

Modelagem Bayesiana e Análise de Dados classificados

Stephen Richard Johnson

Tese apresentada para o grau de
Doutor em Filosofia



Escola de Matemática, Estatística e Física
Universidade de
NewcastleNewcastle upon Tyne
Reino Unido

Janeiro de 2019

Agradecimentos

Gostaria de agradecer aos meus orientadores, Daniel Henderson e Richard Boys, sem os quais esta tese não teria sido possível. Serei eternamente grato não apenas por seus conselhos inestimáveis, mas também por seu apoio e paciência duradouros, principalmente em tempos de turbulência. Além disso, devo agradecer à minha mãe e ao meu pai, junto com o resto da minha família, por todo o amor, bondade e incentivo ao longo de meus estudos de pós-graduação. Finalmente, a todos os meus amigos e colegas da Escola de Matemática, Estatística e Física que tornaram meu tempo na Universidade de Newcastle tão agradável.

Abstrair

Os dados classificados são centrais para muitas aplicações na ciência e nas ciências sociais e surgem quando os rankers (indivíduos) usam algum critério para ordenar um conjunto de entidades. Tais classificações são, portanto, equivalentes a permutações dos elementos de um conjunto. A maioria dos modelos de dados classificados baseia-se num forte pressuposto de homogeneidade, como todos os classificadores partilham a mesma visão sobre as preferências das entidades. O objetivo desta tese é desenvolver uma classe mais rica de modelos que possam revelar qualquer estrutura de subgrupo plausível dentro dos dados, tanto para classificadores quanto para entidades.

Começamos examinando o modelo de Plackett-Luce, uma extensão do modelo de Bradley-Terry para comparações pareadas. Primeiro, esse modelo é estendido para atender quando os classificadores não relatam uma classificação completa de todas as entidades. Por exemplo, eles podem relatar apenas suas cinco principais entidades classificadas depois de ver algumas ou todas as entidades. Outra questão é que a maioria dos trabalhos nessa área pressupõe que todos os classificadores sejam igualmente informados sobre as entidades que estão classificando. Muitas vezes, essa suposição será questionável e, portanto, desenvolvemos um modelo que permite que os classificadores tenham confiabilidade diferente. Este modelo, o modelo ponderado de Plackett-Luce, permite essa heterogeneidade por meio de um novo modelo de mistura de dois componentes definido pelo aumento do modelo de Plackett-Luce.

A ideia de que os classificadores podem ser heterogêneos em suas crenças sobre entidades não é nova. No entanto, pode haver grupos de classificadores com cada grupo compartilhando a mesma visão sobre entidades. Geralmente, o número de tais grupos não será conhecido e, portanto, investigamos a possibilidade de tal estrutura de grupo usando uma mistura de processo de Dirichlet de modelos ponderados de Plackett-Luce. Também pode ser útil avaliar se algumas entidades são passíveis de troca, ou seja, se também há agrupamento de entidades dentro de cada grupo de rankers, uma questão que tem recebido pouca atenção na literatura. Estendemos o modelo ainda mais para explorar o agrupamento de classificadores e entidades, adaptando o processo Nested Dirichlet. O modelo resultante é uma mistura de processo Weighted Adapted Nested Dirichlet (WAND) dos modelos Plackett-Luce. A inferência posterior é conduzida por meio de um esquema de amostragem de Gibbs simples e eficiente. A riqueza de informações na distribuição posterior permite a inferência sobre muitos aspectos da estrutura de agrupamento tanto entre grupos de rankers quanto entre grupos de entidades (dentro de grupos de rankers), em contraste com muitas outras análises (bayesianas). A metodologia é ilustrada usando vários estudos de simulação e exemplos de dados reais.

Finalmente, relaxamos a suposição de um processo de classificação conhecido subjacente a esses modelos, observando o modelo Extended Plackett-Luce recentemente desenvolvido. Esse modelo permite a inferência para a ordem em que um conjunto homogêneo de classificadores atribui entidades a classificações. A análise deste modelo é desafiadora, mas descobrimos que o uso de métodos de Monte Carlo (MC3) da cadeia de Markov acoplada à Metrópole pode fornecer mistura adequada no espaço de alta dimensão de todas as permutações possíveis quando o número de entidades não é pequeno.

Conteúdo

1 Introdução	1
1.1Introdução . . .	1
1.1.1Objetivos da tese . .	4
1.1.2Esboço da tese . .	5
1.2Inferência bayesiana . . .	7
1.2.1Teorema de Bayes	7
1.3Cadeia de Markov Monte Carlo	8
1.3.1O algoritmo Metropolis-Hastings . . .	8
1.3.2Ajustando algoritmos de Metropolis-Hastings	11
1.3.3O amostrador Gibbs	13
1.3.4Bloquear atualizações	14
1.3.5Convergência	15
1.3.6Análise de amostras posteriores	16
1.4Aumento de dados	16
2 Análise de dados homogêneos classificados	19
2.1Introdução	19
2.2O modelo de Plackett-Luce	20
2.2.1Classificações Top-M	23
2.2.2Problemas de identificabilidade	24
2.2.3Reescalonamento	24
2.2.4Laços	24
2.2.5Simulação de dados do modelo de Plackett-Luce modificado	26

2.3 Inferência bayesiana	27
2.3.1Especificação prévia e variáveis latentes	27
2.3.2Distribuições condicionais completas	29
2.3.3MCMC - Amostragem de Gibbs via variáveis latentes	31
2.4Estudo de simulação	32
2.4.1Análise posterior	33
2.5O modelo Plackett-Luce ponderado	38
2.5.1Simulando dados do modelo Plackett-Luce ponderado	40
2.6Inferência bayesiana	40
2.6.1Especificação prévia e variáveis latentes	41
2.6.2Distribuições condicionais completas	42
2.6.3MCMC - Amostragem de Gibbs via variáveis latentes	43
2.7Estudo de simulação	45
2.7.1Análise posterior	45
2.8Resumo	50
3 Análise de dados heterogêneos classificados	51
3.1Introdução	51
3.2Modelos de misturas finitas	51
3.3O processo Dirichlet	54
3.3.1Representações alternativas do processo de Dirichlet	57
3.3.2Gerando uma realização de um processo de Dirichlet	59
3.3.3Geração de parâmetros de modelo consistentes com um processo de Dirichlet anterior a	61
3.3.4Processo de Dirichlet generalizado	63
3.4Modelos de mistura de processos de Dirichlet	64
3.4.1Inferência bayesiana para DPMM	65
3.4.2Ter em conta a incerteza quanto ao parâmetro de concentração	66
3.5Descobrir a heterogeneidade entre classificadores	67
3.5.1O modelo	68

3.5.2 Simulando dados da mistura do processo de Dirichlet de modelos WeightedPlackett-Luce . . . 68	
3.5.3Especificação prévia e variáveis latentes . . . 70	
3.5.4Distribuições condicionais completas . . 71	
3.5.5MCMC usando o algoritmo de Neal 8 . . . 74	
3.5.6Estudo de simulação – revisitando o conjunto de dados 2 . . . 75	
3.6Descobrir subgrupos de entidades dentro de um grupo de classificação . . . 81	
3.6.1O modelo . . . 82	
3.6.2Simulando dados do modelo de Plackett-Luce ponderado com agrupamento de entidades . . . 83	
3.6.3Especificação de variável prévia e latente . . 84	
3.6.4Distribuições condicionais completas . . 86	
3.6.5MCMC usando o Algoritmo de Neal 8 . . . 88	
3.6.6Estudo de simulação – revisitando o conjunto de dados 1 . . . 89	
3.7Resumo . . . 95	
 4 A VARINHA Bayesianá	97
4.1Introdução . . . 97	
4.2Agrupamento bidirecional . . 97	
4.3O Processo de Dirichlet Aninhado Adaptado (ANDP) antes . . . 99	
4.3.1Geração de amostras anteriores (representação de quebra de bastão) . . . 101	
4.3.2Explorar a ANDP antes . . . 103	
4.4O modelo . . . 104	
4.5Uma abordagem de amostragem condicional . . . 105	
4.5.1Simulando dados do modelo WAND . . . 105	
4.5.2Especificação prévia e variáveis latentes . . . 107	
4.5.3 Distribuições condicionais completas . . 108	
4.5.4Movimentos de comutação de etiquetas . . . 114	
4.5.5Alocações de classificadores . . . 115	
4.5.6Dotações de entidades . . . 118	
4.5.7MCMC – um amostrador condicional . . . 120	

4.5.8	Um breve resumo . . .	122
4.6	Uma abordagem de amostragem marginal . . .	122
4.6.1	Amostragem marginal da ANDP anterior . . .	124
4.6.2	Simulação de dados do modelo WAND . . .	125
4.6.3	Especificação prévia e variáveis latentes . . .	126
4.6.4	Distribuições condicionais completas . .	127
4.6.5	MCMC – um amostrador marginal . .	129
4.7	Estudos de simulação . . .	132
4.7.1	Estudo 1 . . .	132
4.7.2	Estudo 2 . . .	140
4.8	Resumo . . .	144
5	Análises de dados reais	147
5.1	Conjunto de dados da Roskam . . .	147
5.1.1	Análise prévia de sensibilidade . . .	150
5.2	Estudo da NBA . . .	153
5.2.1	Análise prévia de sensibilidade . . .	158
5.3	Resumo . . .	160
6	O modelo Extended Plackett-Luce	161
6.1	Introdução . . .	161
6.2	O modelo Plackett-Luce estendido . . .	162
6.2.1	Simulação de dados do modelo Extended Plackett-Luce . . .	165
6.2.2	Identificabilidade do processo de classificação . . .	166
6.3	Informações de verossimilhança sobre a ordem de escolha . . .	170
6.3.1	MM Algoritmo . . .	171
6.3.2	Estudo de simulação . . .	172
6.4	Inferência – uma abordagem bayesiana	174
6.4.1	Especificação prévia e variáveis latentes . . .	175
6.4.2	Distribuições condicionais completas para λ, Z . . .	176
6.4.3	Distribuição condicional completa para σ . .	177

6.4.4 Propostas de Metropolis-Hastings para σ	177
6.4.5 Considerações adicionais	183
6.4.6A Algoritmo Metropolis-within-Gibbs para o modelo EPL	185
6.4.7 Estudo de simulação – Metrópole dentro de Gibbs	186
6.4.8 Metropolis-Hastings propostas para λ	189
6.4.9 Um algoritmo Metropolis-Hastings para o modelo EPL	190
6.4.10 Estudo de simulação – Metropolis-Hastings	191
6.5 Metrópole acoplada cadeia de Markov Monte Carlo	193
6.5.1 A vantagem de visar densidades temperadas	194
6.5.2 Têmpera paralela	196
6.5.3 Esboço geral do algoritmo	197
6.5.4 Ajustando um esquema de amostragem MC3	198
6.5.5 Metrópole paralela acoplada cadeia de Markov Monte Carlo	199
6.6 Inferência – uma abordagem bayesiana (revisitada)	201
6.6.1 Um algoritmo pMC3 para o modelo EPL	201
6.6.2 Estudo de simulação	202
6.7 Resumo	205
7 Conclusão	207
7.1 Trabalhos futuros	209
Uma miscelânea	213
A.1 Derivação do FCD para o parâmetro de concentração DP α	213
B Conjuntos de dados	217

Lista de Figuras

2.1Gráficos de rastreamento da probabilidade de dados completos do log para os conjuntos de dados 1 e 2 da esquerda para a direita, respectivamente. 34

2.2Gráficos de caixa que resumem as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{20} = 1$. As densidades em cada caso são mostradas em branco e vermelho para os conjuntos de dados 1 e 2, respectivamente. As cruzes azuis representam os valores verdadeiros a partir dos quais esses dados foram simulados (escala logarítmica). . . 35

2.3Gráficos de rastreamento da probabilidade de dados logarítmicos completos para as Análises 1 e 2 ($\pi = 0,5, 0,8$) da esquerda para a direita, respectivamente. . . 46

2.4 $\Pr(w_i = 1|D)$ – Probabilidade posterior de que a classificação i seja informativa em cada análise (Análise 1: $\pi = 0,5$, Análise 2: $\pi = 0,8$). As classificações que são permutações aleatórias (41–50) são mostradas em vermelho. Os números mostrados denotam a distância Kendall-tau entre cada classificação e ‘x. 47

2.5Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{20} = 1$. Os boxplots em cada caso são mostrados em branco e vermelho para as análises 1 e 2 ($\pi = 0,5, 0,8$), respectivamente. As cruzes azuis representam os valores verdadeiros, λ_k , a partir dos quais esses dados foram simulados. 47

2.6Gráficos de caixa que resumem as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{20} = 1$. Os boxplots em cada caso são mostrados em branco para Analysis1 em nosso modelo Weighted Plackett-Luce e em vermelho para a análise de Dataset 1 sob o modelo padrão de Plackett-Luce. As cruzes azuis representam os valores verdadeiros, λ_k , a partir dos quais esses dados foram simulados. 48

3.1Múltiplas realizações de um processo de Dirichlet com $G_0 = N(0, 1)$ e $\alpha = 1, 5, 10, 50, 100$ de cima para baixo, respectivamente. 56

3.2CDF empírico de múltiplas realizações de um processo de Dirichlet com $G_0 = N(0, 1)$ e $\alpha = 1, 5, 10, 50, 100$ de cima para baixo, respectivamente. 57

3.3 Gráficos de rastreamento da probabilidade de dados logarítmicos completos para as Análises 1, 2: $\pi_i = 0,5, 0,8$ (superior esquerdo e direito, respectivamente) e Análise 3: $\pi_i = 1$ (inferior).	76
3.4 Completar os dendrogramas de ligação com base nas diferenças entre cada par de classificadores para as Análises 1–3 de cima para baixo, respectivamente.	78
3.5 $\Pr(w_i = 1 D)$ – Probabilidade posterior de que a classificação i seja informativa em cada análise (Análise 1: $\alpha = 0,5$, Análise 2: $\alpha = 0,8$). As classificações que são permutações aleatórias (41–50) são mostradas em vermelho.	80
3.6 Gráficos de rastreio dos dados logarítmicos completos de probabilidade para as Análises 1 e 2 (canto superior esquerdo, direito) e Análises 3 e 4 (canto inferior esquerdo, direito).	92
3.7 $\Pr(w_i = 1 D)$ – Probabilidade posterior de que a classificação i seja informativa em cada análise (Análise 1: $\alpha = \beta = 1$, Análise 2: $\alpha = \beta = 3$). Cores distinguem entre os diferentes priores em α	92
3.8 $\Pr(N_{e,i} = ij D)$ – Distribuição marginal posterior do número de clusters de entidades para cada análise.	93
3.9 Dendrogramas de agrupamento de entidades para as Análises 1, 2 (superior esquerdo é direito) e Análises 3, 4 (inferior esquerdo e direito).	94
4.1 Comparação de distribuições anteriores não paramétricas para agrupamento bidirecional.	101
4.2 Número de agrupamentos de classificadores e entidades sob o Processo de Dirichlet Aninhado Adaptado anterior para vários valores de parâmetros de concentração α e γ	104
4.3 Gráficos da probabilidade posterior $\Pr(w_i = 1 D)$ que o ranker i é informativo para ambos os cenários de anterior em sua capacidade: $\pi_i = 0,5$ (coluna da esquerda) e $\pi_i = 0,9$ (coluna da direita). A linha superior de gráficos mostra a comparação entre as análises restritas (*) e completas (irrestritas) para o conjunto de dados 3. Os gráficos na linha do meio são aqueles para as análises completas usando o Conjunto de Dados 3, com os gráficos correspondentes usando o Conjunto de Dados 4 na linha inferior.	134
4.4 Dendrogramas para agrupamento de ranker dentro do Conjunto de Dados 4 sob uma análise completa para $\pi_i \geq 0,5$ (gráfico esquerdo) e $\pi_i = 0,9$ (gráfico direito).	136
4.5 Dendrogramas para agrupamento de entidades para o Conjunto de Dados 3 (superior) e Conjunto de Dados 4 (inferior) condicional a um único agrupamento de classificadores sob ambas as especificações anteriores para as análises completas.	138
4.6 Probabilidades posteriores P5 para todas as análises, P10 para todas, exceto o caso dos 5 primeiros e P15 para os 15 primeiros e análises completas. As análises do Conjunto de Dados 3 e do Conjunto de Dados 4 são mostradas na linha superior e inferior, respectivamente, para cada escolha anterior de p	139

4.7 Dendrogramas de agrupamento de entidades (condicional a 2 clusters ranker) no rankercluster 1 (esquerda) e cluster ranker 2 (direita) para a análise do Conjunto de Dados 4 withpi = 0,9.	140
4.8 Gráfico da probabilidade posterior $\Pr(w_i = 1 D)$ que o ranker i é informativo (esquerda), as cores distinguem entre os clusters ranker "verdadeiros". Dendrograma (ligação completa) calculado usando a dissimilaridade Δ_{ij} entre os classificadores i e j (direita).	142
4.9Dendrogramas mostrando a dissimilaridade entre entidades dentro dos agrupamentos 1 (esquerda) e 2 (direita), condicionada a dois agrupamentos de classificadores ($N_r = 2$).	143
5.1Conjunto de dados de Roskam: Dendrograma (à esquerda) mostrando a estrutura do cluster ranker junto com a probabilidade posterior, $\Pr(w_i = 1 D)$, para cada classificador i (direita). 149	
5.2Densidades anteriores e posteriores marginais para o número de clusters de entidades dentro de cada cluster de classificadores (condicional a dois clusters de classificadores).	149
5.3Conjunto de dados de Roskam: Dendrogramas mostrando a estrutura de agrupamento de entidades dentro do cluster ranker 1 e 2 (esquerda e direita, respectivamente) condicional a dois clusters ranker.	150
5.4Conjunto de dados de Roskam: Dendrograma (à esquerda) mostrando a estrutura de agrupamento dos rankers junto com a probabilidade posterior $\Pr(w_i = 1 D)$ para cada rankingi (direita) para $\pi_i = 0,85, 0,75, 0,65$ (de cima para baixo, respectivamente).	151
5.5Conjunto de dados de Roskam: Densidades anteriores e posteriores marginais para o número de clusters de rankers (gráfico à esquerda) e o número de clusters de entidades dentro de cada cluster de rankers, condicionada a dois clusters de ranker, (gráfico à direita) para $\pi_i =$ 0,85, 0,75, 0,65	152
5.6Conjunto de dados de Roskam: Dendrogramas da estrutura de agrupamento de entidades dentro rankercluster 1 (esquerda) e ranker cluster 2 (direita). Estes são mostrados para cada especificação anterior, $\pi_i = 0,85, 0,75, 0,65$, de cima para baixo, respectivamente.	152
5.7Conjunto de dados da NBA: Dendrograma (à esquerda) mostrando a estrutura de agrupamento de classificadores e destacando os classificadores com $\Pr(w_i = 1 D) < 0,25$. Gráfico (à direita) das probabilidades posteriores $\Pr(w_i = 1 D)$ para cada classificador, com linhas verticais separando os grupos autocertificados.	154
5.8Densidades anteriores e posteriores marginais para o número de clusters de entidades dentro de cada cluster de classificação (condicional a dois clusters de classificação).	155

- 5.9 Conjunto de dados da NBA: Dendrogramas mostrando a estrutura do cluster de entidades dentro dos clusters 1 e 2 (esquerda e direita, respectivamente) condicionada a clusters de dois rankers. 155
- 5.10 A probabilidade P16 de que cada entidade esteja no top-16 sob o modelo WAND(x) e as probabilidades de que cada entidade seja uma entidade relevante sob BARD (\cdot). A linha vertical separa as equipes que realmente chegaram aos 16 primeiros playoffs. 157
- 5.11 Conjunto de dados da NBA: Dendrograma (à esquerda) mostrando a estrutura de agrupamento de classificadores e destacando esses classificadores com $Pr(w_i = 1 | D) < 0.25$. Gráfico (à direita) das probabilidades posteriores $Pr(w_i = 1 | D)$ para cada classificador, com linhas verticais separando os grupos autocertificados. A linha superior mostra os resultados para a escolha "escalonada" de π e a linha inferior mostra os resultados correspondentes quando $\pi = 0.5$ para todos os classificadores. 159
- 5.12 Conjunto de dados da NBA: Densidades anteriores e posteriores marginais para o número de clusters de ofrankers (gráfico da esquerda) e o número de clusters de entidades dentro de cada cluster de ranker, condicionada a dois clusters de ranker, (gráfico da direita) para escolha "escalonada" de π e $\pi = 0.5$ 159
- 5.13 Conjunto de dados da NBA: Dendrogramas mostrando a estrutura do cluster de entidades dentro dos clusters de ranker 1 e 2 (esquerda e direita, respectivamente) condicionada a dois clusters de ranker para escolha "escalonada" de π e $\pi = 0.5$, superior e inferior, respectivamente. 160
- 6.1 $\pi(D|\lambda_j, \sigma_j)$: log-verossimilhança maximizada dada cada ordem de escolha σ_j e o respectivo MLE $\hat{\lambda}_j$ para $j = 1, \dots, K$ 173
- 6.2 Gráficos de rastreamento dos dados logarítmicos completos verossimilhança (esquerda) e do pos-terior marginal $\pi(\sigma|D)$ (direita) para as cadeias 1 a 5 de cima para baixo, respectivamente (abordagem Metropolis-within-Gibbs). 187
- 6.3 Gráficos de rastreamento dos dados logarítmicos completos verossimilhança (esquerda) e do pos-terior marginal $\pi(\sigma|D)$ (direita) para cadeias 1 a 5 de cima para baixo, respectivamente (abordagem Metropolis-Hastings). 192
- 6.4 Subconjunto do π marginal posterior ($\sigma|D$) mostrando as 25 ordens de escolha com maior suporte posterior das cadeias 1 a 5 (lidas da esquerda para a direita). 193
- 6.5 $\pi(\theta)$: gráfico de densidade de uma mistura normal de dois componentes igualmente ponderada com médias de componentes de ± 2 e desvios-padrão de 0,05. 194
- 6.6 Densidades temperadas $\pi(\theta)/T$ para $T \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ 195

- 6.7 Traçado do log-verossimilhança (esquerda) e do $\pi(\sigma|D)$ (direita) . . . 203
- 6.8 Subconjunto do π marginal posterior ($\sigma|D$) mostrando as 25 ordens de escolha com maior suporte posterior (vermelho denota ordem de escolha usada para simular esses dados) . . . 204
- 6.9Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{10} = 1$. As densidades em cada caso são mostradas em branco e vermelho para aquelas obtidas sob a ordem de escolha com o maior suporte posterior e a ordem de escolha verdadeira, respectivamente. As cruzes azuis representam os valores verdadeiros a partir dos quais esses dados foram simulados (escala logarítmica). . . . 205
- 6.10 Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{10} = 1$. As densidades em cada caso são mostradas em branco e verde para aquelas obtidas sob o modelo EPL (para a ordem de escolha com o maior suporte posterior) e aquelas obtidas sob o modelo padrão de Plackett-Luce ($\sigma = (1, . . . , K)$) respectivamente. As cruzes azuis representam os valores verdadeiros a partir dos quais esses dados foram simulados (escala logarítmica). . . . 205

Lista de Tabelas

1.1 Representações vetoriais de ordem e ordem da mesma classificação	2
2.1 Tipos de classificação	27
2.2 Classificações agregadas em nossa análise dos conjuntos de dados 1 e 2, juntamente com as médias posteriores correspondentes (denotadas λ). O valor de λ que foi usado para simular esses dados também é reproduzido para facilitar a comparação	36
2.3 Classificações agregadas sob o modelo de Plackett-Luce ponderado para a análise do conjunto de dados 2 (para ambas as análises $\pi = 0.5, 0.8$) juntamente com as médias posteriores correspondentes. Para facilitar a comparação, os resultados da Tabela 2.2 (análises padrão de Plackett-Luce) também são fornecidos. A tabela também contém a probabilidade relativa das classificações agregadas em comparação com uma classificação uniforme, $r_i = K! \Pr(X = x_{aggi} -\lambda_i)$	49
3.1 Probabilidades posteriores do número de clusters de rankers, $\Pr(N_r = i D)$, para cada uma das três análises. A expectativa e o desvio padrão da distribuição marginal posterior também são mostrados junto com a distribuição anterior. Os valores modais são destacados em negrito.	77
3.2 Classificações agregadas sob a mistura infinita do modelo ponderado de Plackett-Luce para a análise do conjunto de dados 2 (para análises 1–3; $\pi = 0.5, 0.8, 1$) juntamente com as médias posteriores correspondentes. Os resultados da Tabela 2.2 (análises padrão de Plackett-Luce homogêneas) também são fornecidos para facilitar a comparação.	81
3.3 Probabilidades prévias, $\Pr(N_e = j)$, do número de clusters de entidades para cada análise (topo) e as expectativas prévias e desvios-padrão do número de clusters de entidades e do parâmetro de concentração (inferior). Os valores modais são destacados em negrito.	91

3.4 Médias posteriores marginais dos parâmetros de habilidade para cada análise.
A classificação agregada é a mesma em todas as análises. 95

4.1 Distribuição posterior do número de clusters de classificação N r para análises restritas
(*) e completas (irrestritas). Os números em negrito indicam valores modais. 135

4.2Distribuição posterior do número de clusters de entidades, condicionada a um
cluster de classificação única, para análises restritas (*) e completas (sem
restrições). Números em negrito indicam valores modais. 137

4.3Distribuição posterior do número de clusters de entidades, condicionada a
clusters tworanker, para cada análise do Conjunto de Dados 4 com $\pi = 0,9$. Os
números em negrito indicam valores modais. 140

4.4Alocação verdadeira de entidades em clusters, juntamente com o valor
trueparameter correspondente para cada um dos clusters de entidades. 141

4.5Distribuição posterior do número de agrupamentos de entidades, condicionada a
agrupamentos de duas fileiras. 142

4.6Ordenação das preferências posteriores nos agrupamentos de classificadores 1 e 2
(condicional em dois agrupamentos de classificadores) e a classificação geral/agregada,
com média (e desvio padrão) de seus parâmetros de habilidade. 144

5.1Distribuição prévia e posterior do número de clusters de ranker (até 2 d.p.). 148

5.2Conjunto de dados de Roskam: classificações de entidades por média posterior
dentro do cluster de classificação (condicional a dois clusters de classificação). A
classificação 1 corresponde à entidade mais preferida dentro de cada cluster. 150

5.3Análise da NBA: Ordenações de preferência posteriores dentro dos clusters de
classificação 1 e 2 (condicionadas a dois clusters de classificação) e a classificação
geral/agregada, com média (e desvio padrão) de seus parâmetros de habilidade. As
linhas horizontais indicam o agrupamento de entidades MAP dentro de clusters de
classificadores. Os números na parte inferior são o número de ocorrências em que
o MAPclustering foi observado (de 8038 iterações com dois clusters de
classificadores). 156

6.1Probabilidades de que cada entidade seja atribuída a uma classificação
específica para o modelo padrão de Plackett-Luce com $\lambda = (3, 4, 3, 2, 1)$ 168

6.2 Probabilidades cumulativas de cada entidade ser classificada não inferior a
(esquerda) e não superior a (direita) ou igual a cada posição. A entrada i, j
corresponde a $Pr(j \in x1:i)$ (esquerda) e $Pr(j \in x:iK)$ (direita). 169

6.3 Probabilidades de cada entidade ser classificada em cada posição. A entrada i, j corresponde a $\Pr(x_i = j)$, ou seja, a entidade de probabilidade j recebe a classificação i . . . 170
6.4 Um subconjunto da classificação das ordens de escolha (permutações) com base no valor do log-verossimilhança avaliado no MLE correspondente para os parâmetros de habilidade ($\pi(D ^{*}x_j, \sigma_j)$) 174
6.5 Permutações (ordens de escolha) usadas para a inicialização de cada cadeia 186
6.6 Taxas de aceitação para cada um dos 5 mecanismos de proposta para a ordem de escolha 188
6.7 Alocação teórica ideal de trabalho entre núcleos com tempo de execução total e relativo (assumindo que o trabalho leva uma unidade de tempo) 200
B.1 Conjunto de dados 1 usado na análise PL padrão 218
B.2 10 classificações não informativas adicionais usadas para formar o conjunto de dados 2 219
B.3 Conjunto de dados 3 usado no estudo de simulação 1 sob WAND 220
B.4 10 classificações não informativas adicionais usadas para formar o conjunto de dados 4 221
B.5 Conjunto de dados 5 usado no estudo de simulação 2 sob WAND. Linhas verticais separadas as classificações dentro de cada grupo de classificação diferente 222
B.6 Dados de psicologia de Roskam 223
B.7 Dados da NBA 224
B.8 Conjunto de dados 6 usado nos estudos de simulação para a análise bayesiana do Modelo de Plackett-Luce estendido 225
B.9 20 classificações adicionais usadas para formar o conjunto de dados 6 226

Capítulo 1

Introdução

1.1 Introdução

Embora muitas vezes despercebidos, os rankings aparecem em muitos aspectos da vida cotidiana. As pessoas classificam os objetos o tempo todo, seja com base em preferências pessoais ou experiências passadas; Talvez sejamossos filmes favoritos, jogos online, times esportivos ou até mesmo qual cafeteria preferimos visitar, a lista continua. Na era dos dados em que vivemos, a capacidade de analisar dados classificados está se tornando cada vez mais importante. Por exemplo, as grandes organizações estão interessadas nas preferências dos consumidores para fins publicitários e os motores de pesquisa em linha visam classificar os seus resultados de uma forma óptima. Os dados classificados também são um resultado comum de experimentos que visam descobrir as atitudes ou preferências de uma coorte em relação a um determinado conjunto de itens (Vigneau et al., 1999; Yu et al., 2005; Gormley e Murphy, 2006; Vitelli et al., 2018). Os eventos desportivos também podem dar origem a classificações, sendo particularmente comuns em corridas de cavalos/automóveis ou em torneios round-robin em que o resultado é uma ordenação das equipas ou dos concorrentes individuais; ver, por exemplo, Henery (1981), Stern (1990) e Caronand Doucet (2012).

Esta tese está preocupada com ordenações de preferência que surgem quando os classificadores fornecem uma classificação ou ordenação para um conjunto de entidades de acordo com algum critério. Normalmente, os classificadores serão indivíduos, embora essa estrutura seja bastante geral e grupos, organizações ou mesmo sensores possam ser considerados classificadores, entre outros. Existem possibilidades quase infinitas para as entidades, sejam candidatos políticos, cozinhas mundiais, universidades e assim por diante. Existem duas representações comuns de dados classificados que, usando a terminologia de Marden (1995), chamamos de vetor de classificação e ordem. Ambas as representações retratam as mesmas informações e, portanto, ao modelar dados classificados, deve-se ter cuidado com qual representação é usada. Formalmente, um vetor de classificação $y = (y_1, \dots, y_K)$ de K entidades é uma lista, onde a entrada y_i indica a classificação dada à i -ésima entidade. Em contraste, um vetor de ordem

Vetor de classificação		Vetor de ordem	
Entidade e	Entidade x		
Britânico	3	Índio	1
Chinês	2	Chinês	2
Italiano	4	Britânico	3
Índio	1	Italiano	4
Tailandês	5	Tailandês	5

Tabela 1.1: Representações vetoriais de classificação e ordem da mesma classificação

$x = (x_1, \dots, x_K)$ de K entidades pode ser pensado como uma lista de preferências onde a entrada x_i contém o rótulo da entidade na posição i (com a posição 1 sendo a mais preferida). Por exemplo, suponha que peçamos a um ranker para pedir 5 cozinhas populares do mundo, britânica, chinesa, italiana, Indiana e tailandesa em termos de sua preferência. Se o classificador preferir Indiano, chinês, britânico e italiano com tailandês o menos preferido, a Tabela 1.1 mostra os vetores de classificação e ordem correspondentes - observe que o vetor de classificação depende da ordem em que as entidades são listadas. Adotamos a representação do vetor de ordem e usamos os termos classificação e ordenação de forma intercambiável para denotar um vetor de ordem x . Independentemente do formato adotado, as classificações são observações multivariadas e, além disso, qualquer classificação particular pode ser pensada como uma permutação dos inteiros de 1 a K ; Essa perspectiva pode ser útil para o desenvolvimento de modelos para esses dados. Marden (1995) e, mais recentemente, Alvo e Yu (2014) fornecem uma visão geral dos modelos e da literatura estatística para dados classificados. Existem muitos tipos de modelos para dados classificados, incluindo modelos paramétricos, em estágios e baseados em distância. Os modelos baseados na distância baseiam-se na suposição de que existe uma classificação modal (das entidades) e que se espera que os classificadores relatem classificações que são, em certo sentido, "próximas" dessa ordenação modal. Embora a distância entre duas permutações não esteja bem definida, é necessário fazer uma escolha para se adequar a este tipo de modelo. Duas escolhas comuns da distância são a distância de Kendall e Spearman e estas dão origem aos modelos φ e θ de Mallows (1957), respectivamente; detalhes desses modelos (entre outros) podem ser encontrados em Flinger e Verducci (1986). A inferência bayesiana para modelos baseados em distância pode ser problemática, especialmente quando se supõe que a ordenação modal seja desconhecida, devido à natureza proibitiva de uma constante normalizadora intratável que deve ser aproximada na maioria dos casos. Além disso, os modelos baseados em distância tornam-se cada vez mais difíceis de ajustar à medida que o número de entidades aumenta devido à explosão do tamanho do espaço de permutação. Por esses motivos, não consideraremos modelos baseados em distância e, em vez disso, consideraremos modelos de classificação paramétricos.

Várias distribuições paramétricas sobre o conjunto de permutações foram desenvolvidas. Os chamados modelos de classificação por etapas são exemplos de modelos de classificação paramétrica. Palco-

modelos sábios são sustentados pela ideia de que o processo de classificação, ou seja, como um ranker constrói sua ordenação, pode ser decomposto em $K - 1$ estágios (dependentes). Nesta tese, nos concentraremos predominantemente no popular modelo de Plackett-Luce (Luce, 1959; Plackett, 1975), que pressupõe a ordem a termo, ou seja, a atribuição de entidades a posições no ranking procede sequencialmente do item mais preferido para o menos preferido. Se a suposição do processo de classificação avançada não for plausível, o modelo Reverse Plackett-Luce fornece uma escolha alternativa e Graves et al. (2003) e Henderson e Kirrane (2018) descobriram que esse modelo era mais apropriado ao modelar as corridas da NASCAR e da Fórmula 1, respectivamente. Além disso, Mollica e Tardella (2014) propuseram o modelo Extended Plackett-Luce, que permite que a suposição de um processo de classificação explícito seja relaxada. A inferência para este modelo é desafiadora, particularmente quando o número de entidades não é pequeno, pois a distribuição posterior é estendida para também estar acima do espaço de permutação. A solução bayesiana atual proposta por Mollica e Tardella (2018) depende de um espaço amostral restrito.

