protegidos. A análise da privacidade diferencial e a avaliação de técnicas de proteção de dados são cruciais para garantir que a informação sensível não seja revelada (Hu et al. 2023).

Uma abordagem promissora para investigar o risco de divulgação em dados binários é o uso de *propensity scores*. O *propensity score* é uma medida que estima a probabilidade de uma unidade de análise ser atribuída a um determinado grupo, dado um conjunto de variáveis observadas. Ao utilizar esses *scores* para gerar dados sintéticos, é possível ajustar os dados gerados de forma a preservar a estrutura de dependência entre as variáveis, enquanto se mantém a privacidade dos indivíduos. Isso ajuda a garantir que as relações entre as variáveis sejam preservadas, mas a exposição direta de dados individuais seja minimizada, reduzindo o risco de identificação (Rosenbaum & Rubin 1983).

Para calcular o *propensity scores*, podemos utilizar um modelo de regressão logística, que é uma abordagem eficaz para estimar a probabilidade de uma observação pertencer a um determinado grupo com base em covariáveis observadas. No contexto da comparação entre dados sintéticos e dados reais, o *score* permite avaliar a semelhança entre os dados sintéticos e o banco de dados original.

Seja $\mathbf{Y_{obs}}$ o banco de dados original e $\mathbf{Y_{syn}}$ o banco de dados sintético. A ideia é combinar ambos os bancos de dados em um único conjunto para treinar o modelo de regressão logística. O objetivo é estimar a probabilidade de que cada observação pertença ao banco de dados original $\mathbf{Y_{obs}}$ em vez de ao banco de dados sintético $\mathbf{Y_{syn}}$. Para isso, definimos uma variável binária Y, onde $Y = 1$ indica que a observação pertence ao banco de dados original e $Y = 0$ indica que pertence ao banco de dados sintético.

O modelo de regressão logística pode ser formulado como:

$$\log\left(\frac{P(Y=1 \mid \mathbf{X})}{1-P(Y=1 \mid \mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k, \tag{2.5}$$

em que $P(Y = 1 \mid \mathbf{X})$ representa a probabilidade estimada de que uma observação pertença ao banco de dados original $\mathbf{Y_{obs}}$, dado um vetor de covariáveis $\mathbf{X}$. Os parâmetros β_i são estimados a partir dos dados combinados e refletem a influência das covariáveis sobre a probabilidade de uma observação ser original.

A partir deste modelo, o *propensity score* é obtido como a probabilidade prevista de $Y = 1$ para cada observação. Valores próximos a 1 indicam alta probabilidade de a

observação pertencer ao banco de dados original, enquanto valores próximos a 0 indicam alta probabilidade de pertencimento ao banco de dados sintético.

Analisando a distribuição dos *propensity scores* dos dados sintéticos em comparação com os dados reais, é possível avaliar a qualidade e a segurança dos dados sintéticos. Se os esses valores dos dados sintéticos se distribuírem de maneira semelhante aos dos dados reais, isso sugere que os dados sintéticos replicam bem a estrutura dos dados originais. Caso contrário, discrepâncias significativas podem indicar que os dados sintéticos não são adequados para análise e podem apresentar riscos de divulgação de informações.

Para avaliar a qualidade dos dados sintéticos em relação aos dados originais, é útil calcular a média dos *propensity scores* obtidos a partir do modelo de regressão logística. A média dos fornece uma visão geral da probabilidade média das observações de serem classificadas como pertencentes ao banco de dados original. Se a média dos estiver próxima de 1, isso indica que, em média, as observações sintéticas têm uma alta probabilidade de serem confundidas com as observações do banco de dados original. Esse cenário pode ser preocupante, pois sugere que os dados sintéticos são muito semelhantes aos dados originais, potencialmente comprometendo a segurança e a privacidade das informações. Idealmente, os *scores* dos dados sintéticos devem ser distribuídos de forma que reflitam a variabilidade dos dados reais, sem se concentrar excessivamente em valores próximos a 1. Uma média muito próxima de 1 pode indicar uma sobreposição significativa entre os dados sintéticos e os dados reais, levantando preocupações sobre o risco de divulgação de informações sensíveis (Rosenbaum & Rubin 1983, Cox 2018).

Portanto, ao aplicar métodos para a criação de dados sintéticos, é fundamental integrar tanto medidas de utilidade quanto de risco de divulgação para assegurar que os dados gerados sejam úteis e seguros.

Outro aspecto importante de todo e qualquer processo de inferência é quantificação da incerteza atrelada ao processo de estimação. Embora divulgar uma pequena quantidade de bancos de dados seja o comum, se faz necessário propor outras metodologias robustas para a divulgação de dados sintéticos, levando em consideração toda a variabilidade envolvida no processo. Por esse motivo, apresentamos uma proposta para a divulgação mais robusta dos bancos de dados sintéticos no Capítulo 3 levando em consideração toda a variabilidade dos processo de inferência.

Outro aspecto importante de todo e qualquer processo de inferência é considerar que a precisão das estimativas é frequentemente acompanhada por um nível de incerteza inerente ao processo de estimação. Quando se divulga um número reduzido de bancos de dados sintéticos, como recomendado tradicionalmente, pode-se não capturar completamente a variabilidade do processo de estimação, o que pode levar a conclusões imprecisas ou enviesadas (Rubin 1993, Raghunathan et al. 2003).

A incerteza na inferência é uma preocupação central porque qualquer estimativa derivada de dados sintéticos pode ser afetada por variabilidade adicional introduzida pelo

processo de geração dos dados. Em particular, o número de bancos de dados sintéticos utilizado pode impactar diretamente a confiabilidade das inferências feitas (Rubin 1993). Portanto, ao lidar com dados sintéticos, é fundamental adotar metodologias que considerem toda a variabilidade do processo de estimação para garantir que as estimativas sejam robustas e representativas.

A proposta apresentada no Capítulo 3 busca abordar essas questões ao introduzir uma metodologia mais robusta para a divulgação de dados sintéticos. Essa abordagem considera toda a variabilidade do processo de inferência e oferece um meio mais eficaz de avaliar a qualidade dos dados sintéticos gerados. A utilização de um maior número de bancos de dados sintéticos e o desenvolvimento de técnicas mais avançadas para quantificação da incerteza são passos importantes para melhorar a precisão e a confiabilidade das análises baseadas em dados sintéticos (Rubin 1993, Little & Rubin 2019).

2.2 Redes Bayesianas

Nesta seção, discutiremos o modelo de redes Bayesianas com intuito de estimar a distribuição conjunta de X. Essa informação será utilizada posteriormente para a geração dos dados sintéticos.

Uma rede Bayesiana pode ser descrita como um modelo probabilístico que representa relações entre variáveis. Essa relações podem ser apresentadas em forma de um *directed acyclic graph* (DAG) , no qual os "nós" representam as variáveis, e os arcos/setas representam a dependência condicional entre elas (Heckerman 2008).

Seja $(\Omega, \mathcal{F}, P)$ o espaço de probabilidade no qual estão definidas as p variáveis aleatórias unidimensionais $X = (X_1, \ldots, X_p)$. O objetivo deste trabalho é estimar a medida de probabilidade P e, neste contexto, gerar dados sintéticos. Para isso, definimos um modelo probabilístico $\mathcal{P}$ que é um conjunto de medidas de probabilidade que admite-se conter a verdadeira medida p. Em particular, $\mathcal{P}$ será assumida como sendo uma rede Bayesiana com uma forma específica para as respectivas distribuições condicionais, no nosso caso, Bernoulli. A rede Bayesiana nos permite escrever a distribuição conjunta de X utilizando relações de independência condicional (Heckerman 2008). A princípio, queremos considerar todas as possíveis fatorações de $p(X)$ – a função de probabilidade/densidade conjunta de X. A inferência sobre p será feita sob o paradigma Bayesiano no contexto desse modelo probabilístico.

A rede Bayesiana $\mathcal{G}$ define as relações de independência condicional das variáveis X_j. Por fim, definimos como θ o conjunto de todos os parâmetros que possam indexar as possíveis distribuições condicionais definidas pela rede Bayesiana. O G mais complexo é

dado pela rede completa, ou seja,

$$p(X|G,\theta) = p(X_1)p(X_2|X_1)\dots p(X_p|X_1\dots X_{p-1}). \tag{2.6}$$

A rede Bayesiana $\mathcal{G}$ para um conjunto de variáveis X_j, $j = 1,\dots,p$, pode ser dividida em duas partes: (1) um grafo direcionado, no qual as ligações entre as variáveis representam a relação de dependência entre as variáveis selecionadas; (2) o conjunto das respectivas distribuições condicionais definidas no grafo (Heckerman 2008).

No contexto em que existem independências condicionais, a Equação (2.6) pode ser simplificada. A partir de uma determinada ordenação das variáveis X_j, é possível encontrar algumas relações de independência condicional que simplificam termos da distribuição $p(X)$.

Seja A uma matriz $p \times p$ que representa as relações de independência condicional entre as variáveis X_j, tal que $A = A_{ij}$, onde $A_{ij} = 1$ se X_i é pai de X_j, e $A_{ij} = 0$ caso contrário. Por exemplo, em uma configuração $p(X_1 \mid X_2, X_3)$, X_2 e X_3 são pais de X_1. Da mesma forma, filhos são as variáveis que condicionam a variável principal; nesse caso, X_1 é filho de X_2 e X_3. Denotamos ainda por $A_{i\cdot}$ a i-ésima linha de A, cujos elementos não-nulos indicam os pais da i-ésima variável, e por $A_{\cdot j}$ a j-ésima coluna, que indica os filhos da j-ésima variável.

$$A = \begin{array}{c} \\ X_1 \\ X_2 \\ \vdots \\ X_p \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & X_2 & \dots & X_p \end{array} \\ \begin{bmatrix} 0 & A_{12} & \dots & A_{1p} \\ 0 & \ddots & \dots & A_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \end{array}$$

Seja $pa(X_i)$ o vetor ordenado dos pais da i-ésima variável. Dessa forma, dada as relações de independência condicional definidas pela matriz A, podemos reescrever a Equação (2.6) da seguinte forma:

$$p(X) = p(X_1)\, p\big(X_2|pa(X_2)\big) \dots p\big(X_p|pa(X_p)\big). \tag{2.7}$$

Note que o caso mais simples para a matriz A determina que todas as variáveis são independentes mas, no contexto de estimar $p(X)$, não seria informativo. Já o caso completo, em que não se tem nenhuma independência condicional, o modelo é complexo e, dependendo do número de variáveis, pode ser inviável. Dessa forma, a rede Bayesiana simplificará a matriz A fazendo com que a estimação de $p(X)$ possa ser mais simples do que o modelo sem nenhuma independência.

Considere um exemplo com $p = 6$ variáveis aleatórias, em que o objetivo é modelar $p(X)$ utilizando uma rede Bayesiana. Suponha que, para um determinado indivíduo, as variáveis $X_1,\dots,X_6$ são variáveis binárias que indicam a presença ou ausência das

seguintes características: Espirro, Alergia, Arranhões, Rinite, Resfriado e Gato, respectivamente. Por exemplo, $X_1 = 1$ indica que o indivíduo tem espirros, enquanto $X_1 = 0$ indica que não tem. A Figura 2.1 ilustra uma possível configuração de rede Bayesiana para este cenário.

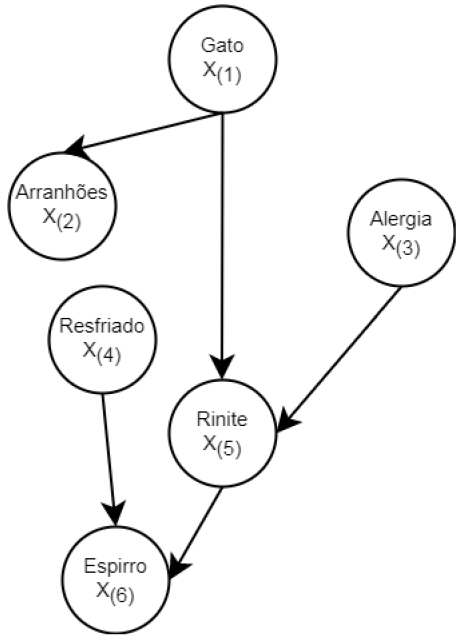


Figura 2.1: Rede Bayesiana para o exemplo do banco de dados de informações sobre donos de gatos.

A rede apresentada na Figura 2.1 indica as relações condicionais entre as variáveis, com as distribuições de probabilidade descritas na Tabela 2.1. Os valores atribuídos às probabilidades condicionais definem um valor dos parâmetros θ.

Tabela 2.1: Tabela de distribuições condicionais exemplo Gato.

$p(X_1 = 1) = 0,20$
$p(X_3 = 1) = 0,45$
$p(X_4 = 1) = 0,45$
$p(X_2 = 1 \mid X_1 = 1) = 0,95$
$p(X_2 = 1 \mid X_1 = 0) = 0,15$
$p(X_5 = 1 \mid X_3 = 1, X_1 = 1) = 0,01$
$p(X_5 = 1 \mid X_3 = 1, X_1 = 0) = 0,90$
$p(X_5 = 1 \mid X_3 = 0, X_1 = 1) = 0,85$
$p(X_5 = 1 \mid X_3 = 0, X_1 = 0) = 0,95$
$p(X_6 = 1 \mid X_4 = 1, X_5 = 1) = 0,01$
$p(X_6 = 1 \mid X_4 = 1, X_5 = 0) = 0,15$
$p(X_6 = 1 \mid X_4 = 0, X_5 = 1) = 0,10$
$p(X_6 = 1 \mid X_4 = 0, X_5 = 0) = 0,99$

Visto que todas as variáveis possuem distribuição Bernoulli, a distribuição completa contém 63 parâmetros a serem estimados, enquanto na rede Bayesiana definida pela

Figura 2.1 tem-se apenas 13. Isso ilustra claramente como a estrutura de rede Bayesiana pode contribuir para trazer parcimônia ao modelo, ou seja, podemos ter modelos menos complexos.

Note que, embora a Figura 2.1 represente o resultado de uma rede Bayesiana, conseguimos facilmente descrever as relações de independência condicional entre as variáveis. Assim, para o exemplo descrito, a distribuição conjunta será da forma:

$$\begin{aligned} p(X) &= p(X_{\text{Gato}})p(X_{\text{Alergia}})p(X_{\text{Resfriado}}) \\ &\quad \times p(X_{\text{Arranhoes}}|X_{\text{Gato}}) \\ &\quad \times p(X_{\text{Rinite}}|X_{\text{Alergia}}, X_{\text{Gato}}) \\ &\quad \times p(X_{\text{Espirro}}|X_{\text{Resfriado}}, X_{\text{Rinite}}) \\ &= p(X_1)p(X_2|X_1)p(X_3)p(X_4)p(X_5|X_3, X_1)p(X_6|X_4, X_5) \end{aligned}$$

Logo, temos que $pa(X_1) = \varnothing$, $pa(X_2) = X_1$, $pa(X_3) = \varnothing$, $pa(X_4) = \varnothing$, $pa(X_5) = (X_1, X_3)$ e $pa(X_6) = (X_4, X_5)$. Com isso, a rede Bayesiana pode descrever de maneira simplificada a distribuição conjunta de $p(X)$. Note que para esse modelo a matriz A será da forma:

$$A = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array},$$

O modelo de rede Bayesiana, embora pareça que esteja ligado apenas à inferência com abordagem Bayesiana, está relacionada à teoria de probabilidades de Bayes. Sendo assim, essa abordagem não se restringe apenas à inferência Bayesiana, a abordagem clássica também é vastamente utilizada na literatura.

Friedman & Koller (2003) apresentam um estudo com abordagem Bayesiana para a criação da rede. Nesse estudo, utiliza-se a ideia de que existe uma ordenação nos dados, e que essa ordenação faz com que o modelo seja mais simples. Logo, o algoritmo fica mais rápido, pois o espaço de possíveis relações entre as variáveis é menor. Os autores utilizam uma distribuição *a priori* para a ordem com base no número de possíveis pais (nós principais). Dessa maneira, é necessário, para que o algoritmo funcione, que exista uma limitação no possível número de pais para as redes construídas. Uma vez que a ordem é estimada, a rede é construída baseada nessa ordenação, o que facilita na criação do MCMC. Os autores constroem um escore baseado no número de pais da rede e em

quais variáveis ainda podem ser utilizadas no modelo com respeito à ordenação dessas variáveis. Os resultados também sofrem uma série de restrições, como o tamanho do banco de dados, número de pais, número máximo de candidatos a pais, e no número máximo de nós presente na rede. Os resultados sugerem que as aproximações feitas em alguns passos do algoritmo MCMC fazem com que as estimações sejam feitas mais rapidamente, mas o estudo não apresenta o impacto dessas aproximações nos resultados.

Koivisto & Sood (2004) apresentam uma extensão do trabalho de Friedman & Koller (2003), agora construindo um modelo exato com abordagem Bayesiana. Da mesma forma, os autores constroem o modelo baseado primeiramente na estimação da ordem dos dados, e depois na construção da rede. O algoritmo proposto busca encontrar estruturas de rede mais prováveis, fazendo com que o tempo de execução do algoritmo seja menor. O tempo do algoritmo é calculado com base no tamanho do banco de dados e na complexidade do algoritmo. O estudo apresenta resultados para alguns bancos de dados sintéticos de tamanho 17, 22, 37 e 100. Os resultados sugerem que é possível fazer uma estimação exata da distribuição *a posteriori* dos dados, mas a complexidade do modelo deve ser levada em consideração. Em casos de tamanho de banco de dados muito grande e com grande número de variáveis, essa metodologia tende a não ser viável.

