UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelo Hierárquico Bayesiano Não Paramétrico Aplicado em Modelagem de Tópicos

Robson Ortz Oliveira Cunha

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

ICMC USP SÃO CARLOS

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

Robson Ortz Oliveira Cunha

Modelo Hierárquico Bayesiano Não Paramétrico Aplicado em Modelagem de Tópicos

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

Orientador: Prof. Dr. Rafael Bassi Stern

São Carlos
Março de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C972m

Cunha, Robson Ortz Oliviera
Modelo Hierárquico Bayesiano Não Paramétrico Aplicado em Modelagem de Tópicos / Robson Ortz Oliviera Cunha; orientador Rafael Bassi Stern. -- São Carlos, 2024.
89 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Estatística Bayesiana. 2. Modelo Hierárquico Não Paramétrico. 3. Modelo de Tópico. I. Bassi Stern, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

FEDERAL UNIVERSITY OF SÃO CARLOS
CENTER FOR EXACT SCIENCES AND TECHNOLOGY
INTERAGENCY PROGRAM GRADUATE IN STATISTICS UFSCar-USP

Robson Ortz Oliveira Cunha

Nonparametric Bayesian Hierarchical Model Applied to Topic Modeling

Thesis presented to the Department of Statistics - Des/UFSCar and the Institute of Mathematics and Computer Sciences - ICMC-USP, as part of the requirements for obtaining the title of Master or Doctor in Statistics - Interagency Program Graduate in Statistics UFSCar-USP.

Advisor: Prof. Dr. Rafael Bassi Stern

São Carlos
March 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

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

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Robson Ortz Oliveira Cunha, realizada em 19/02/2024.

Comissão Julgadora:

Prof. Dr. Rafael Bassi Stern (IME-USP)

Prof. Dr. Marcos Oliveira Prates (UFMG)

Prof. Dr. Luis Gustavo Esteves (IME-USP)

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

Dedico este trabalho a todos aqueles que, em algum momento da vida, subjugaram seus próprios corpos por medo da condenação do seu ser.

AGRADECIMENTOS

Contei com o apoio de muitas pessoas que me fizeram mais forte durante esta trajetória e tornaram-na mais leve. Começo agradecendo imensamente meu melhor amigo Patrick de Oliveira e minha madrinha Uilsiene Vigneron, por me incentivarem a seguir em frente, mesmo diante as adversidades e, sobre tudo, por me mostrarem meu valor, mesmo diante do fracasso.

Ao meu orientador, Prof. Dr. Rafael Bassi Stern, agradeço por sempre demonstrar-se paciente durante todo processo de orientação, por compreender os desafios que enfrentei ao conciliar meu trabalho com o projeto de pesquisa e, especialmente, por desestigmatizar meu próprio modo de articular ideias.

O desenvolvimento deste trabalho também só foi possível graças ao meu até então gestor, Helton Alponti, que além do incentivo em realizá-lo, sugeriu ideias que motivaram o tema do mesmo.

Por fim, agradeço todos meus amigos que, de alguma forma, mostraram-se atenciosos e presentes nesta jornada.

“Sic gorgiamus allos subjectatos nunc.”

(Addams Family - 1991)

“Miracles aren’t something you wish for.
You seize them using your own strength!”

(Claymore, Vol. 23: Mark of the Warrior)

RESUMO

CUNHA, R. O. **Modelo Hierárquico Bayesiano Não Paramétrico Aplicado em Modelagem de Tópicos**. 2024. 89 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, São Carlos – SP, 2024.

Dada a crescente necessidade e importância da análise de dados textuais no ramo da inteligência artificial, modelos que possam compreender melhor a linguagem humana e lidar com dados não estruturados têm ganhado cada vez mais relevância. Neste trabalho, desenvolvemos um estudo sobre o Processo Hierárquico de Dirichlet (HDP) na modelagem de tópicos textuais, explorando seus aspectos práticos ao aplicá-lo em um conjunto de dados (*corpus*) de processos jurídicos, compostos por três tipos de procedimentos distintos. Discorremos sobre as principais propriedades do HDP, sobre a ótica Bayesiana, assumindo que os dados sejam oriundos de uma distribuição de probabilidade Multinomial, baseados no modelo de representação textual de *bag-of-words*, comumente utilizado em processamento de linguagem natural. Procedemos ainda com algumas técnicas de pré-processamento textual, que resultaram em documentos (dados) mais parcimoniosos, e com um estudo de simulação para verificar a performance do modelo. Ao fim do trabalho, apresentamos os resultados das aplicações realizadas e discutimos sobre a problemática da análise de dados em jurimetria.

Palavras-chave: Modelo Não Paramétrico Bayesiano, Processo Hierárquico de Dirichlet, Jurimetria, Modelagem de Tópicos Textuais..

ABSTRACT

CUNHA, R. O. **Nonparametric Bayesian Hierarchical Model Applied to Topic Modeling**. 2024. 89 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, São Carlos – SP, 2024.

Given the growing need and importance of analyzing textual data in the field of artificial intelligence, models that can better understand human language and deal with unstructured data are increasingly relevant gains. In this work, we developed a study on the Hierarchical Dirichlet Process (HDP) in modeling textual topics, exploring its practical aspects by applying it to a data set (*corpus*) of legal processes, composed of three types of different procedures. We will discuss the main properties of HDP, from a Bayesian perspective, assuming that the data comes from a Multinomial probability distribution, based on the *bag-of-words* textual representation model, commonly used in natural language processing . We also proceeded with some textual pre-processing techniques, which resulted in more parsimonious documents (data), and with a simulation study to verify the model's performance. At the end of the work, we present the results of the applications carried out and discuss the issues of data analysis in jurimetry.

Keywords: Non-Parametric Bayesian Model, Hierarchical Dirichlet Process, Topic Modeling, Jurimetry..

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O estudo da linguística é uma área que apresenta inúmeros desafios para análise, processamento e modelagem de dados, uma vez que a língua (seja ela falada, escrita, ou composta por uma coleção de símbolos visuais) consiste em uma fonte não estruturada de informação. Enquanto na Computação o Processamento de Linguagem Natural (NLP, do inglês *Natural Language Processing*) tem explorado técnicas para tratar, processar e minerar tais dados, na Estatística modelos bayesianos não paramétricos têm sido uma excelente alternativa na solução dos principais problemas da linguística. Neste trabalho, iremos discorrer sobre o modelo hierárquico não paramétrico por Processo de Dirichlet (conhecido simplesmente como Processo Hierárquico de Dirichlet) como método para modelagem de tópicos textuais.

Processos Hierárquicos de Dirichlet (HDP, do inglês *Hierarchical Dirichlet Process*) consiste em um modelo de associação mista para análise de agrupamentos, não supervisionada, de dados textuais. O mesmo pressupõe um número infinito de tópicos (grupos), que são compartilhados sobre uma estrutura de dependência hierárquica entre seus processos. Sua estrutura é formada de processos aleatórios denominados Processos de Dirichlet, desenvolvidos por Ferguson (1973) e Sethuraman (1994), e estendida por Teh e Jordan (2009) para sua estrutura hierárquica conjugada.

Devido a técnicas computacionais derivadas do método de Monte Carlo para Cadeias de Markov (MCMC), aliadas a estrutura conjugada das densidades envoltas no processo, o HDP tornou-se computacionalmente conveniente. Através de um amostrador de *Gibbs*, por exemplo, é possível obter amostras da sua distribuição a posteriori que, diferentemente de outros processos, realiza sorteios sobre uma distribuição de distribuições, Neal (2000). Em outras palavras, a medida aleatória nos processos do modelo efetua sorteios sobre uma distribuição de probabilidade, cujas massas pontuais também são densidades. O ajuste do HDP por *Gibbs* também pode ser representado por abstrações como o processo de quebra-bastão (*stick-breaking*) e pelos processos de restaurantes de chineses, fomentados em Teh (2010) sobre o Processo

de Dirichlet e estendido para o HDP, posteriormente, em Teh e Jordan (2009). Quanto aos parâmetros envoltos no HDP, Liu e Nandram (2022) apresentou alguns métodos de determiná-los, todos envolvendo de alguma forma amostragem por rejeição, possibilitando atualizá-los ao longo do processo.

Outra contribuição importante está presente no trabalho de Frank, Greenberg e Lindner (2020). Nele, o autor não só explora as diferentes estruturas conjugadas da distribuição a priori sobre a verossimilhança, como expõe a construção de suas respectivas funções preditivas. Tais distribuições são importantes, uma vez que elas atualizam as camadas latentes dos processos do HDP, responsáveis pelos agrupamentos, em cada nível hierárquico e a cada iteração do amostrador de *Gibbs*.

Das principais característica do HDP, apresentadas por Teh e Jordan (2009), destacamos a sua capacidade de representar o processo, sobre uma distribuição contínua conjugada, por uma forma discretizada do mesmo. Esta característica, além de nos auxiliar na compreensão das realizações sobre uma distribuição probabilística de distribuições, nos revela a forma com que os elementos textuais, agrupados pelos tópicos, são intercambiáveis entre os mesmos.

Quanto aos dados, motivados pela aplicação de técnicas estatística no campo jurídico (disciplina esta conhecida Jurimetria, Nunes (2019)), usamos um *corpus* de processos jurídicos, composto por 18799 documentos, para aplicarmos o HDP como modelo de tópicos. Os documentos foram divididos em três grupos, segundo o procedimento aplicado em cada processo: procedimento ordinário (10471 documentos); procedimento sumário (2051 documentos) e procedimento especial (6277 documentos). Dentre as etapas de tratamento dos dados, o pré-processamento textual mostrou-se extremamente importante, não só pelas particularidades da língua portuguesa, como pelas características dos textos jurídicos, fazendo com que os métodos apresentados por Palmer (2010) fossem adequados a todo o contexto. Além disso, o método de incorporação textual escolhido, TF-IDF (*Term Frequency Inverse Document Frequency*), apresentado por Gibbons e Chakraborti (2014), mantém a mensuração dos elementos textuais através de frequências de palavras (em NLP denominadas de *tokens*).

Com a aplicação do modelo nos dados, analisamos os tópicos obtidos e mensuramos sua qualidade através da métrica de coerência, apresentada em Röder, Both e Hinneburg (2015), que tem como princípio mensurar o quanto um tópico é consistente em relação aos seus termos. Sobre as análises, percebemos que os tópicos mostravam-se bastante interessante quanto aos temas indicados pelos *tokens*, além do modelo ter apresentado bons resultados de convergência quanto a *log-verossimilhança* calculada em cada iteração. Porém, consideramos a coerência calculada sobre cada um dos tópicos como indício de eventuais desafios que o modelo encontra em relação ao tipo de dado na área jurídica, nos levando a apresentar propostas de representação textual capazes de considerar o contexto e não somente a frequência relativa das palavras.

Para isso, estruturamos o trabalho da seguinte forma. No Capítulo 2 apresentamos o arcabouço teórico: uma breve revisão sobre Estatística Bayesiana; seguido da construção do

Processo de Dirichlet (DP) e como geramos o modelo HDP a partir de DP's; e finalizando com estratégias computacionais com o amostrador de *Gibbs*. Em seguida, no Capitulo 3, contextualizamos o uso do HDP na modelagem de tópicos; apresentamos métodos de pré-processamento e representação textual sobre o uso da frequências das palavras; e tratamos das preditivas, usadas na atualização das amostras de tópicos pelo processo, sobre a estrutura probabilística conveniente para os dados. Já no Capítulo 4, tratamos do uso do modelo nos dados jurídicos e analisamos seus resultados. Finalmente, tomando os resultados analisados, propomos uma extensão do trabalho aqui desenvolvido, Capítulo 5.

CAPÍTULO

2

METODOLOGIA

2.1 Estatística Bayesiana

Fazer inferência sobre um conjunto de dados requer quantificar as incertezas envoltas nos experimentos que geram tais dados. Dessa forma, a Inferência Bayesiana utiliza de uma estrutura de modelos probabilísticos como arcabouço para representar diferentes graus de incerteza sobre as quantidades de interesse (parâmetros) do processo.

Considerando um conjunto de $n > 0$ elementos em X, onde X é uma variável aleatória pertencente ao espaço amostral χ, sendo $X_1, X_2, \ldots, X_n$ ensaios independentes e identicamente distribuídos (*i.i.d.*) conforme X. Tomando quantidades de interesse desconhecidas (parâmetros) $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, para $0 < p \leq n$, pertencentes ao espaço paramétrico Θ do modelo. Podemos codificar as incertezas sobre esses parâmetros através da sua probabilidade conjunta especificada por dois elementos Esteves, Izbicki e Stern (2021):

- A probabilidade marginal, ou distribuição a priori, dos parâmetros, denotada por $f(\theta)$. Sendo esta responsável por sumarizar, probabilisticamente, as informações conhecidas em torno desses parâmetros (hipóteses).
- A distribuição amostral condicionada θ e denotada por $f(\mathbf{x}|\theta)$. Isto é, a função de verossimilhança que atualiza as hipóteses quantificadas pela priori, codificando a informação sobre θ presente nos dados.

Ao serem combinados, esses elementos estabelecem a estrutura do Teorema de Bayes.

$$f(\theta|\mathbf{x} = (x_1, x_2, \ldots, x_n)) = \frac{\prod_{i=1}^{n} f(\theta) f_X(x_i|\theta)}{\int_\Theta \prod_{i=1}^{n} f_X(x_i|\theta) f(\theta) d\theta} \tag{2.1}$$

Onde $f(\theta|\mathbf{x})$ é denominada como distribuição a posteriori e contém toda a incerteza que desejamos mensurar sobre θ .

Segundo Gelman *et al.* (2021), a quantidade $f_X(\mathbf{x}) = \int_\Theta \prod_{i=1}^n f_X(x_i|\theta) f(\theta) d\theta$ trata-se da distribuição marginal de X, denominada como *distribuição preditiva a priori* (ou simplesmente, *distribuição preditiva*). Por sua vez, esta não depende de quaisquer θ e pode ser tratada como uma constante, já que a integração em $f(\theta|X = x) = 1$. No entanto, a distribuição preditiva pode ser bastante útil uma vez que, ao observamos uma amostra $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, podemos prever observações futuras $\widetilde{\mathbf{x}} = (x_{n+1}, x_{n+2}, \ldots, x_{n+m})$, para $m > 0$, ao condicioná-las em $\mathbf{x}$. Assim obtemos,

$$
\begin{aligned}
f(\widetilde{\mathbf{x}}|\mathbf{x}) &= \frac{f(x_1, x_2, \ldots, x_{n+m})}{f(x_1, x_2, \ldots, x_n)} \\
&= \frac{\int_\Theta \prod_{i=1}^{n+m} f_X(x_i|\theta) f(\theta) d\theta}{\int_\Theta \prod_{i=1}^{n} f_X(x_i|\theta) f(\theta) d\theta} \\
&= \int_\Theta \prod_{i=n+1}^{n+m} f_X(x_i|\theta, \mathbf{x}) \frac{\prod_{i=1}^{n} f_X(x_i|\theta) f(\theta)}{\int_\Theta \prod_{i=1}^{n} f_X(x_i|\theta) f(\theta) d\theta} d\theta \\
&= \int_\Theta \prod_{i=n+1}^{n+m} f_X(x_i|\theta, \mathbf{x}) f(\theta|\mathbf{x}) d\theta \qquad \text{(por 2.1)} \\
&= \int_\Theta f(\widetilde{\mathbf{x}}|\theta) f(\theta|\mathbf{x}) d\theta. \qquad (2.2)
\end{aligned}
$$

Onde $\int_\Theta f(\widetilde{\mathbf{x}}|\theta) f(\theta|\mathbf{x}) d\theta$ trata-se da *distribuição preditiva a posteriori*. Futuramente, essa distribuição nos será valiosa ao estimar quantidades pelo processo hierárquico, via algoritmo de *Gibbs* em sua etapa de atualização.

Apesar da utilidade da distribuição preditiva ao prever novas observações, vimos que ela depende exclusivamente dos dados observados e, em alguns casos, é possível determiná-la sem calcular diretamente seu valor. Podemos simplificar a equação (2.1) da seguinte forma:

Definição 1. Sejam $f(\cdot)$ e $g(\cdot)$ funções de distribuições de probabilidade e seja $C \in \mathbb{R}$ uma constante, tal que $f(x) = Cg(x)$, para algum x observado. Dizemos que:

$$f(x) \propto g(x).$$

Portanto,

$$f(\theta|\mathbf{x}) \propto \prod_{i=1}^{n} f(\theta) f_X(x_i|\theta). \qquad (2.3)$$

2.1.1 Família de Distribuições Conjugadas

Anteriormente vimos as implicações a cerca das distribuições que constituem o Teorema de Bayes, sobre tudo a distribuição preditiva. Discutimos brevemente que esta pode ser útil no cálculo de predições de dados futuros e, para uma gama de modelos estatísticos, ela pode ser facilmente obtida permitindo a simplificação do Teorema em (2.3).

A proporcionalidade da equação (2.3) é determinada pela seguinte definição:

Definição 2. Considere P uma família de funções, não-negativas e integráveis, sobre Θ, tome ainda a densidade condicional $f(x|\theta)$. Dizemos que,

$$P \text{ é conjugada para } f(x|\theta) \text{ se, para todo } x \in \chi,\ f(x|\theta)f(\theta) \in P.$$

Isto é, se os dados possuem uma distribuição tal qual sua verossimilhança $f(x|\theta)$ e a sua priori em θ está contida no conjunto P, tal que P é a família de distribuições conjugadas para $f(x|\theta)$, então sua posteriori também estará em P, Esteves, Izbicki e Stern (2021). Esse fato nos permite calcular diretamente a posteriori, uma vez que conhecemos a integral de todas as funções em P, fazendo de P um recurso computacionalmente conveniente.

Podemos estender a Definição 2 para uma amostra *i.i.d.* pelo Lema que se sucede:

Lema 1. Dado θ um vetor de parâmetros de interesse e sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória *i.i.d.* da variável X, com densidade $f_X(x|\theta)$. Se P é conjugada para $f(x_1|\theta)$ então P também será conjugada para $f(x_1, x_2, \ldots, x_n|\theta)$.

Demonstração. Demonstrado, por indução, em Esteves, Izbicki e Stern (2021). □

A seguir, apresentaremos algumas prioris conjugadas que serão fundamentais ao decorrer do nosso estudo (o Anexo A contém a definição das distribuições de probabilidade utilizadas).

BETA CONJUGADA

Seja $X \sim Bin(n, \theta)$ e $f(\theta)$ dada por

$$f(\theta) = M\theta^{a-1}(1-\theta)^{b-1}, \qquad a, b, M > 0.$$

Se tomarmos $M = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$, percebemos que $\theta \sim Beta(a, b)$. Ainda se $f(\theta) \in P$, então $f(\theta)$ é conjugada de $f(x|\theta)$, uma vez que $f(\theta|x)$ também pertence a P.

Portanto, para

$$\begin{aligned} X|\theta &\sim Bin(n, \theta) \\ \theta &\sim Beta(\alpha, \beta), \quad \text{para} (\alpha, \beta) = (a, b) \end{aligned}$$

obtemos que

$$
\begin{aligned}
f(\theta|x) &= \frac{f(\theta)f(x|\theta)}{f(x)} \\
&= \frac{B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{x}\theta^{x}(1-\theta)^{n-x}}{\int_\Theta B(\alpha,\beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{x}\theta^{x}(1-\theta)^{n-x}d\theta}, && \text{para } B(\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}{\int_\Theta \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}d\theta} \\
&= \frac{B(\alpha^{'},\beta^{'})\theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}}{B(\alpha^{'},\beta^{'})\int_\Theta \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}d\theta}, && \text{para } (\alpha^{'},\beta^{'})=(\alpha+x,\beta+n-x) \\
&= B(\alpha^{'},\beta^{'})\theta^{\alpha^{'}-1}(1-\theta)^{\beta^{'}-1}, && \text{pois } \int_\Theta B(\alpha^{'},\beta^{'})\theta^{\alpha^{'}-1}(1-\theta)^{\beta^{'}-1}d\theta=1.
\end{aligned}
$$

Finalmente,

$$\theta|X \sim Beta(\alpha^{'}=\alpha+x,\beta^{'}=\beta+n-x).$$

Além disso, considerando uma manipulação análoga a executada anteriormente, podemos determinar trivialmente que

$$f(x)=B(\alpha,\beta)\binom{n}{x}\frac{1}{B(\alpha^{'},\beta^{'})}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{n}{x}\frac{\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha+\beta+n)}.$$

DIRICHLET CONJUGADA

Seja $\mathbf{X}=(X_1,X_2,\ldots,X_k)\in\chi$ um vetor aleatório e $\theta=(\theta_1,\theta_2,\ldots,\theta_k)\in\Theta$ um vetor de parâmetros, para $k\in\mathbb{N}^*$. Se $\mathbf{X}|\theta\sim Mult(n,\theta)$ e $f(\theta)$ é dada por

$$f(\theta)=C\prod_{j=1}^{k}\theta_j^{(\alpha_j-1)},\forall\alpha_j>0 \text{ e } \sum_{j=1}^{k}\theta_j=1.$$

Ao tomarmos $C=\frac{\prod_{j=1}^{k}\Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{k}\alpha_j)}$, notamos que $\theta\sim Dir(\alpha)$. Dizemos então que $f(\theta)\in P$ é conjugada para $f(\mathbf{X}|\theta)$ se para

$$
\begin{aligned}
\mathbf{X}|\theta &\sim Mult(n,\theta) \\
\theta &\sim Dir(\alpha)
\end{aligned}
$$

temos, analogamente ao caso *Beta-Binomial*,

$$
\begin{aligned}
f(\theta|\mathbf{X}=(x_1,x_2,\ldots,x_k)) &= \frac{f(\theta)f(\mathbf{x}|\theta)}{f(\mathbf{x})} \\
&= \frac{\prod_{j=1}^{k}\theta_j^{(\alpha_j+x_j-1)}}{\int_\Theta\prod_{j=1}^{k}\theta_j^{(\alpha_j+x_j-1)}d\theta} \\
&= \frac{B(\alpha^{'})\prod_{j=1}^{k}\theta_j^{(\alpha_j+x_j-1)}}{B(\alpha^{'})\int_\Theta\prod_{j=1}^{k}\theta_j^{(\alpha_j+x_j-1)}d\theta}, && \text{para } B(\alpha^{'})\frac{\prod_{j=1}^{k}\Gamma(\alpha_j^{'})}{\Gamma(\sum_{j=1}^{k}\alpha_j^{'})} \text{ e } \alpha_j^{'}=\alpha_j+x_j \\
&= B(\alpha^{'})\prod_{j=1}^{k}\theta_j^{(\alpha_j^{'}-1)}.
\end{aligned}
$$

Isto é,

$$f(\theta|\mathbf{X}) \in P : \theta|\mathbf{X} \sim Dir(\alpha^{'} = \alpha + \mathbf{x})$$

2.2 Estatística Bayesiana Não Paramétrica

Em estatística, o significado de "não paramétrico" é em geral atribuído a técnicas e modelos cuja estrutura paramétrica não é fixa. Isto nos faz atribuir distribuições de probabilidade acerca desses parâmetros e como eles se relacionam com as covariáveis consideradas.