A maioria dos modelos de dados classificados trata as informações fornecidas por cada classificador igualmente, ou seja, eles assumem que cada classificador é igualmente informativo. Esta é uma suposição bastante forte. É fácil imaginar uma situação em que alguns classificadores estejam significativamente mais informados sobre as entidades em comparação com outros classificadores. Derry et al. (2014) tiveram como objetivo resolver esse problema e usaram a confiabilidade do classificador como parte de sua solução BARD (Bayesian Aggregation of Ranked Data). Além disso, a maioria dos modelos de dados classificados depende de fortes suposições sobre a homogeneidade dos dados classificados; A ideia de que existe uma visão geral de consenso é um exemplo. Modelos mais flexíveis foram propostos com Gormley e Murphy (2008a, b, 2009) e Mollica e Tardella (2014, 2016) considerando misturas finitas de modelos relacionados a Plackett-Luce e para permitir diferentes preferências entre os classificadores. Essa abordagem também foi adotada por Vitelli et al. (2018). No entanto, adoptaram um modelo baseado na distância — nomeadamente o de Mallows (1957) — em vez do modelo de Plackett-Luce. Modelos de mistura infinita mais flexíveis também foram propostos e essas abordagens permitem que o número de grupos seja inferido em vez de ser fixado pelo analista; ver, por exemplo, Caron et al. (2014). Embora esses tipos de modelos permitam que os classificadores expressem crenças diferentes, eles também assumem que cada grupo de classificadores pode distinguir entre cada uma das entidades. No entanto, é possível que um grupo (homogêneo) de classificadores não seja capaz de distinguir entre algumas entidades, ou seja, eles podem acreditar que algumas entidades são trocáveis. Este é um aspecto que muitas vezes é esquecido na literatura e embora métodos tenham sido propostos (de Leeuw e Mair, 2009; Choulakian, 2016), eles geralmente dependem de resumos ad-hoc, em oposição a uma abordagem baseada em modelo.

1.1.1 Objetivos da tese

O principal objetivo desta tese é fornecer modelos flexíveis que permitam a exploração da (possível) estrutura de subgrupos dentro de dados classificados. Mais especificamente, pretendemos identificar grupos homogêneos de indivíduos que compartilham crenças semelhantes, além de descobrir como, ou mesmo todos, esses grupos podem ter dificuldade em distinguir entre certas entidades. Além disso, pretendemos construir modelos que permitam uma potencial heterogeneidade entre as habilidades dos classificadores. A ênfase será colocada em esquemas de inferência eficientes que nos permitam ajustar nossos modelos, sob o paradigma bayesiano, em uma quantidade razoável de tempo computacional. Nas seções posteriores, aumentamos ainda mais a flexibilidade da modelagem, relaxando a suposição de um processo de classificação explícito. Isso é conseguido considerando o modelo Extended Plackett-Luce, que possui um parâmetro (representando o processo de classificação) que é um elemento do conjunto de todas as permutações. Construir esquemas de inferência bayesiana que possam efetivamente explorar grandes espaços discretos não é simples. Este problema é ainda mais desafiador quando tais espaços não exibem uma medida de distância natural (espaço de permutação) e, portanto, nesta tese, pretendemos fornecer uma solução eficaz para esse problema para espaços razoavelmente grandes.

Como ponto de partida, consideramos o modelo padrão de Plackett-Luce e mostramos como as inferências desse modelo podem ser afetadas até mesmo por uma quantidade modesta de classificações espúrias. Este resultado motiva a ideia de que um modelo adequado para dados classificados deve ser flexível o suficiente para permitir a (potencial) heterogeneidade entre as habilidades dos classificadores. Propomos o modelo WeightedPlackett-Luce (WPL), que é formado pelo aumento do modelo padrão de Plackett-Luce, com um parâmetro adicional que nos permite lidar com diferentes habilidades de classificação. Tomando o modelo WPL como nosso bloco de construção, relaxamos a suposição comum de que os dados vêm de um grupo homogêneo de classificadores, no qual cada classificador tem apenas diferenças bastante pequenas de uma visão geral de consenso. Isso é conseguido apelando para não-paramétricos bayesianos; especificamente, propomos uma mistura de processos de Dirichlet de modelos WeightedPlackett-Luce. Em seguida, consideramos a noção de que um grupo (homogêneo) de classificadores pode não ser capaz de distinguir entre algumas entidades, ou seja, eles acreditam que algumas entidades são trocáveis. Para permitir isso, consideramos uma distribuição prévia alternativa não paramétrica que permite que as entidades se agrupem. Combinar esses dois aspectos em um único modelo requer uma técnica de agrupamento bidirecional e nos concentrarmos no Nested DirichletProcess (NDP) (Rodriguez et al., 2008). O NDP não é muito adequado para agrupamentos de entidades e, portanto, propomos o processo Adaptado, Nested Dirichlet (ANDP) anteriormente. A inferência bayesiana prossegue sob abordagens marginais e condicionais, cada uma das quais tem prós e contras associados. Numerosas análises são realizadas em dados simulados e reais e mostram que nosso modelo tem um bom desempenho em diferentes cenários. Esses estudos também destacam a rica informação (posterior) disponível para o analista como resultado do ajuste de nossa

modelo.

Até agora, cada um dos modelos propostos baseou-se no pressuposto subjacente ao processo de classificação a prazo. No capítulo final, relaxamos a suposição de um processo de classificação conhecido, observando o modelo Extended Plackett-Luce (Mollica e Tardella, 2014). Após a introdução do modelo Extended Plackett-Luce, surge uma questão natural: "é possível identificar o processo de classificação dado um conjunto de classificações?". Motivamos a identificabilidade do processo de classificação por meio de vários exemplos e consideramos uma abordagem de máxima verossimilhança para cimentar ainda mais essa ideia. A construção de esquemas de amostragem posteriores (bayesianos) adequados para este modelo é desafiadora e, até onde sabemos, a única solução atual é dada por Mollica e Tardella (2018), mas isso depende de um espaço de parâmetros restrito. Nosso objetivo é desenvolver métodos MCMC capazes de explorar todo o espaço paramétrico e vários algoritmos de amostragem são apresentados, cada um com vários graus de sucesso. Descobrimos que nosso algoritmo final, que usa a cadeia de Markov acoplada a MetropolisMonte Carlo (MC3), funcionou bem e tanto a metodologia quanto o algoritmo MC3 são descritos em detalhes.

1.1.2 Esboço da tese

O restante desta tese está organizado da seguinte forma. Nas seções a seguir, fornecemos uma breve introdução à inferência bayesiana e às técnicas de amostragem de Monte Carlo da cadeia de Markov. Um algoritmo genérico de Metropolis-Hastings é descrito e o amostrador de Gibbs é mostrado como um caso especial. Métodos para diagnosticar a convergência de uma cadeia de Markov são discutidos e também consideramos estratégias sensatas para lidar com a saída de MCMC para garantir que obtenhamos um número razoável de amostras da densidade de interesse. Este capítulo conclui com uma breve discussão sobre o aumento de dados, pois esta é uma técnica particularmente útil que nos permite usar distribuições condicionais completas de forma fechada quando a probabilidade não é padrão.

No Capítulo 2, consideramos a análise de dados homogêneos classificados, ou seja, assumimos que todos os classificadores compartilham crenças semelhantes sobre a preferência das entidades. O modelo de Plackett-Luce (Luce, 1959; Plackett, 1975) e seus pressupostos subjacentes são descritos em detalhes. O modelo básico de Plackett-Luce é então estendido para atender quando os classificadores não relatam uma classificação completa de todas as entidades (classificações superiores e parciais) e o mecanismo subjacente de geração de dados associado é delineado. Um esquema MCMC eficiente é construído e consideramos um breve estudo de simulação para dar uma ideia de como as inferências posteriores podem ser feitas. A segunda metade do Capítulo 2 está preocupada com a noção de confiabilidade do classificador. A maioria dos trabalhos nesta área pressupõe que todos os classificadores sejam igualmente informados sobre as entidades que estão classificando. Muitas vezes, essa suposição será questionável e, portanto, desenvolvemos o novo modelo WeightedPlackett-Luce, pois isso nos permite modelar classificadores com confiabilidade diferente por meio de um

modelo de mistura de dois componentes. Vários estudos de simulação são considerados e comparamos inferências posteriores dos modelos padrão e ponderado de Plackett-Luce.

O Capítulo 3 se concentra em aumentar a flexibilidade de modelagem para que possamos lidar efetivamente com dados classificados heterogêneos. Para fazer isso, apelamos para os modelos de mistura de processos de Dirichlet, que são explorados em detalhes usando duas representações bem conhecidas. Dois modelos são apresentados. O primeiro considera a ideia de que os rankers podem ser heterogêneos em suas crenças sobre entidades. Essa ideia não é nova e as misturas de processos finitos e de Dirichlet dos modelos (padrão) de Plackett-Luce são bem exploradas na literatura. Estendemos um pouco essa abordagem construindo um modelo que comprehende uma mistura de processos de Dirichlet de modelos de Plackett ponderado. O segundo modelo que apresentamos é novo e visa explorar se algumas entidades são intercambiáveis, ou seja, se os classificadores acham difícil (ou impossível) distinguir entre certos grupos de entidades. Esta questão tem recebido pouca atenção na literatura e exploramos essa ideia considerando uma mistura de processos de Dirichlet sobre os parâmetros de habilidade em que o modelo Weighted Plackett-Luce é considerado a distribuição de classificação. A eficácia do nosso modelo para detectar grupos de entidades é avaliada por meio de estudos de simulação e estes concluem o capítulo.

No Capítulo 4, combinamos os aspectos de cada modelo apresentados no capítulo anterior, ou seja, incorporamos o agrupamento de classificadores e entidades em um único modelo, apelando para técnicas de agrupamento bidirecional. Focamos no processo Nested Dirichlet (Rodriguez et al., 2008) e fazemos uma adaptação necessária para que essa distribuição prévia possa ser usada dentro de um contexto de dados classificados. O modelo resultante é uma mistura de processo de Dirichlet aninhado adaptado ponderado (WAND) de modelos de Plackett-Luce. Tanto uma abordagem condicional quanto uma marginal para inferência posterior são consideradas com algoritmos eficientes fornecidos em cada caso. A metodologia é ilustrada usando vários estudos de simulação e no Capítulo 5 consideramos dois exemplos de dados reais. Para a primeira análise de dados reais, usamos um conjunto de dados originalmente coletado em 1968 por Roskam e mais recentemente estudado por de Leeuw (2006). Esses dados consistem em classificações obtidas de psicólogos do Departamento de Psicologia da Universidade de Nijmegen (Holanda). Cada psicólogo foi solicitado a classificar cada uma das 9 subáreas de acordo com o quanto apropriadas elas são para seu trabalho. A segunda análise de dados reais considera um conjunto de dados retirado de Deng et al. (2014) que envolveu classificações de equipes da NBA (National Basketball Association). Em seu artigo, Deng et al. propõem um modelo denominado "Agregação Bayesiana de Dados Classificados" (BARD) e comparamos as inferências do modelo WAND com as do BARD.

No Capítulo 6, relaxamos a suposição de um processo de classificação conhecido, observando o modelo Extended Plackett-Luce recentemente desenvolvido (Mollica e Tardella, 2014). Este modelo contém um parâmetro livre adicional (que é uma permutação) que nos permite aprender sobre a ordem em que um grupo homogêneo de classificadores atribui entidades a

Fileiras. A construção de um esquema de amostragem posterior (bayesiano) adequado para este modelo provou ser desafiadora. No entanto, descobrimos que o uso da cadeia de Markov acoplada a Metropolis Monte Carlo (MC3) funcionou bem e tanto a metodologia quanto o algoritmo MC3 são descritos em detalhes.

Finalmente, nossas conclusões são tiradas no Capítulo 7 e também fornecemos alguns tópicos sugeridos para trabalhos futuros.

1.2 Inferência bayesiana

Nesta tese, trabalhamos quase exclusivamente dentro do quadro bayesiano (Bernardo e Smith, 1994). Nesta configuração, todas as quantidades desconhecidas (parâmetros, variáveis latentes e assim por diante) são consideradas variáveis aleatórias. Uma distribuição de probabilidade conjunta descreve a relação entre as quantidades desconhecidas e os dados (observados). A distribuição posterior é a distribuição (condicional) obtida pelo condicionamento dos dados observados e é essa distribuição que nos permite fazer inferências sobre as quantidades desconhecidas (dados os dados). O posterior é o resultado de nossas crenças anteriores sobre as quantidades desconhecidas sendo atualizadas pelos dados observados por meio da função de verossimilhança.

1.2.1 Teorema de Bayes

Suponha que temos alguns dados $D = \{x_1, \dots, x_n\}$ e estamos interessados em aprender sobre uma coleção de K quantidades desconhecidas $\Lambda = (\lambda_1, \dots, \lambda_K)$. A probabilidade $L(\Lambda | D) = \pi(D|\Lambda)$ é a densidade de probabilidade dos dados dados os parâmetros, mas considerada como uma função dos parâmetros para dados conhecidos. Observe que um "modelo" é normalmente especificado por uma função de verossimilhança particular e isso descreve como os dados estão relacionados aos parâmetros. Dado que estamos trabalhando dentro da estrutura bayesiana, devemos também resumir nossas crenças (prévias) sobre as quantidades desconhecidas por meio da escolha de uma distribuição prévia adequadamente definida $\pi(\Lambda)$. O π posterior $(\Lambda|D)$ é a densidade que reflete nossas crenças atualizadas sobre os parâmetros Λ tendo observado os dados D e segue do Teorema de Bayes como

$$p(L|D) = \frac{\text{_____}}{\pi(D|L)\pi(L)\pi(D)} \quad (1.1)$$

Note-se que o denominador, $\pi(D) = \int \pi(D|\Lambda)\pi(\Lambda)d\Lambda$, é a verossimilhança marginal e é obtido integrando os parâmetros. Claramente, a marginalverossimilhança não depende dos parâmetros Λ e, portanto, esta é simplesmente uma constante de normalização que garante que a densidade posterior se integra a um. Segue-se que o Teorema de Bayes pode

ser escrito como

$$p(L|D) \propto p(D|L)\pi(L) \quad (1.2)$$

e assim o posterior é proporcional ao produto do anterior e da probabilidade.

Normalmente, a probabilidade marginal $\pi(D)$ e, portanto, a densidade posterior $\pi(\Lambda|D)$, não está disponível na forma fechada, ou seja, a posterior não é uma distribuição padrão que pode ser escrita analiticamente com distribuições e momentos marginais conhecidos. Nesses casos, devemos recorrer a outros métodos que nos permitem calcular a distribuição posterior. Um dos métodos mais populares, e aquele em que nos concentraremos, é o MonteCarlo da cadeia de Markov – este é o tópico da próxima seção.

1.3 Cadeia de Markov Monte Carlo

A cadeia de Markov Monte Carlo (MCMC) é uma técnica computacional que pode ser usada para obter realizações a partir da densidade posterior quando ela não está disponível na forma fechada. De fato, o MCMC pode ser usado para obter realizações de qualquer distribuição de interesse e, portanto, também pode ser útil para extrair amostras de distribuições genéricas de alta dimensão. A metodologia que sustenta o MCMC é bem estudada com inúmeros livros e artigos publicados na literatura; ver, por exemplo, Chib e Greenberg (1995), Brooks (1998) e Gamerman e Lopes (2006). A ideia geral é construir uma cadeia de Markov que tem distribuição estacionária $\pi(\cdot)$, que também é conhecida como distribuição alvo. Então, dado qualquer ponto de partida inicial, desde que executemos a cadeia por tempo suficiente para que ela converja (para a distribuição de destino), podemos atualizar repetidamente a cadeia para gerar amostras (dependentes) do π de destino (\cdot). Claramente, dentro de uma configuração de inferência bayesiana, desejamos que o alvo seja o π posterior($\Lambda|D$). A seguir, discutimos um algoritmo fundamental, e um caso especial associado, que nos permite construir cadeias de Markov onde a distribuição alvo é a distribuição posterior.

1.3.1 O algoritmo de Metropolis-Hastings

O algoritmo Metropolis-Hastings, proposto pela primeira vez por Metropolis et al. (1953) e depois generalizado por Hastings (1970), é considerado o algoritmo fundamental usado para construir esquemas MCMC que visam o π posterior ($\Lambda|D$). A noção de um kernel de transição, ou densidade de propostas, é uma ideia-chave por trás do algoritmo Metropolis-Hastings. A densidade (arbitrária) da proposta é denotada $q(\Lambda^*|\Lambda)$ e descreve, probabilisticamente, como passar de um estado atual Λ para um estado proposto Λ^* . O algoritmo (Metropolis-Hastings) a seguir gerará sucessivamente uma sequência de valores $\Lambda(1), \Lambda(2), \dots$, que formam uma cadeia de Markov com distribuição alvo $\pi(\Lambda|D)$.

1. Seja o contador de iteração $t = 1$ e initialize a cadeia para $\Lambda(0) = (\lambda(0)_1, \dots, \lambda(0)_K)$ que cai em algum lugar no suporte de $\pi(\Lambda|D)$, isto é, de modo que $\pi(\Lambda(0)|D) > 0$.

2. Desenhe Λ^* da densidade proposta $q(\Lambda^* | \Lambda(t-1))$.

3. Avalie a probabilidade de aceitação, $p = \min(1, A)$, onde

$$A = \frac{p(L^*|D)\pi(L^*|D) \times}{q(\Lambda(t-1)|\Lambda^*)q(\Lambda^*|\Lambda(t-1))}.$$

4. Seja $\Lambda(t) = \Lambda^*$ com probabilidade p ; caso contrário, seja $\Lambda(t) = \Lambda(t-1)$.

5. Deixe $t \rightarrow t + 1$ e retorne à etapa 2.

Normalmente, não seremos capazes de avaliar a constante normalizadora $\pi(D)$ da densidade posterior (1.1) e, portanto, podemos pensar que também não podemos avaliar a taxa de aceitação A . No entanto, como a densidade posterior aparece tanto no numerador quanto no denominador da taxa de aceitação, precisamos apenas conhecer a distribuição posterior até uma constante de proporcionalidade. Segue-se que, usando (1.2), podemos expressar de forma equivalente a probabilidade de aceitação da Etapa 3 como $p = \min(1, A)$ onde

$$A = \frac{\pi(D|L^*)\pi(L^*)\pi(D|L(t-1))\pi(L(t-1)) \times}{q(L(t-1)|L^*)q(L^*|L(t-1))}.$$

Observe que a escolha da densidade da proposta $q(\Lambda^*|\Lambda)$ é completamente arbitrária e a cadeia de Markov terá como alvo o posterior correto, independentemente da escolha feita (assumindo que o suporte da distribuição da proposta não é menor que o suporte do posterior, ou seja, $q(\cdot|\Lambda) > 0 \forall \Lambda$ onde $\pi(\Lambda|D) > 0$). No entanto, algumas distribuições propostas são melhores do que outras, no sentido de que conduzem a uma cadeia que converge rapidamente e se mistura bem, ou seja, uma cadeia que explora eficazmente o apoio de $\pi(\Lambda|D)$. Agora (brevemente) descrevemos algumas das escolhas mais comuns de distribuição de propostas antes de passar a discutir o "ajuste" dos algoritmos de Metropolis-Hastings na Seção 1.3.2.

Passeio aleatório Metropolis-Hastings

O algoritmo de Metropolis-Hastings de passeio aleatório é onde o valor proposto é da forma $\Lambda^* = \Lambda + \omega$ onde ω é um (vetor) de inovações aleatórias. Geralmente ω é escolhido para seguir uma distribuição normal (multivariada) com 0 média e estrutura de covariância diagonal e assim $q(\Lambda^* | \Lambda) \sim N(\Lambda, \sigma^2 I_K)$.

Propostas simétricas

Uma proposta simétrica é qualquer distribuição de proposta em que $q(\Lambda^*|\Lambda) = q(\Lambda|\Lambda^*)$ para todo Λ^* , Λ em seu suporte. Segue-se que, para este tipo de distribuição de propostas, as taxas de aceitação se simplificam para $A = \pi(\Lambda^*|D)/\pi(\Lambda|D)$ e, portanto, a probabilidade de aceitação é independente da densidade da proposta. Observe que a proposta de passeio aleatório (acima) é um exemplo de distribuição assimétrica da proposta.

Caminhada aleatória log-normal

Uma proposta de passeio aleatório log-normal é particularmente útil quando os parâmetros de interesse são restritos a serem estritamente positivos. O valor proposto é da forma $\Lambda^* = \exp(\log \Lambda + \omega)$ onde ω é uma inovação aleatória. Segue-se que, neste caso, a distribuição proposta é $q(\Lambda^*|\Lambda) \sim LNK(\log \Lambda, \sigma^2 K)$ onde LNK denota a distribuição log-normal K -dimensional. A distribuição log-normal não é simétrica em torno de sua média, portanto, a razão proposta $q(\Lambda^*|\Lambda) / q(\Lambda|\Lambda^*)$ deve ser calculada para avaliar a taxa de aceitação A . Segue-se que

$$A = \frac{p(L^*|D)\pi(L^*|D)}{\prod_{k=1}^K \frac{1}{L_k}} \times \frac{\prod_{k=1}^K \frac{1}{L_k}}{p(L|D)\pi(L|D)}.$$

Propostas de independência

Uma proposta de independência é um mecanismo para gerar valores propostos Λ^* que são independentes do estado atual da cadeia Λ . Segue-se que $q(\Lambda^*|\Lambda) = q(\Lambda^*)$ e a razão proposta simplifica para $q(\Lambda^*)/q(\Lambda)$. Além disso, a taxa de aceitação na Etapa 3 do algoritmo MH pode ser escrita como

$$A = \frac{p(L^*|D)\pi(L(t-1)|D)}{q(\Lambda(t-1))q(\Lambda^*)} \times$$

sob este mecanismo de proposta e, portanto, é claro que podemos aumentar a probabilidade de aceitação escolhendo $q(\Lambda^*)$ para ser o mais semelhante possível a $\pi(\Lambda^*|D)$. Observe que se $q(\Lambda^*) = \pi(\Lambda^*|D)$ então $A = 1$ e o valor proposto sempre será aceito - talvez não seja surpreendente, dado que o valor proposto é do posterior. Essa ideia leva ao Amostrador de Gibbs; Para mais pormenores, ver o ponto 1.3.3.

Atualizações do Componentwise

No que foi discutido até agora, consideramos a proposta Λ^* para conter valores propostos para todas as quantidades desconhecidas K . No entanto, na prática, pode ser difícil

construir uma distribuição de proposta adequada (K -dimensional), particularmente quando o número de parâmetros (desconhecidos) é grande. Uma solução para esse problema é considerar atualizações componentes, ou seja, atualizar cada quantidade desconhecida uma de cada vez (dependendo das quantidades desconhecidas restantes permanecendo fixas em seus valores atuais). Seja $\Lambda - k = (\lambda_1, \dots, \lambda_{k-1}, \lambda_k+1, \dots, \lambda_K)$ seja a coleção de todas as quantidades desconhecidas excluindo λ_k e então um algoritmo de Metropolis-Hastings para atingir o π posterior($\Lambda|D$) usando atualizações do componentwise é o seguinte.

1. Seja t o contador de iteração $t = 1$ e inicialize a cadeia para $\Lambda(0) = (\lambda(0)_1, \dots, \lambda(0)_K)$ que cai em algum lugar no suporte de $\pi(\Lambda|D)$, isto é, de modo que $\pi(\Lambda(0)|D) > 0$.
2. Seja $\Lambda' = \Lambda(t-1)$ o estado atual da cadeia, então para $k = 1, \dots, K$

(a) extrair λ'_k da densidade proposta $q_k(\lambda_k|\lambda')$. (b) avaliar a probabilidade de aceitação, $p_k = \min(1, A_k)$, onde

$$A_k = \frac{\pi(\lambda_k, \Lambda')}{\pi(\lambda_k|D)} \times \frac{q_k(\lambda'_k|\lambda)}{q_k(\lambda_k|\lambda')}.$$

(c) seja $\lambda'_k = \lambda_k$ com probabilidade p_k .

3. A letra $\Lambda(t) = \Lambda'$.

4. Deixe $t \rightarrow t + 1$ e retorne à etapa 2.

Novamente, cada distribuição de proposta (agora univariada) $q_k(\cdot)$ pode ser escolhida arbitrariamente e, portanto, uma combinação dos mecanismos de proposta discutidos anteriormente pode ser usada dependendo das quantidades desconhecidas de interesse.

1.3.2 Ajustando algoritmos de Metropolis-Hastings

A chave para a implementação de um algoritmo Metropolis-Hastings eficiente é a escolha da(s) distribuição(ões) proposta(s). Um algoritmo eficiente é aquele que resulta em uma cadeia que converge rapidamente (para a distribuição alvo) e também se mistura bem, ou seja, uma cadeia que explora efficientemente o suporte de $\pi(\Lambda|D)$. Para distribuições de propostas que dependem do estado atual da cadeia (ou seja, não de propostas de independência), deve ficar claro que a variância da proposta determinará como a cadeia de Markov explora o espaço amostral. Se a variância for muito pequena, a cadeia explorará o espaço lentamente, pois, embora os valores propostos provavelmente sejam aceitos, eles apenas moverão a cadeia a uma pequena distância do estado atual. Em contraste, se a variância for muito grande, então, embora os valores propostos estejam a uma grande distância do estado atual, apenas relativamente poucos dos valores propostos

valores serão aceitos – a cadeia permanecerá, portanto, "presa" no mesmo valor por muitas iterações, o que é ineficiente.

Claramente, existe uma troca entre a probabilidade de aceitação dos valores propostos e a distância que eles nos permitem mover pelo espaço amostral. Diante disso, seria útil se pudéssemos construir distribuições de propostas de modo que os valores propostos estejam a uma distância razoável do estado atual e também sejam provavelmente aceitos. Roberts e Rosenthal (2001) sugerem que, se a distribuição alvo for gaussiana, a probabilidade de aceitação ótima (aquela que maximiza a distância de "salto" ao quadrado esperada) é de 0,234. Este resultado foi estendido para alvos elípticos simétricos por Sherlock e Roberts (2009) com Sherlock (2013) posteriormente fornecendo um conjunto geral de condições suficientes para as quais a probabilidade de aceitação ideal é de 0,234. Além disso, se a distribuição proposta é um passeio aleatório normal, então foi sugerido por Gelman et al. (1996), entre outros, que a variância de ω deve ser

$$\frac{2.382 \text{Var}(\Lambda|D)}{K}.$$

É claro que, em geral, não conhceremos a matriz de variância posterior $\text{Var}(\Lambda|D)$, portanto, uma estimativa obtida por várias execuções piloto do algoritmo pode ser usada.

Infelizmente, não existe uma regra rígida e rápida que descreva a melhor forma de construir distribuições de propostas adequadas em geral. A forma da distribuição de destino afeta o desempenho dos mecanismos de proposta e, portanto, propostas personalizadas são normalmente necessárias para cada cenário. A estratégia que sugerimos é primeiro escolher uma distribuição de proposta que pareça sensata a priori e, em seguida, realizar várias iterações do algoritmo MH para calcular a taxa de aceitação empírica (# propostas aceitas/# iterações). Se a taxa de aceitação for baixa/alta, diminuir/aumentar a variância da distribuição da proposta até que a taxa de aceitação seja de 23%. No início desta seção, observamos que é somente em cenários em que a distribuição da proposta depende do estado atual da cadeia que a variância (da distribuição da proposta) afeta a forma como a cadeia de Markov explora o espaço amostral. Por construção, as propostas de independência não dependem do estado atual da cadeia e, portanto, a probabilidade ótima de aceitação para esse tipo de proposta não é de 0,234. De fato, para propostas de independência, é vantajoso tornar a probabilidade de aceitação o maior possível, ou seja, a probabilidade de aceitação ideal é uma neste caso. Em outras palavras, devemos ter como objetivo construir uma distribuição de proposta que seja o mais próxima possível da distribuição alvo (posterior).

1.3.3 O amostrador de Gibbs

O amostrador de Gibbs é um caso especial do algoritmo Metropolis-Hastings (componente), onde cada um dos valores propostos $\lambda \times k$ é extraído de sua distribuição condicional completa correspondente. Esta técnica foi proposta pela primeira vez por Geman e Geman (1984) no contexto do processamento de imagens e só mais tarde foi levada ao conhecimento dos estatísticos por Gelfand e Smith (1990). A distribuição condicional completa da k -ésima quantidade desconhecida é $\pi(\lambda_k|\lambda_{-k}, D)$ e é a distribuição condicional de λ_k dadas todas as outras quantidades desconhecidas e os dados. É frequente que, embora a π posterior ($\lambda|D$) pode ser intratável, podemos obter a distribuição condicional completa para cada quantidade desconhecida na forma fechada (e, portanto, amostra deles).

Suponha que estejamos no cenário em que temos um conjunto completo de distribuições condicionais completas, ou seja, $\pi(\lambda_k|\lambda_{-k}, D)$ está disponível na forma fechada para $k = 1, \dots, K$. Seja Λ denotar o estado atual da cadeia de Markov e lembrar da Etapa 2 (b) no algoritmo MH (componente) que a taxa de aceitação para cada quantidade desconhecida k é

$$\frac{A_k = \pi(\lambda_{*k}, \lambda_{-k}|D)}{\lambda_{-k}|D) \pi(\lambda_{*k}|D)} \times q_k(\lambda_k| \lambda_{-k})$$

que podemos expressar de forma equivalente como

$$\begin{aligned} A_k &= \frac{\pi(\lambda_{*k}|\lambda_{-k}, D) \pi(\lambda_{-k}|D) \pi(\lambda_k|\lambda_{-k}, D)}{D \pi(\lambda_{-k}|D) \times q_k(\lambda_k|\lambda_{-k}) q_k(\lambda_{*k}|\lambda_k)} = \\ &= \frac{\pi(\lambda_{*k}|\lambda_{-k}, D) \pi(\lambda_k|\lambda_{-k}, D) \times q_k(\lambda_k|\lambda_{-k})}{\pi(\lambda_{*k}|\lambda_k) q_k(\lambda_{*k}|\lambda_k)}. \end{aligned}$$

Agora, se construir uma proposta de independência onde o valor proposto para cada quantidade desconhecida é extraído de sua distribuição condicional completa correspondente, ou seja, tomar $q_k(\lambda_{*k}) = \pi(\lambda_{*k}|\lambda_{-k}, D)$, segue-se que

$$A_k = \frac{\pi(\lambda_{*k}|\lambda_{-k}, D) \pi(\lambda_k|\lambda_{-k}, D) \times \pi(\lambda_k|\lambda_{-k}, D)}{\pi(\lambda_{-k}, D) \pi(\lambda_{*k}|\lambda_{-k}, D)} = 1$$

e, portanto, qualquer valor proposto λ_{*k} é garantido para ser aceito (para $k = 1, \dots, K$). Usando esse resultado, obtemos o caso especial do algoritmo de Metropolis-Hastings conhecido como amostrador de Gibbs. Se as condicionais completas $\pi(\lambda_k|\lambda_{-k}, D)$ estão disponíveis na forma fechada $\text{garfo} = 1, \dots, K$ então uma cadeia de Markov que tem como alvo o π posterior($\lambda|D$) é o seguinte.

1. Seja o contador de iteração $t = 1$ e inicialize a cadeia para $\Lambda(0) = (\lambda(0)_1, \dots, \lambda(0)_K)$ que cai em algum lugar no suporte de $\pi(\Lambda|D)$, isto é, de modo que $\pi(\Lambda(0)|D) > 0$.
2. Obter uma nova realização $\Lambda(t) = (\lambda(t)_1, \dots, \lambda(t)_K)$ de $\Lambda(t-1)$ por amostragem do

distribuições condicionais

$$\begin{aligned}\lambda(t)1 &\sim \pi(\lambda_1|\lambda(t-1)2, \lambda(t-1)3, \dots, \\&\lambda(t-1)K, D), \lambda(t)2 \sim \pi(\lambda_2|\lambda(t)1, \lambda(t-1)3, \\&\dots, \lambda(t-1)K, D), \dots, \lambda(t)K \sim \pi(\lambda_K|\lambda(t)1, \\&\lambda(t)2, \dots, \lambda(t)K-1, D).\end{aligned}$$

3. Deixe $t \rightarrow t + 1$ e retorne à etapa 2.

O amostrador de Gibbs (acima) é particularmente útil quando é inviável amostrar diretamente de $\pi(\lambda|D)$, mas a amostragem de $\pi(\lambda_k|\lambda-k, D)$ é direta. Além disso, ao contrário do algoritmo padrão de Metropolis-Hastings, não precisamos construir distribuições de propostas adequadas (o que pode ser difícil na prática), pois são simplesmente as distribuições condicionais completas. O algoritmo que descrevemos é conhecido como amostrador Gibbs de varredura fixa e é frequentemente usado na prática, pois é simples de implementar. Outras generalizações, como o amostrador Gibbs de varredura aleatória, também existem; consulte o Capítulo 5 de Gamerman e Lopes (2006) para obter detalhes completos.

Metrópole dentro de Gibbs

Claro, não há razão para que precisemos nos restringir a implementar um algoritmo de amostragem de Metropolis-Hastings ou Gibbs. Os dois métodos de amostragem podem ser combinados e isso dá origem ao chamado algoritmo Metropolis-within-Gibbs. Este algoritmo é simplesmente o algoritmo MH componente da Seção 1.3.1, onde uma distribuição totalmente condicional é usada como a distribuição da proposta para algumas quantidades desconhecidas com distribuições de proposta (arbitrárias) sendo usadas para o restante. O algoritmo Metropolis-within-Gibbs é útil quando distribuições condicionais completas estão disponíveis apenas na forma fechada para um subconjunto das quantidades desconhecidas de interesse.

1.3.4 Bloquear atualizações

Na prática, é comum atualizar as quantidades desconhecidas dentro de uma cadeia MCMC uma de cada vez, ou seja, usando os algoritmos de amostragem MH ou Gibbs componentwise. Embora os amostradores deste tipo sejam normalmente mais fáceis de implementar, a utilização de actualizações de um único componente pode dar origem a problemas de convergência e mistura, especialmente quando algumas das quantidades desconhecidas têm uma elevada correlação posterior. Intuitivamente, se duas quantidades desconhecidas λ_i e λ_j ($i \neq j$) estão altamente correlacionadas, então a construção de uma proposta para λ_i precisa levar em conta o valor atual de λ_j . Isso pode ser alcançado usando uma proposta com um

pequena variação, mas isso levará a uma cadeia de mistura ruim. Uma solução prática para esse problema é usar uma atualização de bloco de tais parâmetros altamente correlacionados, ou seja, atualizar várias quantidades desconhecidas simultaneamente. No exemplo acima, pode ser vantajoso gerar o "valor" proposto ($\lambda^{*i}, \lambda^{*j}$) a partir de uma distribuição de proposta (bivariada) $q(\lambda^{*i}, \lambda^{*j} | \lambda_i, \lambda_j)$ e aceitar ou rejeitar a atualização para ambas as quantidades desconhecidas. É claro que essa ideia se generaliza diretamente para tamanhos de bloco maiores que 2 e mais detalhes e discussões podem ser encontrados em Gamerman e Lopes (2006) e Gelman et al. (2014), entre outros.