Sun & Erath (2015*a*), Deeva et al. (2020) utilizam a função escore AIC (*Akaike information criterion*) para estimar a qualidade da estrutura da rede Bayesiana. Após estimar a melhor rede baseada nesse critério, os dados sintéticos são gerados a partir dos valores de $\mathbf{X}$ dada a probabilidade conjunta fatorada da distribuição de $p(X)$, que foi determinada pela rede. Zhang et al. (2017), em contrapartida, criam um algoritmo que, após estimar a rede Bayesiana, acrescenta um ruído em cada marginal de $p(X)$ para gerar os dados sintéticos e assim garantir uma melhor privacidade.

Zhang et al. (2017), em contrapartida, criam um algoritmo que, após estimar a rede Bayesiana, acrescenta um ruído em cada marginal de $p(X)$ para gerar os dados sintéticos e assim garantir uma melhor privacidade. No entanto, essa abordagem apresenta desvantagens significativas. A adição de ruído pode comprometer a utilidade dos dados sintéticos ao distorcer relações e estruturas originais, tornando as análises menos precisas. A escolha do nível de ruído é crítica: muito ruído pode causar perda excessiva de informação, enquanto pouco pode não garantir privacidade adequada. Além disso, a técnica pode não se adequar a todos os tipos de dados ou relações complexas, limitando sua aplicabilidade. Portanto, é essencial equilibrar a privacidade e a utilidade dos dados sintéticos ao aplicar essa abordagem.

Sun & Erath (2015*a*) estimam a rede para um banco de dados com 4 variáveis categóricas, com no máximo 3 categorias cada. Deeva et al. (2020) utilizam um banco de dados com informações sobre seguro contendo 7 variáveis categóricas com até 11 categorias. Por fim, Zhang et al. (2017) utilizam 2 bancos de dados distintos, sendo os dois com 10 variáveis categóricas. Uma das componentes desconhecidas da rede, que deve ser

estimada, é a ordem das variáveis para a distribuição conjunta.

Além disso, algumas propostas na literatura inserem algumas hipóteses e/ou restrições nas escolhas da distribuição *a priori* para a ordem. Por exemplo, Friedman & Koller (2003) assumem uma distribuição *a priori* uniforme sobre as $p!$ possíveis ordenações. Essa hipótese é utilizada num contexto em que se possui uma pequena quantidade de variáveis, e essas são categóricas e com poucas categorias. Ellis & Wong (2008) e Niinimäki et al. (2016) também assumem uma distribuição uniforme. Já Sun & Erath (2015*a*) fazem a estimação da rede com abordagem clássica, em que fixam a primeira variável da rede utilizando algum conhecimento prévio do problema. Sob a abordagem Bayesiana, essa restrição consiste em atribuir uma distribuição de probabilidade *a priori* igual a zero para todas as ordenações nas quais a variável escolhida não é a primeira.

Para a estrutura da rede, alguns trabalhos utilizam a quantidade de pais de cada variáveis para estipular a distribuição *a priori* (Friedman & Koller 2003, Koivisto & Sood 2004). Nesses trabalhos os autores definem a distribuição *a priori* da rede $\mathcal{G}$ baseada no número de pais de cada variável X_j. Basicamente, se assume que cada variável nó X_j possui k pais. Dessa forma existem $\binom{p-1}{k}$ possíveis conjuntos de pais. Assumindo que podemos escolher uniformemente os possíveis conjuntos de pais, a distribuição *priori* sob a estrutura da rede será:

$$p(\mathcal{G}) \propto \prod_{j=1}^{p} \binom{p-1}{|pa(X_j)|}^{-1},$$

no qual $|pa(X_j)|$ representa o número de pais da variável X_j.

Condicionado à ordenação, a estrutura da distribuição *a priori* será a mesma, agora condicionada no número de pais de cada ordenação. Para que as estimações sejam feitas, Friedman & Koller (2003) assumem que cada variável X_j possui apenas três pais, e o número máximo de potenciais pais é 20. Dessa forma, a estimação fica mais rápida. Em um cenário em que o banco de dados possui muitas variáveis, essa restrição pode não ser a ideal. Limitar o número de pais faz com que as redes criadas sejam pequenas e podem não refletir corretamente a distribuição conjunta dos dados.

Um conjunto interessante de distribuições *a priori* para a rede são as distribuições *a priori* modulares. De acordo com Friedman & Koller (2003), distribuições *a priori* modulares são aquelas em que a distribuição da *a priori* pode ser fatorada em componentes mais simples. Ela pode ser representada como um produto de termos fatoráveis. A escolha dessa distribuição é motivada pelo fato de que as distribuições *a priori* são bastante abrangentes e podem ser expressas de várias formas. Além disso, modularidade é necessária para se obter função de escore que pode ser decomposta. Vários algoritmos de rede Bayesiana apresentam a utilização de distribuição *a priori* modular por este fato. A distribuição *a priori* proposta por Friedman & Koller (2003) é modular e portanto oferece a propriedade de fatoração de seus termos.

Eggeling et al. (2019) apresentam algumas distribuições *a priori* modulares para a rede. A distribuição uniforme é a mais simples dentre elas, mas ela pode não ser apropriada em alguns casos, uma vez que a suposição que todas as redes tem a mesma probabilidade pode ser muito forte. Outra distribuição *a priori* discutida é a distribuição de Edge, que é equivalente ao modelo grafo aleatório proposto por Heckerman, Mamdani & Wellman (1995). Ela basicamente propõe uma nova estrutura de rede, apenas modificando a rede estimada anteriormente. A distribuição *a priori* proposta por Pensar et al. (2016) depende diretamente do número de variáveis do estudo.

Para a distribuição *a priori* sob os parâmetros (θ) das distribuições condicionais de $\mathbf{X}$, usualmente são escolhidas distribuições Dirichlet independentes para o vetor de probabilidades de cada multinomial, quando o espaço da rede é discreto. Friedman & Koller (2003), Koivisto & Sood (2004) e Ellis & Wong (2008) utilizam essa distribuição, uma vez que todos os experimentos são feitos com variáveis discretas. Em trabalhos como o de Chen et al. (2017), o banco de dados analisado possui variáveis contínuas, mas o trabalho apenas fornece uma metodologia para discretizar essas variáveis para que o modelo funcione. Quando os dados contínuos possuem distribuição normal, é comum a utilização de uma rede Bayesiana Gaussiana para a estimação da rede (Grzegorczyk 2010). Portanto, a distribuição *priori* sob os parâmetros do modelo será uma normal-Wishart.

Para fazer a estimação da rede, Friedman & Koller (2003) desenvolvem um MCMC que estima separadamente a ordenação dos dados, e depois, a estrutura da rede. Um algoritmo Metropolis-Hastings (Metropolis et al. 1953) é criado para amostrar a ordenação δ a partir da distribuição *a priori* escolhida. Para ordenação, os autores apenas mencionam que o algoritmo se inicia com uma ordenação aleatória e os pares de variáveis são trocados a fim de construir a ordenação. Os movimentos são propostos pela troca de um ou mais elementos na ordenação. Nesse estudo, não é mencionado se os parâmetros das distribuições são estimados ou fixados nesse MCMC. Depois, dada essa ordenação gerada anteriormente, é amostrada uma estrutura de rede.

Diferentes abordagens com relação à metodologia de estimação da rede também são encontradas na literatura. Por exemplo, Friedman & Koller (2003) utilizam um algoritmo com dois estágios. Primeiro, estima-se a ordenação dos dados, e depois, com base nessa informação, estima-se a estrutura da rede G no espaço compatível com a ordenação estimada de acordo com a distribuição de probabilidade condicional dada pela ordenação.

Com o mesmo objetico, Goudie & Mukherjee (2016) propõem um algoritmo MCMC para a estimação Bayesiana da rede. O algoritmo é um *random scan Gibbs sampling* com blocos definidos pelas colunas da matriz A. A cada iteração do algoritmo, d (geralmente 3) colunas são escolhidas aleatoriamente e amostradas da respectiva distribuição condicional completa. Essa amostragem é feita em duas etapas, de forma a viabilizar o seu custo computacional. O conjunto de todas as redes admissíveis, dada a configuração das demais colunas, é dividido de acordo com os possíveis filhos das variáveis pais representadas nas

colunas.

Uma das componentes da partição é amostrada e, na segunda etapa, a rede é amostrada dentre as que estão contidas nesta componente da partição. Esta estratégia reduz drasticamente o custo computacional para amostrar da distribuição condicional completa quando comparada ao algoritmo que amostra diretamente da mesma, calculando o vetor de probabilidades de todas as redes admissíveis. O algoritmo exige que a distribuição *a priori* da rede seja do tipo modular, ou seja, que tenha a seguinte forma:

$$p(G) \propto \prod_j p(pa(X_j)). \tag{2.8}$$

O algoritmo proposto por Goudie & Mukherjee (2016) será empregado neste trabalho para realizar a estimação da rede, visando determinar o modelo que melhor se ajusta aos dados. Posteriormente, utilizando essa informação, será possível gerar os dados sintéticos. Os autores disponibilizam o algoritmo completo para a implementação do Método de Cadeias de Markov Monte Carlo (MCMC) em um pacote `strucmcmc` (Goudie & Mukherjee 2016) no *R* (R Core Team 2023). Esse pacote oferece uma utilização simplificada, permitindo a definição das distribuições *a priori* e a entrada de dados de maneira direta. O pacote fornece algumas configuração de distribuições para a rede, bem como uma distribuição uniforme.

Após a execução do MCMC, obtemos como resultado uma série de redes estimadas pelo modelo, juntamente com outras estatísticas relevantes. Utilizando essa informação, torna-se possível gerar dados sintéticos através de modelos de simulação. O processo completo, incluindo mais detalhes sobre esse modelo, é abordado de forma mais aprofundada no Capítulo 3.

2.3 Análise de Dados Sintéticos via Redes Bayesianas

Como visto na Seção 2.2, o modelo de redes Bayesianas pode ser ajustado usando inferência clássica ou Bayesiana. Se o intuito é gerar dados sintéticos, algumas metodologias são aplicadas para esse fim (Deeva et al. 2020).

O uso da metodologia de geração de dados sintéticos é uma ferramenta importante na pesquisa atual. Ela nos permite explorar bancos de dados confidenciais, que muita vezes não seriam acessíveis ao público. Essa nova abordagem pode ajudar a preencher lacunas na disponibilidade de dados protegidos. Dentre as várias metodologias disponíveis para a análise e geração de dados sintéticos o modelo de rede Bayesianas se torna interessante.

Ao adotar a rede Bayesiana como o modelo para estimar a distribuição conjunta de $p(X)$, podemos entender as relações de independência condicional intrínseca aos dados e com essas informações podemos gerar bancos de dados mais informativos, e ao mesmo tempo proteger os dados reais. Assim, a metodologia de geração de dados sintéticos, especialmente quando baseada em Redes Bayesianas, se torna ferramenta promissora no estudo de dados simulados.

Na literatura sobre redes Bayesianas, é comum tanto o uso de inferência clássica como inferência Bayesiana para fazer a estimação da rede. Nos trabalhos em que o foco é a geração de dados sintéticos, é comum estimar a rede sob uma abordagem de inferência clássica (Deeva et al. 2020). Para escolher a melhor estrutura de rede, geralmente utiliza-se uma abordagem baseada em uma função *score*, que mostra o quão bem a estrutura se adequa aos dados.

Sun & Erath (2015*b*), Deeva et al. (2020), Zhang et al. (2017) utilizam redes Bayesianas para imputação de dados sintéticos. Nesses trabalhos, é utilizada a inferência clássica para a estimação da rede. O problema de estimação de uma rede pode ser dividido em dois tipos: (1) construção da estrutura da rede, e (2) identificação da melhor estrutura e parâmetros do modelo. Para a identificação da rede, é comum o uso de heurísticas gulosas (*greedy heuristics*) para encontrar as estruturas. Basicamente, esse tipo de metodologia se inicia com uma rede vazia, e a seguir se adiciona, exclui ou reverte ligações de independências condicionais para encontrar uma estrutura de rede com pontuação mais alta. Para escolher a melhor estrutura, é utilizada uma abordagem baseada em uma função *score* para verificar o quão bem a estrutura corresponde aos dados. Essa função é calculada baseada no log da função de máxima verossimilhança dos dados para uma dada estrutura, e geralmente possui algum critério de penalização para a complexidade da rede (Sun & Erath 2015*b*).

Embora os estudos mencionados visem a estimativa da rede e a geração de dados sintéticos, as metodologias empregadas para a estimativa da rede frequentemente recorrem a algoritmos de otimização. Embora essas abordagens frequentemente demonstrem resultados promissores, elas tendem a estar condicionadas a várias suposições, como a definição do número de variáveis e das relações entre essas variáveis. Diante desse cenário, esta pesquisa concentra-se na adoção de modelagens inteiramente Bayesianas para a estimação da rede. Essa abordagem destaca-se ao oferecer uma perspectiva mais flexível na estimativa da rede. Dessa forma, essas limitações podem ser adequadas a modelos mais reais. Logo, podemos ter uma contribuição valiosa usando redes para a simulação de dados.

Sun & Erath (2015*b*) apresentam um trabalho que utiliza redes Bayesianas para a geração dos dados sintéticos. O estudo apresenta a construção de um modelo Bayesiano para a rede, em que a distribuição *a posteriori* resultante é a distribuição conjunta dos dados de interesse, que é utilizada para a geração dos dados sintéticos.

Nesse estudo, os autores apresentam também o desempenho dessa abordagem na geração de dados simulados para um banco de dados de Cingapura. Esse banco contém dados de Censo e informações sobre donos de imóveis, como renda, número de moradores, idade, dentre outras. Os resultados mostram que a abordagem de rede Bayesiana proposta pode ser eficiente na caracterização da distribuição conjunta dos dados, que é o objetivo principal da inferência utilizando modelos de rede.

Os autores também comparam a metodologia de geração dos dados simulados através da rede Bayesiana com outras metodologias para o mesmo fim. Todo o estudo é construído para dados categorizados, sendo que o máximo de categorias é definido como 5. Essa característica presente nos dados facilita a estimação da rede, pois já sabemos previamente todas as possíveis relações entre as variáveis. Uma vez que o estudo é feito com o censo completo (os autores possuíam todos os dados do estudo), todas as distribuições condicionais verdadeiras do banco de dados foram calculadas previamente, ou seja, os autores puderam comparar os resultados com as verdadeiras distribuições dos dados.

Como resultado, o trabalho demonstra que, em um cenário onde temos apenas uma amostra do censo, a rede Bayesiana parece não sofrer grandes alterações quando esse tamanho da amostra varia, ou seja, os resultados são muito parecidos para diferentes tamanhos amostrais.

Embora o estudo de Sun & Erath (2015*b*) tenha apresentado bons resultados, eles estão limitados a uma série de restrições e os métodos são aplicáveis apenas a um cenário controlado. Apenas uma rede Bayesiana é estimada para os dados, sendo que a primeira variável de inicialização da rede foi fixada. Os autores não apresentam um estudo sobre o impacto da rede no sigilo dos dados, embora os resultados tenham sido satisfatórios.

Como discutido na Seção 2.1, divulgar apenas alguns bancos de dados pode não ser o melhor quando pensamos em quantificação de incerteza. Portanto, é necessário que a análise e geração dos dados sintéticos seja feita de maneira mais robusta possível.

Capítulo 3

Metodologia

Neste capítulo, apresentaremos a metodologia para simulação e análise de dados sintéticos através da rede Bayesiana. Para a estimação da rede usaremos o algoritmo de MCMC proposto por Goudie & Mukherjee (2016). Será apresentada na Seção 3.2 uma proposta de uma nova classe de distribuições *a priori* modulares para a rede Bayesiana. Na Seção 3.3, apresentaremos uma proposta de geração de dados sintéticos com o foco na distribuição preditiva *a posteriori* de estatísticas dos dados sintéticos.

3.1 Modelo e Notação

Considere o contexto e notação introduzidos na Seção 2.2. Como descrito brevemente no capítulo de introdução, nossa metodologia consiste de uma abordagem para geração e análise de dados sintéticos via redes Bayesianas sob o paradigma Bayesiano. Iremos propor uma análise Bayesiana completa visando uma quantificação de incerteza mais robusta e eficiente.

Considere a notação $\pi(.)$ para representar distribuições de probabilidade ou densidade das variáveis de interesse. Em uma rede Bayesiana, o aprendizado da estrutura da rede é baseado na distribuição *a posteriori* $\pi(G|\mathbf{X})$. O suporte desta distribuição é o espaço $\mathcal{G}$, que inclui todos os DAGs possíveis com p vértices. O número de possíveis DAGs cresce exponencialmente com o número de variáveis p. Kjaerulff & Madsen (2008) fornecem uma fórmula para calcular o número de DAGs possíveis com base no número de vértices, conforme apresentado na Tabela 3.1.