A aplicabilidade de métodos estatísticos não paramétricos deve-se, principalmente, ao fato de não fazerem uso de fortes pressupostos ou necessitarem de estruturas rígidas em relação as dependências a respeito dos parâmetros, Gibbons e Chakraborti (2014). No entanto, a flexibilidade que os modelos não paramétricos incorporam em sua estrutura faz com que a sua complexidade seja quase ilimitada. Isso estabelece um *trade-off* entre dimensionalidade, uma vez que esta cresce em relação a complexidade dos dados, e suaviza problemas de subajuste, Briscoe e Feldman (2011).

Em aprendizado de máquinas, modelos bayesianos não paramétricos ganham destaque, uma vez que sua abordagem, além de estabelecer condições que naturalmente flexibilizam as suposições a respeito dos parâmetros através de sua distribuição a priori, possibilita estimar os parâmetros a posteriori mitigando o sobreajuste do modelo, Teh (2010). Por sua vez, as distribuições a priori em modelos bayesianos não paramétricos possuem uma estrutura ligeiramente diferente. Geralmente, essas distribuições são escolhidas sobre um suporte contendo distribuições de probabilidade pertencentes a famílias paramétricas, delimitando o escopo e a inferência realizada pelo modelo. Em vez disso, modelos bayesianos não paramétricos expandem esse suporte ao espaço que contenha todas as distribuições de probabilidade possíveis, mas ainda mantendo os cálculos de suas posterioris tratáveis.

Dentre os modelos e métodos bayesianos não paramétricos disponíveis, apresentamos o que talvez seja o mais popular, o Processo de Dirichlet. Além de sua popularidade, o Processo de Dirichlet nos será de grande valia para o entendimento do modelo hierárquico que discutiremos na Seção 2.3.

2.2.1 Processo de Dirichlet

Segundo Teh (2010), o Processo de Dirichlet trata-se de um processo estocástico definido em um espaço de funções, onde cada passo é dado por uma distribuição aleatória extraída desse espaço. Em outras palavras, trata-se de um espaço de probabilidades Θ definido por uma distribuição sobre medidas de probabilidade que contêm certas propriedades especiais. Assim, uma amostra sobre um Processo de Dirichlet é interpretada como um sorteio de distribuições aleatórias.

Formalmente, sobre Ferguson (1973), temos:

Definição 3. Sejam Θ um espaço mensurável, $\mathscr{A}$ uma σ-álgebra em Θ e H uma medida não-nula, não-negativa e finitamente aditiva sobre $(\Theta, \mathscr{A})$. Tomando α uma constante positiva e G uma medida aleatória de probabilidade sobre o espaço de probabilidades $(\Theta, \mathscr{A}, H)$, dizemos que G é um Processo de Dirichlet sobre $(\Theta, \mathscr{A})$ com parâmetro α se, para $k = 1, 2, \ldots$ e para todo $A_k \in \mathscr{A}$, a distribuição conjunta dos $G(A_k)_{k \geq 1}$ segue uma Distribuição de Dirichlet:

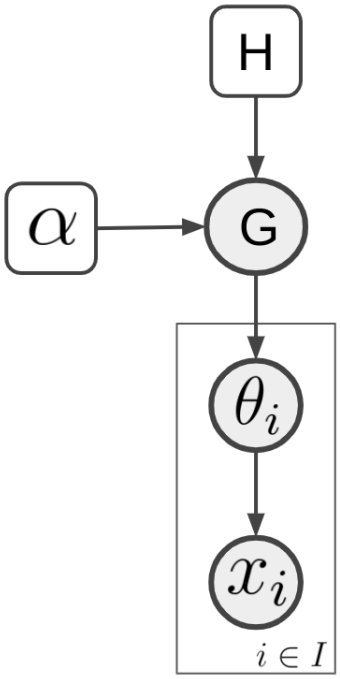
$$(G(A_1), \ldots, G(A_k)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_k)).$$

Em notação, $G \sim DP(\alpha, H)$

A partir da Definição 3, Ferguson (1973) e Sethuraman (1994) mostraram, através do Teorema de Consistência de Kolmogorov, a existência de um processo estocástico que a satisfaça e que G é, de fato, uma probabilidade sobre $\mathbb{R}$, com probabilidade 1.

Além disso a estrutura do processo nos mostra que o conjunto de medidas aleatórias, indexados por G, compartilham de uma medida aleatória base comum, H, sobre um nível de concentração α. A Figura 1 apresenta o esquema do Processo de Dirichlet, onde x_i representa o dado observacional e I o conjunto de índices de i.

Figura 1 – Esquema do Processo de Dirichlet



Fonte: Elaborada pelo autor.

A seguir, mostramos algumas propriedades importantes que nos auxilia no entendimento do Processo bem como estabelecem uma relação entre os parâmetros α e H do mesmo. As demonstrações de todos os resultados apresentados encontram-se no Anexo B.

Proposição 1. Tomando $B_0 = A_j$ e $B_1 = A_1 \cup A_2 \cup \cdots \cup A_{j-1} \cup A_{j+1} \cup \cdots \cup A_k$, $\forall \{A_j\}_{j=1}^k \in \mathscr{A}$ e $k \in \mathbb{N}^*$. Pela definição de um Processo de Dirichlet temos que

$$(B_0, B_1) \sim Dir(\alpha_0 H(B_0), \alpha_1 H(B_1))$$

mas também

$$(B_0, B_1) \sim Beta(\alpha_0 H(B_0), \alpha_1 H(B_1)).$$

Lema 2. Considerando $X_1, X_2, \ldots, X_k$ variáveis aleatórias independentes, tais que $X_j \sim Gama(\alpha_j), \forall j = 1, 2, \ldots, k$ e $k \in \mathbb{N}^*$, então

$$\left(\frac{X_1}{\sum_{j=1}^k X_j}, \ldots, \frac{X_k}{\sum_{j=1}^k X_j} \right) \sim Dir(\alpha_1, \ldots, \alpha_k)$$

Proposição 2. Sejam G um Processo de Dirichlet em $(\Theta, \mathscr{A}, H)$, com parâmetro α, e $\forall A \in \mathscr{A}$. Se $H(A) = 0$, então $G(A) = 0$ com probabilidade 1. Por sua vez, se $H(A) > 0$, então $G(A) > 0$ com probabilidade 1. Além disso,

$$\mathbb{E}[G(A)] = H(A)$$

Proposição 3. Sobre um Processo de Dirichlet, G, definido no espaço de probabilidades $(\Theta, \mathscr{A}, H)$, com parâmetro α, e $\forall A \in \mathscr{A}$, temos que,

$$Var[G(A)] = \frac{H(A)(1 - H(A))}{(\alpha + 1)}$$

Lema 3. Se $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$, então

$$(Y_1, \ldots, Y_{k-2}, Y_{k-1} + Y_k) \sim Dir(\alpha_1, \ldots, \alpha_{k-2}, \alpha_{k-1} + \alpha_k),$$

para $k \in \mathbb{N}^*$.

Lema 4. Se $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$ e π é uma permutação de $\{1, 2, \ldots, k\}$, então

$$(Y_{\pi(1)}, \ldots, Y_{\pi(k)}) \sim Dir\left(\alpha_{\pi(1)}, \ldots, \alpha_{\pi(k)}\right),$$

para $k \in \mathbb{N}^*$.

A Proposição 1 e o Lema 2 estabelecem relações entre o Processo de Dirichlet com outras distribuições de probabilidade (o caso especial do Processo de Dirichlet quando temos apenas dois conjuntos, resultando no Processo Beta, e a construção do Processo de Dirichlet pela normalização de um processo gama, respectivamente). Enquanto as Proposições 2 e 3

estabelecem uma relação entre H e α. Percebemos, a partir das duas últimas preposições, que valores elevados de α implica na diminuição da variância do processo, fazendo com que α seja entendido como uma parâmetro de concentração. Dessa forma, se H é uma medida de probabilidade aleatória, definida sobre um domínio de distribuições, intuitivamente temos que quando $\alpha \to \infty$ então $G(A_j) \to H(A_j), \forall A_j \in \mathscr{A}$, $j = 1, \ldots, k$ e $k \in \mathbb{N}^*$ (discutiremos mais sobre o parâmetro α na Subseção 2.2.4).

Por fim, quanto aos Lemas 3 e 4, ambos apontam comportamentos importantes da distribuição de Dirichlet que culminam na aditividade e permutabilidade do processo, Teh (2010), presentes, por exemplo, em problemas de clusterização discutidos futuramente neste trabalho.

2.2.2 Inferência Bayesiana sobre um Processo de Dirichlet

Teh (2010) nos apresenta a construção de um Processo de Dirichlet sobre a ótica Bayesiana. Seja um Processo de Dirichlet

$$(G(A_1), \ldots, G(A_k)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_k)),$$

onde G é uma distribuição aleatória de probabilidade sobre Θ e $A_1, \ldots, A_k$ são partições em Θ, tal como na Definição 3. Considere $\{\theta_1, \ldots, \theta_i\}$, $\forall \theta_N \in \Theta$ e para $N \in \mathbb{N}^*$, uma sequência de realizações independentes de G.

Sabemos que G é uma distribuição sobre Θ e portanto $\{\theta_i\}_{i=1}^N \in \Theta$. Estamos interessados em determinar a distribuição de G dada a sequência de realizações $\{\theta_i\}_{i=1}^N$. Isto é, desejamos obter a distribuição a posteriori do processo.

DISTRIBUIÇÃO A POSTERIORI DE UM PROCESSO DE DIRICHLET

Para determinarmos a distribuição a posteriori de G, vamos considerar a distribuição a priori $G \sim DP(\alpha, H)$. Então tomemos $n_k = \#\{i : \theta_i \in A_k\}$, onde n_k representa o número de vezes que $\{\theta_i \in A_k\}$ ocorre, de modo que

$$\sum_{j=1}^{k} n_j = N \qquad \sum_{j=1}^{k} G(A_j) = 1$$

Se considerarmos

$$(n_1, n_2, \ldots, n_k) \,|\, \big(G(A_1), G(A_2), \ldots, G(A_k)\big) \sim Mult\,(N; G(A_1), \ldots, G(A_k))$$

e uma vez que

$$\begin{aligned} p\left(\theta_1,\ldots,\theta_N|G(A_1),\ldots,G(A_k)\right) &= \prod_{i=1}^{N} p\big(\theta_i|G(A_1),\ldots,G(A_k)\big) \\ &= \prod_{i=1}^{N} p\big(\theta_i \in A_j|G\big) \\ &\text{mas se } p\big(\theta_i \in A_j|G\big) = \prod_{j=1}^{k} G(A_j)^{\delta_{\theta_i}(A_j)}, \forall i, \\ &= \prod_{i=1}^{N}\prod_{j=1}^{k} G(A_j)^{\delta_{\theta_i}(A_j)} \\ &= \prod_{j=1}^{k} G(A_j)^{\sum_{i=1}^{N}\delta_{\theta_i}(A_j)} \\ &= \prod_{j=1}^{k} G(A_j)^{n_j}, \end{aligned}$$

onde $\delta_{\theta_i}(A_j) = \mathbb{1}_{\{\theta_i \in A_j\}}$. Pela conjugação entre as distribuições de Dirichlet e Multinomial, visto na Seção 2.1, temos

$$\big(G(A_1),\ldots,G(A_k)\big)\,|\,(\theta_1,\ldots,\theta_N) \sim Dir(\alpha H(A_1)+n_1,\ldots,\alpha H(A_k)+n_k).$$

A partir do resultado anterior, vamos reescrever os parâmetros da distribuição em função de uma distribuição $G^{'}$ e uma parâmetro $\alpha^{'}$, obtendo $\alpha^{'}G^{'}(A_j) = \alpha H(A_j)+n_j$.

Se tomarmos

$$\begin{aligned} G^{'}(A_j) &= \mathbb{E}\left[G(A_j)|\theta_1,\ldots,\theta_N\right] \\ &= \frac{\alpha H(A_j)+\sum_{i=1}^{N}\delta_{\theta_i}(A_j)}{\alpha H(\Theta)+N} \\ &= \frac{\alpha H(A_j)+n_j}{\alpha+N}, \forall_j \end{aligned}$$

e, por sua vez, $\alpha^{'} = \alpha+N$, então

$$\big(G(A_1),\ldots,G(A_k)\big)\,|\,(\theta_1,\ldots,\theta_N) \sim Dir(\alpha^{'}G^{'}(A_1),\ldots,\alpha^{'}G^{'}(A_k)). \tag{2.4}$$

Uma vez que o resultado 2.4 é válido para qualquer partição finita e mensurável de Θ, como mostrado em Sethuraman (1994), a distribuição a posteriori de G, observados $\{\theta_i\}_{i=1}^{N} \subseteq \Theta$, é também um processo de Dirichlet dado por

$$G|\,(\theta_1,\ldots,\theta_N) \sim DP\left(\alpha+N, \frac{\alpha H}{\alpha+N}+\frac{\sum_{i=1}^{N}\delta_{\theta_i}}{\alpha+N}\right) \tag{2.5}$$

DISTRIBUIÇÃO PREDITIVA DE UM PROCESSO DE DIRICHLET

Novamente, vamos considerar um Processo de Dirichlet, $G \sim Dir(\alpha,H)$, e tomemos a mesma sequência de realizações extraída desse processo, $\{\theta_1,\theta_2,\ldots,\theta_N\}$, tal como definida

anteriormente. Vimos na Seção 2.1 que podemos obter a distribuição preditiva por meio de um processo estocástico. Dessa forma, seja θ_{N+1} uma observação futura do Processo de Dirichlet sobre uma amostra de tamanho N e seja $\theta_i|G \sim G$, para $i = 1,\ldots,N$, temos que

$$\begin{aligned} f(\theta_{N+1} \in A|\theta_1,\ldots,\theta_N) &= \int f(\theta_{N+1} \in A|G)\, df_G(G|\theta_1,\ldots,\theta_N), && \text{para } A \subset \Theta \\ &= \int G(A) df_G(G|\theta_1,\ldots,\theta_N) \\ &= \mathbb{E}[G(A)|\theta_1,\ldots,\theta_N], && \forall \theta_i \in A \\ &= \frac{1}{\alpha+N}\left(\alpha H(A) + \sum_{i=1}^{N} \delta_{\theta_i}(A)\right), && \text{por 2.5,} \end{aligned}$$

Assim, por Teh (2010),

$$\theta_{N+1}|\theta_1,\ldots,\theta_N \sim \frac{1}{\alpha+N}\left(\alpha H + \sum_{i=1}^{N} \delta_{\theta_i}\right). \tag{2.6}$$

Em suma, pelo resultado acima, toda vez que o parâmetro de concentração $\alpha \to 0$, ou se o tamanho da amostra N é suficientemente maior que α, a preditiva se aproxima da distribuição empírica sobre os dados.

Ainda sobre Teh (2010), a distribuição preditiva, diferentemente de modelos contínuos, possui massa pontual localizada na amostra observada $\{\theta_1,\ldots,\theta_N\}$. Além disso, a relação 2.6 nos indica que, com probabilidade $\frac{1}{\alpha+N}$, sorteios em G poderão assumir valores idênticos aos já observados, de modo que a própria distribuição G possui massas pontuais. No caso de uma sequência de observações de G suficientemente longa, os valores dos sorteios subsequentes serão dados por uma ponderação das massas pontuais, implicando a G uma distribuição discreta. Esse último fato pode ser entendido melhor ao discutirmos as abstrações do processo que serão discutidas na Subseção 2.2.3.

2.2.3 Abstrações do Processo de Dirichlet

Na literatura, encontramos algumas abstrações sobre o Processo de Dirichlet que melhoram a compreensão do mesmo. Nesta Seção trataremos de duas delas, o processo de quebra-bastão (*Strick-Breaking*) e o processo de restaurantes chineses. Veremos que a primeira nos oferece um melhor entendimento sobre a característica discreta sobre G. Já a segunda, fomenta o uso do Processo de Dirichlet como uma técnica/modelo de agrupamento.

STICK-BREAKING

Anteriormente, havíamos comentado como o Processo de Dirichlet pode ser visto, por intuição, como uma soma ponderada de massas pontuais em G. Tal fato é construído segundo Sethuraman (1994) por um processo nomeado de “quebra de bastão"(*Stick-Breaking Process*).

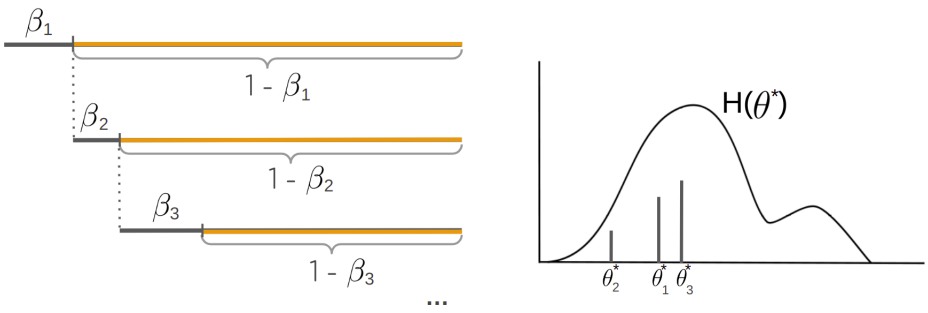
Tomemos a seguinte estrutura:

$$\beta_j|a \sim Beta(1,a), \text{ para } j = 1,2,\ldots,k$$
$$\pi_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j).$$

A abstração consiste em considerarmos um bastão de comprimento unitário. Em seguida partimos o bastão no ponto β_1, atribuindo a π_1 o comprimento restante do bastão. Recursivamente, partimos o bastão no próximo ponto β_j e até obtermos as π_k proporções ao consumir todas as β_k partições do bastão a cada iteração.

A Figura 2 ilustra o processo descrito, onde a cada iteração, ou quebra do bastão, associa um π_k a uma realização θ_k^*[1] em H, sendo H é uma função suave.

Figura 2 – Processo Stick-Breaking



Fonte: Elaborada pelo autor.

Assim, tomando

$$\theta_k^* \sim H$$
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*},$$

as k proporções são utilizadas como probabilidade de sorteio para as respectivas realizações θ_k^* em H. Isto é, a existência de θ_k^* (indicada pela função $\delta_{\theta_k^*}$) que tem probabilidade associada ao seu respectivo π_k, nos permite obter realizações independentes em G. Isso nos mostra que as realizações de um Processo de Dirichlet constituem em um processo discretizado.

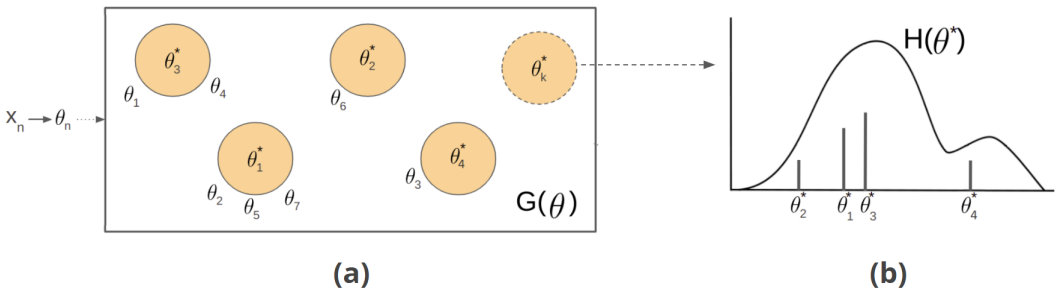
PROCESSO DE RESTAURANTE CHINÊS

Vimos que o processo de *Stick-Breaking* abstrai e evidencia a propriedade de discretização do Processo de Dirichlet. Tal característica é bastante útil no uso do Processo para agrupamentos e uma abstração análoga pode ser feita por uma representação de um Processo de

[1] Nesse ponto do texto, é importante recobrarmos a ideia de que θ e θ^* são medidas aleatórias extraídas de suas respectivas distribuições, G e H.

Restaurante Chinês (CRP, do inglês *Chinese Restaurant Process*). Nessa representação possuímos um restaurante com um número infinito de mesas que podem comportar um número infinito de clientes (denotados por θ), esquematizado pela Figura 3(a). Cada mesa servirá um único prato (indexado por θ^*), definido pelo primeiro cliente que se sentar nela. O cardápio do restaurante também contém um número infinito de pratos a serem servidos, tal como representado por uma distribuição na Figura 3(b). Cada cliente que entra no restaurante escolherá uma mesa para se sentar, caso a mesa já exista ele desfrutará do prato que a mesma estará servindo; caso contrário, uma nova mesa será posta e este poderá selecionar um novo prato para ser servido. Esse processo, presente como um todo na Figura 3, ocorre iterativamente ao longo do tempo.

Figura 3 – Processo de Restaurante Chinês



Fonte: Elaborada pelo autor.

A abstração descreve como clientes $x_1,\ldots,x_N$ que se associam aos seus respectivos assentos ou suas respectivas extrações de G, $\theta_1,\ldots,\theta_N$; as mesas como as extrações sobre H, $\theta_1^*,\ldots,\theta_k^*$; e o cardápio como a própria distribuição H. Além disso, o CRP propõe que a decisão dos clientes sobre as escolhas de uma mesa e consequentemente do prato ocorrerá segundo a popularidade das mesas existentes (número de assentos ocupados) e do parâmetro α do Processo de Dirichlet. Por fim, em relação ao número de *clusters*, Teh (2010) apresenta resultados que mostram o crescimento logarítmico do número de agrupamentos no processo quando N aumenta, condizendo com a ideia de que *clusters* populares tendem a ficar cada vez mais populares ao longo do tempo.