1.3.5 Convergência

Lembre-se de que a ideia geral por trás do MCMC é construir uma cadeia de Markov que tenha a π posterior ($\Lambda|D$) como sua distribuição estacionária (alvo). Portanto, dado que estamos interessados apenas em obter realizações posteriores, devemos garantir que a cadeia tenha atingido sua distribuição estacionária antes de usar outras amostras geradas. Uma vez que uma cadeia de Markov atingiu sua distribuição estacionária, diz-se que ela convergiu. É bem sabido que, à medida que o número de iterações aumenta, a distribuição da cadeia de Markov tende para a distribuição posterior (estacionária), ou seja, $\Lambda(t)|D \rightarrow \Lambda|D$ as $t \rightarrow \infty$. Obviamente, não podemos realizar um número infinito de iterações e, portanto, é útil considerar quantas iterações são necessárias para que $\Lambda(t)|D \approx \Lambda|D$; Isso é conhecido como período de burn-in. O período de burn-in necessário depende muito da forma da distribuição posterior e, em menor grau, do local onde a cadeia é inicializada. É evidente que isso vai depender da situação de interesse. Dito isso, existem métodos para detectar quando uma cadeia de Markov não convergiu. Normalmente, isso é feito por inspeção visual de gráficos de rastreamento, mostrando como as quantidades desconhecidas mudam ao longo das iterações. Se as quantidades desconhecidas mostrarem uma tendência clara ao longo das iterações, isso indica que a cadeia não atingiu sua distribuição estacionária - neste caso, o número de iterações (o período de queima) deve ser aumentado. Em contraste, se os gráficos de rastreamento mostram as quantidades desconhecidas movendo-se em torno do suporte da distribuição de maneira estável, isso sugere que a cadeia de Markov convergiu. Gelfand e Smith (1990) sugerem alguns controles adicionais (informais) que podem ser úteis na avaliação da convergência e alguns controles mais formais foram sugeridos por Geweke (1992), Raftery e Lewis (1992, 1996) e Gelman (1996), entre outros.

Embora essas verificações sejam úteis, ainda é possível (e infelizmente bastante fácil) assumir incorretamente que uma cadeia de Markov convergiu, particularmente quando a distribuição estacionária (posterior) é multimodal. Pode acontecer que, embora as quantidades desconhecidas mostrem sinais de estacionariedade, elas estejam de fato "presas" em um modo local e, portanto, não estejam explorando o suporte da distribuição posterior. Na tentativa de evitar o diagnóstico incorreto da convergência, é útil executar várias cadeias de Markov simultaneamente –

cada um dos quais deve ser inicializado a partir de um valor inicial diferente. Se os gráficos de rastreamento de cada cadeia não se sobreponem, isso é indicativo de que as cadeias ainda não convergiram para a distribuição de destino e é necessário um período de burn-in mais longo.

1.3.6 Análise de amostras posteriores

Uma vez que a cadeia de Markov tenha收敛ido para sua distribuição estacionária, segue-se que todas as amostras geradas serão da π posterior($\Lambda|D$) por construção. No entanto, como observamos ao introduzir o MCMC pela primeira vez no início desta seção, as amostras geradas a partir da cadeia de Markov serão dependentes e os sorteios sucessivos são considerados autocorrelacionados. Se os valores sucessivos são altamente correlacionados, então a quantidade de informação (sobre a distribuição posterior) contida em amostras consecutivas é muito menor do que se esses valores fossem independentes. Um gráfico de autocorrelação pode ser útil para avaliar a quantidade de dependência entre amostras consecutivas. A coda do pacote R (Plummer et al., 2006) fornece uma função útil para gerar um gráfico de autocorrelação que é simplesmente a função de autocorrelação em diferentes tempos de atraso (além de muitos outros diagnósticos de convergência MCMC). Se as amostras geradas forem altamente autocorrelacionadas, pode ser útil diluir a saída do MCMC, o que é feito considerando apenas cada i -ésima iteração.

Quando obtemos um número razoável de realizações posteriores (quase) não autocorrelacionadas, é simples calcular estimativas de estatísticas resumidas posteriores, como as médias marginais e variâncias das quantidades desconhecidas de interesse. Além disso, podemos facilmente obter gráficos de distribuições posteriores marginais (ou conjuntas) usando uma estimativa de densidade de kernel.

1.4 Aumento de dados

Concluímos este capítulo com uma breve visão geral do aumento de dados (Tanner e Wong, 1987) e destacamos as principais vantagens de usar essa abordagem. Começamos com a estrutura padrão. Suponha que a forma de nossa probabilidade seja fora do padrão ou mesmo intratável. A aplicação do Teorema de Bayes geralmente resultará em nossa distribuição posterior, $\pi(\Lambda|D)$, tomando uma forma não padronizada. Isso é um tanto inconveniente, pois desejamos obter amostras dessa distribuição. Claro, poderíamos apelar para o algoritmo Metropolis-Hastings descrito na Seção 1.3.1 para obter realizações posteriores. No entanto, apelar para o aumento de dados pode nos permitir fazer melhor.

A ideia geral por trás do aumento de dados é introduzir algumas variáveis latentes Z para que a densidade posterior conjunta dessas variáveis, juntamente com as quantidades desconhecidas de interesse (Λ), seja de uma forma conveniente, ou seja, $\pi(\Lambda, Z|D)$ é uma distribuição de probabilidade bem conhecida.

A distribuição posterior conjunta de Λ e Z é

$$p(L, Z|D) \propto p(D|L, Z)\pi(Z|L)\pi(L),$$

de onde a densidade posterior de interesse é

$$p(L|D) \propto \int_Z \pi(D|L, Z)\pi(Z|L)\pi(\Lambda)dZ,$$

isto é, a distribuição marginal de nosso posterior aumentado. Segue-se que, se pudermos gerar amostras da distribuição posterior aumentada $\pi(\Lambda, Z|D)$ então podemos obter trivialmente a distribuição posterior sobre as quantidades desconhecidas de interesse Λ .

Infelizmente, na prática, muitas vezes é difícil construir variáveis latentes Z de modo que a densidade posterior da articulação $\pi(\Lambda, Z|D)$ é uma distribuição de probabilidade bem conhecida. No entanto, em alguns cenários, pode ser razoavelmente simples introduzir variáveis latentes que resultam nas distribuições condicionais completas $\pi(\Lambda|D, Z)$ e $\pi(Z|D, \Lambda)$ sendo de forma padrão. De fato, muitas vezes as variáveis latentes são introduzidas definindo sua distribuição condicional completa $\pi(Z|D, \Lambda)$. Se todas as distribuições condicionais completas forem conhecidas, então, a partir dos resultados da Seção 1.3.3, deve ficar claro que podemos usar um amostrador de Gibbs para obter realizações da junta posterior, ou seja, repetidamente

- atualizar Λ dado D e Z por amostragem de $\pi(\Lambda|D, Z)$
- atualizar Z dado D e Λ por amostragem de $\pi(Z|D, \Lambda)$. Segue-se que uma escolha criteriosa de variáveis latentes pode nos permitir evitar a necessidade de implementar um algoritmo MH (que precisa que construirmos e ajustemos as distribuições propostas) e, em vez disso, usar uma abordagem de amostragem de Gibbs mais direta se as distribuições condicionais completas estiverem disponíveis.

Capítulo 2

Análise de dados classificados homogêneos

2.1 Introdução

Consideraremos o popular modelo de Plackett-Luce (Luce, 1959; Plackett, 1975), que é uma extensão dos dados de comparação múltipla (classificados) do modelo para comparações pareadas proposto por Bradley e Terry (1952). Este modelo baseia-se em fortes pressupostos sobre a homogeneidade dos dados classificados, como a ideia de que todos os classificadores partilham uma visão geral consensual acordada em relação à preferência das entidades. Essa suposição talvez não seja bem justificada em muitos cenários do mundo real. Neste capítulo, começamos assumindo que todos os classificadores compartilham visões semelhantes e desenvolvem modelos mais flexíveis que permitem a heterogeneidade entre as crenças dos classificadores nos Capítulos 3 e 4. Os dados normalmente consistem em classificações completas e parciais (a serem definidas na Seção 2.2) e detalharemos como o modelo de Plackett-Luce pode ser modificado para permitir uma classe muito mais rica de classificações, como classificações parciais top-M e top-M (definidas na Seção 2.2.1). Ao longo da maior parte da literatura, presume-se que qualquer classificador em particular não tem mais (ou menos) probabilidade de compartilhar uma preferência semelhante (das entidades) à visão expressa pelo (assumido) grupo de consenso geral. Na Seção 2.5, questionamos essa suposição e propomos que alguns indivíduos talvez estejam significativamente mais informados sobre as entidades que estão classificando e, portanto, sua opinião deve ter mais peso. A confiabilidade do ranker é introduzida no modelo por meio de um indicador binário latente dentro da probabilidade de Plackett-Luce.

2.2 O modelo Plackett-Luce

Assumimos que nossos dados (classificações) são observações do modelo de Plackett-Luce (Luce, 1959; Plackett, 1975). Definimos o conjunto de todas as entidades como $K = \{1, \dots, K\}$ com $K = |\mathcal{K}|$. Cada entidade tem uma "classificação de habilidade" $\lambda_k > 0$ para $k = 1, \dots, K$. As classificações individuais não precisam conter todas as entidades e, portanto, deixamos $n_i \leq K$ ser o número de entidades contidas na classificação i . Assim, uma observação típica deste modelo é $x_i = (x_{i1}, \dots, x_{in_i})$, onde x_{ij} é a entidade que tem classificação j na classificação i . A probabilidade de tal observação é

$$\begin{aligned} \Pr(X_i = | \lambda) &= \prod_{j=1}^{\bar{n}_i} \frac{\lambda_j \sum_{m=1}^{n_i} \lambda_m}{x_{im}} \\ &= \prod_{j=1}^{\bar{n}_i} \frac{\lambda_j \sum_{m=1}^{n_i} \lambda_m}{x_{im}} . \end{aligned} \quad (2.1)$$

A probabilidade de Plackett-Luce acima é considerada um modelo de vários estágios (Marden, 1995) devido à forma como é construído. Esta construção em vários estágios é naturalmente destacada se considerarmos um exemplo simples. Suponha que temos uma classificação completa de $K = 4$ entidades, ou seja $x = (1, 2, 3, 4)$, e parâmetros de habilidade $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. A probabilidade dessa classificação no modelo de Plackett-Luce é

$$\Pr(x = (1, 2, 3, 4) | \lambda) = \frac{\lambda_1 \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{x} \frac{\lambda_2 \lambda_2 + \lambda_3 + \lambda_4}{x} \frac{\lambda_3 \lambda_3 + \lambda_4}{x} .$$

A partir disso, observamos que a probabilidade de uma determinada classificação é construída como o produto das probabilidades individuais (condicionais) de que cada entidade seja classificada dentro da posição independente. Seja $|c|_n$ um operador que recebe um vetor de comprimento arbitrário c e normaliza (mas mantém a proporcionalidade) os valores tais que $\sum_i c_i = 1$. Deve então ficar claro que a probabilidade de qualquer entidade receber a classificação 1 é dada pela entrada correspondente dentro de $|\lambda|_n$. A probabilidade condicional de uma entidade ser alocada classificação 2 é dada pela entrada correspondente em $|\lambda| \setminus |\lambda_1|_n$, ou seja, os valores normalizados, dado que a entidade à classificação 1 não está mais disponível para seleção. Esse condicionamento iterativo (em posições que já foram atribuídas) continua até que apenas uma única entidade permaneça, momento em que essa entidade é classificada por último. Muitas vezes é útil pensar nessa construção em termos de uma corrida contendo cavalos K . Naturalmente, o cavalo que cruza a linha primeiro recebe a classificação/posição 1, ou seja, a entidade mais forte ou preferida dentro de um contexto de classificação. O nível 2 é concedido ao cavalo que teria vencido a corrida se o cavalo que terminou em primeiro lugar não participasse. Da mesma forma, a classificação 3 é concedida ao cavalo que teria vencido, dado que nem o vencedor nem o cavalo premiado com a classificação 2 apareceram na corrida. Este processo continua até que a classificação K seja a única classificação restante a ser atribuída, neste ponto a corrida contém apenas um único cavalo que vencerá por definição.

Como a probabilidade de Plackett-Luce é construída usando o método descrito acima, diz-se que segue o chamado "processo de classificação direta" (Mollica e Tardella, 2014). Embora esse processo de classificação seja intuitivo, ele resulta em uma limitação significativa para o modelo de Plackett-Luce. Ao optar por modelar classificações sob a probabilidade de Plackett-Luce, também assumimos inherentemente que cada classificação é formada usando o processo de classificação direta. Essa suposição é frequentemente esquecida e talvez não seja bem justificada em um cenário do mundo real. Podemos imaginar um cenário em que um determinado classificador esteja mais confiante (ou equivalentemente mais certo) sobre as classificações das entidades que são suas mais e menos preferidas e, portanto, opte por alocar essas classificações primeiro. Assim, neste cenário, um classificador específico pode alocar entidades para classificar 1, depois para classificar K, depois para classificar K - 1 e assim por diante. Este não é o processo de classificação avançada. Uma consequência necessária é que, neste caso, a verdadeira distribuição de classificação subjacente não seguirá o modelo de Plackett-Luce. Com efeito, como veremos no capítulo 6, a escolha do modelo de Plackett-Luce (classificação a prazo) em tal cenário resulta normalmente numa má aproximação à verdadeira distribuição de classificação subjacente e, consequentemente, em inferências potencialmente enganosas.

Se a suposição do processo de classificação futura não for plausível, poderíamos optar por modelar classificações usando o modelo Reverse Plackett-Luce. Este modelo foi sugerido por Marden (1995) e usa uma construção de vários estágios semelhante ao modelo padrão de Plackett-Luce, mas no ranking inverso. A probabilidade de uma determinada classificação sob este modelo é

$$\Pr(X_i = | \lambda) = \prod_{j=1}^K \frac{\lambda x_i K - j + 1 \sum K - j}{\lambda x_i K - 1 \sum K - j}$$

Portanto, para a classificação completa $x = (1, 2, 3, 4)$, a probabilidade é

$$\Pr\{x = (1, 2, 3, 4)|\lambda\} = \frac{L_4 L_4 + L_3 + L_2 +}{L_1 x} \frac{\lambda_3 l_3 + \lambda_2 + \lambda_1}{x} \frac{\lambda_2 \lambda_2 + \lambda_1}{\lambda_1 l_1} \dots$$

Talvez não seja de surpreender que esse modelo siga o chamado "processo de classificação para trás" (Mollica e Tardella, 2014), ou seja, cada classificação é formada pela alocação da entidade que é menos preferida (classificação K), seguida pela segunda menos preferida (classificação K - 1) e assim por diante. Em termos de nosso cenário de corrida de cavalos, o cavalo que termina primeiro é aquele que recebe a classificação K, então o cavalo que teria vencido com a condição de o cavalo "vencedor" não estar na corrida recebe a classificação K - 1, o último cavalo restante é o mais preferido e recebe a classificação 1. Graves et al. (2003) usaram este modelo no contexto das corridas da NASCAR e, mais recentemente, Henderson e Kirrane (2018) usaram o modelo ReversePlackett-Luce ao analisar as corridas de Fórmula 1. Ambas as análises descobriram que a suposição do processo de classificação reversa resultou em um melhor ajuste do modelo do que assumir o processo de classificação avançada. Talvez seja difícil justificar a priori qual suposição é mais plausível. De fato, um ranker pode optar por alocar suas fileiras em qualquer um dos

$K!$ pedidos incluídos no SK, o conjunto de todos os $K!$ permutações de elementos K. Mollica e Tardella (2014) exploraram essa ideia e desenvolveram o modelo Extended Plackett-Luce. Este modelo relaxa as suposições sobre qualquer processo de classificação explícito e a "ordem de escolha" é inferida a partir das próprias classificações. No que se segue, assumimos o processo de classificação progressiva e, portanto, usamos a forma padrão do modelo de Plackett-Luce. Revisaremos o modelo de Plackett-Luce estendido no Capítulo 6.

Uma outra limitação do modelo de Plackett-Luce é que ele define apenas uma probabilidade para certos tipos de classificação. O modelo exige que cada classificador relate uma posição para cada uma das entidades que considera. Isso permite dois tipos de classificação: (a) Classificações completas, que ocorrem quando um classificador considera e atribui uma classificação a todas as entidades possíveis e (b) Classificações parciais, que ocorrem quando um classificador considera um subconjunto de todas as entidades, mas ainda relata uma classificação para cada entidade considerada, e assim $n < K$ neste cenário. Comparação emparelhada, ou seja, uma lista ordenada de duas entidades é equivalente a uma classificação parcial com $n = 2$. Isto não deve ser surpreendente, dado que o modelo de Plackett-Luce é uma generalização (para dados de comparação múltipla) do modelo de Bradley-Terry, que é definido para dados de comparação emparelhados (Bradley e Terry, 1952). Também consideraremos um caso especial conhecido comumente como rankings top-M. Aqui, os indivíduos relatam uma classificação apenas para aquelas entidades que classificam como sendo posicionadas de 1 a M (onde consideraram mais do que M entidades). O modelo de Plackett-Luce não captura adequadamente as informações contidas em dados desse tipo. Por exemplo, se ingenuamente escolhermos modelar as classificações top-M ($withn = M$) usando o modelo Plackett-Luce em (2.1), a entidade atribuída à classificação M seria tratada como se estivesse classificada em último lugar. Além disso, o modelo também se comportaria como se todas as entidades que não aparecessem no ranking não fossem consideradas pelo classificador, ou seja, o modelo trataria isso como um ranking parcial. No entanto, no caso de um ranking M-superior, temos a informação adicional de que, embora não recebam uma classificação específica, as entidades que não aparecem no ranking são consideradas como tendo pelo menos classificação M + 1. Deixamos, portanto, claro que, dentro desta tese, os termos rankings parciais e superiores são usados para se referir aos tipos de ranking conforme definido acima.

A princípio, parece que permitir classificações top-M pode ser problemático, pois desejamos marginalizar todas as posições possíveis (desconhecidas) das entidades não classificadas. No entanto, o modelo de Plackett-Luce (mais especificamente a distribuição que induz sobre os rankings) é internamente consistente, ou seja, a probabilidade de um determinado ranking é independente do subconjunto de entidades a partir do qual o ranking foi formado; ver Hunter (2004) para um esboço de prova. Daqui resulta que, no modelo de Plackett-Luce, é trivial combinar consistentemente classificações incompletas (superiores, parciais). A seção a seguir detalha como a probabilidade de Plackett-Luce pode ser estendida para usar todas as informações contidas nas classificações top-M.

2.2.1 Classificações Top-M

Em cenários do mundo real, muitas vezes encontramos uma classe muito mais ampla de classificações que podem ser classificadas como classificações parciais top-M e top-M. Uma classificação top-M é obtida quando um determinado indivíduo considera todas as entidades K, mas apenas atribui entidades às classificações de 1 a M (suas entidades M preferidas) e deixa as entidades restantes "não classificadas". Uma classificação superior-parcial é um caso especial de uma classificação superior-M que é obtida quando um indivíduo considera apenas um subconjunto de todas as entidades (dai $n < K$) e novamente prossegue para atribuir apenas entidades às classificações de 1 a M (deixando as restantes que consideravam não classificadas).

Uma modificação na probabilidade de Plackett-Luce é necessária para permitir esses tipos de classificação adicionais. Caron et al. (2014) detalham uma modificação para permitir classificações top-M e seu resultado pode ser trivialmente estendido para permitir classificações parciais top-M. Lembre-se de que o conjunto de todas as entidades é definido por $K = \{1, \dots, K\}$. Agora suponha que o ranker i considere $K_i \subseteq K$ e denote o conjunto dessas entidades como $K_i \subseteq K$. Também deixe $U_i = K_i \setminus \{x_i\}$ ser a coleção de entidades "não classificadas" (para o ranker i), com o que queremos dizer as entidades que eles consideraram, mas não apareceram em sua classificação. A partir da definição de U_i , fica claro que qualquer entidade $k \in U_i$ é considerada classificada pelo menos $(n_i + 1)^0$. A probabilidade de uma classificação particular sob o modelo de Plackett-Luce (agora modificado) é

$$\Pr(X_i = | \lambda) = \prod_{j=1}^K \frac{\lambda_{xim=j}}{\lambda_m} \quad (2.2)$$

Observe que

$$\Pr(H_a = | \lambda) = \prod_{j=1}^{n_i-1} \frac{\lambda_{xij}}{\sum_{m \in U_i} \lambda_m}$$

a menos que $U_i = \emptyset$, o conjunto vazio/nulo. Portanto, em uma situação em que temos apenas (a) classificações completas ou (b) parciais (portanto, $U_i = \emptyset \forall i$), as probabilidades de Plackett-Luce modificadas simplificam e recuperamos (2.1). Doravante, manteremos a generalidade total e prosseguiremos assumindo que pelo menos um classificador fornece uma classificação superior ou parcial superior, ou seja, existe um i tal que $U_i \neq \emptyset$. Também notamos que, ao contrário do modelo padrão de Plackett-Luce, simplesmente analisar as classificações reversas não é mais equivalente a analisar as classificações padrão (diretas) sob o modelo de Plackett-Luce reverso. Com efeito, a análise das classificações inversas no modelo 2.2 seria particularmente difícil, uma vez que não dispomos de classificações explícitas para as entidades consideradas classificadas pelo menos na posição $M + 1$. Isso resulta em essas entidades efetivamente empataadas em "primeiro" lugar (sob as classificações reversas). É claro que os dados dessa forma podem ser analisados implementando métodos para lidar com empates em dados classificados – esses

são discutidos na Seção 2.2.4.

2.2.2 Problemas de identificação

O modelo de Plackett-Luce sofre de um problema fundamental de identificação de parâmetros: a probabilidade (2.2) é invariante à multiplicação escalar estritamente positiva dos parâmetros de habilidade. Mais formalmente, se deixarmos $\lambda^*k = C\lambda_k$ para $k = 1, \dots, K$ com $C > 0$, então (devida à normalização dentro da construção da probabilidade) temos $\Pr(X = x|\lambda) = \Pr(X = x|\lambda^*)$ para qualquer classificação x . Podem ser adoptadas várias abordagens para ultrapassar esta questão. Por exemplo, poderíamos escolher restringir λ de modo que ele esteja no $K - 1$ simplex dimensional. Alternativamente, poderíamos adotar uma abordagem semelhante à de uma restrição de canto e corrigir $\lambda_1 \equiv 1$. Ambos os métodos reduzem o número de parâmetros livres para $K - 1$. Caron and Doucet (2012) observou que o problema de identificação também pode resultar em má mistura dentro das cadeias MCMC. Resolver o problema de mistura (e identificabilidade) é obviamente desejável e, dentro da solução bayesiana que consideramos, isso pode ser facilmente alcançado por meio de uma estratégia de redimensionamento adequada dentro do esquema de inferência. Passamos agora a lidar com a questão do reescalonamento.

2.2.3 Reescalonamento

Vamos considerar $\Lambda^\dagger = \sum_{Kk=1} \lambda_k$, a soma de todos os parâmetros de habilidade K . Conforme discutido na Seção 2.2.2, a probabilidade de Plackett-Luce é invariante à multiplicação escalar dos parâmetros λ , portanto, Λ^\dagger não é identificável por verossimilhança. De fato, se deixarmos $\lambda^*k = \lambda_k/\Lambda^\dagger$ para $k = 1, \dots, K$, temos que $\pi(\lambda^*, \Lambda^\dagger | D) = \pi(\lambda^* | D) \pi(\Lambda^\dagger)$. Caron and Doucet (2012) observaram que, sem a adição de uma etapa de redimensionamento, os esquemas MCMC para modelos Plackett-Luce podem sofrer de má mistura. A ideia é redimensionar os parâmetros para que a distribuição posterior de Λ^\dagger seja igual à sua distribuição anterior. Isso é obtido executando uma etapa de redimensionamento apropriada em cada iteração do esquema MCMC. Obviamente, o reescalonamento necessário dependerá da situação, pois a distribuição a priori em Λ^\dagger é induzida pela escolha a priori para cada um dos parâmetros de habilidade, λ_k .

2.2.4 Laços

Os vínculos nos dados classificados podem apresentar desafios específicos de modelagem. A maioria dos modelos para dados classificados, incluindo o modelo de Plackett-Luce, baseia-se no pressuposto de que apenas uma única entidade pode ser atribuída a uma determinada classificação, ou seja, várias entidades não podem ser "vinculadas" à mesma posição. No entanto, é possível superar esse problema e, de fato, vários métodos para incorporar laços em uma análise sob o modelo de Plackett-Luce

têm sido discutidas na literatura. Talvez o método mais intuitivo seja avaliar a contribuição exata que as entidades vinculadas fazem para a probabilidade, ou seja, a média sobre as probabilidades de todas as classificações possíveis formadas pela permutação das classificações das entidades vinculadas.

Infelizmente, essa abordagem pode aumentar significativamente a quantidade de computação necessária, mesmo em cenários em que relativamente poucas entidades estão vinculadas. Por exemplo, quando seis entidades estão empatadas para uma posição específica, a probabilidade de tal classificação contém um adicional $6! = 720$ termos. Além disso, se imaginarmos um cenário em que várias entidades estão vinculadas a vários níveis diferentes, fica claro que o cálculo da probabilidade logo se tornará inviável. Uma abordagem alternativa discutida por Breslow et al. (1974) usa uma aproximação da probabilidade exata, assumindo que cada uma das entidades vinculadas (dentro de uma determinada classificação) é preferida a todas as outras entidades classificadas na mesma posição ou inferiormente. Essa abordagem reduz significativamente a carga computacional (em comparação com o cálculo exato da probabilidade). No entanto, isso tem o custo de uma função de verossimilhança não exata. Baker e McHale (2015) discutem uma aproximação de verossimilhança adicional (mais exata), onde consideram a probabilidade de todas as classificações possíveis (formadas pela permutação das entidades vinculadas). No entanto, os parâmetros de habilidade para as entidades vinculadas são definidos como a média dos respectivos parâmetros de habilidade da entidade vinculada, com algum ruído aleatório adicional. Por exemplo, se houver t entidades vinculadas, então a probabilidade sob o θ ! As classificações possíveis seriam avaliadas de acordo com $\lambda_i = \mu + i$ ($i = 1, \dots, t$) onde $\mu = \sum i / t$. Detalhes completos deste esquema e detalhes de inferência em λ_i são discutidos no Apêndice A de Baker e McHale (2015).

Nossa solução para o problema dos laços é baseada em nossa solução MCMC para o problema da inferência. Essencialmente, em cada iteração do MCMC, simulamos uma classificação sem empates de uma distribuição uniforme em todas as classificações consistentes com as classificações com empates. Por exemplo, suponha que temos uma classificação $x = (1, 2, 3, 4, 5)$ onde as entidades 2 e 3 estão empatadas na segunda posição (indicada pela barra). Para incorporar essa classificação em nossa análise, let $x = (1, 2, 3, 4, 5)$ com probabilidade de 0,5 e, caso contrário, let $x = (1, 3, 2, 4, 5)$ em cada iteração de nosso esquema MCMC. Este esquema pode ser estendido trivialmente para incorporar classificações com empates envolvendo mais de duas entidades, por exemplo, classificações como $x = (\overline{1}, 2, 3, 4, 5)$. Além disso, podemos permitir a possibilidade de mais de um único conjunto de entidades vinculadas dentro de uma única classificação, ou seja, uma classificação da forma $x = (\overline{1}, \overline{2}, 3, \overline{4}, 5)$. Em todos os cenários, amostramos, uniformemente ao acaso, a partir da distribuição discreta em todas as classificações possíveis formadas por permutações das entidades amarradas. Este método para incorporar empates na análise de dados classificados é descrito mais detalhadamente por Glickman e Hennessy (2015).

2.2.5 Simulando dados do modelo de Plackett-Luce modificado

Tendo definido nossa probabilidade de Plackett-Luce modificada (2.2), estamos agora em posição de descrever o processo que nos permite simular dados sob este modelo. Começamos especificando valores "verdadeiros" para os parâmetros de habilidade, ou seja, $\lambda_k > 0$ para $k = 1, \dots, K$. Esses valores podem ser simulados a partir de uma distribuição apropriada, se desejado. Na Seção 2.2.2, discutimos como a probabilidade de Plackett-Luce é invariante à multiplicação escalar (estritamente positiva) dos parâmetros de habilidade; Deveremos, portanto, ser cautelosos ao especificar valores para nossos parâmetros de habilidade. É importante lembrar que os parâmetros de habilidade para cada entidade só podem ser comparados em relação uns aos outros. Portanto, a escolha de $\lambda = (3, 2, 5)$ especifica a distribuição equivalente sobre as classificações como a escolha de $\lambda = (0.3, 0.2, 0.5)$ e, de fato, a mesma distribuição que $C\lambda$ para qualquer $C > 0$.

Descrevemos o processo de geração de dados usando a conhecida representação de variável latente exponencial do modelo de Plackett-Luce (Diaconis (1988), Marden (1995)). Nesta representação, introduzimos variáveis latentes v_j , que são interpretadas como o tempo de chegada (latente) da entidade j (em um processo de Poisson homogêneo). Essas variáveis seguem distribuições independentes com o parâmetro de taxa λ_j . Os tempos de chegada latentes são então trivialmente convertidos em rankings, atribuindo o rank 1 à entidade que chegou primeiro, o rank 2 à entidade que chegou em segundo e assim por diante. Formalmente, a classificação completa x é gerada por meio do seguinte processo.

1. Amostra $v_j \sim \text{indep} \sim \text{Exp}(\lambda_j)$ para $j = 1, \dots, K$.

2. Defina $x_j = \arg\min_{q \in S_j} q$ onde $S_j = K \setminus \{x_1, \dots, x_{j-1}\}$ para $j = 1, \dots, K$.

Observamos que o processo descrito acima é projetado apenas para gerar classificações completas. Depois de obter classificações completas, é possível convertê-las em classificações parciais, top-M ou top-M parciais, conforme necessário. As classificações parciais são formadas removendo (da classificação completa) as entidades que não foram consideradas pelo classificador e preservando a ordem de preferência das entidades que permanecem. As classificações Top-M são obtidas trivialmente considerando apenas as primeiras posições M da classificação completa. As classificações parciais Top-M são obtidas a partir da classificação completa usando um processo de duas etapas. Começamos primeiro obtendo a classificação parcial correspondente (como acima) e, em seguida, a classificação necessária é dada tomando as primeiras posições M dentro da classificação parcial (recém-formada). Por exemplo, suponha que geramos a classificação completa $x = (1, 7, 3, 5, 2, 6, 8, 4)$ usando o processo descrito acima. A Tabela 2.1 mostra os rankings parciais dos 5 primeiros, parciais e 5 primeiros formados a partir desse ranking completo. As classificações parciais foram formadas assumindo que as entidades 2 e 8 não foram consideradas. A tabela também fornece o número de entidades dentro de cada classificação, n_i , e o número total de entidades que cada classificador considerou, K_i .

Tipo de classificação	Classificar				É	Para	Ui	K \ Ki
	12	34	56	78				
Completar	17	35	26	84	88	00		
Topo-5	17	35	25		8	{4, 6, 8}		0
Parcial	17	35	64		66		0	{2, 8}
Top 5 parcial	17	35	65		6	{4}	{2, 8}	

Tabela 2.1: Tipos de classificação

Os conjuntos Ui e $K \setminus Ki$ contendo as entidades não classificadas e as entidades que não foram consideradas pelo ranker i respectivamente também são fornecidos.

2.3 Inferência bayesiana

Agora descrevemos como a inferência bayesiana pode ser realizada usando classificações que seguem o modelo de Plackett-Luce modificado (2.2). Para fazer inferências significativas, esses dados devem conter várias classificações (n). Para formular nossa probabilidade de forma concisa, vamos $D = \{x_i\}_{i=1}^n$ ser a coleção de todas essas classificações. A probabilidade sob o modelo de Plackett-Luce é então dada como o produto das respectivas probabilidades de cada classificação e , portanto, assume a forma

$$\begin{aligned} p(D|\lambda) &= \prod_{i=1}^n \Pr(X_i = ha|\lambda) \\ &= \prod_{i=1}^n \prod_{j=1}^{\text{É}} \frac{\lambda \sum_{m=j}^n \lambda x_{im}}{\lambda m}. \end{aligned} \quad (2.3)$$

A inferência pode prosseguir usando uma abordagem de máxima verossimilhança, como o algoritmo MM (Hunter, 2004), para maximizar essa probabilidade e obter uma estimativa para nossos parâmetros de habilidade λ . Aqui, no entanto, adotamos a abordagem bayesiana para inferência e, portanto, definimos uma distribuição anterior adequada junto com um esquema de amostragem posterior apropriado.

2.3.1 Especificação prévia e variáveis latentes

A escolha de distribuições a priori adequadas é um problema bem discutido na literatura bayesiana (Bernardo e Smith, 1994). Aqui, nossa escolha de distribuição anterior é motivada principalmente pela conveniência matemática. No entanto, acreditamos que nossa escolha é suficientemente flexível para permitir que crenças prévias informativas sejam retratadas, se desejado. Os parâmetros de habilidade λ devem ser estritamente positivos e, portanto, parece sensato escolher distribuições Gamma prior independentes, ou seja, λ_k indep~ $Ga(ak, bk)$ para $k = 1, \dots, K$. Foi demonstrado que o

os parâmetros de taxa b_k não são identificáveis por verossimilhança (Caron e Doucet, 2012) e, portanto, deixamos $b_k = b = 1$, pois nossos parâmetros de habilidade são invariante à multiplicação escalar (estritamente positiva). Isso é feito amplamente na literatura. Portanto, nossa distribuição anterior para os parâmetros de habilidade é λ_k indep~ $Ga(ak, 1)$ para $k = 1, \dots, K$ e nossa especificação de modelo completa é

$$\begin{aligned} X|&\lambda \text{ indep} \sim PL(\lambda) & i = 1, \dots, n, \\ \lambda_k \text{ indep} &\sim Ga(ak, 1) & k = 1, \dots, K, \end{aligned}$$

onde $X|\lambda \sim PL(\lambda)$ denota que a classificação $X = x$ segue o modelo de Plackett-Luce com probabilidade definida em (2.2).

Sob esta escolha prévia, a forma da distribuição posterior é altamente fora do padrão e, portanto, adotamos a abordagem baseada em amostragem da cadeia de Markov Monte Carlo (MCMC). Pode ser desejável implementar um amostrador de Gibbs (se possível) em vez de um amostrador de Metropolis-Hastings, especialmente se isso resultar em maior eficiência de amostragem. Condicional a uma especificação prévia independente do Gamma, Caron e Doucet (2012) mostraram que, ao aplicar técnicas de aumento de dados, é possível facilitar uma atualização conjunta para os parâmetros de habilidade. Nossa espaço amostral é aumentado pela introdução de variáveis latentes apropriadas (coletivamente denotadas Z) que são interpretadas como os tempos hipotéticos (exponenciais) entre eventos das entidades nos processos homogêneos de Poisson referidos na Seção 2.2.5. Essas variáveis latentes são definidas por meio de sua distribuição condicional completa e são dadas por

$$z_{ij}|D, \lambda \text{ indep} \sim \text{Exp} \left(\sum_m \lambda_{xim} + \sum_{m \in U_i} \lambda_m \right), \quad (2.4)$$

pois $i = 1, \dots, n$ e $j = 1, \dots, n_i$.