Portanto, os algoritmos de MCMC são especialmente valiosos porque, na prática, uma pequena parte do espaço $\mathcal{G}$ concentra quase toda a massa de probabilidade *a posteriori*. Isso indica que a maior parte das probabilidades *a posteriori* está localizada em um subconjunto relativamente pequeno do espaço de parâmetros. Como resultado, os algoritmos MCMC podem focar nesse subconjunto de alta probabilidade, em vez de explorar todo o espaço de forma exaustiva. Isso torna o processo de amostragem mais

Tabela 3.1: Número de possíveis DAGs em função do número de vértices na rede.

N° de vértices	N° de possíveis DAGs
1	1
2	2
3	25
4	543
5	29,281
6	3,781,503
7	1,138,779,265
8	783,702,329,343
9	1,213,442,454,842,881
10	4,175,098,976,430,598,143

eficiente, reduzindo o número de amostras necessárias para obter estimativas precisas e aumentando a eficiência computacional do algoritmo.

Temos como objetivo estimar a rede Bayesiana condicionada aos dados originais. Após a estimação desse modelo, iremos obter a distribuição preditiva *a posteriori* dos dados sintéticos e de suas respectivas estatísticas de interesse. Definimos θ como sendo todos os parâmetros que indexam todas as possíveis distribuições condicionais ou marginais da rede e θ_G como aqueles que indexam a rede G. Seja $\mathbf{X}$ o conjunto de dados original, consistindo em uma amostra i.i.d. de tamanho n do vetor aleatório $X = (X_1, \ldots, X_d)$ de variáveis Bernoulli. G é a rede Bayesiana que define a distribuição de X, juntamente com as respectivas distribuições condicionais, que são indexadas por parâmetros θ. Defina $\mathbf{Y}$ como os dados sintéticos, consistindo em n replicações i.i.d. de $(X_1, \ldots, X_d)$. O modelo Bayesiano completo sobre $(\mathbf{X}, G, \theta, \mathbf{Y})$ pode ser fatorado da seguinte forma:

$$p(\mathbf{X}, G, \theta\mathbf{Y}) = \pi(\mathbf{X}|G, \theta)\pi(\mathbf{Y}|G, \theta)\pi(\theta|G)\pi(G). \tag{3.1}$$

Matematicamente, o objetivo da análise é obter a distribuição *a posteriori* de $(G, \theta, \mathbf{Y})$, ou seja, $p(G, \theta, \mathbf{Y}|\mathbf{X})$. Com base em nosso objetivo final e para otimizar o custo computacional, elaboramos um algoritmo que obtém as seguintes distribuições (nesta ordem específica):

1. a distribuição marginal *a posteriori* de G;
2. a distribuição condicional *a posteriori* de $(\theta|G, \mathbf{X})$;
3. a distribuição marginal *a posteriori* $p(\mathbf{Y}|\mathbf{X})$, também chamada de distribuição preditiva *a posteriori* de $\mathbf{Y}$.

A distribuição preditiva *a posteriori* de $\mathbf{Y}$ será dada por:

$$\pi(\mathbf{Y}|\mathbf{X}) = \int \pi(\mathbf{Y}|G,\theta)\pi(G,\theta|\mathbf{X})d\theta dG, \tag{3.2}$$

no qual fazemos a suposição razoável de independência condicional de $\mathbf{X}$ e $\mathbf{Y}$.

Considere $\mathbf{X}$ uma matriz $n \times p$, e assuma que as linhas $X_{i.}$ são condicionalmente independentes e identicamente distribuídos, dado (G,θ), para todo i. Em particular, vamos considerar que cada X_i é uma variável aleatória binária, portanto, as entradas do parâmetro θ são probabilidades. O número de pais de cada variável define a quantidade de distribuições condicionais e, consequentemente, a quantidade de parâmetros que indexam a distribuição da rede.

Para que seja possível utilizar o algoritmo de Goudie & Mukherjee (2016), iremos sempre considerar distribuições *a priori* modulares para a rede, ou seja, $\pi(G) = \prod_j \pi_j(pa(X_j))$. O fato das variáveis da rede serem binárias e a distribuição *a priori* da rede ser modular, possibilita uma simplificação substancial do processo de inferência. Especificamente, *a posteriori* da rede $\pi(G|X)$ é obtida utilizando o algoritmo MCMC de Goudie & Mukherjee (2016). A distribuição *a posteriori* condicional $\pi(\theta|X,G)$ tem forma fechada, sendo esta uma distribuição Beta para cada componente de θ, quando uma distribuição *a priori* Beta(α_j,β_j) é utilizada para cada θ_j que indexam a rede G.

Para que o algoritmo de MCMC seja aplicado a uma cadeia de Markov apenas em G, é necessário realizar a seguinte marginalização Goudie & Mukherjee (2016):

$$\begin{aligned}
\pi(\mathbf{X}|G) &= \int \pi(\mathbf{X}|G,\theta_G)\pi(\theta_G)d\theta_G \\
&= \int \prod_i^n \pi(X_{i.}|G,\theta_G) \prod_{j=1}^{J_G} \pi(\theta_j)d\theta_j \\
&= \prod_{j=1}^{J_G} \int_0^1 \theta_j^{z_j}(1-\theta_j)^{n_j-z_j} \frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j-1}(1-\theta_j)^{\beta_j-1}d\theta_j \\
&= \prod_{j=i}^{J_G} \frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \frac{\Gamma(\alpha_j+z_j)\Gamma(\beta_j+n_j-z_j)}{\Gamma(\alpha_j+\beta_j+n_j)} \\
&\quad \times \int_0^1 \frac{\Gamma(\alpha_j+\beta_j+n_j)}{\Gamma(\alpha_j+z_j)\Gamma(\beta_j+n_j-z_j)} \theta_J^{\alpha_j+z_j-1}(1-\theta_j)^{\beta_j+n_j-z_j-1}d\theta_j \\
&= \prod_{j=1}^{J_G} \frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \frac{\Gamma(\alpha_j+z_j)\Gamma(\beta_j+n_j-z_j)}{\Gamma(\alpha_j+\beta_j+n_j)},
\end{aligned} \tag{3.3}$$

em que J_G é o total de parâmetros θ_j que indexam a rede G, n_j é o número de observações, dentre as n, nas quais ocorre o evento no qual a distribuição indexada por θ_j é condicionada, e z_j o número de sucessos dentre essas n_j observações.

A expressão da verossimilhança marginal em (3.3) mostra explicitamente como a complexidade da rede Bayesiana é penalizada pela dimensão de seu espaço paramétrico. A informação a posteriori sobre os parâmetros β será dominada pela informação da verossimilhança (comparada à anterior), o que significa que cada termo do produto em (3.3) é dominado pela segunda razão. Em particular, se $\alpha_j = 1$ e $\beta_j = 1$, para todo j, a primeira razão é igual a 1. Agora, observe que, como todas as segundas razões estão entre $(0, 1)$ e ficam menores para valores maiores de n_j, a verossimilhança marginal equilibra o ajuste do modelo e a complexidade da seguinte maneira: quanto mais complexa a rede Bayesiana, mais termos a verossimilhança tem, mas esses termos têm valores mais altos (menores n_j).

Finalmente, note que a distribuição a posteriori condicional $p(\theta|G, X)$ é independente para todos os θ_j que indexam G, com $(\theta_j|G, X) \sim \text{Beta}(\alpha_j + z_j, \beta_j + n_j - z_j)$. O algoritmo para obter a distribuição preditiva de Y é apresentado na Seção 3.3.

Observe que podemos ter p ou mais parâmetros θ que indexam a rede G pois, se uma variável tiver mais de um pai, sua condicional terá mais de um θ. Por exemplo, a rede $\pi(\mathbf{X}|G) = \pi(X_1|X_2)\pi(X_2)\pi(X_3)$ tem quatro parâmetros. Note que, se todas as variáveis forem independentes, então $n_j = n$.

No problema descrito na Seção 2.2, suponha uma amostra de tamanho $n = 1000$, em que a Tabela 3.2 apresenta novamente as distribuições condicionais do exemplo.

Tabela 3.2: Tabela de distribuições condicionais exemplo Gato.

$p(X_1 = 1) = 0,20$
$p(X_3 = 1) = 0,45$
$p(X_4 = 1) = 0,45$
$p(X_2 = 1\|X_1 = 1) = 0,95$
$p(X_2 = 1\|X_1 = 0) = 0,15$
$p(X_5 = 1\|X_3 = 1, X_1 = 1) = 0,01$
$p(X_5 = 1\|X_3 = 1, X_1 = 0) = 0,90$
$p(X_5 = 1\|X_3 = 0, X_1 = 1) = 0,85$
$p(X_5 = 1\|X_3 = 0, X_1 = 0) = 0,95$
$p(X_6 = 1\|X_4 = 1, X_5 = 1) = 0,01$
$p(X_6 = 1\|X_4 = 1, X_5 = 0) = 0,15$
$p(X_6 = 1\|X_4 = 0, X_5 = 1) = 0,10$
$p(X_6 = 1\|X_4 = 0, X_5 = 0) = 0,99$

Considere a rede Bayesiana:

$$\pi(\mathbf{X}|G) = \pi(X_1)\pi(X_3)\pi(X_4)\pi(X_2|X_1)\pi(X_5|X_3, X_1)\pi(X_6|X_4, X_5).$$

Neste caso, temos $J_G = 13$ parâmetros a serem estimados. Para a variável X_1, temos que $n_j = n = 1000$ e $z_j = 200$. A mesma lógica se aplica para X_3 e X_4. Para

$\pi(X_2 = 1|X_1 = 1)$, temos $n_j = 200$, que é o número de observações em que $X_1 = 1$, e $z_j = 190$.

É fundamental observar que, em nosso modelo, é necessário levar em consideração possíveis casos de redundâncias, ou seja, redes distintas no espaço de DAGs mas que definem a mesma distribuição conjunta. Nestes casos, é razoável identificar tais redundâncias e tratar essas redes como uma só. Formalmente, este é um problema de identificabilidade. Por exemplo, as redes $\pi(\mathbf{X}) = \pi(X_1)\pi(X_2|X_1)\pi(X_3)$ e $\pi(\mathbf{X}) = \pi(X_2)\pi(X_1|X_2)\pi(X_3)$ são redundantes. As matrizes A_1 e A_2 representam esses dois modelos respectivamente:

$$A_1 = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array},$$

$$A_2 = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}.$$

Casos em que o número de variáveis e pais é maior, a identificação de redundâncias não é trivial. Note, porém, que no caso em que a rede é utilizada para a geração e análise de dados sintéticos, a metodologia aqui proposta não é afetada por esta questão. Já metodologias que utilizam somente a rede mais provável *a posteriori*, pode ter sérios problemas. Em particular, o problema de acabar utilizando uma rede que não define o modelo (após remover as redundâncias) com maior probabilidade *a posteriori*.

O algoritmo de MCMC a ser utilizado está definido no espaço de DAGs e, portanto, suscetível à questão de redes redundantes. Para reportar as estatísticas *a posteriori* da rede, iremos remover todas as redundância simples, como a do exemplo acima. No caso mencionado, temos três variáveis binárias X_1, X_2 e X_3. A redundância simples pode ser descrita comparando duas expressões condicionais diferentes em que somente uma variável é condicionada, mas que resultam na mesma dependência ou independência entre as variáveis. Por exemplo as duas representações fornecidas são: $X_1, X_1|X_2, X_3|X_1, X_2$ e $X_1, X_2|X_1, X_3|X_1, X_2$ levam a mesma distribuição de probabilidade.

3.2 Distribuições *a priori*

A distribuição *a priori* de G tem um papel muito importante na estimação da rede. O modelo de redes Bayesianas tem uma estrutura de aninhamento entre as possíveis redes, de forma que redes mais simples (com mais independência condicional) são casos particulares de redes mais complexas. Por exemplo, a rede completa é o modelo mais geral que tem todas as demais redes como casos particulares, para determinadas escolhas dos parâmetros θ. Dessa forma, se uma rede G_0 é um caso particular de uma rede mais geral G_1, o valor da função de verossimilhança em G_1 sempre será maior ou igual que o valor dessa função em G_0, para os respectivos valores maximizadores de θ. Dessa forma, se uma distribuição *a priori* uniforme for adotada para a rede, a única fonte de penalização de redes mais complexas, sob a abordagem Bayesiana, é o maior número de parâmetros θ. Dependendo do número de variáveis e do tamanho do conjunto de dados, esta penalização pode não ser o suficiente para balancear adequadamente o *trade-off* entre ajuste e parcimônia. Note que a justificativa para se utilizar redes Bayesianas é exatamente permitir uma modelagem parcimoniosa da distribuição de $\mathbf{X}$. Dessa forma, se faz necessário penalizar de forma mais eficiente e robusta a complexidade da rede o que, sob a abordagem Bayesiana, é feito através da distribuição *a priori* da rede.

Como dito anteriormente, iremos adotar uma distribuição *a priori* modular, como exigido pelo algoritmo de Goudie & Mukherjee (2016). Dadas as poucas opções deste tipo de distribuição *a priori* na literatura, propomos aqui uma classe geral de distribuição *a priori* penalizadoras para a rede. Especificamente,

$$\begin{aligned}\pi(G) &\propto exp\Big[-\sum_j h_j(pa(X_j);\gamma_j)\Big] \\ &= \prod_j exp\Big[-h_j(pa(X_j);\gamma_j)\Big], \end{aligned} \tag{3.4}$$

para funções não-negativas h_j que dependem da rede apenas através do conjunto de pais de cada variável.

Os termos α e γ desempenham papéis importantes no controle da complexidade e das penalizações aplicadas à estrutura. O parâmetro γ_j controla a intensidade da penalização que é aplicada à função de penalidade h_j. Pode ser interpretado como um fator que ajusta o custo de adicionar mais pais a uma variável X_j. Se γ_j for grande, a penalização será maior, o que desencoraja a adição de muitos pais para X_j, resultando em uma rede mais simples. Por outro lado, valores menores de γ_j permitem redes mais complexas, pois a penalização pela adição de pais a X_j será mais suave.

Esta especificação permite o uso de diversas estruturas de penalização. Um caso

particular simples e intuitivo é assumir

$$h_j = \gamma |pa(X_j)|^{\alpha}. \tag{3.5}$$

Na Seção 4.1.1, apresentamos um estudo de calibração do hiperparâmetro γ em função do número de variáveis e do tamanho da amostra, quando $\alpha = 1$.

3.3 Simulação dos dados sintéticos

A escolha da metodologia para a geração de dados sintéticos afeta tanto a qualidade das análises realizadas quanto a segurança dos dados originais. Em geral, quanto maior for a proporção preservada dos dados originais, mais fiel será a análise dos dados sintéticos em relação à análise com os dados originais. No entanto, isso implica uma menor segurança dos dados, e vice-versa.

Uma abordagem coerente para lidar com esse *trade-off* entre qualidade da análise e segurança é substituir todo o conjunto de dados original por um conjunto de dados sintético, utilizando uma metodologia que seja eficiente e robusta para aproximar os resultados obtidos diretamente dos dados originais. Nesse sentido, uma abordagem promissora é utilizar a distribuição preditiva *a posteriori* dos dados sintéticos. O paradigma Bayesiano não só utiliza a informação contida nos dados de forma eficiente e robusta, como também permite a quantificação da incerteza envolvida de maneira eficaz. Esta é a ideia principal da metodologia proposta nesta tese, que pode ser formalizada como se segue.

O problema estatístico tratado aqui consiste em fazer análises estatísticas sem utilizar o conjunto de dados originais, mas que levem a resultados que se aproximem ao máximo dos resultados que seriam obtidos se os dados originais fossem utilizados. De forma geral, podemos definir o resultado de uma análise estatística de um conjunto de dados $\mathbf{X}$ como um conjunto $h_1(\mathbf{X}), h_2(\mathbf{X}), \ldots$ de estatísticas. Essas podem ser, por exemplo:

1. estatísticas descritivas (média, mediana, variância, etc.);
2. estimadores pontuais e intervalares de parâmetros que indexam o modelo assumido para os dados;
3. valor-p de testes de hipóteses sobre o modelo.

A metodologia proposta nesta tese consiste em obter a distribuição preditiva *a posteriori* dessas estatísticas para os dados sintéticos Y, ou seja,

$$\pi(h_1(Y), h_2(Y), \ldots | \mathbf{X}). \tag{3.6}$$

No contexto de MCMC considerado aqui e descrito na Seção 3.1, uma amostra da distribuição preditiva de interesse é obtida a partir de uma amostra de MCMC da distribuição *a posteriori* preditiva de Y. Esta, por sua vez, é obtida gerando diretamente da distribuição $\pi(Y|G,\theta)$, para cada par (G,θ) gerado da distribuição *a posteriori* via MCMC. Este algoritmo é fundamentado pela expressão em (3.2).

O algoritmo abaixo descreve todo o processo de inferência proposto.