Novamente, assumindo H uma função contínua suave e considerando $\theta_1^*,\ldots,\theta_k^*$, para $k \in \mathbb{N}^*$, realizações extraídas de H. Se $\theta_1,\ldots,\theta_N$, para $N \geq k$, realizações sobre G e n_k o número de realizações em G associadas a um θ_k^*, como definido na Subseção 2.2.2. Então, para $G \sim DP(\alpha,H)$, as realizações de G estão submetidas a $\theta_1^*,\ldots,\theta_k^*$, de modo que a distribuição preditiva do processo possa ser reescrita como

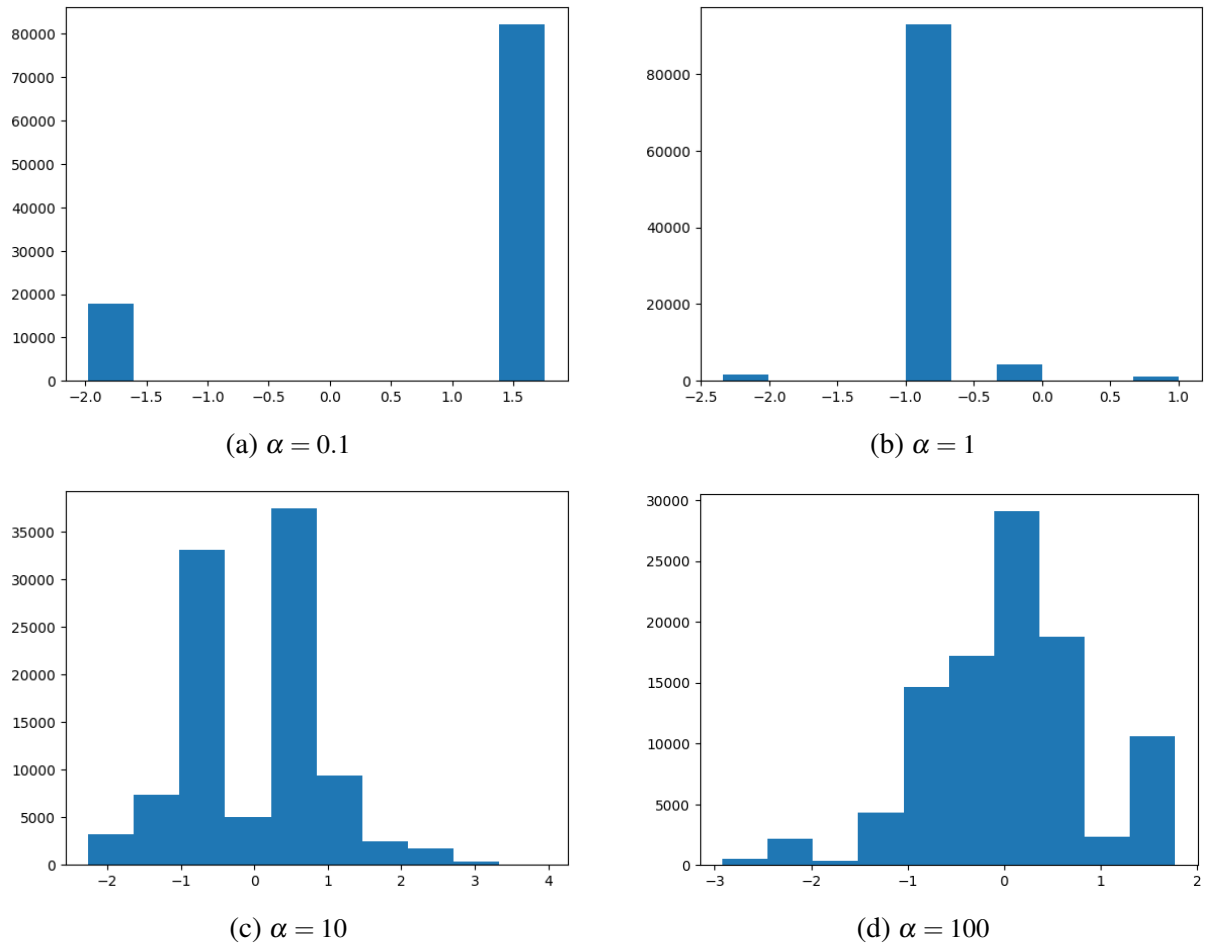
$$\theta_{N+1}|\theta_1,\ldots,\theta_N \sim \frac{1}{\alpha+N}\left(\alpha H + \sum_{j=1}^{k} n_j \delta_{\theta_j^*}\right). \tag{2.7}$$

A estrutura 2.7 evidencia que a seleção de um θ_k^*, a partir de um θ_{N+1} é proporcional a n_k. A medida que algum n_k cresce, percebemos que a probabilidade de se obter uma nova

realização sobre H se torna cada vez menor, já que a mesma é fixa e proporcional a α, e que a probabilidade de seleção do *k-ésimo* elemento em H aumenta (em outras palavras, um *cluster* θ_k^* com muitos elementos θ_i associados, tende a ficar maior e mais provável).

2.2.4 Análise do Parâmetro de Concentração α

Na Subseção 2.2.1, vimos que o parâmetro α está associado a variância de um Processo de Dirichlet, pela Preposição 3, em relação a distribuição base H, fazendo com que α seja interpretado como um parâmetro de concentração sobre H, Teh (2010). A Figura 4 nos ajuda a perceber o efeito do parâmetro sobre o processo para um tamanho fixo de amostra. Baseando-nos no método de *Stinck-Breaking* e tomando $H \sim N(0;1)$ como função base de um Processo de Disrichlet, variamos alguns valores de α para simulação do processo focadas nas realização de θ_i em G, fixando $N = 100000$ e $k = 100$.



(a) $\alpha = 0.1$ (b) $\alpha = 1$

(c) $\alpha = 10$ (d) $\alpha = 100$

Figura 4 – Simulação de um Processo de Dirichlet usando o método de *Stick-Breaking*, sobre $H \sim N(0;1)$ e variações do parâmetro de concentração α.

Fonte: Elaborada pelo autor.

Notamos que o aumento de α faz com que as realizações de G reflitam melhor as distribuição base H, isto é, para $\alpha = 100$, como mostrado na Figura 4d, a distribuição de G se

aproxima a de $H \sim N(0;1)$. Por sua vez, quando α é pequeno, como na Figura 4a, percebemos um afastamento da distribuição base H.

Dado o comportamento e o efeito de α sobre o Processo, podemos voltar nossa atenção em uma forma adequada de determinar tal parâmetro. Liu e Nandram (2022) discutiu três métodos amostrais para estimação de α. O primeiro deles é denominado de Método de Amostragem por Rejeição Adaptativa. Baseado em Gilks e Wild (1992), neste método determinamos uma função logarítmica sobre a densidade de α, digamos $\phi = \ln \alpha$, e tomamos uma distribuição a priori $\pi(\phi)$ com posteriori $\pi(\phi|k)$, onde k indexa o número de realizações distintas sobre H, como denotado anteriormente. A partir disso realizamos amostragem por rejeição, tal qual em Ripley (1987), que consiste em selecionar aleatoriamente pontos subsequentes sobre um domínio ou uma região cartesiana bidimensional. A cada seleção verificamos se o ponto está sob a região da densidade em questão, $\pi(\cdot)$, e caso isso não ocorra o rejeitamos. Logo, a cada unidade amostral bem sucedida a posteriori é atualizada aproximando-se da verdadeira densidade logarítmica, reduzindo assim a chance de rejeição de pontos subsequentes. Há algumas limitações em torno deste método, segundo Liu e Nandram (2022), uma delas é a garantia da log-concavidade de $\pi(\phi)$ e $\pi(\phi|k)$, o que não é problemático por si só (o autor, inclusive, apresenta duas prioris convenientes para o processo), mas implica que estas distribuições sejam estritamente unimodais. Além disso, as avaliações do método de rejeição são normalmente muito caras computacionalmente em determinadas aplicações (Amostragem de *Gibbs*, por exemplo).

O segundo método consiste em amostragem por Misturas de Distribuições Gamma. Inicialmente proposto por Escobar e West (1995), o método consiste na adição de uma variável latente que associa α a algum dos k elementos de mistura. A partir das distribuição a priori e a posteriori sobre essa variável latente e utilizando alguma técnica amostral, geralmente baseadas em Monte Carlo em Cadeia de Marcov (MCMC), obtemos realizações de α sobre uma distribuição Gamma do modelo de misturas[2]. No entanto, segundo Liu e Nandram (2022), uma preocupação em relação ao Método Gamma é que este forneça uma distribuição de amostragem pelo menos bimodal a α, enquanto uma densidade unimodal é preferida, além do mesmo exigir o uso de uma priori Gamma informativa, o que necessitaria de mais validações no método.

O terceiro e último método apresentado por Liu e Nandram (2022) é um dos mais conhecidos métodos de geração de variáveis aleatórias, a *Razão de Uniformes*. A partir de uma observação gerada uniformemente sobre um domínio, podemos obter uma amostra a posteriori para α. Em linhas gerais, determinamos duas variáveis aleatórias independentes U e V, tais que $U,V \sim U(0;1)$. A razão $\frac{V}{U}$ é calculada para cada sorteio obtido das variáveis aleatórias e então usada como observação da distribuição de interesse (α). Neste caso, a regra usada para determinar o valor de α segundo o método é baseada na priori e posteriori $\pi(\alpha)$ e $\pi(\alpha|k)$, respectivamente.

[2] Liu e Nandram (2022) traz uma abordagem do método onde a variável latente inserida no processo é configurada como $\rho = \frac{1}{1+\alpha}$, correspondendo a $Cor(x_i, x_j)$, para $i \neq j$, de um Processo de Dirichlet.

Liu e Nandram (2022) ainda apresenta um estudo de simulação onde compara e analisa os três métodos citados. Em suas análises, apesar de todos os métodos apresentarem resultados razoáveis para α, baseados na convergência à distribuição base, a método de Razão de Uniformes foi o mais acurado dentre eles. Sendo assim, escolhemos este método para determinar α neste trabalho (os detalhes de como definimos k e quais distribuições foram atribuídas aos parâmetros de concentração, envolvidos no método final, serão discutidos na Seção 3.2).

2.3 Modelos Bayesianos Hierárquicos Não Paramétricos

Aplicações estatísticas envolvendo múltiplos parâmetros que são relacionados de alguma maneira pela estrutura do processo (experimento), exigem um modelo de probabilidade conjunta para que estes reflitam as dependências envolvidas. Estruturas hierárquicas são uma excelente alternativa na modelagem de tal problema, uma vez que os resultados desse processo podem ser modelados condicionalmente em certos parâmetros, que por sua vez recebem um desenho probabilístico em termos de novos parâmetros, Gelman *et al.* (2021). A modelagem hierárquica nos auxilia a compreender e encapsular problemas multiparamétricos, desenvolver saídas computacionais eficazes e, sobre tudo, evitar problemas de *overfitting*, pois buscam uma distribuição populacional para estruturar alguma dependência nos parâmetros.

Outra motivação no uso de modelos hierárquicos está na sua independência condicional, na qual um conjunto de parâmetros é acoplado fazendo com que suas distribuições dependam de um parâmetro subjacente compartilhado. Essas distribuições são muitas vezes consideradas idênticas, com base em uma afirmação de permutabilidade garantidas pelo Teorema de Finetti, Teh e Jordan (2009).

Vimos na Seção 2.2 que modelos bayesianos não paramétricos incluem parâmetros com dimensão não fixa, além dos parâmetros de escala e localização, de modo que valemos da modelagem hierárquica para especificar as distribuições sobre esses parâmetros. Podemos ver esta aplicação sobre um Processo de Dirichlet

$$G \sim DP(\alpha, H),$$

notamos que além do parâmetro de concentração α (que geralmente parte de uma distribuição a priori para sua obtenção, como vimos na Subseção 2.2.4), a distribuição base H é muitas vezes considerada uma distribuição paramétrica e seus parâmetros também são dotados de distribuições a priori.

Nessa Seção discutiremos a modelagem hierárquica em Estatística Bayesiana não paramétrica, mais especificamente sobre um Processo de Dirichlet. Em suma, trataremos a distribuição base como uma distribuição não paramétrica, vista agora como um sorteio aleatório de alguma distribuição de medidas. A essa estrutura damos o nome de Processo Hierárquico de Dirichlet (HDP, do inglês *Hierarchical Dirichlet Process*).

2.3.1 Processo Hierárquico de Dirichlet

Até agora discutimos sobre o Processo de Dirichlet e sua aplicação como modelo de agrupamento quando sua componente θ_k^* é uma variável aleatória discreta de cardinalidade desconhecida, uma vez que o Processo possui uma característica de discretização. Dentre os modelos de agrupamentos com esta componente característica, destacamos o Processo Hierárquico de Dirichlet (HDP). Neste modelo detemos de inúmeros grupos, cada qual incorpora uma variável discreta, de cardinalidade desconhecida, e das quais desejamos vincular entre os grupos (em outras palavras, compartilhar *clusters*), Teh e Jordan (2009).

Definição 4. Tome uma coleção de Processo de Dirichlet indexados por G_0 e $G_{j\in\mathfrak{J}}$, onde $\mathfrak{J}$ é um conjunto contável de índices, definidos em um espaço de probabilidade comum (Θ,Ω). Considere H como uma medida base do Processo. Denominamos como Processo Hierárquico de Dirichlet (HDP) a seguinte estrutura

$$\begin{aligned} &G_0|\gamma,H \sim DP(\gamma,H) \\ &G_j|\alpha,G_0 \sim DP(\alpha,G_0) \qquad\qquad \text{para } j\in\mathfrak{J}, \end{aligned} \tag{2.8}$$

para α e γ os respectivos parâmetros de concentração dos Processo de Dirichle indexados.

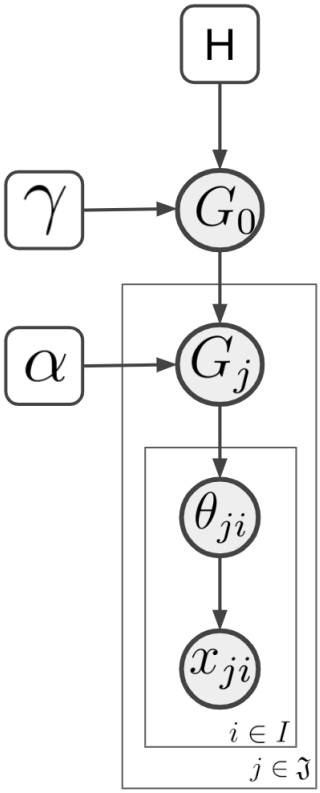
Pela Definição 4, percebemos que o modelo HDP vincula medidas de probabilidades aleatórias (G_0 vinculada as medidas G_j, para todos os grupos dentro do conjunto de índices $\mathfrak{J}$), permitindo que elas compartilhem uma medida base (H), também aleatória. Em outras palavras, o modelo induz o compartilhamento de realizações, ou átomos, entre as medidas aleatórias G_j, uma vez que cada uma herda seu conjunto de átomos do mesmo G_0.

Ainda pela Definição, ao considerarmos uma hierarquia de dois níveis o conjunto de átomos no nível superior ($G_0|\gamma,H$), é compartilhado por toda a hierarquia, isto é, cada G_j recebe um Processo de Dirichlet como distribuição a priori do processo com medida base G_0, que se modifica ao longo das iterações. Enquanto um modelo de multinível hierárquico permite uma estrutura de dependência mais rica dos pesos dos átomos (ou seja, o conjunto raiz não é simplesmente replicado para seus nós filhos). A partir dessa leitura, Teh e Jordan (2009) nos mostra que o Processo Hierárquico de Dirichlet pode ser usado como distribuição a priori dos fatores de agrupamento com a seguinte estrutura

$$\begin{aligned} &\theta_{ji}|G_j \overset{i.i.d.}{\sim} G_j \\ &x_{ji}|\theta_{ji} \sim f_{\theta_{ji}} \qquad\qquad \text{para } j\in\mathfrak{J} \text{ e } i\in I, \end{aligned} \tag{2.9}$$

onde θ_{ji} é o fator correspondente a i-ésima observação do j-ésimo grupo (x_{ji}), sendo I o conjunto de índices de i, e $f_{\theta_{ji}}$ é uma função densidade de probabilidade. A Figura 5 ilustra o HDP com 2 níveis de hierarquia sobre esta estrutura.

Figura 5 – Esquema gráfico do Processo Hierárquico de Dirichlet



Fonte: Elaborada pelo autor.

2.3.2 Inferência sobre o Processo Hierárquico de Dirichlet

Vimos na Subseção 2.2.2 como obtivemos a distribuição a posteriori de um Processo de Dirichlet através da conjugada da distribuição de Dirichlet. A mesma ideia se aplica no modelo hierárquico, porém vamos abordá-la de uma forma um pouco diferente.

Inicialmente, vamos assumir f_θ com priori conjugada em H, tal como em Teh e Jordan (2009), e tomemos as seguintes notações:

- $\theta = \{\theta_{ji}\}_{j\in\mathfrak{J},i\in I}$ são uma sequência de realizações do j-ésimo Processo (G_j), para cada observação x_{ji};
- $\theta^* = \left\{\theta^*_{jl}\right\}_{j\in\mathfrak{J},l\in\mathfrak{L}}$ são uma sequência de realizações distintas sobre G_0, onde $\mathfrak{L}$ denota o conjunto de índices destas realizações;
- $\theta^{**} = \{\theta^{**}_k\}_{k=1,\ldots,K}$ são uma sequência de realizações distintas sobre H, onde K é o número total destas realizações;
- $\mathbf{n} = \{n_{jlk}\}$ é a sequência do número de observações em cada um dos $\#\mathfrak{J}$ Processos, associados aos seus respectivos $\#\mathfrak{L}$ subgrupos compartilhados pelos seus respectivos K grupos e $\sum_{\substack{j\in\mathfrak{J},l\in\mathfrak{L}\\k=1,\ldots,K}} n_{jlk} = N$;

- $\mathbf{m} = \{m_{jk}\}$ é a sequência do número de observações agrupadas em cada um dos seus $\#\mathfrak{J}$ Processos, compartilhados pelos seus respectivos K grupos e $\sum_{\substack{j\in\mathfrak{J}\\k=1,\ldots,K}} m_{jk} = M$.

A partir disso, vamos discutir duas abstrações do processo, tal como realizado para o Processo de Dirichlet, para que então determinarmos a posteriori do mesmo.

STICK-BREAKING

Discutimos que a abstração do Processo de Dirichlet via processo de 'quebra-bastão' nos dá uma representação concreta sobre a característica discreta do Processo. A mesma ideia e representação são estendidas para o HDP. Vamos considerar inicialmente a representação do nível superior do modelos hierárquico, onde $G_0 \sim DP(\gamma, H)$. Como a distribuição a priori de G_0 é dada por um Processo de Dirichlet, temos uma estrutura similar a apresentada na Subseção 2.2.3,

$$G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^{**}},$$

tal que

$$\begin{aligned}
&\beta_k|\gamma \sim Beta(1,\gamma), \text{ para } k = 1,2,\ldots\\
&\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l)\\
&\theta_k^{**}|H \sim H.
\end{aligned}$$

A grande diferença encontra-se na construção de G_j. Como G_j condicionalmente também é distribuída segundo um Processo de Dirichlet, teremos a replicação da estrutura anterior, porém com dependência aos π_k,

$$G_j = \sum_{k=1}^{\infty} \eta_{jk} \delta_{\theta_k^{**}},$$

tal que

$$\begin{aligned}
&\nu_{jk}|\alpha,\pi_1,\ldots,\pi_k \sim Beta\left(\alpha\pi_k, \alpha - \alpha\sum_{l=1}^{k}\pi_l\right), \text{ para } k = 1,2,\ldots\\
&\eta_{jk} = \nu_{jk}\prod_{l=1}^{k-1}(1-\nu_{jl}).
\end{aligned}$$

Neste caso, a abstração consiste em considerar um bastão de comprimento unitário que será partido no ponto β_k, a cada iteração. Da quebra do bastão, o restante do seu comprimento, indexado por π_k, sera então compartilhado ao j-ésimo jogador. Este bastão compartilhado se tornará um novo bastão de referência para o jogador em questão e será quebrado no ponto $\nu_j k$, obtendo assim um novo comprimento η_{jk}. O processo se repete iterativamente para todos os jogadores envolvidos.

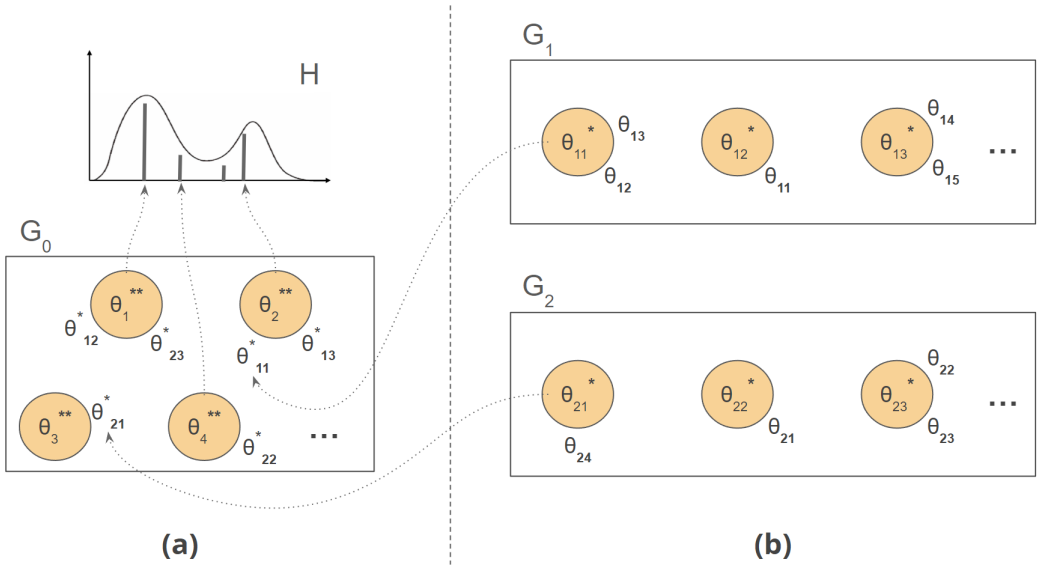
PROCESSO DE RESTAURANTES CHINESES

Na Subseção 2.2.3 vimos que as probabilidades marginais do Processo de Dirichlet podiam ser descritas como um processo de restaurante chinês. Esta mesma abstração existe para o caso do HDP, ela é denominada de Franquia de Restaurantes Chineses (CRF, do inglês *Chinese Restaurant Franchise*).

Neste caso, ocorre um processo de restaurante chinês em cada um dos restaurantes considerados: cada cliente que chegar ao j-ésimo restaurante pode sentar-se em uma mesa já existe, que serve um determinado prato, ou solicitar uma nova mesa e escolher um prato novo que ainda não está sendo servido no restaurante, como representado na Figura 6(b). Cada restaurante, por sua vez, responde a uma matriz que orquestra os pratos compartilhados entre eles de acordo com o número de mesas que servem o determinado prato, a partir de um cardápio em comum, Figura 6(a).

No caso de $G_{j\in\mathfrak{J}}$ restaurantes, o acoplamento entre eles é feito por meio de uma matriz representada por G_0. Nela, ocorre o gerenciamento dos pratos extraídos do cardápio H e compartilhados entre todos os restaurantes. Segundo Teh e Jordan (2009), rotule o i-ésimo cliente do j-ésimo restaurante com uma variável aleatória θ_{ji} que é distribuída de acordo com G_j. Da mesma forma, seja θ^*_{jl} uma variável aleatória correspondente à l-ésima mesa do j-ésimo restaurante; a realização dessas variáveis é feita de forma aleatória, independente e identicamente distribuída (*i.i.d.*) conforme G_0. Finalmente, os pratos são variáveis *i.i.d.* θ^{**}_k, distribuídas de acordo com a medida base H. Acoplamos essas variáveis da seguinte forma $\theta_{ji} = \theta^*_{jl_{ji}} = \theta^{**}_{k_{jl}}$[3].

Figura 6 – Processos de Franquia de Restaurante Chinês



Fonte: Elaborada pelo autor.

[3] Adicionamos os subíndices ji em jl para indicar que não estamos considerando realizações distintas em G_0, mas sim as realizações obtidas para cada uma das θ_{ji}.

Griffiths *et al.* (2003) mostra que o CRF captura as probabilidade marginais do HDP ao integrar as medidas aleatórias G_j e G_0 do processo. Dessa forma estaremos produzindo um conjunto de distribuições condicionais sobre as variáveis θ_{ji} e θ^*_{jl}, respectivamente, como apresentadas pelo autor,

$$\theta_{ji}|\theta_{j1},\ldots,\theta_{j,i-1},\alpha,G_0 \sim \sum_{l=1}^{m_{j\cdot}} \frac{n_{jl\cdot}}{\alpha+n_{j\cdot\cdot}}\delta_{\theta^*_{jl}} + \frac{\alpha}{\alpha+n_{j\cdot\cdot}}G_0,$$
$$\theta^*_{jl}|\theta^*_{11},\ldots,\theta^*_{1m_{1\cdot}},\ldots\theta^*_{j,l-1},\gamma,H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{\gamma+m_{\cdot\cdot}}\delta_{\theta^{**}_k} + \frac{\gamma}{\gamma+m_{\cdot\cdot}}H. \tag{2.10}$$

Lembrando que n_{jlk} é a quantidade de clientes no j-ésimo restaurante, que sentaram-se na l-ésima mesa e pediram o k-ésimo prato, m_{jk} é a quantidade de mesas no j-ésimo restaurante servindo o k-ésimo prato e K é o maior índice dentre a coleção de pratos já pedidos.