À parte

Se primeiramente deixarmos Λ denotar a coleção de todos os parâmetros de habilidade, então, usando os resultados da Seção 1.4, podemos verificar que realmente obtemos a parte posterior desejada de nossos parâmetros de habilidade quando integramos essas variáveis latentes da parte posterior da articulação da seguinte forma

$$p(L|D) = \int Z \pi(\Lambda, Z|D)dZ$$

$$\propto \int Z p(Z|L, D)\pi(L|D)\pi(L)dZ$$

$$\begin{aligned}
 &= p(L) \frac{\int_0^\infty \cdots \int_0^\infty n_i}{\prod_{i=1}^n \prod_{j=1}^m \sum_{m=j}^n x_{ij}} \lambda x_{ij} \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda \\
 &\quad \times \text{XP} \left(\sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda \right) dZ \\
 &= p(L) \prod_{i=1}^n \prod_{j=1}^m \lambda x_{ij} \times \prod_{i=1}^n \prod_{j=1}^m \frac{\int_0^\infty \text{Exp} \left(\sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda \right) dZ}{\sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda} \text{Este} \\
 &= p(L) \prod_{i=1}^n \prod_{j=1}^m \lambda x_{ij} \times \prod_{i=1}^n \prod_{j=1}^m \frac{\prod_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda}{\prod_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda} \\
 &\quad \times \text{XP} \left(\sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda \right) dZ \\
 &= p(L) \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda x_{ij}}{\sum_{m=j}^n x_{ij} + \sum_{m \in U_i} 1} \prod_{i=1}^n \prod_{j=1}^m e^{0 -} e^{-\infty} \\
 &= \pi(L)\pi(D|L).
 \end{aligned}$$

Como veremos na seção a seguir, também podemos obter a distribuição condicional completa para Λ na forma fechada sob esta especificação de variável latente.

2.3.2 Distribuições condicionais completas

Nossa distribuição posterior é formada pela aplicação do Teorema de Bayes. Como aumentamos nosso espaço amostral, a distribuição posterior resultante é uma distribuição conjunta contendo as variáveis aleatórias latentes Z e os parâmetros de habilidade λ . Antes de iniciar nossa derivação, é útil primeiro construir a densidade de todas as grandezas estocásticas no modelo; isso é dado por

$$\begin{aligned}
 \pi(\lambda, D, Z) &= \pi(Z|D, \lambda)\pi(D|\lambda)\pi(\lambda) \\
 &= \prod_{i=1}^n \prod_{j=1}^m \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda \text{-eles} \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda \\
 &\quad \times \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda \sum_{m=j}^n x_{im} + \sum_{m \in U_i} 1}{\prod_{k=1}^K \frac{\lambda^{ak-1} k!}{\Gamma(ak)}} \\
 &= \prod_{k=1}^K \frac{\lambda^{ak-1} k!}{\Gamma(ak)} \times \prod_{i=1}^n \prod_{j=1}^m \frac{\text{Experi}^{-\text{eles}}}{\text{ência}} \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda . \quad (2.5)
 \end{aligned}$$

Agora podemos obter as distribuições condicionais completas (FCDs) construindo a distribuição condicional de cada quantidade desconhecida, dadas todas as outras quantidades estocásticas e os dados. Se começarmos com as variáveis latentes Z , deve ficar claro que

$$p(Z|D, l) \propto \prod_{i=1}^n \prod_{j=1}^m \frac{\text{Exp}^{-\sum_{m=j}^n \lambda x_{im}}}{\sum_{m \in U_i} \lambda} \prod_{m \in U_i} \lambda^m,$$

e assim a distribuição condicional completa do z_{ij} é como em (2.4) (por construção). A distribuição condicional completa para as quantidades aleatórias restantes, λ , é

$$\begin{aligned} p(l|D, Z) &\propto \prod_{k=1}^K \frac{\lambda^{ak}}{1k} e^{-\lambda} \prod_{i=1}^n \prod_{j=1}^m \frac{\text{Experi}^{-\sum_{m=j}^n \delta_{ij}(k)z_{ij}}}{\text{ência}} \prod_{m=j}^n \lambda x_{im} + \sum_{m \in U_i} \lambda^m \\ &= \prod_{k=1}^K \frac{\lambda^{ak+qk}}{1k} \text{Exp}^{-\sum_{i=1}^n \sum_{j=1}^m \delta_{ij}(k)z_{ij}} \lambda^k \end{aligned}$$

onde

$$qk = \sum_{i=1}^n I(k \in \{x_i\}),$$

e

$$\delta_{ij}(k) = I(k \in \{x_{ij}, \dots, x_{ni}\} \cup U_i), \quad (2.6)$$

são o número de vezes que a entidade k aparece em uma classificação e uma variável indicadora sobre o evento em que a entidade k recebe uma classificação não melhor que j na classificação i , respectivamente. Segue-se então que

$$\lambda_k|D, Z \text{ indep } \sim \text{Ga} \quad \lambda_k \sim \frac{1}{ak + qk + 1} \sum_{i=1}^n \sum_{j=1}^m \delta_{ij}(k)z_{ij},$$

para $k = 1, \dots, K$.

Como agora temos um conjunto completo de distribuições condicionais completas, estamos em posição de construir um esquema de amostragem para gerar realizações a partir de nossa distribuição posterior. Notamos que esta é uma modificação direta do amostrador de Gibbs de Caron e Doucet (2012), onde aqui a definição de $\delta_{ij}(k)$ em (2.6) mudou para que possamos lidar com classificações top-M. Este amostrador também é um pouco semelhante ao detalhado em Caron et al. (2014), mas aqui consideraremos um número finito (fixo) de entidades $K < \infty$.

2.3.3MCMC - Amostragem de Gibbs via variáveis latentes

Na Seção 2.3.2, derivamos um conjunto completo de distribuições condicionais completas assumindo a especificação de variável anterior e latente como na Seção 2.3.1. Agora podemos construir um esquema de Monte Carlo de Markovchain para gerar realizações a partir de nossa distribuição posterior: este é um amostrador de Gibbs. O esboço do algoritmo é o seguinte.

1. Inicialize o contador de iteração para $t = 1$. Inicialize o estado da cadeia, uma opção é a seguinte

- Para $k = 1, \dots, K$, amostra $\lambda_{(0)k} \sim \text{Ga}(ak, 1)$.
- Para $i = 1, \dots, n, j = 1, \dots, n_i$, amostra $\delta_{ij}(k) \sim \text{Exp} \left(\frac{\sum_{m=j}^n \lambda_{(0)m} x_{im} + L_{(0)}}{\sum_{m \in U_i} m} \right)$.

2. Obtenha novas realizações de $\lambda(t)$, $Z(t)$ de $\lambda(t-1)$, $Z(t-1)$ da seguinte forma:

- Para $k = 1, \dots, K$, amostra

$$\begin{aligned} \lambda_{(t)k} | D, Z(t-1) &\sim \text{Ga} \left(\frac{ak + QK}{\sum_{m=j}^n \lambda_{(t)m} x_{im} + \sum_{m \in U_i} m} \right) \\ &+ \frac{\sum_{m=j}^n \delta_{ij}(k) z_{im}}{\sum_{m \in U_i} m}. \end{aligned}$$

- Para $i = 1, \dots, n, j = 1, \dots, n_i$, amostra

$$z_{(t)ij} | D, \lambda(t) \sim \text{Exp} \left(\frac{\sum_{m=j}^n \lambda_{(t)m} x_{im}}{\sum_{m \in U_i} m} \right).$$

3. Redimensionar:

- Amostra $\lambda^\dagger \sim \text{Ga} \left(\frac{ak}{(K \sum_{k=1}^K k)} \right)$.
- Calcule $\Sigma = \frac{\lambda^\dagger}{K \sum_{k=1}^K k}$.
- Para $k = 1, \dots, K$, seja $\lambda_{(t)k} \rightarrow \lambda_{(t)k} \Sigma / \lambda^\dagger$.

4. Defina $t = t + 1$ e retorne à etapa 2.

Com a eficiência computacional em mente, notamos que q_k e $\delta_{ij}(k)$ dependem apenas dos dados e, portanto, permanecem constantes em todo o esquema de Monte Carlo da cadeia de Markov. Esses valores podem, portanto, ser calculados na etapa 1 e reutilizados em cada iteração. Também notamos que $q_k = n$ para todo k se nossos dados consistirem inteiramente em classificações completas.

O reescalonamento na etapa 3 decorre da discussão na Seção 2.2.3, onde foi observado que, sem a adição de uma etapa de reescalonamento, os esquemas MCMC para modelos Plackett-Luce podem sofrer de má mistura (Caron e Doucet, 2012). A ideia era redimensionar os parâmetros para que a distribuição posterior da soma de todos os parâmetros de habilidade K seja a mesma que sua distribuição anterior (já que os dados não são informativos sobre essa soma). Seja $\Lambda \dagger = \sum_{k=1}^K \lambda_k$ a soma de todos os parâmetros de habilidade K . Então, como $\lambda_k \text{ indep} \sim \text{Ga}(\alpha_k, 1)$ para $k = 1, \dots, K$, prior segue-se que o prior (induzido) para $\Lambda \dagger$ é uma distribuição $\text{Ga}(\sum_{k=1}^K \alpha_k, 1)$. O posterior para $\Lambda \dagger$ pode, portanto, ser mantido o mesmo que o anterior, extraíndo uma realização de $\Lambda \dagger$ a partir de uma distribuição $\text{Ga}(\sum_{k=1}^K \alpha_k, 1)$ e, em seguida, multiplicando os valores λ atuais (posteiros) por um fator de $\Lambda \dagger / \Sigma$, onde $\Sigma = \sum_{k=1}^K \lambda_k$ denota a soma atual (posterior) dos parâmetros de K -skill.

2.4 Estudo de simulação

Neste estudo, realizamos inferência bayesiana em dados que são simulados (gerados) sob o modelo de Plackett-Luce. O benefício de realizar inferência em dados simulados a partir do modelo verdadeiro é que conhecemos os valores dos parâmetros a partir dos quais esses dados foram gerados. Podemos, portanto, avaliar o desempenho do nosso modelo nessas condições antes de realizar a inferência em um cenário do mundo real. Aqui, consideraremos dois conjuntos de dados, ambos contendo (apenas) classificações completas de $K = 20$ entidades. O conjunto de todas as entidades é, portanto, dado por $K = \{1, \dots, 20\}$. Nossa primeiro conjunto de dados (Conjunto de dados 1) contém $n = 40$ classificações que foram simuladas usando o processo descrito na Seção 2.2.5 sujeito aos valores de parâmetro "verdeadeiros"

$$\lambda_1 = 20, \quad \lambda_k = \lambda_{k-1} - 1, \quad \text{para } k = 2, \dots, K,$$

isto é, $\lambda = (20, 19, \dots, 1)$. De acordo com essa especificação do parâmetro, as entidades indexadas por números menores são mais preferidas; Estas entidades são, portanto, mais propensas a figurar no início de uma classificação (a receber uma classificação numérica baixa) em comparação com as entidades indexadas por grandes números. Isso fica claro se considerarmos a noção de uma classificação ótima. A classificação ótima, denotada \hat{x} , é definida como a classificação tal que a probabilidade de Plackett-Luce é maximizada (condicional a alguns parâmetros de habilidade fixos). Matematicamente, tal classificação é dada por

$$\hat{x}|\lambda = \arg\max_{x \in SK} \Pr(X = x|\lambda), \quad (2.7)$$

de onde deve ficar claro que a classificação ideal é $\hat{x} = (1, 2, \dots, 20)$ sob nossa escolha atual de parâmetros de habilidade.

O segundo conjunto de dados (Conjunto de Dados 2) é composto por $n = 50$ classificações, as primeiras 40 das quais são as classificações do Conjunto de Dados 1 e as 10 classificações adicionais (numeradas de 41 a 50) são permutações aleatórias das K entidades. O conjunto de dados 2 é, portanto, uma extensão do conjunto de dados 1 e deixamos claro que as classificações comuns entre esses conjuntos de dados mantêm os mesmos rótulos. As permutações aleatórias devem ser chamadas de não informativas ou spamrankings. As classificações não informativas são geradas sob o processo usual de geração de dados (como na Seção 2.2.5) com $\lambda_k = c$ para $k = 1, \dots, K$, onde c é uma constante positiva arbitrária. Notamos que este método de simulação é equivalente a amostrar um elemento uniformemente ao acaso do conjunto de todas as permutações SK. O objetivo de analisar um conjunto de dados como este é investigar como nossa distribuição posterior (e, portanto, nossa inferência) é afetada por esses rankings de spam. Em certo sentido, esta é uma análise de sensibilidade para determinar o quanto robusto é o modelo de Plackett-Luce para a adição de classificações espúrias. As classificações usadas neste estudo podem ser encontradas nos apêndices; A Tabela B.1 contém o Conjunto de Dados 1 e as 10 classificações adicionais incluídas no Conjunto de Dados 2 são fornecidas na Tabela B.2.

Antes de podermos realizar inferência bayesiana sobre esses dados, devemos primeiro especificar distribuições anteriores adequadas. Para manter a conjugação (e, portanto, usar o amostrador de Gibbs descrito na Seção 2.3.3), escolhemos distribuições gama anteriores independentes para cada um de nossos parâmetros de habilidade, conforme discutido na Seção 2.3.1. Nesse cenário, também conhecemos os valores trueparameter a partir dos quais esses dados foram simulados, no entanto, optamos por realizar inferência assumindo que não temos conhecimento prévio sobre a força de cada entidade. Desejamos, portanto, uma especificação prévia de modo que cada classificação seja igualmente provável a priori. Isso é conseguido escolhendo $\lambda_k \text{ indep} \sim \text{Ga}(a, 1)$, isto é, definindo $\lambda_k = a$ para $k = 1, \dots, K$. Sem perda de generalidade, tomamos $a = 1$.

2.4.1 Análise posterior

Antes de começarmos nossa investigação sobre a distribuição posterior, daremos primeiro alguns detalhes computacionais. Nosso algoritmo MCMC foi inicializado usando um sorteio aleatório da distribuição anterior. Em seguida, passamos a realizar iterações de 11K; o primeiro 1K dos quais foi descartado como um período de queima. Isso nos deixou com 10K amostras (quase) não autocorrelacionadas de nossa distribuição posterior. O tempo computacional necessário para realizar inferências sobre esses dados é de aproximadamente 1 e 1,3 segundos para os conjuntos de dados 1 e 2, respectivamente. Este esquema de inferência é implementado em C e a computação é realizada em um único thread de uma CPU Intel Core i7-4790S (velocidade de clock de 3,20 GHz).

Nossa distribuição posterior é de alta dimensão, ou seja, $(n \times K) + K$ nessas análises. Avaliar a convergência e a mistura de cada parâmetro individual é, portanto, problemático; especialmente porque é fácil ver que a dimensão do nosso espaço de parâmetros aumentará

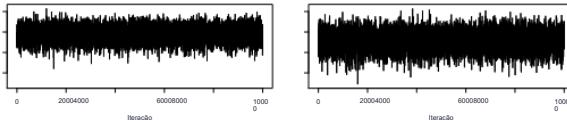


Figura 2.1: Gráficos de rastreamento da probabilidade de dados completos do log para os conjuntos de dados 1 e 2 da esquerda para a direita, respectivamente.

significativamente para conjuntos de dados maiores. Consequentemente, é desejável obter um método para avaliar de forma conveniente a convergência e mistura de uma cadeia de Markov para espaços amostrais de alta dimensão como este. Em vez de considerar cada variável aleatória, propomos considerar um resumo geral de nossas variáveis aleatórias, ou seja, a probabilidade de dados completos, $\pi(Z, D|\lambda) = \pi(Z|D, \lambda)\pi(D|\lambda)$. Gelman et al. (2014) defendem essa abordagem para avaliar a convergência (especialmente ao implementar modelos de mistura que consideramos no Capítulo 3). A Figura 2.1 mostra gráficos de rastreamento da semelhança de dados completos de log (após burn-in) para as análises dos conjuntos de dados 1 e 2. Observamos que nossas cadeias parecem estar se misturando bem e, além disso, cada cadeia parece estar amostrando de sua distribuição estacionária. A convergência (para a distribuição estacionária) também foi verificada inicializando várias cadeias em diferentes valores iniciais e verificando se as distribuições posteriores são equivalentes (até ruído estocástico) em todos os casos.

Dado que estamos convencidos de que nosso esquema MCMC está gerando realizações a partir da distribuição posterior (para ambas as análises), podemos agora começar nossa investigação sobre as inferências sobre nossos parâmetros de habilidade λ . Para facilitar a comparação, realizamos o redimensionamento (offline) deixando $\lambda_k \rightarrow \lambda_k/\lambda_{20}$ para $k = 1, \dots, 20$ em cada iteração de nossa saída MCMC. Como λ_{20} agora assume seu valor verdadeiro, podemos comparar nossos marginais posteriores (para os parâmetros de habilidade restantes) em relação aos valores verdadeiros escolhidos no início deste estudo, $\lambda = (20, 19, \dots, 1)$. A Figura 2.2 mostra os boxplots da distribuição marginal posterior para cada $\log \lambda_k$. As distribuições correspondentes às análises dos conjuntos de dados 1 e 2 são mostradas em branco e vermelho, respectivamente. As cruzes azuis denotam os valores verdadeiros a partir dos quais as classificações (informativas) foram simuladas. Também deixamos claro que, devido à nossa redimensionação, λ_{20} é constante e, portanto, omitido do gráfico. Além disso, os valores atípicos, definidos como as observações mais avançadas do que 1,5 vezes o intervalo interquartil (IQR) dos quartis superior e inferior, também foram omitidos.

Sob a análise do Conjunto de Dados 1, observamos que as distribuições marginais posteriores normalmente têm suporte significativo para os valores dos parâmetros verdadeiros. Isso não é particularmente surpreendente, dado que esses dados foram simulados a partir do modelo verdadeiro. Existem, é claro, algumas exceções; As distribuições marginais posteriores para as entidades 3, 5 e 16 mostram

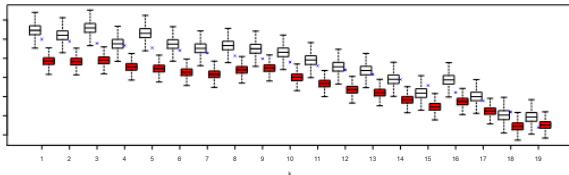


Figura 2.2: Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda 20 = 1$. As densidades em cada caso são mostradas em branco e vermelho para os conjuntos de dados 1 e 2, respectivamente. As cruzes azuis representam os valores verdadeiros a partir dos quais esses dados foram simulados (escala logarítmica).

suporte para valores maiores de λ do que os valores a partir dos quais esses dados foram gerados. Em outras palavras, a análise sugere que essas entidades são "mais fortes" (ou mais preferidas) do que sabemos que realmente são. Acreditamos que esta é uma característica desses dados e, se tivéssemos analisado um conjunto de dados maior consistindo em, digamos, $n = 1000$ classificações, esperaríamos que nossos posteriores marginais fossem um pouco mais focados nos valores verdadeiros. A partir da análise do conjunto de dados 1, podemos concluir que o modelo de Plackett-Luce (e nosso esquema de amostragem associado) é capaz de fazer inferências razoáveis a partir de um conjunto de dados classificados.

Agora consideraremos a análise do Conjunto de Dados 2, é interessante ver como a introdução de 10 classificações não informativas adicionais tem um efeito significativo em nossas distribuições marginais de posterior; ver figura 2.2. Nesta análise, as distribuições marginais posteriores geralmente mostram pouco suporte para os valores verdadeiros a partir dos quais os 40 rankings informativos foram simulados. No entanto, não vale a pena que, pelo menos para esta análise, o modelo ainda seja capaz de detectar uma tendência semelhante (descendente) na preferência das entidades quanto à análise do Conjunto de Dados 1. A tendência, no entanto, parece menos significativa e os rankings não informativos parecem ter induzido um efeito de "achatamento", ou seja, as diferenças (relativas) entre nossas distribuições posteriores marginais são menos convincentes. O resultado disso é que há mais incerteza (posterior) sobre a ordem de preferência das entidades. Isso talvez não seja surpreendente, dado que cada classificação não informativa expressa uma preferência aleatória e, portanto, nosso modelo leva isso em consideração aumentando a incerteza.

Muitas vezes, o objetivo / propósito da análise de dados classificados é obter a chamada classificação agregada. Uma classificação agregada é uma classificação única que resume as preferências em todas as classificações contidas em um determinado conjunto de dados; nessa medida, pode ser interpretado como uma classificação "média". Existem inúmeras maneiras de obter essa classificação. Aqui, escolhemos formar nossa classificação agregada ordenando as entidades com base em sua média marginal posterior. A classificação agregada, que denotamos x_{agg} , é, portanto, equivalente a

\hat{x}	L	Conjunto de dados 1		Conjunto de dados 2	
		xagg1	λ1	xagg2	λ2
1	20.00	3	27.47	3	11.67
2	19.00	1	25.83	1	11.44
3	18.00	5	23.90	2	11.29
4	17.00	2	22.66	4	9.84
5	16.00	4	18.11	9	9.54
6	15.00	6	17.98	5	9.41
7	14.00	8	17.31	8	9.11
8	13.00	7	16.12	6	8.55
9	12.00	9	15.91	7	8.10
10	11.00	10	14.50	10	7.48
11	10.00	11	11.84	11	6.36
12	9.00	12	9.95	12	5.42
13	8.00	13	9.00	13	5.02
14	7.00	14	7.17	14	4.17
15	6.00	16	7.08	16	3.98
16	5.00	15	4.99	15	3.47
17	4.00	17	4.58	17	3.10
18	3.00	18	2.81	19	2.16
19	2.00	19	2.68	18	2.08
20	1.00	20	1.00	20	1.00

Tabela 2.2: Classificações agregadas em nossa análise dos conjuntos de dados 1 e 2, juntamente com as médias posteriores correspondentes (denotadas λ). O valor de λ que foi usado para simular esses dados também é reproduzido para facilitar a comparação.

a classificação "ótima" dada λ onde $\lambda = (\lambda_1, \dots, \lambda_K)$ é o vetor parâmetro que contém as médias marginais posteriores para cada entidade. Formalmente $xagg = \hat{x}|\lambda$ onde $\hat{x}|\lambda$ é como em (2.7).

A Tabela 2.2 fornece as classificações agregadas para as análises dos conjuntos de dados 1 e 2 (denotados xagg1 e xagg2, respectivamente), juntamente com as médias posteriores marginais correspondentes (λ_1 e λ_2). Para facilitar a comparação, os valores verdadeiros a partir dos quais esses dados foram simulados, juntamente com a classificação "ideal" com base nos valores verdadeiros (\hat{x}), também são fornecidos. Observamos como as classificações agregadas em ambas as análises são coerentes com a classificação ideal; particularmente para as entidades que são classificadas pelo menos em décimo. A distância de Kendall-tau, $K_t(a, b)$, é uma medida de distância que é definida para quaisquer duas ordenações, a e b. O valor da distância de Kendall-tau é equivalente ao número de swaps adjacentes (classificação por bolha) que devem ser realizados para b de modo que se alinhe com a. É, portanto, uma distância útil a ser usada ao comparar classificações (Marden, 1995). Podemos calcular a distância de Kendall-tau entre a classificação ótima e as classificações agregadas em cada análise, dando $K_t(\hat{x}, xagg1) = 6$ e $K_t(\hat{x}, xagg2) = 10$. Portanto, concluímos que o

A análise do conjunto de dados 1 resulta em uma classificação agregada que é, em certo sentido, mais semelhante à verdadeira ordenação de preferência em comparação com os resumos equivalentes baseados no conjunto de dados 2.

Outra característica interessante das distribuições posteriores é que, embora as classificações agregadas sejam um pouco semelhantes, as médias marginais posteriores das entidades são significativamente diferentes dentro de cada análise; consulte a Tabela 2.2 e os boxplots mostrados na Figura 2.2. Em outras palavras, embora λ_1 e λ_2 definam distribuições diferentes sobre as classificações, a classificação modal é semelhante em cada vetor de parâmetro. No entanto, como as médias marginais posteriores (dos parâmetros de competência) são significativamente menos dispersas na análise do conjunto de dados 2, tal sugere um nível acrescido de incerteza quanto à posição de uma entidade. Para resumir essa incerteza, analisamos a probabilidade da classificação modal (x_{agg}), calculada usando as médias posteriores dos parâmetros de habilidade (λ), em relação à mesma probabilidade calculada sob a distribuição uniforme. A probabilidade de qualquer classificação (completa) x sob a distribuição uniforme é $1/K!$ e assim a quantidade de juros é $r = K! \Pr(X = x_{agg}|\lambda)$. A ideia é que grandes valores de r indicam que a classificação modal tem um suporte muito maior (posterior) do que uma classificação uniforme e, portanto, podemos concluir que a distribuição de classificação (definida por λ) é, em certo sentido, mais concentrada em torno da classificação modal. Em contraste, pequenos valores de r , portanto, indicam níveis crescentes de incerteza sobre a posição das entidades dentro do ranking. Observe que $r = 1$ corresponde à distribuição uniforme sobre as classificações e, portanto, neste caso, cada classificação é igualmente provável. Para as análises consideradas aqui, obtemos $r_1 = 103625$ e $r_2 = 123364$ para os conjuntos de dados 1 e 2 e, portanto, a probabilidade da classificação agregada na análise do conjunto de dados 1 é mais de cem mil vezes maior do que a probabilidade de uma classificação uniforme, com isso reduzindo para cerca de doze mil vezes para a análise do conjunto de dados 2. Segue-se que, embora as classificações agregadas em ambas as análises sejam semelhantes, há muito mais suporte posterior para a classificação agregada na análise do conjunto de dados 1. Isso novamente destaca como as classificações não informativas no Dataset 2 enfraqueceram nossas inferências sobre os parâmetros de habilidade.

Para concluir, este estudo mostrou como as classificações periféricas podem ter um efeito significativo em nossa distribuição posterior. Esta é uma característica do nosso modelo que não é particularmente desejável. Um modelo mais robusto seria aquele que fosse mais flexível e tivesse a capacidade de permitir a heterogeneidade potencial entre a força de classificações específicas. Na próxima seção, detalharemos uma extensão do modelo de Plackett-Luce que nos permite explicar essa heterogeneidade potencial.

2.5 O modelo de Plackett-Luce ponderado

Aludimos e, de fato, mostramos por meio de nosso estudo de simulação na Seção 2.4 que a consunção de nossos dados com permutações aleatórias pode ter um efeito significativo nas crenças posteriores. Esta é uma característica do modelo Plackett-Luce que não é particularmente desejável. Idealmente, gostaríamos de um modelo em que nossas inferências posteriores não fossem significativamente afetadas por algumas classificações / observações espúrias. Nesta seção, descrevemos uma nova extensão do modelo padrão de Plackett-Luce, que visa permitir uma heterogeneidade potencial entre a quantidade de informações contidas em classificações específicas.

Antes de delinearmos tal modelo, é natural reformular esse problema em termos de confiabilidade do ranker. A partir da forma da probabilidade de Plackett-Luce (2.3), fica claro como cada classificação faz uma contribuição igualmente ponderada para a probabilidade geral. Em outras palavras, o modelo considera cada classificador igualmente informativo / confiável. Esta é uma suposição bastante forte. É fácil conceber um cenário em que alguns classificadores são significativamente mais informados sobre as entidades que estão classificando em comparação com outros classificadores. No restante desta seção, propomos uma extensão do modelo de Plackett-Luce para que as classificações não contribuam mais igualmente para a probabilidade geral. O modelo resultante é aquele que permite que alguns rankings tenham uma influência maior sobre nossas inferências de parâmetros do que outros e, portanto, permite uma heterogeneidade potencial entre os frankers de habilidade.

Optamos por modelar essa heterogeneidade potencial entre os rankings por meio de um modelo misto com dois componentes: um para "rankings informativos" e outro para "rankings não informativos". O modelo de mistura é definido usando variáveis indicadoras binárias latentes $W_i \in \{0, 1\}$ for $i = 1, \dots, n$. Deixamos $W_i = 0$ se a classificação i não for informativa e $W_i = 1$ caso contrário. A probabilidade de uma determinada classificação (condicional aos parâmetros de habilidade e à variável indicadora binária latente) sob este "modelo de Plackett-Luce ponderado" é

$$\Pr(X_i = x_i | \lambda, W_i = w_i) = \prod_{j=1}^k \frac{\lambda w_i x_{ij}}{\sum_{m=1}^{n_i} \lambda w_i x_{im} + \sum_{m \in U_i} \lambda w_m}, \quad (2.8)$$

de onde, para uma classificação informativa ($w_i = 1$), recuperamos (2.2), a Plackett-Luce probabilidade padrão. No entanto, para uma classificação não informativa ($w_i = 0$) temos

$$\Pr(X_i = x_i | \lambda, W_i = 0) = \frac{\lambda \prod_{j=1}^{n_i} \sum_{m=1}^{n_i} \lambda x_{im} + \sum_{m \in U_i} \lambda}{\lambda \prod_{j=1}^{n_i} \sum_{m=1}^{n_i} \lambda x_{im} + \sum_{m \in U_i} \lambda} = \frac{(\lambda)^{(n_i)}}{(\lambda)^{(n_i)} + \sum_{m \in U_i} \lambda} = \frac{P(K_i)}{P(K_i) + P(U_i)} \quad (2.9)$$

isto é, o recíproco do número de permutações ordenadas de entidades n de um conjunto de tamanho K . A implicação de uma classificação ser considerada pouco informativa sob este modelo resulta em sua contribuição para a probabilidade de ser constante, independentemente dos valores dos parâmetros de habilidade - isso deve ficar claro, pois $\Pr(X_i = x_i | \lambda, w_i = 0)$ não depende de λ . Escrevemos este modelo de probabilidade usando a notação $X_i|\lambda, w_i \sim PLW(\lambda, w_i)$. À primeira vista, tomar w_i como binário pode parecer bastante restritivo, pois, com $w_i = 0$, o modelo de Plackett-Luce ponderado assume que todo o ranking x_i é completamente não informativo. Para permitir mais flexibilidade, consideramos permitir que cada classificador tenha uma variável binária diferente em cada posição de sua classificação, introduzindo indicadores binários dependentes de posição com $i = 1, \dots, n, j = 1, \dots, n$. No entanto, julgamos que isso introduziria muitos parâmetros e levaria a problemas de identificabilidade, principalmente quando consideramos agrupar classificadores e entidades nos Capítulos 3 e 4. Também consideramos permitir que os parâmetros de peso sejam contínuos, como no intervalo unitário. No entanto, isso também é problemático, pois não apenas torna os pesos ininterpretáveis (em termos de confiabilidade mais franca), mas também introduz problemas de identificabilidade. Para ver esse problema, é útil considerar um exemplo simples. Começamos observando que a probabilidade ponderada de Plackett-Luce é simplesmente a probabilidade padrão de Plackett-Luce avaliada nos parâmetros de habilidade Awi e, portanto, as probabilidades de diferentes classificações de entidades para o ranker i são descritas pela distribuição PL (λw_i). Agora considere dois rankers i e j e deixe o vetor de parâmetro de habilidade ser λ . Suponha que esses classificadores tenham parâmetros de peso $w_i = 0,8$ e $w_j = 0,5$ e, portanto, aqui as distribuições de classificação são $PL(\lambda w_i = 0,8)$ e $PL(\lambda w_j = 0,5)$. No entanto, distribuições de classificação equivalentes (e, portanto, a mesma probabilidade ponderada de Plackett-Luce) podem ser obtidas usando $\lambda^* = \lambda 0,8$, com $w_i = 1$ e $w_j = 0,5 / 0,8 = 0,625$. Este exemplo simples mostra um problema de identificabilidade para (λ, w) . Além disso, o valor de w_i não é significativo, pois o classificador i seria classificado como bastante informativo no cenário λ , mas extremamente informativo no cenário λ^* . Esses problemas não ocorrem se escolhermos w_i para ser binário. Além disso, essa escolha tem o benefício de que $w_i = 1$ recupera a distribuição padrão de Plackett-Luce e $w_i = 0$ é significativo na medida em que representa uma distribuição de classificação uniforme.

É igualmente evidente que o modelo de Plackett-Luce ponderado (WPL) (2.8) difere do modelo proposto por Benter (1994). O modelo WPL considera pesos específicos do classificador (w_i , $i = 1, \dots, n$) para permitir a heterogeneidade potencial entre as habilidades dos classificadores, enquanto o modelo Benter considera os parâmetros do "peso" dependentes da posição (w_j , para $j = 1, \dots, K$) que são comuns a todos os classificadores e representam a "importância" de cada estágio no processo de classificação. Em teoria, também seria possível introduzir ambos os conjuntos de parâmetros de peso e, portanto, considerar um modelo de Benter ponderado. No entanto, é provável que isso dê origem a problemas identificáveis e, portanto, esse modelo não é considerado mais nesta tese.

2.5.1 Simulando dados do modelo de Plackett-Luce ponderado

Nosso modelo de probabilidade subjacente mudou como resultado da introdução de variáveis indicadoras binárias para refletir a capacidade (latente) dos classificadores. Daqui resulta que o mecanismo de geração de dados também deve ser adaptado para ter em conta esta (potencial) heterogeneidade adicional. Generalizamos a conhecida representação de variável latente exponencial do modelo padrão Plackett-Luce e, condicionada aos parâmetros de habilidade λ e à variável indicadora $w \in \{0, 1\}$, os tempos de chegada latentes correspondentes sob o modelo Weighted Plackett-Luce

$$v_j \text{ indep} \sim \text{Exp}(\lambda w_j), \quad (2.10)$$

para $j = 1, \dots, K$.

Um ranking completo x pode então ser simulado sob o modelo de Plackett-Luce ponderado, conforme descrito na Seção 2.2.5, onde v_j é extraído de (2.10) na etapa 1. Também observamos que, embora o conjunto de dados 2 do estudo de simulação anterior (Seção 2.4) tenha sido simulado sob o modelo padrão de Plackett-Luce, esses dados seguem a distribuição subjacente definida pelo modelo ponderado de Plackett-Luce com $\lambda = (20, 19, \dots, 1)$ e $w_i = 1$ para $i = 1, \dots, 40$; $w_i = 0$ para $i = 41, \dots, 50$.

2.6 Inferência bayesiana

A inferência bayesiana para os parâmetros do modelo ponderado de Plackett-Luce é obtida da maneira semelhante à do modelo padrão de Plackett-Luce (ver Seção 2.3). No entanto, aqui assumimos que os indicadores binários latentes são desconhecidos (observe que assumimos essencialmente $w_i = 1 \forall i$ para o modelo padrão de Plackett-Luce). Segue-se que temos quantidades aleatórias adicionais no modelo, ou seja, $w = (w_i)_{i=1}^n$, que também devem ser inferidas a partir dos dados. Tal como acontece com o modelo padrão de Plackett-Luce, a probabilidade é formada tomando o produto das respectivas probabilidades para cada uma das n classificações, e assim

$$\pi(D|\lambda, w) = \prod_{i=1}^n \prod_{j=1}^K \frac{\sum_{m=j}^K \lambda w_i x_{im}}{\sum_{m \in U_i}}, \quad (2.11)$$

onde $D = (x_{ij})_{i,j=1}^n$ denota a coleção de todas as classificações. Mais uma vez, como no modelo padrão de Plackett-Luce, uma abordagem de máxima verossimilhança poderia ser implementada, se desejado. Para este modelo, é claro que teríamos que fazer adaptações apropriadas aos esquemas de otimização dentro da literatura para permitir nossas variáveis indicadoras adicionais latentes. No entanto, aqui procedemos dentro da estrutura bayesiana, definimos uma distribuição anterior adequada e aplicamos o Teorema de Bayes para obter nossa distribuição posterior.