Algoritmo 1: Algoritmo para geração $\pi(h_1(Y), h_2(Y), \ldots|\mathbf{X})$

Entrada: Conjunto de dados original (tamanho n); rede inicial para o MCMC.
Rode o algoritmo de MCMC de Goudie & Mukherjee (2016) que converge para a distribuição *a posteriori* $\pi(G|\mathbf{X})$;
Guarde uma amostra de tamanho M de G - $(G^{(1)}, \ldots, G^{(M)})$, do MCMC, para escolhas razoáveis de *burn-in* e *lag*;
Faça $m = 1$
enquanto $m \leq M$, **faça**
    Gere $\theta^{(m)} \sim \pi(\theta|G^{(m)}, \mathbf{X})$;
    Gere um conjunto de dados sintéticos $Y^{(m)} \sim \pi(Y|G^{(m)}, \theta^{(m)})$ de tamanho n;
    Calcule $h_1(Y^{(m)}), h_2(Y^{(m)}), \ldots$;
    Faça $m = m + 1$.
fim

Os valores do *burn-in*, *lag* e M devem ser escolhidos de forma a obter o maior tamanho efetivo de amostra possível para as estatísticas h com o menor custo computacional possível. Basicamente, o *lag* deve ser escolhido de forma a mitigar a autocorrelação da cadeia no sentido de os tamanhos efetivos de amostra das h's ser aproximadamente M e este, por sua vez, ser pelo menos 500.

O algoritmo acima pode ser utilizado de formas diferentes com relação ao produto/*output* a ser entregue ao usuário final (que não pode ter acesso aos dados originais). Alguns exemplos desses produtos são os seguintes, em ordem crescente de complexidade.

1. Distribuição *a posteriori* de G representada pelo distribuição empírica (redes e frequências relativas) da amostra $(G^{(1)}, \ldots, G^{(M)})$ obtida no Algoritmo 1.

2. Os M conjuntos de dados sintéticos gerados no Algoritmo 1.

3. Apenas alguns (5 a 10) dos M conjuntos de dados sintéticos do item anterior.

4. A amostra de $(h_1(Y), h_2(Y), \ldots)$ gerada no Algoritmo 1.

5. A média e intervalo de credibilidade de $(h_1(Y), h_2(Y), \ldots)$ obtidos a partir da amostra gerada no Algoritmo 1.

A escolha do produto a ser fornecido depende do objetivo e expertise do usuário final. Por exemplo, utilizando o produto 1 acima, que é o mais simples, o usuário final

pode reproduzir todo o Algoritmo 1, substituindo o MCMC pela geração de uma amostra i.i.d. da distribuição empírica fornecida.

Capítulo 4

Análise de Dados

Neste capítulo serão apresentados estudos com dados simulados e reais para investigar a eficiência e aplicabilidade da metodologia proposta no capítulo anterior.

As análises com dados simulados tem dois objetivos principais, verificar a eficiência da estimação da rede Bayesiana usando o algoritmo de Goudie & Mukherjee (2016) e, principalmente, analisar a eficiência da metodologia proposta para geração e análise de dados sintéticos. Para isso serão gerados cenários variados em relação ao número de variáveis, tamanho do conjunto de dados e distribuições *a priori* para a rede. Os valores verdadeiros dos respectivos parâmetros θ variam entre 0.2 e 0.8. Para cada cenário determinado por estes valores, serão geradas $m = 10$ replicações (conjuntos de dados). Os cenários simulados estão descritos na Tabela 4.1.

Tabela 4.1: Cenários considerados para os dados simulados.

Qtd Variáveis	**Tamanho Amostra**	**Rede**
3	500 1000 5000	$X_1, X_2 \vert X_1, X_3$
4	1000 5000	$X_1, X_2 \vert X_1, X_3 \vert X_4, X_4$
7	2000 5000	$X_1, X_2 \vert X_1, X_3, (X_4 \vert X_3, X_5), X_6, (X_5 \vert X_6, X_7), X_7$

Para investigar a eficiência do algoritmo MCMC para estimação da rede, serão analisadas estatísticas *a posteriori*, como a probabilidade *a posteriori* da rede mais provável e da rede verdadeira. Sempre que possível, redes redundantes serão identificadas para o cálculo dessas probabilidades.

Para investigar a sensibilidade da distribuição *a priori* da rede, todas as análise serão feitas considerando duas distribuições *a priori*, uma distribuição uniforme e priori penalizadora apresentada na Equação (3.5).

No contexto da análise dos dados sintéticos, realizaremos uma comparação entre três métodos distintos de geração, denominados aqui por S_1, S_2 e S_3. S_1 se refere à abordagem proposta no Capítulo 3, ou seja, considerando a distribuição preditiva *a posteriori*

dos dados sintéticos. S_2 se refere a uma metodologia amplamente utilizada na literatura, consistindo na geração de 5 conjuntos de dados sintéticos utilizando a rede Bayesiana mais provável a *posteriori* e o EMV para θ. Esses conjuntos serão posteriormente combinados conforme a metodologia detalhada no Capítulo 2.

O método S_3 utiliza a técnica de simulação de dados sintéticos proposta por Nowok et al. (2022), disponível no pacote *synthpop* do software R, para gerar um único banco de dados sintético. Nesse pacote, os autores empregam um modelo não Bayesiano para a geração dos dados, utilizando o método de Ajuste Proporcional Iterativo (IPF). Essa abordagem busca estabelecer as distribuições marginais presentes no banco de dados original. Em cada iteração, pesos são calculados para cada registro no conjunto de dados original. A cada passo, o IPF avalia se as distribuições do conjunto de dados ajustado se aproximam das distribuições marginais desejadas. Se a convergência for alcançada, o processo é encerrado, caso contrário, as iterações prosseguem.

A escolha de empregar os três métodos distintos de geração de dados sintéticos se justifica pela busca de uma abordagem abrangente e comparativa para avaliar a qualidade e a representatividade dos dados gerados. Em particular, pretende-se investigar a eficiência e robustez da metodologia proposta com relação à: 1. aproximação das análises feitas com os dados originais; 2. quantificação de incerteza envolvida no processo de inferência.

Consideraremos três estatísticas dos dados sintéticos para realizar as análises - h_1, h_2 e h_3. A primeira estatística é uma função univariada de comparação entre os intervalos de confiança de 95%, para um determinado parâmetro θ, obtidos com os dados originais e com os dados sintéticos. Especificamente, essa função é uma medida de sobreposição entre os dois intervalos. A estatística h_2 é o EMV de um determinado parâmetro θ. Por fim, h_3 é o valor-p do um teste qui-quadrado de independência entre duas variáveis X_j escolhidas.

As análises consistirão de dois procedimentos principais: 1. comparar os resultados obtidos com as três metodologias de dados sintéticos com os respectivos resultados obtidos com os dados originais; 2. comparar os resultados obtidos com os métodos S_2 e S_3 com aqueles obtidos por S_1 para investigar o quanto os dois primeiros comprometem a quantificação de incerteza.

A medida h_1 de sobreposição de intervalos de confiança é definida como:

$$O_{IC} = \frac{2 \times (\min\{u_1, u_2\} - \max\{l_1, l_2\})}{(u_1 - l_1) + (u_2 - l_2)}, \tag{4.1}$$

em que (l_1, u_1) é o intervalo de confiança com os dados originais e (l_2, u_2) é o intervalo de confiança com os dados sintéticos. Esta medida será calculada para vários θ. Quanto maior for esta medida, melhor é o método para aproximar os resultados da análise correspondente feita com os dados originais.

Os métodos S_2 e S_3 retornam valores pontuais para as estatísticas h. No caso de S_2, a estatística é obtida para cada um dos 5 bancos sintéticos e combinados, utilizando as medidas de utilidade apresentadas na Seção 2.1.1, para obter o valor final. No caso de S_3, apenas 1 banco sintético é considerado, retornando o valor correspondente da estatística.

Já o método proposto S_1, retorna a distribuição *a posteriori* preditiva de cada h. Essas serão apresentadas através de sua média *a posteriori* e intervalo de credibilidade de 98%.

Apresentamos, na Seção 4.1.1, uma abordagem robusta para escolher o hiperparâmetro γ para a classe de distribuições *a priori* para a rede proposta no Capítulo 3, em função do número de variáveis e do tamanho da amostra. Essa abordagem será então utilizada para escolher os valores de γ nas análises com os dados simulados para duas escolhas distintas de α, a saber, 0 (uniforme) e 1.

A análise do banco de dados real será feita de maneira simular à análise dos dados simulados. Serão considerados os mesmos três métodos S_1, S_2 e S_3, as mesmas três estatísticas h e as mesmas duas escolhas de α para a distribuição *a priori* da rede.

4.1 Estudo com dados simulados

A Tabela 4.2 apresenta o tempo computacional necessário para gerar os dados sintéticos em diferentes cenários simulados. Os tempos computacionais são fornecidos para cada combinação de quantidade de variáveis e tamanho da amostra. Esses valores representam o tempo necessário para gerar todos os dados sintéticos correspondentes a cada configuração, refletindo a complexidade e o custo computacional associados ao processamento de diferentes volumes e estruturas de redes.

Tabela 4.2: Tempo Computacional para cada cenário simulado

Qtd Variáveis	**Tamanho Amostra**	**Tempo Computacional**
3	500	10 min
	1000	12 min
	5000	20 min
4	1000	120 min
	5000	270 min
7	2000	349 min
	5000	723 min

4.1.1 Calibração da distribuição a priori da rede

Com o objetivo de selecionar valores razoáveis para o hiperparâmetro γ na distribuição a priori na Equação (3.5), conduzimos um estudo utilizando os cenários descritos na Tabela 4.1.

O objetivo principal é entender a relação entre γ e o nível de penalização da priori, em função do número de variáveis e do tamanho do conjunto de dados, quando $\alpha = 1$. Note que, quanto maior for γ, mais forte é a penalização de redes mais complexas (com mais pais).

A priori uniforme na rede corresponde à priori em (3.5) com $\gamma = 0$. Como os cenários testados aqui consideram redes verdadeiras consideravelmente mais simples que a rede completa, espera-se que, partindo de $\gamma = 0$, a medida que o valor de γ aumente, também aumente a probabilidade *a posteriori* da rede verdadeira. Dessa forma, a calibração deste hiperparâmetro será feita buscando o seu menor valor para o qual a probabilidade *a posteriori* da rede verdadeira excede 0,85.

Espera-se que os valores encontrados para γ, em função do número de variáveis e do tamanho de amostra, sejam escolhas robustas para tratar o *trade-off* ajuste versus parcimônia sempre que redes consideravelmente mais simples que a completa sejam uma boa solução.

A Figura 4.1 mostra a relação entre γ e a probabilidade *a posteriori* da rede verdadeira.

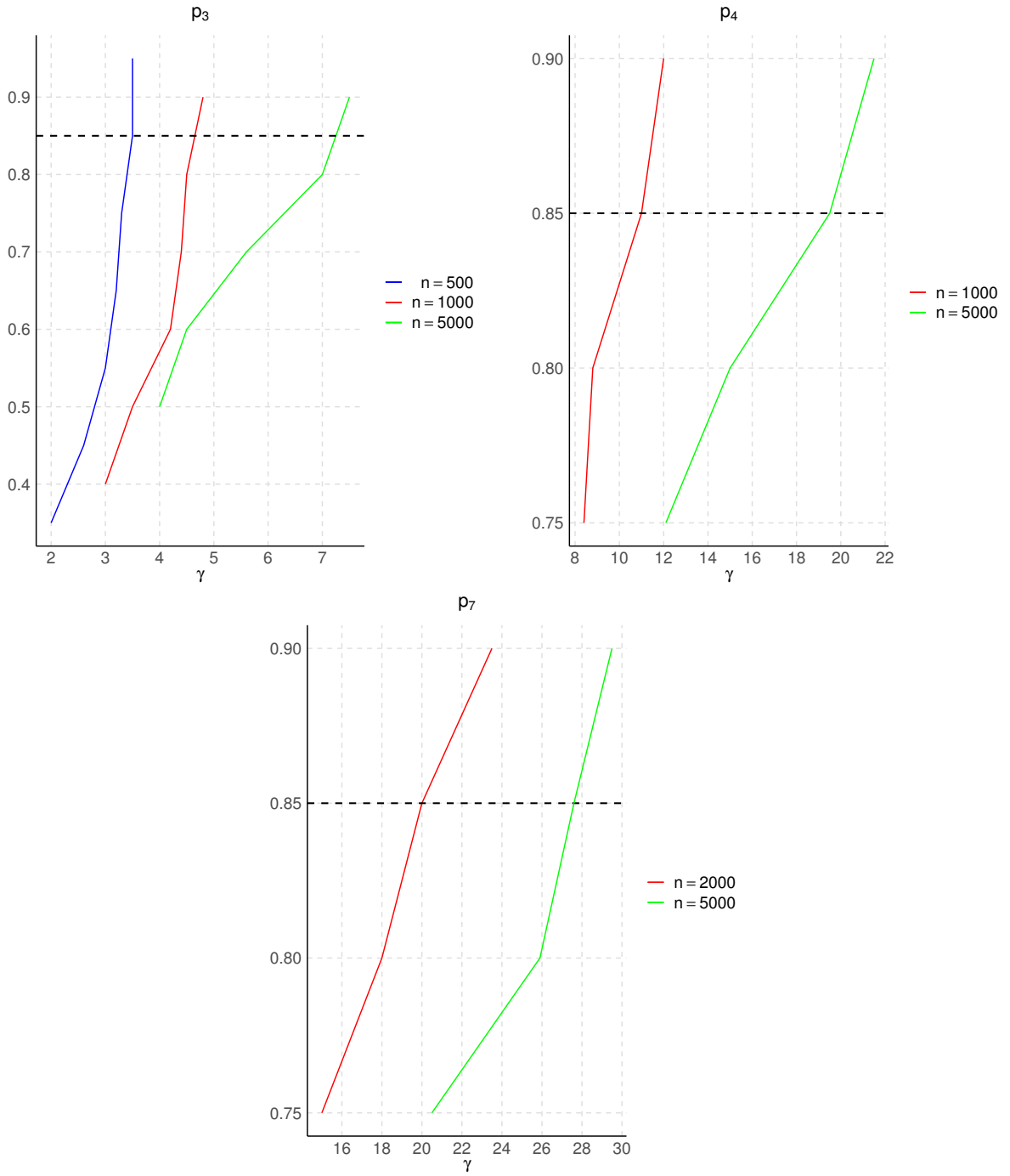


Figura 4.1: Gráfico dos valores de γ versus a probabilidade *a posteriori* da rede verdadeira.

4.1.2 Estimação da rede

Para cada um dos cenários na Tabela 4.1 são geradas 10 replicações com o objetivo de analisar a precisão dos resultados e sua confiabilidade. Os valores reais dos parâmetros θ são escolhidos entre 0.2 e 0.8 e *prioris* $U(0,1)$ são adotadas para esses. São consideradas

duas distribuições *a priori* para a rede, uma uniforme (P_2) e a distribuição proposta em (3.5) (P_1) e calibrada como explicado na Seção 4.1.1. Essas prioris impactam nos resultados dos métodos S_1 e S_2, que utilizam o modelo de rede Bayesiana. No caso do método S_3 os rótulos P_1 e P_2 nos gráficos se referem a duas replicações (dois conjuntos de dados sintéticos) do método.

Sob o método proposto, S_1, redes redundantes simples serão consideradas como sendo a mesma rede, já que definem a mesma estrutura de dependência. Para as demais redundâncias, apesar de potencialmente comprometer o cálculo da rede mais provável *a posteriori*, a não-junção destas redes não compromete a análise dos dados sintéticos uma vez que a mesma é baseada na distribuição preditiva *a posteriori*.

Os tamanhos de cadeia MCMC para a rede variaram entre 5.000 e 80.000 iterações, modelos com mais variáveis e mais dados necessitam de mais iterações. Os tamanhos de *burn-in* variaram entre 1.000 e 10.000. Para o método proposto S_1, considera-se um *lag* razoável para selecionar a amostra *a posteriori* da rede, de forma a mitigar o efeito da autocorrelação da cadeia e poder trabalhar com tamanhos de amostra que não comprometam o custo computacional. Os *lags* escolhidos variam entre 8 e 140 e visam retornar um tamanho efetivo de amostra de pelo menos 500 para as estatísticas h.

Para avaliar a precisão da estimação da rede, calculamos o número de replicações em que a rede verdadeira foi a mais provável *a posteriori* e a média (sob essas replicações). Apresentamos também a média dessa probabilidade entre as demais replicações. Esses resultados são apresentados na Tabela 4.3.