Notamos que, similar ao processo de restaurante chinês da Subseção 2.2.3, temos aqui as probabilidades de sorteio para uma mesa e um prato servido em algum momento no processo, bem como as probabilidade de se efetuar um sorteio inédito através das distribuições G_0 e H, compartilhadas em seus respectivos níveis. Isso mostra a característica de discretização no processo, de modo que as probabilidades envolvidas são proporcionais a frequência dos clientes e das mesas entre cada restaurante em relação a toda franquia e entre cada pedido em relação a toda franquia, respectivamente.

2.3.3 Estrutura a Posteriori do Processo Hierárquico de Dirichlet

Vamos tomar a Franquia de Restaurantes Chineses (CRF) anterior de modo que H seja uma distribuição contínua, por suposição, fazendo com que as realizações θ^{**} sejam únicas. A partir da integração 2.10 das medidas aleatórias G_j produzimos um Restaurante Chinês para cada j-ésimo grupo, bem como uma sequência de sorteios *i.i.d.* da medida base G_0, que são usados recursivamente na integração da mesma. Assim, o estado do CRF consiste nos rótulos dos pratos θ^{**}_k, a mesa l_ji, em que o i-ésimo cliente se senta, e o prato k_jl servido na l-ésima mesa. Como funções de estado do CRF, temos também o número de clientes e mesas n e m, respectivamente, e os rótulos de clientes θ_{ji} e mesas θ^*_{jl}, todos definidos no início da Subseção 2.3.2.

Sobre Teh e Jordan (2009), tomando a distribuição de G_0 condicionada as realizações *i.i.d.* $\theta^* = \left\{\theta^*_{jl_{ji}}\right\}$, fazendo com que G_0 seja independente do resto do CRF, pois os restaurantes interagem com G_0 apenas através dos θ^* *i.i.d.*. Assim a posteriori do processo sobre G_0 segue, como visto na Subseção 2.2.2, para um Processo de Dirichlet:

$$G_0|\gamma,H,\theta^* \sim DP\left(\gamma+\sum_{j,k} m_{jk}, \frac{\gamma H+\sum_k m_{\cdot k}\delta_{\theta^{**}_k}}{\gamma+\sum_{j,k} m_{jk}}\right) \tag{2.11}$$

Observamos que os valores para m e θ^{**} são determinados dado θ^*, uma vez que são o valor de contagem e rótulos únicos (pois, por suposição, H é contínua), respectivamente, sobre θ^*. Pela

Definição 4, uma amostra de 2.11 pode ser extraída da seguinte forma:

$$\begin{aligned}
&\beta_0, \beta_1, \ldots, \beta_k | \gamma, G, \theta^* \sim Dir(\gamma, m_{\cdot 1}, \ldots, m_{\cdot k}) \\
&G_{0_{\mathbf{new}}} | \gamma, H \sim DP(\gamma, H), \\
&G_0 = \beta_0 G_{0_{\mathbf{new}}} + \sum_{l=1}^{k} \beta_l \delta_{\theta_l^{**}}.
\end{aligned}$$

Isto é, a posteriori para G_0 é uma mistura de rótulos θ^{**} de extrações independentes de um $DP(\gamma, H)$.

Por fim, a posteriori para G_j segue a posteriori de um Processo de Dirichlet usual, Subseção 2.2.2, dado a sua medida base G_0 e amostras *i.i.d.* θ_j.

$$G_j | \alpha, G_0, \theta_j \sim DP\left(\alpha + \sum_{l,k} n_{jlk}, \frac{\alpha G_0 + \sum_k n_{j\cdot k} \delta_{\theta_k^{**}}}{\alpha + \sum_{l,k} n_{jlk}}\right) \tag{2.12}$$

Analogamente a G_0, $n_{j\cdot\cdot}$ e θ^{**} são o valor de contagem e rótulos únicos, respectivamente, sobre θ_j, podemos amostrar de 2.12 tal como:

$$\begin{aligned}
&\pi_{j0}, \pi_{j1}, \ldots, \pi_{jk} | \alpha, \theta_j \sim Dir\left(\alpha\beta_0, \alpha\beta_1 + n_{j\cdot 1}, \ldots, \alpha\beta_k + n_{j\cdot k}\right) \\
&G_{j_{\mathbf{new}}} | \alpha, G_0 \sim DP(\alpha\beta_0, G_{0_{\mathbf{new}}}), \\
&G_j = \pi_{j0} G_{j_{\mathbf{new}}} + \sum_{l=1}^{k} \pi_{jl} \delta_{\theta_l^{**}}.
\end{aligned}$$

De modo que o processo para obtenção da posteriori é válido para todo $j \in \mathfrak{J}$. Observamos também que G_j é uma mistura de rótulos θ^{**} e uma extração independente de um DP, que herda β_0 e o indexa ao seu parâmetro de concentração α.

Um outro ponto discutido por Teh e Jordan (2009) consiste na divisão de um HDP em uma "parte discreta" e uma "parte contínua" devido a esse processo de integrações sobre G_0 e G_j mostrado pelos autores. A "parte discreta" consiste em rótulos com valores únicos (θ^{**}), associando pesos diferentes a estes rótulos para cada DP. A "parte contínua", por sua vez, é um sorteio separado de um HDP com a mesma estrutura hierárquica do HDP original e medida de base global H, mas com parâmetros de concentração alterados. Isto é, ela consiste na estrutura de distribuições relacionadas entre a série infinita de rótulos sorteados *i.i.d.* sobre H e a verossimilhança dos dados.

2.3.4 Aplicações do Modelo Hierárquico de Dirichlet

O Modelo HDP tem uma ampla gama de aplicações, no entanto três delas se destacam. A primeira consiste no uso do modelo como método de busca, ou Recuperação da Informação (IR, do inglês *Information Retrieval*). A importância dos motores de busca modernos tem colocado em foco um problema clássico no campo da IR, em como deve ser representada uma coleção de

documentos para que documentos relevantes possam ser devolvidos em resposta a uma consulta. Cowans (2004) desenvolveu um estudo onde mostrou que o HDP fornece justificativas estatística para a intuição por trás da representação textual conhecida como frequência do termo–inverso da frequência nos documentos *tf-idf* (do inglês *term frequency–inverse document frequency*). A intuição geral é que a relevância de um documento para uma consulta deve ser proporcional à frequência dos termos de consulta que ele contém ("frequência do termo"), mas que os termos de consulta que aparecem em muitos documentos devem ser reduzidos, uma vez que são menos informativos ("frequência inversa nos documentos"). Em seu estudo, o autor denota x_{ji} como a i-ésima palavra no j-ésimo documento para algum *corpus* de documentos, onde o intervalo de x_{ji} é um vocabulário discreto Θ e H é a medida base de probabilidade sobre este vocabulário. A partir disso, ele define um *score* de relevância capaz de ordenar documentos mais relevantes dado um termo de busca.

Uma outra aplicação muito relevante do HDP consiste em fornecer uma distribuição baseada em parâmetros latentes e não nos dados observados. O faseamento de haplótipos é um problema em genética que pode ser formulado como um modelo estatístico de mistura, como apresentado por Stephens, Smith e Donnelly (2001). O problema consiste em um conjunto de M marcadores binários ao longo de um cromossomo. Os cromossomos humanos vêm em pares, então denotamos θ_{i1} e θ_{i2} como os vetores de marcadores de valor binário para um par de cromossomos de um i-ésimo indivíduo. Esses vetores são chamados de haplótipos e os elementos desses vetores são chamados de alelos. Um genótipo x_i, por sua vez, é um vetor que registra o par não ordenado de alelos para cada marcador; isto é, a associação dos alelos ao cromossomo é perdida. O problema do faseamento de haplótipos é restaurar haplótipos (que são úteis para prever associações de doenças) de genótipos (que são prontamente testados experimentalmente, enquanto os haplótipos não são). Assim, dado um conjunto de haplótipos de uma população população, cujo conjunto tem cardinalidade desconhecida, este problema pode ser formulado como um problema de modelagem de mistura DP onde um *cluster* é um haplótipo, tal como mostra o autor.

Finalmente, sua terceira aplicação concentra-se em modelos de tópicos ou agrupamentos textuais. No decorrer do trabalho, já havíamos citado esse tipo de aplicação quando tratamos dos métodos de abstração para um DP e um HDP. Agora, mais especificamente sobre Blei *et al.* (2003), o HDP pode ser utilizado e visto como uma generalização de um modelo de mistura finita em que cada ponto de dados está associado a múltiplas retiradas de um modelo de mistura, e não a uma única retirada. Para motivar a formulação do modelo de tópico, considere o problema de modelar as ocorrências de palavras em um conjunto de processos jurídicos (por exemplo, para fins de classificação de processos futuros). Uma metodologia simples de agrupamento pode tentar colocar cada processo em um único agrupamento. Mas pareceria mais útil poder classificar processos de acordo com "tópicos". Por exemplo, um processo pode ser principalmente sobre causa trabalhista, mas também pode referir-se à problemas de saúde do requerente e a assedio moral e psicológico ao mesmo. Além disso, como este exemplo sugere, seria útil poder atribuir

valores numéricos ao grau em que um artigo trata cada tópico. É exatamente esta aplicação, incluindo o contexto de Jurimetria tomado como exemplo, que concentraremos nossos esforços e trataremos nos Capítulos seguintes.

2.3.5 Método Computacional para um Modelo Hierárquico de Dirichlet

O processo de inferência sobre o HDP pode ser estruturado sobre métodos computacionais de amostragem via algoritmo de Monte Carlo via Cadeias de Markov (MCMC). Dentre os algoritmos baseados em MCMC, a amostragem de Gibbs é o mais adequado para modelos hierárquicos, dado a sua facilidade em avaliar as distribuições condicionais envolvidas no HDP, de modo a trabalhar com suas conjugadas, tornando-o computacionalmente conveniente, Sucar (2015). Em termos gerais, a ideia do amostrador de Gibbs é tornar um problema multivariado em uma sequência de problemas univariados baseados nas distribuições condicionais, de modo que ao iteramos o processo produziremos uma Cadeia de Markov. Isso nos permite obter amostras de uma distribuição a posteriori desconhecida e, através destas amostras, obter estimativas das características da distribuição objetivo.

Baseando-nos em Neal (2000) e Heinrich (2011), tomemos a abstração da franquia de restaurantes chineses como plano de fundo da representação do HDP. Vamos assumir a sequência de variáveis $\{L_{ji}\}_{j\in\mathfrak{J},i=1,\ldots,n_j}$ que indexam a mesa na qual o cliente x_{ji} se sentou e $\{K_{jl}\}_{j\in\mathfrak{J},l=1,\ldots,m_j}$ a variável que indexa o prato que o mesmo pediu.[4] Note que pelo processo de franquia de restaurantes chineses temos três alternativas para cada i-ésimo cliente *i*ao chegar em seu respectivo j-ésimo restaurante:

- Sentar-se em uma mesa já ocupada e, consequentemente, se servir do prato servido na mesma;
- Sentar-se em uma nova mesa e solicitar um prato que já está sendo servido em algumas das outras mesas do restaurante;
- Sentar-se em uma nova mesa e pedir um prato que ainda não foi servido no restaurante.

[4] Enquanto L,K são variáveis usadas no processo para atribuir amostras da mesa e do prato pedido pelo cliente θ_{ji}, θ^* e θ^{**} são as representações das massas pontuais (amostrada sobre G_0,H, respectivamente) dessas variáveis, que indicam quais distribuições estamos considerando para este cliente em questão.

Estes três casos ocorrem dispondo das seguintes probabilidades:

$$\begin{cases} L_{ji} = l & \text{com probabilidade } \propto \frac{n_{jl\cdot}^{-x_{ji}}}{n_{j\cdot\cdot}^{-x_{ji}}+\alpha} f_{k_{jl}}^{-x_{ji}}(x_{ji}) \\ L_{ji} = l_{new}, K_{jl_{new}} = k & \text{com probabilidade } \propto \frac{\alpha}{n_{j\cdot\cdot}^{-x_{ji}}+\alpha} \left(\frac{m_{\cdot k}^{-x_{ji}}}{m_{\cdot\cdot}^{-x_{ji}}+\gamma}\right) f_k^{-x_{ji}}(x_{ji}) \\ L_{ji} = l_{new}, K_{jl_{new}} = k_{new} & \text{com probabilidade } \propto \frac{\alpha}{n_{j\cdot\cdot}^{-x_{ji}}+\alpha} \left(\frac{\gamma}{m_{\cdot\cdot}^{-x_{ji}}+\gamma}\right) f_{k_{new}}^{-x_{ji}}(x_{ji}) \end{cases} \tag{2.13}$$

onde o subscrito *new* indica a execução de um novo sorteio ainda não visto por ambos os níveis, o sobrescrito $-x_{ji}$ indica que estamos considerando todos os clientes exceto o x_{ji} e n, m são as quantidades tal qual definidas na Subseção 2.3.2.

Teh e Jordan (2009) introduziu atualizações para o amostrador de *Gibbs* sobre uma estrutura geral de HDP, onde a distribuição de x_{ji} para algum *cluster* k, dado todas as outras observações, é definida como

$$f_k^{-x_{ji}}\left(x_{ji}\right) = \frac{\int f_{\theta_k^{**}}\left(x_{ji}\right) \prod_{j'i' \neq ji} f_{\theta_k^{**}}\left(x_{j'i'}\right) h(\theta_k^{**}) d\theta_k^{**}}{\int \prod_{j'i' \neq ji} f_{\theta_k^{**}}\left(x_{j'i'}\right) h(\theta_k^{**}) d\theta_k^{**}}, \tag{2.14}$$

para $h(\cdot), f_{\theta_k^{**}}(\cdot)$ densidades conjugadas de H e f_θ, respectivamente, tal como apresentado em 2.8 e 2.9. No Capitulo 3, daremos detalhes sobre esta distribuição (preditiva) $f^{-x_{ji}}$ ao indexarmos distribuições de probabilidade específicas para h e $f_{\theta_k^{**}}$.

Da mesma forma, as atualizações para algum prato k_{jl}, associado ao j-ésimo restaurante e a l-ésima mesa, ocorrem sobre as seguintes probabilidades:

$$\begin{cases} K_{jl} = k & \text{com probabilidade } \propto \frac{m_{\cdot k}^{-x_{ji}}}{m_{\cdot\cdot}^{-x_{ji}}+\gamma} f_k^{-x_{ji}}(x_{ji}|L_{ji} = l) \\ K_{ji} = k_{new} & \text{com probabilidade } \propto \frac{\gamma}{m_{\cdot\cdot}^{-x_{ji}}+\gamma} f_{k_{new}}^{-x_{ji}}(x_{ji}|L_{ji} = l) \end{cases} \tag{2.15}$$

Heinrich (2011) descreveu o algoritmo do processo de modo que o adaptamos para a abstração de CRF como mostra o Algoritmo 1.

No que diz respeito do custo computacional das atualizações de *Gibbs* sobre o processo de CRF, Teh e Jordan (2009) nos apresenta que, em geral, o custo é dado pela recorrência do cálculo das probabilidades condicionais marginais $f_k()$. Isto ocorre cerca de $(1 + M_{j' \neq j} + \#K_{j' \neq j})$, onde $M_{j' \neq j}$ e $\#K_{j' \neq j}$ são o número total de mesas e pratos, respectivamente, em todos os restaurantes exceto o j-ésimo. O que resulta em uma desvantagem em casos onde o número de níveis hierárquicos é grande. Uma outra desvantagem do método pelo CRF, apontada pelo autor, está no acoplamento da amostragem nos vários restaurantes (uma vez que todos os DP's estão integrados). Este acoplamento dificulta a obtenção de um amostrador CRF para certos modelos (por exemplo, o HDP-HMM). Sendo necessário construir amostradores que utilizem uma representação mista de CRP e *stick-breaking* para dissociar os DP's envolvidos.

Por fim, existem inúmeros métodos de avaliação e análises a cerca do amostrador de Gibbs e sua convergência. Neste sentido, focaremos nossos esforços na análise de diagnóstico de convergência do método, Capitulo 4, como tratado Gamerman (1996). Através de um estudo simulado, iremos verificar sede fato atingimos a convergência com relação a função objetivo H, utilizando para isso a *log-verossimilhança*, e ao número de restaurantes K estipulados pelo processo. Como o processo é Markoviano esperado que à medida em que o número de iterações aumenta, o método se aproxima da condição de equilíbrio convergindo, aproximadamente e a partir de uma determinada iteração, para a métrica desejada.

Algoritmo 1 – Amostrador de *Gibbs* para HDP sobre CRF

Requer: Vetor de clientes $\{\overrightarrow{x}\}$
INICIALIZAÇÃO
$\#K \in \mathbb{N}^*$
zerar as quantidades associadas a n e m.
laço todos os restaurantes $j \in \mathfrak{J}$:
laço todos os clientes $i = 1, 2, \ldots, n_j$:
amostrar um prato indexado por $z_{j,i} = k^*$, por 2.15
incrementar as quantidades m_{jk}, n_{jlk}
fim laço
fim laço
amostrar L_{ji}, por 2.13

AMOSTRADOR GIBBS
enquanto critério de parada **faça**
laço todos os restaurantes $j \in \mathfrak{J}$:
laço todos os clientes $i = 1, 2, \ldots, n_j$:
decrementar as quantidades m_{jk}, n_{jlk}
amostrar um prato k^*, por 2.15
se $k^* \in \{1, 2, \ldots, K\}$ **então**
incrementar as quantidades m_{jk}, n_{jlk}
amostrar L_{ji}, por 2.13
senão
criar e incrementar as quantidades $m_{jk_{new}}, n_{jlk_{new}}$
amostrar $L_{ji} = l_{new}$, por 2.13
fim se
fim laço
fim laço
fim enquanto
verifique a convergência
retorne os resultados

CAPÍTULO

3

MODELOS DE TÓPICOS

A comunicação humana tem sido um tópico de grande interesse na computação, uma vez que a mesma ocorre, majoritariamente, pelo uso de línguas naturais. Na análise linguística de um texto em linguagem natural digital é preciso definir a unidade de formação do documento (letras, palavras e sentenças) para que possamos realizar uma boa interpretação de seus sinais. No entanto essa tarefa é árdua, já que a língua natural dispõe de uma estrutura bastante complexa, não raramente apresentando ambiguidades dependendo do contexto tratado, sem contar os diversos idiomas e variações linguísticas regionais.

O Processamento de Linguagem Natural (NLP, do inglês *Natural Language Processing*) é a área da ciência da computação focada na geração e compreensão destas línguas, sendo grande parte de seu desafio tratar tais ambiguidades, Palmer (2010). Atualmente encontramos uma vasta coleção de técnicas, capaz de de coletar e preparar corporas de forma automática, convertendo-as em estruturas mensuráveis. Dentre as técnicas mais difundidas para processamento textual, traremos as baseadas em frequência de palavras, ou *tokens*, que são mais simples quanto as regras que constituem a estrutura dos dados e tendem a gerar representações esparsas de dados numéricos, Qader, Ameen e Ahmed (2019) e Das e Chakraborty (2018).

Além de técnicas de processamento textual, o campo da modelagem deste tipo de dados é bastante vasto e desejado por áreas que, até então, não utilizavam de grande aparato estatístico e computacional para desenvolvimento de suas tecnologias. Dentre os modelos de NLP, temos interesse especifico em modelos de tópicos ou de associação mista. Segundo Griffiths *et al.* (2003) e Erosheva e Fienberg (2005), este tipo de modelo é uma generalização de um modelo de mistura finita em que cada ponto de dados está associado a múltiplas retiradas de um modelo de mistura, e não a uma única retirada. Como vimos, o HDP é uma excelente ferramenta para esse tipo de modelagem, sobre tudo por sua estrutura não paramétrica que flexibiliza a determinação prévia do número de tópicos.

Neste Capítulo, apresentaremos algumas técnicas populares de processamento textual

que nos ajudam a simplificar a complexidade dos dados e nos garante uma boa representação numérica (mensurável), Seção 3.1, de modo que as distribuições desejadas para modelá-los sobre HDP sejam bastante adequadas. E por fim, retomamos ao HDP e apresentamos como sua estrutura, dada a modelagem de tópicos, é formada na Seção 3.2.

3.1 Pré-Processamento e Elementos Textuais

Segundo Palmer (2010) o pré-processamento textual pode ser dividido em duas etapas. A primeira consiste na triagem de documentos, cujo processo consiste na coleta dos textos e, em alguns casos, a conversão dos mesmos em estruturas bem definidas. Atualmente a triagem é bastante facilidade, pois não raro encontramos APIs capazes de acessar *datasets* e extrair deles documentos textuais bem definidos. No entanto, em alguns casos onde os textos podem apresentar alguma restrição, código de ética ou informações sensíveis (por exemplo, textos na jurimetria) que dificultam seu acesso ou sua definição.

A segunda etapa consiste na segmentação textual, nela temos o trabalho de converter o conjunto de documentos, denominado em NLP de *corpus* ou *corpora*, em segmentos bem definidos e limitados. Cada uma destes elementos, representados pelas palavras, correspondem ao que denominamos de *token* e seu processo de segmentação é denominado de *tokenização*.

Podemos ainda aplicar etapas correlacionadas a segmentação que são importantes no pré-processamento para a obtenção de documentos mais parcimoniosos. Vamos destacar quatro delas:

- **Normalização -** envolve a mesclagem de diferentes formas de escrita de um único *token* em uma forma canônica normalizada. Nesta etapa, além do *case-sensitive* (maiúsculas e minúsculas) e eliminação de caracteres especiais, também é realizada a correspondência de abreviaturas e acrônimos a seus *tokens*.

- **Eliminação de *stop words* -** consiste em *tokens* de ligação, identificados gramaticalmente como artigos, conjunções e proposições. No entanto, as *stop words* podem variar segundo o contexto empregado. Na recuperação de informação (*Information Retrieval*), palavras muito curtas tendem a ser classificadas como *stop words*, assim alguns verbos, os pronomes e alguns termos de negócio (como unidades de medida, por exemplo) podem ser considerados como tal.

- **Lematização -** consiste no processo de deflexionar uma palavra/*token*, empregado no contexto onde a flexão da palavra não é relevante (processo de *tagueamento* e uso de índices, por exemplo).

- **N-grama -** em suma, é uma técnica de agrupamento de *n tokens* segundo a correlação e/ou frequência que esses aparecem (sequencialmente) no texto. Em português, por exemplo,

palavras conjuntas como “guarda-roupa"(que perderá seu hífen no processo de normalização) e “bem humorado"podem gerar os bi-gramas: “guarda_roupa"e “bem_humorado". Há muitas formas quantitativas de determinar se um conjunto de *tokens* pode tornar-se um n-grama, o método mais comum, no entanto, consiste na informação mútua normalizada pontual (nPMI, do inglês *Normalized Pointwise Mutual Information*).