2.6.1 Especificação prévia e variáveis latentes

Optamos por usar a mesma especificação prévia para nossos parâmetros de habilidade que na Seção 2.3.1 pelas razões discutidas nela. No entanto, para este modelo, também devemos especificar uma distribuição prévia adequada sobre nossas variáveis indicadoras binárias latentes w . Como $w_i \in \{0, 1\}$ nós escolhemos $\text{indep} \sim \text{Bern}(p_i)$ onde $p_i \in (0, 1)$ é a probabilidade de que a classificação i seja informativa a priori. Observe que omitimos a escolha de $p_i = 0$, pois isso implica que $\Pr(w_i = 0 | D) = 1$; de onde a probabilidade de classificação i é constante independentemente de λ . A contribuição para a verossimilhança da classificação i pode, portanto, ser absorvida pela constante de proporcionalidade ao aplicar o Teorema de Bayes. Segue-se que, se realmente acreditamos que uma classificação não é informativa (com probabilidade 1), é suficiente simplesmente omitir essa classificação de nossa análise. Se definir probabilidades p_i a priori não é desejado, então uma estrutura de modelo hierárquica poderia ser construída; dado que p_i é uma probabilidade, uma distribuição Beta seria uma escolha sensata, no entanto, isso não é considerado aqui. Nossa especificação completa do modelo é, portanto,

$$\begin{aligned} X_{ij} | \lambda, w &\text{ indep} \sim \text{PLW}(\lambda, w_i) & i = 1, \dots, n, \\ \lambda_k &\text{ indep} \sim G\alpha(ak, 1) & k = 1, \dots, K, \\ w_i &\text{ indep} \sim \text{Berna}(p_i) & i = 1, \dots, n. \end{aligned}$$

Tal como acontece com o modelo padrão de Plackett-Luce, é possível aumentar nosso espaço amostral para que uma atualização conjunta para nossos parâmetros de habilidade λ possa ser alcançada. A forma da probabilidade mudou desde que introduzimos nossos pesos de classificação latentes e , portanto, devemos também modificar nossa especificação de variável latente. As variáveis latentes apropriadas a serem introduzidas para este modelo são

$$z_{ij} | D, \lambda, w \text{ indep} \sim \frac{\text{Exp}}{\sum_{m=j}^n \sum_{m \in U_i} w_m}, \quad \text{Vou levá-lo + } \sum_{m \in U_i} w_m, \quad (2.12)$$

pois $i = 1, \dots, n$ e $j = 1, \dots, n_i$.

2.6.2 Distribuições condicionais completas

Mais uma vez, passamos a derivar as distribuições condicionais completas para cada uma de nossas variáveis aleatórias, construindo primeiro a densidade de todas as quantidades estocásticas. Isso é dado por

$$\begin{aligned}
 p(\lambda, D, Z, w) &= \pi(Z|D, \lambda, w)\pi(D|\lambda, w)\pi(\lambda)\pi(w) \\
 &= \prod_{i=1}^n \prod_{j=1}^n \sum_{m=j}^n \frac{\text{Vou levá-lo} + \sum_m \lambda w}{\text{im}} \cdot \frac{\text{-eles} \sum_{m=j}^n \text{Vou levá-lo} + \sum_m \lambda w}{\text{im}} \\
 &\quad \times \prod_{i=1}^n \prod_{j=1}^n \frac{\lambda w_{ij}}{\lambda w_{im}} \times \prod_{i=1}^n \frac{\lambda \lambda k - 1 - k - e^{-\lambda}}{\lambda k \Gamma(\lambda k)} \times \prod_{i=1}^n p_{wi} (1 - p_i)^{1 - w_i} \\
 &= \prod_{k=1}^K \frac{\lambda \lambda k - 1 - k - e^{-\lambda}}{\lambda k \Gamma(\lambda k)} \times \prod_{i=1}^n p_{wi} (1 - p_i)^{1 - w_i} \\
 &\quad \times \prod_{i=1}^n \prod_{j=1}^n \frac{\lambda w_{ij}}{\exp} \cdot \frac{\text{-eles} \sum_{m=j}^n \text{Vou levá-lo} + \sum_m \lambda w}{\text{im}} \quad . \quad (2.13)
 \end{aligned}$$

Sem surpresa, dado que as variáveis latentes introduzidas são definidas por meio de suas distribuições condicionais completas, observamos a partir de (2.13) que o FCD para o z_{ij} é como em (2.12) para $i = 1, \dots, n, j = 1, \dots, n_i$. A distribuição condicional completa para λ é

$$\begin{aligned}
 p(\lambda|D, Z, w) &\propto \prod_{k=1}^K \frac{\lambda \lambda k - e^{-\lambda}}{\lambda k} \prod_{i=1}^n \prod_{j=1}^n \frac{\lambda w_{ij}}{\exp} \cdot \frac{\text{-eles} \sum_{m=j}^n \text{Vou levá-lo} + \sum_m \lambda w}{\text{im}} \\
 &= \prod_{k=1}^K \frac{\lambda \lambda k - e^{-\lambda}}{\lambda k} \cdot \frac{\exp}{\lambda k - 1} \cdot \frac{1}{\lambda k} + \sum_{i=1}^n \sum_{j=1}^n \frac{W \sum_k \delta_{ij}(k) z_{ij}}{\lambda k} ,
 \end{aligned}$$

onde

$$\tilde{q}_k = \sum_{i=1}^n w_i I(k \in \{x_i\}),$$

é o número de vezes que a entidade k aparece em um ranking informativo. Como em nossa análise anterior, $\delta_{ij}(k)$ é uma variável indicadora sobre o evento em que a entidade k recebe uma classificação não melhor que j na classificação i e é dada por (2.6). A partir disso, podemos obter os FCDs para nossos parâmetros de habilidade como

$$\lambda_k|D, Z, w \text{ indep} \sim \frac{\lambda}{\lambda k - 1} \tilde{q}_k, \quad 1 + \sum_{i=1}^n \sum_{j=1}^n \frac{W \sum_k \delta_{ij}(k) z_{ij}}{\lambda k} , \quad \text{para } k = 1, \dots, K.$$

O modelo Weighted Plackett-Luce também contém as variáveis indicadoras binárias adicionais. Essas variáveis seguem uma distribuição discreta com dois componentes. Seja $w-i = (w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n)$ denotar o vetor que contém todos os valores do indicador, exceto aquele associado ao ranker i . Segue-se que

$$\Pr(w_i = 1 | D, \lambda, Z, w-i) \propto \Pr(w_i = 1) \pi(Z|w_i = 1, D, \lambda, w-i) \Pr(D|w_i = 1, \lambda, w-i)$$

$$= p_i \prod_{i=1}^n \prod_{j=1}^n \text{Experi} \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i \setminus \{i\}} \lambda$$

$$\propto p_i \prod_{j=1}^n \text{Experi} \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i \setminus \{i\}} \lambda$$

e

$$\Pr(w_i = 0 | D, \lambda, Z, w-i) \propto \Pr(w_i = 0) \pi(Z|w_i = 0, D, \lambda, w-i) \Pr(D|w_i = 0, \lambda, w-i)$$

$$= (1 - p_i) \prod_{i=1}^n \prod_{j=1}^n \lambda x_{ij} \exp \sum_{m=j}^n \lambda x_{im} + \sum_{m \in U_i \setminus \{i\}} \lambda$$

$$\propto (1 - p_i) \prod_{j=1}^n \text{Exp} \sum_{m=j}^n 1 + \sum_{m \in U_i \setminus \{i\}} 1$$

$$= (1 - p_i) \prod_{j=1}^n \exp \{-\text{side}(K_i - j + 1)\}.$$

Portanto, a distribuição condicional completa (discreta) para w_i é

$$w_{ij}|D, \lambda, Z \text{ indep} \sim \text{Bern}(p_i),$$

onde

$$p_i = \frac{\Pr(w_i = 1 | D, \lambda, Z, w-i) \Pr(w_i = 1 | D, \lambda, Z, w-i)}{\Pr(w_i = 0 | D, \lambda, Z, w-i)}, \quad (2.14)$$

é a probabilidade de que a classificação i seja informativa (dadas as outras quantidades).

2.6.3 MCMC - Amostragem de Gibbs via variáveis latentes

Na seção anterior, derivamos um conjunto completo de distribuições condicionais completas para cada quantidade aleatória dentro de nosso modelo. Tal como acontece com o modelo padrão de Plackett-Luce, podemos implementar um amostrador de Gibbs para gerar realizações a partir de nossa distribuição posterior. Para brevidade e clareza de leitura, omitimos o contador de iteração t ao delinejar este algoritmo. No entanto, as atualizações prosseguem da mesma maneira que o algoritmo descrito

na seção 2.3.3. O algoritmo procede da seguinte forma.

1. Inicialize o estado da cadeia. Uma possibilidade é a seguinte.

- Para $k = 1, \dots, K$, amostra λ_k indep~ $Ga(\alpha_k)$,
- 1).• Para $i = 1, \dots, n$, amostra w_i indep~ $Bern(p_i)$.
- Para $i = 1, \dots, n$, $j = 1, \dots, n_i$, amostra

$$\text{z}_{ij} | \lambda, w \text{ indep} \sim \frac{\lambda}{\sum_{m=j}^n w_m} \text{ Vou levá-lo + } \sum_{m \in U_i} \lambda_m w_m .$$

2. Atualize o estado da cadeia executando repetidamente as etapas a seguir.

- Para $k = 1, \dots, K$, amostra

$$\lambda_k | D, Z, w \text{ indep} \sim \frac{\lambda_k + q_k}{1 + \sum_{i=1}^n w_i \sum_{j=1}^{n_i} \delta_{ij}(k) z_{ij}} .$$

- Para $i = 1, \dots, n$, $j = 1, \dots, n_i$, amostra

$$z_{ij} | D, \lambda, w \text{ indep} \sim \frac{\lambda}{\sum_{m=j}^n w_m} \text{ Vou levá-lo + } \sum_{m \in U_i} \lambda_m w_m .$$

- Para $i = 1, \dots, n$, amostra

$$w_i | D, \lambda, Z \text{ indep} \sim Bern(p_i),$$

onde p_i é dado por (2.14).

- Redimensionar:– Amostra

$$\lambda \dagger \sim Ga(K \sum k = 1, \lambda_k, 1) .$$

$$- \text{ Calcule } \Sigma = \sum_{k=1}^K \lambda_k .$$

– Para $k = 1, \dots, K$, seja $\lambda_k \rightarrow \lambda_k L \dagger / \Sigma$.

Notamos aqui que, ao contrário da análise padrão de Plackett-Luce,⁷ q_k não é mais uma função apenas dos dados. O valor agora depende das variáveis aleatórias w e, portanto, o cálculo de q_k é necessário em cada iteração do nosso esquema MCMC; Especificamente após novas realizações de W terem sido feitas. Por outro lado, os indicadores $\delta_{ij}(k)$ continuam a ser apenas uma função dos dados e, por conseguinte, podem ser calculados no passo 1 e utilizados em todo o esquema MCMC.

2.7 Estudo de simulação

Neste estudo, realizamos inferência bayesiana assumindo que nossas classificações seguem o modelo WeightedPlackett-Luce. Principalmente, nos concentramos em saber se esse modelo é capaz de identificar (corretamente) classificações não informativas contidas em um determinado conjunto de dados. Em nosso estudo anterior (Seção 2.4), observamos como nossas crenças posteriores foram significativamente alteradas pela contenção de nosso conjunto de dados com permutações aleatórias. Com isso em mente, será interessante ver como nosso modelo de Plackett-Luce ponderado se comporta, uma vez que tem a flexibilidade de reduzir a influência de classificações específicas em nossa inferência de parâmetros.

Dados esses objetivos, revisitamos um conjunto de dados de nosso estudo de simulação anterior no modelo padrãoPlackett-Luce; ou seja, o conjunto de dados 2 da Seção 2.4. Lembre-se de que o conjunto de dados 2 contém = 50 classificações completas de $K = 20$ entidades. Quarenta dos rankings (1–40) são informativos e, portanto, seguem o modelo Plackett-Luce ponderado sujeito a $w_i = 1$. As classificações restantes (41–50) são classificações não informativas, ou seja, permutações aleatórias das Kentidades. Considera-se que seguem o modelo de Plackett-Luce ponderado sujeito a $w_i = 0$. Lembramos também ao leitor que a classificação ótima, ou seja, a classificação que maximiza a probabilidade de Plackett-Luce, é $\mathbf{x} = (1, 2, \dots, 20)$ sob os parâmetros usados para simular esses dados. Observe que, estritamente falando, esses dados não foram gerados a partir do processo de geração de dados para o modelo de Plackett-Luce ponderado. No entanto, conforme discutido na Seção 2.5.1, o processo usado para gerar o Conjunto de Dados 2 é equivalente ao uso do processo de geração de dados para o modelo de Plackett-Luce ponderado sujeito à escolha de $\lambda = (20, 19, \dots, 1)$, $w_i = 1$ para $i = 1, \dots, 40$ e $w_i = 0$ para $i = 41, \dots, 50$.

Neste estudo, optamos por usar a especificação anterior, conforme descrito na Seção 2.6.1, ou seja, distribuições anteriores Gamma independentes em nossos parâmetros de habilidade e distribuições independentes de Bernoulli em nossos pesos de classificação latentes. Além disso, também desejamos fazer a mesma suposição em relação aos nossos parâmetros de habilidade que em nossa análise anterior, ou seja, que cada classificação é igualmente provável a priori. Dado isso, deixamos $a_k = a = 1$ para $k = 1, \dots, K$, que dá $A_k \text{ indep} \sim Ga(1, 1)$. Vamos agora considerar duas análises separadas desses dados, cada uma das quais tem uma especificação prévia alternativa em nossas variáveis indicadoras binárias latentes. Na Análise 1, optamos por assumir que cada classificação tem a mesma probabilidade de ser informativa, pois não é informativa; portanto, $p_i = 0.5$. Observamos que essa escolha não está de acordo com esses dados, pois a verdadeira proporção de classificações informativas dentro desse conjunto de dados é $40/50 = 0.8$. Assim, deixamos $p_i = 0.8$ a priori na Análise 2.

2.7.1 Análise posterior

As realizações de nossa distribuição posterior (para ambas as análises) são obtidas implementando o amostrador de Gibbs detalhado na Seção 2.6.3. Para cada análise, a cadeia de Markov

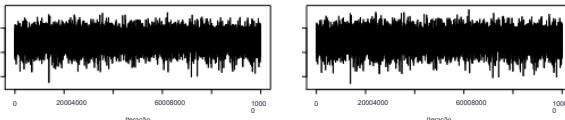


Figura 2.3: Gráficos de rastreamento da probabilidade de dados completos logarítmicos para as Análises 1 e 2 ($\pi = 0,5, 0,8$) da esquerda para a direita, respectivamente.

é inicializado em um sorteio aleatório da distribuição anterior. Cada cadeia funciona para 11K iterações; os primeiros 1K dos quais são descartados como burn-in. Isso resulta em realizações (quase) não autocorrelacionadas de 10K de nossa distribuição posterior. Como na análise anterior, nosso esquema de inferência é implementado em C e a computação é realizada em um único threadde uma CPU Intel Core i7-4790S (velocidade de clock de 3,20 GHz). O tempo computacional necessário para realizar a inferência sobre esses dados é de aproximadamente 1,9 segundos para ambas as análises. A mistura de nossas cadeias MCMC é avaliada inspecionando os gráficos de rastreamento da probabilidade de dados completos do log; ver figura 2.3. A partir disso, observamos que nossas cadeias parecem estar se misturando bem no espaço amostral das quantidades aleatórias. A convergência foi verificada inicializando cada cadeia em vários valores iniciais e verificando se as realizações posteriores são equivalentes (até ruído estocástico) em todos os casos.

Começamos nossa investigação sobre a distribuição posterior resumindo as distribuições marginalposterior para cada um de nossos pesos de classificação w_i . A Figura 2.4 mostra a probabilidade posterior de que cada classificador seja informativo, ou seja, $\Pr(w_i = 1|D)$. Essa probabilidade é obtida tomando a média posterior de π e não simplesmente a expectativa posterior de w_i . Nossa probabilidade é, portanto, resultado de um estimador Rao-Blackwellizado que normalmente nos fornece uma estimativa melhor do que simplesmente tomar a média (posterior) de w_i (Casella e Robert, 1996). A distância de Kendall-tau, $K_T(x, \bar{x})$, entre cada uma de nossas classificações simuladas e a classificação ótima também é dada na Figura 2.4. Observamos que, com exceção da classificação 42, todas as classificações não informativas (41-50) recebem uma probabilidade posterior menor de serem informativas do que a especificada a priori em cada análise respectiva. Como esperado, essas classificações também são aquelas que normalmente têm uma distância maior da classificação ideal. É encorajador ver que os rankings que são informativos (1-40) quase sempre obtêm grandes probabilidades posteriores de serem informativos, particularmente na Análise 2, onde $\pi = 0,8$. Dito isso, as probabilidades posteriores para o classificador (informativo) 8 não estão próximas de 1 em nenhuma das análises. Uma inspeção mais detalhada dessa classificação revela que ela é um tanto atípica desse conjunto de dados: as entidades 2 e 4 aparecem nas 5 últimas posições e as entidades 13, 11 e 10 aparecem nas 5 primeiras posições. Esses recursos estão um pouco em desacordo com os verdadeiros valores de parâmetros dos quais esses dados foram

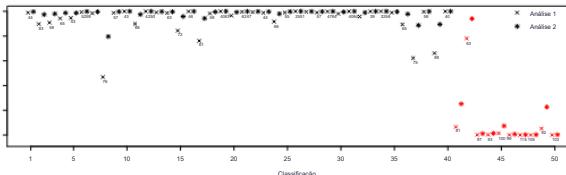


Figura 2.4: $\Pr(w_i = 1|D)$ – Probabilidade posterior de que a classificação i seja informativa em cada análise (Análise 1: $\pi_1 = 0,5$, Análise 2: $\pi_2 = 0,8$). As classificações que são permutações aleatórias (41–50) são mostradas em vermelho. Os números mostrados denotam a distância de Kendall-tau entre cada classificação e "x".

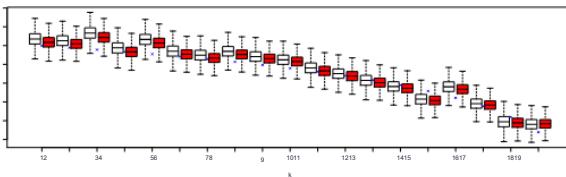


Figura 2.5: Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{20} = 1$. Os boxplots em cada caso são mostrados em branco e vermelho para as Análises 1 e 2 ($\pi_1 = 0,5, 0,8$, respectivamente). As cruzes azuis representam os valores verdadeiros, λ_k , a partir dos quais esses dados foram simulados.

simulado. Portanto, parece que o modelo ponderado de Plackett-Luce é capaz de identificar corretamente as classificações que são um tanto espúrias e reduzir sua contribuição para a probabilidade de forma adequada.

A Figura 2.5 mostra os boxplots da distribuição marginal posterior de $\log \lambda_k$ para as Análises 1 e 2 (mostradas em branco e vermelho, respectivamente). Como em nossas análises sob o modelo padrão Plackett-Luce, redimensionamos nossas realizações posteriores para que λ_{20} assuma seu valor verdadeiro, ou seja, defina $\lambda_k \rightarrow \lambda_k/\lambda_{20}$. É claro que as distribuições marginais posteriores são comparáveis entre as duas análises; Isso não é uma surpresa, uma vez que as probabilidades posteriores de cada classificador ser informativo são semelhantes em ambas as especificações anteriores. Além disso, a ordem de preferência das entidades também foi identificada pelo modelo; isso é claramente visto através da tendência de queda na Figura 2.5 à medida que k aumenta.

Curiosamente, quando comparamos as distribuições marginais posteriores dos parâmetros de habilidade sob o modelo ponderado de Plackett-Luce com aquelas sob a análise do conjunto de dados 1, assumindo o modelo padrão de Plackett-Luce, vemos semelhanças significativas; ver figura 2.6.

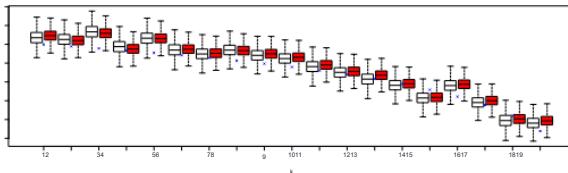


Figura 2.6: Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{20} = 1$. Os boxplots em cada caso são mostrados em branco para a Análise 1 em nosso modelo Plackett-Luce ponderado e em vermelho para a análise do Conjunto de Dados 1 sob o modelo padrão de Plackett-Luce. As cruzes azuis representam os valores verdadeiros, λ_k , a partir dos quais esses dados foram simulados.

É evidente que, ao adotar o modelo de Plackett-Luce ponderado, nossa distribuição posterior é significativamente menos afetada pela incorporação de classificações não informativas em nosso conjunto de dados. Isto se torna um pouco mais claro se considerarmos os rankings agregados posteriores. A Tabela 2.3 fornece as classificações agregadas para as análises do Conjunto de Dados 2 em ambas as especificações anteriores para nosso modelo de Plackett-Luce ponderado; denotados $xagg3$ e $xagg4$, respectivamente. As médias marginais posteriores correspondentes, λ_3 e λ_4 , sobre as quais esses agregados são formados, também são fornecidas. Para facilitar a comparação, as classificações agregadas ($xagg1$, $xagg2$) e as médias marginais posteriores correspondentes (λ_1, λ_2) para as análises dos conjuntos de dados 1 e 2 sob o modelo padrão de Plackett-Luce também são reproduzidas aqui. Os valores verdadeiros a partir dos quais esses dados foram simulados, juntamente com a classificação "ótima" com base nos valores verdadeiros, também aparecem na Tabela 2.3. Considerando nossas duas análises separadas sob o modelo Weighted Plackett-Luce, observamos como as classificações agregadas são equivalentes em ambas as especificações anteriores. Além disso, essa classificação agregada também é equivalente àquela formada a partir da análise do conjunto de dados 1 sob o modelo padrão de Plackett-Luce, ou seja, a análise sem classificações de spam. Isso é particularmente interessante, pois mostra como o efeito das classificações de spam em nossas crenças posteriores foi negado pela introdução de nossos indicadores binários na probabilidade de Plackett-Luce.

Na Seção 2.4.1, observamos que pode ser útil observar a probabilidade relativa da classificação agregada em comparação com uma classificação uniforme ($r = K! \Pr(X = xagg|\lambda)$) para cada análise. Essa quantidade fornece informações sobre o quanto concentrada é a distribuição de classificação em torno da classificação modal (posterior). A Tabela 2.3 fornece o r_i para cada análise considerada aqui, juntamente com aqueles para as análises dos conjuntos de dados 1 e 2 sob o modelo padrão Plackett-Luce. Observamos que os valores do r_i em cada uma das análises de WeightedPlackett-Luce são semelhantes aos da análise do Conjunto de Dados 1 sob o modelo PL padrão. Mais uma vez, isso destaca como a adoção do modelo de Plackett-Luce ponderado

x̂	L	PLW				PL			
		pi = 0,5		pi = 0,8		Conjunto de dados 2		Conjunto de dados 1	
		xagg3	λ3	xagg4	λ4	xagg1	λ1	xagg2	λ2
1	20.00	3	28.83	3	25.55	3	27.47	3	11.67
2	19.00	1	24.60	1	22.41	1	25.83	1	11.44
3	18.00	5	24.19	5	21.96	5	23.90	2	11.29
4	17.00	2	23.40	2	21.43	2	22.66	4	9.84
5	16.00	4	19.35	4	17.32	4	18.11	9	9.54
6	15.00	6	17.73	6	16.19	6	17.98	5	9.41
7	14.00	8	17.62	8	16.17	8	17.31	8	9.11
8	13.00	7	15.93	7	14.70	7	16.12	6	8.55
9	12.00	9	15.31	9	14.46	9	15.91	7	8.10
10	11.00	10	14.06	10	13.40	10	14.50	10	7.48
11	10.00	11	11.37	11	10.44	11	11.84	11	6.36
12	9.00	12	9.71	12	9.08	12	9.95	12	5.42
13	8.00	13	8.16	13	7.63	13	9.00	13	5.02
14	7.00	14	6.91	14	6.57	14	7.17	14	4.17
15	6.00	15	6.88	16	6.42	16	7.08	16	3.98
16	5.00	15	4.93	15	4.74	15	4.99	15	3.47
17	4.00	17	4.38	17	4.19	17	4.58	17	3.10
18	3.00	18	2.71	18	2.62	18	2.81	19	2.16
19	2.00	19	2.52	19	2.54	19	2.68	18	2.08
20	1.00	20	1.00	20	1.00	20	1.00	20	1.00
Re		129768	93809	103625	12364				

Tabela 2.3: Classificações agregadas sob o modelo Weighted Plackett-Luce para a análise do conjunto de dados 2 (para ambas as análises $\pi = 0,5, 0,8$) juntamente com as médias posteriores correspondentes. Para facilitar a comparação, os resultados da Tabela 2.2 (análises padrão de Plackett-Luce) também são fornecidos. A tabela também contém a probabilidade relativa das classificações agregadas em comparação com uma classificação uniforme, $r_i = K! \Pr(X = x_{aggi} | \lambda_i)$.

resulta em nossa distribuição posterior sendo significativamente menos afetada pela inclusão de classificações não informativas no conjunto de dados.

2.8 Resumo

Este capítulo delineou o modelo de Plackett-Luce e discutiu as suposições subjacentes, juntamente com algumas de suas limitações. Problemas de identificabilidade e várias soluções possíveis foram descritos com nosso método preferido de resolver isso sendo empregar uma estratégia de redimensionamento adequada em nosso algoritmo de amostragem posterior. A probabilidade de Plackett-Luce foi estendida para lidar com uma classe muito mais rica de classificações (classificações parciais superiores e superiores). Ao apelar para técnicas de aumento de dados, uma estratégia eficiente de amostragem de Gibbs foi possibilitada (sujeita a certas especificações prévias) e um esboço detalhado do algoritmo para amostrar a partir da distribuição posterior foi fornecido.

Numerosos estudos de simulação foram considerados, revelando como, sob o modelo padrão de Plackett-Luce, nossas inferências (posteriore) podem ser substancialmente diferentes se os dados contiverem classificações incomuns (spam). Esta é uma característica indesejável do modelo e, portanto, na Seção 2.5, propusemos o modelo Weighted Plackett-Luce, que permite a noção de confiabilidade do ranker por meio de um indicador binário latente. Vimos por meio de estudos de simulação como as inferências sob o modelo Weighted Plackett-Luce são mais robustas para a adição de classificações de spam. Além disso, nosso modelo foi capaz de identificar corretamente os rankings que eram, em certo sentido, incomuns. Embora o modelo de Plackett-Luce ponderado permita um certo nível de heterogeneidade do classificador – ou seja, que um pequeno grupo de classificadores que parecem ter uma preferência alternativa pode ser ponderado para baixo – o modelo não é suficiente para lidar efetivamente com um cenário em que vários grupos de classificadores expressam preferências diferentes. Segue-se que, mesmo ao modelar classificações usando o modelo Weighted Plackett-Luce, ainda devemos fazer a suposição subjacente de que todos os classificadores compartilham crenças/preferências semelhantes sobre as entidades que estão classificando. Acreditamos, no entanto, que essa suposição talvez seja implausível em cenários do mundo real e, no próximo capítulo, construiremos modelos mais flexíveis que permitem que a suposição de crenças homogêneas seja relaxada.

Capítulo 3

Análise de dados heterogêneos classificados

3.1 Introdução

Até agora, assumimos que um único vetor de parâmetro λ é suficiente para resumir as crenças de todos os classificadores que contribuem para um conjunto de dados. Além disso, também assumimos que cada parâmetro de habilidade λ_k ($k = 1, \dots, K$) é único. Agora supomos que talvez existam grupos de classificadores, cada grupo com suas próprias crenças sobre a verdadeira ordem de preferência das entidades. Para implementar essa estrutura, cada classificador tem seu próprio parâmetro $\lambda_{k,i}$. No entanto, todos os classificadores dentro do mesmo grupo de classificadores compartilham as mesmas crenças sobre as entidades e, portanto, todos os classificadores dentro do grupo têm os mesmos valores de parâmetros de habilidade. Propomos ainda que determinados grupos de classificadores podem não ser capazes de distinguir entre certas entidades, ou seja, pode haver estrutura de grupo nas entidades dos grupos de classificadores. Para permitir isso, exigimos a possibilidade de que os valores dentro de cada vetor de parâmetro λ não sejam necessariamente únicos. Apelamos para métodos de agrupamento não paramétricos bayesianos para implementar essa estrutura, especificamente usando processos de Dirichlet. Primeiro, no entanto, revisamos métodos para misturas finitas.

3.2 Modelos de misturas finitas

Uma abordagem comum ao analisar dados heterogêneos é apelar para modelos de mistura. Essa rica classe de modelos nos permite inferir subgrupos contidos nos dados sem informações (prévias) sobre a pertença a subgrupos de observações individuais. Um subgrupo pode ser pensado como um conjunto de observações individuais que formam um grupo "homogêneo". Presume-se que as observações individuais dentro de cada subgrupo sigam o mesmo subjacente

distribuição.

Os modelos mais simples dentro desta classe são modelos de mistura finita; ver, por exemplo, Everitt and Hand (1981) e Lindsay (1995). Observe que, para facilitar a notação e a exposição, escrevemos $f(\cdot|\cdot)$ para uma função de densidade (ou probabilidade) (dependendo se a quantidade é contínua ou discreta), mas simplesmente nos referimos a essas funções como densidades. Em modelos de mistura finitos, a densidade paramétrica que define o modelo, denotada $f(x)$, é composta por um número fixo (finito) N de componentes de mistura. Mais formalmente, dizemos uma mistura de componentes de N de densidade $f(x)$ se ela assumir a forma

$$f(x|\psi, \lambda) = \sum_{c=1}^N \psi c f_c(x|\lambda_c) \quad (3.1)$$

onde $f_1(x), \dots, f_N(x)$ são densidades de componentes e ψ são pesos de mistura para cada componente. A densidade de cada componente também é parametrizada por um valor único λ_c . Para que essa densidade seja bem definida, as seguintes restrições devem ser mantidas: (a) as densidades de componentes $f_c(x|\lambda_c)$ devem ser todas funções de densidade válidas, ou seja, exigimos $f_c(x|\lambda_c) \geq 0$ para todos e $\int f_c(x|\lambda_c) dx = 1$ para $c = 1, \dots, N$, e (b) os pesos da mistura ψ devem estar no simplex ($N - 1$) dimensional, isto é, $\psi_c \geq 0$ para $c = 1, \dots, N$ e $\sum \psi_c = 1$. Assumindo que essas condições sejam válidas, essa distribuição de mistura é definida para qualquer escolha de densidades de componentes, sejam elas contínuas ou discretas. Na prática, no entanto, essas densidades são frequentemente escolhidas da mesma família.

Se tivermos n observações, denotadas $x = (x_1, \dots, x_n)$, a probabilidade (dados observados) é

$$\pi(x|\psi, \lambda) = \prod_{i=1}^n \left\{ \sum_{c=1}^N \psi c f_c(x_i|\lambda_c) \right\}, \quad (3.2)$$

o que, em geral, é muito complicado. No entanto, é possível tornar a forma da probabilidade substancialmente mais direta apelando para métodos de aumento de dados - especificamente introduzindo variáveis indicadoras de componente / cluster latentes que discutimos agora.

Uma abordagem comum ao implementar modelos de mistura é introduzir variáveis indicadoras de cluster latentes, aqui denotadas $c = (c_1, \dots, c_n)$, onde $c_i = j$ denota que a observação i pertence ao componente / cluster j . Condicionado às variáveis indicadoras de agrupamento latentes, o modelo é simplificado significativamente, pois a densidade condicional para observação x_i é simplesmente $f_i(x_i | \lambda_c)$. Essas variáveis aleatórias (não observadas) seguem uma distribuição categórica definida como $Pr(c_i = c) = \psi_c$ para $i = 1, \dots, n$, $c = 1, \dots, N$ e denotou $c_i | \psi \sim \text{Cat}(\psi)$. Portanto, a probabilidade conjunta (dados completos) dos dados x e as variáveis indicadoras de cluster latente c

É

$$\pi(x, c|\psi, \lambda) = \prod_{i=1}^n \psi c_i f_{ci}(x_i | \lambda c_i),$$

Como, dados os parâmetros λ , ψ , os pares (x_i, c_i) são independentes. Esta forma de probabilidade é substancialmente mais direta do que (3.2) e é a razão pela qual os indicadores de cluster latentes são normalmente introduzidos ao ajustar modelos de mistura.

Segue-se que as implementações bayesianas de modelos de mistura finita, dados os indicadores latentes, são geralmente da forma

$$\begin{aligned} x_i|\lambda, c_i, \psi &\sim f_{ci}(x_i|\lambda c_i) \\ c_i|\psi &\sim \text{Gato}(\psi) \\ \psi &\sim \text{Dir}(\alpha) \end{aligned} \quad (3.3)$$

onde $i = 1, \dots, n$ e $\text{Dir}(\alpha)$ denota a distribuição de Dirichlet com parâmetros de concentração $\alpha = (\alpha_1, \dots, \alpha_N)$ onde $\alpha_i > 0$. Observe que, na definição do modelo acima, os componentes da mistura (e os indicadores de cluster latentes) são intercambiáveis, ou seja, podem ser renomeados arbitrariamente, mantendo a especificação do modelo equivalente (Stephens, 2000). Portanto, dentro de um contexto de inferência, talvez não seja sensato favorecer um componente de mistura específico a priori. Isso é conseguido escolhendo os parâmetros de concentração para serem $\alpha_i = \alpha = 1$, o que dá $\psi \sim \text{Dir}(1)$, ou seja, os pesos dos componentes da mistura ψ seguir uma distribuição uniforme sobre o simplex dimensional $(N - 1)$. Naturalmente, poderíamos escolher formar uma mistura de N componentes dos modelos de Plackett-Luce, deixando as distribuições de componentes serem da forma de Plackett-Luce, com $X_i | \Lambda, c_i \sim PL(\lambda c_i)$ onde $\lambda c = (\lambda c_1, \dots, \lambda c_K)$ é o vetor de parâmetro associado ao componente (clus-ter) c e $\Lambda = \{\lambda c\}_{N=1}$ é a coleção de todos esses vetores de parâmetro. De fato, Gormley e Murphy (2008a, b, 2009) e Mollica e Tardella (2014) propõem misturas finitas de Plackett-Luce e modelos relacionados para permitir preferências diferentes entre os classificadores. Essa abordagem também foi adotada por Vitelli et al. (2018), mas em vez disso eles escolheram um modelo baseado na distância, ou seja, o de Mallows (1957). É claro que essa abordagem poderia ser trivialmente estendida para formar uma mistura de N componentes de modelos ponderados de Plackett-Luce por letting $X_i|\Lambda, c_i \sim PLW(\lambda c_i, w)$. Sob essa configuração, os pesos do classificador w seriam comuns em todos os componentes, com apenas o vetor de parâmetro λ sendo específico do cluster. Em certo sentido, os modelos descritos no Capítulo 2 podem ser considerados um caso trivial dentro da estrutura do modelo de mistura (finita), com componentes de mistura $N = 1$, ou seja, um único subgrupo homogêneo que contém toda a população de classificadores.

