UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÁTICA

JUAN PABLO ARGOTE OSORIO

ALOCAÇÃO LATENTE DE DIRICHLET PARA MODELAGEM DE TÓPICOS EM DISSERTAÇÕES DE MESTRADO EM ESTATÍSTICA E ÁREAS CORRELATAS NO BRASIL

RIO DE JANEIRO

2025

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE MATEMÂTICA

JUAN PABLO ARGOTE OSORIO

ALOCAÇÃO LATENTE DE DIRICHLET PARA MODELAGEM DE TÓPICOS EM DISSERTAÇÕES DE MESTRADO EM ESTATÍSTICA E ÁREAS CORRELATAS NO BRASIL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Orientador: Prof. Dr. Carlos Tadeu Pagani Zanini

Rio de Janeiro

2025

CIP - Catalogação na Publicação

A693a

Argote Osorio, Juan Pablo
Alocação Latente de Dirichlet para Modelagem de Tópicos em Dissertações de Mestrado em Estatística e Áreas Correlatas no Brasil / Juan Pablo Argote Osorio. -- Rio de Janeiro, 2025.
80 f.

Orientador: Carlos Tadeu Pagani Zanini.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Programa de Pós-Graduação em Estatística, 2025.

1. Alocação Latente de Dirichlet. 2. Modelagem de Tópicos. 3. Dissertações de Mestrado. 4. Método de Monte Carlo via cadeias de Markov. 5. Inferência variacional. I. Zanini, Carlos Tadeu Pagani, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

JUAN PABLO ARGOTE OSORIO

ALOCAÇÃO LATENTE DE DIRICHLET PARA MODELAGEM DE TÓPICOS EM DISSERTAÇÕES DE MESTRADO EM ESTATÍSTICA E ÁREAS CORRELATAS NO BRASIL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Aprovada em: 24 de fevereiro de 2025

Banca Examinadora:

Documento assinado digitalmente
gov.br
CARLOS TADEU PAGANI ZANINI
Data: 22/04/2025 11:50:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Carlos Tadeu Pagani Zanini - Universidade Federal do Rio de Janeiro

Presidente

Documento assinado digitalmente
gov.br
HUGO TREMONTE DE CARVALHO
Data: 21/04/2025 11:49:39-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Hugo Tremonte de Carvalho - Universidade Federal do Rio de Janeiro

Avaliador interno

Documento assinado digitalmente
gov.br
DAIANE APARECIDA ZUANETTI
Data: 20/04/2025 14:41:09-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Daiane Aparecida Zuanetti - Universidade Federal de São Carlos

Avaliadora externa

Prof. Dr. João Batista de Morais Pereira - Universidade Federal do Rio de Janeiro

Avaliador interno (suplente)

Dedico esta dissertação de mestrado a Deus, aos meus pais e aos meus irmãos; por tudo.

Agradecimentos

Agradeço a Deus; pela minha família, pela minha vida, por estar sempre comigo, por tudo.

Agradeço aos meus pais, Marino e Alicia; pelo amor, pelo apoio, pela compreensão, pelos conselhos, por tudo.

Agradeço aos meus irmãos, Felipe e Andrés; pelo amor, pelo apoio, pela compreensão, pelos conselhos, por tudo.

Agradeço ao meu orientador, professor Carlos; pela pedagogia ao compartilhar seus conhecimentos, pelo apoio, pela compreensão, pela disposição.

Agradeço ao Programa de Pós-Graduação em Estatística; pela oportunidade de estudar na Universidade Federal do Rio de Janeiro, e por todo o apoio e compreensão.

Agradeço à Capes; pelo apoio financeiro para estudar enquanto vivi na cidade do Rio de Janeiro.

Muito obrigado.

Atenciosamente,

Juan Pablo

"...estudia, y no serás cuando crecido
ni el juguete vulgar de las pasiones,
ni el esclavo servil de los tiranos."

– Elías Calixto Pompa

Resumo

Esta dissertação aborda a modelagem de tópicos presentes em dissertações de mestrado em estatística e áreas correlatas no Brasil, através de modelos de Alocação Latente de Dirichlet. O principal objetivo é inferir os tópicos latentes abordados nessas dissertações. Primeiramente, discute-se e apresenta-se a construção de um *corpus* de documentos composto pelas dissertações mais recentes em distintas Instituições de Ensino Superior do Brasil, extraídas manualmente a partir dos endereços eletrônicos de cada um dos programas de mestrado analisados. O procedimento inferencial adotado para o modelo de Alocação Latente de Dirichlet consiste em métodos de Monte Carlo via cadeias de Markov e inferência variacional. Discute-se ainda diferentes métodos para escolha do número de tópicos incluindo critérios de informação como o de Akaike, o Bayesiano, o de Deviância, o de Watanabe-Akaike e métricas baseadas na coerência dos tópicos latentes inferidos. A metodologia adotada fornece uma compreensão aprofundada dos tópicos predominantes nesse *corpus*.

Palavras-chave: Alocação Latente de Dirichlet, Modelagem de Tópicos, Dissertações de Mestrado, Método de Monte Carlo via cadeias de Markov, inferência variacional.

Abstract

This master's thesis addresses the topic modeling of master's theses in statistics and related areas in Brazil, through Latent Dirichlet Allocation models. The main objective of the work is to infer the latent topics covered in these theses. First, the construction of a *corpus* of documents is discussed and presented, composed of the most recent theses from different Higher Education Institutions in Brazil, manually extracted from the web pages of each of the analyzed master's programs. The inferential procedure adopted for the Latent Dirichlet Allocation model consists of Markov chain Monte Carlo methods and variational inference. Different methods for choosing the number of topics are also discussed, including information criteria such as Akaike, Bayesian, Deviance, Watanabe-Akaike, and metrics based on the coherence of the inferred latent topics. The adopted methodology provides an in-depth understanding of the predominant topics in this *corpus*.

Keywords: Latent Dirichlet Allocation, Topic Modeling, Master's Theses, Markov chain Monte Carlo methods, variational inference.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A modelagem de tópicos é uma área em constante evolução no campo da análise de textos e mineração de dados, cujo objetivo em geral é inferir tópicos latentes em coleções de documentos. Em particular, no modelo de Alocação Latente de Dirichlet (LDA, do inglês *Latent Dirichlet Allocation* [Blei et al., 2003]), os tópicos são modelados como distribuições de probabilidade latentes sobre o conjunto de palavras em um dicionário (também chamado de *vocabulário*). Basicamente, cada tópico representa um conceito subjacente que pode ser encontrado nos documentos analisados. No caso de documentos de dissertações de mestrado em estatística, tais como os analisados neste trabalho, por exemplo, é muito provável a ocorrência de tópicos de teoria estatística, tópicos de aplicações em diferentes áreas (por exemplo: agricultura, medicina, engenharia, entre outras), e tópicos relacionados com a matemática. Neste exemplo, espera-se uma dependência entre os tópicos abordados nas dissertações e sua proveniência, isso é, a qual universidade o texto pertence, entre outros fatores.

Cada documento, por sua vez, é considerado como uma combinação de tópicos, cada um deles sendo uma distribuição de probabilidade distinta sobre as palavras do vocabulário. No modelo LDA, as palavras são sorteadas aleatoriamente de acordo com um dos tópicos latentes, também escolhido de forma aleatória e independente para cada palavra do documento. Isso significa que um documento pode tratar de diferentes tópicos simultaneamente, e a mistura desses tópicos é o que define seu conteúdo temático.

Tanto os pesos (relevâncias) de cada tópico nos documentos quanto as distribuições de probabilidade sobre o vocabulário (os próprios tópicos) são parâmetros latentes do modelo LDA e que ao serem estimados possibilitam entender melhor a estrutura temática dos documentos da base de dados e como eles se relacionam entre si.

O LDA é dito, portanto, um *modelo probabilístico generativo* de um *corpus* [Blei et al., 2003]. Especificamente, assumimos que K tópicos estão associados a uma coleção, e que cada documento exibe esses tópicos com diferentes proporções. Essa é frequentemente uma suposição natural, pois os documentos em um *corpus* tendem a ser heterogêneos, combinando um subconjunto de ideias principais ou temas que permeiam a coleção como um todo [Blei, 2009].

Esta dissertação de mestrado concentra-se na aplicação de modelos LDA para inferências sobre os tópicos abordados em dissertações de mestrado em estatística e áreas correlatas, em dissertações de Instituições de Ensino Superior (IES) do Brasil. Adota-se o paradigma Bayesiano de inferência sendo utilizados métodos de Monte Carlo via cadeias de Markov (MCMC, do inglês *Markov Chain Monte Carlo*) e de inferência variacional. Alguns dos trabalhos na literatura nos quais utiliza-se o modelo LDA fazendo inferência via MCMC e inferência variacional são por exemplo: Griffiths and Steyvers [2002], Blei et al. [2003], Blei [2009], Hoffman et al. [2013], Kim [2020].

MCMC é um procedimento para obter amostras de distribuições de probabilidade complexas, permitindo que uma cadeia de Markov convirja para a distribuição alvo e, em seguida, extraia amostras dessa cadeia [Gilks et al., 1996]. Cada estado da cadeia é uma atribuição de valores às variáveis sendo amostradas, e as transições entre os estados são Markovianas. No contexto de MCMC para modelos LDA utilizamos o método de amostragem de Gibbs [Geman and Geman, 1984], onde o próximo estado é obtido amostrando-se sequencialmente cada um dos parâmetros de suas respectivas distribuições condicionais completas, que são todas conhecidas [Blei et al., 2003].

A ideia básica da inferência variacional, por outro lado, é minimizar a divergência

de Kullback-Leibler (KL) entre a distribuição de probabilidade a posteriori e a distribuição variacional pertencente a uma classe de distribuições aproximantes. Essencialmente, considera-se uma família de distribuições variacionais, indexadas por um conjunto de parâmetros variacionais. Os parâmetros variacionais são inferidos por um procedimento de otimização que busca encontrar a melhor aproximação para a posteriori dentre as distribuições pertencentes à família variacional.

Como mencionado anteriormente, as condicionais completas do modelo LDA estão disponíveis analiticamente, o que possibilita o uso de um amostrador de Gibbs eficiente para a amostragem das variáveis latentes do modelo. Em contraste, o procedimento de inferência variacional adotado neste trabalho baseia-se na hipótese de *mean-field*, assumindo uma família variacional de distribuições que presume independência entre as variáveis latentes, e utiliza um procedimento de inferência via coordenadas ascendentes para estimação dos parâmetros variacionais. Esse procedimento envolve a otimização iterativa das distribuições variacionais de modo a aproximar a posteriori de forma mais eficiente. Enquanto métodos MCMC tendem a ser mais precisos na aproximação da distribuição a posteriori, pode ser significativamente custoso em termos de tempo computacional no contexto de modelos LDA, especialmente quando o *corpus* contém muitos documentos. Por outro lado, a inferência variacional, embora ofereça uma aproximação mais rápida mesmo para *corpus* com muitos documentos pode não capturar características importantes da distribuição a posteriori a depender do grau de simplicidade da família variacional estipulada em comparação com a complexidade da posteriori a ser aproximada. Dessa forma, a escolha entre métodos MCMC e inferência variacional envolve um balanço entre a precisão da aproximação da posteriori e os recursos computacionais disponíveis.

A base de dados dos documentos (*corpus*) utilizada neste trabalho, foi construída obtendo algumas das dissertações de mestrado mais recentes disponíveis online nos sites dos programas de pós-graduação relacionados com a área da estatística, das várias IES do Brasil. Somente são analisados os resumos em inglês (os *abstracts*), uma vez que nessa seção

concentram-se as palavras e as ideias principais das dissertações. A escolha pelos resumos em inglês justifica-se pelo fato de que muitas das terminologias utilizadas no contexto estatístico têm origem no inglês. Utilizar exclusivamente os resumos poderia facilitar o pré-processamento do *corpus*, uma vez que possibilitaria o uso de funções automatizadas já disponíveis.

1.1 Objetivos

O objetivo geral desta pesquisa é fornecer uma metodologia estatística capaz de inferir sobre os tópicos latentes predominantemente abordados nas dissertações de mestrado no campo da estatística no Brasil, ao longo dos últimos anos. A seguir lista-se os objetivos específicos desta dissertação.

1. Construir a base de dados, composta por dissertações de mestrado relevantes no campo da estatística no Brasil. Esta base de dados servirá como o *corpus* de texto sobre o qual o modelo LDA será aplicado.

2. Inferir os assuntos abordados nas dissertações em estatística e áreas correlatas, com a utilização do modelo LDA para identificar e categorizar os tópicos tratados nas dissertações de mestrado em estatística. Para isso, será necessária a aplicação de técnicas de análise de texto para processar e preparar os dados das dissertações para entrada no modelo LDA.

3. Implementação do modelo LDA por métodos de MCMC e via inferência variacional como uma alternativa computacionalmente mais eficiente para a implementação da LDA em que se pese a grande quantidade de parâmetros latentes a serem estimados.

4. Apresentação e discussão dos resultados obtidos por meio do modelo LDA. Os tópicos identificados são analisados e interpretações serão fornecidas sobre os assuntos abordados nas dissertações de mestrado em estatística no Brasil nos últimos anos.

1.2 Procedimentos e ferramentas de implementação

A leitura dos arquivos no formato PDF para compor o *corpus* de documentos, o pré-processamento do *corpus*, a construção da base de dados, a inferência do modelo LDA e outras implementações; serão realizadas na linguagem de programação R [R Core Team, 2023]. Os arquivos dos scripts e de tabelas com informação da base de dados e do *corpus* de documentos, serão disponibilizados no enlace: `https://github.com/JuanPabloAO/Codigos_no_R.git`.

1.3 Estrutura do texto

O restante desta dissertação apresenta a seguinte estrutura: no capítulo 2, apresenta-se a construção da base de dados; no capítulo 3, fala-se de inferência Bayesiana, com enfoque em métodos de MCMC e inferência variacional; no capítulo 4, aborda-se a modelagem de tópicos por meio de LDA, utilizando MCMC e inferência variacional; no capítulo 5, detalha-se os resultados; e no capítulo 6, são expostas as conclusões e trabalhos futuros.

Capítulo 2

Construção da Base de Dados

Neste capítulo, descreve-se o processo de construção da base de dados provenientes das IES no Brasil que oferecem cursos de mestrado na área de estatística, detalhando-se a metodologia empregada para coletar os dados de interesse.

A fonte de informação primária utilizada para compor a lista das IES que oferecem cursos de mestrado em estatística no Brasil foi o site do Conselho Regional de Estatística 3° Região (CONRE-3) através da URL https://www.conre3.org.br/portal/instituicoes-de-ensino/ (acesso em 18/06/2023). Inicialmente, identificou-se um total de 26 IES listadas no site do CONRE-3. Para cada uma destas instituições, acessou-se seus respectivos sites com o objetivo de extrair as dissertações mais recentes nos cursos de mestrado em estatística que ofereciam. Em cada site foram extraídas até 20 dissertações mais recentes, selecionadas como amostras representativas das temáticas dos programas de mestrado em estatística oferecidos por cada instituição nos últimos anos.

Por outro lado, 3 das 26 IES foram descartadas da análise, por razões que incluem a falta de informações relevantes em seus sites, a inexistência de programas de mestrado em estatística ou a falta de acesso às dissertações mais recentes. As IES em questão são:

- **Instituto de Matemática Pura e Aplicada (IMPA):** possui Mestrado Acadêmico em Matemática (a linha de Probabilidade está em Doutorado); há poucas dissertações no site. As dissertações variam de 2011 a 2022. Não foi encontrado um

repositório disponível até a data de setembro de 2023, em que se concluiu a coleta dos dados textuais das IES.

- **Universidade Federal de Pelotas (UFPEL):** possui um programa de Mestrado em Epidemiologia pertencente à Faculdade de Medicina. Desconsiderou-se a Universidade porque considerou-se que os temas eram muito específicos em tópicos de Medicina, com baixa ocorrência de palavras relacionadas a estatística nos *Abstracts* e capítulos introdutórios.

- **Universidade Federal de São João del-Rei (UFSJ):** não consta no site o Mestrado em Estatística Aplicada e Biometria (nem em qualquer outro tópico relacionado à Estatística). O Repositório Institucional da UFSJ (RI-UFSJ) está em processo de implementação. Foi aprovado pelo Conselho Universitário (CONSU), em 8 de novembro de 2021 conforme o Regimento Interno do RI-UFSJ.

Na Tabela 2.1, observa-se a distribuição dos anos das dissertações, sendo a maioria no ano 2022 (com 133 documentos), seguido pelo ano 2021 (com 119 documentos); e com um intervalo de anos desde 2013 até 2023.

Tabela 2.1: Distribuição de frequências dos anos das dissertações que compõem a base de dados.

2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
1	1	2	1	4	14	32	75	119	133	35

Na Tabela 2.2 apresenta-se o número de dissertações que foram aceitas para compor a base de dados, tendo sido inicialmente selecionadas as 20 dissertações mais recentes disponíveis nos sites de cada IES. Há uma exceção neste sentido: a Universidade Federal do Rio Grande do Sul (UFRGS) possui apenas 17 documentos de dissertações acessíveis até a data de conclusão da coleta de dados. Finalmente, algumas das dissertações foram excluídas da análise devido ao fato de terem sido digitalizadas como imagens, em vez de

serem arquivos de texto no formato PDF, dificultando assim a extração automatizada das palavras no processo de construção do vocabulário. No total, foram analisadas 417 dissertações.

Tabela 2.2: Número de dissertações aceitas para a construção da base de dados em cada uma das Instituições de Ensino Superior (IES).

	NOME DA IES	NÚMERO DE DISSERTAÇÕES
1	Escola Nacional de Ciências Estatísticas (ENCE)	20
2	Escola Nacional de Saúde Pública Sérgio Arouca (ENSP)	20
3	Escola Superior de Agricultura Luiz de Queiroz- USP (ESALQ)	20
4	Universidade Estadual de Maringá (UEM)	16
5	Universidade Federal do Amazonas (UFAM)	17
6	Universidade Federal do Ceará (UFC)	20
7	Universidade Federal de Campina Grande (UFCG)	11
8	Universidade Federal de Lavras (UFLA)	20
9	Universidade Federal de Minas Gerais (UFMG)	19
10	Universidade Federal do Pará (UFPA)	20
11	Universidade Federal da Paraíba (UFPB)	20
12	Universidade Federal de Pernambuco (UFPE)	20
13	Universidade Federal do Rio Grande do Sul (UFRGS)	17
14	Universidade Federal do Rio de Janeiro (UFRJ)	15
15	Universidade Federal do Rio Grande do Norte (UFRN)	16
16	Universidade Federal Rural de Pernambuco (UFRPE)	19
17	Universidade Federal de São Carlos (UFSCar)	20
18	Universidade Federal de Viçosa (UFV)	19
19	Universidade de Brasilia (UnB)	19
20	Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)	12
21	Universidade Estadual de Campinas (UNICAMP)	19
22	Universidade Federal de Alfenas (UNIFAL)	20
23	Universidade de São Paulo (USP)	18

Apenas os resumos em inglês (*abstracts*) das dissertações foram considerados para compor o *corpus* de documentos. Como pré-processamento, foram removidas as *stopwords* (palavras sem significado relevante para caracterização dos tópicos latentes, como conectivos e preposições por exemplo). Além disso, as palavras foram submetidas a um processo de *lematização* e *stemização*. O primeiro processo consiste em, dada uma forma flexionada de uma palavra (ou seja, no plural, no feminino, conjugada, entre outras), encontrar o lema correspondente, isso é, a forma que, por convenção, é aceita como representante de todas as formas flexionadas dessa palavra. O segundo processo consiste em remover prefixos e sufixos das palavras extraindo-se apenas seu radical, para ter uma maior

consolidação das palavras finais no vocabulário. Por exemplo:

- A palavra *models* é transformada em *model* com a lematização, retirando-se o plural. A stemização também reduz a palavra *models* à *model*, porque a raiz é a mesma.
- A palavra *distributions*, é transformada em *distribution* com a lematização, enquanto *distributed*, fica *distribute* também com a lematização. Aplicando stemização nas duas palavras obtém-se *distribut*.

A base de dados final é composta por três atributos, codificadas como:

- **word:** são as palavras do vocabulário, codificadas numericamente como $\{1, \ldots, V\}$, com $V = 5935$.
- **doc:** são os documentos que compõem o *corpus*, codificados numericamente como $\{1, \ldots, D\}$, com $D = 417$.
- **freq:** é a frequência com que a palavra $v \in \{1, \ldots, V\}$ aparece no documento $d \in \{1, \ldots, D\}$. Na base de dados tem-se um total de 61892 palavras com repetição (observações) das quais 61375 são de fato palavras válidas, uma vez que 107 palavras foram excluídas do vocabulário (equivalente a 517 observações) por não serem interpretáveis (como por exemplo: *<U+0001D4A2>*, *<U+03B3>k<U+03B3>*, *<U+03B7>*~).

Uma representação abreviada da estrutura da base de dados consta na Tabela 2.3 em que se destacam as primeiras cinco linhas e a última, sendo composta por 39236 linhas e ordenada pela coluna 2 ("**doc**") e, em seguida, pela coluna 1 ("**word**"). Por exemplo, poderíamos fazer a interpretação da primeira linha da seguinte forma: *"a palavra de número* 96*, aparece uma vez no documento* 1*"*. Fazendo uso das tabelas de conversão de números em palavras e documentos, tem-se: *"a palavra 'advisor' ('orientador', em português), aparece uma única vez na dissertação com nome: 'A reconfiguração dos territórios da pesca artesanal na ilha de Paquetá, Baía de Guanabara, RJ- 2023', da ENCE"*.

Tabela 2.3: Estrutura da base de dados.

word	**doc**	**freq**
96	1	1
208	1	1
227	1	2
233	1	1
345	1	8
⋮	⋮	⋮
5867	417	1

Por fim a Figura 2.1 apresenta as distribuições do número de palavras contidas nos resumos em inglês das dissertações, dentro de cada IES. Pode-se observar que a UFPB é a IES com as dissertações com maior número de palavras (com assimetria positiva), sendo o número máximo de 351 palavras. A UFCG é a IES que apresenta o menor número de palavras nas dissertações, sendo o número mínimo de 33 palavras. Também observa-se na Figura 2.1, valores atípicos elevados nas IES: ENSP, UEM, UFMG, UFPE, UFRJ, e UFRPE; e valores atípicos baixos somente na UFV.

A seguir, no capítulo 3, apresenta-se a teoria geral de inferência Bayesiana, em que se faz uma breve definição de termos gerais como: distribuição a priori, distribuição a posteriori, conjugação, MCMC e inferência variacional. Os detalhes do uso dos métodos inferenciais na estimação dos parâmetros do modelo LDA, à luz da base de dados construída serão discutidos no capítulo 4.

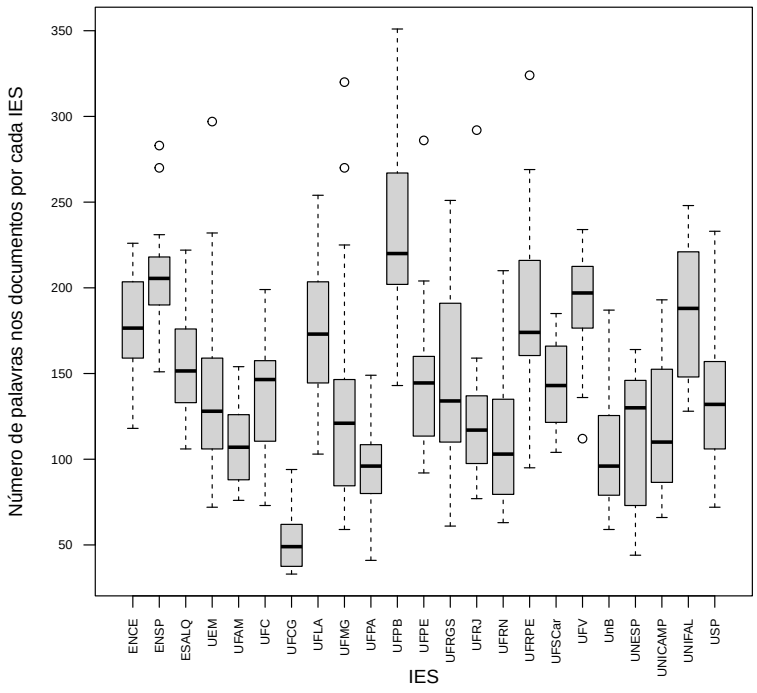


Figura 2.1: Distribuições do número de palavras nos *abstracts* das dissertações, por Instituição de Ensino Superior (IES); apos da remoção das *stopwords*.

Capítulo 3

Inferência Bayesiana

Este capítulo aborda conceitos gerais de inferência Bayesiana e métodos computacionais, MCMC e inferência variacional em linhas gerais. Primeiramente define-se: y, como uma observação da variável aleatória Y; $\boldsymbol{\vartheta}$, como os parâmetros da distribuição da variável aleatória Y, ou seja, $y \sim p(y \mid \boldsymbol{\vartheta})$; e o vetor $\mathbf{y} = (y_1, \ldots, y_n)$, como uma amostra ou um conjunto de n observações independentes da variável aleatória Y, ou seja, $y_1, \ldots, y_n \sim p(y \mid \boldsymbol{\vartheta})$.

A *distribuição a priori* é a distribuição do(s) parâmetro(s) antes de qualquer dado ser observado, denotada por $p(\boldsymbol{\vartheta})$. A *distribuição amostral* é a distribuição dos dados observados condicionada aos seus parâmetros, ou seja, $p(\mathbf{y} \mid \boldsymbol{\vartheta})$. Essa quantidade também é chamada de *verossimilhança*, se vista como uma função do(s) parâmetro(s), por vezes escrita como $\mathrm{L}(\boldsymbol{\vartheta}; \mathbf{y}) = p(\mathbf{y} \mid \boldsymbol{\vartheta})$.