Apesar de ser uma etapa recorrentemente utilizada, a segmentação não deve ser uma regra. Luhn (1960) já havia discutido a cautela de sua utilização, visto que o uso de eliminação de *stop words* e da lematização podem impactar métodos de processamento textual que levam em consideração o contexto das palavras e não somente sua posição no documento. Isso ocorre, por exemplo, em textos pequenos, onde qualquer fragmento textual é importante para a captação do contexto e da informação do documento como um todo. Então, textos curtos podem se beneficiar do uso de *stop words*, já que muitas vezes estas são empregadas como ligações semânticas entre as unidades textuais. Nesse mesmo sentido, no que diz respeito a língua portuguesa, a lematização apresenta péssimos resultados na modelagem textual, isso porque boa parte dos lematizadores são funcionais para o inglês. E aqueles que são mais adequados para a língua portuguesa brasileira produzem lemas estranhos ou ambíguos (isso ocorre, geralmente, quando um mesmo radical gera palavras que ora se apresenta como verba, ora como adjetivo, por exemplo).

3.1.1 Representação Textual Baseado em Frequência

Discutimos um pouco sobre técnicas de pré-processamento que nos auxiliam na obtenção de um dado textual mais conciso. Agora precisamos determinar técnicas que transformará esse texto um uma estrutura mensurável e conveniente para então serem modelados. Em NLP, encontramos uma vasta gama de modelos responsáveis por esse morfismo, transformando uma unidade textual em um vetor numérico de modo que haja algum tipo preservação de estrutura. Neste caso, a unidade textual pode variar (letras, palavras, sentenças ou frases inteiras) e que o texto final, obtido após o pré-processamento, deve estar estruturado tal que essas unidades sejam identificadas como *tokens*.

Dado a sua simplicidade e sua facilidade de implementação, iremos tratar de duas técnicas consideradas de *one-hot encoding*. Em aprendizado de máquina, *one-hot encoding* é um processo que transforma unidades de dados em sinais positivos, dado sua existência sobre alguma conjunto de regras, ou em sinais negativos, caso ele não seja observado sobre as mesmas regras, McTear, Callejas e Griol (2016). As técnicas baseadas em *one-hot encoding* que são mais comumente usadas em NLP são o modelo de *Bag of Words* e *Term Frequency–Inverse Document Frequency* (TF-IDF), sendo esta última apresentada anteriormente na Subseção 2.3.4.

BAG OF WORDS

A primeira representação textual por incorporação é dada por um modelo bastante simples e que baseia-se na frequência dos *tokens* ao longo de todo o documento, McTear, Callejas e Griol (2016). Denominada *bag of words* (BoW), essa representação é computacionalmente fácil de se obter e, geralmente, exige pouco pré-processamento textual (*tokenização* e remoção de *stop words*[1]).

Segundo Qader, Ameen e Ahmed (2019), o BoW é comumente aplicado em problemas de detecção de objetos, classificação de imagens, reconhecimento de eventos e, para nosso caso, em problemas de classificação de documentos. Em todos estas aplicações, no entanto, o BoW contabiliza a frequência de cada palavra (*token*) válida em cada documento, segundo o vocabulário do *corpus*, construindo um vetor com a frequência de todos os *tokens* do vocabulário naquele documento. Uma importante característica desse tipo de representação textual, segundo o autor, é que ela se baseia em uma coleção não ordenada de *tokens*, de modo que seja ignorado ou perdido qualquer noção não trivial de gramática (construções semânticas e contexto, por exemplo), capturando apenas a multiplicidade das palavras.

Ainda sobre Qader, Ameen e Ahmed (2019), sua estrutura é baseada na união disjunta de dois documentos ("sacos"), somando as multiplicidades de cada elemento ("palavra"). O exemplo a seguir modela dois documentos segundo um modelo de BoW:

Nem todos aqueles que vagam estão perdidos.

***Bag* 1:** {nem: 1, vagam: 1, todos: 1, estão: 1, aqueles: 1, perdidos: 1}

Você pode encontrar coisas que perdeu, mas nunca as coisas que abandonou.

***Bag* 2:** {você: 1, pode: 1, coisas: 2, perdeu: 1, abandonou: 1, nunca: 1, encontrar: 1 }

No entanto, em implementações mais atuais, é comum atribuir índices aos *tokens* que compõe o vocabulário do *corpus* como alternativa ao uso do dicionário. Essa estratégia, conhecida como *hashing trick*, foi apresentada por Weinberger *et al.* (2010) e é muito utilizada no âmbito computacional para poupar uso de memória na armazenagem do dicionário.

TF-IDF

Uma outra técnica de representação textual bastante similar ao *bag of words* é denominada de TF-IDF (*Term Frequency Inverse Document Frequency*), comentada brevemente na Subseção 2.3.4. Enquanto o BoW olha unicamente para as frequências dos *tokens* nos documentos, dando pesos altos a palavras muito frequentes, o TF-IDF é quase uma extensão natural do BoW já

[1] A remoção de *stop words* é quase obrigatória no uso do modelo de *Bag of Words*, uma vez que a frequência destas é, geralmente, muito alta, implicando em problemas em modelos de agrupamento, por exemplo, que gerará tópicos apenas com *stop words* dada sua frequência nos documentos analisados.

que, apesar de ainda considerar a frequência de *tokens* nos documentos, o mesmo pondera tais frequências pela raridade dos *tokens* no corpus. Em outras palavras, os documentos ainda serão representados por vetores (sejam eles baseados no vocabulário, ou no sistema de índices, citados anteriormente), mas a frequência dos *tokens* será ponderada pelo inverso de suas frequências em todos os documentos.

Tomando Das e Chakraborty (2018), consideremos I o conjunto de índices que compõe o vocabulário de um *corpus*, contendo $j = 1,2,\ldots,J$ documentos; w como a variável que indexa a frequência absoluta das palavras/*tokens* dos textos; n_j o número de palavras do j-ésimo documento e $m_{i\in I}$ o número de documentos que contenham a i-ésima palavra presente no vocabulário. Temos que,

$$TF = \frac{w_{ji}}{n_j} \qquad IDF = \ln\left(1 + \frac{J}{m_{i\in I}}\right)$$

e portanto

$$\text{TF-IDF} = TF \times IDF.$$

Notamos que o TF-IDF de fato pode ser entendido por uma extensão do BoW, já que a parcela TF de seu cálculo nada mais é que a normalização das frequências do BoW. Além disso a ponderação nos apresenta uma outra interpretação sobre a importância de cada *token* para cada texto. Digamos que a palavra $i \in I$ é muito frequente em diversos documentos, isso implica que esta palavra diz muito pouco de algum desses documentos, já que ela é muito comum em todo *corpus*. Em contrapartida, se a frequência desta palavra é alta para o documento j, mas ela aparece em quase nenhum documento além dele, dizemos então que a palavra i traz muita informação contextual ao este documento.

3.2 Processo Hierárquico de Dirichlet na Modelagem de Tópicos Textuais

Vamos retomar ao modelo HDP visto na Seção 2.3, dado por

$$\begin{aligned} G_0|\gamma,H &\sim DP(\gamma,H) \\ G_j|\alpha,G_0 &\sim DP(\alpha,G_0) \end{aligned}$$

$$\begin{aligned} \theta_{ji}|G_j &\sim G_j \\ x_{ji}|\theta_j i &\sim F_{\theta\, ji}, \end{aligned} \tag{3.1}$$

para cada documento $j \in \mathfrak{J}$ e para cada palavra/*token* $i = 1,\ldots,n_j$. Comentamos anteriormente que uma das aplicações do HDP consiste no agrupamento textual e, devido a sua característica não-paramétrica, ele o faz inferindo a quantidade de grupos pelo processo e compartilhando informação intra e entre eles.

Em NLP é comum tratarmos o processo de agrupamento textual por modelagem de tópicos, onde entendemos por tópicos os rótulos atribuídos a cada grupo que pretendem explicar o conteúdo textual contido neles. Na Subseção 2.3.4, motivamos o processo de modelagem de tópicos considerando o problema de identificar as temáticas por trás de processos jurídicos nos baseando nas ocorrências de palavras dos mesmos (resultando, por exemplo, na catalogação de novos processos). Enquanto uma metodologia simples de agrupamento tenta colocar cada artigo em um único agrupamento, o HDP propõe uma classificação cruzada de artigos de acordo com tópicos, Teh e Jordan (2009). Isto é, um artigo pode ser principalmente sobre um tema, mas compartilha de outros tópicos afins.

Este efeito de compartilhamento de tópicos, nos permite mensurar o grau em que um documento trata cada tópico obtido. Para isso, vamos definir um tópico como uma distribuição de probabilidade de palavras, extraídas de um vocabulário, ou dicionário, V. Na Seção 2.3 discutimos sobre a especificação completa do HDP que, não por acaso, nos fornece uma estrutura que compartilha os pesos atribuídos aos tópicos θ_k^{**} em todos os níveis, de forma que o nível filho herde os pesos do nível mãe. Além disso, os tópicos devem ser vinculados para que os mesmos possam aparecer em documentos diferentes. Percebemos que na especificação do HDP, em 3.1, os sorteios feitos sobre G_0 são compartilhados entre as distribuições aleatórias G_j. Isso resulta em uma coleção (ou mistura) de modelos para cada documento.

Ainda na Seção 2.3, ao aludirmos sobre a inferência do HDP, percebemos que o processo de sorteios sobre as distribuições em cada nível hierárquico sofrem atualizações de acordo com a frequência da unidade amostral em cada nível. Em outras palavras, a posteriori do processo leva em conta a frequência das palavras em cada documento, bem como a frequência com que cada uma é alocada no conjunto de distribuições de cada documento. Analogamente, o processo também considera a frequência com que cada tópico aparece no *corpus* no DP do nível G_0. Sobre esse perspectiva, notamos que a forma com que o HDP representa suas unidades textuais coincide com os métodos de *BoW* e TI-IDF apresentados anteriormente, baseando-se em frequências.

Para a aplicação do HDP sobre a modelagem de tópicos, em NLP, consideremos: $\mathfrak{J}$ o conjunto de índices de documentos; x_{ji} a i-ésima unidade textual do j-ésimo documento; V como o tamanho do vocabulário de todo o *corpus*; k o índice do tópico, onde θ_k^{**} é sua massa pontual que representa sua distribuição sobre o vocabulário (isto é, θ_k^{**} é um sorteio em H); finalmente, θ_{ji} e θ_{jk}^{*} são, respectivamente, as variáveis latentes de agrupamento dos níveis das unidades textuais e dos documentos, onde ambas seguem um Processo de Dirichlet com seus respectivos parâmetros.

Por fim, retomando a estrutura na função $f_k^{-x_{ji}}$ (equação 2.14), Frank, Greenberg e Lindner (2020) apresenta uma solução para a mesma ao definirmos a estrutura de distribuições de probabilidade a cerca da aplicação do HDP em tópicos textuais. Esse resultado nos possibilita usar 2.14 como atualização das amostras dentro do processo de *Gibbs*. Já vimos que H (distribuição

base do processo) pode ser conjugada em $f_{\theta_k^{**}}$, na Subseção 2.3.2, então ao definirmos

$$f_{\theta_k^{**}}(\mathbf{x}_{ji}) \sim Mult(n, \theta_k^{**})$$
$$h(\theta_k^{**}) \sim Dir(\gamma),$$

a probabilidade em que $\mathbf{x}_{ji}$ é gerada por um tópico k, baseando-nos em Frank, Greenberg e Lindner (2020), é tal que

$$f_k^{-\mathbf{x}_{ji}}(\mathbf{x}_{ji}) \propto \frac{\Gamma(\gamma + n_{kv}^{-ji} + 1)\Gamma\left(N\gamma + n_k^{-ji}\right)}{\Gamma\left(\gamma + n_{kv}^{-ji}\right)\Gamma\left(N\gamma + n_k^{-ji} + 1\right)} \tag{3.2}$$

Algumas notações consideradas até agora foram adaptadas para o resultado anterior a fim de presar uma notação mais parcimoniosa. Desse modo, $v = x_{ji}$ é tido como índice associado a variável observada na iteração em questão, k é o índice do k-ésimo tópico associado a i-ésima observação, do j-ésimo documento, agrupada sobre o l-ésimo conjunto e w é o índice de associação a todas as observações $x_{j'i'}$, para $j', i' \neq j, i$. Detalhes do desenvolvimento de 3.2 podem ser encontrados no Anexo C.

Da mesma forma, precisamos atualizar a amostra para o caso de x_{ji} ser gerado por um novo tópicos, k_{new}. Analogamente ao processo anterior, podemos obtê-lo através de

$$f_{k_{new}}^{-\mathbf{x}_{ji}}(\mathbf{x}_{ji}) \propto \frac{\Gamma(N\gamma)\Gamma(\gamma+1)\prod_{w \neq v}\Gamma(\gamma)}{\Gamma(N\gamma+1)\prod_w \Gamma(\gamma)} \tag{3.3}$$

Também sobre Frank, Greenberg e Lindner (2020), é importante salientar que é computacionalmente recomendável o calculo de 3.2 e 3.3 aplicando-se a função logarítmica netas, uma vez que isso torna-se conveniente a tratabilidade das funções Gamma.

3.2.1 Método de Determinação dos Parâmetros de Concentração α e γ

Anteriormente, abordamos três métodos de estimação do parâmetro de concentração de um DP (Subseção 2.2.4). Baseando-nos no que foi discutido e nos resultados obtidos pelas análises de Liu e Nandram (2022), iremos obter α e γ a partir do método de razão de Uniformes. Para isso, consideramos

$$U, V \sim U(0,1),$$

tais que U e V são *i.i.d.*

Consideremos $\tau > 0$ um parâmetro de concentração arbitrário, tal que

$$\pi(\tau) = \frac{1}{(1+\tau)^2}$$

é sua priori não informativa, segundo Liu e Nandram (2022), a tal passo que

$$h(\tau) = \pi(\tau|k) \propto \frac{\tau^k \Gamma(\tau)}{\Gamma(\tau+n)(\tau+1)^2}$$

é descrita como sua distribuição a posteriori para o processo.

Logo, para amostrar um valor de τ seguimos com o Método da Razão de Uniformes sobre o seguinte algoritmo:

Algoritmo 2 – Razão de Uniformes para Obtenção de um Parâmetro de Concentração τ

Inicie $b = 1$ e $d = 1$
Gere u e v independentemente de $U(0,b)$ e $U(0,d)$
Tome $\tau^* = \frac{v}{u}$
se $u^2 \leq h(\tau^*)$ **então**
    $\tau = \tau^*$
senão
    Atualize $b = \sup_{\tau^*} \sqrt{h(\tau^*)}$ e $d = \sup_{\tau^*} \tau^* \sqrt{h(\tau^*)}$
fim se

Percebemos que o método em questão tem dependência em k, isso o torna particularmente interessante pois nos possibilita atualizar os parâmetros de concentração do HDP a cada indexação de novo tópico ou a cada extração de um tópico existente. É muito comum, Petitjean F. (2018), definirmos um k inicial igual ao número total de documentos do corpus para inicializarmos o HDP e, consequentemente, para obtermos α e γ. Logo, seguimos com esta estrutura em nossas aplicações.

CAPÍTULO

4

APLICAÇÃO DO PROCESSO HIERÁRQUICO DE DIRICHLET NA MODELAGEM DE TÓPICOS EM PROCESSOS JURÍDICOS

Anteriormente, discutimos sobre como o HDP é uma excelente ferramenta para modelar tópicos textuais graças a sua estrutura não paramétrica. Nos últimos anos, o campo jurídico tem se beneficiado com a Estatística, uma vez que esta fornece metodologias sólidas de análise descritiva e inferência, sobre tudo com seus modelos baseados em probabilidade para extrapolação de resultados e mensuração de incertezas. Desse modo, a modelagem de tópicos por HDP encontra um campo fértil na Jurimetria, disciplina resultante da aplicação de métodos estatísticos na análise de processos e fatos jurídicos, como motivamos na Subseção 2.3.4. Assim o modelo atua no agrupamento dos processos jurídicos da seguinte maneira: definimos um tópico como uma distribuição de probabilidade entre um conjunto de palavras retiradas do algum vocabulário que compõe o corpus de processos. Cada processo, por sua vez, é um documento modelado como uma distribuição de probabilidade entre tópicos (em geral, um documento estará associado a vários tópicos).

Neste Capítulo, faremos uma rápida introdução ao conceito de Jurimetria e discutiremos um pouco mais sobre a aplicação da modelagem de tópicos na área, Seção 4.1. Nas Seções 4.2, 4.3 e 4.4 partiremos para descrição dos dados de aplicação do modelo, seguido da descrição dos métodos de processamento textual empregados e da discussão sobre a métrica de análise de resultados, respectivamente. E finalmente, na Seção 4.5 analisaremos os resultados da aplicação.

4.1 Conceito de Jurimetria

É notório que a estatística como ferramenta analítica, descritiva e inferencial, tem grande importância nas mais diversas áreas científicas. Dentro do contexto jurídico, além destes im-

portantes fatores, a estatística ainda ganha outras funcionalidades. Nunes (2019) define como Jurimetria a disciplina do conhecimento que utiliza a metodologia estatística para investigar o funcionamento de uma ordem jurídica. Nesse sentido, o que se refere à análise descritiva de populações, a Jurimetria permite o estudo do comportamento coletivo dos agentes jurídicos, possibilitando o isolamento das características destes sujeitos de Direito do comportamento de cada indivíduo. Isto é, se no âmbito jurídico um "sujeito de direito" é utilizado para se referir ao cidadão não apenas como pessoa física, mas incluindo-o em entidades coletivas, como empresas, associações civis e organizações não-governamentais. Dessa forma, a estatística possibilita dissociar, em sua forma descritiva, o comportamento dos indivíduos do fato social.

Já em seu fator inferencial, a Jurimetria, em todo seu arcabouço estatístico, possibilita prever reações coletivas diante de alterações no ambiente social através de modelos probabilísticos. Tornando-se possível não só estabelecer fatores causais e generalizar resultados, como controlar incertezas no Direito. Isto é, contribui para a mitigação da dubiez jurídica por parte dos operadores do Direito (por exemplo, os juízes trabalhando para controlar e prever os efeitos da sua decisão de forma a fazer justiça). Como resultado, a Jurimetria ao estabelecer a análise dos comportamentos coletivos em função das normas jurídicas, permite a compreensão sobre o funcionamento do Direito e viabiliza a criação de modelos capazes de aproximar os resultados produzidos pela ordem jurídica das expectativas e aspirações sociais, Nunes (2019).

Portanto, o objetivo da Jurimetria é a investigação do funcionamento da ordem jurídica, isto é, do conjunto de normas jurídicas que busca influenciar o comportamento humano através da aplicação de sanções. Neste sentido, modelos de tópicos são ferramentas eficazes no estudo destas ordens através de seus resultados e do agrupamento lógico de seus temas e comportamentos. Poudyal, Gonçalves e Quaresma (2019), por exemplo, apresentaram uma proposta para identificar automaticamente argumentos em documentos legais. Em sua abordagem, utilizaram um algoritmo de agrupamento *Fuzzy c-means* (FCM) com proposta de avaliar com um conjunto de decisões jurisprudenciais do Tribunal Europeu dos Direitos Humanos (CEDH), revelando resultados bastante promissores.

Da mesma forma, pretendemos aplicar o modelo de tópicos por HDP no contexto jurídico, afim de obter tópicos informativos que possibilite acessar de forma mais objetiva e informativa processos segundo o procedimento jurídico aplicado. Esse tipo de agrupamento nos permite, por exemplo, associar temas recorrentes sobre cada procedimento, bem como entender quais temas tendem a parecer juntos a partir do processo analisado. Além deste perfil analítico, através do modelo poderíamos entender a partir de um novo processo em andamento, qual seria(m) seu(s) tópico(s) correlato(s), o que ajudaria operadores de Direito a mitigar suas incertezas e controlar o resultado do processo com base naqueles encontrados pelo(s) tópico(s).

4.2 Datasets

Consideramos um conjunto de dados composto por um *corpus* sobre sentenças judiciais separadas em três classe de procedimentos: ordinário, sumário e juizado especial cível. Geralmente os procedimentos se diferenciam pelo tempo processual e pela pena aplicada sobre cada um, dado o julgamento. Processos Ordinários e Sumários são casos de procedimentos comuns na jurisdição, Badaró (2019). Geralmente tratam de casos com maior celeridade processual, caracterizadas por um ritmo menos formal e por soluções rápidas que lidem em benefício das partes (os valores envolvidos nas causas desse tipo de processo são baixos). Diferenciam-se de acordo com os prazos de audiências e recursos aplicados no processo. Em contrapartida, Processos de Juizado Especial Cível geralmente são casos mais complexos com ganhos maiores. Tais processos têm como atribuição conciliar, processar e julgar as causas cujos valores não ultrapassem 40 salários-mínimos. Suas causas são mais específicas e mais longas de serem apuradas, geralmente envolvendo crimes contra honra, propriedade imaterial e, em alguns casos, crimes funcionais e dolosos.

Além dos tipos de procedimentos, o conjunto de dados conta com as seguintes covariáveis:

- `magistrado` - nome dos juízes responsáveis pelos respectivos processos.
- `assunto` - área na qual o processo foi realizado.
- `txt` - consiste no documento propriamente dito. Cada texto possui um sintetização do processo, informações sobre recursos (se houver) e a sentença do mesmo dada pelo seu respectivo juiz. Os textos apresentam um cabeçalho e não são raros a presença de informações sensíveis (como nome e documento do requerente, por exemplo).

A partir disto, temos por objetivo obter agrupamentos textuais tais que:

- Nos permita explorar os temas identificados pelos tópicos que compões o corpus, sumarizando assim os tipos de processos por assunto.
- Possibilitem extrair, previamente, tendências sobre a sentença de um novo processo, baseada nas análises dos tópicos obtidos pelo *corpus* jurídico atual.
- Cataloguem de forma automática o corpus, facilitando sua consulta e compreensão.

4.3 Pré-Processamento

Dado as técnicas de pré-processamento textual, discutidas na Seção 3.1, tomamos os seguintes passos para tratar os dados textuais:

- **Normalização -** eliminamos caracteres especiais, como símbolos e acentuação das palavras, e eliminamos *case-sensitive*, deixando todos os *tokens* em minúscula.

- ***Stop words* -** retirada de *tokens* de ligação (artigos, conjunções e proposições) e unidades de medidas.

- **Informações sensíveis -** também eliminamos tokens que contenham informações sensíveis, como dados pessoais, nomes dos envolvidos e afins.

- **Bigrama -** o processo de bigrama foi aplicado de forma sistemática para construir *tokens* especiais. Dentro do contexto jurídico, citações de Leis e Artigos são recorrentes em textos e aplicação de sentenças. Afim de conservar a informação contida nesses elementos, aplicamos regras baseadas em *regex* (expressão regular) que primeiro, identifica a palavra "*lei*" e suas variações no documento e, dado sua identificação, busca pelo número ou acrônimo referido a ela. Em seguida, aplicamos o mesmo processo para a palavra "*artigo*" e suas variações. Por fim, com os *tokens* de "`lei`"e "`artigo`" gerados, unimos ambos em um único *token* já que os Artigos, no Direito, só fazem sentido se acompanhados da Lei que os executa (por sua vez, as Leis têm sentido próprio e determinantes). Como resultado, geramos um *token* "`lei_artigo`" e um *token* "`lei`". Abaixo temos um exemplo:

> *[...] sentença nos termos do parágrafo 4º do art 20 do código de processo civil condicionado aos termos do art 12 da lei 1060.*

***Tokens* Especiais:**

1 - {'`lei_cpc_artigo_20`', '`lei_cpc`'}

2 - {'`lei_1060_artigo_12`', '`lei_1060`'}

- **Remoção de repetição de estruturas léxicas -** também aplicamos *regex* para remover repetições léxicas correspondentes as cabeçalhos e expressões padronizadas dos textos jurídicos.