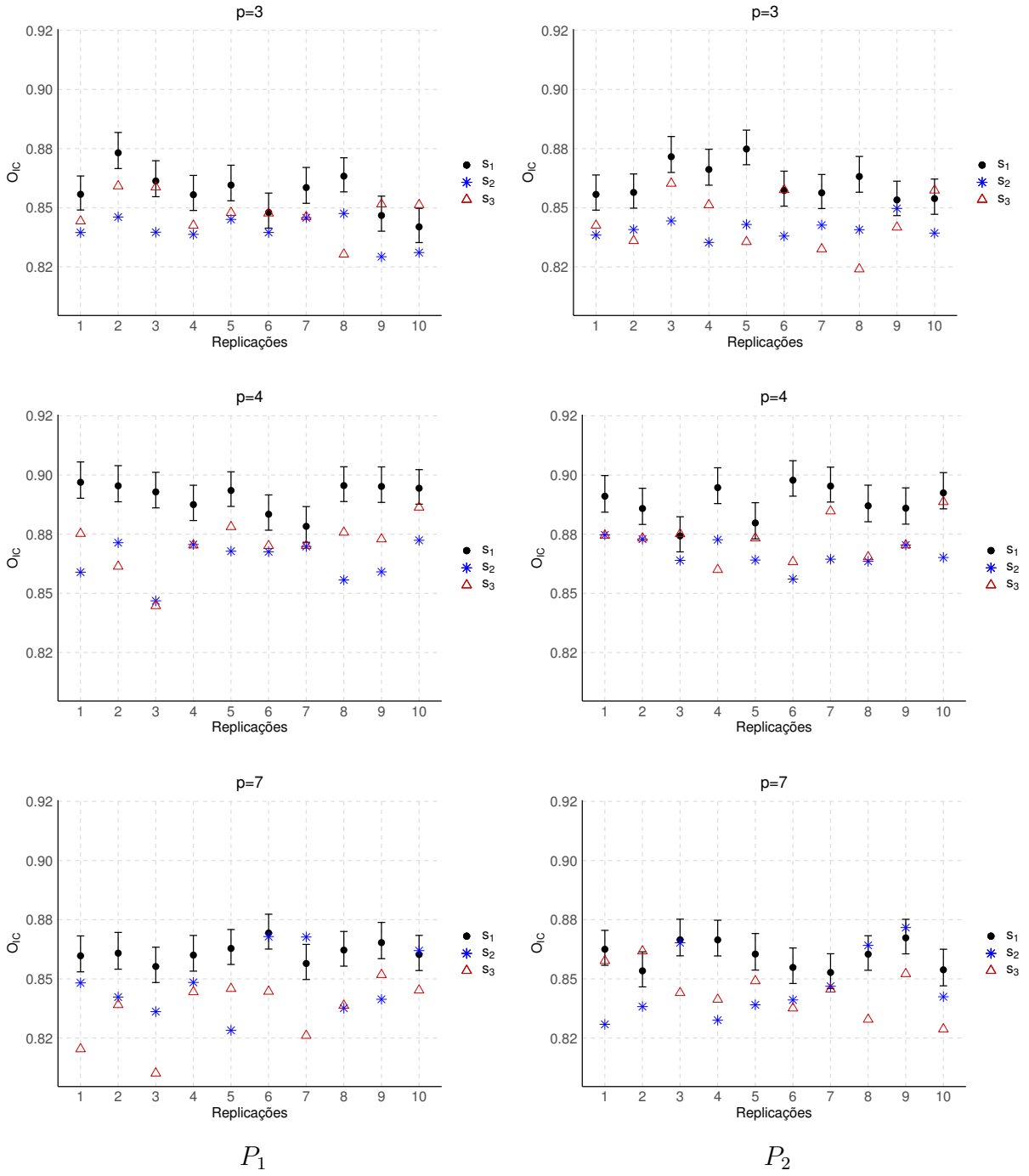
Tabela 4.3: Estatísticas da distribuição *a posteriori* da rede. Número de replicações nas quais a rede verdadeira é a mais provável - média (sob essas replicações) da prob. *a posteriori* da rede verdadeira / mesma média (sob as demais replicações).

Qtde Var.	**n=500**	**n=1000**	**n=2000**	**n=5000**
3 (P_1)	9 - .93 / .28	9 - .93 / .23	-	9 - .94 / .15
3 (P_2)	10 - .87 /	10 - .88 /	-	9 - .85 / .42
4 (P_1)	-	8 - .89 / .19	-	8 - .86 / .17
4 (P_2)	-	7 - .82 / .16	-	8 - .86 / .20
7 (P_1)	-	-	5 - .69 / .23	7 - .70 / .22
7 (P_2)	-	-	6 - .71 / .30	7 - .72 / .15

Observamos que, conforme esperado, em quase todos os casos, a rede verdadeira foi a mais provável *a posteriori*, indicando que o algoritmo de Goudie & Mukherjee (2016) é eficiente para aproximar a distribuição *a posteriori* da rede. Além disso, notamos resultados similares ao analisar *a priori* uniforme P_2 em comparação com *a priori* proposta P_1. Maiores diferenças são esperadas para conjuntos de dados menores e modelos com mais variáveis (mais redes).

4.1.3 Análise dos dados sintéticos

A Figura 4.2 apresenta os resultados para a medida O_{IC} do parâmetro $\theta = P(X_2 = 1|X_1 = 0)$ com tamanho de amostra $n = 5000$.



P_1 P_2

Figura 4.2: Gráfico da medida O_{IC} para $\theta = P(X_2 = 1|X_1 = 0)$ com $n = 5000$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.

De forma geral, o método proposto apresenta maior similaridade com os resultados obtidos com os dados originais por apresentar valores mais elevados para a % de sobre-

posição. Na maioria das replicações, os valores de O_{IC} obtidos pelos outros dois métodos estão fora (abaixo) do intervalo de credibilidade obtido pelo método proposto.

Além de apresentar um *overlap* maior com o IC obtido com os dados originais, o método S_1 possui a vantagem adicional de incorporar a incerteza do processo em sua análise. Essa característica confere-lhe uma escolha mais robusta e informativa na interpretação dos resultados.

Os resultados são similares entre as duas distribuições *a priori* utilizadas para a rede, para ambos os métodos S_1 e S_2.

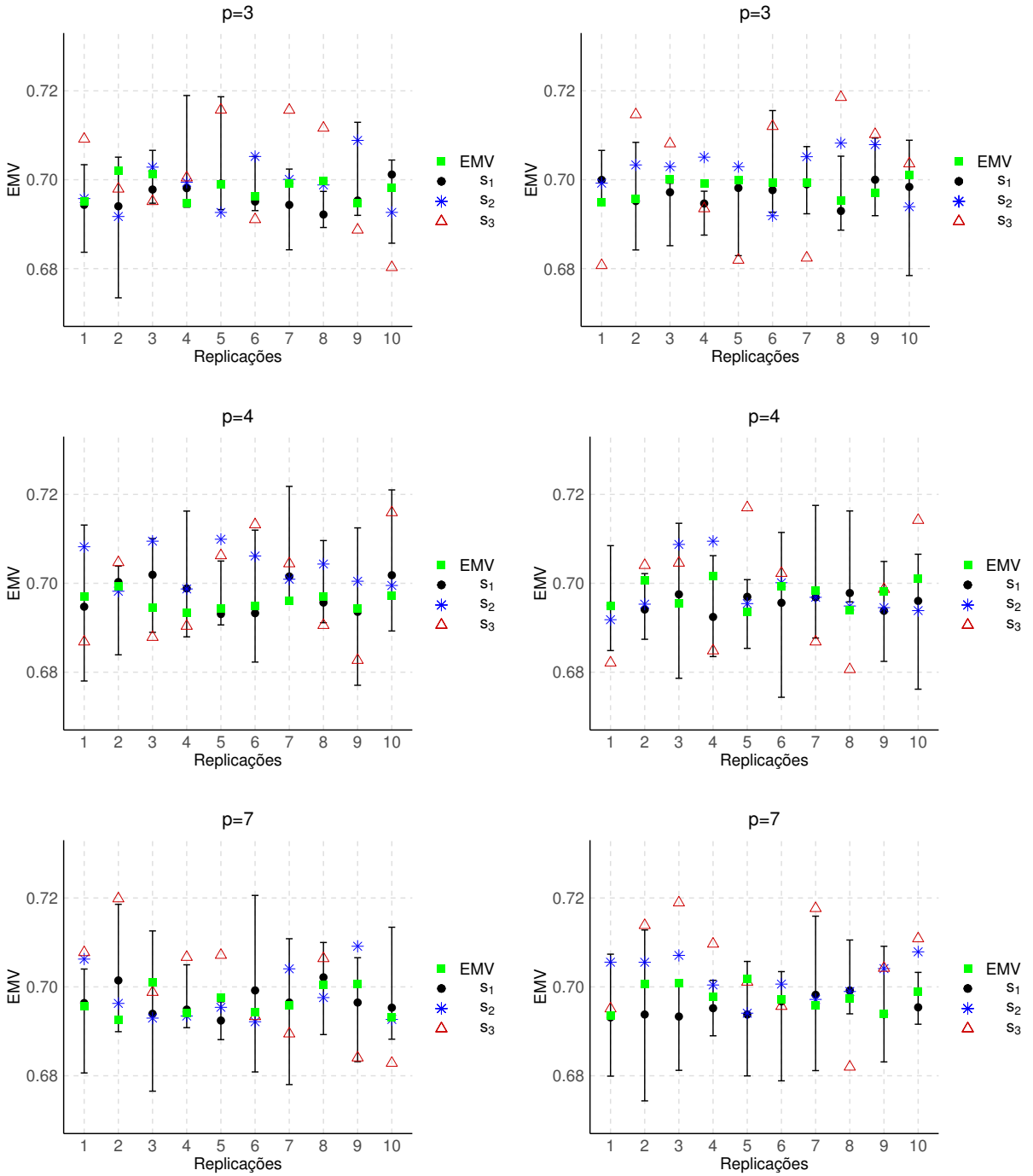
Resultados para o tamanho de amostra 1000 são apresentados no Apêndice A. A única diferença relevante com relação aos resultados para $n = 5000$ é que os valores de O_{IC} são menores para $n = 1000$, refletindo o fato de uma amostra menor conter menos informação.

A Figura 4.3 apresenta os resultados para o EMV do parâmetro $\theta = P(X_2 = 1|X_1 = 0)$, para $n = 5000$. As diferenças entre os resultados para as duas distribuições *a priori* são pequenas.

De forma geral, o método S_3 apresenta uma performance pior que os outros dois métodos. No caso do proposto método S_1, o EMV verdadeiro cai dentro do intervalo de credibilidade em quase todos os casos.

As Figuras 4.4 e 4.5 apresentam os resultados para o valor-p do teste qui-quadrado de independência entre X_1 e X_2 e mostram uma performance muito superior do método proposto S_1. A média da distribuição preditiva desta estatística está muito mais próxima do valor verdadeiro e sempre concordando com a conclusão do teste original, ao contrário dos outro dois métodos.

Uma coleção de resultados para outros parâmetros θ e tamanhos de amostra são apresentados no Apêndice A. Os resultados seguem o mesmo padrão dos apresentados nesta seção.



P_1 P_2

Figura 4.3: Gráfico do EMV de $\theta = P(X_2 = 1|X_1 = 0)$ com $n = 5000$. EMV verdadeiro em verde. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.

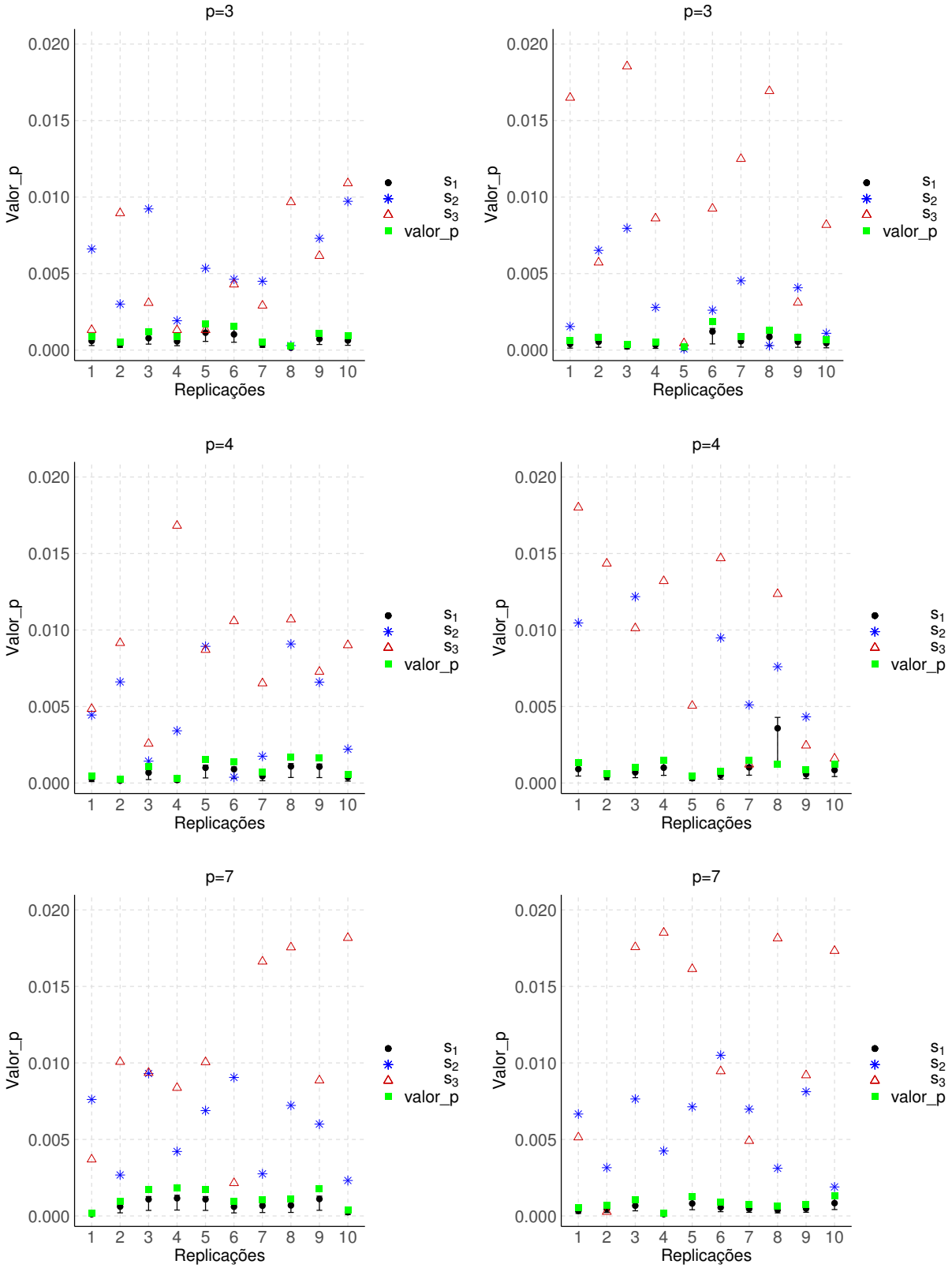
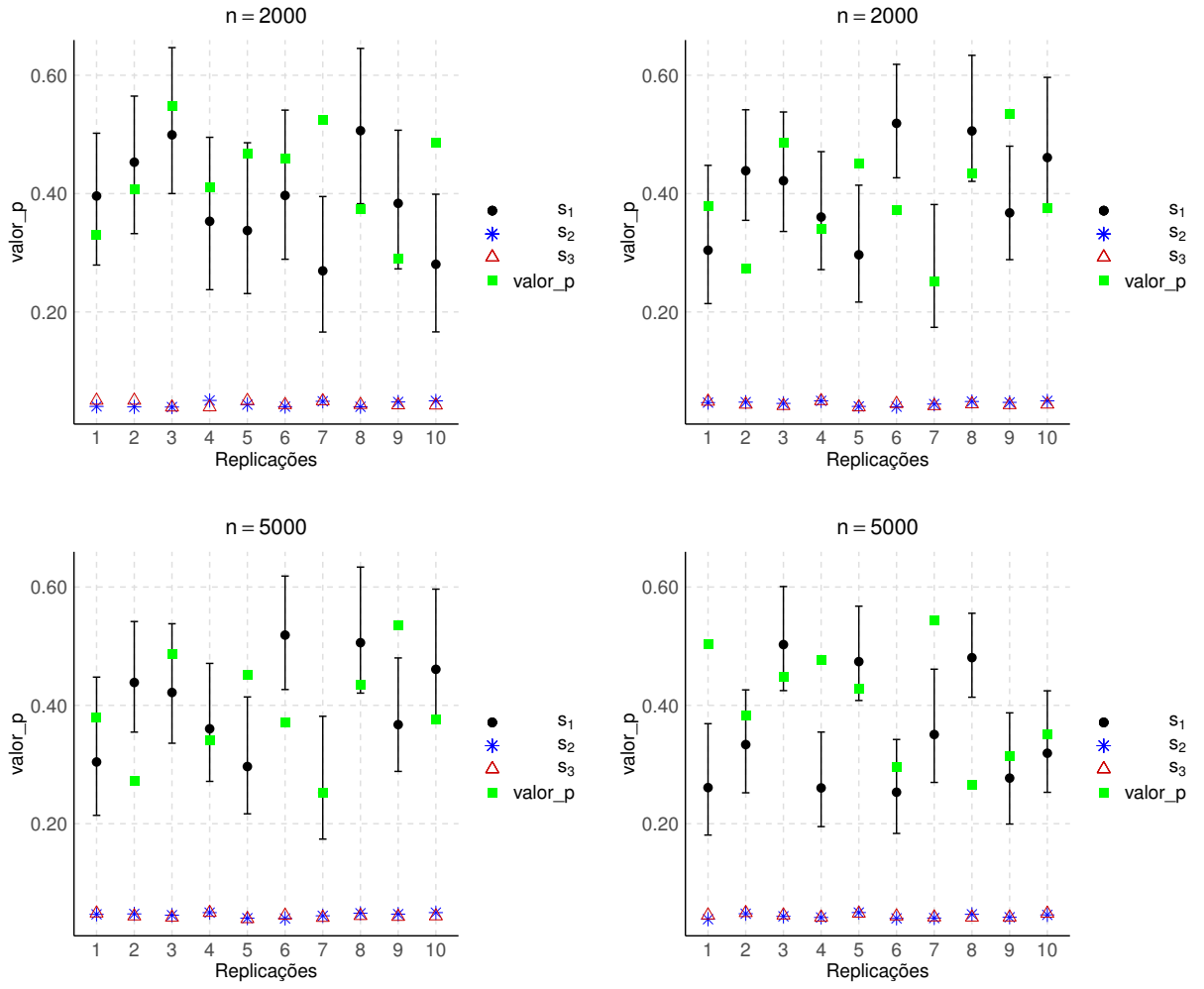


Figura 4.4: Gráfico do valor-p do teste de independência qui-quadrado para X_2 e X_1 com $n = 5000$. Valor-p verdadeiro em verde. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.