A *distribuição a posteriori* é a distribuição do(s) parâmetro(s) após a observação dos dados, determinada pelo teorema de Bayes,

$$p(\boldsymbol{\vartheta} \mid \mathbf{y}) = \frac{p(\boldsymbol{\vartheta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta})}{p(\mathbf{y})} \propto p(\mathbf{y} \mid \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}).$$

Na prática, para quase todos os modelos Bayesianos complexos usados em aprendizado de máquina, a distribuição a posteriori $p(\boldsymbol{\vartheta} \mid \mathbf{y})$ não é obtida em uma forma fechada. Nessas situações, é necessário recorrer a técnicas computacionais, como MCMC

via amostrador de Gibbs ou Metropolis–Hastings, e o algoritmo de inferência variacional via coordenadas ascendentes (CAVI, do inglês *Coordinate Ascent Variational Inference*) [Lee, 2021].

A distribuição a priori costuma ser, quando possível, assumida como proveniente de uma família de distribuições chamadas *prioris conjugadas*. A utilidade de uma priori conjugada é que a distribuição a posteriori correspondente estará na mesma família, e o cálculo de sua condicional completa pode ser expresso em forma fechada.

A *distribuição preditiva a posteriori* é a distribuição de um novo dado y^* condicionado aos dados observados, marginalizada sobre a posteriori:

$$p(y^* \mid \mathbf{y}) = \int p(y^* \mid \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta} \mid \mathbf{y}) d\boldsymbol{\vartheta}.$$

A teoria Bayesiana utiliza a distribuição preditiva a posteriori para realizar inferências preditivas, isto é, em vez de somente fornecer estimativas pontuais para a previsão, tem-se uma distribuição sobre todos os pontos possíveis; somente dessa forma é que toda a distribuição a posteriori dos parâmetros é utilizada. A distribuição a posteriori não se limita a fornecer um único valor estimado para os parâmetros, mas sim uma distribuição completa que reflete todas as possíveis estimativas dos parâmetros, ponderadas por suas probabilidades. Isso é fundamental para as previsões, pois permite capturar a incerteza associada aos parâmetros e fornecer uma previsão mais robusta e realista.

Em comparação, a previsão na estatística frequentista geralmente envolve encontrar uma estimativa pontual ótima do(s) parâmetro(s) (por exemplo, por máxima verossimilhança) e em seguida substituir essa estimativa no lugar do verdadeiro valor do(s) parâmetro(s) que caracteriza(m) a distribuição dos dados a serem preditos. A previsão frequentista nestes moldes não leva em consideração a incerteza quanto ao valor do parâmetro, e, portanto, subestima a variância das previsões [Gelman et al., 2013].

3.1 MCMC

MCMC é uma classe de métodos de simulação baseados em cadeias de Markov para obter amostras aproximadas de uma distribuição alvo $\pi(\boldsymbol{\vartheta})$, sendo esta a distribuição estacionária da cadeia de Markov em questão. A amostragem da cadeia é feita de forma sequencial, com a distribuição condicional de transição satisfazendo a propriedade Markoviana: para qualquer iteração $t \in \mathbb{N}$, a distribuição de $\boldsymbol{\vartheta}_t$, dado todos os valores $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_{t-1}$ anteriores, depende apenas do valor mais recente, $\boldsymbol{\vartheta}_{t-1}$.

Em contextos de inferência Bayesiana em que a distribuição alvo da qual se deseja amostrar é a posteriori $\pi(\boldsymbol{\vartheta}) = p(\boldsymbol{\vartheta} \mid \mathbf{y})$, os métodos MCMC costumam ser usados quando não é possível (ou não é computacionalmente eficiente) amostrar $\boldsymbol{\vartheta}$ diretamente de $p(\boldsymbol{\vartheta} \mid \boldsymbol{y})$. Em vez disso, os métodos MCMC produzem amostras de forma iterativa, de modo que, a cada passo do processo, esperamos extrair de uma distribuição que se aproxima de $p(\boldsymbol{\vartheta} \mid \boldsymbol{y})$ desde que a simulação seja realizada por tempo suficiente para que a distribuição das amostras de $\boldsymbol{\vartheta}$ esteja suficientemente próxima da distribuição estacionária $\pi(\boldsymbol{\vartheta})$ [Gelman et al., 2013].

3.1.1 Amostrador de Gibbs

Um algoritmo particular de MCMC que tem sido útil em muitos problemas multidimensionais é o *amostrador de Gibbs* [Gelman et al., 2013]. Suponha que o vetor de parâmetros $\boldsymbol{\vartheta}$ tenha sido dividido em d componentes ou subvetores, $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_d)$. A cada iteração do amostrador de Gibbs, percorre-se os componentes de $\boldsymbol{\vartheta}$, amostrando cada um condicionalmente ao valor de todos os outros. Existem assim d etapas de amostragem para cada iteração do algoritmo. A cada iteração t, amostra-se $\boldsymbol{\vartheta}_j$ $(j = 1, \ldots, d)$ da distribuição condicional dada todas as outras componentes de $\boldsymbol{\vartheta}$:

$$\boldsymbol{\vartheta}_j^t \sim p(\boldsymbol{\vartheta}_j \mid \boldsymbol{\vartheta}_{-j}^{t-1}, \boldsymbol{y}), \quad j = 1, \ldots, d,$$

onde $\boldsymbol{\vartheta}_{-j}^{t-1}$ representa todos os componentes de $\boldsymbol{\vartheta}$, exceto para $\boldsymbol{\vartheta}_j$, nos seus valores atuais:

$$\boldsymbol{\vartheta}_{-j}^{t-1} = (\boldsymbol{\vartheta}_1^t, \ldots, \boldsymbol{\vartheta}_{j-1}^t, \boldsymbol{\vartheta}_{j+1}^{t-1}, \ldots, \boldsymbol{\vartheta}_d^{t-1}).$$

Assim, cada subvetor $\boldsymbol{\vartheta}_j$ é atualizado condicionalmente aos valores mais recentes das outras componentes de $\boldsymbol{\vartheta}$, que são os valores da iteração t para as componentes já atualizadas e os valores da iteração $t-1$ para as demais.

Para muitos problemas envolvendo modelos estatísticos, é possível amostrar diretamente de muitas ou todas as distribuições condicionais completas dos parâmetros. Este é o caso, por exemplo, dos modelos LDA cujas distribuições condicionais completas tem forma fechada conhecida, o que possibilita a aplicação do amostrador de Gibbs. Embora existam outros algoritmos de MCMC, como o de Metropolis-Hastings ou o de Monte Carlo Hamiltoniano, optou-se por utilizar o amostrador de Gibbs, uma vez que este será empregado para treinar os modelos LDA neste trabalho. A escolha do amostrador de Gibbs é justificada pela sua simplicidade, eficiência e pela facilidade de implementação nos modelos LDA.

3.2 Inferência variacional

Os métodos de inferência variacional costumam ser usados para realizar inferência em modelos probabilísticos complexos, onde a inferência exata é muitas vezes inviável ou demasiadamente custosa. A inferência variacional aproxima uma distribuição complexa, no nosso caso a posteriori por uma distribuição mais simples $q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})$ proveniente de uma família de distribuições variacionais $\mathcal{Q} = \{q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}); \boldsymbol{\psi} \in \mathcal{V}\}$ pré especificada [Jordan et al., 1999].

A escolha da família de distribuições variacionais $\mathcal{Q}$ deve ser tal que seja composta de distribuições fáceis de se manipular e amostrar, bem como realizar procedimentos de otimização nos seus parâmetros, como os descritos mais adiante. Geralmente, escolhe-se

uma distribuição de uma família de distribuições parametrizadas respeitando o suporte de $\boldsymbol{\vartheta}$ à posteriori, como por exemplo uma distribuição normal multivariada no caso em que o suporte de $\boldsymbol{\vartheta}$ é o conjunto $\mathbb{R}^d$ [Blei et al., 2017; Kucukelbir et al., 2017].

A implementação de métodos de inferência variacional requer a obtenção de uma cota inferior sobre a verossimilhança marginal dos dados observados, isto ocorre porque calcular a verossimilhança marginal exata pode ser computacionalmente ineficiente ou até mesmo impraticável, especialmente em modelos complexos. Define-se a *verossimilhança marginal* (às vezes também chamada de *evidência*) como a distribuição dos dados observados marginalizada com respeito ao(s) parâmetro(s), ou seja,

$$p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\vartheta}) p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}.$$

No contexto de inferência variacional, uma vez especificada a família variacional de distribuições de probabilidade parametrizada por $\boldsymbol{\psi}$, o próximo passo é estabelecer um problema de otimização que determine os valores ótimos dos parâmetros variacionais $\boldsymbol{\psi}$. Seguindo Jordan et al. [1999], começamos limitando inferiormente a log-verossimilhança marginal (ou o logaritmo da evidência) usando a desigualdade de Jensen:

$$\begin{aligned}
\log p(\mathbf{y}) &= \log \int p(\boldsymbol{\vartheta}, \mathbf{y}) d\boldsymbol{\vartheta} \\
&= \log \int \frac{p(\boldsymbol{\vartheta}, \mathbf{y}) q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})}{q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})} d\boldsymbol{\vartheta} \\
&\geq \int q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \log p(\boldsymbol{\vartheta}, \mathbf{y}) d\boldsymbol{\vartheta} - \int q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} \\
&= \mathbb{E}_{\boldsymbol{\vartheta} \sim q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})} \left[\log p(\boldsymbol{\vartheta}, \mathbf{y}) \right] - \mathbb{E}_{\boldsymbol{\vartheta} \sim q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})} \left[\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \right]. \qquad (3.1)
\end{aligned}$$

Assim, vemos que a desigualdade de Jensen nos fornece um limite inferior para a log-verossimilhança marginal em função de uma distribuição variacional arbitrária $q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})$. Originalmente, Jordan et al. [1999] definem como objetivo da inferência variacional a maximização desta cota inferior denominada limite inferior de evidência (ELBO, do inglês

Evidence Lower Bound) denotada por $\mathcal{L}(\boldsymbol{\psi}; \mathbf{y}) = \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})]$ (expressão do final da Equação (3.1)) em função dos parâmetros variacionais. Pode-se verificar facilmente que a diferença entre o primeiro e o último termo da Equação (3.1), ou seja, $\log p(\mathbf{y}) - \{\mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})]\}$, é a divergência de Kullback-Leibler (KL) entre a distribuição variacional e a distribuição a posteriori, definida como:

$$KL(q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \parallel p(\boldsymbol{\vartheta} \mid \mathbf{y})) = \mathbb{E}_q \left[\log \frac{q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})}{p(\boldsymbol{\vartheta} \mid \mathbf{y})} \right].$$

A divergência de KL quantifica a diferença entre a distribuição a posteriori e a distribuição variacional [Kullback and Leibler, 1951; Kullback, 1959], ou seja, o excesso esperado de surpresa ao utilizar a distribuição a posteriori como modelo em vez da distribuição variacional, quando esta última é considerada a distribuição real para o propósito de aproximação. A divergência de KL é sempre um número real não negativo, com valor 0 se e somente se as duas distribuições em questão forem idênticas. Temos:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\psi}; \mathbf{y}) + KL(q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \parallel p(\boldsymbol{\vartheta} \mid \mathbf{y})) &= \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})] + \mathbb{E}_q \left[\log \frac{q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})}{p(\boldsymbol{\vartheta} \mid \mathbf{y})} \right] \\
&= \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})] + \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \\
&\quad - \log p(\boldsymbol{\vartheta} \mid \mathbf{y})] \\
&= \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})] \\
&\quad + \mathbb{E}_q \left\{ \log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) - \log \left[\frac{p(\boldsymbol{\vartheta}, \mathbf{y})}{p(\mathbf{y})} \right] \right\} \\
&= \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})] + \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \\
&\quad - \log p(\boldsymbol{\vartheta}, \mathbf{y}) + \log p(\mathbf{y})] \\
&= \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] - \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})] + \mathbb{E}_q [\log q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})] \\
&\quad - \mathbb{E}_q [\log p(\boldsymbol{\vartheta}, \mathbf{y})] + \log p(\mathbf{y}) \\
&= \log p(\mathbf{y}), \qquad (3.2)
\end{aligned}$$

onde $KL(q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \parallel p(\boldsymbol{\vartheta} \mid \mathbf{y})) = \log p(\mathbf{y}) - \mathcal{L}(\boldsymbol{\psi}; \mathbf{y})$. Sendo assim, do fato de que o último termo da Equação (3.2) não depende de $\boldsymbol{\psi}$, decorre que a maximização do ELBO $\mathcal{L}(\boldsymbol{\psi}; \mathbf{y})$ é equivalente à minimização da divergência de Kullback-Leibler entre a distribuição variacional e a posteriori. Portanto, conclui-se que os procedimentos de inferência variacional buscam obter a distribuição $q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})$ em $\mathcal{Q}$ que melhor aproxima a distribuição a posteriori em termos da divergência de Kullback-Leibler.

A obtenção do ponto de ótimo $\boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\psi}} KL(q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) \parallel p(\boldsymbol{\vartheta} \mid \mathbf{y})) = \arg\max_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\psi}; \mathbf{y})$ pode ser feita por meio de métodos iterativos de coordenada ascendente, ou por métodos baseados em gradientes. O algoritmo de inferência variacional via coordenadas ascendentes (*CAVI*, do inglês *Coordinate Ascent Variational Inference*) [Bishop, 2006; Blei et al., 2017] minimiza iterativamente a divergência de Kullback-Leibler (KL) entre $q_{\boldsymbol{\psi}_j}(\boldsymbol{\vartheta}_j)$ e $p(\boldsymbol{\vartheta} \mid \mathbf{y})$, enquanto fixa os outros $q_{\boldsymbol{\psi}_i}(\boldsymbol{\vartheta}_i)$ (onde $i \neq j$) nos valores mais recentemente atualizados [Lee, 2021]; supondo que a distribuição variacional possa ser decomposta como um produto de distribuições independentes, uma para cada variável $\boldsymbol{\vartheta}_j$, ou seja,

$$q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta}) = \prod_{j=1}^{n} q_{\boldsymbol{\psi}_j}(\boldsymbol{\vartheta}_j),$$

onde $q_{\boldsymbol{\psi}_j}(\boldsymbol{\vartheta}_j)$ é a densidade variacional associada à variável $\boldsymbol{\vartheta}_j$, e $\boldsymbol{\psi}_j$ é o parâmetro variacional que deve ser otimizado. A hipótese de *mean-field* assume que as variáveis $\boldsymbol{\vartheta}_j$ são independentes entre si, ou seja, sua distribuição conjunta pode ser aproximada como o produto de distribuições marginais independentes [Jordan et al., 1999], como indicado anteriormente. Esta suposição simplifica o problema, pois permite tratar cada variável separadamente, sem considerar a dependência entre eles.

A otimização dos parâmetros $\boldsymbol{\psi}_j$ é realizada de forma iterativa. Em cada iteração, os valores de $\boldsymbol{\psi}_i$ (para $i \neq j$) são considerados fixos, pois já foram atualizados nas iterações anteriores. Ou seja, em cada passo, a distribuição $q_{\boldsymbol{\psi}_j}(\boldsymbol{\vartheta}_j)$ é otimizada enquanto as distribuições das outras variáveis permanecem inalteradas.

Para atualizar o valor de $\boldsymbol{\psi}_j$, maximiza-se a função do ELBO $\mathcal{L}(\boldsymbol{\psi}; \mathbf{y})$. Esse processo é feito derivando $\mathcal{L}(\boldsymbol{\psi}; \mathbf{y})$ em relação a $\boldsymbol{\psi}_j$ e igualando a derivada a zero, ou seja,

$$\frac{\partial}{\partial \boldsymbol{\psi}_j} \mathcal{L}(\boldsymbol{\psi}; \mathbf{y}) = 0.$$

Esse procedimento garante que o valor de $\boldsymbol{\psi}_j$ seja o que maximiza o ELBO para aquele parâmetro, considerando que os outros parâmetros estão fixos.

Os passos de otimização (otimização iterativa e maximização) são repetidos até que os parâmetros $\boldsymbol{\psi}_j$ se estabilizem, ou seja, até que a função do ELBO $\mathcal{L}(\boldsymbol{\psi}; \mathbf{y})$ não apresente mais melhorias significativas. A repetição do processo continua até que o algoritmo convirja, o que significa que os parâmetros $\boldsymbol{\psi}_j$ atingem um valor que não muda substancialmente entre as iterações subsequentes. Vale ressaltar que, embora o algoritmo CAVI garanta a convergência para um ótimo local da função ELBO, ele não assegura que esse ótimo seja o global. Portanto, a distribuição variacional encontrada pelo algoritmo será uma boa aproximação da distribuição a posteriori $p(\boldsymbol{\vartheta} \mid \mathbf{y})$, mas com a possibilidade de não ser a melhor aproximação possível no espaço de soluções.

No capítulo 4, serão feitos os cálculos detalhados para o modelo LDA na abordagem do MCMC e inferência variacional.

Capítulo 4

Inferência Bayesiana para modelos LDA

Modelos Bayesianos para tópicos em documentos de texto, como o modelo de alocação de Dirichlet latente (LDA), são ferramentas importantes para a análise estatística de coleções de documentos [Blei and Lafferty, 2007]. O modelo LDA assume que as palavras de cada documento surgem de uma mistura de tópicos, cada um dos quais sendo uma distribuição de probabilidade sobre o vocabulário.

Neste capítulo será abordada a modelagem de tópicos via LDA e a estimação dos seus parâmetros utilizando MCMC e inferência variacional.

4.1 Modelagem de tópicos via LDA

Em problemas envolvendo modelagem de tópicos para documentos de texto, os dados $\mathbf{w} = \{w_{dn}; d \in \{1, \ldots, D\}; n \in \{1, \ldots, N_d\}\}$ correspondem a palavras w_{dn}, que ocorrem em cada documento d, de modo que D denota o número de documentos na base de dados e N_d o número de palavras no documento d. Modelos estocásticos para aprendizagem de tópicos descrevem a ocorrência das palavras nos documentos através de um processo aleatório caracterizado por parâmetros desconhecidos cujos valores numéricos precisam ser estimados à luz dos dados. As palavras w_{dn} no documento d são geradas da seguinte forma: para cada espaço em branco no documento (representado pelo índice $n \in \{1, \ldots, N_d\}$) amostra-se aleatoriamente um tópico $\boldsymbol{\beta}_k$ da lista de K tópicos disponí-

veis $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ com probabilidades $\theta_{d1}, \ldots, \theta_{dK}$, respectivamente. Em seguida, escolhe-se uma palavra aleatoriamente utilizando a distribuição de probabilidades $\boldsymbol{\beta}_k$ correspondente ao tópico selecionado. Por representarem distribuições de probabilidade sobre um vocabulário com, digamos, V palavras no total, denota-se os tópicos $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kV})^\top$ onde $k \in \{1, \ldots, K\}$ como vetores de probabilidade em dimensão V. Isso significa que $\boldsymbol{\beta}_k \in [0,1]^V$ com $\sum_{v=1}^{V} \beta_{kv} = 1 \ \forall \ k \in \{1, \ldots, K\}$. Os vetores $\boldsymbol{\beta}_k$ e $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK})^\top$ são parâmetros a serem estimados e portanto, sob a ótica Bayesiana, requerem uma distribuição a priori, no caso dada por $\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \text{Dirichlet}(\eta_1, ..., \eta_V)$ e $\boldsymbol{\theta}_d \overset{\text{iid}}{\sim} \text{Dirichlet}(\alpha_1, ..., \alpha_K)$ com $\boldsymbol{\beta}_k$ e $\boldsymbol{\theta}_d$ independentes para todo $d = 1, \ldots, D$ e $k = 1, \ldots, K$. Para fins de obtenção analítica das condicionais completas necessárias para implementação de MCMC, define-se variáveis latentes $z_{dn} \in \{1, \ldots, K\}$ que indicam o tópico sorteado para a n-ésima palavra do documento d, i.e., $z_{dn} = k$ indica que o tópico k será utilizado para sortear a n-ésima palavra do documento d. A Figura 4.1 ilustra o processo gerador de palavras nos documentos segundo o modelo LDA.

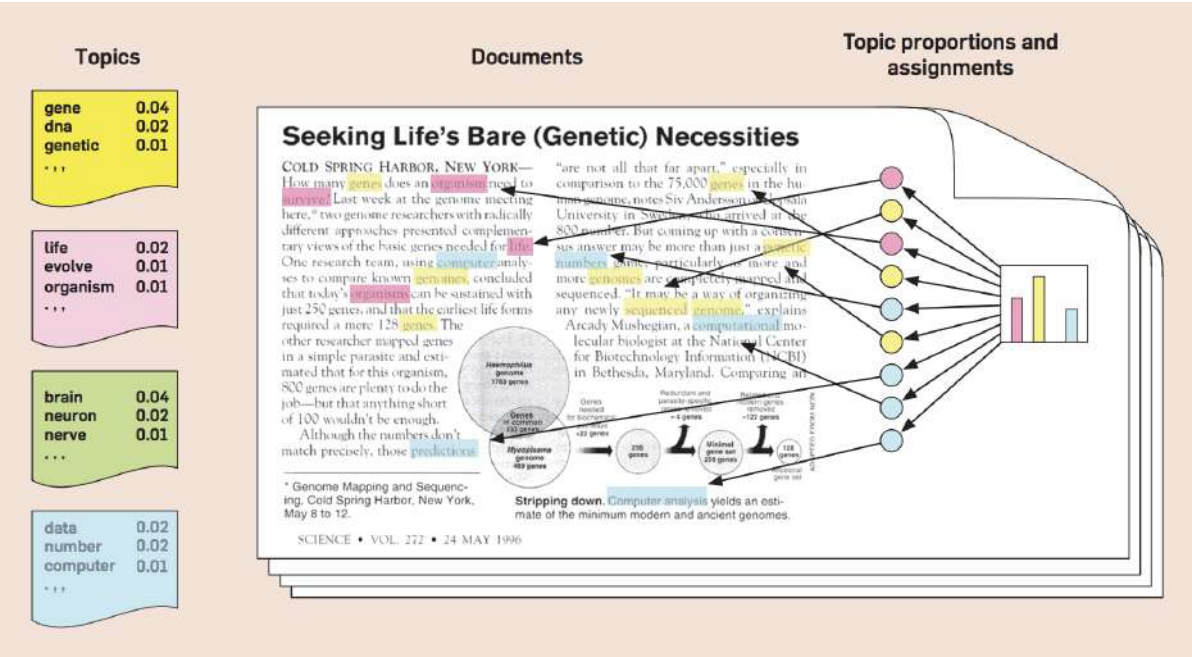


Figura 4.1: Representação ilustrativa do procedimento para geração de documentos de acordo com o modelo LDA. Figura extraída de Blei [2012].

O modelo probabilístico descrito no parágrafo anterior e ilustrado na Figura 4.1 é definido hierarquicamente da seguinte forma:

1. Amostre K tópicos: $\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \text{Dirichlet}(\eta_1, ..., \eta_V)$, para $k \in \{1, ..., K\}$;

2. Para cada documento $d \in \{1, ..., D\}$,

   - Selecione as proporções dos tópicos: $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$;
   - Para cada palavra w_{dn} a ser amostrada no documento d:
     (a) Amostre o indicador do tópico do qual a n-ésima palavra será sorteada:
     $z_{dn} \mid \boldsymbol{\theta}_d \sim \text{Categórica}(\boldsymbol{\theta}_d) \iff \boldsymbol{z}_{dn} \mid \boldsymbol{\theta}_d \sim \text{Multinomial}(1; \boldsymbol{\theta}_d)$;
     (b) Sorteie uma palavra de acordo com o tópico selecionado:
     $w_{dn} \mid \boldsymbol{\beta}_{z_{dn}} \sim \text{Categórica}(\boldsymbol{\beta}_{z_{dn}}) \iff \boldsymbol{w}_{dn} \mid \boldsymbol{\beta}_{z_{dn}} \sim \text{Multinomial}(1; \boldsymbol{\beta}_{z_{dn}})$.

Voltando a exemplo ilustrado na Figura 4.1, observa-se no lado esquerdo, uma coluna com quatro tópicos e as suas três palavras mais prováveis: tópico 1 ($\boldsymbol{\beta}_1$; *gene* [0,04], *dna* [0,02], *genetic* [0,01],...), tópico 2 ($\boldsymbol{\beta}_2$; *life* [0,02], *envolve* [0,01], *organism* [0,01],...), tópico 3 ($\boldsymbol{\beta}_3$; *brain* [0,04], *neuron* [0,02], *nerve* [0,01],...), e tópico 4 ($\boldsymbol{\beta}_4$; *data* [0,02], *number* [0,02], *computer* [0,01],...). Logo, cada documento d será composto por uma proporção ($\boldsymbol{\theta}_d$) específica de palavras advindas de cada um dos tópicos. Pode-se ver no lado direito da Figura 4.1, o gráfico de barras dessas proporções no documento do exemplo ($\boldsymbol{\theta}_{d^*}$; sendo d^* o documento com nome: *"Seeking Life's Bare [Genetic] Necessities"*); em que o tópico 1 tem a maior proporção (prevalência), seguido pelo tópico 2, e o tópico 4 (respectivamente), mas o tópico 3 não teria relevância no documento do exemplo (d^*). Depois, para a geração documento, se faz o sorteio das palavras (w_{d^*n}) que vão formá-lo. Considera-se um documento como um conjunto de espaços (sem importar a sua ordem), os quais vão a ser preenchidos com as palavras amostradas; para a n-ésima palavra a ser amostrada, primeiro amostra-se o tópico (z_{d^*n}) segundo as proporções no documento do exemplo $\boldsymbol{\theta}_{d^*} = (\theta_{d^*1}, \ldots, \theta_{d^*K})$. Digamos que o k-ésimo tópico seja sorteado, ou seja,

$z_{dn} = k$. Em seguida, amostra-se a palavra a ser atribuída nesse espaço, segundo a distribuição das palavras do vocabulário nesse tópico $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kV})$. Observa-se como isso acontece na Figura 4.1, lembrando que algumas palavras podem ter a mesma raiz depois da lematização e stemização:

1. Para o primeiro espaço (círculo pequeno), foi amostrado o tópico 2 ($z_{d^*1} = 2$), do qual foi escolhida aleatoriamente a palavra *life*.