4.4 Métrica de Avaliação: Coerência

Uma das principais métricas de qualidade de tópicos é conhecida, em NLP, como coerência. Dizemos que um conjunto de fatos, em nosso caso de elementos textuais, que constituem um tópico é coerente se eles se sustentam mutuamente. Isto é, se o tópico apresenta um contexto capaz de correlacionar todos seus elementos. A grande questão nesse caso está em como quantificar a coerência de um conjunto de elementos textuais. Röder, Both e Hinneburg (2015) apontam alguns aspectos para estabelecer tal medida:

- Comparar cada elemento com todos os outros;

- Comparação par a par de elementos;

- Comparar subconjuntos disjuntos de elementos entre si.

O interesse na medida de coerência surgiu com a mineração de texto, uma vez que modelos de tópicos e outros métodos de aprendizado não supervisionados não oferecem garantias sobre a interpretabilidade de seus resultados. Em termo gerais, modelos de tópicos são representados, geralmente, por conjuntos de palavras (elementos textuais) rotuladas de forma não supervisionada. Nesse sentido a medida de coerência classifica cada tópico quanto à sua compreensibilidade.

Complementar aos possíveis aspectos apresentados por Röder, Both e Hinneburg (2015), que nos oferece esquemas para comparar elementos textuais, podemos considerar também métodos úteis para estimar probabilidades de palavras e normalizar comparações numéricas. Neste sentido, o autor explora as principais estruturas que unificam ambos aspectos e define métricas de coerência.

Dentre as principais as principais métricas de coerência, destacam-se:

- C_{UCI} - *score* de coerência baseado em janelas deslizantes e na informação mútua pontual de todos os pares de *tokens*, usando as principais *n* palavras por tópicos. As janelas são construídas a partir de um processo pré-treinado sobre um *corpus* de mais de 2 milhões de artigos da *Wikipedia*.

- C_{UMass} - este *score* calcula a coerência simplesmente pela frequência com que duas palavras aparecem juntas no corpus.

- C_{nPMI} - neste caso, a coerência é calculada baseada na distância entre vetores de contexto de cada palavra, ou por sua informação pontual mútua (PMI), e normalizando tal medida, a fim de obter valores no intervalo $[-1;1]$.

Os dois primeiros *scores*, dadas as aplicações sobre os *corpus* em questão, se mostram inadequados. O primeiro, devido ao método pré-treinado sobre documentos de língua inglesa; e o segundo por basear-se em frequências conjuntas entre palavras, sendo influenciado por textos com estruturas léxicas repetidas.

Portanto, tomamos como métrica de avaliação para os nossos modelos a coerência baseada em informação pontual mútua normalizada. Essa baseia-se em vetores de contexto (v) para cada palavra principal do tópico. Um vetor de contexto de uma palavra x é definido usando contagens de coocorrências de determinados *tokens* em janelas de tamanho 10, selecionadas simetricamente ao redor de x. Segundo o autor, a maior correlação com classificações de coerência de tópicos avaliados por humanos (padrão ouro) foi encontrada ao definir os elementos desses vetores como nPMI,

$$v_i = \big(nPMI(x_i,x_1), nPMI(x_i,x_2), \ldots, nPMI(x_i,x_{i-1}), nPMI(x_i,x_{i+1}), \ldots, nPMI(x_i,x_n)\big)$$

Após a obtenção dos vetores, a C_{nPMI} é calculada segundo a soma das medidas de similaridade por cosseno entre os vetores de cada palava:

$$C_{nPMI} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \cos\left(v_i, v_j\right).$$

Dado sua normalização e a restrição de seus valores no intervalo $[-1;1]$, interpretamos seus sinais em três intervalos. Valores de coerência que tendem a -1, imprimem uma completa dissociação de palavras que descrevem o tópico, gerando tópicos sem ou com quase interpretabilidade. Em contra partida, quando a coerência tende a 1, observamos uma alta interpretabilidade dos tópicos e uma fácil identificação dos assuntos. Por fim, quando os valores de coerência ficam em torno de zero, isso implica independência entre as palavras que compõe o tópico em questão, em geral, os tópicos podem apresentar alguma interpretabilidade, mas são acarretados de palavras com alta frequência e pouco significado de contexto.

4.5 Resultados

Ajustamos os modelos de HDP, após o processamento de dados, sobre cada um dos *corpus* estratificados, segundo o procedimento jurídico aplicado, dos textos processuais. Para isso, contamos com a biblioteca `tomotopy`, em *Python*, desenvolvida por Fenstermacher, Schneider e bab2mi (2021), suficientemente otimizada para o ajuste dos modelos.

No geral, para cada estrato, consideramos um k inicial igual ao número de documentos presentes em cada estrato. Durante o processo medimos o número de tópicos e calculamos a distribuição marginal pelo método livre de verossimilhança, Rodrigues, Nott e Sisson (2020) para analisar a convergência do processo. Em suma, segundo os autores, o valor do logaritmo de verossimilhança para cada iteração pode ser interpretado como uma pontuação de convergência do algoritmo. Uma vez que ele se estabiliza a partir de uma iteração, então temos que o número de iterações é suficiente para extrair os tópicos do conjunto de dados. Os valores das métricas em questão eram obtidos a cada 10 iterações do processo. Sobre o processamento computacional, tomamos um *burn-in*, definido analiticamente, de 600 iterações (onde, a partir deste, observamos certa estabilidade do processo para todos os estratos), e atribuímos um limite mínimo de iterações válidas do processo igual a 1000 para que pudéssemos avaliar a convergência do modelo sobre as métricas em questão.

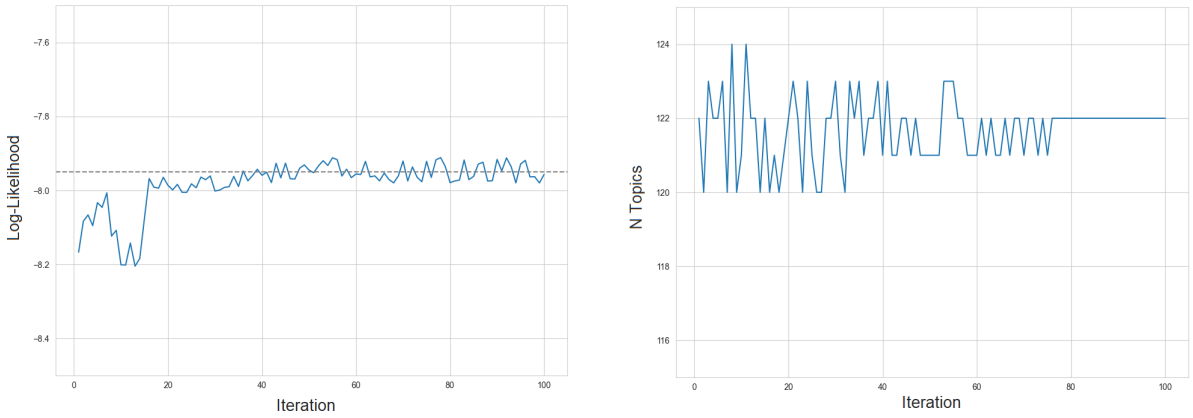
A respeito da análise e interpretação dos tópicos, para cada estrato apresentamos um gráfico de dispersão do tópicos e as tabelas contendo as palavras chaves de cada um deles. Para

melhorar a visualização e interpretação dos mesmos, adicionamos a regra de considerar tópicos com, ao menos, 1% de proporção de cobertura do corpus (o Apêndice D trás os gráficos de dispersões ao considerar todos os tópicos ajustados). Neste mesmo sentido, removemos as top 50 *tokens* mais frequentes em cada um dos ajustes (listamos tais *tokens* no Apêndice E) no intuito de elimina expressões repetitivas que geralmente não agregam informação ao texto e pode interferir negativamente na construção de um tópico. Por fim, apresentamos o valor médio de coerência, calculado segundo o método discutido anteriormente.

No geral, apesar de uma coerência próxima de zero (indicando independência nas palavras que compõe os tópicos), em todos os três casos, o modelo apresentou certa consistência na obtenção de tópicos e gerou um bom entendimento sobre as temáticas envoltas em cada tipo de procedimento.

PROCESSOS POR PROCEDIMENTO DE JUIZADO ESPECIAL CÍVIL

Neste estrato contamos com 6277 documentos e um vocabulário contendo 847296 palavras distintas. Ao final do processo obtivemos os parâmetros de concentração $\alpha = 0.0280$ e $\gamma = 20.5050$, com um número total de 122 tópicos encontrados, e uma coerência média entre os tópicos de $\overline{C}_{nPMI} = -0.0444$. A Figura 7 apresenta a convergência sobre a log-verossimilhança e sobre o número de tópicos, considerando o *burn-in* de 600 iterações, mostrando que em ambos os casos houve estabilidade nos valores.



(a) Convergência pelo método livre de verossimilhança. (b) Convergência pelo número de tópicos.

Figura 7 – Estudo de convergência sobre o *corpus* de processos com aplicação de procedimento especial cível.

Fonte: Elaborada pelo autor.

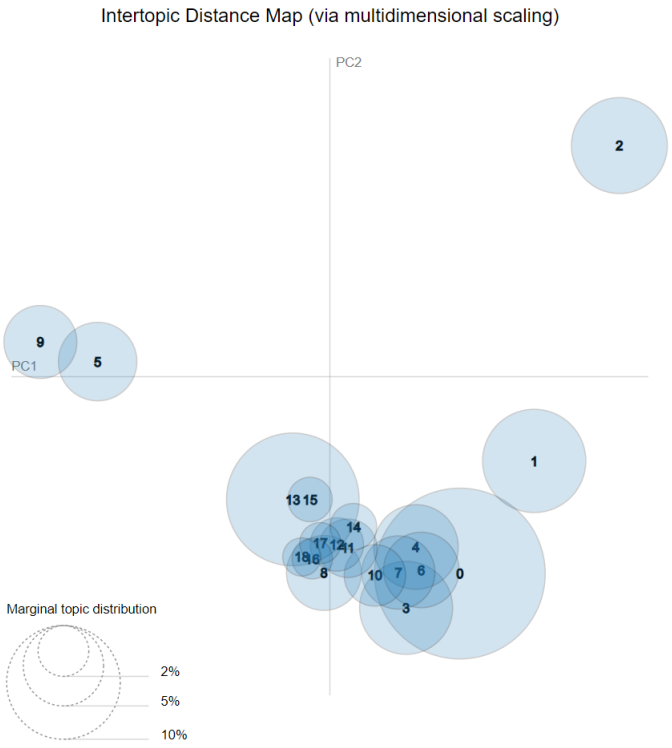
Na Figura 8, encontramos o gráfico de dispersão dos tópicos para aqueles com cobertura maior ou igual a 1% do *corpus*. Resumidamente, cada ponto/"bolha" representa um tópico, onde o seu tamanho capta a cobertura ou a concentração de informação baseada no *corpus* considerado. A distância entre os pontos/"bolhas" indicam o quão distintos os tópicos são uns dos outros. No geral, é desejado que o tópicos apresentem tanho similar e que esteja bem distribuídos sobre as componentes do plano, isto é, que não ocorram interseções e/ou sobreposições entre eles. Podemos notar que neste caso um emaranhado de tópicos se sobrepuseram enquanto os tópicos $1, 2, 5, 9$ apontam uma boa distinção de de informação entre eles, presando pelo comportamento desejável do gráfico.

Finalmente, a Tabela 1 trás a lista de palavras chave para cada tópico. Aqui percebemos alguns tópicos bastante relevantes quanto ao nível de informação. O Tópico 15, por exemplo, parece trazer o cerne do problema nos processos, vinculado a plano de saúde e sua falta de cobertura em casos de reembolso. Da mesma forma, os Tópicos 6 e 18 parecem apresentar como o requerente se sente diante o problema, o que indica prejuízo financeiro ao mesmo. Dos os Tópicos $1, 2, 5, 9$, citados anteriormente, destacam-se:

- **Tópico 1 -** temas relacionados a instituições publicas.

- **Tópico 2 -** temas relacionados a instituições financeiras e cobrança de tarifas.
- **Tópico 5 -** execução de bens do devedor pelo fiador (segundo Código de Processo Civil artigo 794).

Figura 8 – Representação dos tópicos para documentos de processos sobre procedimentos de juizado especial cível segundo suas componente.



Fonte: Elaborada pelo autor.

Tabela 1 – Palavras chave dos principais tópicos para processos sobre procedimentos de juizado especial cível.

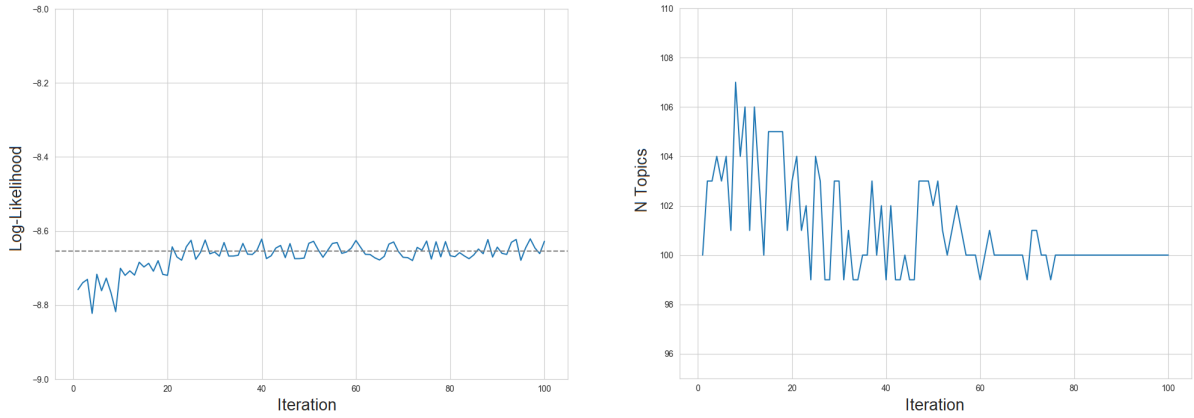
Tópicos	**Palavras Chave**
Tópico 0	debito, linha, banco, servicos, qualquer
Tópico 1	cento, cinco, porque, processuais, negativacao
Tópico 2	tarifa, cmn, tac, abertura, instituicao
Tópico 3	etapa, primeira, recolhido, minimo, calcular
Tópico 4	devera, jurisdicao, fins, primeiro, advogado
Tópico 5	cautelas, praxe, lei_cpc_artigo_794, inciso, oportunamente
Tópico 6	existencia, devedores, cadastro, inscricao, ofendido
Tópico 7	pedidos, apontamento, vitima, passara, alude
Tópico 8	recursal, levandose, consideracao, tecnica, corrigido
Tópico 9	desentranhamento, constantes, constante, egregia, fichamemoria
Tópico 10	cartao, senha, saque, saques, banco
Tópico 11	prescricao, condenado, lei_1160_artigo_4, pagar, lei_1160
Tópico 12	banco, art55, originais, desentranhandose, pertencentes
Tópico 13	acordo, homologo, produza, juridicos, efeitos
Tópico 14	observado, sobre, minimo, lei_cc_lei_1160_artigo_4, contar
Tópico 15	saude, cobertura, plano, clausula, reajuste
Tópico 16	individuo, normalidade, psicologico, antecipado, sofrimento
Tópico 17	conflito, colegio, capital, encontro, contratual
Tópico 18	lesao, tristeza, personalissimo, desconforto, filho

Fonte: Dados da pesquisa.

PROCESSOS POR PROCEDIMENTO ORDINÁRIO

Quanto aos processos ordinários contamos com um total de 10471 documentos e um vocabulário contendo 4327807 *tokens* distintos. Encontramos os valores dos parâmetros de concentração $\alpha = 0.0085$ e $\gamma = 15.1870$, ao final do processo, bem como um total de 100 tópicos e coerência média de $\overline{C}_{nPMI} = 0.0323$.

A Figura 9 mostra a convergência do processo, considerando o *burn-in* de 600 iteração para o método livre de verossimilhança e para o número de tópicos.



(a) Convergência pelo método livre de verossimilhança. (b) Convergência pelo número de tópicos.

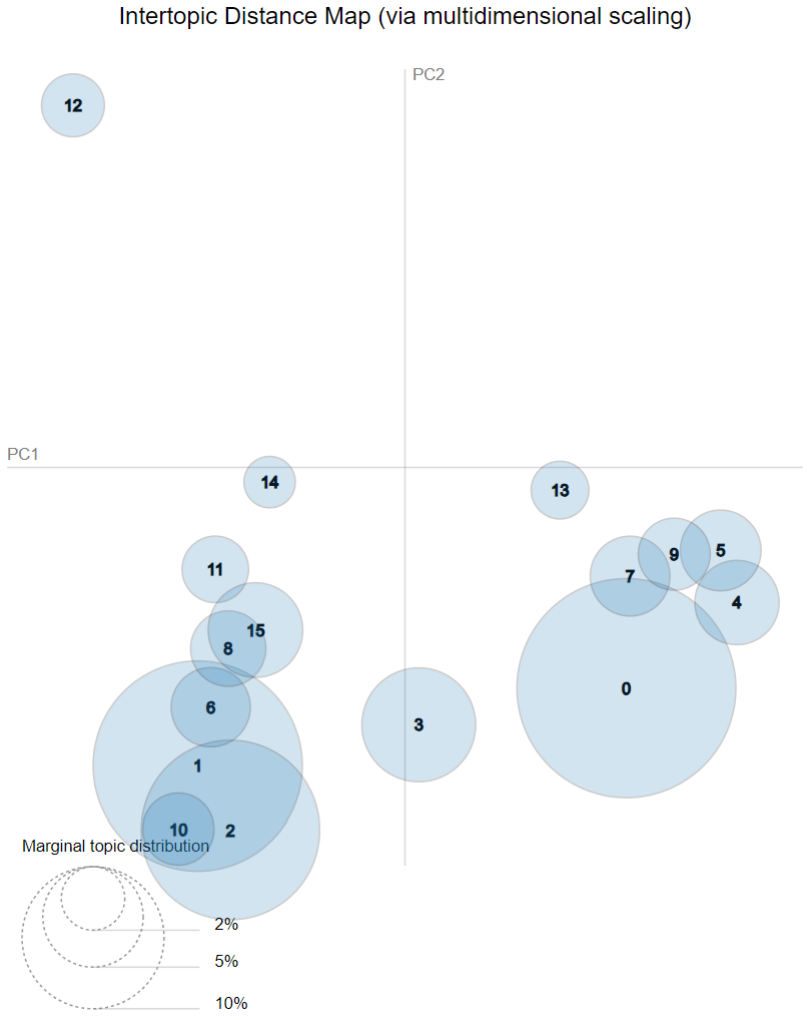
Figura 9 – Estudo de convergência sobre o *corpus* de processos com aplicação de procedimento ordinário.

Fonte: Elaborada pelo autor.

Quando a dispersão dos principais tópicos no gráfico, apresentado na Figura 10. notamos um comportamento mais próximo ao esperado (tópicos com relativamente um mesmo tamanho e separados).

Por fim, a Tabela 2 traz as palavras chave para os tópicos em questão. O Tópico 3, por exemplo, trata de problemas trabalhistas em seus direitos. Outro tópico que se destaca, o Tópico 0, traz consigo temas ligados a financiamento e pagamento por amortização de dívidas. E o Tópico 12, o mais afastado dentro todos, carrega o emprego do artigo 794 (Código de Processo Civil), de execução de bens, sobre leis de transito (ao se avaliar as demais palavras do Tópico). Esse último tópico analisado levanta um caráter importante dos *tokens* especias. Como geralmente Leis e Artigos são usados na execução de um processo, muitos deles se repetiram em diferentes tópicos desde que partilhem de um contexto semelhante. A primeiro momento isto parece um problema, porém é um reflexo na prática de como as palavras podem intercambiarem-se sobre os diferentes tópicos.

Figura 10 – Gráfico de dispersão dos tópicos, para processos com procedimento ordinário adotado, segundo suas componentes.



Fonte: Elaborada pelo autor.

Tabela 2 – Palavras chave dos principais tópicos para processos sobre procedimentos ordinário.

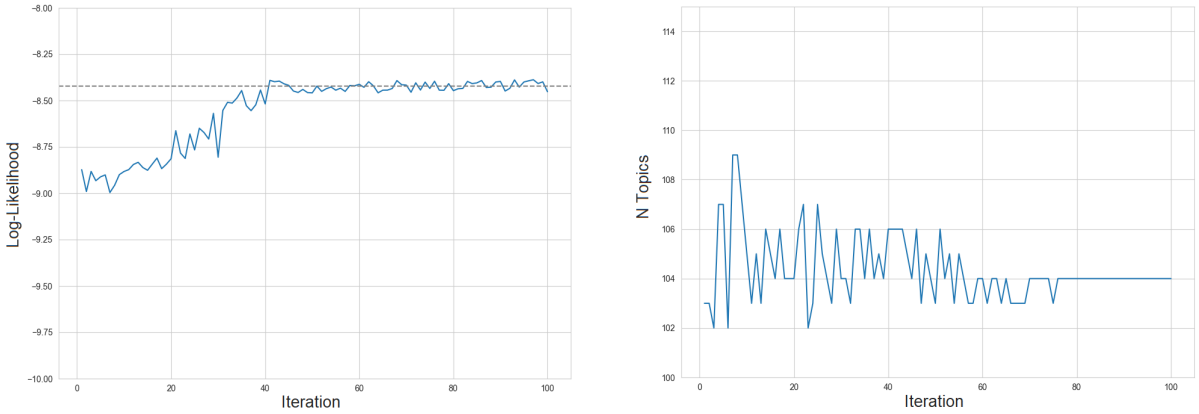
Tópicos	**Palavras Chave**
Tópico 0	comissao, tarifa, remuneratorios, permanencia, encargos
Tópico 1	cartao, requerida, inexigibilidade, responsabilidade, cadastros
Tópico 2	responsabilidade, culpa, inscricao, requerida, vitima
Tópico 3	descontos, desconto, folha, vencimentos, salario
Tópico 4	tarifa, abertura, tac, agrg, bancario
Tópico 5	comissao, permanencia, nacional, grifei, financeiro
Tópico 6	saude, cobertura, tratamento, medico, plano
Tópico 7	tarifas, tarifa, amortizacao, capital, comissao
Tópico 8	saude, plano, lei_9656, lei_9656_artigo_31, aposentado
Tópico 9	: boafe, encargos, objetiva, capital, norma
Tópico 10	responsabilidade, culpa, risco, agente, civil
Tópico 11	acoes, integralizacao, participacao, dividendos, subscricao
Tópico 12	exequente, homologo, lei_cpc_artigo_794, levantamento, extinta
Tópico 13	tarifa, abusividade, confira, provisoria, ano
Tópico 14	poupanca, caderneta, plano, ipc, indice
Tópico 15	aposentadoria, previdencia, complementacao, economus, beneficio

Fonte: Dados da pesquisa.