P_1 P_2

Figura 4.5: Gráfico do valor-p do teste de independência qui-quadrado para X_1 e X_5 com $p = 7$. Valor-p verdadeiro em verde. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.

4.2 Dados Reais

A aplicação da metodologia de geração de dados sintéticos em bancos de dados confidenciais apresenta desafios significativos. Isso decorre da natureza sensível desses bancos, os quais são protegidos por medidas de sigilo. Portanto, o objetivo é realizar uma análise em um banco de dados real disponível para estudo, mas tratá-lo com a devida confidencialidade. Neste contexto, vamos aplicar a metodologia proposta a uma base de dados real.

A Pesquisa Nacional por Amostra de Domicílios (PNAD) é uma pesquisa domiciliar de âmbito nacional realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A pesquisa tem o objetivo de coletar informações sobre as características socioeconômicas da população brasileira, incluindo informações sobre trabalho, educação, renda, saúde e habitação. Os dados dessa pesquisa podem ser acessados através do pacote *PNADcIBGE* disponível no *software* R. Basta selecionar o ano desejado juntamente com as variáveis necessárias (Braga & Assuncao 2018).

Para esta análise consideramos um banco de dados da PNAD de 2023 com informações sobre 5651 residências no Brasil. Esse banco de dados possui várias variáveis com informações de residentes brasileiros em relação a vários fatores socioeconômicos. Consideramos 5 variáveis selecionadas dentre as 237 disponíveis. Essas variáveis foram escolhidas para uma exemplificação de informações que podem ser confidenciais. A Tabela 4.4 apresenta as variáveis selecionadas bem como suas categorias.

Tabela 4.4: Variáveis PNAD 2023.

Variável	Categorias
X_1 : Tipo de moradia	Rural ou Urbana
X_2 : Gênero	Feminino ou Masculino
X_3 : Idade	>50 ou <=50
X_4 : Cor	Branca ou Parda/Preta
X_5 : Renda	> Salário Mínimo 2023 ou <= Salário Mínimo 2023

Obs: Salário Mínimo 2023 = $R\$1.320,00$.

Tabela 4.5: Proporções observadas no conjunto de dados da PNAD 2023.

Variável	%
X_1 = Urbana	0.79
X_2 = Masculino	0.59
X_3 = <= 50	0.75
X_4 = Parda/Preta	0.74
X_5 = <= Salário Mínimo	0.34

Seguindo as diretrizes obtidas na Seção 4.1.1, adotamos $\gamma = 26$ para a priori P_1. A Tabela 4.6 apresenta a probabilidade das duas redes mais prováveis *a posteriori*.

Tabela 4.6: Probabilidades *a posteriori* das duas redes mais prováveis para os dados da PNAD.

Priori	**Rede Estimada**	**Prob.**
P_1	$X_1, X_3 \mid X_1, X_4 \mid X_3, (X_5 \mid X_1, X_4), (X_2 \mid X_1, X_5)$	0,256
P_2		0,131
P_1	$X_5, X_4 \mid X_5, X_3 \mid X_4, (X_1 \mid X_3, X_5), (X_2 \mid X_1, X_5)$	0,057
P_2		0,007

Conforme esperado, as probabilidades *a posteriori* das duas redes mais prováveis são maiores para a priori informativa, indicando uma maior concentração *a posteriori*.

O fato de a rede mais provável *a posteriori* não ter alta probabilidade provoca a discussão de pontos importantes relacionados ao problema estudado nesta tese. Primeiramente, isto indica um razoável nível de incerteza sobre a rede e que um número grande de diferentes redes é necessário para englobar a maior parte da massa de probabilidade da distribuição *a posteriori*. Portanto, um método que utilize todas essas redes, propriamente ponderadas por suas respectivas probabilidades, será mais eficiente e robusto para quantificar a incerteza do sistema. Este é exatamente o caso do método S_1, proposto nesta tese. O segundo ponto importante é que as particulares relações de independência condicional da rede mais provável não devem ser muito valorizadas/interpretadas dada a baixa probabilidade.

As Figuras 4.6 e 4.7 apresentam os resultados para as mesmas três estatísticas consideradas nos estudos de dados simulados para os mesmos três métodos. Os resultados são consideravelmente mais próximos do que se observa por meio da aplicação de métodos usuais aos dados observados para o método proposto (S_1) nas três estatísticas. Além disso, para o método S_1, a priori informativa (P_1) teve uma performance muito melhor que a priori uniforme (P_2) no caso do EMV.

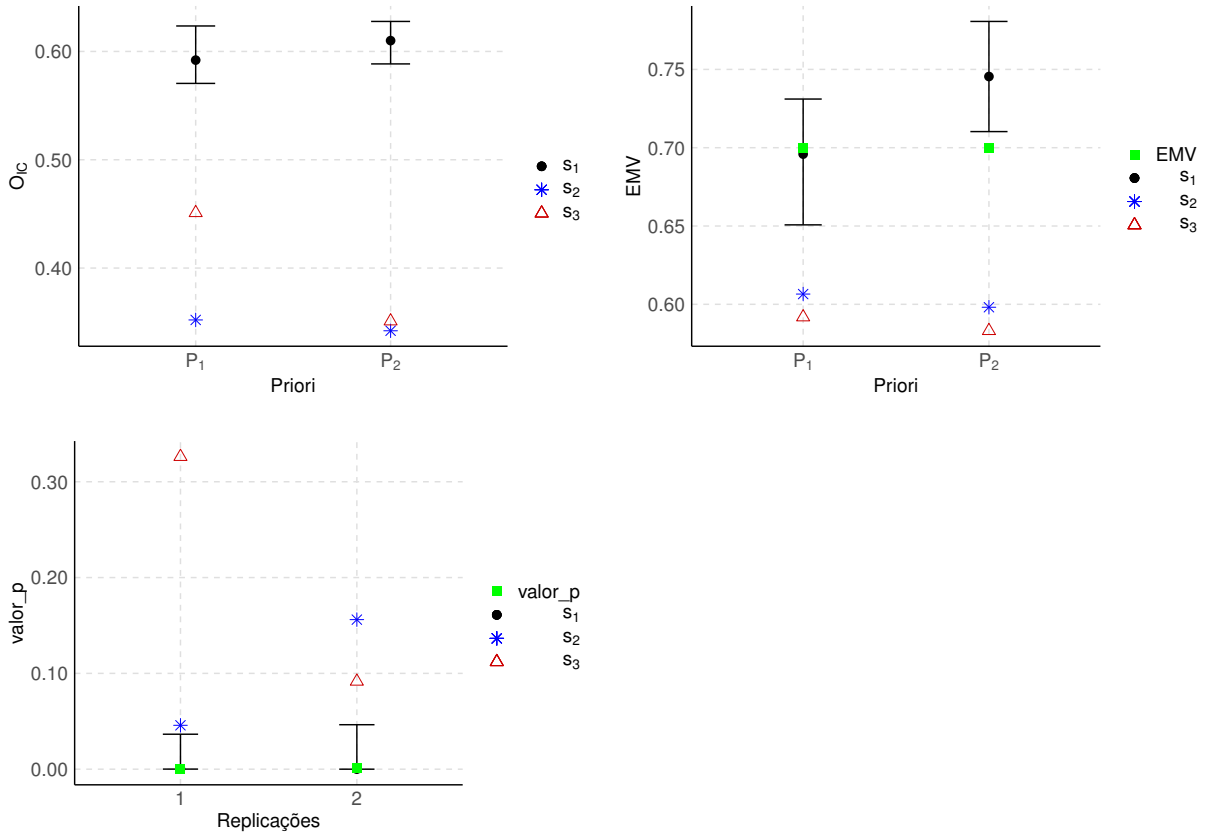


Figura 4.6: Gráfico das estatísticas O_{IC}, EMV para $\theta = P(X_3 = 1|X_1 = 1)$ e valor-p do teste de independência entre X_1 e X_3. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva.

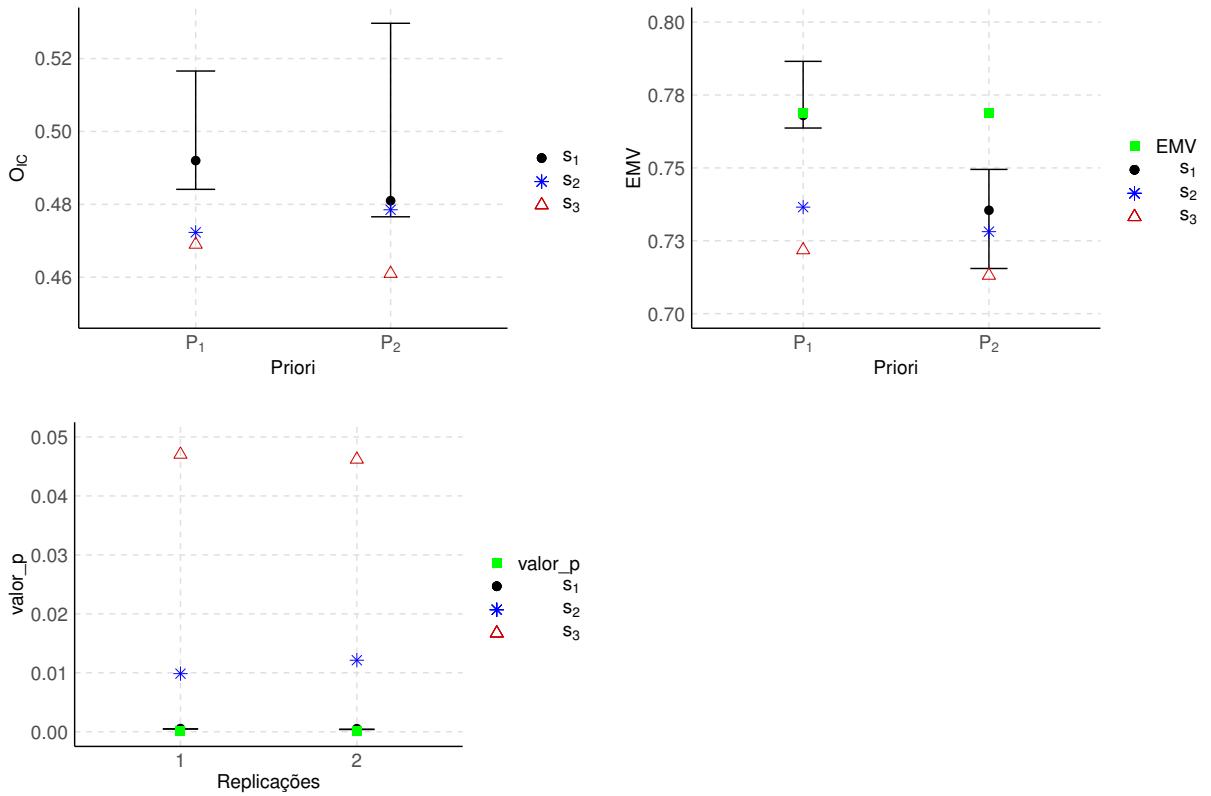


Figura 4.7: Gráfico das estatísticas O_{IC}, EMV para $\theta = P(X_4 = 1|X_3 = 1)$ e valor-p do teste de independência entre X_3 e X_4. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva.

Capítulo 5

Considerações Finais

O objetivo principal desta tese foi propor uma análise mais completa para a geração de dados sintéticos via redes Bayesianas sob o enfoque do paradigma Bayesiano. A metodologia proposta envolve desde a estimação da rede Bayesiana para ajustar os dados observados até a geração e análise de dados sintéticos. Uma nova classe de *prioris* penalizadoras modulares foi proposta com o objetivo de fornecer mais flexibilidade às análises.

Como discutido, a rede Bayesiana é uma ferramenta importante na estimação de distribuições conjuntas. No contexto de dados sintéticos, essa metodologia traz grandes vantagens em relação a outras abordagens, como sua flexibilidade a vários tipos de dados e fácil interpretação dos resultados. A metodologia proposta considera uma abordagem completamente Bayesiana para a geração e análise dos dados sintéticos, sendo assim mais robusta e eficiente que as abordagens mais comuns na literatura existente. Além disso, o maior custo computacional da abordagem proposta não compromete sua utilização em exemplos nos quais as outras abordagens são computacionalmente inviáveis.

O MCMC proposto por Goudie & Mukherjee (2016) se mostrou eficiente para aproximar a distribuição *a posteriori* da rede. Esse algoritmo é então utilizado para propor um método de gerar dados sintéticos e obter a distribuição preditiva de suas estatísticas de interesse. Esta abordagem permite a produção de diferentes produtos/outputs para serem entregues ao usuário final, dependendo de seus objetivos e expertise para lidar com dados.

A comparação entre a metodologia proposta e outras duas muito comuns na literatura foi feita através da análise de três estatísticas dos dados sintéticos, sendo o objetivo principal, aproximar o melhor possível a mesma estatística obtida com os dados originais. Além de ter apresentado uma melhor performance, o método proposto permite a quantificação da incerteza envolvida no fato de se utilizar os dados sintéticos.

Foi também apresentada uma análise de sensibilidade da distribuição a *a priori* da rede, comparando uma *priori* informativa (penalizadora) com uma uniforme. Os resultados mostraram que a primeira tem um desempenho melhor ou igual à segunda, dependendo do cenário e da estatística considerada.

A análise e geração de dados sintéticos foi realizada com base em um banco de dados real, especificamente a PNAD de 2023. Este banco de dados contém informações

sobre residentes brasileiros, incluindo gênero, renda, cor, idade e tipo de moradia. Embora o banco de dados não seja classificado como confidencial, optamos por tratá-lo como tal devido à complexidade envolvida na obtenção de dados sigilosos e à possibilidade de sua natureza sensível. O estudo demonstrou claramente as vantagens da metodologia proposta, especialmente em situações de maior incerteza *a posteriori* em relação à rede geradora dos dados.

Diante desse cenário, destacamos novamente a relevância de se considerar e quantificar a incerteza associada à inferência feita com dados sintéticos. Esse procedimento se revela crucial para uma abordagem mais robusta e confiável.

Como trabalhos futuros à tese, temos o interesse de explorar a inclusão de outros tipos de variáveis, desde multinomiais até variáveis contínuas, visando aprimorar a abrangência e aplicabilidade dos modelos univariados. Outra objetivo futuro é ampliar a análise, considerando bancos de dados que possuam variáveis de diferentes tipos, proporcionando uma visão mais completa e realista das relações entre diferentes tipos de dados.

Visando melhorar as análises, queremos considerar diferentes configurações de bancos de dados em relação ao tamanho da amostra e ao número de variáveis, explorando cenários mais desafiadores, pois apenas um número limitado de cenários foi explorado no estudo e restringindo as diretrizes em aplicações reais mais gerais.

Como o trabalho considerou apenas cenários com dados completos, pretende-se inserir no modelo a estimação com dados faltantes. Com isso teremos um algoritmo mais abrangente para a maioria das configurações de bancos de dados reais.

Por fim, pretendemos criar um pacote do R que seja o mais amigável possível para utilizar a metodologia proposta.

Referências

Bishop, C. M. (2006), 'Pattern recognition and machine learning', *Springer google schola* 2, 1122–1128.

Braga, D. & Assuncao, G. (2018), 'Pnadcibge: downloading, reading and analysing pnadc microdata', *R package version 0.4* 3.

Caiola, G. & Reiter, J. P. (2010), 'Random forests for generating partially synthetic, categorical data.', *Trans. Data Priv.* 3(1), 27–42.

Cano, R., Sordo, C. & Gutiérrez, J. M. (2004), Applications of Bayesian networks in meteorology, *in* 'Advances in Bayesian networks', Springer, pp. 309–328.

Chen, Y.-C., Wheeler, T. A. & Kochenderfer, M. J. (2017), 'Learning discrete Bayesian networks from continuous data', *Journal of Artificial Intelligence Research* 59.

Cooper, G. F. & Herskovits, E. (1992), 'A Bayesian method for the induction of probabilistic networks from data', *Machine learning* 9, 309–347.

Cox, D. R. (2018), *Analysis of binary data*, Routledge.

Cox, L. H. (1980), 'Suppression methodology and statistical disclosure control', *Journal of the American Statistical Association* 75(370), 377–385.

Cui, W. (2019), 'Visual analytics: A comprehensive overview', *IEEE access* 7, 81555–81573.

Deeva, I., Andriushchenko, P. D., Kalyuzhnaya, A. V. & Boukhanovsky, A. V. (2020), Bayesian Networks-based personal data synthesis, *in* 'Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good', pp. 6–11.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O. & Ponti, A. (2004), 'Bayesian networks for imputation', *Journal of the Royal Statistical Society Series A: Statistics in Society* 167(2), 309–322.