Embora os modelos de mistura finita ofereçam flexibilidade para modelar dados heterogêneos, especificar uma forma apropriada de tal modelo não é uma tarefa trivial. Uma das principais questões que

surge quando o ajuste de modelos de mistura finita é a restrição de que um número fixo de componentes de mistura deve ser escolhido a priori. Isso requer que o analista decida quantos subgrupos estão contidos em uma população antes de realizar sua análise. Na tentativa de superar esse problema, muitos optam por ajustar vários modelos, cada um com números diferentes de componentes, e então apelam para técnicas da seleção de modelos (como o critério de informação de Akaike (AIC) ou o critério de informação bayesiano (BIC)) para determinar qual modelo melhor se ajusta aos dados. Esta solução, no entanto, tem o custo de realizar inúmeras análises. O analista também é obrigado a escolher o número (diferente) de componentes a serem considerados. Idealmente, o modelo de mistura seria definido de modo que o número de componentes não fosse fixado a priori e, em vez disso, permitisse que o número de componentes fosse inferido usando, por exemplo, métodos de salto reversível (Richardson e Green, 1997). Alternativamente, podemos apelar para uma classe mais flexível de modelos, ou seja, modelos de mistura infinita. Como o nome sugere, os modelos de mistura infinita contêm um número "infinito" de componentes e, portanto, a densidade subjacente $f(x | \psi, \lambda)$ pode ser pensada como o caso limitante como $N \rightarrow \infty$ de uma mistura finita (3.1). Observe que um número "infinito" de componentes só existe na teoria e, na prática, o número de componentes não vazios pode ser no máximo o número de observações. Dada a forma de um modelo de mistura finita (3.3), é claro que precisamos de uma distribuição de Dirichlet de dimensão infinita para definir um modelo de mistura infinita. A versão generalizada (para dimensão infinita) da distribuição de Dirichlet é o processo de Dirichlet - este é o tópico da próxima seção.

3.3 O processo Dirichlet

Na seção anterior, discutimos a necessidade de aumentar a flexibilidade da modelagem e relaxar a exigência de um número fixo de componentes a priori, apelando para modelos de mistura infinitos. Tais modelos permitem uma generalidade total e, além disso, permitem inferir o número de componentes da mistura a partir dos dados. Por sua natureza, os modelos de mistura infinita induzem um espaço de parâmetros dimensionais infinitos e, portanto, se enquadram na área de não-paramétricos bayesianos (Hjort et al., 2010).

O processo de Dirichlet é um conjugado anterior para distribuições categóricas de dimensão infinita - uma generalização (para dimensão infinita) do resultado de que a distribuição de Dirichlet é um conjugado anterior para a distribuição categórica (multinomial). Agora fornecemos uma visão geral e descrevemos as representações comuns do processo de Dirichlet antes de considerar modelos de mistura finita. A visão geral fornecida aqui é um tanto breve e remetemos o leitor a Ferguson (1973) e Antoniak (1974) ou ao livro mais recente de Hjort et al. (2010) para obter mais detalhes sobre a teoria da medida subjacente aos processos de Dirichlet.

Usamos a notação $G \sim DP(\alpha, G_0)$ para denotar que uma distribuição G segue um Dirichlet

processo, onde α e G_0 denotam o parâmetro de concentração e a distribuição de bases, respectivamente. Como o nome sugere, G_0 é uma distribuição em si e pode ser uma distribuição contínua ou discreta. Uma realização de um processo de Dirichlet, no entanto, é quase certamente uma distribuição discreta, independentemente da forma de G_0 . As realizações (distribuições) são extraídas em torno da distribuição básica de uma maneira conceitualmente semelhante a como as realizações da distribuição normal são extraídas em torno da média. Além disso, a expectativa de um processo de Dirichlet também é a distribuição de base, ou seja, $E(G) = G_0$. O parâmetro de concentração controla o desvio das realizações da distribuição de base e, nesse sentido, comporta-se de forma semelhante a um desvio padrão (inverso) e, portanto, não surpreendentemente, $G \rightarrow G_0$ como $\alpha \rightarrow \infty$, ou seja, as realizações ficam cada vez mais semelhantes à distribuição de base à medida que o parâmetro de concentração aumenta. Como $\alpha \rightarrow 0$, as realizações G são distribuições discretas concentradas em um único ponto de massa. Segue-se que a probabilidade de dois componentes distintos (de G) serem iguais, $\Pr(\lambda_i = \lambda_j)$ para $i \neq j$, tende a 0 como $\alpha \rightarrow \infty$ (assumindo que G_0 é contínuo), enquanto $\Pr(\lambda_i = \lambda_j) \rightarrow 1$ como $\alpha \rightarrow 0$. A Figura 3.1 ilustra isso representando (em cada linha) três realizações independentes de um processo de Dirichlet com $\alpha = 1, 5, 10, 50, 100$ de cima para baixo, respectivamente. A distribuição básica escolhida neste caso é $G_0 = N(0, 1)$. Observamos que, para os valores menores de α nossas realizações (distribuição discreta), G são definidas em menos átomos, com alguns tendo grande massa. Segue-se que (teoricamente) se $\alpha = 0$ teríamos uma única massa pontual em algum valor $\lambda \in \mathbb{R}$. Para valores maiores de α , observamos que as realizações (distribuições) têm cada vez mais átomos únicos, cada um com peso relativamente pequeno. Em teoria, se $\alpha = \infty$ então essa distribuição seria definida sobre infinitos átomos, cada um dos quais tem massa 0, ou seja, a distribuição seria $G_0 = N(0, 1)$. A convergência de G para a distribuição de base à medida que α aumenta é talvez mais facilmente vista através da Figura 3.2, que mostra a função de distribuição cumulativa empírica (CDF) para cada uma das respectivas realizações mostradas na Figura 3.1. É claro que, à medida que o α aumenta, o CDF dessas realizações se torna mais parecido com o de uma distribuição $N(0, 1)$, ou seja, $G \rightarrow G_0$ como $\alpha \rightarrow \infty$. Sethuraman (1994) mostrou que cada processo de Dirichlet tem uma representação correspondente de quebra de bastão: escrever $G \sim DP(\alpha, G_0)$ é equivalente a

$$G(\cdot) = \sum_{j=1}^{\infty} \psi_j \delta_{\lambda_j}(\cdot) \quad (3.4)$$

$$\psi_j = v_j \prod_{i \neq j}^{< \infty} (1 - v_i)$$

$$v_j \text{ indep} \sim \text{Beta}(1, \alpha) \lambda_j \text{ indep} \sim G_0$$

onde $\delta_{X(j)}$ denota a medida de probabilidade de Dirac concentrada em x e $j \in N$. A construção dos pesos determina que $\sum_j u_{jj} = 1$ e, portanto, segue-se que a representação de quebra de bastão define uma distribuição discreta G com átomos λ_j que têm respectivas habilidades de problema (pesos) u_{jj} . Também notamos que os pesos estão diminuindo estocasticamente, ou seja, $E(u_{jj}) \rightarrow 0$ como $j \rightarrow \infty$. Embora a representação de quebra de bastão forneça a visão mais intuitiva de como o processo de Dirichlet é definido, existem representações alternativas. Na próxima seção, consideraremos duas alternativas comuns à representação de quebra de pau - o processo do restaurante chinês e o esquema de urna P'olya (relacionado).

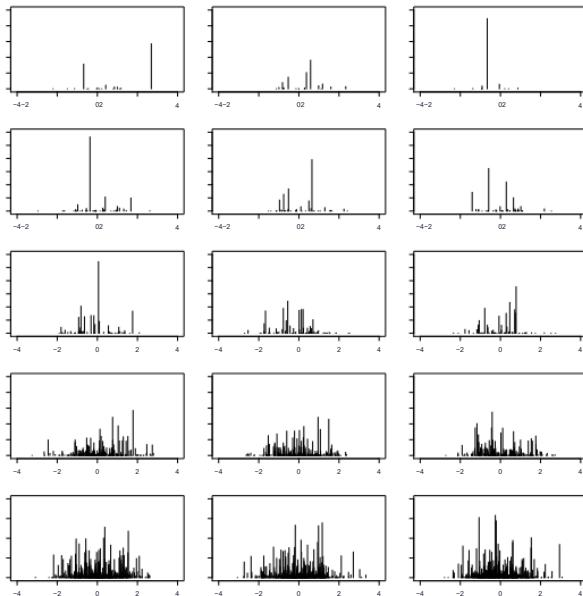


Figura 3.1: Múltiplas realizações de um processo de Dirichlet com $G_0 = N(0, 1)$ e $\alpha = 1, 5, 10, 50, 100$ de cima para baixo, respectivamente.

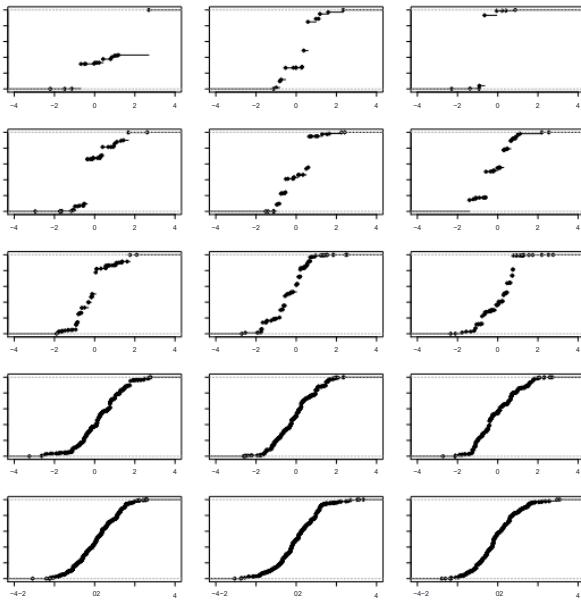


Figura 3.2: CDF empírico de múltiplas realizações de um processo de Dirichlet com $G_0 = N(0, 1)$ e $\alpha = 1, 5, 10, 50, 100$, de cima para baixo, respectivamente.

3.3.1 Representações alternativas do processo de Dirichlet

Representação do processo de restaurante chinês

Uma maneira alternativa de visualizar um processo de Dirichlet é através do processo do restaurante chinês (Aldous, 1985). Essa analogia, atribuída a Jim Pitman e Lester Dubins, descreve a distribuição sobre as alocações de cluster (mais formalmente a distribuição sobre partição) induzida pelo processo de Dirichlet e procede da seguinte forma. Suponha que haja um restaurante chinês que contenha um número infinito de mesas, cada uma com assentos infinitos

capacidade. O primeiro cliente a entrar no restaurante senta-se na mesa número 1. O próximo cliente então tem uma escolha: (a) eles se sentam em uma mesa ocupada com probabilidade proporcional ao número de pessoas atualmente nessa mesa ou (b) eles se sentam em uma nova mesa (desocupada) com probabilidade proporcional a α . Este processo continua até que todos os n clientes tenham sido atendidos. Neste ponto, $N \leq n$ mesas serão ocupadas, e os indivíduos em cada mesa serão interpretados como agrupados. Segue-se que N_c é o número de clusters únicos. A partir dessa metáfora, deve ficar claro que um cliente tem maior probabilidade de se sentar a uma mesa com um grande número de clientes do que com apenas alguns clientes. Esta é uma característica do processo de Dirichlet, que muitas vezes é resumida pela frase "os ricos ficam mais ricos". Desenhar realizações independentes de G_0 e atribuir um valor a cada mesa resulta em uma distribuição discreta com probabilidades proporcionais ao número de pessoas sentadas em cada mesa respectiva. Este processo é permutável, ou seja, a ordem em que os n clientes chegam não afeta a distribuição de probabilidade final.

Representação da urna Pólya

Outra maneira de visualizar um processo de Dirichlet (e o processo de restaurante chinês associado) é por meio de um esquema de urna Pólya (Blackwell e MacQueen, 1973). Para esta analogia, é útil considerar $\alpha \in N$, embora as probabilidades de as observações serem atribuídas a cada agrupamento (a ser definido) sejam válidas para qualquer $\alpha \in R^>0$. Suponha que temos uma urna cheia de bolas brancas α . Uma realização do processo de Dirichlet é obtida desenhando repetidamente bolas da urna, sujeito às seguintes regras. Se a bola sorteada for branca, geramos uma bola adicional de cor única antes de colocar a bola branca que selecionamos e a nova bola colorida de volta na urna. Se a bola sorteada não for branca, geramos uma bola adicional da mesma cor da sorteada e devolvemos ambas de volta à urna. Este processo continua até que n bolas tenham sido sorteadas. Uma vez que a enésima bola tenha sido retirada (e a ação apropriada tomada), as bolas brancas são descartadas da urna. Neste ponto, haverá $N_c \leq n$ bolas de cores únicas dentro da urna e a distribuição sobre essas cores é equivalente à distribuição sobre as mesas dentro do processo de restaurante chinês. Desenhar realizações independentes de G_0 e atribuir um valor a cada cor única resulta em uma distribuição discreta com probabilidades proporcionais ao número de cada bola colorida. Novamente, esse processo também é intercambiável, ou seja, se n pessoas selecionam uma bola, a ordem em que as pessoas estão dispostas não afeta a distribuição final de probabilidade.

A probabilidade de atribuir a i -ésima pessoa/bola à cor da mesa/bola j em ambas as representações alternativas é

$$\Pr(c_i = j | c_1, \dots, c_{i-1}) = \frac{\text{ncij} + i - 1}{\text{Nc} + i - 1}$$

onde ncij denota o número atual de observações atribuídas ao cluster j (na iteração i), e N_c é o número de clusters únicos que existem atualmente ($N_c = 0$ quando $i = 1$).

3.3.2 Gerando uma realização de um processo de Dirichlet

Nesta seção, descrevemos como obter realizações de um processo de Dirichlet, ou seja, como obter uma realização da distribuição discreta G onde $G|\alpha, G_0 \sim DP(\alpha, G_0)$. A representação de quebra de bastão, embora direta, vem com problemas inerentes. Na prática, não é possível amostrar um número infinito de pesos para cada um dos átomos correspondentes para definir G como em (3.4). Infelizmente, apelar para representações alternativas do processo de Dirichlet, como os esquemas de restaurante chinês / urna Pólya, produz problemas semelhantes. Esses processos são projetados para permitir a amostragem direta de G , em vez de gerar uma realização do próprio G . Talvez, dada uma realização particular de alocações de agrupamento c e parâmetros de agrupamento correspondentes (λ_j) indep~ G_0 para $j = 1, \dots, N_c$ de G , possamos pensar que uma realização da distribuição discreta G é uma com átomos λ_j cujos pesos são proporcionais ao número de observações dentro do cluster j , isto é, com $\Pr(\lambda = \lambda_j) \propto \#(c_i = j)$ para $i = 1, \dots, n$ e $j = 1, \dots, N_c$. No entanto, esta não é uma realização verdadeira de G , pois tal distribuição é definida apenas sobre os átomos que estão atualmente atribuídos a uma das observações. O número infinito restante de átomos (e pesos correspondentes) permanece indefinido até que uma observação seja atribuída a eles. Diante disso, uma verdadeira realização de G só pode ser gerada usando esse método se considerarmos o limite como $n \rightarrow \infty$, o que é claramente inviável. Sem uma solução óbvia para esse problema, em vez disso, voltamos nosso foco para a representação de quebra de bastão. A questão principal aqui é como amostrar o número infinito de pesos para cada um dos átomos correspondentes, de modo que possamos definir G como em (3.4). Isto se torna viável se escolhermos um parâmetro de truncamento N_1 adequado $< \infty$ para que a distribuição

$$G_* = \sum_{j=1}^{N_1} \psi_j \delta_{\lambda_j}(\cdot)$$

é uma aproximação razoável de G , de modo que precisamos apenas de pesos N_1 amostrais (e átomos). Claramente $*d \rightarrow G$ como $N_1 \rightarrow \infty$, portanto, o nível de aproximação diminui

à medida que N_1 aumenta. Este parâmetro de truncamento precisa ser escolhido para que $\sum j > N_1 \psi_j$ ' 0, uma restrição equivalente a $\sum j \leq N_1 \psi_j$ ' 1. Ishwaran e Zarepour (2002) descrevem como se pode escolher um parâmetro de truncamento adequado. Observamos, no entanto, que este método de amostragem pode se tornar inviável para grandes α . Lembre-se de que os pesos são definidos como

$$\psi_j = v_j \prod_{j' < j} (1 - v'), \quad \text{onde} \quad v_j \text{ indep} \sim \text{Beta}(1, \alpha),$$

E assim, para aumentar α temos que $\psi_j > 0$ para valores cada vez maiores de j ; isso decorre do resultado da que $E(v_j) \rightarrow 0$ como $\alpha \rightarrow \infty$. Portanto, na situação em que α é grande, pode se tornar inviável escolher um parâmetro de truncamento adequadamente grande N_1 para que a restrição de $\sum j \leq N_1 \psi_j$ ' 1 seja satisfeita. No entanto, na prática, muitas vezes é possível escolher um valor adequado de N_1 de modo que $G^* \approx G$ mesmo para valores modestamente grandes de α .

- Escolha N_1 suficientemente grande.
- Escolha $\alpha > 0$ ou simule a partir de uma distribuição adequada.
- Simule $\lambda \mid j$ indep $\sim G_0$ para $j = 1, \dots, N_1$.
- Simule v_j indep $\sim \text{Beta}(1, \alpha)$ para $j = 1, \dots, N_1$.
- Definir $\psi_j = v_j / (1 - v')$ para $j = 1, \dots, N_1$.

$\prod_{j' < j}$

A distribuição discreta (realização de G) é aquela definida sobre os átomos $\lambda \mid j$ com pesos ψ_j , isto é, $\Pr(\lambda = \lambda \mid j) \propto \psi_j$ para $j = 1, \dots, N_1$. Observe que G^* só define uma distribuição de mistura se os pesos ψ somam um; ver Seção 3.2. No entanto, os pesos acima só satisfazem isso no caso limite quando $N_1 = \infty$. É claro que poderíamos simplesmente redimensionar esses pesos para que $\sum N_1 j = 1$, no entanto, muitas vezes é vantajoso, particularmente dentro de um contexto de inferência, construir os pesos de modo que eles somem um, independentemente da escolha do parâmetro de truncamento (finito) N_1 . Isso pode ser alcançado simulando v_j indep $\sim \text{Beta}(1, \alpha)$ para $j = 1, \dots, N_1 - 1$ e fixando $V_{N_1} = 1$. Os pesos são então construídos da maneira usual, no entanto, nesta configuração, temos $\psi_{N_1} = \prod_{j' < N_1} (1 - v')$ e, portanto, ψ_{N_1} contém toda a massa restante e isso garante que $\sum N_1 j = 1 \psi_j = 1$. Segue-se que, neste caso, G^* é uma distribuição de mistura válida (discreta) definida pelas probabilidades $\Pr(\lambda = \lambda \mid j) = \psi_j$ para $j = 1, \dots, N_1$.

3.3.3 Gerando parâmetros de modelo consistentes com um processo de Dirichlet anterior

Na seção anterior, consideramos como obter realizações de G onde $G|\alpha, G_0 \sim DP(\alpha, G_0)$. Agora supomos que os dados são n observações e denotam o parâmetro para a i -ésima observação por x_i . O foco desta seção é como obter uma realização de $\lambda = (\lambda_1, \dots, \lambda_n)$ dado $\lambda|G \sim G$, isto é, como tirar realizações (dos parâmetros) de G , onde G segue um processo de Dirichlet. Tais realizações são razoavelmente simples de obter sob a representação de quebra de bastão. No entanto, qualquer realização obtida usando este método será apenas de uma aproximação à distribuição verdadeira definida pelo processo de Dirichlet. Essa aproximação se deve à necessidade de truncar G para que seja de dimensão finita (veja acima). Por outro lado, as representações chinesas de restaurantes / urnas P'olya permitem que realizações exatas sejam extraídas diretamente de G (não precisamos obter uma realização de G explicitamente). Além disso, essas amostras permanecem da distribuição verdadeira definida pelo processo de Dirichlet, independentemente do valor de α .

Para gerar uma realização de $\lambda = (\lambda_1, \dots, \lambda_n)$, apelamos para os indicadores de agrupamento latente c , onde $c_i = c$ denota que a observação i está dentro do agrupamento c . Note-se que, tendo em conta os parâmetros de agrupamento (únicos) $\lambda_{\dagger j}$, os indicadores de agrupamento permitem-nos identificar completamente λ e, por conseguinte, obter uma realização para os indicadores de agrupamento equivale a tirar uma realização para λ . Descrevemos agora como obter realizações de c (e, portanto, λ) que são consistentes com o processo de Dirichlet anterior sob as representações de quebra de pau e de urna de restaurante chinês / P'olya.

Representação de quebra de bastão

Para gerar uma realização das alocações de cluster consistente com um processo de Dirichlet antes de usar a representação de quebra de bastão, devemos primeiro obter uma realização (aproximada) da distribuição discreta G . Isso pode ser obtido usando o método descrito na seção anterior, que fornece uma realização de G definida por $\Pr(\lambda = \lambda_{\dagger j}) = u_j$ para $j = 1, \dots, N_1$. Dada uma realização de G , podemos gerar uma realização das alocações de cluster para n observações da seguinte forma.

- Amostra $c_i \sim \text{Cat}(N_1, \psi)$ para $i = 1, \dots, n$. O vetor parâmetro λ é então dado por $\lambda_i = \lambda_{\dagger c_i}$ para $i = 1, \dots, n$.

Representação de restaurante chinês/urna P'olya

Sob a representação de quebra de bastão, precisamos primeiro obter uma realização de (a distribuição) G antes de extrair amostras das alocações de cluster (de G). No entanto, sob a representação do restaurante chinês/urna P'olya, esta etapa não é mais necessária e, em vez disso, podemos extrair realizações (de alocações de cluster c e, portanto, λ) diretamente de Gusing o seguinte processo:

- Escolha $\alpha > 0$ ou simule a partir de uma distribuição adequada.
- Defina $c_1 = 1$ e o número (atual) de clusters como $N_c = 1$.
- Para $i = 2, \dots, n$ simular a alocação da observação i para um cluster de acordo com a distribuição discreta

$$\Pr(c_i = j | c_1, \dots, c_{i-1}) = \frac{n_{c_j} + i - 1}{N_c + i - 1}, \quad \text{para } j = 1, \dots, N_c,$$

$$\Pr(c_i = N_c + 1 | c_1, \dots, c_{i-1}) = \frac{\alpha}{N_c + i - 1}$$

$$= \frac{\alpha}{N_c + i - 1},$$

onde n_{c_j} denota o número de pontos atualmente dentro do cluster j (na iteração i), e $N_c \rightarrow N_c + 1$ se $c_i = N_c + 1$.

- Simule λ_j indep~G0 para $j = 1, \dots, N_c$. Novamente, quanto à representação de quebra de bastão, o parâmetro associado à observação é dado por $\lambda_j c_i$. Segue-se que o vetor parâmetro λ é dado por $\lambda_i = \lambda_j c_i$ para $i = 1, \dots, n$. Agora destacamos uma diferença sutil, mas importante, entre os dois métodos. Suponha que estejamos no cenário em que $n < N_1$, ou seja, o número de observações é significativamente menor que o parâmetro de truncamento. Neste caso, o erro resultante da aproximação utilizada na abordagem de quebra de varas será razoavelmente pequeno (condicionado a uma escolha adequada do parâmetro de concentração). No entanto, na abordagem de quebra de bastão, a realização de G é definida sobre um número finito (fixo) de átomos (N_1), que, como resultado, restringe o número máximo de clusters a N_1 . Em outras palavras, independentemente do número de observações n , o vetor parâmetro λ pode conter no máximo N_1 valores únicos. Segue-se que este método de geração de realizações de parâmetros pode resultar em uma aproximação da distribuição verdadeira definida pelo processo de Dirichlet no limite $n \rightarrow \infty$. No entanto, o método do restaurante chinês / urna P'olya permite a possibilidade de que cada uma das observações i possa se juntar a um novo cluster e, portanto, é atribuída a um parâmetro (único) que é um sorteio independente da distribuição de base. Segue-se que, neste caso, o limite superior do número de clusters é teoricamente infinito quando se considera o limite como $n \rightarrow \infty$.

3.3.4 Processo Dirichlet generalizado

O processo de Pitman-Yor é uma versão generalizada do processo de Dirichlet. Este processo é creditado a Pitman e Yor (1997) por seu trabalho na distribuição de Poisson-Dirichlet de dois parâmetros. No entanto, o nome foi cunhado por Ishwaran e James (2001) em sua revisão de antecedentes que quebram bastões. Aqui deixamos PY(α, d, G_0) denotar o processo Pitman-Yor com parâmetros governantes α ($> -d$), conhecido como parâmetro de força, um parâmetro de desconto $0 \leq d < 1$ e uma distribuição de base G_0 . Quanto ao processo de Dirichlet, uma realização do processo de Pitman-Yor é uma distribuição discreta sobre um conjunto infinito de átomos; também esses átomos são (independentes) extraídos da distribuição de bases G_0 . No entanto, em contraste com o processo de Dirichlet, o peso (probabilidade) associado a cada átomo é extraído de uma distribuição de Poisson-Dirichlet de dois parâmetros. Isso resulta no processo de Pitman-Yor sendo mais flexível do que o processo de Dirichlet no que diz respeito ao comportamento da cauda e é frequentemente o modelo preferido para analisar dados com caudas de lei de potência (o processo de Dirichlet tem caudas exponenciais).

Para visualizar a relação entre os processos de Pitman-Yor e Dirichlet, considere a representação de quebra de bastão do primeiro

$$G(\cdot) = \sum_{j=1}^{\infty} \psi_j \delta_{\lambda_j}(\cdot) \quad (3.5)$$

$$\psi_j = v_j \prod_{i < j} (1 - v_i)$$

$$v_j \text{ indep} \sim \text{Beta}(1 - d, \alpha + jd) \lambda_j \text{ indep} \sim G_0.$$

Claramente, o caso $d = 0$ produz uma distribuição G de (3.5) que é equivalente à do processo de Dirichlet (3.4), ou seja, $PY(\alpha, 0, G_0) \equiv DP(\alpha, G_0)$. Por esta razão, o processo de Dirichlet é considerado um caso especial do processo Pitman-Yor. O processo Inverso-Gama Normalizado é outro caso especial dado por $d = 0,5$ e $\alpha = 0$.

Precisamos $\sum_j \psi_j = 1$ para que G seja bem definido, ou equivalentemente, os pesos dos átomos devem estar no simplex. Se deixarmos $a = 1 - d$ e $b_j = \alpha + jd$, o Lema 1 de Ishwaran e James (2001) mostra que $\sum_{j=1}^{\infty} \psi_j = 1$ quase certamente se e somente se $\sum_{j=1}^{\infty} \lambda_j = \infty$.

$$\sum_{j=1}^{\infty} \psi_j = 1 \text{ quase certamente} \iff \sum_{j=1}^{\infty} \lambda_j = \infty. \quad (3.6)$$

É trivial verificar se a condição (3.6) é válida para o processo de Dirichlet. Lembre-se de que o processo de Dirichlet é um caso especial do processo de Pitman-Yor com $d = 0$ e, portanto, $a = 1$

e $b_j = \alpha$. Diante disso, temos

$$\begin{aligned} u_m > 0 &\implies 1 + \frac{1}{u_m} > 1 \\ &\implies \sum_{a=1}^{\infty} (1 + \frac{1}{a}) > 0 \\ &\implies \sum_{j=1}^{\infty} \sum_{a=1}^{\text{tor}} (1 + \frac{1}{a}) = \infty \\ &\implies \sum_{j=1}^{\infty} \psi_j = 1 \text{ quase certamente} \end{aligned}$$

e assim a distribuição em (3.4) é bem definida.

A seguir, nos concentraremos no processo de Dirichlet e observamos que esta é uma escolha comum de quebrar o bastão antes; principalmente devido à disponibilidade de esquemas de amostragem eficientes. Muitos desses esquemas de inferência (eficientes) fazem uso da representação do processo de restaurante chinês. Infelizmente, tais representações normalmente não estão disponíveis para o processo de Pitman-Yor, portanto, a representação de quebra de bastão deve ser usada e, sob essa representação, só é possível obter uma distribuição posterior aproximada (embora a aproximação possa ser feita arbitrariamente pequena com poder de computação suficiente) - isso é discutido mais adiante na Seção 3.4.1.

3.4 Modelos de mistura de processo Dirichlet

O desenvolvimento do modelo de mistura de processos de Dirichlet (DPMM) é creditado a Ferguson (1973) e Antoniak (1974). Desde a sua concepção, os DPMMs tornaram-se populares na literatura bayesiana, pois permitem que modelos complexos e flexíveis sejam construídos com relativa facilidade. Um modelo típico de mistura de processo de Dirichlet é uma mistura de uma distribuição F sobre seus parâmetros. Por exemplo

$$x_i | \lambda_i \sim F(\lambda_i),$$

$$\begin{aligned} \lambda_i | G &\sim G, G | \alpha, G_0 \sim \\ &DP(\alpha, G_0), \end{aligned}$$

onde DP denota um processo de Dirichlet (formalmente definido pela representação de quebra de bastão em (3.4)) com parâmetro de concentração α e distribuição de bases G_0 .

3.4.1 Inferência bayesiana para DPMM

Existem várias maneiras de realizar inferência bayesiana para modelos de mistura de processos de Dirichlet. A maioria dos métodos pode ser classificada como adotando uma abordagem condicional ou marginal, conforme resumido, por exemplo, em Papaspiliopoulos e Roberts (2008). As abordagens condicionais normalmente usam truncamento para aproximar o aspecto de dimensão infinita do anterior de quebra de bastão, conforme pioneiro de Ishwaran e James (2001). No entanto, evitar aproximações é benéfico e os amostradores de fatia e retrospectiva de Walker (2007) e Papaspiliopoulos e Roberts (2008) fornecem métodos para alcançar isso. Infelizmente, esses métodos podem sofrer de má mistura e convergência ao tentar amostrar distribuições posteriores multimodais. Uma solução é a adição de movimentos apropriados de troca de rótulos (Hastie et al. (2015), Papaspiliopoulos e Roberts (2008)), embora, em geral, seja necessário mais trabalho empírico para determinar o número e os tipos de movimentos que fornecem uma solução eficaz.

Por esses motivos, evitamos métodos condicionais aqui e, em vez disso, implementamos um marginal sampler. Esses amostradores normalmente envolvem um esquema de urna P'olya e marginalizam a distribuição dimensional finita (Escobar e West (1995), MacEachern e Müller (1998)) e, assim, evitam a necessidade de aproximação. O Algoritmo 8 de Neal (2000), doravante referido como o Algoritmo 8 de Neal, é um desses amostradores. Este algoritmo demonstrou ser um dos métodos de amostragem mais eficientes para misturas do Processo Dirichlet; ver, por exemplo, Papaspiliopoulos e Roberts (2008). Além disso, não há necessidade de movimentos adicionais de troca de rótulos. A eficiência é alcançada pelo algoritmo que executa apenas atualizações para os componentes exclusivos que estão atualmente atribuídos a uma observação. Cada observação é então atribuída a: (a) um componente que está atualmente em uso ("ativo") ou (b) a um dos m componentes auxiliares cujos parâmetros são (independentes) extraídos da distribuição da base. O número de componentes auxiliares, m , é escolhido pelo analista. Observamos que isso coloca pouco fardo sobre o analista, pois a escolha de m é feita apenas para fins de eficiência - a distribuição de equilíbrio da cadeia de Markov permanece exata para todas as escolhas de $m \geq 1$ (Neal, 2000). De nossa experiência, descobrimos que tomar $m = 2$ ou 3 componentes auxiliares normalmente produz uma cadeia de Markov bem misturada. Geralmente, a mistura melhora à medida que o número de componentes auxiliares aumenta devido às observações terem mais oportunidade de se juntar a um cluster alternativo. Aumentar m , no entanto, vem com custo computacional adicional, pois m (independente) desenho são necessários a partir da distribuição de base para cada observação em cada iteração de nosso algoritmo. Além disso, a distribuição discreta (condicional completa) sobre as alocações de cluster também aumentará em dimensão.

O Algoritmo 8 de Neal está intimamente relacionado a outros esquemas de amostragem (marginais) na literatura. Quando $m = 1$, o Algoritmo 8 de Neal se assemelha muito ao algoritmo "sem lacunas" de MacEachern e Müller (1998). No entanto, a probabilidade de que uma observação se move de

seu cluster atual (ativo) em um cluster auxiliar é maior sob o Algoritmo 8 de Neal. À primeira vista, isso pode não parecer útil, uma vez que o objetivo é agrupar as observações. No entanto, o processo de Dirichlet tem a infeliz propriedade de "mascarar" pequenos clusters, ou seja, a penalidade por criar um cluster adicional para abrigar algumas observações é maior do que a penalidade por colocá-las em um cluster ativo atual - mesmo que essas observações sejam um pouco diferentes do grupo existente. É evidente que devem ser encorajadas observações para formar novos agrupamentos (juntar os auxiliares) sempre que estes difiram suficientemente dos agrupamentos existentes.

3.4.2 Ter em conta a incerteza quanto ao parâmetro de concentração

Central para a implementação de um modelo de mistura de processo de Dirichlet é a escolha do parâmetro de concentração α . A escolha de α resulta em um prévio implícito no número de clusters ou valores únicos (N_c). Antoniak (1974) fornece uma forma implícita da distribuição a priori condicional, $\pi(N|c, n)$, quando temos n observações; ver secção A.1 nos apêndices para mais pormenores.

Escolher um valor de α é um pouco difícil, a menos que tenhamos conhecimento prévio substancial do número de subgrupos nos dados. Portanto, pode ser útil expressar crenças (incertas) sobre α em termos de uma distribuição anterior e, assim, inferir sua distribuição posterior. Escobar e West (1995) e West (1992) mostram que, ao usar um esquema de amostragem marginal e uma mistura finita de distribuições Gama como um prior para α , é possível derivar uma distribuição condicional completa de forma fechada para α . É bastante simples simular a partir dessa distribuição condicional completa e, portanto, isso pode ser incorporado ao esquema de amostragem de Gibbs.

No caso mais simples em que α tem uma distribuição gama (mistura de componente único de), ou seja, $\alpha \sim Ga(a\alpha, b\alpha)$, a distribuição condicional completa é a mistura de dois componentes

$$\alpha | \dots \sim \pi \text{Ga}(a\alpha + N_c, b\alpha - \log h) + (1 - \pi) \text{Ga}(a\alpha + N_c - 1, b\alpha - \log h),$$

onde os pesos da mistura são dados por

$$\frac{\pi(1 - \pi)}{1n(b\alpha - \log h)},$$

e

$$h | \dots \sim Beta(\alpha + 1, n).$$

Aqui n uma variável aleatória latente que facilita a atualização conjugada. Este resultado é obtido na Seção A.1 dos apêndices.