2. Para o segundo espaço, foi amostrado o tópico 1 ($z_{d^*2} = 1$), do qual foi escolhida aleatoriamente a palavra *genomes* (a qual não é uma das três palavras mais prováveis).

3. Para o terceiro espaço, foi amostrado o tópico 2 ($z_{d^*3} = 2$), do qual foi escolhida aleatoriamente a palavra *organism*.

4. Para o quarto espaço, foi amostrado o tópico 1 ($z_{d^*4} = 1$), do qual foi escolhida aleatoriamente a palavra *genes*.

5. Para o quinto espaço, foi amostrado o tópico 4 ($z_{d^*5} = 4$), do qual foi escolhida aleatoriamente a palavra *numbers*.

6. Para o sexto espaço, foi amostrado o tópico 1 ($z_{d^*6} = 1$), do qual foi escolhida aleatoriamente a palavra *genetic*.

7. Para o sétimo espaço, foi amostrado o tópico 4 ($z_{d^*7} = 4$), do qual foi escolhida aleatoriamente a palavra *predictions* (a qual não é uma das três palavras mais prováveis).

8. Para o oitavo espaço, foi amostrado o tópico 4 ($z_{d^*8} = 4$), do qual foi escolhida aleatoriamente a palavra *computacional*.

9. Para o nono espaço, foi amostrado o tópico 4 ($z_{d^*9} = 4$), do qual foi escolhida aleatoriamente a palavra *computer*.

Na Equação (4.1), definimos $p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})$ a distribuição conjunta dos parâmetros $\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}$ e das observações $\mathbf{w}$ do modelo LDA:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(\mathbf{z}_{dn} \mid \boldsymbol{\theta}_d) p(\mathbf{w}_{dn} \mid \boldsymbol{\beta}_{\boldsymbol{z}_{dn}}) \tag{4.1}$$

onde,

$\boldsymbol{\beta}_k \sim \text{Dirichlet}(\eta_1, \ldots, \eta_V), \text{com } k \in \{1, \ldots, K\}$;

$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K), \text{com } d \in \{1, \ldots, D\}$;

$z_{dn} \sim \text{Categórica}(\boldsymbol{\theta}_d) \iff \boldsymbol{z}_{dn} \sim \text{Multinomial}(1; \boldsymbol{\theta}_d), \text{com } d \in \{1, \ldots, D\} \text{ e } n \in \{1, \ldots, N_d\}$;

$w_{dn} \sim \text{Categórica}(\boldsymbol{\beta}_{z_{dn}}) \iff \boldsymbol{w}_{dn} \sim \text{Multinomial}(1; \boldsymbol{\beta}_{z_{dn}}), \text{com } d \in \{1, \ldots, D\} \text{ e } n \in \{1, \ldots, N_d\}$,

sendo as quantidades $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_V)$ e $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ hiperparâmetros fixados ao elicitar as distribuições a priori de $\boldsymbol{\beta}_k$ e $\boldsymbol{\theta}_d$, respectivamente. No caso particular das implementações realizadas neste trabalho, utiliza-se sempre $\eta_v = 1$ e $\alpha_k = 1$, $\forall\ v \in \{1, \ldots, V\}$ e $\forall\ k \in \{1, \ldots, K\}$. Neste caso, temos distribuições de Dirichlet simétricas, as quais podem ser úteis, por exemplo, quando não há conhecimento prévio que favoreça um componente em relação a outro (distribuições a priori não informativas). Um exemplo disso é a obtenção das distribuições de probabilidade do vocabulário em cada um dos K tópicos ($\boldsymbol{\beta}_k$) e das proporções dos tópicos em cada um dos D documentos que compõem o *corpus* ($\boldsymbol{\theta}_d$). Mais especificamente, quando todas as componentes do vetor paramétrico da distribuição Dirichlet são iguais a 1, a mesma se torna equivalente a uma distribuição uniforme sobre o simplex padrão aberto ($V - 1$ para $\boldsymbol{\beta}_k$, ou $K - 1$ para $\boldsymbol{\theta}_d$). Ou seja, essa distribuição é uniforme sobre todos os pontos em seu suporte. Esta distribuição particular é conhecida como distribuição Dirichlet plana.

Sendo um modelo Bayesiano, temos a distribuição a priori sobre os parâmetros $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z})$ e a função de verossimilhança para os dados observados $\mathbf{w}$. A distribuição a

posteriori, que representa nossa crença sobre $\boldsymbol{\vartheta}$ dado $\mathbf{w}$, é dada pela seguinte expressão:

$$
\begin{aligned}
p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w}) &= \frac{p(\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}) p(\boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\mathbf{z} \mid \boldsymbol{\theta})}{p(\mathbf{w})} \\
&\propto p(\mathbf{w} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}) p(\boldsymbol{\beta}) p(\boldsymbol{\theta}) p(\mathbf{z} \mid \boldsymbol{\theta}) \\
&= p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}) \\
&= \prod_{k=1}^{K} p(\boldsymbol{\beta}_k) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(z_{dn} \mid \boldsymbol{\theta}_d) p(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}}) \\
&\propto \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1} \right] \left[\prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1} \right] \left[\prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \theta_{dk}^{\mathbb{1}(z_{dn}=k)} \right] \\
&\quad \left[\prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)} \right].
\end{aligned}
$$

Calcular a distribuição a posteriori exata $p(\boldsymbol{\vartheta} \mid \mathbf{w})$ é computacionalmente inviável no caso do modelo LDA. Para isso, é necessária a utilização de métodos computacionais.

4.2 MCMC em modelos LDA

Os métodos MCMC se baseiam no algoritmo de Metropolis-Hastings (M-H) desenvolvido por Metropolis et al. [1953] e Hastings [1970], tendo sua popularidade na estatística crescido vertiginosamente nos anos 90 a partir de trabalhos como Müller [1991] e Chib and Greenberg [1995] por exemplo. Em aplicações Bayesianas, esta classe de métodos, conforme mencionado anteriormente, permite criar uma cadeia de Markov que converge para a distribuição à posteriori $p(\boldsymbol{\vartheta} \mid \mathbf{w})$. Um caso particular de algoritmo de M-H é o amostrador de Gibbs. Desenvolvido por Geman and Geman [1984] e Gelfand and Smith [1990], o amostrador de Gibbs define uma cadeia de Markov que sorteia amostras das condicionais completas $p(\vartheta_\ell \mid \boldsymbol{\vartheta}_{-\ell}, \mathbf{w}),\ \ell \in \{1, \ldots, L\}$ em que $\boldsymbol{\vartheta}_{-\ell}$ denota as entradas do vetor $\boldsymbol{\vartheta}$ excetuando-se a ℓ-ésima.

A seguir, descreve-se o cálculo das condicionais completas. Primeiramente, será apresentado o cálculo para a definição da verossimilhança, das distribuições a priori e da

distribuição conjunta (novamente):

Verossimilhança:

$$\begin{aligned} p(\mathbf{w} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}) &= \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{dn} \mid \boldsymbol{\beta}, z_{dn}) \\ &= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \beta_{z_{dn}, w_{dn}} \\ &= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \beta_{k, w_{dn}}^{\mathbb{1}(z_{dn}=k)} \\ &= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{k,v}^{\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)} \end{aligned}$$

Então, $(w_{dn} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, z_{dn}) \sim \text{Categórica}(\boldsymbol{\beta}_{z_{dn}}) \iff (\boldsymbol{w}_{dn} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, z_{dn}) \sim \text{Multinomial}(1; \boldsymbol{\beta}_{z_{dn}})$.

Distribuições a priori:

- $\boldsymbol{\theta}_d = (\theta_{d1}, \ldots, \theta_{dK}) \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$
- $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kV}) \sim Dirichlet(\eta_1, \ldots, \eta_V)$
- $P(z_{dn} = k \mid \boldsymbol{\theta}_d) = \theta_{dk}$, com $k = 1, \ldots, K \iff z_{dn} \sim \text{Categórica}(\boldsymbol{\theta}_d) \iff \boldsymbol{z}_{dn} \sim \text{Multinomial}(1; \boldsymbol{\theta}_d)$

Com as seguintes expressões matemáticas:

- $\boldsymbol{\theta}_d \sim Dirichlet(\alpha_1, \ldots, \alpha_K)$, com a seguinte função de densidade conjunta de $\boldsymbol{\theta}$ de ordem $D \times K$:

$$p(\boldsymbol{\theta}) = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d)$$

$$= \prod_{d=1}^{D} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1}$$
$$= \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \right]^{D} \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1}$$

- $\boldsymbol{\beta}_k \sim Dirichlet(\eta_1, \ldots, \eta_V)$, com a seguinte função de densidade conjunta de $\boldsymbol{\beta}$ de ordem $K \times V$:

$$p(\boldsymbol{\beta}) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k)$$
$$= \prod_{k=1}^{K} \frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1}$$
$$= \left[\frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \right]^{K} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1}$$

- $z_{dn} = k \mid \boldsymbol{\theta}_d \sim$ Categórica$(\boldsymbol{\theta}_d) \iff \boldsymbol{z}_{dn} = \boldsymbol{e}_k \mid \boldsymbol{\theta}_d \sim$ Multinomial$(1; \boldsymbol{\theta}_d)$, onde $\boldsymbol{e}_k$ é o vetor canônico da coordenada k de uma base canônica de dimensão K (com $k \in \{1, \ldots, K\}$), e $z_{dn} = k \mid \boldsymbol{\theta}_d$ com função de massa de probabilidade conjunta de $\mathbf{z} \mid \boldsymbol{\theta}$:

$$p(\mathbf{z} \mid \boldsymbol{\theta}) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} P(z_{dn} = k \mid \boldsymbol{\theta}_d)$$
$$= \prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \theta_{dk}^{\mathbb{1}(z_{dn}=k)},$$

com $k = 1, \ldots, K$.

Cálculo da Distribuição Conjunta:

$$\begin{aligned}
p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) &= p(\mathbf{w} \mid \mathbf{z}, \boldsymbol{\beta}) p(\mathbf{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}) \\
&= \prod_{k=1}^{K} p(\boldsymbol{\beta}_k) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(z_{dn} \mid \boldsymbol{\theta}_d) p(w_{dn} \mid \boldsymbol{\beta}_{\boldsymbol{z}_{dn}}) \\
&\propto \left[\prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1} \right] \left[\prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1} \right] \left[\prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \theta_{dk}^{\mathbb{1}(z_{dn}=k)} \right] \\
&\left[\prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)} \right]
\end{aligned}$$

Cálculo das Condicionais Completas:

- Para $(\boldsymbol{\theta}_d \mid \mathbf{w}, \boldsymbol{\beta}, \mathbf{z})$, temos que:

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{w}, \boldsymbol{\beta}, \mathbf{z}) &\propto \left[\prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{k=1}^{K} \theta_{dk}^{\mathbb{1}(z_{dn}=k)} \right] \left[\prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1} \right] \\
&\propto \left[\prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\sum_{n=1}^{N_d} \mathbb{1}(z_{dn}=k)} \right] \left[\prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1} \right] \\
&\propto \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{dk}^{\sum_{n=1}^{N_d} \mathbb{1}(z_{dn}=k) + \alpha_k - 1}
\end{aligned}$$

logo,

$\boldsymbol{\theta}_d \mid \mathbf{w}, \boldsymbol{\beta}, \mathbf{z} \sim \mathit{Dirichlet}(\sum_{n=1}^{N_d} \mathbb{1}(z_{dn} = 1) + \alpha_1, \ldots, \sum_{n=1}^{N_d} \mathbb{1}(z_{dn} = K) + \alpha_K)$, com $d = 1, \ldots, D$ independentes entre si.

- Para $(\boldsymbol{\beta}_k \mid \boldsymbol{\theta}, \mathbf{w}, \mathbf{z})$, temos que:

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \mathbf{w}, \mathbf{z}) &\propto \left[\prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)}\right]\left[\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\eta_v-1}\right] \\ &\propto \left[\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)}\right]\left[\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\eta_v-1}\right] \\ &\propto \prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)+\eta_v-1} \end{aligned}$$

logo,

$$\begin{aligned} \boldsymbol{\beta}_k \mid \boldsymbol{\theta}, \mathbf{w}, \mathbf{z} \sim Dirichlet(&\textstyle\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=1)+\eta_1, \ldots, \\ &\textstyle\sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=V)+\eta_V), \end{aligned}$$

com $k = 1, \ldots, K$ independentes entre si.

- Para $(z_{dn} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta})$, temos que:

$$\begin{aligned} p(\mathbf{z} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}) &\propto \left[\prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{kv}^{\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)}\right]\left[\prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{k=1}^{K}\theta_{dk}^{\mathbb{1}(z_{dn}=k)}\right] \\ &\propto \prod_{d=1}^{D}\prod_{n=1}^{N_d}\beta_{z_{dn},w_{dn}}\theta_{d,z_{dn}} \end{aligned}$$

logo,

$$P(z_{dn} = k \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}) \propto \beta_{k,w_{dn}}\theta_{dk}$$

então,

$$P(z_{dn} = k \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{\beta_{k,w_{dn}}\theta_{dk}}{\sum_{j=1}^{K}\beta_{j,w_{dn}}\theta_{dj}}, \qquad k = 1, \ldots, K$$

assim,

$$z_{dn} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta} \sim \text{Categórica}\left(\frac{\beta_{1,w_{dn}}\theta_{d1}}{\sum_{j=1}^{K}\beta_{j,w_{dn}}\theta_{dj}}, \ldots, \frac{\beta_{K,w_{dn}}\theta_{dK}}{\sum_{j=1}^{K}\beta_{j,w_{dn}}\theta_{dj}}\right) \iff$$
$$\boldsymbol{z}_{dn} \mid \mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta} \sim \text{Multinomial}\left(1; \frac{\beta_{1,w_{dn}}\theta_{d1}}{\sum_{j=1}^{K}\beta_{j,w_{dn}}\theta_{dj}}, \ldots, \frac{\beta_{K,w_{dn}}\theta_{dK}}{\sum_{j=1}^{K}\beta_{j,w_{dn}}\theta_{dj}}\right).$$

Em resumo, no caso do modelo LDA, as condicionais completas estão disponíveis analiticamente nas Equações (4.2), (4.3) e (4.4).

$$P(z_{dn} = k \mid \boldsymbol{\theta}_d, \boldsymbol{\beta}, w_{dn}) = \frac{\beta_{k,w_{dn}}\theta_{dk}}{\sum_{\ell=1}^{K}\beta_{\ell,w_{dn}}\theta_{d\ell}}, \qquad k = 1, \ldots, K. \tag{4.2}$$

$$p(\boldsymbol{\theta}_d \mid \mathbf{z}_d) = \text{Dirichlet}\left(\alpha_1 + \sum_{n=1}^{N_d}\mathbb{1}(z_{dn} = 1), \ldots, \alpha_K + \sum_{n=1}^{N_d}\mathbb{1}(z_{dn} = K)\right). \tag{4.3}$$

$$p(\boldsymbol{\beta}_k \mid \mathbf{z}, \mathbf{w}) = \text{Dirichlet}\left(\eta_1 + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}(z_{dn} = k, w_{dn} = 1), \ldots, \eta_V + \sum_{d=1}^{D}\sum_{n=1}^{N_d}\mathbb{1}(z_{dn} = k, w_{dn} = V)\right). \tag{4.4}$$

No Algoritmo 4.1 do amostrador de Gibbs para o modelo LDA, se faz uso das Equações (4.2), (4.3) e (4.4).

Algorithm 4.1 Algoritmo do amostrador de Gibbs para o modelo LDA

inicializar $z_{dn}^{(0)} := 1, \forall\, d, n$

inicializar $\theta_{dk}^{(0)} := \dfrac{1}{K}, \forall\, d, k$

inicializar $\beta_{kv}^{(0)} := \dfrac{1}{V}, \forall\, k, v$

fixar $\boldsymbol{\eta} := (\eta_1, \ldots, \eta_V)$, onde $\eta_i > 0$

fixar $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_K)$, onde $\alpha_i > 0$

fixar $T :=$ número de iterações.

tem-se as observações $w_{dn} :=$ (n-ésima palavra do documento d)

repetir para $t = 1$ até T

    para $d = 1$ até D

        para $n = 1$ até N_d

            para $k = 1$ até K

                calcular: $P_k = P : (z_{dn} = k \mid \boldsymbol{\theta}_d^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, w_{dn})$

$$= \frac{\beta_{k,w_{dn}}^{(t-1)} \theta_{dk}^{(t-1)}}{\sum_{\ell=1}^{K} \beta_{\ell,w_{dn}}^{(t-1)} \theta_{d\ell}^{(t-1)}}$$

            obter $z_{dn}^{(t)}$, amostrando um $k \in \{1, \ldots, K\}$, com vetor de probabilidades associado $(P_1, \ldots, P_K)$.

        amostrar: $\boldsymbol{\theta}_d^{(t)} \sim \text{Dirichlet}\Big(\alpha_1 + \sum_{n=1}^{N_d} \mathbb{1}(z_{dn}^{(t)} = 1), \ldots,$

$$\alpha_K + \sum_{n=1}^{N_d} \mathbb{1}(z_{dn}^{(t)} = K)\Big)$$

    para $k = 1$ até K

        amostrar: $\boldsymbol{\beta}_k^{(t)} \sim \text{Dirichlet}\Big(\eta_1 + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{1}(z_{dn}^{(t)} = k, w_{dn} = 1), \ldots,$

$$\eta_V + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{1}(z_{dn} = k, w_{dn} = V)\Big)$$

4.3 Modelos LDA na abordagem variacional

Essa é a propriedade de *mean-field* que define a família variacional que será exibida adiante que assume que as variáveis do modelo na distribuição são independentes, permitindo uma simplificação significativa no processo de inferência. Com essa suposição, a distribuição variacional conjunta dos parâmetros e variáveis latentes pode ser fatorada da seguinte forma:

$$q(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\gamma}) = \prod_{k=1}^{K} q(\boldsymbol{\beta}_k \mid \boldsymbol{\lambda}_k) \left[\prod_{d=1}^{D} q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d) \prod_{n=1}^{N_d} q(\mathbf{z}_{dn} \mid \boldsymbol{\phi}_{dn}) \right]$$

onde, de acordo com a distribuição variacional q,

$\boldsymbol{\beta}_k \overset{q}{\sim} \text{Dirichlet}(\lambda_{k1}, \ldots, \lambda_{kV}), \text{com } k \in \{1, \ldots, K\}$;
$\boldsymbol{\theta}_d \overset{q}{\sim} \text{Dirichlet}(\gamma_{d1}, \ldots, \gamma_{dK}), \text{com } d \in \{1, \ldots, D\}$;
$z_{dn} \overset{q}{\sim} \text{Categórica}(\boldsymbol{\phi}_{dn}) \iff \boldsymbol{z}_{dn} \overset{q}{\sim} \text{Multinomial}(1; \boldsymbol{\phi}_{dn})$, onde $\boldsymbol{\phi}_{dn} = (\phi_{dn1}, \ldots, \phi_{dnK})$, e com $d \in \{1, \ldots, D\}$ e $n \in \{1, \ldots, N_d\}$.

Obtemos a cota inferior de evidência (ELBO) para a preditiva a priori do modelo original com respeito à distribuição variacional $q(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{\theta} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\gamma})$, conforme definido pela Equação (4.5):

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w}) = \mathbb{E}_q \left\{ \log[p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})] - \log[q(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi})] \right\}. \tag{4.5}$$

Logo, temos o problema de otimização na Equação (4.6).

$$\begin{aligned}
(\boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) &= \arg\min_{(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi})} KL\left(q(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}) \parallel p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w})\right) \\
&= \arg\min_{(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi})} \mathbb{E}_q \left\{ \log \left[\frac{q(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} \mid \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi})}{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z} \mid \mathbf{w})} \right] \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \arg\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})} \mathbb{E}_q \left\{\log[q(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})] - \log[p(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \mathbf{w})]\right\} \\
&= \arg\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})} \mathbb{E}_q \left\{\log[q(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})] - \log\left[\frac{p(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z},\mathbf{w})}{p(\mathbf{w})}\right]\right\} \\
&= \arg\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})} \mathbb{E}_q \left\{\log[q(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})] - \log[p(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z},\mathbf{w})] + \cancel{\log p(\mathbf{w})}\right\}.
\end{aligned}
\tag{4.6}
$$

Note que, $\log[p(\mathbf{w})]$ pode ser omitido pelo fato de que em sua expressão analítica não estão contidos os parâmetros variacionais $(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})$. Como foi explicado em parágrafos anteriores, o objetivo da inferência variacional é minimizar a divergência de Kullback-Leibler entre $q_{\boldsymbol{\psi}}(\boldsymbol{\vartheta})$ e $p(\boldsymbol{\vartheta} \mid \mathbf{y})$, e isso equivale à maximizar o ELBO, como é mostrado na Equação (4.7).

$$
\begin{aligned}
(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*,\boldsymbol{\phi}^*) &= \arg\min_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})} \mathbb{E}_q \left\{\log[q(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})] - \log[p(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z},\mathbf{w})]\right\} \\
&= \arg\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})} \underbrace{\mathbb{E}_q \left\{\log[p(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z},\mathbf{w})] - \log[q(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})]\right\}}_{\text{ELBO}:=\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi};\mathbf{w})}.
\end{aligned}
\tag{4.7}
$$

Expandindo os termos que aparecem no cálculo do ELBO na Equação (4.5), temos,

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi};\mathbf{w}) &= \mathbb{E}_q \left\{\log[p(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z},\mathbf{w})] - \log[q(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{z} \mid \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})]\right\} \\
&= \sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\beta}_k)] + \sum_{d=1}^{D} \mathbb{E}_q[\log p(\boldsymbol{\theta}_d)] \\
&+ \sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathbb{E}_q[\log p(z_{dn} \mid \boldsymbol{\theta}_d)] + \sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}})] \\
&- \sum_{k=1}^{K} \mathbb{E}_q[\log q(\boldsymbol{\beta}_k \mid \boldsymbol{\lambda}_k)] - \sum_{d=1}^{D} \mathbb{E}_q[\log q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d)] \\
&- \sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathbb{E}_q[\log q(z_{dn} \mid \boldsymbol{\phi}_{dn})].
\end{aligned}
$$

Para fazer o cálculo da maximização do ELBO e obter,

$$
(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*,\boldsymbol{\phi}^*) = \arg\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})} \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi};\mathbf{w}),
$$

vamos calcular as derivadas parciais com respeito aos parâmetros variacionais,

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \lambda_{kv}} = \frac{\partial}{\partial \lambda_{kv}} \left\{ \sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\beta}_k)] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}})] - \sum_{k=1}^{K} \mathbb{E}_q[\log q(\boldsymbol{\beta}_k \mid \boldsymbol{\lambda}_k)] \right\} \tag{4.8}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \gamma_{dk}} = \frac{\partial}{\partial \gamma_{dk}} \left\{ \sum_{d=1}^{D} \mathbb{E}_q[\log p(\boldsymbol{\theta}_d)] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(z_{dn} \mid \boldsymbol{\theta}_d)] - \sum_{d=1}^{D} \mathbb{E}_q[\log q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d)] \right\} \tag{4.9}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \phi_{dnk}} = \frac{\partial}{\partial \phi_{dnk}} \left\{ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(z_{dn} \mid \boldsymbol{\theta}_d)] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}})] - \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log q(z_{dn} \mid \boldsymbol{\phi}_{dn})] \right\} \tag{4.10}$$

Nos cálculos a seguir, para a obtenção das equações de atualização dos parâmetros variacionais, é importante ter claro o seguinte:

1. Alguns termos e suas somatórias vão sumir da expressão, por serem constantes no cálculo das derivadas parciais

2. Seja $\mathbf{X} = (X_1, \ldots, X_K)$ um vetor aleatório tal que $\mathbf{x} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, onde $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ é o vetor de parâmetros positivos da distribuição de Dirichlet; então a $\mathbb{E}[\log(X_i)] = \Psi(\alpha_i) - \Psi\left(\sum_{i=1}^{K} \alpha_i\right)$, onde $\Psi(\cdot)$ é a função digama.