PROCESSOS POR PROCEDIMENTO SUMÁRIO

Finalmente sobre os processos sumários temos um total de 2051 documentos e um vocabulário com 505413 *tokens* distintos. Encontramos os valores dos parâmetros de concentração $\alpha = 0.0261$ e $\gamma = 22.7480$, um total de 104 tópicos distintos e coerência média de $\overline{C}_{nPMI} = -0.0649$.

A convergência do processo é mostrada na Figura 11, aproximadamente, a partir da iteração de número 400 pelo o método livre de verossimilhança (considerando o *burn-in* de 600 iterações).



(a) Convergência pelo método livre de verossimilhança. (b) Convergência pelo número de tópicos.

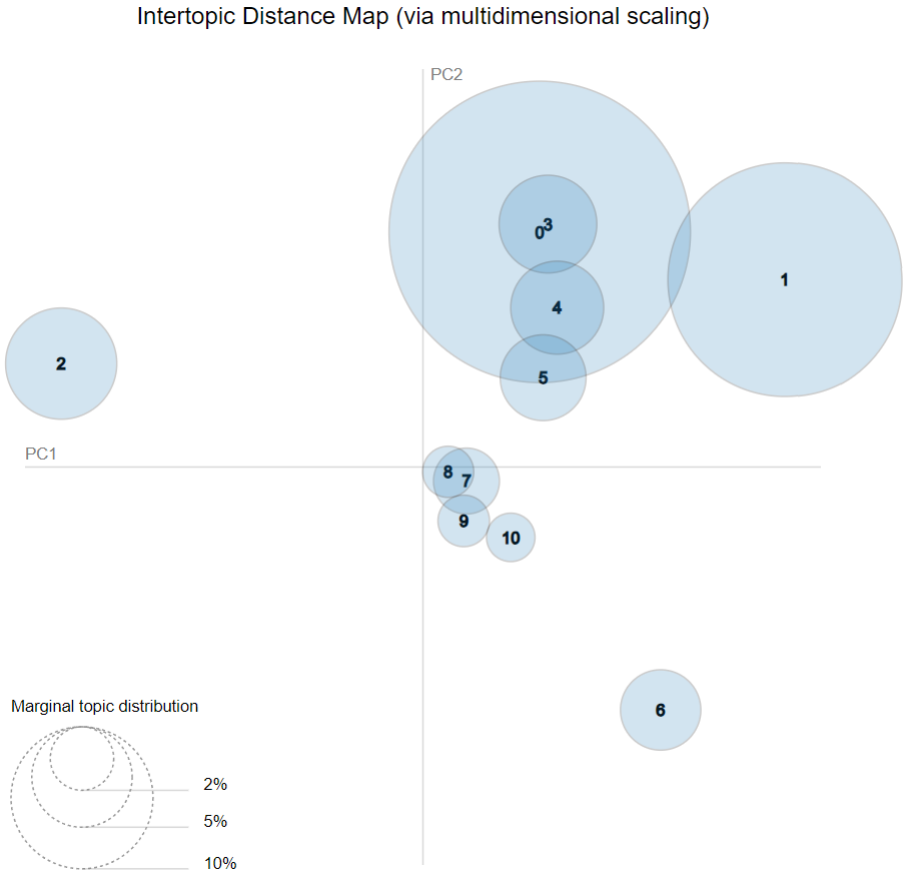
Figura 11 – Estudo de convergência sobre o *corpus* de processos com aplicação de procedimento ordinário.

Fonte: Elaborada pelo autor.

Neste estrato foi onde nos deparamos com tópicos "dominantes", isto é, tópicos que tem um tamanho desproporcionalmente maior que os demais, caso dos Tópicos $0, 1$, no gráfico de dispersão da Figura 12. Isso pode ocorrer quando temos um excesso de *tokens* com alto frequência, mas com "pouco valor" (entenda-se por "pouco valor", expressões que são utilizadas por praxe em um determinado contexto, como o jurídico), além de variações de uma mesma palavra devido a sua flexão de lexical (feminino, masculino, singular, plural).

Apesar disso, o Tópico 5, segundo a Tabel 4, apresenta uma temática claramente ligada a ajuste de juros e correções de inflação pelo Índice de Preços ao Consumidor (IPC) sobre cardenetas. Aqui, novamente encontramos um tópico ligado a execução de bens (Tópico 2). Tal tópico foi sistematicamente encontrado em todos os três estratos e representado de forma muito característica no gráfico de dispersão (sempre isolado dos demais). Isso é especialmente interessante, pois mostra a consistência do modelo em *corpus* diferentes.

Figura 12 – Distribuição dos tópicos de processos sumários segundo componentes.



Fonte: Elaborada pelo autor.

Tabela 3 – Palavras chave dos principais tópicos para processos sobre procedimentos sumário.

Tópicos	**Palavras Chave**
Tópico 0	responsabilidade, requerida, cartao, protecao, conta
Tópico 1	capitalizacao, comissao, tarifa, encargos, permanencia
Tópico 2	exequente, levantamento, expecase, lei_cpc_artigo_794, execucao
Tópico 3	plano, saude, lei_9656, mesmas, lei_9656_artigo_31
Tópico 4	saude, tratamento, cobertura, medico, plano
Tópico 5	poupanca, caderneta, indice, ipc, correcao
Tópico 6	equilibrio, clausulas, exigidos, contratuais, regime
Tópico 7	responsabilidade, culpa, risco, atividade, requerida
Tópico 8	acoes, integralizacao, participacao, prescricao, subscricao
Tópico 9	agrg, rel, cerceamento, min, eresp
Tópico 10	seguranca, boafe, juridica, nacional, ano

Fonte: Dados da pesquisa.

CAPÍTULO

5

CONSIDERAÇÕES FINAIS

Após discorrer sobre a estrutura por trás do modelo de Processos Hierárquicos de Dirichlet (HDP) e levantar as problemáticas sobre o processamento e análise de dados textuais, exploramos o desempenho do HDP aplicado em um corpus jurídico. Neste contexto, os documentos tinham por características *tokens* especiais que identificam fortemente o tema do processo em questão; estruturas contendo repetição léxicas que consistiam em padrões repetidos em quase todos os documentos (como um cabeçalho contendo informações sensíveis do requerente); e uma falta de padronização entre a estrutura do texto quanto a suas seções e modos de execução, mesmo tratando-se de um mesmo procedimento conduzido pelo mesmo magistrado. Aplicamos um tratamento que pretendia mitigar efeitos negativos destas características para que então o corpus fosse representado por vetores numéricos baseados na frequência das palavras e finalmente processados pelo modelo.

Apesar da consistência em nosso fluxo de trabalho, recebemos valores de coerência ruins que indicam uma inconsistência nos tópicos, isto é, indica que as palavras que os compões são bastante diversas fazendo com que os tópicos fossem pouco informativos. Contrariando a métrica de avaliação escolhida, percebemos que os principais tópicos, aqueles que cobriam ao menos 1% do vocábulo do *corpus*, apresentaram interpretações interessante sobre as temáticas abordadas por ele. Logo, apesar de considerarmos bastante razoáveis os resultados obtidos, suficientes para contribuir significativamente como ferramenta em Jurimetria, podemos levantar alguns efeitos que consigam minimizar a contradição encontrada.

O primeiro, e o mais importante, seria quanto ao pré-processamento. Ao trabalharmos com os *tokens* especiais vimos uma melhora significativa nos resultados dos testes executados sobre uma amostra do corpus, teste estes que faziam parte da validação do código e do processo em si. O tratamento das Leis e Artigos no contexto jurídico mostrou-se extremamente valioso na identificação de tópicos. Antes de o aplicarmos, observamos cerca de 3 tópicos por estrato e após sua aplicação aumentamos a detecção de tópicos para, aproximadamente, 100 por estrato.

Então, buscar melhorias no sentido de incorporar palavras ou expressões próprias e com valor alto em significado no contexto jurídico (como o casos de Leis e Artigos) é uma boa maneira de assegurar a qualidade dos dados e torná-los mais "consumíveis" por técnicas mais simples de representação textual, como a aplicada no trabalho.

Outro ponto decorre diretamente da representação textual no processo. Sabemos que uma estrutura textual é composta de inúmeros elementos, sendo a coesão e coerência dois dos mais importantes. Enquanto a coerência trata do sentido lógico textual, a coesão trata da estrutura gramatical de sentenças ao longo do texto. Fato é que ambos são imprescindíveis para o entendimento do texto. Logo, uma estrutura de representação textual que respeite e acople esses elementos tende a gerar resultados em estruturas numéricas mais valiosas. O que não ocorre na representação por TF-IDF, já que está se baseia na frequência da palavra, se se quer importar-se com a posição das mesmas no texto.

Por fim, podemos levantar outras métricas e criar estruturas que nos possibilitem avaliar os tópicos de diferentes formas. Uma boa métrica de análise conjunta a coerência é a medida de perplexidade. Trata-se da mensuração da bondade do modelo de tópicos ao prever novos dados, refletindo sua capacidade de generalização. Nesse sentido, aplicar um método supervisionado em parte do conjunto de dados, onde para isso a rotulação de tópicos em cada documento deve ocorrer de forma manual e por um especialista, também pode gerar melhores percepções sobre o comportamento dos tópicos ao longo do modelo HDP.

Trabalhos Futuros

Baseado nas observações anteriores, propomos melhorias no processo para trabalhos futuros. Desejando captar melhor o contexto presente nos documentos e melhorar a representação textual para o modelo de HDP, nossa primeira proposta é substituir a representação TF-IDF por vetores de *embeddings* via modelo *FastText*, apresentado em Zhang *et al.* (2020) e Athiwaratkun, Wilson e Anandkumar (2018). Esse tipo de modelo consegue construir vetores de representação textual baseados no posicionamento das palavras e na correlação entre ela e sua vizinhança, ainda permitindo a construção de vetores por palavras. Com essa nova representação em mãos, precisamos então adequar as distribuições de probabilidade do modelo que reflitam os dados mais adequadamente. Nesse sentido, podemos trabalhar com o modelo normal ao justificarmos o uso dessa distribuição sobre a média dos vetores de *embeddings* das palavras de um documento. Logo, a representação do documento seria um único vetor que é então incorporada a distribuição normal para que conseguimos construir uma conjugada adequada gerando então um modelo HDP-GMM (*Hierarchy of Dirichlet Process Gaussian Mixture Models*), similar ao proposto proposto em Rinaldi e Pozzo (2021). Com isso, esperamos trazer uma melhora no processo gerando tópicos com uma coerência significativamente melhor.

REFERÊNCIAS

ATHIWARATKUN, B.; WILSON, A.; ANANDKUMAR, A. Probabilistic FastText for multi-sense word embeddings. In: GUREVYCH, I.; MIYAO, Y. (Ed.). **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 1–11. Citado na página 72.

BADARÓ, G. Dos procedimentos: Procediment ordinário, sumário, sumaríssimo e procedimentos especiais. In: ____. **Processo Penal**. [S.l.]: Editora Revista dos Tribunais, 2019. cap. 13. Citado na página 57.

BLEI, D. M.; JORDAN, M. I.; GRIFFITHS, T. L.; TENENBAUM, J. B. Hierarchical topic models and the nested chinese restaurant process. In: **Proceedings of the 16th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 2003. (NIPS'03), p. 17–24. Citado na página 42.

BRISCOE, E.; FELDMAN, J. Conceptual complexity and the bias/variance tradeoff. **Cognition**, Volume 118, Issue 1, 2011. Citado na página 25.

COWANS, P. J. Information retrieval using hierarchical dirichlet processes. In: **Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2004. p. 564–565. Citado na página 42.

DAS, B.; CHAKRABORTY, S. An improved text sentiment classification model using tf-idf and next word negation. **ArXiv**, abs/1806.06407, 2018. Citado nas páginas 47 e 51.

EROSHEVA, E. A.; FIENBERG, S. E. Bayesian mixed membership models for soft clustering and classification. In: WEIHS, C.; GAUL, W. (Ed.). **Classification, the Ubiquitous Challenge**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. p. 11–26. Citado na página 47.

ESCOBAR, M. D.; WEST, M. Bayesian density estimation and inference using mixtures. **Journal of the American Statistical Association**, American Statistical Association, Taylor Francis, Ltd., v. 90, n. 430, p. 577–588, 1995. Citado na página 34.

ESTEVES, L.; IZBICKI, R.; STERN, R. **Inferência Bayesiana**. 2021. Citado nas páginas 21 e 23.

FENSTERMACHER, D.; SCHNEIDER, J.; BAB2MI. **bab2min/tomotopy:**. Zenodo, 2021. Disponível em: <https://doi.org/10.5281/zenodo.4999089>. Citado na página 60.

FERGUSON, T. S. A Bayesian Analysis of Some Nonparametric Problems. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 1, n. 2, p. 209 – 230, 1973. Disponível em: <https://doi.org/10.1214/aos/1176342360>. Citado nas páginas 17, 26 e 81.

FRANK, Q.; GREENBERG, M.; LINDNER, G. **Replicating Hierarchical Dirichlet Processes**. 2020. Disponível em: <https://github.com/morrisgreenberg/hdp-py/blob/master/Paper.pdf>. Acesso em: 01/09/2022. Citado nas páginas 18, 52, 53 e 83.

GAMERMAN, D. **Simulação estocástica via cadeias de Markov**. [S.l.]: ABE, 1996. Citado na página 45.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; RUBIN, D. B. **Bayesian Data Analysis**. 2nd ed.. ed. New York: Chapman and Hall/CRC, 2021. Citado nas páginas 22 e 35.

GIBBONS, J.; CHAKRABORTI, S. **Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded**. Taylor & Francis, 2014. ISBN 9780203911563. Disponível em: <https://books.google.com.br/books?id=kJbVO2G6VicC>. Citado nas páginas 18 e 25.

GILKS, W. R.; WILD, P. Adaptive rejection sampling for gibbs sampling. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, [Wiley, Royal Statistical Society], v. 41, n. 2, p. 337–348, 1992. Citado na página 34.

GRIFFITHS, T.; JORDAN, M.; TENENBAUM, J.; BLEI, D. Hierarchical topic models and the nested chinese restaurant process. In: THRUN, S.; SAUL, L.; SCHöLKOPF, B. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: MIT Press, 2003. v. 16. Citado nas páginas 40 e 47.

HEINRICH, G. Infinite lda: Implementing the hdp with minimum code complexity. **arbylon.net**, 2011. Citado nas páginas 43 e 44.

LIU, Y.; NANDRAM, B. Sampling methods for the concentration parameter and discrete baseline of the dirichlet process. **Statistics in Transition New Series**, v. 23, p. 21–36, 12 2022. Citado nas páginas 18, 34, 35 e 53.

LUHN, H. P. Key word-in-context index for technical literature (kwic index). **American Documentation**, v. 11, n. 4, p. 288–295, 1960. Citado na página 49.

MCTEAR, M.; CALLEJAS, Z.; GRIOL, D. Spoken language understanding. In: ____. **The Conversational Interface: Talking to Smart Devices**. Cham: Springer International Publishing, 2016. p. 161–185. Citado nas páginas 49 e 50.

NEAL, R. M. Markov chain sampling methods for dirichlet process mixture models. **Journal of Computational and Graphical Statistics**, American Statistical Association, Taylor Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America, v. 9, p. 249–265, 2000. Citado nas páginas 17 e 43.

NUNES, M. N. Conceito de jurimetria. In: ____. **Jurimetria: como a estatística pode reinventar o direito**. 2 ed.. ed. Boston, MA: Revista dos Tribunais, 2019. p. 111–142. Citado nas páginas 18 e 56.

PALMER, D. D. Text preprocessing. In: ____. **Handbook of Natural Language Processing**. [S.l.]: Chapman & Hall/CRC, 2010. cap. 2, p. 09–28. Citado nas páginas 18, 47 e 48.

PETITJEAN F., B. W. W. G. e. a. Accurate parameter estimation for bayesian network classifiers using hierarchical dirichlet processes. **Mach Learn**, Volume 107, p. 1303–1331, 2018. Citado na página 54.

POUDYAL, P.; GONçALVES, T.; QUARESMA, P. Using clustering techniques to identify arguments in legal documents. In: **ASAIL@ICAIL**. [S.l.: s.n.], 2019. Citado na página 56.

QADER, W.; AMEEN, M. M.; AHMED, B. An overview of bag of words;importance, implementation, applications, and challenges. In: . [S.l.: s.n.], 2019. p. 200–204. Citado nas páginas 47 e 50.

RINALDI, S.; POZZO, W. D. (H)DPGMM: a hierarchy of Dirichlet process Gaussian mixture models for the inference of the black hole mass function. **Monthly Notices of the Royal Astronomical Society**, v. 509, n. 4, p. 5454–5466, 2021. Citado na página 72.

RIPLEY, B. D. **Stochastic simulation**. New York, NY, USA: John Wiley & Sons, Inc., 1987. ISBN 0-471-81884-4. Citado na página 34.

RöDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2015. (WSDM '15), p. 399–408. ISBN 9781450333177. Citado nas páginas 18, 58 e 59.

RODRIGUES, G. S.; NOTT, D. J.; SISSON, S. A. Likelihood-free approximate gibbs sampling. **Statistics and Computing**, Kluwer Academic Publishers, USA, v. 30, p. 1057–1073, 2020. Citado na página 60.

SETHURAMAN, J. A constructive definition of Dirichlet priors. **Statistica Sinica**, v. 4, p. 639–650, 1994. Citado nas páginas 17, 26, 29 e 30.

STEPHENS, M.; SMITH, N. J.; DONNELLY, P. A new statistical method for haplotype reconstruction from population data. **The American Journal of Human Genetics**, p. 978–989, 2001. ISSN 0002-9297. Citado na página 42.

SUCAR, L. E. **Probabilistic Graphical Models: Principles and Applications**. 1st ed.. ed. London: Springer London, 2015. Citado na página 43.

TEH, Y. W. Dirichlet process. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 280–287. Citado nas páginas 17, 25, 28, 30, 32 e 33.

TEH, Y. W.; JORDAN, M. I. Hierarchical bayesian nonparametric models with applications. In: ____. **Bayesian Nonparametrics**. [S.l.]: Cambridge University Press, 2009. (Cambridge Series in Statistical and Probabilistic Mathematics), p. 158–207. Citado nas páginas 17, 18, 35, 36, 37, 39, 40, 41, 44 e 52.

WEINBERGER, K.; DASGUPTA, A.; ATTENBERG, J.; LANGFORD, J.; SMOLA, A. **Feature Hashing for Large Scale Multitask Learning**. 2010. Citado na página 50.

ZHANG, F.; GAO, W.; FANG, Y.; ZHANG, B. Enhancing short text topic modeling with fasttext embeddings. In: **2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)**. [S.l.: s.n.], 2020. p. 255–259. Citado na página 72.

ANEXO

A

DISTRIBUIÇÕES DE PROBABILIDADE

Distribuição Binomial

Seja X uma variável aleatória, dizemos que X segue uma distribuição binomial com parâmetros $n \in \mathbb{N}$ e $\theta \in [0,1]$, se, para todo X assumindo valores $\{0,1,2,\ldots,n\}$, temos

$$f_X(x|n,\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \mathbb{1}_{\{x=1,2,\ldots,n\}},$$

onde $\mathbb{1}$ é uma função indicadora, assumindo 1 se sua condição é cumprida e 0 caso contrário. Em notação, $X \sim Bin(n,\theta)$.

Distribuição Beta

Seja X uma v.a. que assume valores no intervalo $[0,1]$. Considere os parâmetros $\alpha, \beta > 0$, dizemos que X tem distribuição Beta se

$$f_X(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}_{\{x\in[0,1]\}}.$$

Em notação, $X \sim Beta(\alpha,\beta)$.

Distribuição Mutinomial

Considere um experimento onde $X_1, X_2, \ldots, X_k$ denotam a ocorrência de $k \in \mathbb{N}$ eventos de interesse com probabilidade $\theta_1, \theta_2, \ldots, \theta_k$, sujeitos a $\sum_{j=1}^{k} \theta = 1$. Considere ainda $n \in \mathbb{N}$ ensaios independentes desse experimento. Dizemos que o vetor aleatório $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ tem distribuição multinomial com parâmetros $n, \theta = (\theta_1, \theta_2, \ldots, \theta_k)$ se

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_k|n,\theta) = \frac{\Gamma(n+1)}{\prod_{j=1}^{k} \Gamma(x_j+1)} \prod_{j=1}^{k} \theta_j^{x_j} \mathbb{1}_{\{\sum_{j=1}^{k} x_j = n\}},$$

onde Γ é a função gama, tal que para todo valor inteiro positivo n definidos $\Gamma(n) = (n-1)!$. Em notação, $\mathbf{X} \sim Mult(n,\theta)$.

Distribuição Dirichlet

Seja $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ um vetor aleatório condicionado a $\sum_{j=1}^{k} x_j = 1$. Considere um vetor de parâmetros $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k), \forall \alpha_j \geq 0$. Dizemos que $\mathbf{X}$ tem distribuição de Dirichlet se

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_k | \alpha) = \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} x_j^{\alpha_j - 1} \mathbb{1}_{\{\sum_{j=1}^{k} x_j = 1\}}.$$

Em notação $\mathbf{X} \sim Dir(\alpha)$.

ANEXO

B

DEMONSTRAÇÕES DE RESULTADOS

Proposição 4. Tomando $B_0 = A_j$ e $B_1 = A_1 \cup A_2 \cup \cdots \cup A_{j-1} \cup A_{j+1} \cup \cdots \cup A_k$, $\forall \{A_j\}_{j=1}^k \in \mathscr{A}$ e $k \in \mathbb{N}^*$. Pela definição de um Processo de Dirichlet temos que

$$(B_0, B_1) \sim Dir(\alpha_0 H(B_0), \alpha_1 H(B_1))$$

mas também

$$(B_0, B_1) \sim Beta(\alpha_0 H(B_0), \alpha_1 H(B_1)).$$

Demonstração. Dado a definição de B_0 e B_1, onde B_1 é complementar a B_0, e se $(B_0, B_1) \sim Dir(\alpha_0 H(B_0), \alpha_1 H(B_1))$, considere que $\theta_0, \theta_1 \in \Theta$ são observações de $H(B_0), H(B_1)$, respectivamente. Então

$$\begin{aligned} f_{H(B_0),H(B_1)}(\theta_0, \theta_1 | \alpha, H) &= \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \theta_0^{\alpha_0 - 1} \theta_1^{\alpha_1 - 1} \mathbb{1}_{\{\sum_{j=0}^1 \theta_j = 1\}} \\ &= \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)} \theta_0^{\alpha_0 - 1} (1 - \theta_0)^{\alpha_1 - 1} \mathbb{1}_{\{\sum_{j=0}^1 \theta_j = 1\}}, \end{aligned}$$

onde $\mathbb{1}$ é a função indicadora que define o domínio da distribuição.

Portanto, $(B_0, B_1) \sim Beta(\alpha_0 H(B_0), \alpha_1 H(B_1))$. □

Lema 5. Considerando $X_1, X_2, \ldots, X_k$ variáveis aleatórias independentes, tais que $X_j \sim Gama(\alpha_j), \forall j = 1, 2, \ldots, k$ e $k \in \mathbb{N}^*$, então

$$\left(\frac{X_1}{\sum_{j=1}^k X_j}, \ldots, \frac{X_k}{\sum_{j=1}^k X_j} \right) \sim Dir(\alpha_1, \ldots, \alpha_k)$$

Demonstração. Tome $Z = \sum_{j=1}^k X_j$, sabemos que $Z \sim Gama(\sum_{j=1}^k \alpha_j)$ pela propriedade da soma de variáveis aleatórias independentes gama.