Drechsler, J. (2011), *Synthetic datasets for statistical disclosure control: theory and implementation*, Vol. 201, Springer Science & Business Media.

Drechsler, J. & Reiter, J. P. (2010), 'Sampling with synthesis: A new approach for releasing public use census microdata', *Journal of the American Statistical Association* 105(492), 1347–1357.

Drechsler, J. & Reiter, J. P. (2011*a*), 'An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets', *Computational Statistics & Data Analysis* 55(12), 3232–3243.

Drechsler, J. & Reiter, J. P. (2011*b*), 'An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets', *Computational Statistics & Data Analysis* 55(12), 3232–3243.

Duncan, G. T., Fienberg, S. E. & Krishnan, R. (2001), 'Confidentiality, disclosure and data access: Theory and practical applications for statistical agencies'.

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006), Calibrating noise to sensitivity in private data analysis, *in* 'Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3', Springer, pp. 265–284.

Eggeling, R., Viinikka, J., Vuoksenmaa, A. & Koivisto, M. (2019), On structure priors for learning Bayesian networks, *in* 'The 22nd International Conference on Artificial Intelligence and Statistics', PMLR, pp. 1687–1695.

Ellis, B. & Wong, W. H. (2008), 'Learning causal Bayesian network structures from experimental data', *Journal of the American Statistical Association* 103(482), 778–789.

Friedman, N., Geiger, D. & Goldszmidt, M. (n.d.), 'Bayesian network classifiers'.

Friedman, N. & Koller, D. (2003), 'Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks', *Machine learning* 50(1), 95–125.

Goudie, R. J. & Mukherjee, S. (2016), 'A gibbs sampler for learning dags', *Journal of Machine Learning Research* .

Grzegorczyk, M. (2010), An introduction to Gaussian Bayesian networks, *in* 'Systems Biology in Drug Discovery and Development', Springer, pp. 121–147.

Grzegorczyk, M. & Husmeier, D. (2008), 'Improving the structure mcmc sampler for bayesian networks by introducing a new edge reversal move', *Machine Learning* 71(2-3), 265–305.

Heckerman, D. (2008), 'A tutorial on learning with Bayesian networks', *Innovations in Bayesian networks* pp. 33–82.

Heckerman, D., Geiger, D. & Chickering, D. M. (1995), 'Learning bayesian networks: The combination of knowledge and statistical data', *Machine learning* 20, 197–243.

Heckerman, D., Mamdani, A. & Wellman, M. P. (1995), 'Real-world applications of Bayesian networks', *Communications of the ACM* 38(3), 24–26.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., Miyachi, S. & Doya, K. (1999), 'Parallel neural networks for learning sequential procedures', *Trends in neurosciences* 22(10), 464–471.

Hruschka, E. R., Hruschka, E. R. & Ebecken, N. F. (2004), Feature selection by Bayesian networks, *in* 'Advances in Artificial Intelligence: 17th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2004, London, Ontario, Canada, May 17-19, 2004. Proceedings 17', Springer, pp. 370–379.

Hu, Y., Wu, F., Li, Q., Long, Y., Garrido, G. M., Ge, C., Ding, B., Forsyth, D., Li, B. & Song, D. (2023), 'Sok: Privacy-preserving data synthesis', *arXiv preprint arXiv:2307.02106* .

Karr, A. F. & Reiter, J. (2014*a*), 'Using statistics to protect privacy', *Privacy, Big Data, and the Public Good: Frameworks for Engagement* 1, 276–295.

Karr, A. F. & Reiter, J. (2014*b*), 'Using statistics to protect privacy', *Privacy, Big Data, and the Public Good: Frameworks for Engagement* 1, 276–295.

Kennickell, A. & Lane, J. (2006), Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances, *in* 'International conference on privacy in statistical databases', Springer, pp. 291–303.

Kirkpatrick, S., Gelatt Jr, C. D. & Vecchi, M. P. (1983), 'Optimization by simulated annealing', *science* 220(4598), 671–680.

Kjaerulff, U. B. & Madsen, A. L. (2008), 'Bayesian networks and influence diagrams', *Springer Science+ Business Media* 200, 114.

Koivisto, M. & Sood, K. (2004), 'Exact Bayesian structure discovery in Bayesian networks', *The Journal of Machine Learning Research* 5, 549–573.

Koller, D. & Friedman, N. (2009), *Probabilistic graphical models: principles and techniques*, MIT press.

Korb, K. B. & Nicholson, A. E. (2010), *Bayesian artificial intelligence*, CRC press.

Little, R. J. & Rubin, D. B. (2019), *Statistical analysis with missing data*, Vol. 793, John Wiley & Sons.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* 21(6), 1087–1092.

Mihaljević, B., Bielza, C. & Larrañaga, P. (2021), 'Bayesian networks for interpretable machine learning and optimization', *Neurocomputing* 456, 648–665.

Neal, R. M. (2012), *Bayesian learning for neural networks*, Vol. 118, Springer Science & Business Media.

Niinimäki, T., Parviainen, P. & Koivisto, M. (2016), 'Structure discovery in Bayesian networks by sampling partial orders', *The Journal of Machine Learning Research* 17(1), 2002–2048.

Nowok, B., Raab, G. M., Dibben, C., Snoke, J., van Lissa, C. & Nowok, M. B. (2022), 'Package 'synthpop''.

Pensar, J., Nyman, H., Lintusaari, J. & Corander, J. (2016), 'The role of local partial independence in learning of Bayesian networks', *International journal of approximate reasoning* 69, 91–105.

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, Vienna, Austria.
URL: *https://www.R-project.org*

Raghunathan, T. E. (2021), 'Synthetic data', *Annual review of statistics and its application* 8, 129–140.

Raghunathan, T. E., Reiter, J. P. & Rubin, D. B. (2003), 'Multiple imputation for statistical disclosure limitation', *Journal of official statistics* 19(1), 1.

Reiter, J. P. (2003), 'Inference for partially synthetic, public use microdata sets', *Survey Methodology* 29(2), 181–188.

Reiter, J. P. (2005), 'Using cart to generate partially synthetic public use microdata', *Journal of Official Statistics* 21(3), 441.

Reiter, J. P. (2023), 'Synthetic data: A look back and a look forward.', *Trans. Data Priv.* 16(1), 15–24.

Reiter, J. P. & Raghunathan, T. E. (2007), 'The multiple adaptations of multiple imputation', *Journal of the American Statistical Association* 102(480), 1462–1471.

Reiter, J. P., Wang, Q. & Zhang, B. (2014), 'Bayesian estimation of disclosure risks for multiply imputed, synthetic data', *Journal of Privacy and Confidentiality* 6(1).

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* 70(1), 41–55.

Rubin, D. B. (1993), 'Discussion statistical disclosure limitation', *Journal of official Statistics* 9(2), 461.

Russell, S. J. & Norvig, P. (2010), *Artificial intelligence a modern approach*, London.

Sun, L. & Erath, A. (2015*a*), 'A bayesian network approach for population synthesis', *Transportation Research Part C: Emerging Technologies* 61, 49–62.

Sun, L. & Erath, A. (2015*b*), 'A bayesian network approach for population synthesis', *Transportation Research Part C: Emerging Technologies* 61, 49–62.

Surendra, H. & Mohan, H. (2017), 'A review of synthetic data generation methods for privacy preserving data publishing', *International Journal of Scientific & Technology Research* 6(3), 95–101.

Woo, M.-J., Reiter, J. P., Oganian, A. & Karr, A. F. (2009), 'Global measures of data utility for microdata masked for disclosure limitation', *Journal of Privacy and Confidentiality* 1(1).

Yancey, W. E., Winkler, W. E. & Creecy, R. H. (2002), 'Disclosure risk assessment in perturbative microdata protection', pp. 135–152.

Young, J., Graham, P. & Penny, R. (2009), 'Using Bayesian networks to create synthetic data', *Journal of Official Statistics* 25(4), 549–567.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. & Xiao, X. (2017), 'Privbayes: Private data release via bayesian networks', *ACM Transactions on Database Systems (TODS)* 42(4), 1–41.

Apêndice

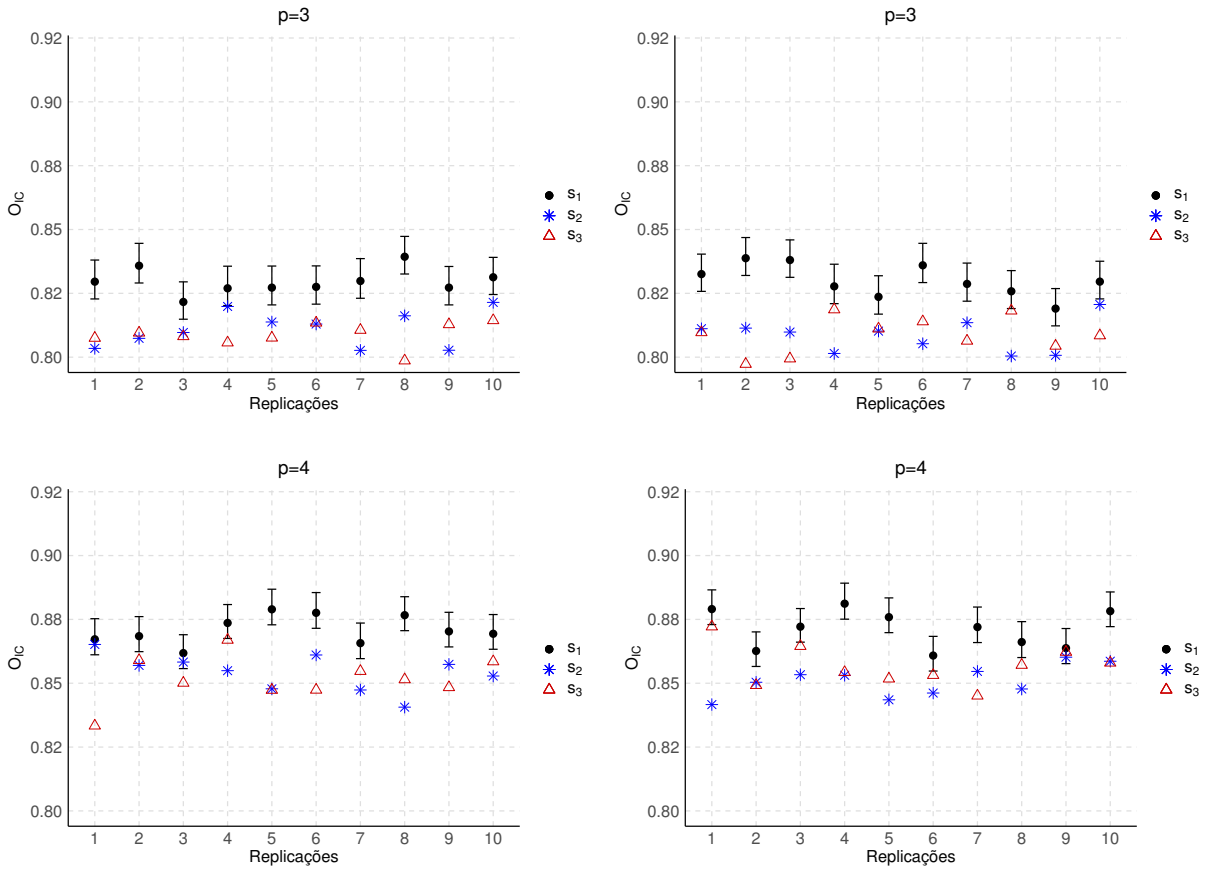
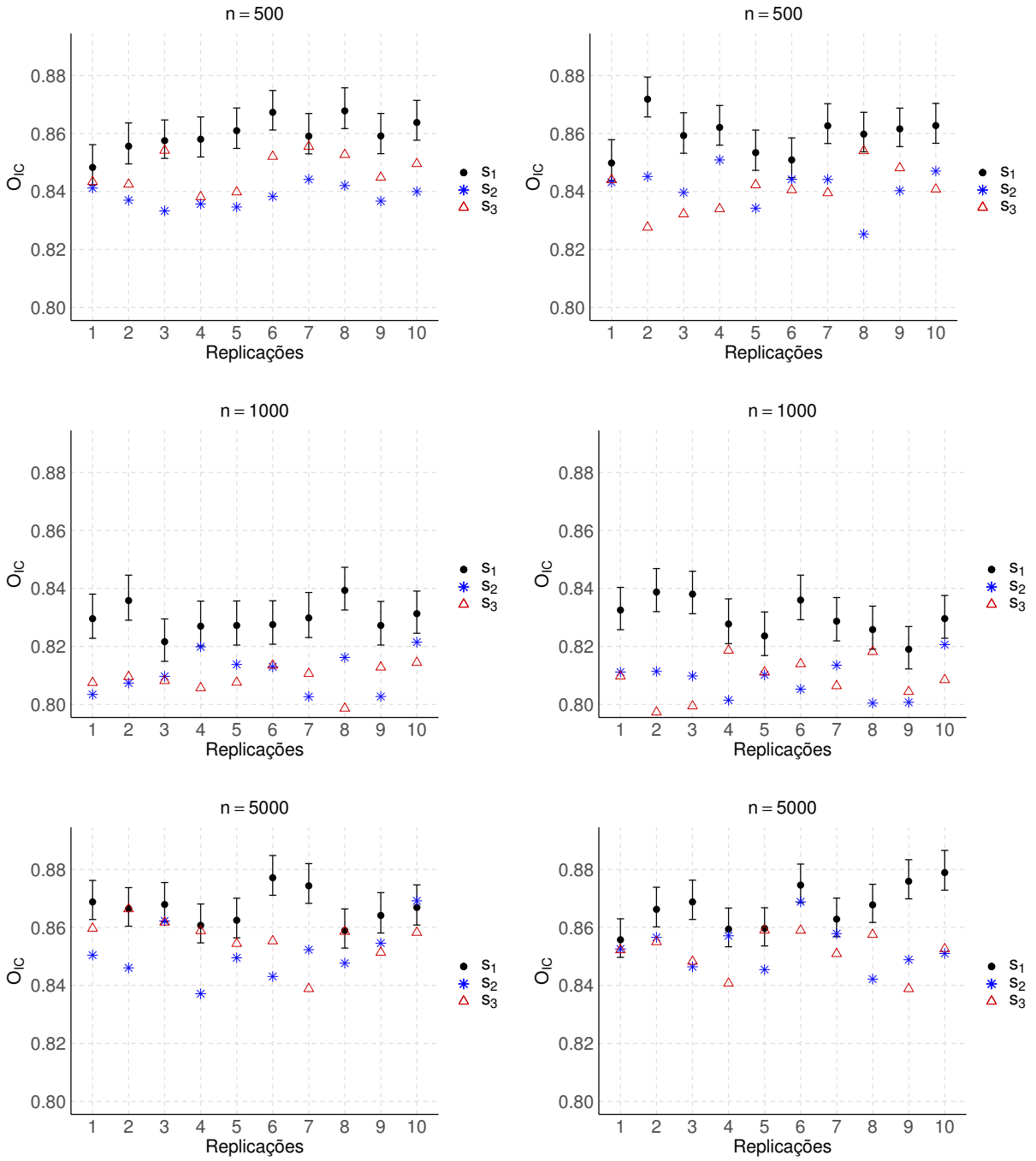


Figura A.1: Gráfico da medida O_{IC} para $\theta = P(X_2 = 1|X_1 = 0)$ com $n = 1000$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.



P_1 P_2

Figura A.2: Gráfico da medida O_{IC} para $\theta = P(X_3 = 1)$ e $p = 3$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.