3. Seja Y uma variável com $y > 0$, então $\dfrac{d}{dy}\Psi(y) = \Psi'(y)$, onde $\Psi'(\cdot)$ é a função trigama.

Agora, desenvolvendo cada uma das Equações (4.8), (4.9), (4.10) temos:

Equação (4.8):

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \lambda_{kv}} = \frac{\partial}{\partial \lambda_{kv}} \left\{ \sum_{k=1}^{K} \mathbb{E}_q[\log p(\boldsymbol{\beta}_k)] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}})]$$

$$
\begin{aligned}
&- \sum_{k=1}^{K} \mathbb{E}_q[\log q(\boldsymbol{\beta}_k \mid \boldsymbol{\lambda}_k)] \Bigg\} \\
&= \frac{\partial}{\partial \lambda_{kv}} \Bigg\{ \sum_{k=1}^{K} \mathbb{E}_q \left[\log \left[\frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1} \right] \right] \\
&+ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log \beta_{z_{dn} w_{dn}}] \\
&- \sum_{k=1}^{K} \mathbb{E}_q \left[\log \left[\frac{\Gamma(\sum_{v=1}^{V} \lambda_{kv})}{\prod_{v=1}^{V} \Gamma(\lambda_{kv})} \prod_{v=1}^{V} \beta_{kv}^{\lambda_{kv} - 1} \right] \right] \Bigg\} \\
&= \frac{\partial}{\partial \lambda_{kv}} \Bigg\{ \sum_{k=1}^{K} \mathbb{E}_q \left[\log \left[\frac{\Gamma(\sum_{v=1}^{V} \eta_v)}{\prod_{v=1}^{V} \Gamma(\eta_v)} \prod_{v=1}^{V} \beta_{kv}^{\eta_v - 1} \right] \right] \\
&+ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log \beta_{kv}^{\mathbb{1}(z_{dn}=k)\mathbb{1}(w_{dn}=v)}] \\
&- \sum_{k=1}^{K} \mathbb{E}_q \left[\log \left[\frac{\Gamma(\sum_{v=1}^{V} \lambda_{kv})}{\prod_{v=1}^{V} \Gamma(\lambda_{kv})} \prod_{v=1}^{V} \beta_{kv}^{\lambda_{kv} - 1} \right] \right] \Bigg\} \\
&= \frac{\partial}{\partial \lambda_{kv}} \Bigg\{ \sum_{v=1}^{V} (\eta_v - 1) \mathbb{E}_q[\log(\beta_{kv})] \\
&+ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\mathbb{1}(z_{dn} = k)] \mathbb{E}_q[\mathbb{1}(w_{dn} = v)] \mathbb{E}_q[\log \beta_{kv}] \\
&- \log[\Gamma(\sum_{v=1}^{V} \lambda_{kv})] + \sum_{v=1}^{V} \log[\Gamma(\lambda_{kv})] - \sum_{v=1}^{V} (\lambda_{kv} - 1) \mathbb{E}_q[\log(\beta_{kv})] \Bigg\} \\
&= \frac{\partial}{\partial \lambda_{kv}} \Bigg\{ \sum_{v=1}^{V} (\eta_v - 1)[\Psi(\lambda_{kv}) - \Psi(\sum_{l=1}^{V} \lambda_{kl})] \\
&+ \sum_{d=1}^{D} \sum_{n=1}^{N_d} [\mathbb{1}(w_{dn} = v) P_q(z_{dn} = k)[\Psi(\lambda_{kv}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})]] \\
&- \log[\Gamma(\sum_{v=1}^{V} \lambda_{kv})] + \sum_{v=1}^{V} \log[\Gamma(\lambda_{kv})] - \sum_{v=1}^{V} (\lambda_{kv} - 1)[\Psi(\lambda_{kv}) - \Psi(\sum_{l=1}^{V} \lambda_{kl})] \Bigg\} \\
&= (\eta_v - 1)[\Psi'(\lambda_{kv}) - \Psi'(\sum_{v=1}^{V} \lambda_{kv})] \\
&+ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \mathbb{1}(w_{dn} = v)[\Psi'(\lambda_{kv}) - \Psi'(\sum_{v=1}^{V} \lambda_{kv})]
\end{aligned}
$$

$$
\begin{aligned}
&- \Psi(\sum_{v=1}^{V} \lambda_{kv}) + \Psi(\lambda_{kv}) \\
&- \left\{ (\lambda_{kv} - 1)[\Psi'(\lambda_{kv}) - \Psi'(\sum_{v=1}^{V} \lambda_{kv})] + [\Psi(\lambda_{kv}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] \right\} \\
&= [\Psi'(\lambda_{kv}) - \Psi'(\sum_{v=1}^{V} \lambda_{kv})] \left\{ (\eta_v - 1) - (\lambda_{kv} - 1) + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \mathbb{1}(w_{dn} = v) \right\} \\
&= [\Psi'(\lambda_{kv}) - \Psi'(\sum_{v=1}^{V} \lambda_{kv})] \left\{ \eta_v - \lambda_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \mathbb{1}(w_{dn} = v) \right\}.
\end{aligned}
$$

Agora igualamos esse resultado à zero,

$$
\begin{aligned}
&\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \lambda_{kv}} = 0 \\
\Rightarrow &[\Psi'(\lambda_{kv}) - \Psi'(\sum_{v=1}^{V} \lambda_{kv})] \left\{ \eta_v - \lambda_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \mathbb{1}(w_{dn} = v) \right\} = 0 \\
\Rightarrow &\lambda_{kv} = \eta_v + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \mathbb{1}(w_{dn} = v).
\end{aligned}
$$

Equação (4.9):

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \gamma_{dk}} &= \frac{\partial}{\partial \gamma_{dk}} \left\{ \sum_{d=1}^{D} \mathbb{E}_q[\log p(\boldsymbol{\theta}_d)] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(z_{dn} \mid \boldsymbol{\theta}_d)] \right. \\
&\quad \left. - \sum_{d=1}^{D} \mathbb{E}_q[\log q(\boldsymbol{\theta}_d \mid \boldsymbol{\gamma}_d)] \right\} \\
&= \frac{\partial}{\partial \gamma_{dk}} \left\{ \sum_{d=1}^{D} \mathbb{E}_q \left[\log \left[\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{dk}^{\alpha_k - 1} \right] \right] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log(\theta_{dz_{dn}})] \right. \\
&\quad \left. - \sum_{d=1}^{D} \mathbb{E}_q \left[\log \left[\frac{\Gamma(\sum_{k=1}^{K} \gamma_{dk})}{\prod_{k=1}^{K} \Gamma(\gamma_{dk})} \prod_{k=1}^{K} \theta_{dk}^{\gamma_{dk} - 1} \right] \right] \right\} \\
&= \frac{\partial}{\partial \gamma_{dk}} \left\{ (\alpha_k - 1) \mathbb{E}_q(\log \theta_{dk}) + \sum_{n=1}^{N_d} \mathbb{E}_q[\mathbb{1}(z_{dn} = k)] \mathbb{E}_q(\log \theta_{dk}) \right. \\
&\quad \left. - \left\{ \mathbb{E}_q[\log \Gamma(\sum_{k=1}^{K} \gamma_{dk})] - \mathbb{E}_q[\log \Gamma(\gamma_{dk})] + (\gamma_{dk} - 1) \mathbb{E}_q[\log \theta_{dk}] \right\} \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \gamma_{dk}} \Bigg\{ (\alpha_k - 1)[\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] + \sum_{n=1}^{N_d} \phi_{dnk}[\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] \\
&- \log \Gamma(\sum_{k=1}^{K} \gamma_{dk}) + \log \Gamma(\gamma_{dk}) - (\gamma_{dk} - 1)[\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] \Bigg\} \\
&= (\alpha_k - 1)[\Psi'(\gamma_{dk}) - \Psi'(\sum_{k=1}^{K} \gamma_{dk})] + \sum_{n=1}^{N_d} \phi_{dnk}[\Psi'(\gamma_{dk}) - \Psi'(\sum_{k=1}^{K} \gamma_{dk})] \\
&- \Psi(\sum_{k=1}^{K} \gamma_{dk}) + \Psi(\gamma_{dk}) - [\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] \\
&- (\gamma_{dk} - 1)[\Psi'(\gamma_{dk}) - \Psi'(\sum_{k=1}^{K} \gamma_{dk})] \\
&= [\Psi'(\gamma_{dk}) - \Psi'(\sum_{k=1}^{K} \gamma_{dk})][(\alpha_k - 1) + \sum_{n=1}^{N_d} \phi_{dnk} - (\gamma_{dk} - 1)].
\end{aligned}
$$

Agora, igualando o anterior resultado à zero,

$$
\begin{aligned}
&\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \gamma_{dk}} = 0 \\
\Rightarrow &[\Psi'(\gamma_{dk}) - \Psi'(\sum_{k=1}^{K} \gamma_{dk})][(\alpha_k - 1) + \sum_{n=1}^{N_d} \phi_{dnk} - (\gamma_{dk} - 1)] = 0 \\
\Rightarrow &\gamma_{dk} = \alpha_k + \sum_{n=1}^{N_d} \phi_{dnk}.
\end{aligned}
$$

Equação (4.10):

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \phi_{dnk}} &= \frac{\partial}{\partial \phi_{dnk}} \Bigg\{ \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(z_{dn} \mid \boldsymbol{\theta}_d)] + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log p(w_{dn} \mid \boldsymbol{\beta}_{z_{dn}})] \\
&- \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}_q[\log q(z_{dn} \mid \boldsymbol{\phi}_{dn})] \Bigg\} \\
&= \frac{\partial}{\partial \phi_{dnk}} \{\mathbb{E}_q[\log(\theta_{dz_{dn}})] + \mathbb{E}_q[\log(\beta_{z_{dn} w_{dn}})] - \mathbb{E}_q[\log(\phi_{dnz_{dn}})]\} \\
&= \frac{\partial}{\partial \phi_{dnk}} \Bigg\{ \phi_{dnk}[\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] \\
&+ \phi_{dnk}[\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] - \phi_{dnk} \log(\phi_{dnk}) \Bigg\}
\end{aligned}
$$

$$= [\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] - \log(\phi_{dnk}) - \frac{\phi_{dnk}}{\phi_{dnk}}.$$

Agora, igualamos esse resultado à zero,

$$\begin{aligned}
&\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \phi_{dnk}} = 0 \\
\Rightarrow &[\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] - \log(\phi_{dnk}) - \frac{\phi_{dnk}}{\phi_{dnk}} = 0 \\
\Rightarrow &\log(\phi_{dnk}) = [\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] - 1 \\
\Rightarrow &\phi_{dnk} = \exp\left\{ [\Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk})] + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] - 1 \right\} \\
\Rightarrow &\phi_{dnk} \propto \exp\left\{ \Psi(\gamma_{dk}) + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] \right\}.
\end{aligned}$$

Depois de fazer o cálculo do $\arg\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\phi})}$ do ELBO, obtemos as Equações (4.11), (4.12) e (4.13) de atualização, para os parâmetros variacionais.

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \lambda_{kv}} = 0 \Rightarrow \lambda_{kv} = \eta_v + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk} \mathbb{1}(w_{dn} = v), \tag{4.11}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \gamma_{dk}} = 0 \Rightarrow \gamma_{dk} = \alpha_k + \sum_{n=1}^{N_d} \phi_{dnk}, \tag{4.12}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\phi}; \mathbf{w})}{\partial \phi_{dnk}} = 0 \Rightarrow \phi_{dnk} \propto \exp\left\{ \Psi(\gamma_{dk}) + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})] \right\}. \tag{4.13}$$

As Equações (4.11), (4.12) e (4.13) de atualização dos parâmetros variacionais, replicam a estrutura das distribuições condicionais completas utilizadas no amostrador de Gibbs. No entanto, ao invés de gerar amostras aleatórias, o método variacional ajusta distribuições aproximadas de forma determinística, minimizando a divergência de KL entre a distribuição variacional e a distribuição a posteriori. Essa correspondência estrutural evidencia que a inferência variacional pode ser interpretada como uma versão determinística

do amostrador de Gibbs, especialmente em modelos LDA cujas distribuições condicionais completas tem forma fechada conhecida.

No Algoritmo 4.2 de inferência variacional para o modelo LDA, portanto consiste em atualizar iterativamente as Equações (4.11), (4.12) e (4.13), até alcançar um critério de convergência pré estabelecido, o qual consiste em observar a log-verossimilhança do modelo aproximado em cada uma das iterações, até não observar uma melhora significativa.

Algorithm 4.2 Algoritmo da inferência variacional para o modelo LDA

inicializar $\phi_{dnk}^{(0)} := \frac{1}{K}, \forall\, d, n, k$
inicializar $\gamma_{dk}^{(0)} := \alpha_k + \frac{N_d}{K}, \forall\, d, k$
inicializar $\lambda_{kv}^{(0)} := \eta_v + \frac{\sum_{d=1}^{D} N_d}{V}, \forall\, k, v$
fixar $\kappa = 0,8$
fixar $\tau = 1$
repetir
calcular $\rho^{(t)} = (t+\tau)^{-\kappa}$
    para $d = 1$ até D
        para $n = 1$ até N_d
            para $k = 1$ até K
                $\phi_{dnk}^{(t)} \propto \exp\left\{\Psi(\gamma_{dk}) + [\Psi(\lambda_{kw_{dn}}) - \Psi(\sum_{v=1}^{V} \lambda_{kv})]\right\}$
            normalizar $\boldsymbol{\phi}_{dn}^{(t)}$ para que a soma das suas componentes seja 1.
        para $k = 1$ até K
            $\gamma_{dk}^{(t)} := \alpha_k + \sum_{n=1}^{N_d} \phi_{dnk}^{(t)}$
    para $k = 1$ até K
        para $v = 1$ até V
            $\lambda = \eta_v + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dnk}^{(t)} \mathbb{1}(w_{dn} = v)$
            $\lambda_{kv}^{(t)} = (1-\rho_t)\lambda_{kv}^{(t-1)} + \rho_t \lambda$
até convergir.

As iterações no Algoritmo 4.2 baseiam-se no CAVI, como foi explicado na secção 3.2. No mesmo Algoritmo, observa-se também que $\rho_t = (t+\tau)^{-\kappa}$ [Hoffman et al., 2013], onde ρ_t é o tamanho do passo (*step-size*, no inglês) na iteração t, $\kappa = 0,8$ é a taxa de esquecimento que controla com que rapidez as informações antigas são esquecidas, e $\tau = 1$ é o atraso que diminui o peso das iterações iniciais.

O uso de ρ_t é essencial na atualização dos parâmetros variacionais globais λ_{kv}, segundo

a equação $\lambda_{kv}^{(t)} = (1-\rho_t)\lambda_{kv}^{(t-1)} + \rho_t \lambda$ (contida no Algoritmo 4.2), pois permite realizar uma média ponderada entre o valor anterior de λ_{kv} e uma nova estimativa baseada em uma subamostra dos dados. Essa abordagem suaviza as flutuações nas atualizações decorrentes da aleatoriedade da subamostragem, garantindo uma convergência mais estável ao longo das iterações. Além disso, o decaimento gradual do tamanho do passo assegura que o algoritmo satisfaça as condições de convergência da otimização estocástica.

Capítulo 5

Resultados

Neste capítulo apresenta-se os resultados obtidos a partir da implementação feita na linguagem de programação R [R Core Team, 2023] dos algoritmos MCMC e inferência variacional, como descritos no capítulo 4, para inferência no modelo LDA aplicado à base de dados descrita no capítulo 2. Discute-se em especial os tópicos latentes aprendidos pelo modelo, a escolha do número de tópicos K e as medidas de coerência dos tópicos, em cada abordagem. Na Seção 5.3, apresentam-se os tempos computacionais requeridos para a execução dos algoritmos de inferência utilizados no treinamento do modelo LDA, bem como as especificações do computador empregado nesse processo.

5.1 Abordagem via MCMC

Inicialmente foram simuladas cinco cadeias para implementação do amostrador de Gibbs aos modelos LDA conforme no descrito no Algoritmo 4.1 cada uma considerando um número de tópicos $K \in \{10, 20, 30, 40, 50\}$, de acordo com as Equações (4.2), (4.3) e (4.4). Foi estipulado o número de iterações igual a 10000, o *burn-in* ou *período de aquecimento* igual a 2500, e o espaçamento igual a 5. Após a simulação das cadeias, obtém-se portanto amostras aproximadas de tamanho 1500 das distribuições a posteriori.

Utilizando as amostras aproximadas das distribuições a posteriori conjuntas de $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ para cada um dos modelos LDA já mencionados, calculam-se os critérios de informação

cujos valores são apresentados na Tabela 5.1. Critério de Informação de Akaike (AIC) [Akaike, 1973], Critério de Informação Bayesiano (BIC) [Schwarz, 1978], Critério de Informação de Deviância (DIC) [Spiegelhalter et al., 2002], e Critério de Informação de Watanabe-Akaike (WAIC) [Watanabe, 2010]; medem a qualidade de um modelo, assumindo que existe um modelo verdadeiro teórico e que qualquer modelo estimado perderá parte do conteúdo informacional [Chen et al., 2021].

Tabela 5.1: Valores calculados dos critérios de informação AIC, BIC, DIC e WAIC para os modelos LDA com 10, 20, 30, 40 e 50 tópicos. Quanto menores os valores, maiores as evidencias à favor do modelo.

	AIC	BIC	DIC	WAIC
Modelo com 10 tópicos	$\mathbf{948724,9}$	$\mathbf{1521978}$	$822776,9$	$3893,969$
Modelo com 20 tópicos	1048660	2195165	$795978,9$	$3638,251$
Modelo com 30 tópicos	1157044	2876801	$777419,9$	$3466,014$
Modelo com 40 tópicos	1273188	3566198	$766564,3$	$3367,382$
Modelo com 50 tópicos	1389148	4255411	$\mathbf{755534,4}$	$\mathbf{3272,519}$

Observa-se na Tabela 5.1 que os critérios AIC e BIC tendem a escolher modelos com menos tópicos ($K = 10$); ao contrário do comportamento nos critérios DIC e WAIC, que tendem a escolher um número de tópicos maior nos modelos LDA. Não observa-se portanto uma concordância entre os quatro critérios.

Visando escolha de um modelo parcimonioso, considerou-se mais pertinente a utilização do AIC e BIC. Da Tabela 5.1, temos indícios que o melhor modelo pode ter número de tópicos próximo de 10 e, por isso, aplica-se MCMC para diferentes números de tópicos, i.e., $K \in \{2, \ldots, 15\} \setminus \{10\}$. Uma vez aprendidos os parâmetros $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ para os novos modelos LDA, também calcula-se os critérios de informação (AIC, BIC, DIC e WAIC), cujos valores consolidados amostram-se na Tabela A.1 que encontra-se no Apêndice. Observa-se na Tabela A.1 o mesmo comportamento que com os modelos treinados inicialmente, sem observar concordância entre os diferentes critérios quanto à escolha do número de tópicos (a escolha do número tópicos para cada critério seria: AIC: $K = 2$, BIC: $K = 2$, DIC: $K = 50$, e WAIC: $K = 50$).

Por fim, ainda buscando concordância entre os critérios, foram retiradas do conjunto de dados as palavras com frequência total entre 1 e 3 na base de dados (nota-se na Figura 5.1 a distribuição da frequência das palavras nas observações da base de dados), uma vez que essas palavras não devem exercer influência na análise dos tópicos nos modelos LDA. Neste procedimento, foram retiradas 3822 palavras do vocabulário (V) ficando com 2113; e correspondentes a retirar 5630 observações, restando 55745 na base de dados. Os critérios de informação foram refeitos visando reduzir o grau de penalização no cálculos do AIC e BIC.

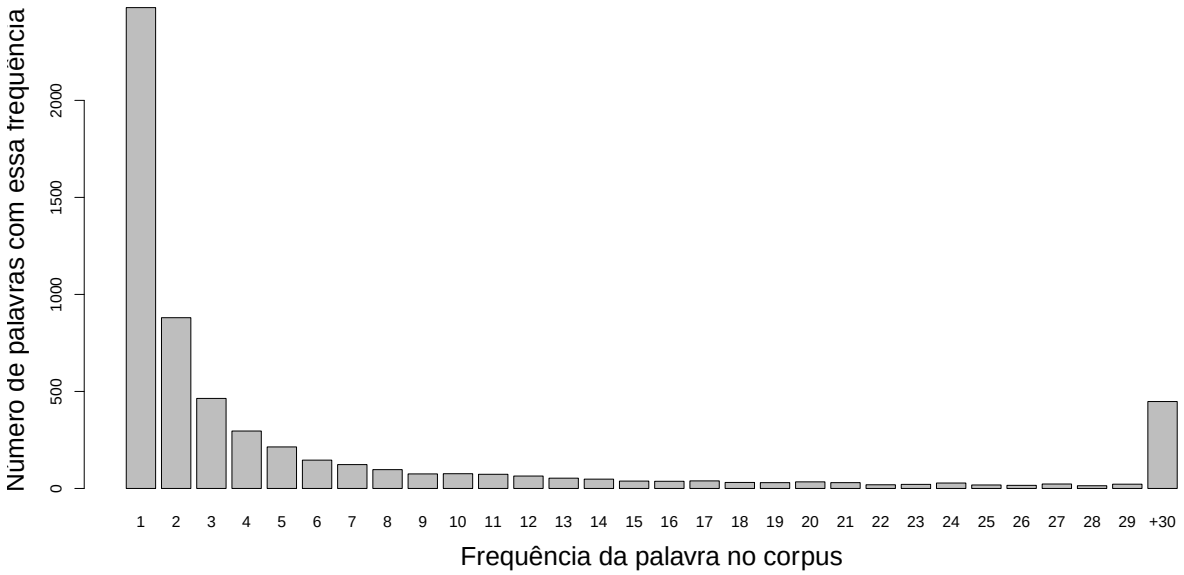


Figura 5.1: Distribuição da frequência das palavras no total de documentos da base de dados. A última barra corresponde a palavras com frequência absoluta igual ou superior a 30 ao longo de toda a base de dados.

Na Figura 5.2, tem-se o novo cálculo do AIC (uma vez que este critério aplica uma menor penalização que o BIC, DIC e WAIC) para os modelos LDA já treinados. Observa-se que os modelos LDA com $K = 6$ até $K = 15$, são identificados pelo AIC como os modelos mais adequados. Entretanto, intuitivamente, espera-se que existam consideravelmente mais do que 6 tópicos latentes presentes nas dissertações de mestrado em estatística publicadas nos últimos anos no Brasil. Portanto, considerou-se que a utilização dos critérios

de informação para seleção de K neste contexto não forneceu uma indicação clara e definitiva. Em consequência, foram consideradas outras métricas para seleção de modelos na base de dados analisada, baseada na perplexidade e na coerência dos tópicos estimados [Blei et al., 2003; Röder et al., 2015].

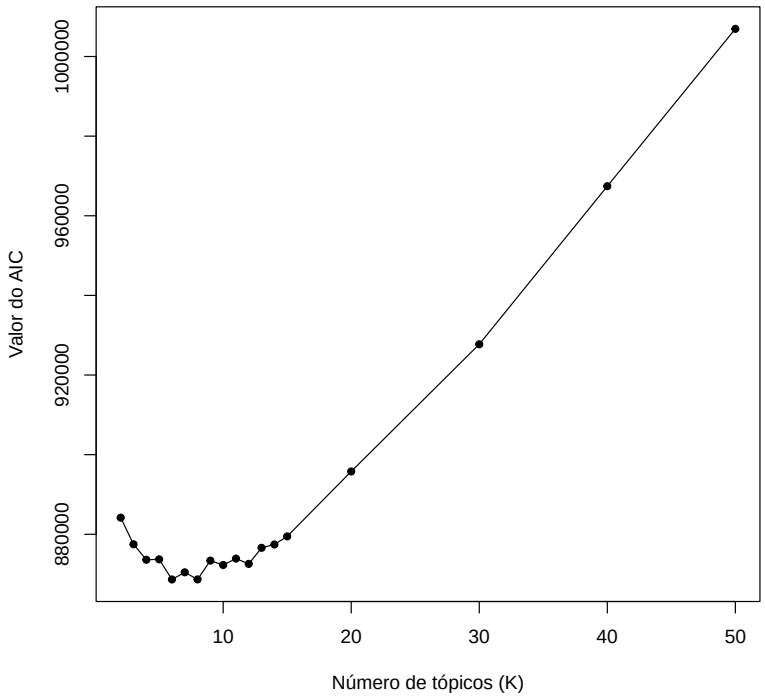


Figura 5.2: Cálculo do AIC dos modelos LDA, tirando as palavras do vocabulário com frequência igual desde 1 até 3.

Calculamos a perplexidade dos modelos treinados. A perplexidade, usada por convenção em modelagem de textos, é uma função monotonamente decrescente da verossimilhança dos dados de teste, sendo algebricamente equivalente ao inverso da média geométrica da verossimilhança por palavra. Um valor mais baixo de perplexidade indica melhor desempenho de generalização [Blei et al., 2003]. Mais formalmente, para um conjunto de teste (mesmo conjunto de dados usado para o treinamento dos modelos, com variação no número de tópicos K; ou seja, não foi feita a divisão dos dados em conjuntos de treino e teste ou um *train-test split*) de D documentos, a perplexidade é dada pela

Equação (5.1).

$$perplexidade(D_{teste}) = \exp\left[-\frac{\sum_{d=1}^{D}\log p(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d}\right], \tag{5.1}$$

onde $\mathbf{w}_d = (\mathrm{w}_{d1}, \ldots, \mathrm{w}_{dN_d})$ representa o vetor de palavras (com repetição) observadas no documento $d \in \{1, \ldots, D\}$. Uma das razões para não ter sido realizada a divisão entre treino e teste no conjunto de dados, foi o baixo número de palavras presente em cada *abstract* das dissertações que compõem o *corpus*. No entanto, ao proceder dessa forma, a perplexidade passa a medir a capacidade do modelo de memorizar os dados, e não de generalizar.

A Figura 5.3 contém os valores da perplexidade calculados para cada um dos modelos LDA treinados. O modelo LDA com 50 tópicos se apresenta como o de menor valor da perplexidade (seguido pelo modelo com 40 tópicos), o que indica que o modelo associa, uma alta probabilidade preditiva sobre as palavras observadas nos documentos.

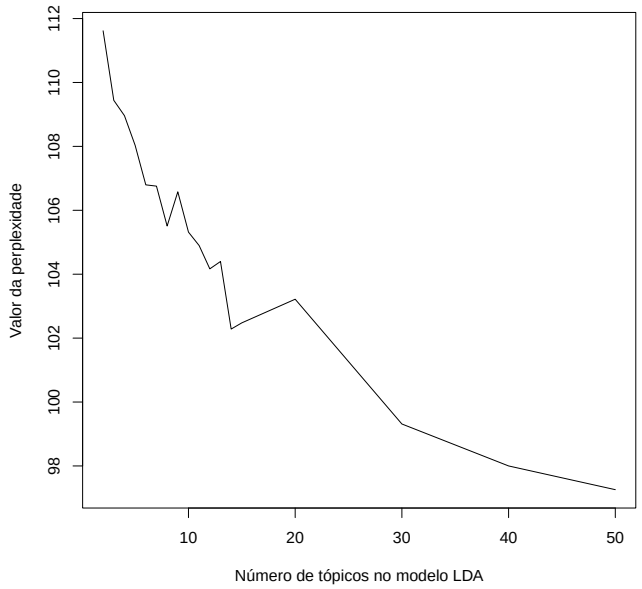


Figura 5.3: Cálculo da medida da perplexidade nos modelos LDA treinados.