Considerando que $Y_i = \frac{X_i}{Z}$, para $i = 1, 2, \ldots, k-1$, vamos tomar a seguinte relação

$$X_i = Y_i Z, \forall i$$
$$X_k = Z\left(1 - \sum_{i=1}^{k-1} Y_i\right).$$

Pelo métodos Jacobiano temos que

$$|J| = \begin{vmatrix} Z & 0 & 0 & \ldots & Y_1 \\ 0 & Z & 0 & \ldots & Y_2 \\ 0 & 0 & Z & \ldots & Y_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -Z & -Z & -Z & \ldots & (1 - \sum_{i=1}^{k-1} Y_i) \end{vmatrix} = Z^{k-1}$$

tal que

$$\begin{aligned} f_{\mathbf{Y},Z}(y_1, y_2, \ldots, y_{k-1}, z) &= f_{\mathbf{X}}(x_1, \ldots, x_k)|J| \\ &= \prod_{i=1}^{k-1} f_{X_i}(y_i z) f_{X_k}\left(z\left(1 - \sum_{i=1}^{k-1} y_i\right)\right)|J| \\ &= \prod_{j=1}^{k} \frac{1}{\Gamma(\alpha_j)} \prod_{i=1}^{k-1} y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{k-1} y_i\right)^{\alpha_k - 1} z^{\left(\sum_{j=1}^{k} \alpha_j - 1\right)} e^{-z}. \end{aligned}$$

Por fim, pela densidade marginal temos

$$\begin{aligned} f_{\mathbf{Y}}(y_1, y_2, \ldots, y_{k-1}) &= \int_0^\infty f_{\mathbf{Y}|Z}(y_1, y_2, \ldots, y_{k-1}, z) dz \\ &= \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{i=1}^{k-1} y_i^{(\alpha_i - 1)} \left(1 - \sum_{i=1}^{k-1} y_i\right)^{(\alpha_k - 1)}, \end{aligned} \tag{B.1}$$

ou seja, B.1 tem distribuição de Dirichlet uma vez que $\sum_{i=1}^{k-1} y_i + \left(1 - \sum_{i=1}^{k-1} y_i\right) = 1$. Reescrevendo $\left(1 - \sum_{i=1}^{k-1} Y_i\right) = \frac{X_k}{Z} = Y_k$, temos então

$$f_{\mathbf{Y}}(y_1, y_2, \ldots, y_k) = \frac{\Gamma(\sum_{j=1}^{k} \alpha_j)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \prod_{j=1}^{k} y_j^{(\alpha_j - 1)} \mathbb{1}_{\{\sum_{j=0}^{1} y_j = 1\}}$$

segue uma Distribuição Dirichlet com parâmetros $\{\alpha_j\}_{j=1}^k$. □

Proposição 5. Sejam G um Processo de Dirichlet em $(\Theta, \mathscr{A}, H)$, com parâmetro α, e $\forall A \in \mathscr{A}$. Se $H(A) = 0$, então $G(A) = 0$ com probabilidade 1. Por sua vez, se $H(A) > 0$, então $G(A) > 0$ com probabilidade 1. Além disso,

$$\mathbb{E}[G(A)] = H(A)$$

Demonstração. Tomando a Proposição 1, sabendo que $(A, A^c) \sim Beta(\alpha H(A), \alpha H(A^c))$, temos que

$$\mathbb{E}[G(A)] = \frac{\alpha H(A)}{\alpha(H(A) + H(A^c))} = \frac{H(A)}{H(\Theta)} = H(A).$$

□

Proposição 6. Sobre um Processo de Dirichlet, G, definido no espaço de probabilidades $(\Theta, \mathscr{A}, H)$, com parâmetro α, e $\forall A \in \mathscr{A}$, temos que,

$$Var[G(A)] = \frac{H(A)(1-H(A))}{(\alpha+1)}$$

Demonstração.

$$\begin{aligned} Var[G(A)] &= \frac{\mathbb{E}[G(A)]\,(1-\mathbb{E}[G(A)])}{\alpha+1} \\ &= \frac{H(A)\,(1-H(A))}{\alpha+1}, \qquad \text{pela Preposição 2.} \end{aligned}$$

□

Lema 6. Se $(Y_1,\ldots,Y_k) \sim Dir(\alpha_1,\ldots,\alpha_k)$, então

$$(Y_1,\ldots,Y_{k-2},Y_{k-1}+Y_k) \sim Dir(\alpha_1,\ldots,\alpha_{k-2},\alpha_{k-1}+\alpha_k),$$

para $k \in \mathbb{N}^*$.

Demonstração. Tomamos $(Y_1,\ldots,Y_k)$ variáveis aleatórias, tais que $Y_j = \frac{X_j}{\sum_{j=1}^k X_j}$, sendo Y_j, X_j definidos segundo o Lema 2, para $j = 1,\ldots,k$. Assim $(Y_1,\ldots,Y_{k-2},Y_{k-1}+Y_k)$ é tal que

$$Y' = Y_{k-1}+Y_k = \frac{X_{k-1}+X_k}{\sum_{j=1}^k X_j}.$$

Sabemos ainda que

$$X_{k-1}+X_k \sim Gama(\alpha_{k-1}+\alpha_k),$$

isto é, ainda mantemos a propriedade do Lema 2, mesmo com a soma das variáveis. Portando, ainda pelo Lema 2, temos que

$$(Y_1,\ldots,Y_{k-2},Y_{k-1}+Y') = \left(\frac{X_1}{\sum_{j=1}^k X_j},\ldots,\frac{X_{k-2}}{\sum_{j=1}^k X_j},\frac{X_{k-1}+X_k}{\sum_{j=1}^k X_j}\right) \sim Dir(\alpha_1,\ldots,\alpha_{k-2},\alpha'),$$

para $\alpha' = \alpha_{k-1}+\alpha_k$. □

Lema 7. Se $(Y_1,\ldots,Y_k) \sim Dir(\alpha_1,\ldots,\alpha_k)$ e π é uma permutação de $\{1,2,\ldots,k\}$, então

$$(Y_{\pi(1)},\ldots,Y_{\pi(k)}) \sim Dir\left(\alpha_{\pi(1)},\ldots,\alpha_{\pi(k)}\right),$$

para $k \in \mathbb{N}^*$.

Demonstração. Ver Ferguson (1973). □

ANEXO

C

ATUALIZAÇÃO DO AMOSTRADOR DE GIBBS SOBRE A DISTRIBUIÇÃO PREDITIVA

Baseados em Frank, Greenberg e Lindner (2020), vamos considerar

$$f(x_{ij}|\phi_k) \propto \phi_{kv} \sim Multi(n, \phi_k)$$
$$h(\phi_k) \sim Dir(\gamma) = \frac{1}{B(\gamma)} \prod_v \phi_v^{\gamma-1} \qquad B(\gamma) = \frac{\prod_{j=1}^k \Gamma(\gamma_j)}{\Gamma(\sum_{j=1}^k \gamma_j)}$$

onde $\phi_k = \theta_k^{**}$, por simplificação; $v = x_{ji}$ indica diretamente as observações $x_j i$ em todo seu domínio, tal que $\#v = N$; ϕ_{kv} é o núcleo da função de verossimilhança de v em função de ϕ_k; finalmente ϕ_v são os parâmetros da distribuição conjugada $h(\cdot)$ sobre todo o domínio das observações (na aplicação de tópicos textuais, sobre todo o vocabulário).

Usando a equação 2.14 como atualização dos sorteios no amostrador de *Gibbs* e substituindo os elementos anteriores na mesma, temos

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\int f_{\phi_k}(x_{ji}) \left[\prod_{j'i' \neq ji} f_{\phi_k}(x_{j'i'}) h(\phi_k)\right] d\phi_k}{\int \left[\prod_{j'i' \neq ji} f_{\phi_k}(x_{j'i'}) h(\phi_k)\right] d\phi_k}$$
$$= \frac{\int f_{\phi_k}(x_{ji}) \left[\prod_{j'i' \neq ji} \phi_{kv} h(\phi_k)\right] d\phi_k}{\int \left[\prod_{j'i' \neq ji} \phi_{kv} h(\phi_k)\right] d\phi_k}$$

Vamos reescrever a relação $j'i' \neq ji$ como w de tal modo que w represente o conjunto de todas

as observações $x_{j'i'}$, exceto x_{ji}. Então,

$$
\begin{aligned}
f_k^{-x_{ji}}\left(x_{ji}\right) &\propto \frac{\int \phi_{kv}\left[\prod_w \phi_{kw}^{n_{kw}^{-ji}} \prod_w \phi_{kw}^{\gamma-1}\right] d\phi_k}{\int \left[\prod_w \phi_{kw}^{n_{kw}^{-ji}} \prod_w \phi_{kw}^{\gamma-1}\right] d\phi_k} \\
&= \frac{\int \phi_{kv}\phi_{kv}^{n_{kv}^{-ji}} \phi_{kv}^{\gamma-1}\left[\prod_{w\neq v} \phi_{kw}^{n_{kw}^{-ji}} \prod_{w\neq v} \phi_{kw}^{\gamma-1}\right] d\phi_k}{\int \left[\prod_w \phi_{kw}^{n_{kw}^{-ji}} \prod_w \phi_{kw}^{\gamma-1}\right] d\phi_k} \\
&= \left\{\int \phi_{kv}^{n_{kv}^{-ji}+\gamma}\left[\prod_{w\neq v} \phi_{kw}^{n_{kw}^{-ji}+\gamma-1}\right] d\phi_k\right\} \frac{1}{\int \left[\prod_w \phi_{kw}^{n_{kw}^{-ji}+\gamma-1}\right] d\phi_k}
\end{aligned}
$$

Reescrevendo as integrais em relação a funções Gamma, temos para o primeiro termo do produto

$$
\begin{aligned}
f_k^{-x_{ji}}\left(x_{ji}\right) &= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\prod_{w\neq v}\Gamma(\gamma+n_{kw}^{-ji})}{\Gamma\left(\sum_{w\neq v}\left[\gamma+n_{kw}^{-ji}\right]+\left[\gamma+n_{kv}^{-ji}+1\right]\right)}\right\} \frac{1}{\int \prod_w \phi_{kw}^{n_{kw}^{-ji}+\gamma-1} d\phi_k} \\
&= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\prod_{w\neq v}\Gamma(\gamma+n_{kw}^{-ji})}{\Gamma\left(\sum_{w}\left[\gamma+n_{kw}^{-ji}\right]+1\right)}\right\} \frac{1}{\int \prod_w \phi_{kw}^{n_{kw}^{-ji}+\gamma-1} d\phi_k} \\
&= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\prod_{w\neq v}\Gamma(\gamma+n_{kw}^{-ji})}{\Gamma\left(N\gamma+n_{k}^{-ji}+1\right)}\right\} \frac{1}{\int \prod_w \phi_{kw}^{n_{kw}^{-ji}+\gamma-1} d\phi_k}
\end{aligned}
$$

Analogamente, para o segundo termo do produto

$$
\begin{aligned}
f_k^{-x_{ji}}\left(x_{ji}\right) &= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\prod_{w\neq v}\Gamma(\gamma+n_{kw}^{-ji})}{\Gamma\left(N\gamma+n_{k}^{-ji}+1\right)}\right\} \frac{\Gamma\left(\sum_w\left[\gamma+n_{kw}^{-ji}\right]\right)}{\prod_w \Gamma\left(\gamma+n_{kw}^{-ji}\right)} \\
f_k^{-x_{ji}}\left(x_{ji}\right) &= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\prod_{w\neq v}\Gamma(\gamma+n_{kw}^{-ji})}{\Gamma\left(N\gamma+n_{k}^{-ji}+1\right)}\right\} \frac{\Gamma\left(N\gamma+n_{k}^{-ji}\right)}{\prod_w \Gamma\left(\gamma+n_{kw}^{-ji}\right)}
\end{aligned}
$$

Finalmente, rearranjando os termos entre o produto, obtemos

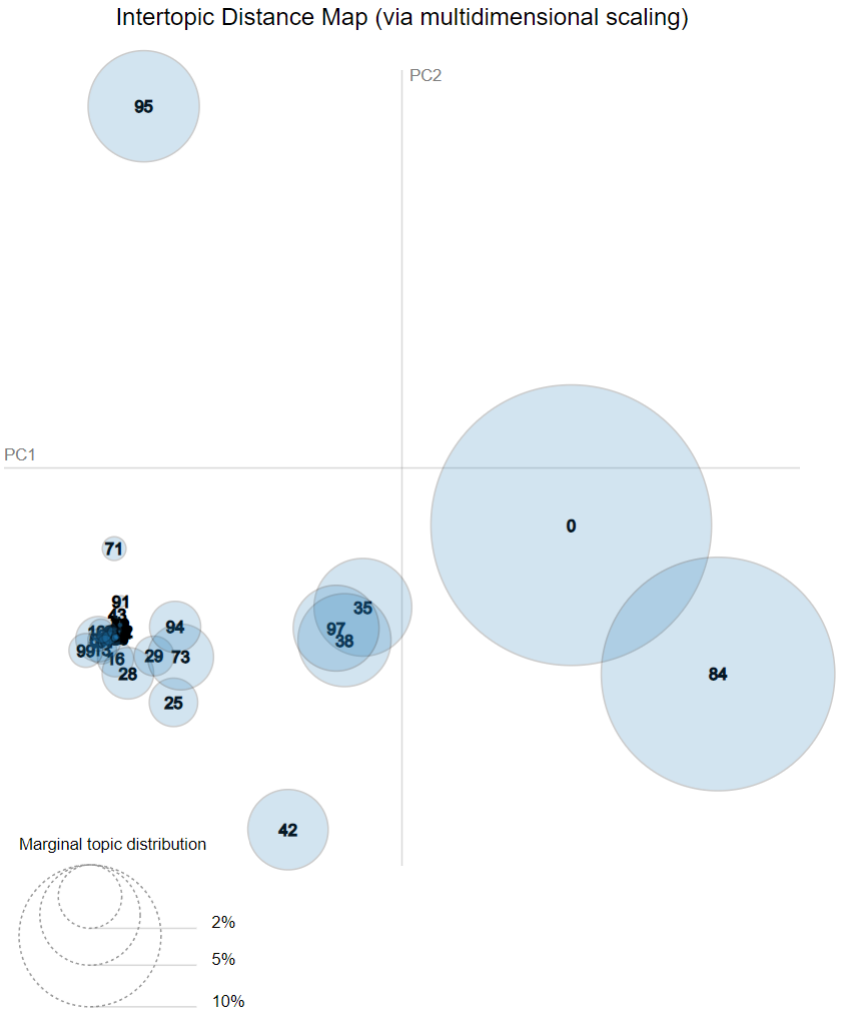
$$
\begin{aligned}
f_k^{-x_{ji}}\left(x_{ji}\right) &= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\Gamma\left(N\gamma+n_{k}^{-ji}\right)}{\Gamma\left(N\gamma+n_{k}^{-ji}+1\right)}\right\} \frac{\prod_{w\neq v}\Gamma(\gamma+n_{kw}^{-ji})}{\prod_w \Gamma\left(\gamma+n_{kw}^{-ji}\right)} \qquad \text{(C.1)} \\
&= \left\{\frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\Gamma\left(N\gamma+n_{k}^{-ji}\right)}{\Gamma\left(N\gamma+n_{k}^{-ji}+1\right)}\right\} \frac{1}{\Gamma\left(\gamma+n_{kv}^{-ji}\right)} \\
&= \frac{\Gamma(\gamma+n_{kv}^{-ji}+1)\Gamma\left(N\gamma+n_{k}^{-ji}\right)}{\Gamma\left(\gamma+n_{kv}^{-ji}\right)\Gamma\left(N\gamma+n_{k}^{-ji}+1\right)}.
\end{aligned}
$$

ANEXO

D

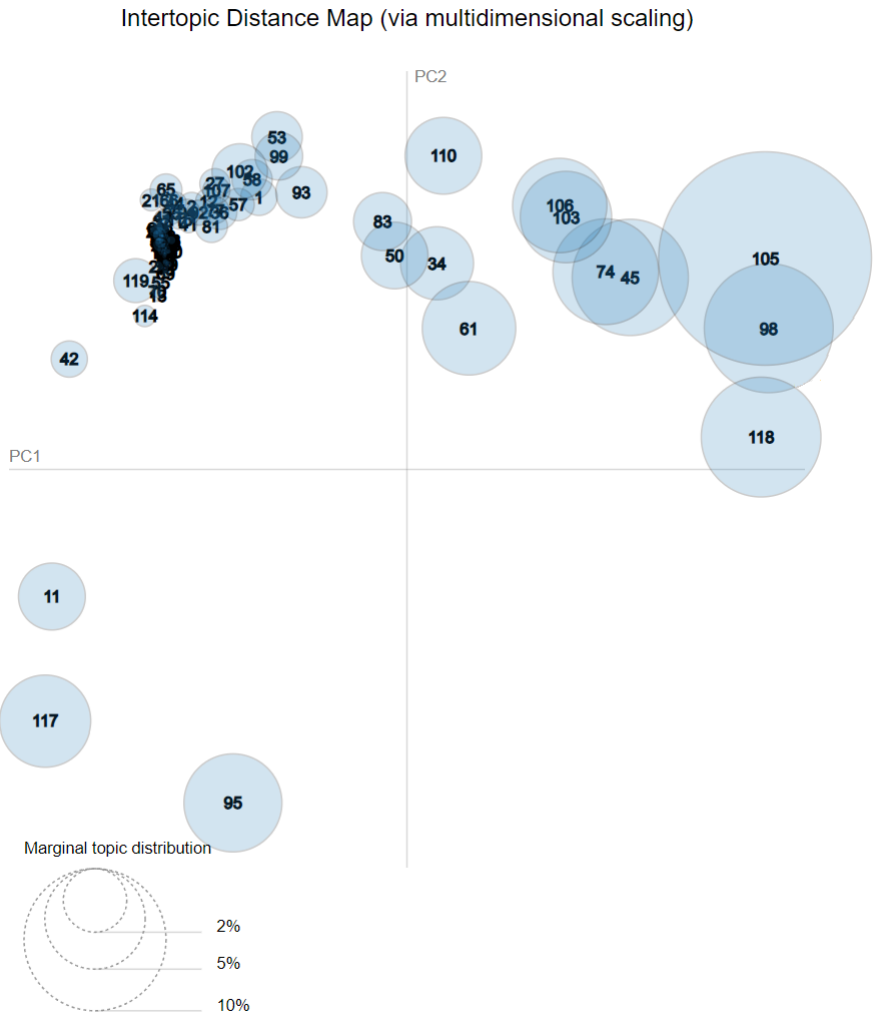
GRÁFICOS DE DISPERSÃO COMPLETOS

Figura 13 – Distribuição dos todos os tópicos de processos sobre procedimento sumário.



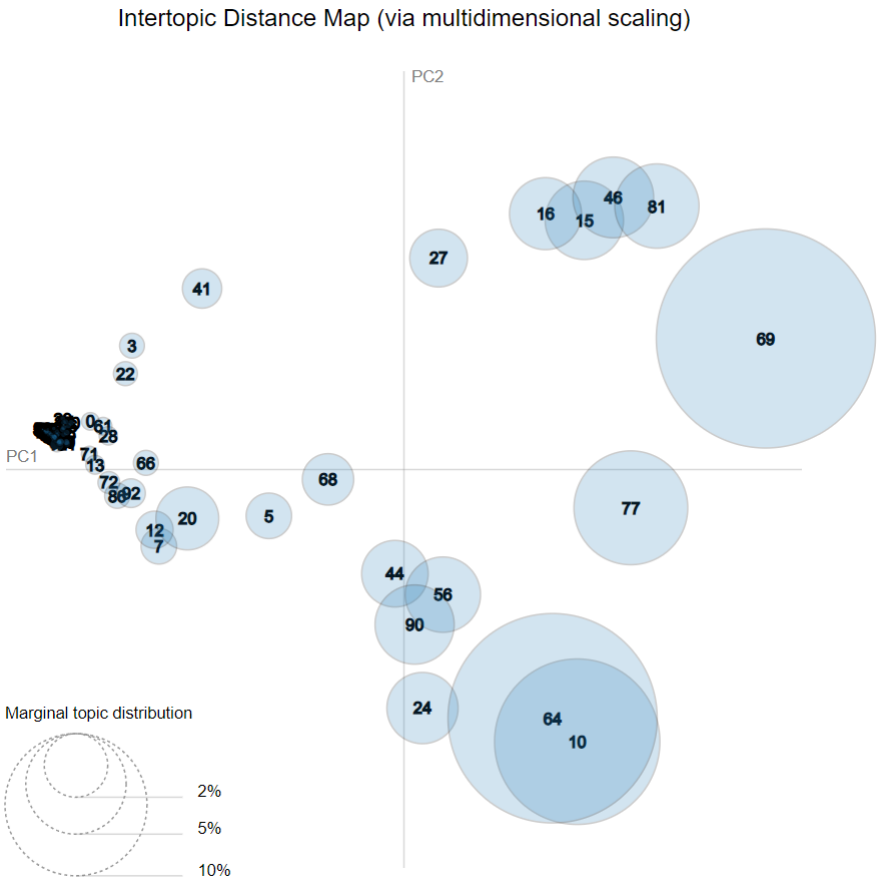
Fonte: Elaborada pelo autor.

Figura 14 – Distribuição dos todos os tópicos de processos sobre procedimento de juizado especial civil.



Fonte: Elaborada pelo autor.

Figura 15 – Distribuição dos todos os tópicos de processos sobre procedimento ordinário.



Fonte: Elaborada pelo autor.

ANEXO

E

TOKENS REMOVIDOS DO PROCESSO DE MODELAGEM DE TÓPICOS POR HDP

Tabela 4 – Lista dos Top 50 *tokens* desconsiderados na modelagem de tópicos por cada estrato considerado.

Estrato	**Palavras Excluídas**
Procedimento Especial Civil	nao, valor, autora, parte, autos, lei_9099, autor, termos, pagamento, lei_cpc, partes, dano, julgo, credito, sentenca, recurso, julgado, dias, prazo, moral, indenizacao, danos, caso, apos, prova, pedido, assim, transito, acao, morais, inicial, deve, forma, bem, contrato, desde, processo, juros, custas, servico, justica, condenacao, preparo, conforme, nome, data, desta, favor, dispensado, honorarios
Procedimento Ordinário	nao, juros, contrato, autor, valor, autora, acao, pagamento, direito, lei_cpc, credito, banco, recurso, dano, parte, danos, pedido, cobranca, autos, caso, indenizacao, sao, moral, prova, taxa, partes, termos, nome, assim, especial, morais, capitalizacao, bem, relacao, forma, tribunal, sobre, inicial, justica, deve, sendo, julgamento, documentos, servicos, contratos, fato, pois, lei_cdc, conta, debito
Procedimento Sumário	nao, juros, autor, contrato, valor, autora, acao, lei_cpc, autos, pagamento, direito, credito, parte, banco, danos, pedido, dano, indenizacao, termos, cobranca, caso, julgo, nome, partes, moral, prova, morais, sao, recurso, relacao, assim, inicial, forma, bem, documentos, deve, inciso, debito, sobre, taxa, pois, servicos, sendo, tutela, tribunal, fato, ainda, justica, julgamento, lei_cdc

Fonte: Dados da pesquisa.


ICMC
USP
SÃO CARLOS