3.5 Descobrindo a heterogeneidade entre os classificadores

Até agora, nosso principal objetivo tem sido realizar inferência bayesiana e obter uma ordenação de preferência única de entidades que resume uma coleção de classificações e é uma forma de agregação de classificação. Esse objetivo só é apropriado se os classificadores forem homogêneos em termos de suas crenças sobre as entidades, ou seja, cada classificação individual (dentro de uma coleção) segue a mesma distribuição de classificação subjacente. Nossa atual modelo ponderado de Plackett-Luce permite heterogeneidade entre as habilidades do classificador, o que, em certo sentido, permite diferentes grupos de classificadores. No entanto, essa heterogeneidade permite apenas uma variabilidade limitada entre grupos informativos e não informativos de classificadores. Este modelo é, portanto, inadequado para lidar com um cenário em que vários grupos de classificadores expressam preferências alternativas em relação às entidades que estão classificando.

Neste capítulo, apresentamos até agora modelos de mistura finita e infinita e discutimos como esses modelos podem ser implementados para modelar dados heterogêneos. Agora apelamos para esses métodos para construir um modelo capaz de lidar com dados heterogêneos classificados. Nesta seção, supomos que pode haver grupos de classificadores, cada um dos quais tem suas próprias crenças sobre a preferência das entidades. Nesse cenário, desejamos construir um modelo que permita que cada grupo de classificadores tenha seu próprio conjunto exclusivo de parâmetros de habilidade que resumem suas crenças sobre as entidades. Anteriormente, nossos modelos consideravam apenas um único vetor de parâmetro que resume todas as classificações, ou seja, consideramos apenas um único grupo de classificadores. Para implementar uma estrutura de agrupamento, agora precisamos de um vetor de parâmetro associado a cada classificador (λ_i para $i = 1, \dots, n$). Esses vetores de parâmetros não precisam ser únicos e, de fato, desejamos que os classificadores com as mesmas crenças sobre as entidades também compartilhem o mesmo vetor de parâmetros. Nossa modelo deve, portanto, ser construído de forma que permita $\Pr(\lambda_i = \lambda_j) \geq 0$ para todo $i \neq j$. Essa estrutura pode ser alcançada implementando uma distribuição anterior do processo de Dirichlet, onde os átomos do DP são vetores de parâmetros (único). Neste cenário (por $\alpha < \infty$), o processo de Dirichlet anterior especifica uma distribuição discreta sobre uma faixa de vetores de parâmetros e, portanto, se extraímos amostras dessa distribuição, teremos $\Pr(\lambda_i = \lambda_j) > 0$ para todo $i \neq j$. In o que se segue, descrevemos um novo modelo que consiste em uma mistura infinita de modelos de Plackett-Luce ponderado com um processo de Dirichlet como a distribuição a priori conjugada. Somos capazes de derivar um conjunto completo de distribuições condicionais completas para cada parâmetro de interesse em nosso modelo usando variáveis latentes. Isso nos permite formar um algoritmo de amostragem de Gibbs no qual apelamos para o Algoritmo 8 de Neal (Neal, 2000) para amostrar as variáveis indicadoras de cluster latente de forma eficiente. Observamos que grande parte da literatura sobre modelagem de dados heterogêneos tem sido limitada a modelos de mistura finita de modelos padrão de Plackett-Luce. No entanto, aqui permitimos uma heterogeneidade adicional entre as habilidades dos classificadores, assumindo que o modelo Weighted Plackett-Luce é o verdadeiro

distribuição de classificação subjacente. Também podemos relaxar a suposição de um número fixo de componentes a priori, construindo o modelo como um modelo de mistura de processos de Dirichlet. A seção é concluída com um estudo de simulação.

3.5.1 O modelo

Suponha que tenhamos n classificadores onde cada classificador relata posições para entidades $i \leq K$. Os principais componentes do nosso modelo completo, a mistura do processo Dirichlet dos modelos WeightedPlackett-Luce (WDP), podem ser escritos como

$$\begin{aligned} X_i | \lambda_i, w_i &\text{indep} \sim PLW(\lambda_i, w_i) & i = 1, \dots, n, \\ \lambda_i | G &\text{indep} \sim \\ G & \\ G | \alpha, G_0 &\sim DP(\alpha, G_0). \end{aligned}$$

Para tornar a forma da distribuição a priori não paramétrica inequívoca, definimos a representação de quebra de bastão que aqui é

$$\begin{aligned} G(\cdot) &= \sum_{s=1}^{\infty} \psi_s \delta_{\lambda_s}(\cdot) \\ \psi_s &= v_s \prod_{i' < s} (1 - v') \\ v_s &\text{indep} \sim Beta(1, \\ \alpha) \lambda_s &\text{indep} \sim G_0 \end{aligned}$$

para $s \in N$ e $k = 1, \dots, K$. Observe que, neste modelo, λ_{sk} agora denota o parâmetro skill, para a entidade k no grupo/cluster s , enquanto anteriormente o parâmetro skill de cada entidade era denotado λ_k (já que tínhamos apenas um único grupo de classificadores). Além disso, não há necessidade de escolher uma única distribuição de base G_0 . Em vez disso, uma distribuição de base única pode ser escolhida para cada um dos parâmetros de habilidade K , ou seja, podemos deixar λ_{sk} indep $\sim G_0 k$ na representação de quebra de bastão acima. Dito isso, a escolha da(s) distribuição(es) base(s) deve(m) ser intercambiável entre rótulos de cluster, dada a permutabilidade inerente dos componentes dentro do processo de Dirichlet anterior, ou seja, $G_0 k$ não deve depender de s .

3.5.2 Simulando dados da mistura do processo de Dirichlet dos modelos WeightedPlackett-Luce

Nesta seção, descrevemos como simular dados de nossa mistura de processos de Dirichlet de modelos ponderados de Plackett-Luce. É útil introduzir variáveis indicadoras de cluster latentes

ao lidar com modelos de mistura. Aqui introduzimos $\mathbf{cr} = (cr_1, \dots, cr_n)$ onde $cri = j$ denota que o ranker i está associado ao vetor de parâmetro λ_j . Por exemplo, se $\mathbf{cr} = (1, 2, 1, 2)$, os classificadores 1 e 3 estão no cluster 1 e os rankers 2 e 4 estão no cluster 2; Cada cluster tem um vetor de parâmetro exclusivo, aqui λ_1 e λ_2 . Além disso, deixamos $N_r = |\{cri=j\}|=1$ denotam o número de clusters de ranker únicos, ou seja, o número de vetores de parâmetros únicos, e seja $\Lambda = (\lambda_1, \dots, \lambda_{N_r})$ denotar a coleção desses vetores de parâmetros únicos.

Dados esses indicadores de cluster latentes, agora podemos descrever como gerar dados no modelo WDP descrito na seção anterior. Primeiro, precisamos especificar uma estrutura de agrupamento de classificação e isso pode ser alcançado (a) definindo explicitamente valores para as variáveis de alocação de cluster latentes cri e rotulando-as $1, \dots, N_r$, ou (b) extrair uma realização (marginalmente) da distribuição anterior do processo de Dirichlet da seguinte forma.

- Escolha $a > 0$ ou simule a partir de uma distribuição adequada.
- Defina $cri = 1$ e o número (atual) de clusters como $N_r = 1$.
- Para $i = 2, \dots, n$ simular a alocação do classificador i a um cluster de acordo com a distribuição discreta dada por

$$\Pr(cri = j | cr_1, \dots, cri-1) = \frac{nrij\alpha + i - 1}{\alpha + i - 1} \quad \text{para } j = 1, \dots, N_r,$$

$$\Pr(cri = N_r + 1 | cr_1, \dots, cri-1) = \dots,$$

onde $nrij$ denota o número de pontos atualmente dentro do cluster j (na iteração i)
e $N_r \rightarrow N_r + 1$ se $cri = N_r + 1$.

Dada uma estrutura de agrupamento dos classificadores, agora precisamos escolher valores para os parâmetros de habilidade (específicos do cluster) λ_{ik} . Novamente, eles podem ser escolhidos explicitamente ou, alternativamente, podem ser extraídos da distribuição anterior por amostragem $\lambda_{ik} \sim G_0 k$ para $s = 1, \dots, N_r, k = 1, \dots, K$. Lembre-se de que aqui temos uma mistura de modelos de Plackett-Luce ponderados e, portanto, precisamos escolher se cada classificação deve ser informativa ou não, ou seja, escolha (ou amostra) um valor de $w_i \in \{0, 1\}$ para $i = 1, \dots, n$. Depois que os parâmetros do modelo forem totalmente especificados, podemos usar a representação exponencial de variável latente do modelo de Plackett-Luce ponderado para gerar classificações. Uma coleção de n classificações completas $\{x_i\}_{i=1}^n$ pode ser gerada através do seguinte processo. Para $i = 1, \dots, n$,

- Amostra $vij \sim \text{Exp}(\lambda_{wicrij})$ para $j = 1, \dots, K$.
- Definir $x_{ij} = \begin{cases} 1 & \text{se } vij \leq S_{ij} \\ 0 & \text{caso contrário} \end{cases}$ onde $S_{ij} = K \setminus \{x_{i1}, \dots, x_{ij-1}\}$ para $j = 1, \dots, K$.

Tipos alternativos de classificações (como uma classificação entre os 5 primeiros) podem ser obtidos a partir das classificações completas simuladas usando o mesmo processo discutido na Seção 2.2.5.

3.5.3 Especificação prévia e variáveis latentes

Ao implementar pela primeira vez o modelo padrão (ou ponderado) de Plackett-Luce no Capítulo 2, discutimos como é vantajoso usar distribuições anteriores de gama nos parâmetros de habilidade, pois isso fornece atualizações conjugadas; ver secção 2.3. Neste modelo, a especificação prévia equivalente é alcançada deixando $G0k = Ga(ak, 1)$ que dá $\lambda_{sk} \sim Ga(ak, 1)$ para $k \in N$ e $k = 1, \dots, K$. Nossas crenças anteriores sobre a força da entidade k (em relação às outras entidades) são então expressas através do parâmetro ak . Lembre-se de que o parâmetro de taxa não é identificável com probabilidade e, portanto, consideramos que é 1. A distribuição anterior nos indicadores de capacidade binária latente permanece como antes, com $wi \sim Bern(pi)$ onde $pi \in (0, 1)$ para $i = 1, \dots, n$. Também desejamos inferir o parâmetro de concentração DP a partir dos dados e, portanto, precisamos especificar uma distribuição anterior: tomamos $\alpha \sim Ga(a\alpha, b\alpha)$. Antes de descrevermos o algoritmo de computação posterior (baseado em aumento de dados), precisamos definir as variáveis indicadoras de cluster latentes. Usamos os indicadores de cluster introduzidos na seção anterior, ou seja, $cr = (cr1, \dots, crn)$ onde $cri = j$ denota que a classificação está associada ao vetor de parâmetro λ_j . Lembre-se de que $Nr = |\{cri\}| = 1, \dots, n$ denota o número de clusters de ranker únicos e $\Lambda = (\lambda_1, \dots, \lambda_{Nr})$ é a coleção dos vetores de parâmetros de habilidade exclusivos.

Estamos agora em posição de definir a distribuição anterior sobre os parâmetros de habilidade neste modelo. O modelo contém Nr clusters ranker, cada um dos quais tem um vetor de parâmetro associado λ . O modelo contém parâmetros de habilidade exclusivos $N \times K$, portanto, condicional aos parâmetros de cluster latentes, a distribuição anterior de Λ é

$$p(\Lambda|cr) = \prod_{c=1}^{Nr} \prod_{k=1}^K \frac{\lambda_{ak} - 1}{\lambda_{ak} + \sum_{l \neq k} \lambda_{al}}.$$

O modelo assume que cada classificação segue a probabilidade ponderada de Plackett-Luce, ou seja, $Xi | \Lambda, w, cr \sim PLW(\lambda_{cri}, wi)$. Portanto, a probabilidade da i -ésima classificação pode ser expressa usando as variáveis indicadoras de cluster latentes como

$$Pr(Xi = ha | \Lambda, w, cr) = \prod_{j=1}^{Nr} \frac{\lambda_{wicri,xij} \sum_{m=1}^{Nm} \lambda_{wicri,color}}{\lambda_{wicri,xij} + \sum_{m \in Uj} \lambda_{wicri,m}},$$

e assim, como os rankings são (condicionalmente) independentes, a probabilidade de todos os n rankings é

$$\pi(D|L, w, cr) = \prod_{i=1}^n \frac{\lambda_{wicri,xij} \sum_{m=1}^{n_i} \lambda_{wicri,color}}{\sum_{m \in U_i} \lambda_{wicri,m}}.$$

Conforme discutido no Capítulo 2, a probabilidade de Plackett-Luce não admite inferência bayesiana conjugada. No entanto, vimos que as atualizações conjugadas para os parâmetros de habilidade podem ser alcançadas aumentando o espaço de parâmetros com variáveis latentes apropriadas. Aqui, as variáveis latentes ainda são definidas em termos dos tempos exponenciais latentes entre chegadas, mas agora são baseadas em parâmetros de habilidade específicos do cluster, ou seja,

$$\text{z}_{ij}|D, \Lambda, w, cr \text{ indep} \sim \frac{\sum_{m=j}^n \lambda_{wicri,xij} + \lambda_{wicr_i}}{\sum_{m \in U_i} \lambda_{wicr_i}}, \quad (3.7)$$

pois $i = 1, \dots, n$ e $j = 1, \dots, n_i$.

3.5.4 Distribuições condicionais completas

A distribuição posterior é formada pela aplicação do Teorema de Bayes. A distribuição posterior $\pi(Z, \Lambda, w, cr|D)$ é agora uma distribuição conjunta das variáveis aleatórias latentes Z, a coleção de parâmetros de habilidade únicos Λ , as variáveis indicadoras binárias w, os indicadores latentes cr e o parâmetro de concentração DP α . Podemos obter realizações da distribuição posterior usando uma estratégia de amostragem de Gibbs que coleta amostras dos FCDs de cada quantidade desconhecida por sua vez. Os indicadores de cluster latente podem ser extraídos de suas respectivas distribuições condicionais completas usando o Algoritmo 8 de Neal (Neal, 2000). Além disso, o FCD do parâmetro de concentração DP α também é conhecido; ver ponto 3.4.2. No restante desta seção, derivamos os FCDs para as quantidades desconhecidas restantes (Z, Λ, w) e um esboço completo do esquema MCMC usado para gerar amostras posteriores pode ser encontrado na Seção 3.5.5.

Antes de iniciar a derivação das distribuições condicionais completas, é útil primeiro construir a densidade de todas as grandezas estocásticas. Note-se, no entanto, que, dado que já temos os FCDs para cr e α , é sensato considerar apenas a densidade condicional de todas as quantidades estocásticas restantes, dadas as variáveis indicadoras de cluster latente e o parâmetro DPconcentration. Esta densidade de junta (condicional) é

$$p(L, D, Z, w|cr, \alpha) = p(L, D, Z, w|cr) = p(Z|D, L, w, cr)\pi(D|L, w, cr)\pi(L|cr)\pi(w)$$

$$\begin{aligned}
 &= \prod_{i=1}^n \prod_{j=1}^m \sum_{m=j}^{\infty} \lambda wicri, xij + \lambda wic \prod_{r, m} \frac{\prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic}{\prod_{r, m} \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic} \\
 &\times \prod_{i=1}^n \prod_{j=1}^m \lambda wicri, xij \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic \prod_{r, m} \frac{\prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic}{\prod_{r, m} \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic} \\
 &= \prod_{i=1}^n \prod_{j=1}^m \exp \frac{\prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic}{\prod_{r, m} \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic} \\
 &\quad \times \prod_{i=1}^n \prod_{k=1}^K \frac{\lambda^{ak-1} c^{ke}}{\lambda^{ck} \Gamma(ak)} \prod_{i=1}^n p_{wi} (1 - p_i)^{1-wi} \\
 &\quad \times \prod_{i=1}^n \prod_{k=1}^K \frac{\lambda^{ak-1} c^{ke}}{\lambda^{ck} \Gamma(ak)} \prod_{i=1}^n p_{wi} (1 - p_i)^{1-wi}.
 \end{aligned} \tag{3.8}$$

As distribuições condicionais completas (FCDs) podem ser obtidas a partir dessa densidade construindo a distribuição condicional de cada quantidade aleatória dadas todas as outras quantidades. Os parâmetros latentes $Z = \{z_{ij}\}$ são definidos por meio de sua distribuição condicional completa e, portanto, não deve ser surpresa que obtenhamos

$$p(Z|D, L, w, cr, a) \propto \prod_{i=1}^n \prod_{j=1}^m \frac{\prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic}{\prod_{r, m} \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic},$$

e, portanto, as distribuições condicionais completas para z_{ij} são como em (3.7) para $i = 1, \dots, n, j = 1, \dots, n_i$.

A distribuição condicional completa para os parâmetros de habilidade λ^{ck} (específicos do cluster) é derivada da seguinte forma:

$$\begin{aligned}
 &p(L|D, Z, w, cr, a) \\
 &\propto \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda wicri, xij}{\exp \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic} \prod_{r, m} \frac{\lambda^{ak-1} c^{ke}}{\prod_{c=1}^K \prod_{k=1}^N \lambda^{ck} \Gamma(ak)} \\
 &\propto \prod_{c=1}^K \prod_{k=1}^N \frac{\lambda^{ak+ck} e^{-\lambda^{ck}}}{\lambda^{ck} \Gamma(ak)} \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda wicri, xij}{\exp \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic} \prod_{r, m} \frac{\lambda^{ak-1} c^{ke}}{\prod_{c=1}^K \prod_{k=1}^N \lambda^{ck} \Gamma(ak)} \\
 &= \prod_{c=1}^K \prod_{k=1}^N \frac{\lambda^{ak+ck}}{\lambda^{ck} \Gamma(ak)} \frac{\prod_{i=1}^n \prod_{j=1}^m \lambda wicri, xij}{\prod_{r, m} \prod_{m=j}^{\infty} \lambda wicri, xij + \lambda wic} \frac{\prod_{c=1}^K \prod_{k=1}^N c^{ke}}{\prod_{c=1}^K \prod_{k=1}^N \lambda^{ck} \Gamma(ak)},
 \end{aligned}$$

onde

$$y_{ck} = \sum_{i=1}^n w_i I(c' = c) I(k \in \{x_i'\})$$

é o número de classificações informativas associadas ao cluster c no qual a entidade k aparece

e

$$\zeta_{ij}(c, k) = I(\text{cri} = c) \times I(k \in \{x_{ij}, \dots, x_{ini}\} \cup U_i),$$

é uma função indicadora de que a entidade k recebe uma classificação não melhor que j na classificação i , onde a classificação i também está associada ao cluster c . Segue-se que o FCD para λck está disponível em formato fechado e é

$$\lambda c k | \dots \quad \square \quad \square a k + y c k, 1 + \sum_{i=1}^n W \sum_{j=1}^c \zeta_{ij}(c, k) z_{ij} \quad \square, \quad (3.9)$$

para c = 1, ..., N r, k = 1, ..., K.

As únicas variáveis aleatórias restantes no modelo são os pesos do classificador w . A distribuição condicional completa para w_i é a distribuição discreta com

$$\Pr(w_i = 1 | D, \Lambda, Z, w - i, cr, \alpha) \propto \Pr(w_i = 1) \pi(D|w_i = 1, \Lambda, w - i, cr) \pi(Z|w_i = 1, \Lambda, D, w - i, cr)$$

$$\propto \frac{\prod_{j=1}^k \lambda c_{ri,xij}}{\exp(-\sum_{m=j}^k \lambda c_{ri,xim})} \sum_{m=j}^k \lambda c_{ri,m}$$

e

$$\Pr(w_i = 0 | D, \Lambda, Z, w_{-i}, cr, \alpha) \propto \Pr(w_i = 0) \pi(D|w_i = 0, \Lambda, w_{-i}, cr) \pi(Z|w_i = 0, \Lambda, D, w_{-i}, cr)$$

$$\propto (1 - p_i) \prod_{j=1}^{\ell} \exp \{-\text{side}(K_i - j + 1)\}.$$

Portanto, para $i = 1, \dots, n$, a distribuição condicional completa é dada por

$$wi \mid \dots \text{indep} \sim \text{Berna}(\pi), \quad (3.10)$$

onde

$$\rho_i = \frac{\Pr(w_i = 1 | D, \Lambda, Z, w_{-i}, cr, \alpha) \Pr(w_i = 1 | D, \Lambda, Z, w_{-i}, cr, \alpha)}{\Pr(w_i = 0 | D, \Lambda, Z, w_{-i}, cr, \alpha)},$$

é a probabilidade de que a classificação seja informativa (dadas as outras quantidades)

Lembre-se de que os indicadores de cluster latente cr podem ser amostrados usando o Algoritmo 8 de Neal (Neal, 2000), que implementa um esquema de urna de Pólya para marginalizar os parâmetros dimensionais infinitos. A distribuição condicional completa resultante para as alocações de cluster é uma distribuição discreta sobre os clusters que são ativos e m componentes auxiliares. O parâmetro de concentração DP também pode ser amostrado a partir da sua distribuição condicional completa

indicado na seção 3.4.2. A seção a seguir fornece um esboço completo do amostrador de Gibbs usado para obter realizações posteriores.

3.5.5MCMC usando o Algoritmo de Neal 8

Estamos agora em posição de descrever o algoritmo usado para amostrar a partir da distribuição posterior $\pi(\Lambda, Z, cr, w, q|D)$. Primeiro, definimos a contribuição para a probabilidade de dados completos do ranker i ser

$$f(x_i, z_{ij}|\Lambda, w, cr, \alpha) = \prod_{j=1}^{\bar{q}} \frac{\lambda_{wicri,xij}}{\exp \left(\sum_{m=j}^{\bar{q}} \lambda_{wicri,xim} + \sum_{m \in U_i} \lambda_{wicri,xi_m} \right)}.$$

Agora podemos descrever o algoritmo de forma concisa. Suponha que tenhamos n classificações de $ni < K$ entidades. O estado da cadeia de Markov tem elementos $\Lambda = (\lambda_c : c \in \{cr1, \dots, cm\})$, $Z = (z_{ij})$, $cr = (cri)$, $w = (wi)$ e q para $i = 1, \dots, n$, $j = 1, \dots, ni$. O algoritmo repetidamente faz as seguintes amostras:

- Para $i = 1, \dots, n$: Seja $q-$ o número de cri_j distintos para $j \neq i$ e $h = q- + m$. Rotule esses valores de cri_j em $\{1, \dots, q-\}$. Se $cri_j = cri_k$ para algum $j \neq i$, desenhe λ_{ck} indep~ $\sim G0k$ for $q- < c \leq h$, $k = 1, \dots, K$. Se $cri_j = cri_k \forall j \neq i$, seja cri_j o rótulo $q- + 1$, e desenhe λ_{ck} indep~ $\sim G0k$ para $q- + 1 < c \leq h$, $k = 1, \dots, K$. Desenhe um novo valor para cri_j de $\{1, \dots, h\}$ usando as seguintes probabilidades: $Pr(cri_j = c|D, Z, \Lambda, cr-i, w, q-) = \frac{\lambda_c}{\sum_{c=1}^{q-} \lambda_c}$ se $c \neq cri_i$, $f(x_i, z_{ij}|\Lambda, w, cri = c, cr-i, q-) = \frac{\lambda_c}{\sum_{c=1}^{q-} \lambda_c} f(x_i, z_{ij}|\Lambda, w, cri = c, cr-i, q-)$ se $c = cri_i$, $0 < c \leq h$, onde $\Lambda = \{\lambda_1, \dots, \lambda_h\}$, $n-i, c = \#\{cri_j = c, j \neq i\}$, e b é a constante de normalização apropriada. Altere o estado para conter apenas aqueles λ_c que agora estão associados a uma ou mais observações, ou seja, seja $\Lambda' = (\lambda_c : c \in \{cr1, \dots, cm\})$.

—

- Amostra λ_{ck} de (3.9) para $c = 1, \dots, N$, $r, k = 1, \dots, K$.
- Amostra z_{ij} de (3.7) para $i = 1, \dots, n$, $j = 1, \dots, ni$.
- Amostra wi de (3.10) para $i = 1, \dots, n$.
- Redimensionar – Amostra Λ'
 $\sim Ga(N | \sum_{k=1}^K \lambda_k)$.

- Calcule $\Sigma = \sum_{c=1}^{\text{Sem K}} \sum_{k=1}^{l_c} l_{ck}$.
- Para $c = 1, \dots, N$ e $k = 1, \dots, K$, seja $\lambda_{ck} \rightarrow \lambda_{ck} L_c^T / \Sigma$.

- Exemplo a como na Seção 3.4.2 com $n = n$ e $N c = N r$. Observe que esta etapa de reescalonamento é uma generalização direta daquela discutida na Seção 2.2.3. A generalização é necessária, pois agora temos K parâmetros de habilidade exclusivos em cada um dos clusters de classificação N .

3.5.6 Estudo de simulação – revisitar o conjunto de dados 2

Para nosso primeiro estudo de simulação, optamos por revisitar o Conjunto de Dados 2 introduzido no Capítulo 2. Lembre-se de que este conjunto de dados contém $n = 50$ classificações, as primeiras 40 das quais são classificações informativas e as 10 restantes (rotuladas de 41 a 50) são permutações não informativas / aleatórias. Cada classificação dentro deste conjunto de dados é uma classificação completa de $K = 20$ entidades.

Antes de podermos realizar a inferência bayesiana, devemos primeiro escolher uma distribuição a priori adequada. Como em nossas análises anteriores desses dados, optamos por deixar cada ordenação das entidades ser igualmente provável a priori, ou seja, deixe $a_k = 1$ para todo k com a distribuição a priori resultante sobre os parâmetros de habilidade sendo $A\theta \sim \text{Ga}(1, 1)$. Especificar uma escolha (anterior) para o parâmetro de concentração do processo de Dirichlet é um tanto difícil e, portanto, colocamos uma distribuição anterior sobre α . Escolhemos $a_\alpha = b_\alpha = 1$ para que $\alpha \sim \text{Ga}(1, 1)$, o que dá uma distribuição prévia bastante fraca sobre o número de clusters ranker. Observe que aqui o número modal anterior de clusters de classificação é 1 (com probabilidade de 0,19) e, portanto, parece razoável, dada a natureza deste conjunto de dados - consulte a Tabela 3.1 para a distribuição anterior completa sobre o número de clusters. Também precisamos escolher probabilidades prévias de que cada ranker seja informativo, ou seja, especificar $p_i = \Pr(w_i = 1)$ para cada ranker i . Aqui consideramos 3 análises, cada uma definida por escolhas particulares do p_i . As Análises 1 e 2 tomam a especificação equivalente aos estudos considerados na Seção 2.7, ou seja, para a Análise 1 deixamos $p_i = 0,5$ (cada classificador tem a mesma probabilidade de ser informativo e não informativo) e na Análise 2 tomamos $p_i = 0,8$ (a verdadeira proporção dentro desses dados). Para a análise final (Análise 3), assumimos o modelo padrão de Plackett-Luce, que é alcançado por taking $p_i = 1$. Essa escolha é usada para avaliar o quanto robusta é nossa análise para assumir que todos os classificadores são informativos quando, na verdade, há classificadores não informativos no conjunto de dados. Intuitivamente, podemos pensar que o modelo de mistura do processo de Dirichlet seria flexível o suficiente para agrupar as classificações informativas e formar um cluster separado para abrigar as classificações não informativas. No entanto, como veremos, esse não é o caso. As análises 1 e 2 nos permitem comparar o desempenho de nosso modelo de mistura DP em comparação com o modelo homogêneo (monocomponente) considerado no Capítulo 2.

Análise posterior

Para gerar realizações a partir da distribuição posterior (para cada análise), implementamos o algoritmo de amostragem descrito na Seção 3.5.5 com $m = 2$ variáveis auxiliares. Cada cadeia de Markov foi inicializada em um sorteio aleatório da distribuição anterior. Para obter realizações 10K (quase) não autocorrelacionadas da distribuição posterior, precisávamos diminuir a saída por fatores de 60, 20 e 5 para as análises 1–3, respectivamente. Portanto, executamos o esquema para iterações de 600K, 200K e 50K para cada análise respectiva e também permitimos a cada cadeia um período de burn-in de 10K iterações após a inicialização - essas amostras foram descartadas. O tempo computacional necessário para realizar a inferência foi de (aproximadamente) 126, 33 e 11 segundos para cada análise. A mistura das cadeias MCMC foi avaliada inspecionando gráficos de rastreamento da probabilidade de dados logarítmicos completos; ver figura 3.3. Isso é conveniente não apenas porque nosso espaço de estados é vasto, mas também porque a dimensão da distribuição posterior pode mudar a cada iteração (dependendo do número de grupos de rankers únicos). Portanto, não é realista inspecionar gráficos de rastreamento de parâmetros individuais dentro da cadeia de Markov, principalmente porque os rótulos de cluster podem ser trocados arbitrariamente. A convergência foi avaliada inicializando várias cadeias em diferentes valores iniciais e verificando se as distribuições posteriores resultantes eram equivalentes (até o ruído estocástico).

Começamos determinando a distribuição posterior formada na Análise 3 ($\pi = 1$) – assumindo uma mistura de processo de Dirichlet de modelos padrão de Plackett-Luce. Nossa intuição a priori nos levou a acreditar que nosso modelo de mistura (descrito na Seção 3.5.1) pode permitir a formação de um cluster que abriga os classificadores informativos e um cluster separado para abrigar os classificadores não informativos. A distribuição marginal posterior para o número de ranker

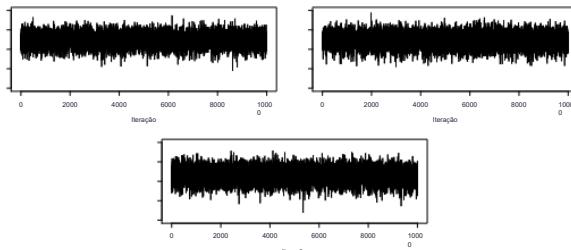


Figura 3.3: Gráficos de rastreamento da probabilidade de dados logarítmicos completos para as Análises 1, 2: $\pi = 0,5, 0,8$ (superior esquerdo e direito, respectivamente) e Análise 3: $\pi = 1$ (inferior)

Análise	eu						E	SD
	12	34	56	78	9	≥ 10		
1	0.86	0.12	0.02	0.00	0.00	0.00	1.16	0.42
2	0.75	0.19	0.05	0.01	0.00	0.00	1.32	0.62
3	0.00	0.01	0.04	0.11	0.20	0.24	0.20	0.12
Prévio	0.19	0.17	0.15	0.12	0.10	0.07	0.06	0.04
					0.03	0.07	4.18	3.00

Tabela 3.1: Probabilidades posteriores do número de clusters de classificadores, $Pr(N = r | D)$, para cada uma das três análises. A expectativa e o desvio padrão da distribuição marginal posterior também são mostrados junto com a distribuição anterior. Os valores modais são destacados em negrito.

grupos (Tabela 3.1) dá $Pr(N = 2 | D) = 0,01$ e, portanto, há pouco suporte posterior para essa sugestão. Talvez surpreendentemente, o número modal posterior de grupos de ranker aqui é seis. No entanto, notamos que há uma grande incerteza sobre isso. Em contraste, para as Análises 1 e 2, vemos suporte posterior significativo para um único grupo homogêneo com $Pr(N = 1 | D) = 0,86$ e 0,75, respectivamente. Permitir claramente a incerteza sobre a confiabilidade do ranker resulta em uma mudança significativa nas crenças posteriores sobre os grupos de rankers contidos nesses dados - essa é uma característica do modelo que será discutida com mais detalhes posteriormente.

A distribuição posterior marginal para o número de clusters de classificação fornece uma visão útil da distribuição posterior; no entanto, não conta a história completa. Um resumo mais aprofundado da distribuição posterior pode ser obtido se considerarmos a estrutura de agrupamento subjacente dos classificadores. A distribuição posterior da alocação de classificadores para grupos de classificadores é, obviamente, bastante complexa. Uma maneira comum de resumir a heterogeneidade do ranker é por meio de uma única alocação resumida para cada grupo de rankers, como a alocação máxima posterior (MAP) ou as melhorias na alocação do MAP propostas por Dahl (2006) e Lau e Green (2007). No entanto, esses resumos podem ser enganosos, a menos que a probabilidade posterior do número modal de grupos seja bastante grande. Observe que, para a distribuição posterior da Análise 3, esse certamente não é o caso. Em vez disso, preferimos resumir a heterogeneidade do ranker usando probabilidades de dissimilaridade $\Delta_{ij} = Pr(\text{cri } i \neq \text{cri } j | D)$, ou seja, a probabilidade posterior de que dois classificadores (i e j) não sejam alocados ao mesmo cluster. A alocação de classificadores para grupos poderia então ser determinada pelo limite dessas probabilidades. No entanto, isso também pode sofrer de alocações inconsistentes de, digamos, rankertriples, particularmente quando suas probabilidades de dissimilaridade estão próximas do limite. Portanto, seguindo Medvedovic e Sivaganesan (2002), usamos um método resumido padrão da análise de agrupamento, ou seja, um dendrograma calculado a partir das probabilidades de dissimilaridade Δ_{ij} . Observe que consideramos os dendrogramas formados usando o método de ligação completa, também conhecido como agrupamento de vizinhos mais distantes. Este método tende a produzir clusters mais densamente compactados e não sofre de "encadeamento"; ver Everitt et al. (2011) para mais informações

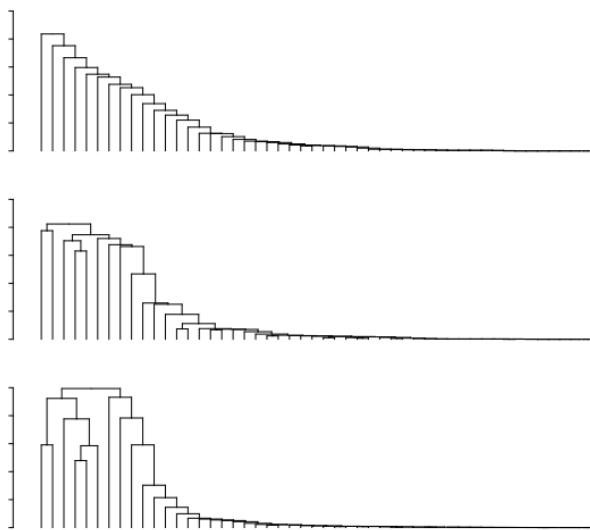


Figura 3.4: Dendrogramas de ligação completa com base nas diferenças entre cada par de classificadores para as análises 1–3 de cima para baixo, respectivamente.

detalhes sobre métodos de ligação. É claro que muitos outros métodos também podem ser usados para resumir a heterogeneidade entre os classificadores, ver, por exemplo, Rastelli e Friel (2017) e as referências neles.