A perplexidade baixa não necessariamente significa que os tópicos são interpretáveis

ou coerentes. Isso ocorre porque a perplexidade pode ser influenciada por fatores como a escolha do número de tópicos, a quantidade de dados e a distribuição de palavras. Em algumas situações, é possível que um modelo com perplexidade baixa produza tópicos que não fazem sentido. Por isso, é recomendado usar a perplexidade em conjunto com outras avaliações qualitativas, como a coerência dos tópicos, para garantir que o modelo LDA seja não apenas estatisticamente bom, mas também útil e interpretável.

O cálculo da Medida de Coerência (C_{UCI}) para um dado tópico $k \in \{1, \ldots, K\}$ no modelo LDA [Röder et al., 2015] é dado pela Equação (5.2).

$$C_{UCI}^{k} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{PMI}_k(w_i, w_j), \tag{5.2}$$

onde $\mathrm{PMI}_k(w_i, w_j) = log\frac{P_k(w_i, w_j) + \epsilon}{P_k(w_i)P_k(w_j)}$. Define-se o PMI (do inglês *Pointwise Mutual Information*), como relação estatística entre duas palavras, avaliando se elas ocorrem juntas mais frequentemente do que seria esperado por acaso. Os termos w_i e w_j, são as palavras no conjunto das 15 melhores palavras (N) em cada tópico (k) que estamos considerando para medir a coerência, com $i, j \in \{1, \ldots, 15\}$ tais que $i \neq j$; $P_k(w_i)$ e $P_k(w_j)$, são as probabilidades marginais de que as palavras w_i e w_j ocorram nos documentos do *corpus*; e $P_k(w_i, w_j)$, é a probabilidade conjunta de que as palavras w_i e w_j ocorram nos documentos do *corpus*, geralmente, refere-se à probabilidade de que ambas as palavras apareçam juntas nesses documentos do *corpus*.

A Figura 5.4 contém a média das C_{UCI}^{k} calculadas nos tópicos de cada um dos modelos LDA treinados. O modelo LDA com 40 tópicos se apresenta como o de maior coerência media dos tópicos.

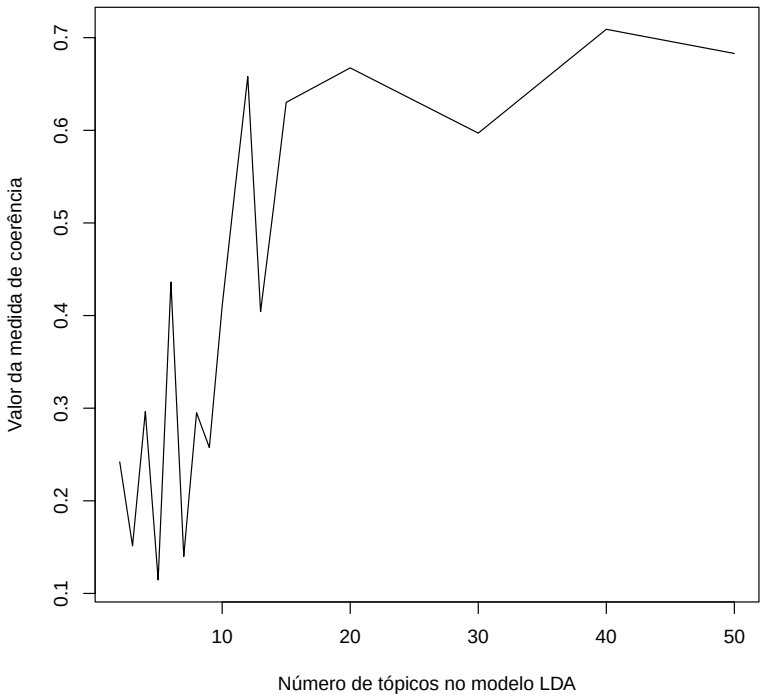


Figura 5.4: Cálculo da medida de coerência (C_{UCI}) nos tópicos dos modelos LDA treinados via MCMC.

Também, observa-se na Figura 5.5 as distribuições por meio de diagramas de caixa, das medidas de coerência (C_{UCI}) calculadas para cada um dos tópicos nos diferentes modelos LDA treinados. Novamente, o modelo LDA com 40 tópicos, apresenta a maior mediana, sem pontos atípicos, mas não tendo o máximo valor global (em comparação com os demais modelos LDA) nos seus valores de C_{UCI}.

Finalmente, escolhe-se o modelo LDA com 40 tópicos ($K = 40$), a partir do qual constrói-se a inferência acerca dos tópicos latentes a partir do vetor $\boldsymbol{\beta}$ à posteriori. A Tabela 5.2 apresenta as 15 palavras mais relevantes de cada tópico com respeito à pontuação do termo (*term-score*, em inglês) [Blei, 2009], definida como,

$$\text{term-score}_{k,v} = \hat{\beta}_{k,v} \log \left(\frac{\hat{\beta}_{k,v}}{\left(\prod_{j=1}^{K} \hat{\beta}_{j,v} \right) \frac{1}{K}} \right)$$
$$= \hat{\beta}_{k,v} \left[\log(\hat{\beta}_{k,v}) - \log \left(\prod_{j=1}^{K} \hat{\beta}_{j,v} \right) \frac{1}{K} \right].$$

A pontuação da v-ésima palavra no tópico k ($\text{term-score}_{k,v}$) é portanto composta pela sua probabilidade a posteriori no tópico k ($\hat{\beta}_{k,v}$), ponderada pela diferença, em escala logarítmica, da mesma probabilidade a posteriori e a média geométrica da probabilidade a posteriori desse termo (v) em todos os tópicos $\left(\log(\hat{\beta}_{k,v}) - \log \left(\prod_{j=1}^{K} \hat{\beta}_{j,v} \right) \frac{1}{K} \right)$. A pontuação do termo penaliza portanto palavras comuns a todos os tópicos simultaneamente.

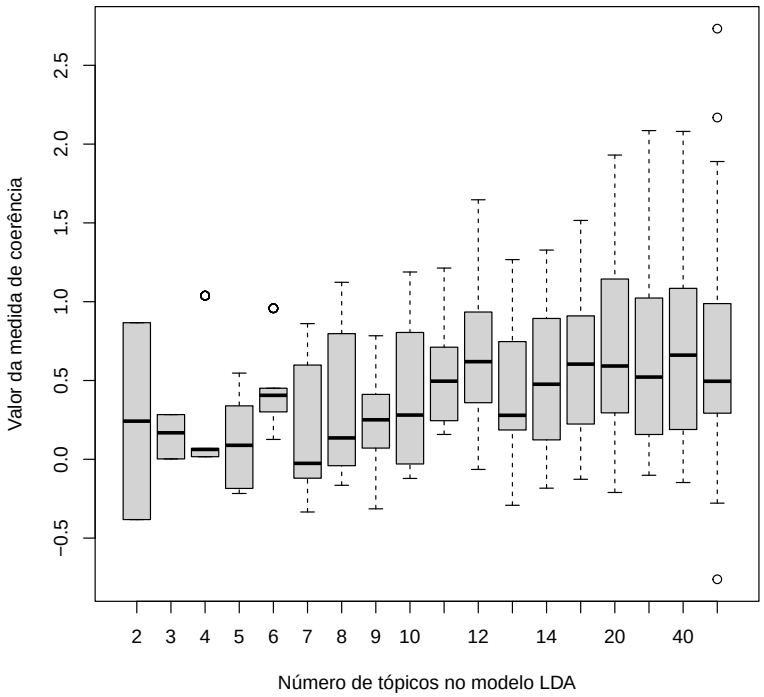


Figura 5.5: Distribuições da medida de coerência (C_{UCI}) nos tópicos dos modelos LDA treinados.

Tabela 5.2: 15 palavras mais prováveis de cada tópico, respeito à pontuação do termo [Blei, 2009], na abordagem via MCMC.

Tópico 1		Tópico 2		Tópico 3		Tópico 4		Tópico 5		Tópico 6		Tópico 7		Tópico 8		Tópico 9		Tópico 10	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
network	0.171	model	0.101	algebra	0.212	ccc	0.052	batch	0.168	genom	0.141	conflict	0.142	enem	0.071	price	0.215	estatística	0.091
neural	0.132	joint	0.088	automorph	0.114	mix	0.037	code	0.058	select	0.109	prison	0.117	insur	0.07	coffe	0.15	group	0.091
artifici	0.066	portfolio	0.049	varieti	0.097	statist	0.037	macaé	0.058	gev	0.088	stabil	0.069	notebook	0.065	commod	0.115	são	0.088
neg	0.065	copula	0.045	free	0.08	urban	0.035	process	0.038	marker	0.088	space	0.052	rural	0.057	sugar	0.09	tourism	0.07
nois	0.064	viscoelast	0.042	porou	0.07	estatística	0.034	gam	0.036	extrem	0.081	matrix	0.045	individu	0.049	market	0.082	sanit	0.056
binomi	0.058	estim	0.04	flow	0.068	são	0.033	weibul	0.034	predict	0.072	resolut	0.045	item	0.047	ammi	0.081	ciência	0.054
machin	0.055	garch	0.04	rehabilit	0.066	cross	0.032	monitor	0.033	trait	0.069	ecumen	0.044	evalu	0.046	soi	0.069	computação	0.048
learn	0.053	surviv	0.038	categori	0.054	mobil	0.032	urban	0.033	prior	0.056	optim	0.044	difficulti	0.039	séri	0.064	graduação	0.047
inflat	0.048	student	0.036	finit	0.048	sustain	0.032	chart	0.03	blup	0.054	africa	0.042	differenti	0.037	crisi	0.06	pó	0.047
wavelet	0.045	flow	0.034	low	0.046	effect	0.031	arma	0.028	snp	0.049	system	0.042	user	0.036	agricultur	0.058	domest	0.046
count	0.043	robust	0.034	prematur	0.042	model	0.031	gener	0.028	maximum	0.045	logic	0.041	candid	0.035	return	0.056	mestrado	0.045
poisson	0.04	fluid	0.032	discharg	0.041	scale	0.031	data	0.027	structur	0.045	territori	0.041	nonrespons	0.035	transfer	0.056	programa	0.045
techniqu	0.036	persist	0.031	group	0.041	search	0.031	privaci	0.027	genet	0.043	cut	0.039	equiti	0.034	chicken	0.051	commun	0.044
complex	0.035	conform	0.03	medium	0.041	dgp	0.03	approach	0.026	estim	0.039	prefer	0.037	apt	0.028	cattl	0.049	flow	0.044
bell	0.033	longitudin	0.03	strongli	0.041	select	0.03	time	0.026	latent	0.039	littl	0.036	exam	0.028	job	0.047	estim	0.043

Tópico 11		Tópico 12		Tópico 13		Tópico 14		Tópico 15		Tópico 16		Tópico 17		Tópico 18		Tópico 19		Tópico 20	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
opm	0.102	seri	0.139	numer	1.067	sampl	0.155	chang	0.137	miss	0.147	normal	0.065	death	0.149	diet	0.102	genotyp	0.084
synthet	0.085	quantil	0.109	solut	0.128	test	0.07	seri	0.086	inpaint	0.082	segment	0.065	color	0.086	adjust	0.044	chart	0.069
beta	0.075	arma	0.104	equat	0.079	size	0.062	biom	0.074	imag	0.081	coordin	0.058	fruit	0.068	eeg	0.042	control	0.064
workshop	0.075	time	0.072	ellipt	0.076	classif	0.051	event	0.073	imput	0.073	imag	0.055	volum	0.063	hypothesi	0.04	monitor	0.06
dispens	0.074	distribut	0.058	problem	0.073	binomi	0.039	hot	0.069	gaussian	0.059	residu	0.048	mortal	0.06	simplifi	0.033	distribut	0.051
limb	0.056	correl	0.054	theorem	0.064	varieti	0.038	recurr	0.067	blend	0.05	bayesian	0.042	game	0.058	densiti	0.032	propos	0.049
auxiliari	0.054	model	0.047	exist	0.055	financi	0.035	advers	0.064	digit	0.047	sampl	0.042	polic	0.055	test	0.031	method	0.047
data	0.052	paramet	0.044	function	0.055	train	0.035	patient	0.063	problem	0.043	correl	0.041	paraiba	0.053	fit	0.03	indic	0.043
orthot	0.045	detrend	0.037	variat	0.049	sequenti	0.029	visibl	0.062	particl	0.042	infer	0.038	militari	0.049	replic	0.029	ewma	0.041
orthoped	0.044	mar	0.036	laplacian	0.045	literatur	0.027	graph	0.06	coffe	0.04	diagnosi	0.033	product	0.049	agnost	0.028	complementari	0.029
valu	0.04	function	0.035	domain	0.039	mosaic	0.027	pixel	0.058	data	0.039	random	0.029	yule	0.047	nonlinear	0.028	water	0.027
suppli	0.038	censor	0.032	topologi	0.038	plastic	0.027	mode	0.047	twitter	0.037	detect	0.028	offic	0.04	lee	0.027	process	0.026
analysi	0.035	propos	0.032	mountain	0.037	secur	0.027	quantif	0.045	homogen	0.036	bernoulli	0.027	simon	0.039	nitrogen	0.027	tocher	0.026
locomot	0.035	return	0.032	bound	0.036	smut	0.027	point	0.041	sensori	0.036	design	0.025	pandem	0.037	rabbit	0.027	arl	0.025
product	0.033	fruit	0.03	discret	0.035	cerrado	0.026	fire	0.037	sph	0.036	gaussian	0.025	formula	0.031	scienc	0.027	gener	0.025

Tópico 21		Tópico 22		Tópico 23		Tópico 24		Tópico 25		Tópico 26		Tópico 27		Tópico 28		Tópico 29		Tópico 30	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
quantil	0.178	knot	0.131	fish	0.079	distribut	0.123	measur	0.088	item	0.114	spatial	0.296	shift	0.074	violenc	0.141	covid	0.131
likelihood	0.107	censor	0.094	esb	0.071	model	0.089	particl	0.076	model	0.057	cluster	0.2	kernel	0.058	municip	0.108	food	0.099
profil	0.067	method	0.078	diseas	0.064	mixtur	0.078	probabl	0.074	vsap	0.054	rate	0.15	hospit	0.053	health	0.105	pandem	0.07
cluster	0.058	spline	0.06	region	0.061	likelihood	0.06	pass	0.06	base	0.045	lethal	0.114	breastfeed	0.049	pmaq	0.101	death	0.057
open	0.055	step	0.045	wait	0.058	zero	0.059	random	0.058	respond	0.041	incid	0.111	smell	0.048	primari	0.1	frequenc	0.055
estim	0.049	regress	0.04	artisan	0.054	mathemat	0.054	retir	0.055	voic	0.041	neighborhood	0.108	tast	0.047	transfer	0.099	anim	0.05
regress	0.044	locat	0.039	leprosi	0.051	estim	0.05	carryov	0.053	valid	0.039	mortal	0.086	surviv	0.042	manag	0.093	period	0.05
data	0.037	detect	0.036	jet	0.05	amazona	0.049	complet	0.047	lomax	0.038	socioeconom	0.084	hour	0.041	resourc	0.087	nest	0.049
propos	0.035	penal	0.036	leishmaniasi	0.049	master	0.047	effect	0.046	remot	0.037	janeiro	0.083	dataset	0.039	care	0.086	year	0.045
smooth	0.033	linear	0.031	flow	0.046	propos	0.046	subject	0.044	mortal	0.036	rio	0.082	preval	0.037	team	0.074	afternoon	0.044
log	0.03	number	0.031	paquetá	0.046	saunder	0.041	durat	0.038	cpp	0.035	sdi	0.079	factor	0.035	usf	0.067	caus	0.04
mont	0.029	outlier	0.03	island	0.045	birnbaum	0.04	space	0.036	influenza	0.033	risk	0.064	syndrom	0.035	famili	0.063	ecolog	0.04
inform	0.027	estim	0.029	user	0.045	paramet	0.037	topolog	0.033	aberr	0.029	citi	0.063	covari	0.03	cartograph	0.056	mortal	0.037
modifi	0.026	nonlinear	0.027	spatiotempor	0.044	postgradu	0.037	theori	0.03	gompertz	0.029	time	0.062	loss	0.03	financi	0.056	household	0.036
maximum	0.025	model	0.026	territori	0.044	bayesian	0.036	treatment	0.03	makeham	0.029	seri	0.055	wet	0.03	outlier	0.056	minut	0.036

Tópico 31		Tópico 32		Tópico 33		Tópico 34		Tópico 35		Tópico 36		Tópico 37		Tópico 38		Tópico 39		Tópico 40	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
program	0.133	twin	0.084	covid	0.073	sporotrichosi	0.104	dry	0.145	home	0.057	design	0.081	mortal	0.095	grade	0.175	matern	0.125
bak	0.093	discharg	0.082	pesticid	0.066	beam	0.061	mass	0.068	fraud	0.053	experiment	0.078	cancer	0.08	ident	0.151	bond	0.087
sneppen	0.091	hospit	0.076	higher	0.05	anim	0.049	accumul	0.062	dirichlet	0.047	mean	0.071	children	0.059	algebra	0.146	fetal	0.081
médico	0.083	blast	0.068	chikungunya	0.049	pragmat	0.049	plant	0.05	credit	0.045	block	0.058	year	0.057	polynomi	0.142	attach	0.076
option	0.083	rice	0.068	educ	0.049	survei	0.049	rainfal	0.046	carbon	0.042	control	0.053	rate	0.052	women	0.104	mother	0.075
care	0.077	preterm	0.067	teacher	0.046	precast	0.048	region	0.046	chapter	0.038	soil	0.052	region	0.045	bill	0.094	elect	0.072
avalanch	0.072	resist	0.058	consumpt	0.038	linear	0.04	precipit	0.045	multinomi	0.038	cluster	0.048	strain	0.042	pregnanc	0.078	dengu	0.059
articl	0.061	clone	0.057	pandem	0.037	oglinum	0.04	poset	0.043	card	0.037	bag	0.04	patient	0.041	hdp	0.073	pregnanc	0.058
critic	0.055	genet	0.049	papaya	0.036	telephon	0.039	hereditari	0.042	distribut	0.037	shape	0.038	covid	0.039	femal	0.065	bimod	0.057
primari	0.054	roc	0.049	laboratori	0.034	channel	0.038	spell	0.041	metric	0.035	life	0.037	pardo	0.036	weight	0.06	vote	0.056
mai	0.052	select	0.046	acut	0.032	hypothes	0.037	gompertz	0.038	class	0.033	forecast	0.035	rmr	0.036	curvatur	0.059	gestat	0.054
caregiv	0.05	breastfeed	0.044	product	0.031	nonparametr	0.029	season	0.037	bay	0.029	analysi	0.034	child	0.034	gain	0.055	infant	0.051
famili	0.049	curv	0.041	epidemiolog	0.03	plan	0.029	count	0.035	discrimin	0.029	synthet	0.034	geograph	0.034	topic	0.05	prenat	0.045
neonat	0.049	milk	0.038	health	0.03	human	0.027	maiz	0.034	model	0.028	treatment	0.033	risk	0.032	will	0.05	risk	0.043
access	0.047	trait	0.037	fear	0.029	rbd	0.027	bertalanffi	0.031	permit	0.028	bootstrap	0.032	treatment	0.031	hypersurfac	0.045	women	0.042

Depois de fazer a análise de cada tópico no modelo LDA escolhido, com base nas pontuações dos termos mais relevantes em cada um dos tópicos, apresenta-se uma breve descrição dos tópicos latentes na Tabela 5.3.

5.1.1 Análise de convergência das cadeias

Com o propósito de analisar o comportamento das cadeias de algumas componentes β_{kv} da matriz $\boldsymbol{\beta} \mid \mathbf{w}$, foram escolhidas as componentes com maior pontuação do termo e maior probabilidade (olhar Figuras 5.6 e 5.7, respectivamente). De modo geral, pode-se observar que não se tem convergência nas cadeias dos $\hat{\beta}_{kv}$, somente tendo-se uma possível convergência (se tivesse mais período de aquecimento, isto é, se esse período fosse de 5000 ao invés de 2500) na cadeia do $\hat{\beta}_{27,970}$ mostrada na Figura 5.6. Também é importante dizer que, as escalas nos eixos verticais são de amplitude bastante pequenas, o que poderia indicar que fazendo mais iterações, a diferença na estimação vai ser pequena, sem contemplar a ideia de convergir à distribuição estacionaria do parâmetro $\boldsymbol{\beta}$.

Agora analisando as cadeias escolhidas das componentes do $\hat{\boldsymbol{\theta}}$, mostradas na Figura 5.8, observa-se convergência em todas, pois os seus valores oscilam de forma estável e numa amplitude quase fixa, ao redor de sua média ($\hat{\theta}_{dk}$).

Ante a não convergência das cadeias dos $\hat{\beta}_{kv}$, mostrada nas Figuras 5.6 e 5.7, primeiramente note-se que estas cadeias não contém os valores iniciais, logo observa-se na Figura A.1 o comportamento da cadeia ao início (as 100 primeiras iterações simuladas) para dois componentes de $\hat{\boldsymbol{\beta}}$; com espaçamento de 5, sem período de aquecimento e com valores iniciais nas duas componentes ($\beta_{13,1}$ e $\beta_{27,970}$) de 0, 0, 5 e 1; do anterior, percebe-se que as cadeias não ficam presas nesses valores iniciais, isto é, a cadeia da um pulo para o possível valor à convergir ou no que vai-se estar movimentando ou oscilando (no caso de não convergência, como está ocorrendo no modelo LDA treinado e escolhido como o melhor), isto dá indícios que o comportamento da cadeia é independente do valor inicial na componente β_{kv}.

Tabela 5.3: Nome dos tópicos, determinados para o modelo LDA escolhido, na abordagem via MCMC.