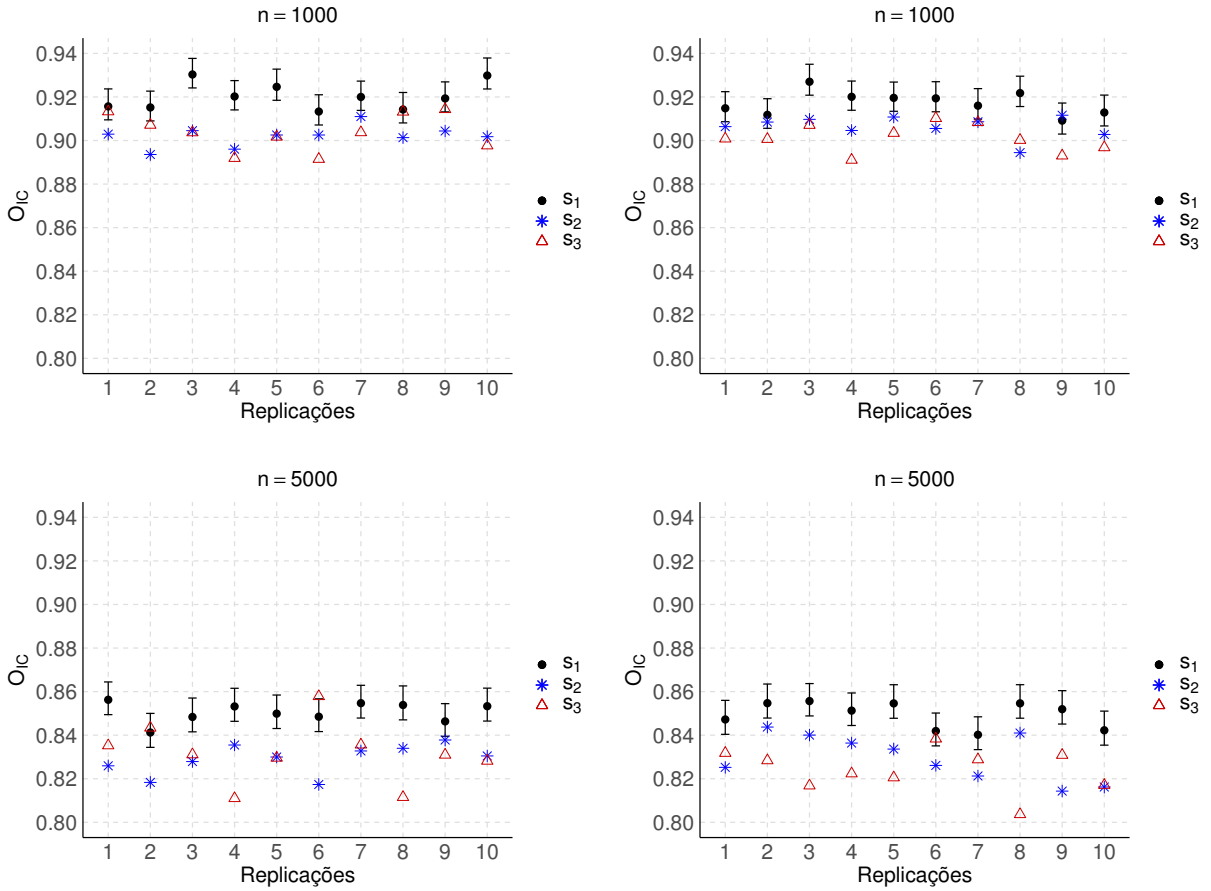
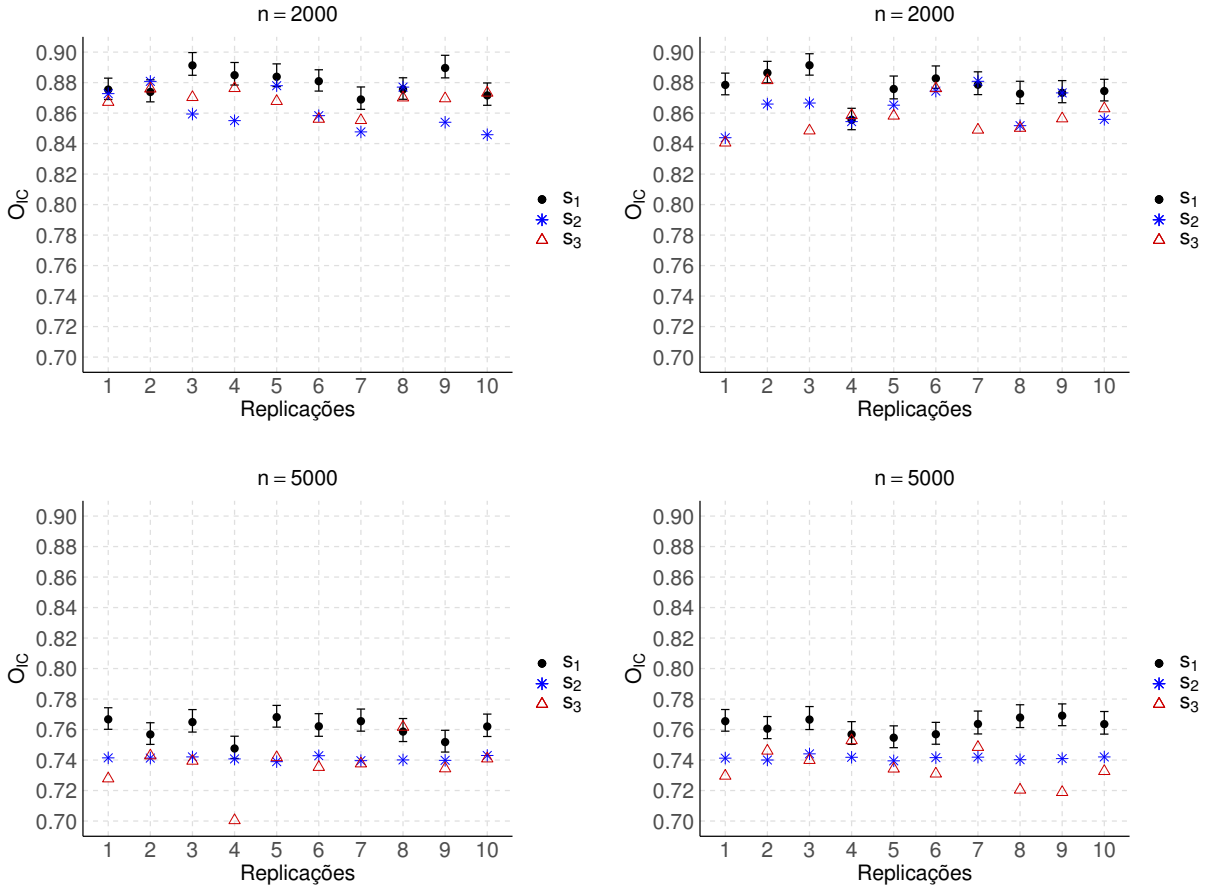
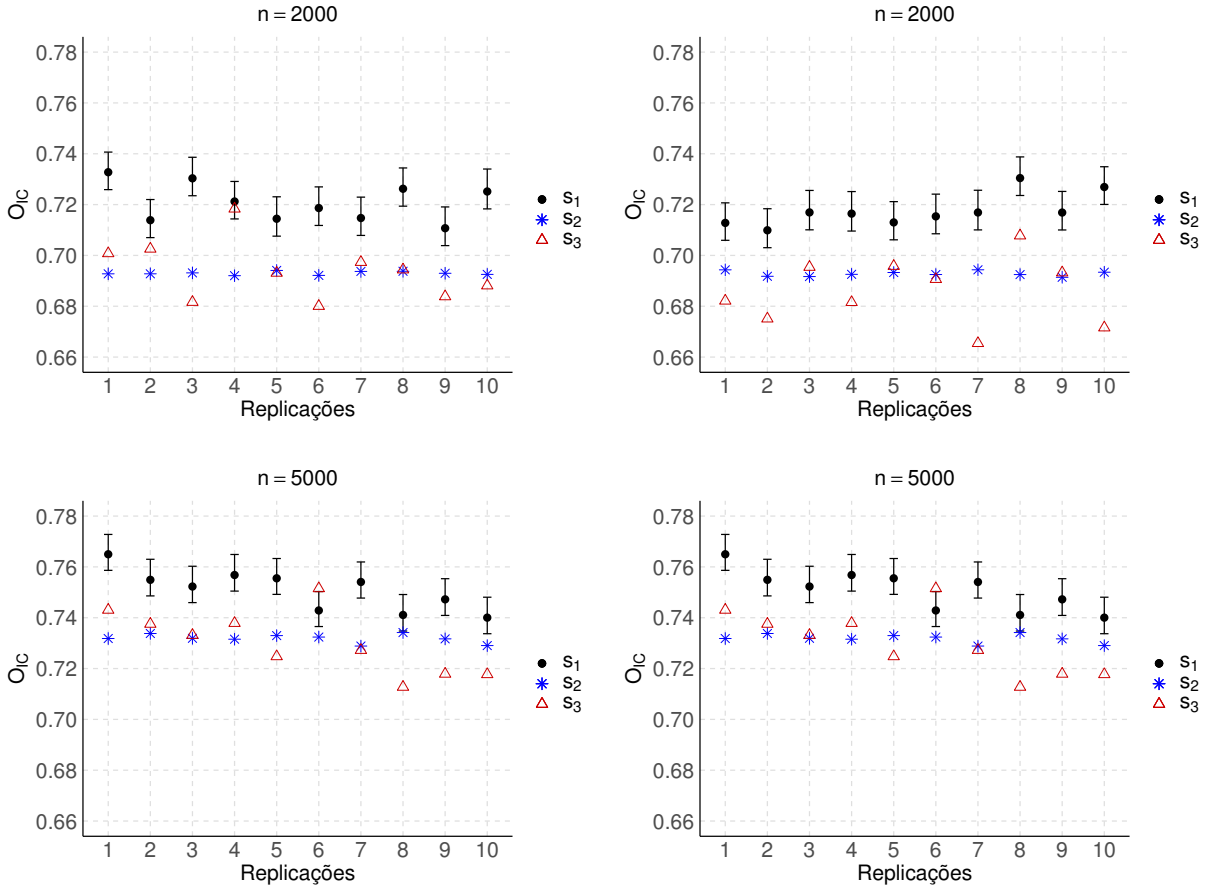


Figura A.3: Gráfico da medida O_{IC} para $\theta = P(X_3 = 1|X_4 = 1)$ e $p = 4$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.



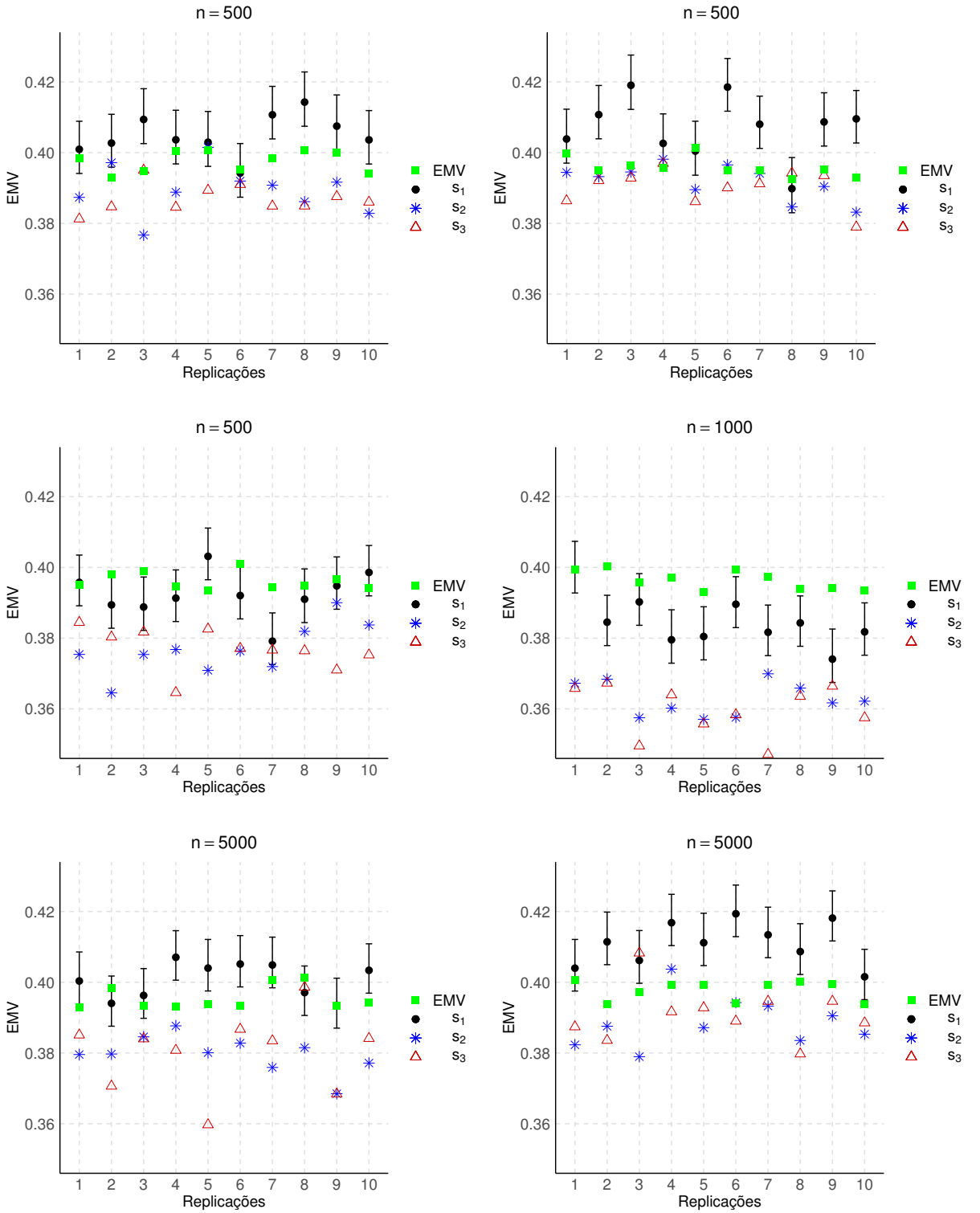
P_1 P_2

Figura A.4: Gráfico da medida O_{IC} para $\theta = P(X_4 = 1|X_3 = 1, X_5 = 1)$ e $p = 7$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.



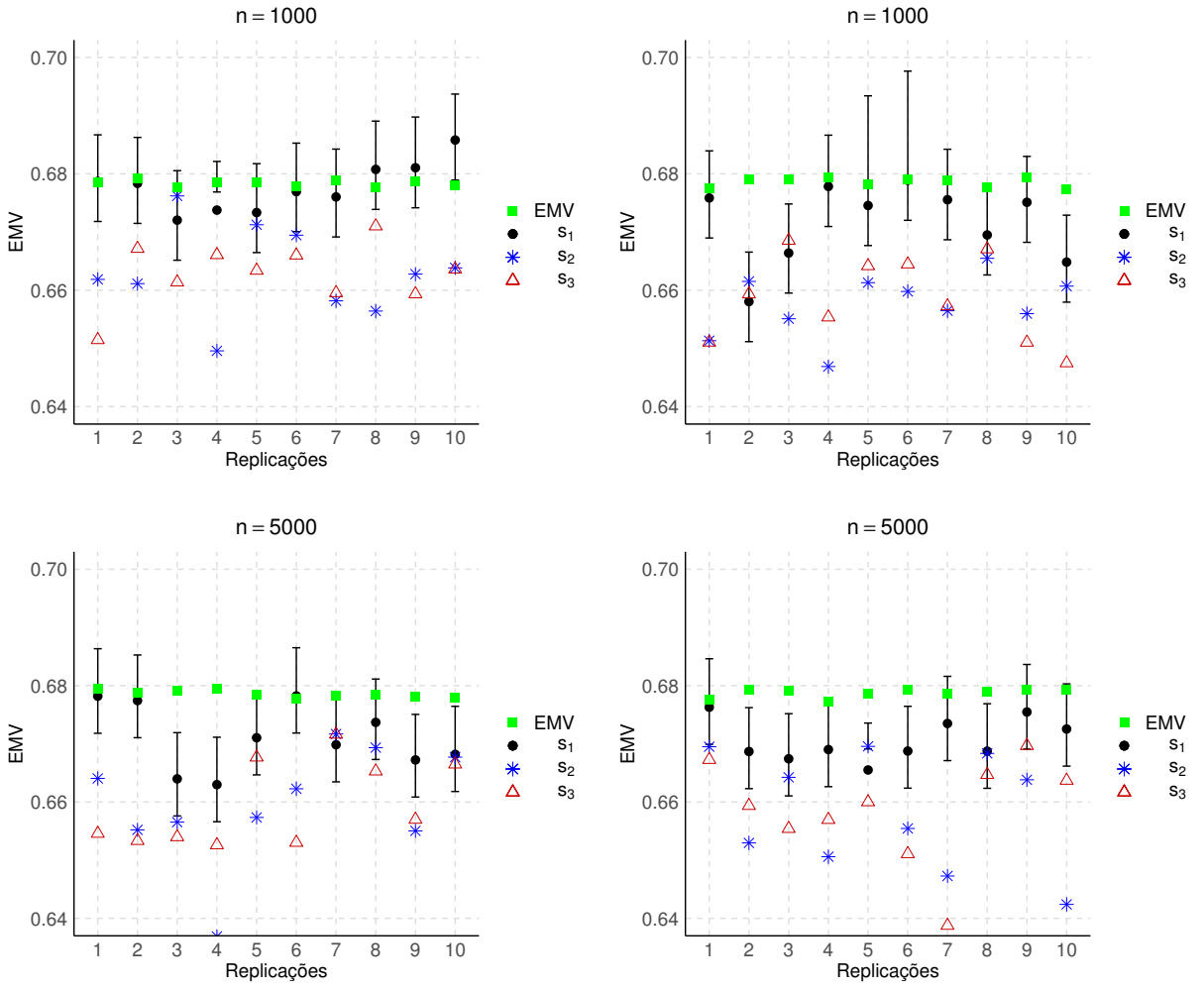
P_1 P_2

Figura A.5: Gráfico da medida O_{IC} para $\theta = P(X_5 = 1|X_6 = 0, X_7 = 1)$ e $p = 7$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. P_1 à esquerda e P_2 à direita.



P_1 P_2

Figura A.6: Gráfico do EMV para $\theta = P(X_1 = 1)$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. EMV verdadeiro em verde. Cenário: $p = 3$. P_1 à esquerda e P_2 à direita.



P_1 P_2

Figura A.7: Gráfico da medida EMV para $\theta = P(x_3|X_4 = 1)$ e $p = 4$. S_1: valor esperado (ponto preto) e $IC_{98\%}$ da distribuição preditiva. EMV verdadeiro em verde. P_1 à esquerda e P_2 à direita.