	NOME DO TÓPICO
TÓPICO 1	Aprendizado de Máquina (Machine Learning)
TÓPICO 2	Modelagem Estatística e Econometria
TÓPICO 3	Matemática Aplicada e Teoria das Categorias
TÓPICO 4	Modelagem Estatística, Mobilidade Urbana e Otimização
TÓPICO 5	Modelagem Estatística e Análise de Dados Temporais
TÓPICO 6	Genética Estatística e Bioinformática
TÓPICO 7	Teoria dos Conflitos, Otimização e Sistemas Territoriais
TÓPICO 8	Avaliação Educacional e Psicometria
TÓPICO 9	Economia Agrícola e Modelagem de Preços de Mercadorias
TÓPICO 10	Gestão de Recursos e Análise de Fluxos no Turismo e Saneamento
TÓPICO 11	Tecnologia Assistiva e Ortopedia
TÓPICO 12	Análise de Séries Temporais e Modelagem Estatística Multivariada
TÓPICO 13	Métodos Numéricos e Equações Diferenciais Parciais
TÓPICO 14	Estatística Aplicada, Pesquisa e Modelagem em Diversos Campos
TÓPICO 15	Análise de Dados Ambientais e Saúde Pública
TÓPICO 16	Processamento de Imagens Digitais e Imputação de Dados
TÓPICO 17	Estatística Bayesiana e Processamento de Imagens
TÓPICO 18	Estatísticas e Análise de Dados em Saúde Pública e Sociedade
TÓPICO 19	Modelagem Estatística e Biológica: Técnicas e Aplicações
TÓPICO 20	Controle Estatístico de Processos e Análise de Dados
TÓPICO 21	Estatística Computacional e Modelagem de Dados
TÓPICO 22	Modelagem Estatística e Regressão Não Linear
TÓPICO 23	Saúde Pública e Territórios: A Interação Entre Fatores Ambientais, Geográficos e Sociais
TÓPICO 24	Modelos Estatísticos e Métodos de Estimação Bayesiana
TÓPICO 25	Processos Estocásticos e Análise de Experimentos
TÓPICO 26	Modelos Estatísticos e Análise de Dados de Sobrevivência
TÓPICO 27	Análise Espacial e Temporal de Indicadores de Saúde Urbana e Socio-econômicos no Rio de Janeiro
TÓPICO 28	Aprendizado de Máquina e Modelagem Estatística para Dados de Saúde
TÓPICO 29	Gestão e Melhoria da Atenção Básica em Saúde Pública
TÓPICO 30	Epidemiologia Ecológica e Análise de Dados sobre a Pandemia de COVID-19
TÓPICO 31	Modelos Dinâmicos e Cuidados de Saúde Primária com Enfoque em Saúde Familiar e Neonatal
TÓPICO 32	Genética, Saúde Pública e Avaliação de Modelos em Saúde
TÓPICO 33	Saúde Pública, Epidemiologia e Impactos da Pandemia
TÓPICO 34	Pesquisa Epidemiológica, Estatística e Planejamento Experimental em Saúde
TÓPICO 35	Modelagem Ecológica e Estatísticas Ambientais
TÓPICO 36	Modelagem Estatística para Análise de Risco e Detecção de Fraude
TÓPICO 37	Delineamento Experimental e Métodos Estatísticos de Previsão
TÓPICO 38	Epidemiologia das Taxas de Mortalidade e Análise de Risco em Saúde Pública
TÓPICO 39	Modelagem Matemática na Saúde Feminina e Políticas Públicas
TÓPICO 40	Saúde Materno-Infantil e Políticas Públicas de Saúde

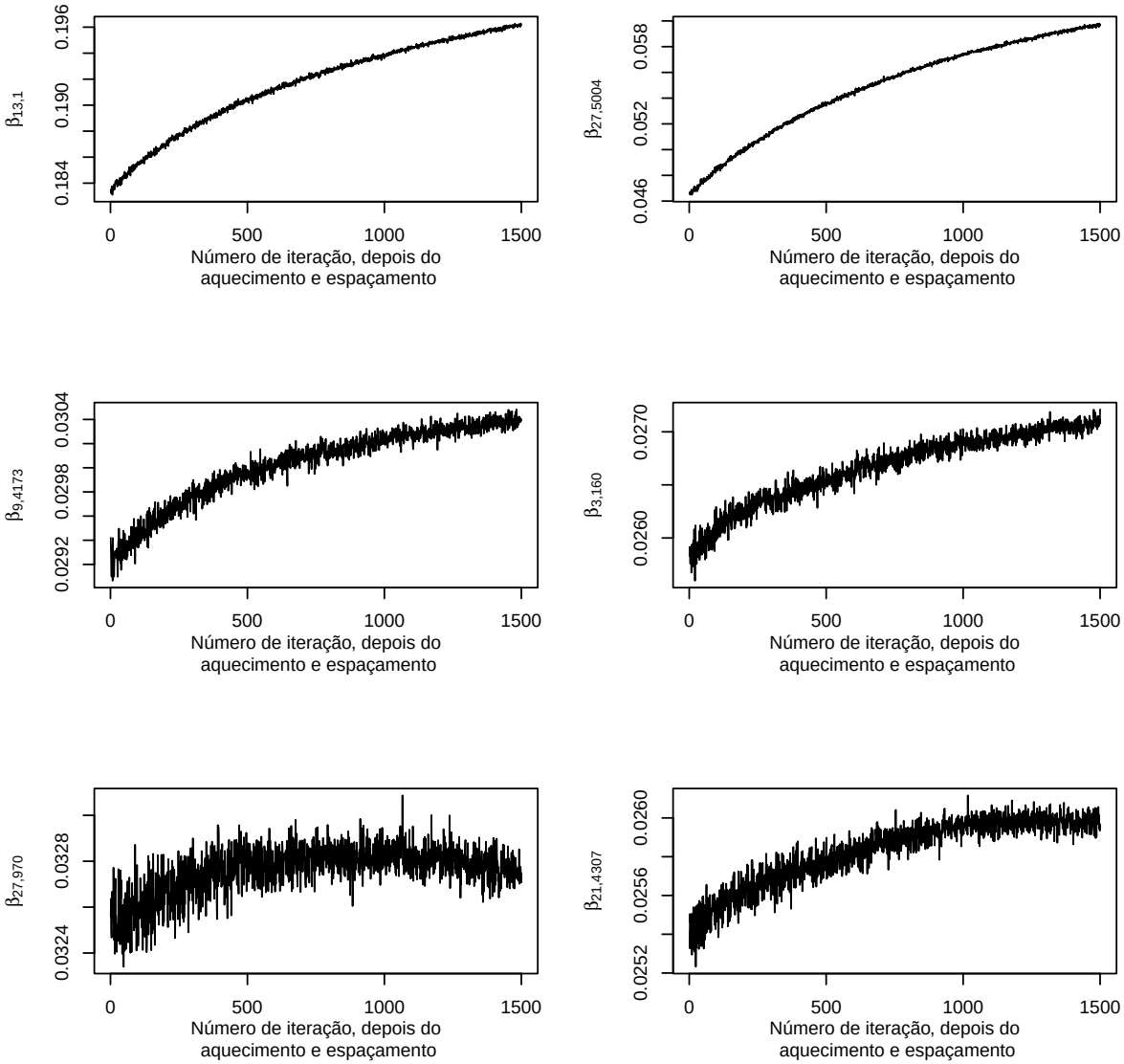


Figura 5.6: Cadeias de algumas das componentes de $\hat{\boldsymbol{\beta}}$, sendo as componentes com maior pontuação do termo; desde o modelo LDA treinado no MCMC com 10000 iterações, $K = 40$, espaçamento de 5 e aquecimento de 500.

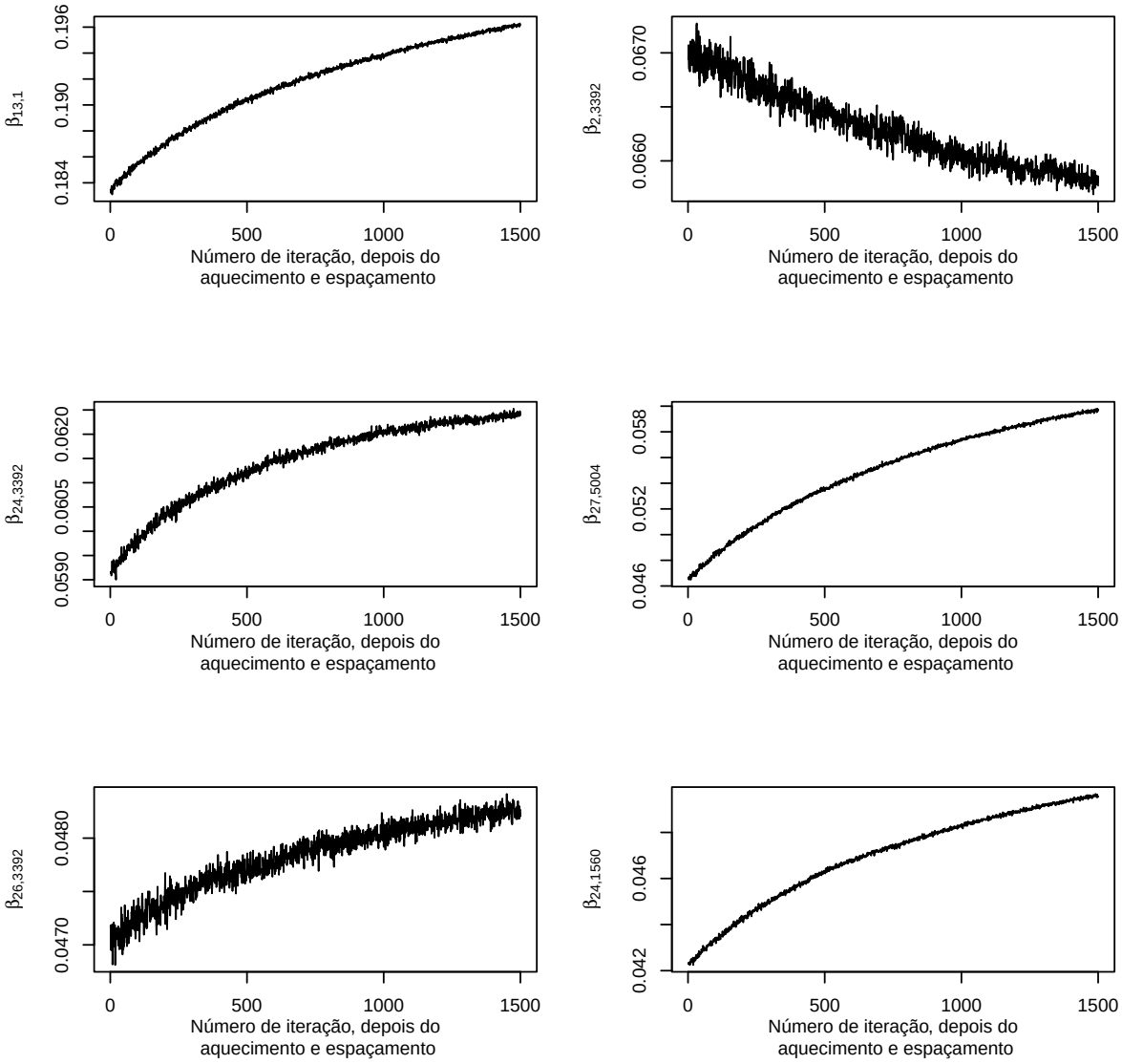


Figura 5.7: Cadeias de algumas das componentes de $\hat{\boldsymbol{\beta}}$, sendo as componentes com maior probabilidade; desde o modelo LDA treinado no MCMC com 10000 iterações, $K = 40$, espaçamento de 5 e aquecimento de 500.

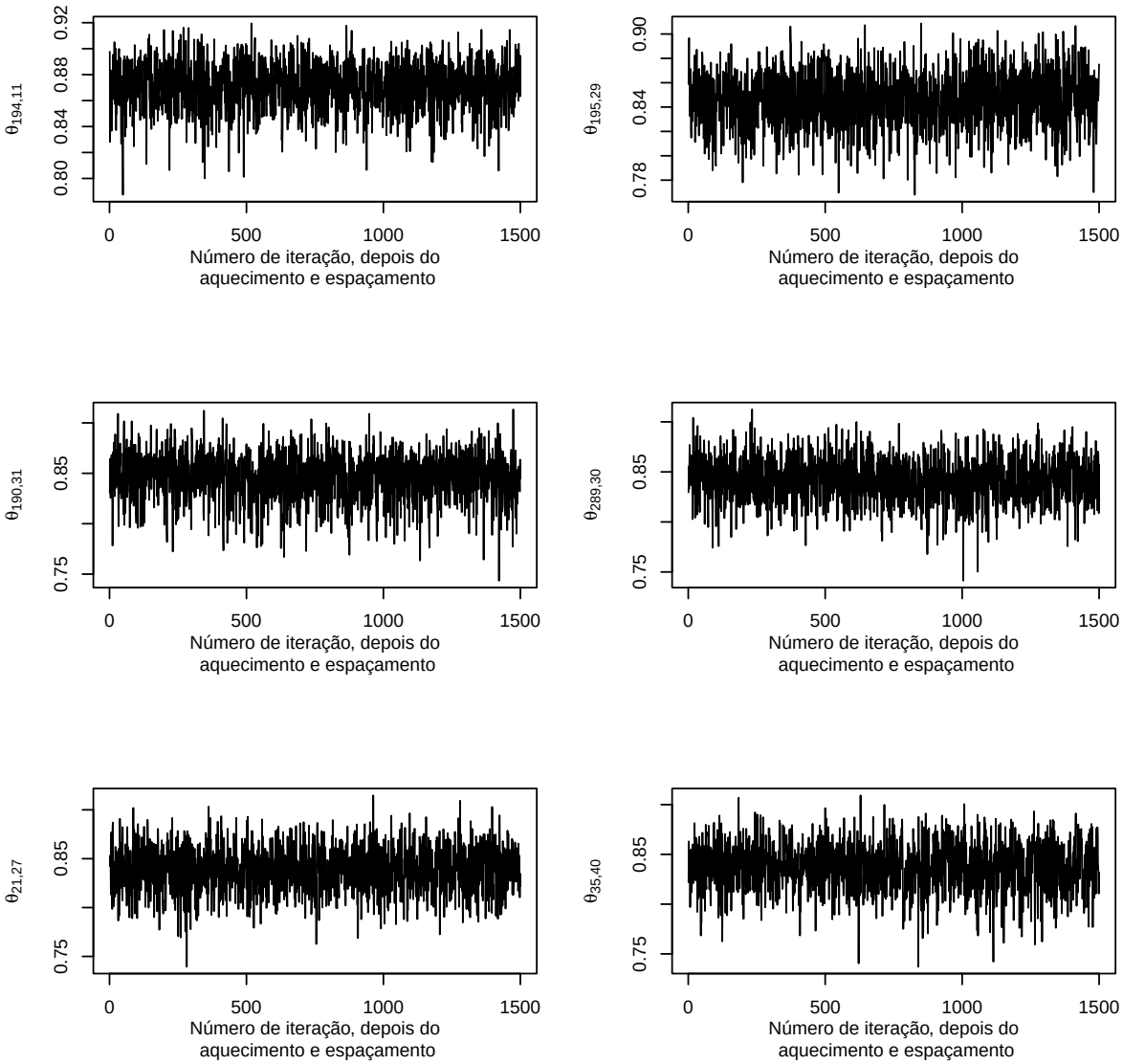


Figura 5.8: Cadeias de algumas das componentes de $\hat{\boldsymbol{\theta}}$, sendo as componentes com maior probabilidade; desde o modelo LDA treinado no MCMC com 10000 iterações, $K = 40$, espaçamento de 5 e aquecimento de 500.

Visando investigar mais à fundo a convergência das cadeias, foi treinado novamente um modelo LDA com 40 tópicos (por ter sido escolhido como o melhor modelo), 500000 iterações, aquecimento de 300000, espaçamento de 100 e tirando inicialmente as observações com *NA* (correspondentes às 517 observações que não são válidas, mencionadas no capítulo 2). Nas Figuras A.2, A.3 e A.4; são mostradas as cadeias do modelo treinado de algumas das componentes de $\hat{\boldsymbol{\beta}}$ com maior pontuação do termo e maior probabilidade, e de algumas das componentes de $\hat{\boldsymbol{\theta}}$ com maior probabilidade, respectivamente. Novamente, tem-se que as cadeias das componentes do $\hat{\boldsymbol{\beta}}$ não tem convergência para sua distribuição estacionária, mas tem-se que, ainda mantêm-se valores muito baixos em termos de amplitude.

Mesmo que não se tenha observado evidências de convergência à distribuição estacionária nos modelos treinados para $\hat{\boldsymbol{\beta}}$, mas sim para $\hat{\boldsymbol{\theta}}$, continuamos com a escolha e análise nesta abordagem de MCMC, do modelo LDA treinado com 40 tópicos. Apesar das questões apontadas, pode-se observar que os tópicos inferidos para o modelo LDA escolhido, tem interpretabilidade nas 15 palavras com maior pontuação do termo que conformam cada um dos 40 tópicos (olhar as Tabelas 5.2 e 5.3).

O modelo LDA escolhido, possibilita ainda a realização de recomendação por similaridade usando as proporções posteriores dos tópicos para definir uma medida de similaridade baseada em tópicos entre os documentos. Pode-se usar a distância de Hellinger $\mathcal{H}(d, f)$ entre dois documentos $d, f \in \{1, \ldots, D\}$ como uma medida de similaridade [Blei, 2009], definida como,

$$\mathcal{H}(d, f) = \sum_{k=1}^{K} \left(\sqrt{\hat{\theta}_{d,k}} - \sqrt{\hat{\theta}_{f,k}} \right)^2 .$$

O procedimento consiste em listar os documentos do *corpus* que mais se assemelham a um dado documento fixado d, o que é feito comparando-se a similaridade entre as proporções de tópicos $\boldsymbol{\theta}_d$ e $\boldsymbol{\theta}_f$ nos diferentes pares de documentos d e f do *corpus*. Por

exemplo, a Figura A.5 mostra os 10 documentos mais similares ao documento de número 194 ($d = 194$), entitulado *"Órteses, Próteses e Meios auxiliares de locomoção (OPM) nas oficinas ortopédicas da rede de cuidados à pessoa com deficiência- 2021- IES: UFPB"*, incluindo também as principais palavras dos principais tópicos no documento d (ordenada pela pontuação do termo [Blei, 2009]), e as proporções de cada tópico estimadas no documento d. Pode-se observar primeiramente o nome do documento, o ano de elaboração, e à qual IES pertence: *"Documento 194: Órteses, Próteses e Meios auxiliares de locomoção (OPM) - 2021- IES: UFPB"*. Também observa-se os 10 documentos mais similares com a mesma estrutura, sendo o documento mais similar do *corpus* o número 151: *"Imputação de Dados Sintéticos Através de Árvores de Classificação- 2019- IES: UFMG"*; também observa-se os três tópicos com mais probabilidade no documento, sendo o tópico 11 o mais provável com um valor do $0,871$, o qual com base na Tabela 5.3 é *"Tecnologia Assistiva e Ortopedia"*.

5.2 Abordagem via inferência variacional

Optou-se por utilizar inferência variacional, através do algoritmo de coordenada ascendente com pequenos lotes (*minibatch*). Este algoritmo otimiza as distribuições aproximadas dos parâmetros do modelo de maneira mais eficiente, utilizando *minibatch* (subconjunto pseudo-aleatório de documentos) para atualizar iterativamente as estimativas, o que melhora significativamente a escalabilidade e reduz o tempo de computação, especialmente em grandes volumes de dados.

Durante o treinamento dos modelos LDA para diferentes quantidades de tópicos ($K \in \{3, 4, \ldots, 14, 15, 20, 30, 40, 50\}$), foram deixadas fixas as seguintes especificações: os *minibatch* possuem tamanho igual a 3 documentos (isto é, a base de dados será dividida em 139 *minibatches*), num total de 6672 iterações; seguindo o Algoritmo 4.2 descrito no capítulo 4.

Após de obter os modelos treinados, a convergência de cada um foi avaliada medi-

ante a sua log-verossimilhança dos dados observados em todos os documentos do *corpus*, calculada ao longo das iterações sob os parâmetros estimados pelo modelo variacional. Observa-se na Figura 5.9 que as curvas dos valores calculados da log-verossimilhança em cada um dos modelos treinados, existem evidências de convergência em todos os casos, embora é possível usar mais iterações para uma maior precisão. Observa-se que o modelo com log-verossimilhança mais elevada é o de 30 tópicos, seguido pelos modelos de 50 e 40 tópicos (respectivamente).

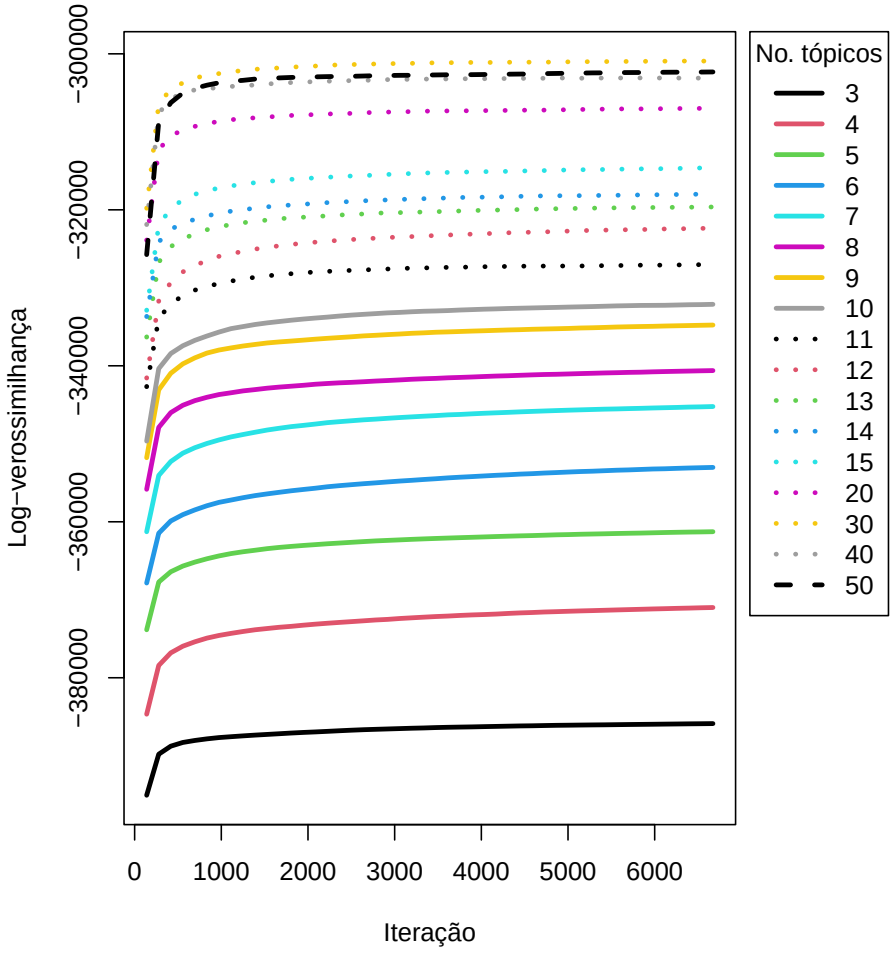


Figura 5.9: Log-verossimilhança dos modelos treinados via inferência variacional.

Pode-se observar na Tabela 5.4 os valores pontuais da log-verossimilhança de cada um

dos modelos LDA treinados, na sua última iteração. Ratificando o que se observou na Figura 5.9, o modelo LDA treinado via inferência variacional com 30 tópicos apresenta a maior log-verossimilhança com um valor de $-300934,5$, seguido pelos modelos de 50 e 40 tópicos, com valores de -302323.4 e -303112.6, respectivamente.

Tabela 5.4: Valores da log-verossimilhança na sua última iteração, dos modelos LDA treinados na inferência variacional.

	VALOR DA LOG-VEROSSIMILHANÇA
Modelo LDA com 3 tópicos	-385877.9
Modelo LDA com 4 tópicos	-370992.5
Modelo LDA com 5 tópicos	-361274.8
Modelo LDA com 6 tópicos	-353028.5
Modelo LDA com 7 tópicos	-345222.7
Modelo LDA com 8 tópicos	-340614.4
Modelo LDA com 9 tópicos	-334774.6
Modelo LDA com 10 tópicos	-332119.2
Modelo LDA com 11 tópicos	-327018.4
Modelo LDA com 12 tópicos	-322345.5
Modelo LDA com 13 tópicos	-319634.2
Modelo LDA com 14 tópicos	-317986.0
Modelo LDA com 15 tópicos	-314624.1
Modelo LDA com 20 tópicos	-306974.4
Modelo LDA com 30 tópicos	**-300934.5**
Modelo LDA com 40 tópicos	-303112.6
Modelo LDA com 50 tópicos	-302323.4

Depois de olhar a log-verossimilhança na sua última iteração dos modelos treinados, a Figura 5.10 contém a média das medidas de coerência C_{UCI}^k calculadas nos tópicos $k = 1, \ldots, K$ de cada um dos modelos LDA treinados. O modelo LDA com 50 tópicos se apresenta como o de maior coerência média dos tópicos.

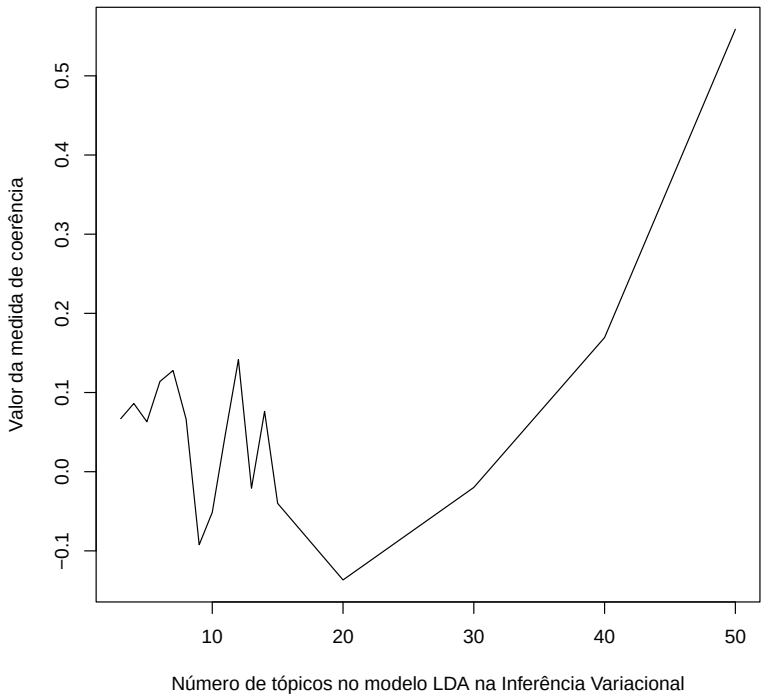


Figura 5.10: Cálculo da medida de coerência (C_{UCI}) nos tópicos dos modelos LDA treinados via inferência variacional.

Também, observa-se na Figura 5.11 as distribuições por meio de diagramas de caixa, das medidas de coerência (C_{UCI}) calculadas para cada um dos tópicos nos diferentes modelos LDA treinados via inferência variacional. Novamente, o modelo LDA com 50 tópicos, apresenta a maior mediana, sem pontos atípicos, e com valores de coerência elevados. Também observa-se a maior variabilidade nas medidas de coerência com $K = 50$ em comparação com os outros modelos.

A análise das medidas de coerência apresentadas nas Figuras 5.10 e 5.11 parecem indicar a escolha de 40 ou 50 tópicos. Entretanto, após a análise dos tópicos inferidos nesses dois modelos demonstrou que, na verdade, os tópicos inferidos apresentam baixo grau de interpretabilidade na prática, sendo muito difícil a identificação dos assuntos abordados nas dissertações. Por outro lado, a escolha de $K = 30$ (motivada pela análise da verossimilhança dos modelos segundo a Figura 5.9 e a Tabela 5.4) levou à identificação

de tópicos com mais interpretabilidade. Ordenando de forma decrescente as entradas de $\hat{\boldsymbol{\beta}}_k$ nos 30 tópicos, a Tabela 5.5 apresenta as 15 palavras mais prováveis de cada tópico, respeito à pontuação do termo. Uma breve descrição dos tópicos latentes encontra-se na Tabela 5.6.

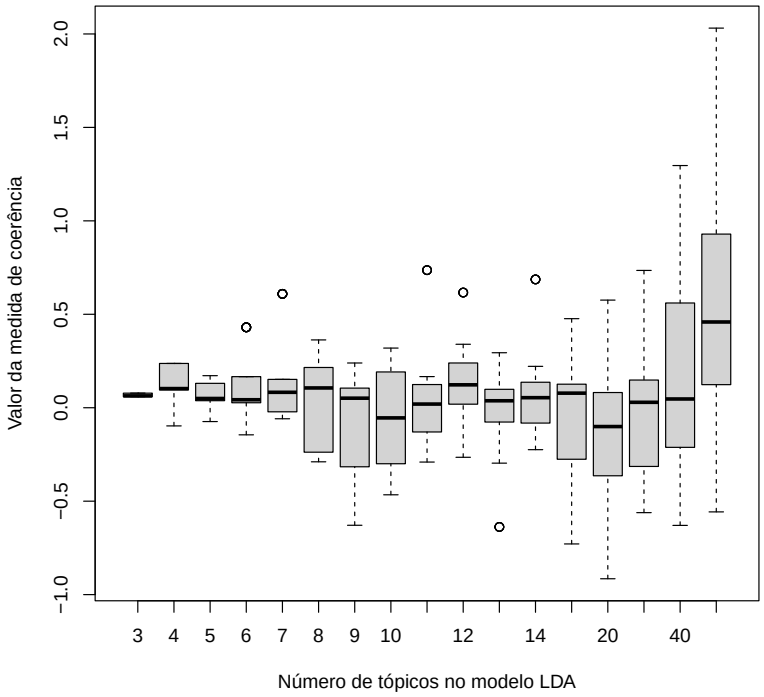


Figura 5.11: Distribuições da medida de coerência (C_{UCI}) nos tópicos dos modelos LDA treinados na inferência variacional.

Também é feita recomendação por similaridade para algum documento d, a partir do modelo escolhido nesta abordagem, além das outras estatísticas como as principais palavras dos principais tópicos (ordenadas pela pontuação do termo), e as proporções de cada tópico estimadas nesse mesmo documento d, a partir dos parâmetros estimados $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\theta}}$. Como exemplo, observa-se na Figura A.6 os 10 documentos mais similares ao documento de número 106 ($d = 106$), intitulado *"Documento 106: Modelo de regressão beta retangular para análise de dados com medidas repetidas- 2019- IES: UFC"*, incluindo também as principais palavras dos principais tópicos no documento d (ordenada pela pontuação do termo), e as proporções de cada tópico estimadas no documento d.

Tabela 5.5: 15 palavras mais prováveis de cada tópico, respeito à pontuação do termo [Blei, 2009], no modelo LDA escolhido na inferência variacional.

Tópico 1		Tópico 2		Tópico 3		Tópico 4		Tópico 5		Tópico 6		Tópico 7		Tópico 8		Tópico 9		Tópico 10	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
test	0.926	chang	0.374	likelihood	0.494	problem	0.289	outlier	0.718	distribut	0.359	flow	0.363	coffe	0.839	matemática	0.501	price	0.506
valu	0.37	return	0.254	zero	0.488	work	0.179	volatil	0.501	data	0.314	finit	0.236	option	0.293	stochast	0.394	probabl	0.403
random	0.318	access	0.248	real	0.289	will	0.157	asymptot	0.429	estim	0.26	calcul	0.226	ellipt	0.287	topolog	0.343	mathemat	0.256
design	0.298	classif	0.217	maximum	0.28	studi	0.156	birnbaum	0.378	method	0.236	genom	0.219	competit	0.24	penal	0.28	degre	0.245
extrem	0.212	market	0.182	saunder	0.237	function	0.138	mont	0.375	model	0.204	polici	0.201	upper	0.219	wavelet	0.235	particl	0.209
mean	0.199	mainli	0.164	version	0.225	posit	0.132	lasso	0.283	propos	0.176	user	0.193	element	0.2	spell	0.215	bootstrap	0.195
imag	0.173	period	0.14	estim	0.2	two	0.13	carlo	0.255	regress	0.163	indic	0.142	compet	0.181	share	0.195	consumpt	0.162
method	0.171	signific	0.137	inflat	0.2	system	0.125	amount	0.184	paramet	0.119	posterior	0.129	nois	0.175	incomplet	0.175	programa	0.145
forest	0.125	structur	0.131	dispers	0.165	present	0.125	constant	0.181	simul	0.117	demand	0.117	rehabilit	0.17	usf	0.173	line	0.144
complet	0.104	life	0.128	poisson	0.154	keyword	0.102	bernoulli	0.167	gener	0.115	game	0.107	text	0.16	heteroscedast	0.172	chart	0.143
miss	0.1	stage	0.119	ratio	0.127	point	0.1	lattic	0.154	sampl	0.101	fluid	0.106	corn	0.143	sari	0.127	real	0.142
vector	0.097	classifi	0.112	wide	0.116	program	0.097	empir	0.148	perform	0.096	increas	0.104	driven	0.142	gam	0.114	articl	0.139
real	0.096	analyz	0.108	coverag	0.112	method	0.092	séri	0.141	studi	0.087	regim	0.103	vertic	0.128	arima	0.101	master	0.137
sensit	0.096	present	0.104	sensori	0.095	number	0.091	littl	0.131	normal	0.081	central	0.102	rabbit	0.126	camila	0.1	pair	0.13
allow	0.089	code	0.098	cours	0.086	space	0.087	simplifi	0.126	function	0.076	viçosa	0.092	smell	0.123	geometri	0.097	amazona	0.128

Tópico 11		Tópico 12		Tópico 13		Tópico 14		Tópico 15		Tópico 16		Tópico 17		Tópico 18		Tópico 19		Tópico 20	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
spatial	0.342	equat	0.644	random	0.461	health	0.325	techniqu	0.359	prior	0.478	statist	0.214	network	1.224	form	0.475	analysi	0.187
detect	0.2	latent	0.482	polynomi	0.454	studi	0.198	algebra	0.309	paulo	0.385	covid	0.128	machin	0.594	ident	0.361	cluster	0.129
risk	0.187	confirm	0.268	mestrado	0.322	mortal	0.159	theori	0.219	arma	0.362	result	0.097	neural	0.407	soil	0.248	group	0.121
women	0.139	biom	0.189	attribut	0.282	year	0.155	decis	0.209	solv	0.351	diseas	0.095	sugar	0.275	open	0.217	são	0.117
associ	0.133	extract	0.18	color	0.259	care	0.127	real	0.197	shift	0.249	sever	0.091	commod	0.25	memori	0.217	municip	0.103
paulo	0.111	intens	0.175	loss	0.237	factor	0.12	student	0.16	kernel	0.218	research	0.09	team	0.245	symptom	0.182	possibl	0.093
observ	0.101	fish	0.171	universidad	0.211	pandem	0.08	recurr	0.146	modifi	0.215	develop	0.09	boundari	0.198	visual	0.173	area	0.089
regress	0.098	solut	0.165	bak	0.186	brazil	0.077	comput	0.134	rice	0.207	data	0.085	topologi	0.177	paraiba	0.153	estatística	0.088
quantit	0.097	part	0.158	best	0.172	death	0.075	gamma	0.133	partit	0.186	individu	0.083	phenomena	0.164	esb	0.139	need	0.079
proport	0.093	roc	0.156	cut	0.154	analyz	0.072	markov	0.126	krige	0.168	can	0.081	articl	0.131	configur	0.127	high	0.079
case	0.089	stock	0.152	pesticid	0.148	educ	0.07	carri	0.123	econom	0.168	work	0.081	show	0.125	intersect	0.123	region	0.075
correl	0.088	interpol	0.143	topic	0.133	first	0.066	usual	0.123	variant	0.146	process	0.08	imbalanc	0.099	horizont	0.114	number	0.073
characterist	0.083	avoid	0.141	chemic	0.12	higher	0.062	bound	0.116	spatio	0.145	associ	0.07	visibl	0.097	dna	0.103	howev	0.071
variabl	0.081	opm	0.137	climat	0.119	caus	0.059	gener	0.105	maxim	0.136	case	0.07	carryov	0.089	list	0.098	locat	0.063
occurr	0.08	comorbid	0.122	call	0.104	grade	0.056	correct	0.104	integ	0.116	qualiti	0.065	overdispers	0.075	matric	0.098	citi	0.063

Tópico 21		Tópico 22		Tópico 23		Tópico 24		Tópico 25		Tópico 26		Tópico 27		Tópico 28		Tópico 29		Tópico 30	
Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.	Palavra	Pont.Term.
financi	0.711	genet	0.674	process	0.543	surviv	0.676	model	1.174	quantil	1.234	item	0.867	valid	0.839	model	0.51	predict	0.444
pmaq	0.232	plant	0.538	treatment	0.421	food	0.626	effect	0.228	discrimin	0.278	fraud	0.362	solut	0.733	base	0.16	interest	0.299
fix	0.189	genotyp	0.531	time	0.348	censor	0.545	linear	0.21	superior	0.267	four	0.288	smooth	0.534	evalu	0.157	dynam	0.27
sneppen	0.185	varieti	0.436	control	0.293	transfer	0.38	fit	0.199	conflict	0.257	enem	0.243	exist	0.347	variabl	0.148	variat	0.259
médico	0.171	speci	0.297	hospit	0.25	marker	0.355	time	0.131	snp	0.229	lomax	0.204	dissertação	0.205	select	0.146	illustr	0.186
clone	0.162	environ	0.232	optim	0.243	categori	0.273	compon	0.113	classifi	0.216	mine	0.16	organ	0.183	two	0.128	combin	0.173
chosen	0.148	akaik	0.2	patient	0.242	northeast	0.189	process	0.109	advers	0.191	spars	0.142	nonparametr	0.149	bayesian	0.128	densiti	0.164
avalanch	0.141	cell	0.184	dataset	0.203	sporotrichosi	0.177	valu	0.076	laplacian	0.18	feasibl	0.121	souza	0.146	product	0.118	universidad	0.144
receiv	0.13	ammi	0.169	batch	0.186	automorph	0.176	first	0.073	plai	0.164	tri	0.103	instanc	0.124	approach	0.107	cost	0.143
januari	0.129	cattl	0.159	monitor	0.179	multipl	0.113	best	0.07	frame	0.16	extern	0.101	voic	0.117	order	0.087	result	0.123
parameter	0.127	gamlss	0.154	approxim	0.15	mountain	0.092	non	0.07	guarante	0.148	statu	0.1	quantif	0.107	statist	0.086	usual	0.098
adult	0.113	height	0.12	converg	0.136	close	0.091	event	0.067	choos	0.147	card	0.096	wait	0.103	present	0.083	concept	0.095
signal	0.102	precipit	0.114	cancer	0.115	principl	0.087	analysi	0.067	teacher	0.147	hydrodynam	0.086	scott	0.093	data	0.082	pass	0.095
perman	0.089	ethanol	0.113	clinic	0.108	carolina	0.086	assess	0.063	beam	0.131	gather	0.086	dispens	0.091	mean	0.078	spline	0.093
pension	0.086	largest	0.101	diagnosi	0.1	censorship	0.085	mix	0.062	turn	0.095	cruz	0.083	involv	0.084	differ	0.067	limit	0.089

Tabela 5.6: Nome dos tópicos, determinados para o modelo LDA escolhido na inferência variacional.

	NOME DO TÓPICO
Tópico 1	Modelagem Estatística e Aprendizado de Máquina
Tópico 2	Econometria e Modelagem Financeira
Tópico 3	Inferência Estatística e Modelos de Regressão
Tópico 4	Métodos Computacionais e Análise de Sistemas
Tópico 5	Métodos Estatísticos, Inferência e Algoritmos Computacionais
Tópico 6	Métodos Estatísticos e Análise de Dados
Tópico 7	Modelagem de Fluxos e Processos Dinâmicos
Tópico 8	Estratégias Competitivas e Processos Cognitivos
Tópico 9	Estatística Computacional e Processos Estocásticos
Tópico 10	Análise de Dados e Métodos Computacionais Aplicados
Tópico 11	Modelagem Estatística e Análise de Risco em Saúde Pública
Tópico 12	Modelagem de Dados Espaciais e Estatísticas em Saúde Pública
Tópico 13	Sistemas Complexos, Dinâmica e Modelagem Ambiental
Tópico 14	Estudos de Mortalidade, Causas e Tendências em Saúde
Tópico 15	Métodos Computacionais em Teoria da Decisão e Processos Estocásticos
Tópico 16	Processos Estocásticos, Modelos ARMA e Técnicas Computacionais
Tópico 17	Estudos Estatísticos sobre a COVID-19 e Severidade da Doença
Tópico 18	Modelagem de Machine Learning com Redes Neurais e Análise de Superdispersão
Tópico 19	Epidemiologia e Saúde Bucal com Análise Geoespacial e Genética
Tópico 20	Análise de Cluster e Localização Geográfica
Tópico 21	Análise Financeira, Modelos Complexos e Saúde Pública
Tópico 22	Modelos de Interação Genética e Ambiental em Agricultura
Tópico 23	Controle de Processos e Análise Temporal no Diagnóstico de Câncer
Tópico 24	Modelagem de Sobrevivência e Estatísticas em Doenças Infecciosas
Tópico 25	Modelagem Estatística de Efeitos e Processos Temporais
Tópico 26	Análise Discriminante e Algoritmos em Genética e Bioinformática
Tópico 27	Modelos Estatísticos e Análise de Dados em Diversos Contextos
Tópico 28	Técnicas Estatísticas e Topológicas em Análise de Dados
Tópico 29	Métodos Estatísticos para Avaliação de Modelos e Seleção de Variáveis
Tópico 30	Análise de Dados, Previsão e Interpolação

No relatório mostrado na Figura A.6, observa-se os 10 documentos mais similares com a mesma estrutura, sendo o documento mais similar do *corpus* o de número 230: *"Estimação de processos com longa dependência na presença de muitos dados faltantes- 2023- IES: UFRGS"*; também observa-se os três tópicos com mais probabilidade no documento, sendo o tópico 6 o mais provável com um valor do $0,673$, o qual com base na Tabela 5.6 se intitula *"Métodos Estatísticos e Análise de Dados"*.

5.3 Comparação das duas abordagens com os resultados obtidos

Ao comparar os tópicos inferidos com as duas abordagens, a inferência variacional e o MCMC, observa-se que foi conseguido identificar temas relevantes da estatística e áreas correlatas (olhar Tabelas 5.3 5.6), mas com algumas diferenças. Na abordagem do MCMC, tem-se tópicos que foram aprendidos que pertencem a áreas que não são observadas na abordagem da inferência variacional, como: Urbanismo (tópico 4), Economia Agrícola (tópico 9), Gestão de Recursos (tópico 10), Tecnologia Médica (tópico 11), Ecologia (tópicos 15 e 35), Biologia (tópico 19), Controle de Qualidade (tópico 20), Gestão de Saúde (tópico 29), Modelagem Matemática (tópico 31), Lógica (tópico 3), Educação (tópico 8), Pesquisa (tópicos 14 e 37), Segurança (tópico 36); assim como na inferência variacional foram aprendidos tópicos pertencentes a áreas que ao parecer não estão compreendidas na abordagem do MCMC, como: Finanças (tópicos 2 e 21), Engenharia (tópicos 4 e 23), Ciências Ambientais (tópicos 13 e 20), Agricultura (tópico 22).

Desde o fato que na abordagem do MCMC tenham-se mais áreas compreendidas nos tópicos aprendidos que não estejam nas áreas compreendidas nos tópicos aprendidos na inferência variacional; poderia-se dizer que os tópicos aprendidos no MCMC são mais precisos que os aprendidos na inferência variacional, lembrando que o modelo LDA escolhido no MCMC tem 40 tópicos e o escolhido na inferência variacional tem 30 tópicos.

Por outro lado, desde os tópicos aprendidos nas duas abordagens, as principais áreas

que estão compreendidas nestas, por ordem de relevância com base na frequência de tópicos em cada abordagem, são mostradas na Tabela 5.7. Observa-se que a Estatística é a principal área, isto é um bom indício lembrando que os documentos que conformam o *corpus* são dissertações de mestrado em Estatística e áreas correlatas de IES do Brasil.

Tabela 5.7: Principais áreas compreendidas pelos tópicos aprendidos em cada abordagem (MCMC e inferência variacional).

	NÚMERO DE TÓPICOS	
ÁREA	MCMC	INFERÊNCIA VARIACIONAL
Estatística	18	14
Saúde Pública	13	6
Ciências da Computação	4	8
Matemática	6	6
Epidemiologia	4	3

Agora são examinados os tempos computacionais que foram necessários para treinar os modelos LDA em cada abordagem, utilizando o software R versão 4.3.1 (2023-06-16 ucrt); executado em um notebook com processador AMD E2-1800 APU com Radeon(tm) HD Graphics 1,70 GHz, 4 GB de RAM e sistema operacional Windows 8 Single Language (64 bits) © 2012 Microsoft Corporation. Mostra-se na Tabela 5.8 que a inferência variacional é mais eficiente que que o MCMC (com 10000 iterações), em todos os modelos, mesmo que na coluna 3 dessa Tabela foi feito o cálculo proporcional à 10000 iterações na abordagem da inferência variacional (lembrando que foram 6672 iterações nesta última).

Tabela 5.8: Tempos computacionais (em segundos) para o treinamento dos modelos LDA na abordagem via MCMC a na abordagem via inferência variacional.

	MCMC	Inferência Variacional	
K	10000 iterações	6672 iterações	10000 iterações proporcionais
3	825	63,195	94,717
4	940	86,650	129,871
5	1048	107,930	161,766
6	1164	128,725	192,933
7	1294	151,795	227,510
8	1464	173,520	260,072
9	1564	197,570	296,118
10	1716	226,030	338,774
11	1855	250,190	374,985
12	2084	272,290	408,109
13	2159	296,960	445,084
14	2308	327,985	491,584
15	2460	357,015	535,094
20	3930	471,990	707,419
30	5680	767,840	1150,839
40	7564	1060,235	1589,081
50	9557	1329,590	1992,791

Capítulo 6

Conclusões e Trabalhos Futuros

Um *corpus* de documentos foi construído a partir das dissertações mais recentes de diferentes IES do Brasil, manualmente extraídas dos endereços eletrônicos de cada programa de mestrado analisado. O *corpus* está composto por 417 documentos correspondentes à 23 IES do Brasil (com um intervalo de anos de elaboração desde 2013 até 2023), dos quais foram analisados os *abstracts*. O vocabulário contém 5935 palavras únicas, num total de 61892 palavras.

Os métodos MCMC foram implementados para aprender os parâmetros ($\boldsymbol{\theta}$ e $\boldsymbol{\beta}$) dos modelos LDA, permitindo a inferência dos tópicos latentes e identificação dos tópicos presentes em cada documento. Foi selecionado o número de tópicos igual a 40, utilizando a medida de coerência (C_{UCI}), garantindo uma representação adequada dos temas abordados nas dissertações, os quais foram amostrados na Tabela 5.3; mesmo que a convergência das cadeias do $\boldsymbol{\beta}$ não foi obtida, mas com uma estimativa da variabilidade e amplitude de valores quase insignificantes, isso depois de observar o modelo LDA treinado com 500000 iterações. Alguns dos tópicos inferidos são da Estatística (como o 2, 5, 12, entre outros), da Saúde Pública (como o 15, 18, 23, entre outros), da Matemática (como o 3, 13, 24, entre outros); as três áreas mais relevantes na abordagem do MCMC, como foi mostrado na Tabela 5.7.

A inferência variacional foi implementada também para aprender os parâmetros ($\boldsymbol{\theta}$ e $\boldsymbol{\beta}$) dos modelos LDA, permitindo a inferência dos tópicos latentes e identificação dos

tópicos presentes em cada documento. Foi selecionado o número de tópicos igual a 30, depois de fazer o cálculo da medida de coerência (C_{UCI}), na qual mostra-se nas Figuras 5.10 e 5.11 que os melhores a escolher seriam os de 50 e 40 tópicos, respectivamente, mas a interpretabilidade nesses tópicos não é boa para fazer a interpretação dos temas, e para fazer a escolha de algum destes. Neste cenário, a escolha final foi baseada no cálculo da log-verossimilhança, mostrada na Figura 5.9, o que levou a fazer a escolha do modelo LDA com 30 tópicos, e depois de observar as palavras que compõem cada um dos tópicos, foi mais fácil inferir os seus assuntos, garantindo uma representação adequada dos temas abordados nas dissertações, os quais foram amostrados na Tabela 5.6. Alguns dos tópicos inferidos são da Estatística (como o 1, 3, 5, entre outros), das Ciências da Computação (como o 4, 9, 10, entre outros), da Saúde Pública (como o 11, 14, 17, entre outros); as três áreas mais relevantes na abordagem da inferência variacional, como foi mostrado na Tabela 5.7.

Observa-se na abordagem via MCMC como na abordagem via inferência variacional que a principal área dos tópicos aprendidos é a Estatística, como teria que ser esperado lembrando que o *corpus* foi conformado por dissertações recentes de mestrado em Estatística e áreas correlatas no Brasil; outras áreas principais presentes foram Saúde Pública, Ciências da Computação, Matemática e Epidemiologia.

Uma vez inferidos os tópicos nas dissertações que compõem o *corpus* nas duas abordagens, é possível fazer uma análise detalhada automatizada de cada um dos documentos, obtendo-se por exemplo: documentos similares, os seus principais tópicos igualmente com as proporções de cada tópico. Além, é claro da interpretação prática do conteúdo de cada tópico latente a partir dos vetores $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$.

A inferência variacional, sendo mais rápida, apresentou uma distribuição mais generalista dos tópicos, o que pode ser útil em grandes volumes de dados, porém, em alguns casos, perdeu detalhes sutis, razão pela qual na abordagem via MCMC tem-se mais áreas que estão compreendidas nos tópicos aprendidos na mesma, que não estão nos tópicos

aprendidos na inferência variacional. Essa comparação permite que destaque-se a eficiência da inferência variacional e a precisão do MCMC, no entanto tendo em conta tudo o acontecido no momento de fazer a escolha do melhor modelo LDA nas duas abordagens.

Ao longo desta dissertação, abordou-se a grande tarefa de escolher o número de tópicos justo para o modelo LDA, uma tarefa que se revelou complexa devido à diversidade de critérios disponíveis para a comparação de modelos. Estes critérios, que incluem medidas como a perplexidade e a medida de coerência dos tópicos, podem levar a escolhas diferentes quanto ao número ideal de tópicos. Essa variabilidade destaca a dificuldade intrínseca na seleção desse parâmetro e sublinha a importância de considerar, não apenas as métricas estatísticas, mas também a interpretabilidade prática dos tópicos inferidos. De fato, a capacidade de interpretar de forma coerente os tópicos extraídos é essencial para a utilidade do modelo em aplicações do mundo real.

Como trabalhos futuros, uma das abordagens a ser considerada é a ampliação do conjunto de dados utilizado, incorporando também o capítulo de introdução das dissertações. Essa ampliação pode resultar em uma melhoria substancial na qualidade dos tópicos gerados, uma vez que os capítulos de introdução vão aumentar a frequência das palavras que poderiam ser relevantes no momento da inferência dos tópicos no modelo LDA. Ao incluir esse tipo de informação adicional, espera-se que os tópicos inferidos sejam mais interpretáveis, pois as palavras que os compõem vão ser mais prováveis ou com melhor pontuação do termo, eliminando a pouca informação (na frequência de palavras relevantes) que pode acontecer com somente considerar para o aprendizado dos tópicos, os resumos em inglês (*abstracts*). Seria também importante, retirar as palavras da base de dados, que poderiam-se considerar de maneira subjetiva que não são significativas para inferir os tópicos no modelo LDA, isto é, depois de já ter removido as *stopwords*.

No uso do algoritmo do amostrador de Gibbs, é recomendável utilizar uma inicialização aleatória das atribuições de tópicos, o que poderia favorecer uma melhor mistura da cadeia de Markov e promove uma exploração mais eficiente do espaço dos parâmetros a posteriori

desde o início das iterações, o que poderia evitar problemas como a introdução de um forte viés no estado inicial do algoritmo; tendo como consequência uma convergência mais lenta, uma exploração inadequada do espaço de tópicos e, em alguns casos, resultados pouco representativos, especialmente quando o número de iterações é limitado.

No contexto da análise de convergência dos parâmetros estimados a partir do amostrador de Gibbs, não é recomendável inspecionar individualmente as cadeias de cada componente das matrizes de $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$, devido à sua alta dimensionalidade levando que muitas sejam próximas de zero. Em vez disso, sugere-se utilizar medidas-resumo globais, como a verossimilhança ou a densidade a posteriori, para avaliar a estabilidade das amostras e inferir a convergência do modelo de forma mais eficaz.

Também fazer a aplicação de métodos de inferência mais avançados, como o método de inferência via Monte Carlo Hamiltoniano (HMC), descrito por Neal [2012]. Esse método, que utiliza o No U-Turn Sampler (NUTS) desenvolvido por Hoffman and Gelman [2014], é conhecido por sua capacidade de oferecer uma convergência mais eficiente e precisa em comparação com o tradicional amostrador de Gibbs. Ao comparar esse enfoque com o amostrador de Gibbs implementado na abordagem do MCMC, espera-se observar uma melhoria significativa na convergência das cadeias, o que pode resultar em estimativas mais robustas e precisas para a inferência dos tópicos.

Os arquivos dos scripts elaborados e utilizados na linguagem de programação R [R Core Team, 2023], e os arquivos com a extensão .csv de tabelas com informação da base de dados e do *corpus* de documentos, estão disponíveis no enlace: `https://github.com/JuanPabloAO/Codigos_no_R.git`. O arquivo com a extensão .csv da base de dados final (correspondente ao objeto no R com o nome "data"), tem um tamanho de 441 KB.

Referências Bibliográficas

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado. 42

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 18

Blei, D. M. (2009). Topic models. Technical Report MIT-CSAIL-TR-2009-008, Massachusetts Institute of Technology. xii, 2, 47, 49, 55, 56, 61

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84. ix, 21

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877. 16, 18

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17 – 35. 20

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022. 1, 2, 44

Chen, J., Rubin, E., and Cornwall, G. (2021). *Data Science for Public Policy*. Springer Series in the Data Sciences. Springer International Publishing. 42

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335. 25

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409. 25

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition. 13, 14

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741. 2, 25

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC. 2

Griffiths, T. and Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 381–386. 2

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. 25

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347. 2, 39

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623. 69

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233. 15, 16, 18

Kim, J. N. (2020). Variational expectation-maximization algorithm in posterior distribution of a latent dirichlet allocation model for research topic analysis. *Expert Systems with Applications*, 150:113274. 2

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474. 16

Kullback, S. (1959). *Information Theory and Statistics*. Wiley publication in mathematical statistics. Wiley. 17

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86. 17

Lee, S. Y. (2021). Gibbs sampler and coordinate ascent variational inference: A set-theoretical review. *Communications in Statistics - Theory and Methods*, 51(6):1549–1568. 13, 18

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092. 25

Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. *Technical Report*, pages 91–09. 25

Neal, R. (2012). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*. 69

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 5, 41, 69

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search*

and Data Mining (WSDM '15), pages 399–408, New York, NY, USA. Association for Computing Machinery. 44, 46

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. 42

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. 42

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594. 42

Apêndice A

Resultados adicionais na implementação da abordagem via MCMC e da abordagem via inferência variacional

Tabela A.1: Valores calculados dos critérios de informação AIC, BIC, DIC e WAIC; para os modelos LDA com 2 à 15, 20, 30, 40 e 50 tópicos, na abordagem via MCMC.

	AIC	BIC	DIC	WAIC
Modelo com 2 tópicos	899461, 9	1014112	874358, 4	4428, 378
Modelo com 3 tópicos	900428, 6	1072404	862823	4308, 41
Modelo com 4 tópicos	904179, 5	1133481	854008, 1	4213, 178
Modelo com 5 tópicos	911929, 7	1198556	849163, 7	4158, 69
Modelo com 6 tópicos	914514, 9	1258466	839190, 4	4063, 783
Modelo com 7 tópicos	923960, 4	1325237	835980, 3	4026, 325
Modelo com 8 tópicos	929818, 6	1388421	829166, 8	3961, 341
Modelo com 9 tópicos	942189, 2	1458117	828964, 1	3953, 772
Modelo com 10 tópicos	948724, 9	1521978	822776, 9	3893, 969
Modelo com 11 tópicos	957979, 1	1588557	819394, 6	3863, 36
Modelo com 12 tópicos	964298, 5	1652202	813052, 6	3805, 17
Modelo com 13 tópicos	975975, 8	1721204	812071, 6	3790, 436
Modelo com 14 tópicos	984458, 8	1787012	807845	3753, 784
Modelo com 15 tópicos	994137, 5	1854016	804896, 8	3726, 095
Modelo com 20 tópicos	104866	2195165	795978, 9	3638, 251
Modelo com 30 tópicos	1157044	2876801	777419, 9	3466, 014
Modelo com 40 tópicos	1273188	3566198	766564, 3	3367, 382
Modelo com 50 tópicos	1389148	4255411	755534, 4	3272, 519

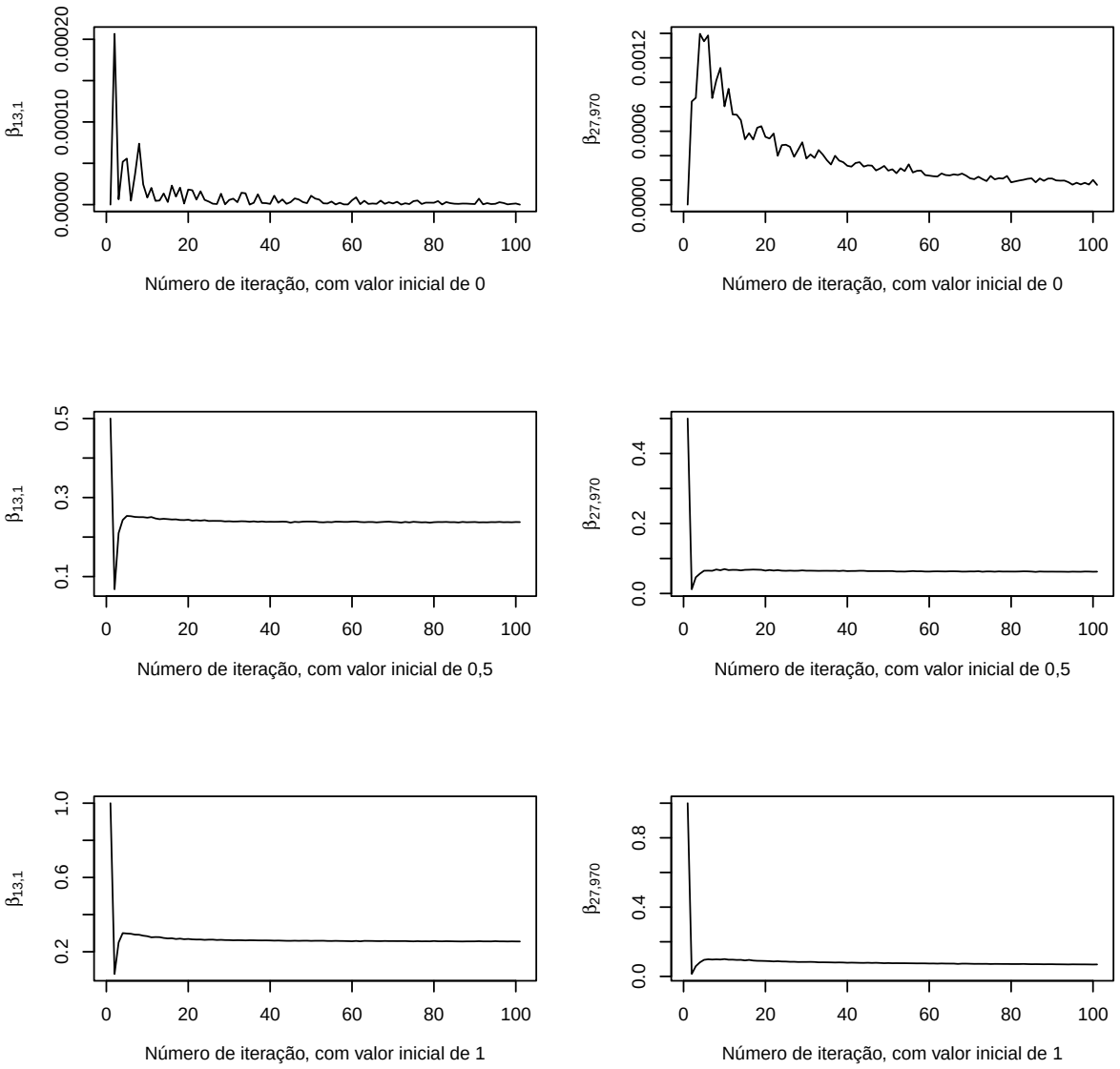


Figura A.1: Comportamento das 100 primeiras iterações (sem aquecimento) de dois componentes (kv) de $\boldsymbol{\beta}$, simuladas para um modelo LDA treinado no MCMC, com $K = 40$, espaçamento de 5; para valores iniciais nessas componentes (β_{kv}) iguais à 0, 0, 5 e 1.

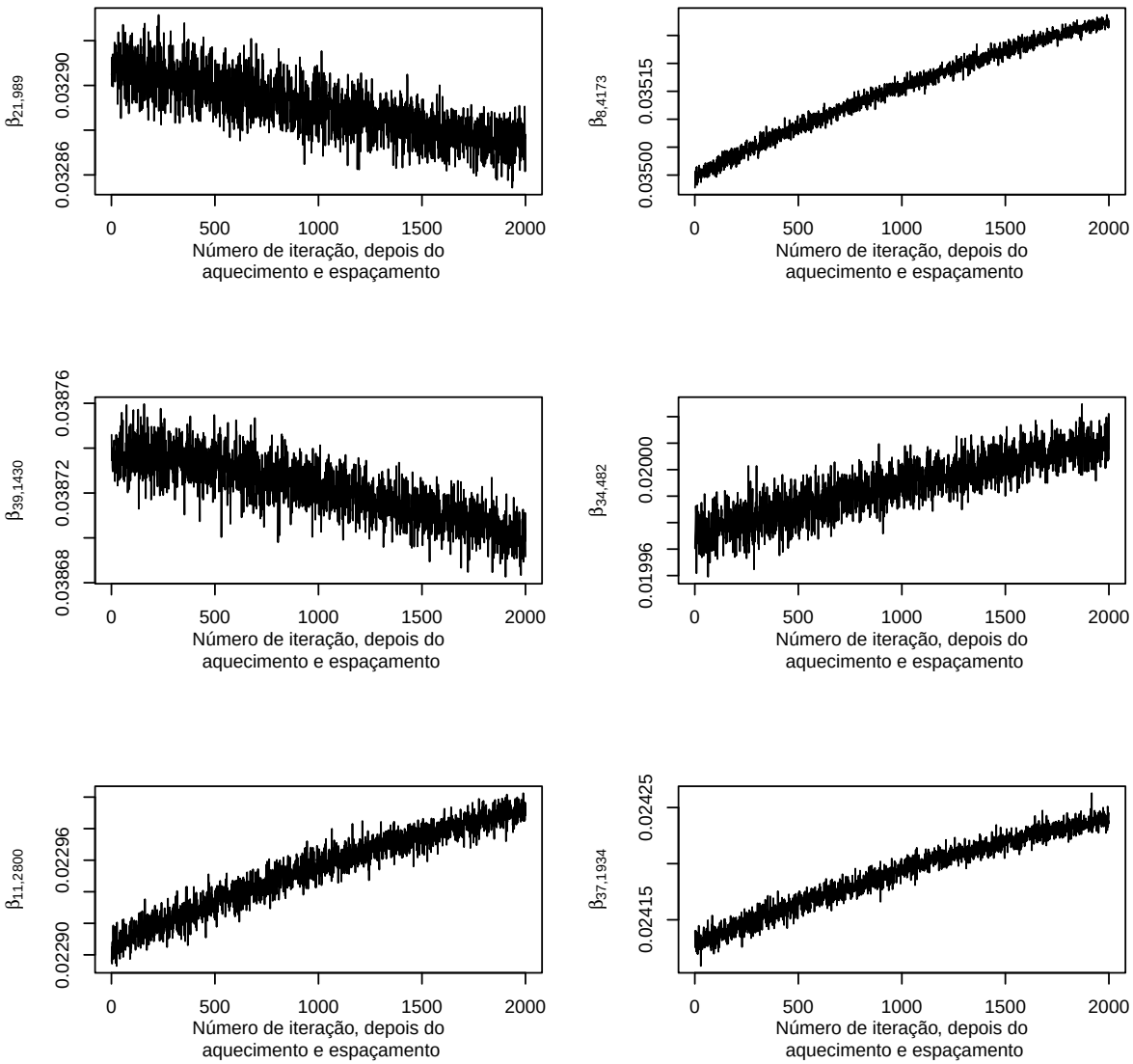


Figura A.2: Cadeias de algumas das componentes de $\hat{\boldsymbol{\beta}}$, sendo as componentes com maior pontuação do termo; desde o modelo LDA treinado no MCMC com 500000 iterações, $K = 40$, espaçamento de 100 e aquecimento de 3000.

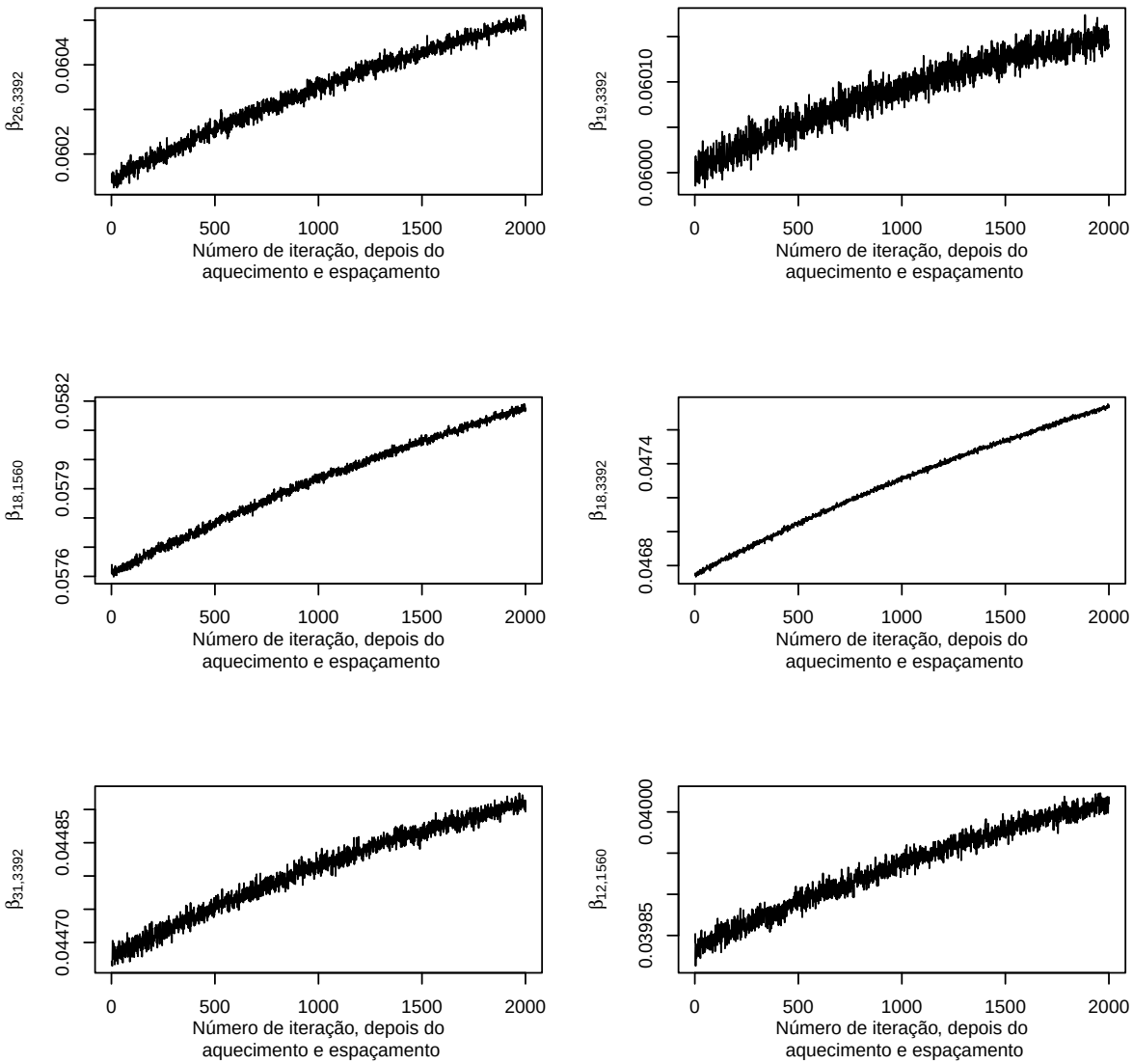


Figura A.3: Cadeias de algumas das componentes de $\hat{\boldsymbol{\beta}}$, sendo as componentes com maior probabilidade; desde o modelo LDA treinado no MCMC com 500000 iterações, $K = 40$, espaçamento de 100 e aquecimento de 3000.

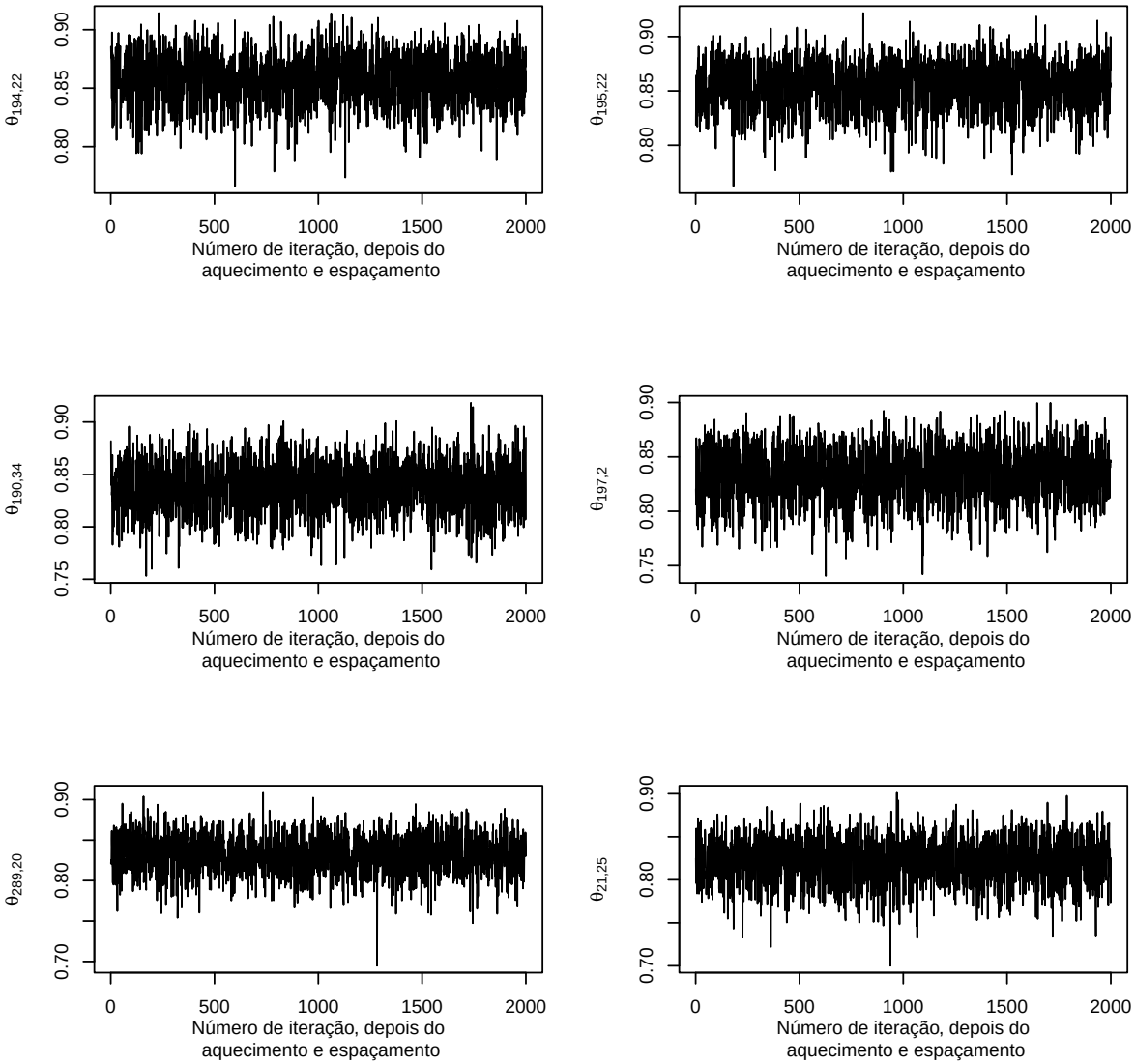


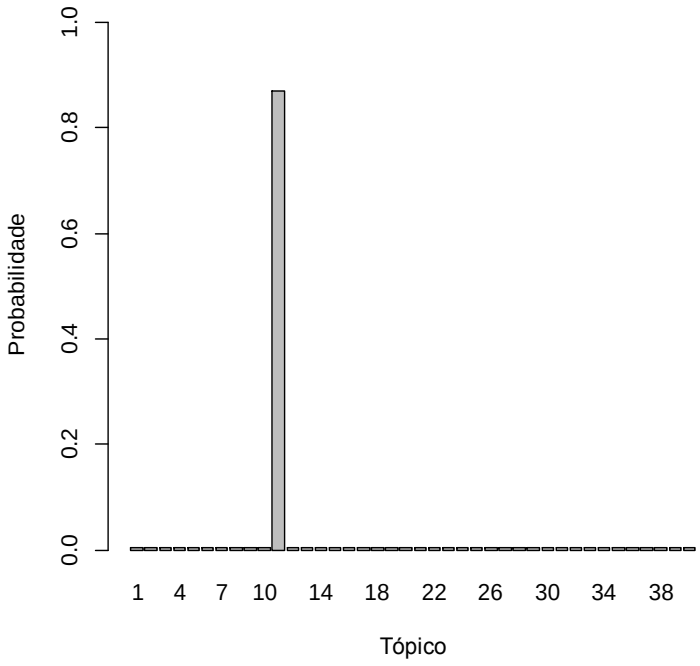
Figura A.4: Cadeias de algumas das componentes de $\hat{\boldsymbol{\theta}}$, sendo as componentes com maior probabilidade; desde o modelo LDA treinado no MCMC com 500000 iterações, $K = 40$, espaçamento de 100 e aquecimento de 3000.

Documento 194: Órteses, Próteses e Meios auxiliares de locomoção (OPM) - 2021- IES: UFPB

Principais palavras dos principais tópicos (por pontuação do termo)

	Tópico 11 (prob. 0.871)	Tópico 29 (prob. 0.005)	Tópico 3 (prob. 0.004)
1	opm	violenc	algebra
2	synthet	municip	automorph
3	beta	health	varieti
4	workshop	pmaq	free
5	dispens	primari	porou
6	limb	transfer	flow
7	auxiliari	manag	rehabilit
8	data	resourc	categori
9	orthot	care	finit
10	orthoped	team	low
11	valu	usf	prematur
12	suppli	famili	discharg
13	analysi	cartograph	group
14	locomot	financi	medium
15	product	outlier	strongli

Proporções de tópico esperadas



Top 10 de documentos semelhantes

	10 DOCUMENTOS MAIS SIMILARES	DISTÂNCIA DE HELLINGER
1	151- Imputação de Dados Sintéticos Através de Árvores de - 2019- IES: UFMG	0.081
2	223- Variabilidade e aderência em modelos - 2022- IES: UFPE	0.092
3	154- Métodos de simulação de dados geográficos sintéticos - 2018- IES: UFMG	0.165
4	89- O Modelo Log-Poly-Weibull Multivariado Aplicado à Predição - 2017- IES: UFAM	0.312
5	86- Modelo de Regressão Beta para Diferença de Proporções- 2022- IES: UFAM	0.596
6	106- Modelo de regressão beta retangular para análise de dados - 2019- IES: UFC	0.650
7	149- Comparação entre Algoritmos MCMC para Inferência - 2019- IES: UFMG	0.674
8	146- A Necessidade de Classificações Repetidas no Modelo - 2019- IES: UFMG	0.675
9	416- Uma Introdução ao Processo de Nascimento e Assassinato- 2021- IES: USP	0.679
10	115- A quantile regression model based on an owen distribution- 2022- IES: UFCG	0.682

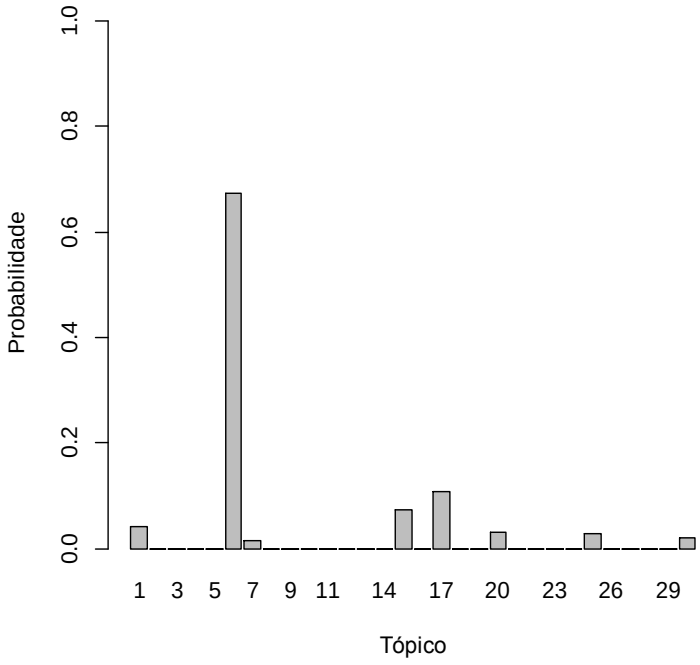
Figura A.5: Resultados obtidos do documento 194, com a função criada na linguagem de programação R, no modelo LDA treinado no MCMC.

Documento 106: Modelo de regressão beta retangular para análise de dados - 2019- IES: UFC

Principais palavras dos principais tópicos (por pontuação do termo)

	Tópico 6 (prob. 0.673)	Tópico 17 (prob. 0.108)	Tópico 15 (prob. 0.074)
1	distribut	statist	techniqu
2	data	covid	algebra
3	estim	result	theori
4	method	diseas	decis
5	model	sever	real
6	propos	develop	student
7	regress	research	recurr
8	paramet	data	comput
9	simul	individu	gamma
10	gener	can	markov
11	sampl	work	carri
12	perform	process	usual
13	studi	associ	bound
14	normal	case	gener
15	function	qualiti	correct

Proporções de tópico esperadas



Top 10 de documentos semelhantes

	10 DOCUMENTOS MAIS SIMILARES	DISTÂNCIA DE HELLINGER
1	230- Estimação de processos com longa dependência na presença- 2023- IES: UFRGS	0.128
2	162- Uma Abordagem via Mistura Finita para Modelos de - 2020- IES: UFMG	0.146
3	325- Modelo de regressão aleatória utilizando o software R- 2022- IES: UFV	0.151
4	358- Objective Bayesian Analysis of Process - 2021- IES: UNESP	0.163
5	154- Métodos de simulação de dados geográficos sintéticos - 2018- IES: UFMG	0.165
6	157- Modelos de Séries Temporais Semiparamétricos com Fator Latente- 2020- IES: UFMG	0.174
7	226- Classificação com inferência para dados de alta dimensão- 2022- IES: UFRGS	0.174
8	335- Intervalos de confiança bootstrap para algumas estatísticas não - 2022- IES: UnB	0.177
9	291- A nonparametric bayesian approach for modeling and - 2022- IES: UFSCar	0.183
10	398- Previsão da temperatura máxima de Lavas-MG comparação da normal- 2022- IES: UNIFAL	0.188

Figura A.6: Resultados obtidos do documento 106, com a função criada na linguagem de programação R, no modelo LDA treinado na inferência variacional.