A Figura 3.4 mostra os dendrogramas calculados a partir das matrizes de dissimilaridade para cada uma das análises consideradas. A alocação de classificadores para grupos de classificadores para as Análises 1 e 2 é um tanto trivial, dado $\Pr(N_r = 1|D) = 0,86$ e $0,75$, respectivamente. Os dendrogramas correspondentes confirmam que todos os classificadores são frequentemente agrupados; evidente através dos valores de dissimilaridade em que os classificadores se juntam ao grupo principal. No entanto, é encorajador ver que os classificadores não informativos (com exceção do classificador 42) são os últimos a ingressar no cluster principal: esses classificadores têm os maiores valores de dissimilaridade. Na Análise 3, a alocação de

classificadores para grupos não é tão simples. O dendrograma correspondente mostra que há um grande aglomerado contendo esses classificadores numerados {1, 2, ..., 40, 42}. Essa conclusão pode ser tirada uma vez que $\Delta_{ij} \leq 0,30 \Rightarrow (1 - \Delta_{ij}) > 0,70$ para $i = j \in \{1, 2, \dots, 40, 42\}$, ou seja, qualquer par de classificadores dentro deste conjunto é agrupado pelo menos 70% do tempo. Dada a grande proporção do tempo em que esses classificadores são co-agrupados, é razoável concluir que eles têm crenças semelhantes sobre as entidades. Os classificadores restantes, aqueles numerados 41, 43, ..., 50, normalmente têm uma dissimilaridade maior que 0,5. Fica claro, olhando para o lado esquerdo do dendrograma, que não há uma estrutura de agrupamento clara entre qualquer um desses classificadores. Isso talvez não seja surpreendente, dado que suas classificações associadas são permutações aleatórias e, portanto, provavelmente expressam preferências contraditórias.

Agora voltamos ao ponto que observamos anteriormente, ou seja, que permitir a incerteza sobre a classificação muda as crenças posteriores sobre o número de grupos de classificadores. Depois de investigar a distribuição posterior para cada análise, talvez esse resultado não seja tão surpreendente quanto parece à primeira vista. Na Análise 3, o modelo padrão de Plackett-Luce não tem flexibilidade para reduzir a contribuição que os classificadores não informativos fazem para a probabilidade geral, e isso leva à formação de clusters adicionais para abrigar os classificadores que não são consistentes com os outros (os rankings não informativos); ver figura 3.4. Além disso, esses rankers nem mesmo formam um único agrupamento homogêneo devido à alta variação nas permutações aleatórias (como mencionado anteriormente). Por outro lado, nas Análises 1 e 2 (mistura de modelos de Plackett-Luce Ponderado) o modelo é capaz de reduzir o peso dos rankers não informativos; veja a Figura 3.5. Lembre-se de que quando $w_i = 0$ a probabilidade de classificação i é constante ($\Pr(X_i = x_i | \lambda, w_i = 0) = 1 / P(K_i, n_i)$) e não depende de λ . Assim, um classificador que é considerado não informativo é livre para ingressar em um cluster, independentemente de suas crenças sobre as entidades, pois a probabilidade não é afetada. De fato, esse classificador normalmente se juntará ao maior cluster "ativo" - isso decorre da noção de enriquecimento que sustenta o processo de Dirichlet (conforme mencionado na Seção 3.3.1). Consequentemente, não é surpreendente que os rankings não informativos (41-50) se juntem ao cluster principal, ou seja, o cluster que abriga os classificadores informativos nas Análises 1 e 2.

Concluímos esta seção com uma breve comparação das Análises 1–3 e aquelas em que assumimos que todos os classificadores eram homogêneos em suas crenças sobre as entidades no Capítulo 2. Existem semelhanças significativas entre as distribuições posteriores dos pesos dos classificadores nas Análises 1 e 2 nesta seção e na Seção 2.7; ver figuras 3.5 e 2.4. Isso talvez não seja surpreendente, dado que os pesos do classificador não são específicos do cluster e temos suporte posterior significativo para um único cluster de ranker nessas análises. Observe que, quando há apenas um único cluster de classificadores, as análises apresentadas aqui são análogas às da Seção 2.7. As classificações agregadas formadas nas Análises 1 e 2 aqui são muito semelhantes às das análises homogêneas correspondentes consideradas na Seção 2.7; ver Tabelas 2.3 e 3.2. Observe que aqui a classificação agregada é determinada ordenando o

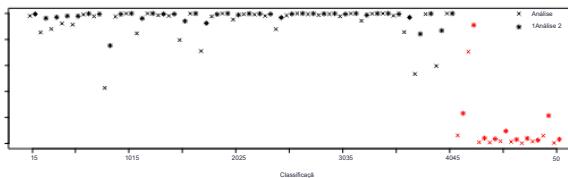


Figura 3.5: $\Pr(w_i = 1|D)$ – Probabilidade posterior de que a classificação i seja informativa em cada análise (Análise 1: $p_i = 0,5$, Análise 2: $p_i = 0,8$). As classificações que são permutações aleatórias (41–50) são mostradas em vermelho.

média da distribuição posterior (totalmente) marginal para cada entidade (marginalizada sobre clusters). Lembre-se de que, para as análises homogêneas no Capítulo 2, observamos que as classificações agregadas para a análise do Conjunto de Dados 2 sob o modelo Weighted Plackett-Luce foram equivalentes àquelas formadas pela análise do Conjunto de Dados 1 sob o modelo padrão de Plackett-Luce. Este é o caso, pois o modelo WPL é capaz de identificar corretamente os rankers não informativos e reduzi-los. A mesma conclusão pode ser tirada aqui; ver Figura 3.5 e Tabela 3.2. Sem surpresa, para a Análise 3 ($p_i = 1$; o modelo padrão de Plackett-Luce), a classificação agregada é afetada pelas informações prévias enganosas que afirmam que os classificadores não informativos são informativos. Isso também foi observado ao considerar a análise homogênea sob o modelo padrão de Plackett-Luce; ver seção 2.4.

A distribuição posterior da Análise 3 sugere claramente que há uma heterogeneidade significativa entre as crenças dos rankers; ver Tabela 3.1. Resumir esses dados heterogêneos por meio de uma classificação agregada geral talvez não seja sensato. As diferenças de preferências entre os grupos de classificação são facilmente vistas através das classificações agregadas dentro do cluster. Tal agregado é formado primeiramente condicionando um número apropriado de grupos de classificação e , em seguida, ordenando as médias marginais posteriores dos parâmetros de habilidade dentro de cada grupo. Para a Análise 3, condicionando em 6 grupos de classificadores (o modo posterior), a classificação agregada dentro do cluster para o cluster de classificadores 1 (aquele que normalmente abriga classificadores informativos) é muito semelhante ao agregado geral nas outras análises. Os agregados dentro do cluster restantes (aqueles para clusters de classificação 2–6) mostram pouca coerência com a verdadeira ordem de preferência da entidade e , e, em vez disso, parecem ser permutações aleatórias das entidades. Isso talvez não seja surpreendente, pois esses clusters normalmente abrigam os rankers não informativos.

Pessoas com deficiência heterogêneas PL homogêneo											
$\pi = 0,5$		$\pi = 0,8$		$\pi = 1$		Conjunto de dados 2		Conjunto de dados 2		Conjunto de dados 2	
λ	L	Xagg	1	Xagg	1	Xagg	1	xagg	11	xagg	2
1	20.00	3	30.91	3	25.94	3	7.203	27.47	3	11.67	
2	19.00	1	26.27	1	22.70	2	6.411	25.83	1	11.44	
3	18.00	5	25.73	5	22.28	1	6.365	23.90	2	11.29	
4	17.00	2	24.94	2	21.59	5	6.102	22.66	4	9.84	
5	16.00	4	20.73	4	17.58	4	5.214	18.11	9	9.54	
6	15.00	6	18.97	6	16.41	9	5.066	17.98	5	9.41	
7	14.00	8	18.72	8	16.35	8	5.028	17.31	8	9.11	
8	13.00	7	16.94	7	14.85	6	4.827	16.12	6	8.55	
9	12.00	9	16.10	9	14.59	7	4.359	15.91	7	8.10	
10	11.00	10	14.7810	10	13.4710	10	4.1410	14.5010	7	7.48	
11	10.00	11	12.1411	11	10.5811	11	3.3811	11.8411	6	6.36	
12	9.00	12	10.2012	12	9.2012	12	2.8612	9.9512	5	5.42	
13	8.00	13	8.6813	13	7.7413	13	2.7213	9.0013	5	5.02	
14	7.00	14	7.2814	14	6.6016	14	2.2514	7.1714	4	4.17	
15	6.00	15	7.2816	15	6.4814	15	2.1916	7.0816	3	3.98	
16	5.00	15	5.1715	15	4.7715	15	1.9015	4.9915	3	3.47	
17	4.00	17	4.5717	17	4.2117	17	1.6117	4.5817	3	3.10	
18	3.00	18	2.8518	18	2.6719	18	1.5718	2.8119	2	2.16	
19	2.00	19	2.6219	19	2.5918	19	1.3419	2.6818	2	2.08	
20	1.00	20	1.0020	20	1.0020	20	1.0020	1.0020	1	1.00	

Tabela 3.2: Classificações agregadas sob a mistura infinita do modelo ponderado de Plackett-Luce para a análise do conjunto de dados 2 (para análises 1–3; $\pi = 0,5, 0,8, 1$) junto com as médias posteriores correspondentes. Os resultados da Tabela 2.2 (análises padrão homogêneas de Plackett-Luce) também são fornecidos para facilitar a comparação.

3.6 Descobrindo subgrupos de entidades em um grupo de classificadores

Começamos observando que, até agora, nesta tese, assumimos que a preferência (força) de cada entidade é resumida por um parâmetro de habilidade único λ . No entanto, nesta seção, agora consideraremos a noção de que um grupo (homogêneo) de classificadores pode não ser capaz de distinguir entre algumas entidades, ou seja, eles acreditam que algumas entidades estão ligadas em força. Para permitir isso, consideramos uma distribuição a priori alternativa não paramétrica que permite que as entidades se agrupem. Para alcançar o agrupamento nas entidades, consideramos o processo de Dirichlet anterior nos parâmetros de habilidade, ou seja, tomamos uma distribuição discreta (de dimensão infinita) sobre λ_k tal que $\Pr(\lambda_i = \lambda_j) = 0$ para $i \neq j$. Observe que no que se segue, consideramos apenas o cenário em que há um único grupo de classificadores ($cr = 1$); relaxamos essa suposição no Capítulo 4. Portanto, ao contrário da seção anterior, assumimos que todos os classificadores compartilham crenças semelhantes sobre as entidades.

3.6.1 O modelo

Suponha que tenhamos classificações de n classificadores onde o classificador i relata posições (classificações) para entidades $n < K$. Aqui consideramos apenas um único grupo de classificadores e, portanto, o parâmetro λ da entidade k é denotado por λ_k . Os classificadores são considerados homogêneos e, portanto, seus rankings seguem a mesma distribuição de classificação subjacente, definida pelo modelo WeightedPlackett-Luce. O modelo aqui descrito é, portanto, semelhante ao considerado na Seção 2.5. No entanto, ao contrário deste cenário anterior (onde cada parâmetro de habilidade seguia uma distribuição contínua única), todos os parâmetros de habilidade K agora seguem uma distribuição discreta dimensional infinita G , onde G segue um processo de Dirichlet. Este modelo pode ser resumido

$$\begin{aligned} X_{ij} | \lambda \text{ indep} &\sim PLW(\lambda, w_i) & i = 1, \dots, n, \\ \lambda_k | G \text{ indep} &\sim G & k = 1, \dots, K, \\ G | \alpha, G_0 &\sim DP(\alpha, G_0). \end{aligned}$$

Para tornar a forma da distribuição a priori não paramétrica inequívoca, definimos sua representação de quebra de bastão, e isso é

$$\begin{aligned} G(\cdot) &= \sum_{s=1}^{\infty} \psi_s \delta_{\lambda_s}(\cdot) \\ \psi_s &= v_s \prod_{i < s} (1 - v_i) \\ v_s &\text{ indep} \sim \text{Beta}(1, \alpha) \lambda_s \\ \text{indep} &\sim G_0 \end{aligned}$$

para $s \in N$. Dada a forma da construção de quebra de bastão, é claro que os átomos do processo de Dirichlet são de fato quantidades escalares e não vetores de parâmetros como na Seção 3.5. No entanto, ao contrário da implementação normal do modelo de Plackett-Luce ponderado, seus elementos não precisam ser únicos. Outra característica importante deste modelo (em comparação com o modelo que leva em conta a heterogeneidade do ranker) é que não temos mais a liberdade de especificar uma distribuição de base única para cada uma das K entidades. Isso decorre da permutabilidade dos átomos dentro do processo de Dirichlet e, portanto, G_0 não deve depender de s . A implicação desta restrição na nossa especificação anterior será discutida mais adiante na Seção 3.6.3.

3.6.2 Simulando dados do modelo de Plackett-Luce ponderado com agrupamento de entidades

Nesta seção, descrevemos como simular dados do modelo de Plackett-Luce ponderado com um processo de Dirichlet antes dos parâmetros de habilidade da entidade. Como mencionado anteriormente, a introdução de variáveis indicadoras de cluster latentes é útil ao lidar com modelos de mistura. Aqui, adotamos indicadores e notações de agrupamento latente semelhantes aos da Seção 3.5, onde consideramos o agrupamento em classificadores. Supomos que existam N e clusters de entidades, ou seja, o vetor parâmetro para os parâmetros de habilidade K contém N e valores únicos. A coleção desses parâmetros únicos é denotada $\Lambda = (\lambda_1, \dots, \lambda_N)$. Além disso, também usamos indicadores de associação de cluster latente $ce = (ce_1, ce_2, \dots, ce_K)$ com $ce_k \in \{1, 2, \dots, N\}$ para $k = 1, 2, \dots, K$. Por exemplo, se $ce = (1, 2, 1, 2)$ então as entidades 1 e 3 estão no cluster 1 e as entidades 2 e 4 estão no cluster 2; Os dois clusters têm os parâmetros únicos $\lambda_1 = \lambda_2$. Isso equivale a dizer que as entidades 1 e 3 são equivalentes e são consideradas vinculadas em força; O mesmo também é verdade para as entidades 2 e 4. Usando esses indicadores de cluster latentes, o vetor de parâmetro de habilidade completo, aqui denotado $\lambda \dagger$, que contém o parâmetro para cada entidade é dado por $\lambda \dagger_k = \lambda cek$ para $k = 1, \dots, K$. Agora podemos descrever como gerar dados sob este modelo (descrito na Seção 3.6.1). Novamente, primeiro devemos especificar uma estrutura de cluster, mas desta vez para as entidades e não para os rankers. Isso pode ser alcançado (a) definindo explicitamente valores para as variáveis de alocação de cluster latentes cek e certificando-se de que elas sejam rotuladas como $1, \dots, N$ e ou (b) extraíndo uma realização (marginalmente) da distribuição anterior do processo de Dirichlet da seguinte forma.

- Escolha $a > 0$ ou simule a partir de uma distribuição adequada.
- Defina $ce_1 = 1$ e o número (atual) de clusters como N e $= 1$.
- Para $k = 2, \dots, K$ simula a alocação da entidade k a um cluster de acordo com a distribuição discreta

$$\Pr(cek = N e + 1 | ce_1, \dots, ce_{k-1}) = \frac{\text{nek}_j + k}{a + k - 1} \quad \text{para } j = 1, \dots, N \text{ e,}$$

onde nek_j denota o número de pontos atualmente dentro do cluster j (na iteração k) e $N e \rightarrow N e + 1$ se $cek = N e + 1$.

Uma vez que temos uma estrutura de clustering das entidades, agora escolhemos valores para os parâmetros de habilidade (específicos do cluster) λs . Estes podem ser escolhidos explicitamente ou alternadamente extraídos da distribuição anterior (base) por amostragem $\lambda s \sim G_0$ para $s = 1, \dots, N$ e. Nós também devemos

escolha se cada classificador é informativo ou não, ou seja, escolha (ou amostra) um valor $\text{dewi} \in \{0, 1\}$ para $i = 1, \dots, n$. Os parâmetros do modelo agora estão totalmente especificados e, portanto, agora podemos usar a representação de variável exponencial do modelo de Plackett-Luce ponderado para gerar classificações. Uma coleção de n classificações completas $\{x_{ij}\}_{j=1}^K$ é gerada através do seguinte processo.

Para $i = 1, \dots, n$

- Amostra $v_{ij} \text{ indep} \sim \text{Exp}(\lambda_{i(j)})$ para $j = 1, \dots, K$.
- Definir $x_{ij} = v_{ij} \text{ onde } S_{ij} = K \setminus \{x_{i1}, \dots, x_{ij-1}\}$ para $j = 1, \dots, K$.
 $\arg\min_{q \in S_{ij}} q$

Observe que o processo descrito acima é equivalente ao processo de geração de dados para o modelo Plackett-Luce ponderado (sem agrupamento) como na Seção 2.5.1, mas aqui o vetor de parâmetro λ é substituído por $\Lambda = (\lambda_{ce1}, \dots, \lambda_{c e K})$. Tipos alternativos de classificações, por exemplo, uma classificação top-5, podem ser obtidos a partir das classificações completas simuladas usando o mesmo processo discutido na Seção 2.2.5.

3.6.3 Especificação de variável prévia e latente

Agora descrevemos nossa distribuição anterior e a probabilidade sob este modelo e apelamos para os indicadores de notação e agrupamento latente introduzidos na seção anterior. Supomos que haja N e clusters de entidades e deixe $\Lambda = (\lambda_1, \dots, \lambda_{Ne})$ denotar a coleção dos parâmetros de habilidade exclusivos. Os indicadores de pertença de agrupamento latente são dados por $ce = (ce_1, ce_2, \dots, ce_K)$ com $ce_k \in \{1, 2, \dots, N\}$ para $k = 1, 2, \dots, K$ onde $ce_k = j$ denota que a entidade k pertence ao cluster j .

Observamos anteriormente que escolher um valor para o parâmetro de concentração do processo de Dirichlet a pode ser difícil. Portanto, quanto a quando consideramos o agrupamento de rankers, tomamos $a \sim Ga(a, b)$ para que possamos inferir esse parâmetro a partir dos dados. Também escolhemos uma distribuição prévia Gamma para os parâmetros de habilidade para que uma atualização conjugada possa ser realizada (após o aumento dos dados). Lembre-se de que, para este modelo, devemos escolher uma distribuição de base única para todos os parâmetros de habilidade, ou seja, G_0 não pode mais depender de k . Aqui tomamos $G_0 = Ga(a, 1)$ que dá λ indep $\sim Ga(a, 1)$ a priori. Observe que não ser capaz de especificar um prior exclusivo para cada uma das k entidades tem um efeito significativo sobre a quantidade de informações que podem ser inseridas na análise por meio da distribuição anterior. De fato, dado que $G_0 = Ga(a, 1)$, a única escolha que temos é colocar uma prévia não informativa nos parâmetros de habilidade; ou seja, que cada ordenação das entidades é igualmente provável. Isso é uma consequência dos parâmetros de habilidade λ_k serem uma amostra aleatória para qualquer escolha de $um \in R^+$ e, portanto, $\Pr(\lambda_i > \lambda_j) = \Pr(\lambda_j > \lambda_i)$ para todo $i \neq j$. Diante disso, podemos pensar que é

suficiente para deixar $a = 1$. No entanto, o valor de a still fornece informações sobre a variância dos parâmetros de habilidade, por exemplo. A flexibilidade limitada da distribuição anterior para os parâmetros de habilidade é uma desvantagem desse modelo. Infelizmente, não há como resolver isso, dada a suposição de permutabilidade do processo de Dirichlet. No entanto, notamos que, em um cenário do mundo real, podemos desejar que os dados sejam a principal força motriz por trás da inferência e, portanto, a incapacidade de colocar informações prévias fortes no modelo talvez não seja muito importante. De fato, vimos em nossas análises anteriores que tomar $a_k = a = 1$ ainda permite inferências informativas sobre os parâmetros de habilidade. Usando $G_0 = Ga(a, 1)$ e lembrando que o modelo contém N e parâmetros de habilidade únicos, a distribuição anterior para Λ (condicional aos indicadores de cluster ce) tem densidade

$$p(L|ce) = \prod_{c=1}^{Nao} \frac{\lambda^{a-1} ce^{-\lambda}}{\lambda C(a)}.$$

Além disso, o modelo assume que cada classificação segue a probabilidade ponderada de Plackett-Luce, ou seja, $X_i | \Lambda, w, ce \sim PLW(\Lambda, w)$. Portanto, também precisamos de uma distribuição prévia para os pesos do classificador latente, w . Aqui escolhemos a especificação anterior usada em análises anteriores, ou seja, $w_i \sim indep \sim Bern(p_i)$ onde $p_i \in (0, 1]$ para $i = 1, \dots, n$. Agora construímos a probabilidade sob este modelo. A probabilidade da i -ésima classificação, expressa em termos das variáveis indicadoras latentescluster, é

$$\Pr(X_i = x_i | \Lambda, ce, w) = \prod_{j=1}^n \frac{\lambda^{w_i c_{ij}}}{\sum_{m \in U_i}^{n_i m=j} \lambda^{w_i c_{im}} + \sum_{m \in U_i}^{n_i m \neq j} \lambda^{w_i c_{im}}},$$

e, portanto, como as classificações são (condicionalmente) independentes, a probabilidade é

$$\pi(D|L, ce, w) = \prod_{i=1}^n \prod_{j=1}^{n_i} \frac{\lambda^{w_i c_{ij}}}{\sum_{m \in U_i}^{n_i m=j} \lambda^{w_i c_{im}} + \sum_{m \in U_i}^{n_i m \neq j} \lambda^{w_i c_{im}}}.$$

Sem surpresa, e como vimos anteriormente, a forma da probabilidade não permite conjugar a inferência bayesiana. Podemos, no entanto, usar técnicas de aumento de dados introduzindo variáveis latentes apropriadas para que as distribuições condicionais completas para os parâmetros de habilidade sejam simples. Aqui, as variáveis latentes necessárias são novamente aquelas que correspondem aos tempos exponenciais latentes entre chegadas e (expressas em termos de nossos indicadores latentes) são

$$\begin{aligned} z_{ij}|D, \Lambda, w, ce &\sim indep \sim Exp \\ \text{Exp} & \quad \sum_{m=j}^{\square} \lambda^{w_i c_{im}} + \sum_{m \in U_i}^{\square} \lambda^{w_i c_{im}}, \end{aligned} \tag{3.11}$$

pois $i = 1, \dots, n$ e $j = 1, \dots, n_i$.

3.6.4 Distribuições condicionais completas

Agora podemos usar o Teorema de Bayes para obter a distribuição posterior. A distribuição posterior $\pi(\Lambda, Z, ce, w, q|D)$ é agora uma distribuição conjunta das variáveis aleatórias latentes Z , a coleção de parâmetros de habilidade únicos Λ , as variáveis indicadoras binárias w , as alocações latentes ce e o parâmetro de concentração DP a . Mais uma vez, podemos amostrar os indicadores de cluster latente usando o Algoritmo 8 de Neal (Neal, 2000) e o FCD do parâmetro DPconcentration a como na Seção 3.4.2. O restante desta seção diz respeito à derivação dos FCDs para as quantidades desconhecidas restantes (Z, Λ, w) e um esboço completo do esquema de amostragem de Gibbs usado para gerar amostras posteriores pode ser encontrado na Seção 3.6.5.

Antes de iniciar a derivação das distribuições condicionais completas, é útil primeiro construir a densidade de todas as grandezas estocásticas. Observe, no entanto, que, dado que já temos os FCDs para ce e a , é sensato considerar apenas a densidade condicional de todas as quantidades estocásticas restantes, dadas as variáveis indicadoras de cluster latente e o parâmetro DPconcentration. Esta densidade de junta (condicional) é

$$p(L, D, Z, w|ce, a) = p(L, D, Z, w|ce) = p(Z|D, L, w,$$

$$ce)\pi(D|L, w, ce)\pi(L|ce)\pi(w)$$

$$\begin{aligned} &= \prod_{i=1}^n \prod_{j=1}^m \frac{\sum_{m=j}^n \lambda_{wic} \exp^{-\sum_{m=j}^n \lambda_{wic}}}{\sum_{m \in U_i} \lambda_{wic}} \times \prod_{i=1}^n \prod_{j=1}^m \frac{\sum_{m=j}^n \lambda_{wic} \exp^{-\sum_{m=j}^n \lambda_{wic}}}{\sum_{m \in U_i} \lambda_{wic}} \\ &\quad \times \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda_{wic} \exp^{-\sum_{m=j}^n \lambda_{wic}}}{\sum_{m \in U_i} \lambda_{wic}} \times \prod_{c=1}^n \frac{\lambda^{a-1} (1-\lambda)^{1-w_i}}{\Gamma(c)} \times \prod_{i=1}^n p_{wi}^{w_i} (1-p_i)^{1-w_i} \\ &= \prod_{i=1}^n \prod_{j=1}^m \frac{\lambda^{a-1} (1-\lambda)^{1-w_i}}{\Gamma(c)} \times \prod_{c=1}^n \frac{\lambda^{a-1} (1-\lambda)^{1-w_i}}{\Gamma(c)} \times \prod_{i=1}^n p_{wi}^{w_i} (1-p_i)^{1-w_i}. \end{aligned} \tag{3.12}$$

A derivação das distribuições condicionais completas (FCDs) para nossos parâmetros segue de maneira semelhante àquela usada para agrupamento de classificadores na Seção 3.5.4. Agora construiremos a distribuição condicional de cada quantidade aleatória dadas todas as outras quantidades estocásticas. Por construção, a distribuição condicional completa para as variáveis latentes Z é como em (3.11) para $i = 1, \dots, n, j = 1, \dots, n_i$. Este resultado também pode ser derivado diretamente de (3.12) como este

Dá

$$p(Z|D, L, w, ce, a) \propto \prod_{i=1}^n \prod_{j=1}^m \frac{\text{Exp} \left(\sum_{m \in U_i} \lambda_{wicexim} + \frac{\lambda_{wic}}{\sum_{m \in U_i} \lambda_{wic}} \right)}{\text{Exp} \left(\sum_{m \in U_i} \lambda_{wicexim} + \frac{\lambda_{wic}}{\sum_{m \in U_i} \lambda_{wic}} \right) + \frac{\lambda_{a-1ce}}{\lambda c C(a)}}.$$

A distribuição condicional completa para os parâmetros de habilidade exclusivos λ_c é derivada da seguinte forma. Temos

$$\begin{aligned} p(L|D, Z, w, ce, a) &\propto \prod_{i=1}^n \prod_{j=1}^m \frac{\text{Exp} \left(\sum_{m \in U_i} \lambda_{wicexim} + \frac{\lambda_{wic}}{\sum_{m \in U_i} \lambda_{wic}} \right) \times \frac{\lambda_{a-1ce}}{\lambda c C(a)}}{\text{Exp} \left(\sum_{m \in U_i} \lambda_{wicexim} + \frac{\lambda_{wic}}{\sum_{m \in U_i} \lambda_{wic}} \right) + \frac{\lambda_{a-1ce}}{\lambda c C(a)}} \\ &\propto \prod_{c=1}^C \prod_{i=1}^n \prod_{j=1}^m \frac{\text{Exp} \left(\sum_{m \in U_i} \lambda_{wicexim} + \frac{\lambda_{wic}}{\sum_{m \in U_i} \lambda_{wic}} \right) \times \frac{\lambda_{a-1ce}}{\lambda c C(a)}}{\text{Exp} \left(\sum_{m \in U_i} \lambda_{wicexim} + \frac{\lambda_{wic}}{\sum_{m \in U_i} \lambda_{wic}} \right) + \frac{\lambda_{a-1ce}}{\lambda c C(a)}} \\ &= \prod_{c=1}^C \frac{\lambda_{a-1ce}}{\lambda c} \frac{\text{Exp} \left(\sum_{i=1}^n \sum_{j=1}^m \Gamma_{ij}(c) z_{ij} \right)}{\sum_{i=1}^n \sum_{j=1}^m \Gamma_{ij}(c) z_{ij}}, \end{aligned}$$

onde

$$\beta_c = \sum_{i=1}^n \sum_{j=1}^m I(cex_{ij} = c),$$

é o número de vezes que uma entidade no cluster c aparece em um ranking informativo

$$\gamma_{ij}(c) = \sum_{m \in U_i} I(cex_{im} = c) + \sum_{m \in U_i} I(cem = c),$$

é o número de vezes que as entidades no cluster c não são classificadas melhor do que j th no i th ranking. Segue-se que a distribuição condicional completa para λ_c está disponível em formas fechadas

$$\lambda_c | \dots, \beta_c + \gamma_{ij}(c), 1 + \sum_{i=1}^n \sum_{j=1}^m \Gamma_{ij}(c) z_{ij} \sim Ga_{\dots}, \quad (3.13)$$

para $c = 1, \dots, N$ e.

As únicas quantidades aleatórias restantes no modelo para as quais não temos atualmente uma distribuição condicional completa são os pesos de classificação latentes w . Lembre-se de que denotamos a coleção de pesos de classificação latentes excluindo aquela associada ao classificador i by $w - i = (w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n)$. A distribuição condicional completa para w_i é a distribuição discreta

$\Pr(w_i = 1|D, \Lambda, Z, w_{-i}, ce, a) \propto \Pr(w_i = 1)\pi(D|w_i = 1, \Lambda, w_{-i}, ce)\pi(Z|w_i = 1, \Lambda, D, w_{-i}, ce)$

$$\propto p_i \prod_{j=1}^{\infty} \frac{\lambda c_{xij}}{\exp \sum_{m=j}^{\infty} \lambda c_{xim}} + \sum_{m \in U_i} \frac{\lambda c_{wm}}{\exp \sum_{m=j}^{\infty} \lambda c_{wm}},$$

$\Pr(w_i = 0|D, \Lambda, Z, w_{-i}, ce) \propto \Pr(w_i = 0)\pi(D|w_i = 0, \Lambda, w_{-i}, ce)\pi(Z|w_i = 0, \Lambda, D, w_{-i}, ce)$

$$\propto (1 - p_i) \prod_{j=1}^{\infty} \exp \{-\text{side}(K_i - j + 1)\}.$$

Portanto, para $i = 1, \dots, n$, a distribuição condicional completa é

$$w_i | \dots \text{indep} \sim \text{Ber}(p_i), \quad (3.14)$$

onde

$$p_i = \frac{\Pr(w_i = 1|D, \Lambda, Z, w_{-i}, ce, a)\Pr(w_i = 1|D, \Lambda, Z, w_{-i}, ce, a) +}{\Pr(w_i = 0|D, \Lambda, Z, w_{-i}, ce, a)},$$

é a probabilidade de que a classificação i seja informativa (dadas as outras quantidades).

Lembre-se de que os indicadores de cluster latente ce podem ser amostrados usando o Algoritmo 8 de Neal (Neal, 2000), que implementa um esquema de urna Pólya para marginalizar os parâmetros dimensionais infinitos. A distribuição condicional completa resultante para as alocações de cluster é uma distribuição discreta sobre os clusters que são ativos e m componentes auxiliares. O parâmetro de concentração DP também pode ser amostrado a partir de sua distribuição condicional completa dada na Seção 3.4.2. A seção a seguir fornece um esboço completo do amostrador de Gibbs usado para obter realizações posteriores.

3.6.5 MCMC usando o Algoritmo de Neal 8

Antes de delinear o esquema de amostragem de Gibbs para gerar realizações a partir da distribuição posterior, $\pi(\Lambda, Z, ce, w, a|D)$, é útil primeiro definir a contribuição para a probabilidade de dados completos do ranker i ser

$$f(x_i, z_i|\Lambda, w, ce, a) = \prod_{j=1}^{\infty} \frac{\exp \left(-\sum_{m=j}^{\infty} \lambda_{wicxim} \right)}{\sum_{m \in U_i} \exp \left(-\sum_{m=j}^{\infty} \lambda_{wicxim} \right)}.$$

Agora podemos descrever o algoritmo de forma concisa. Suponha que tenhamos n classificações de $n_i < K$ entidades. O estado da cadeia de Markov tem elementos $\Lambda = (\lambda_c : c \in \{ce1, \dots, ceK\})$, $Z = (z_{ij})$, $ce = (ce_k)$, $w = (w_i)$ e α para $i = 1, \dots, n$, $j = 1, \dots, n_i$, $k = 1, \dots, K$. O algoritmo faz uma amostragem repetida da seguinte maneira:

- Para $i = 1, \dots, K$: Seja q_- o número de ce_j distinto para $j \neq i$ e $h = q_- + m$. Rotule esses valores ce_j em $\{1, \dots, q_-\}$. Se $ce_i = ce_j$ para algum $j \neq i$, extraia valores independentemente de G_0 para aqueles λ_c para aqueles para os quais $q_- < c \leq h$. Se $ce_i = ce_j \forall j \neq i$, seja ce_i o rótulo $q_- + 1$ e extraia valores independentemente de G_0 para aqueles λ_c para os quais $q_- + 1 < c \leq h$. Desenhe um novo valor para ce_i de $\{1, \dots, h\}$ usando as seguintes probabilidades: $Pr(ce_i = cq|D, Z, \Lambda, w, ce_i, \alpha) = \frac{\lambda_c}{\sum c} b^{n-i} c f(x_i, z_i|\Lambda, w, ce_i = c, ce_{-i}, \alpha)$, $1 \leq c \leq q_-, b$ am $f(x_i, z_i|\Lambda, w, ce_i = c, ce_{-i}, \alpha)$, $q_- < c \leq h$, onde $\Lambda = \{\lambda_1, \dots, \lambda_h\}$, $n-i, c = \#\{ce_j = c : j \neq i\}$ e b é a constante de normalização adequada. Altere o estado para conter apenas aqueles λ_c que agora estão associados a uma ou mais observações, ou seja, seja $\Lambda = (\lambda_c : c \in \{ce1, \dots, ceK\})$.

- Amostra λ_c de (3.13) para $c = 1, \dots, N$ e.
- Amostra z_{ij} de (3.11) para $i = 1, \dots, n$, $j = 1, \dots, n_i$.
- Amostra w_i de (3.14) para $i = 1, \dots, n$.
- Redimensionar– Amostra $\Lambda \dagger \sim Ga(N ea, 1)$ – Calcular $\Sigma = Ne \sum$

- $\sum_{c=1}^{Lc.}$
- Para $c = 1, \dots, N$ e, seja $\lambda_c \rightarrow \lambda_c L \dagger / \Sigma$
- Exemplo a como na Seção 3.4.2 com $n = K$ e $N c = N$ e. A etapa de reescalonamento fornecida aqui é semelhante à mencionada na Seção 2.2.3.

3.6.6 Estudo de simulação – revisitando o conjunto de dados 1

Para nosso primeiro estudo de simulação, revisitamos o Conjunto de Dados 1, apresentado no Capítulo 2. Lembre-se de que este conjunto de dados contém $n = 40$ classificações completas de $K = 20$ entidades, de onde $n_i = K$ para $i = 1, \dots, n$. Observe também que esses dados foram simulados a partir do modelo padrão de Plackett-Luce. Em nossa configuração atual, consideramos, portanto, que esses dados contêm $N = K = 20$

Clusters de entidades com cada entidade dentro de seu próprio cluster. Alternativamente, e equivalente, poderíamos considerar esses dados como sendo do modelo Weighted Plackett-Luce com um processo de Dirichlet anterior nos parâmetros de habilidade usando o processo descrito na Seção 3.5.2 com $c_{ek} = k$ for $k = 1, \dots, K$, $\lambda = (20, 19, \dots, 1)$ e $w_i = 1$ for $i = 1, \dots, n$.

O objetivo de reanalizar esses dados é ver como esse modelo se comporta em um cenário em que a coleção de classificações a serem analisadas contém um grande número de clusters de entidades; especificamente o cenário em que cada entidade está em seu próprio cluster ($N = K$). Será interessante ver se obtemos suporte posterior significativo para 20 clusters de entidades e, se não, como as inferências são afetadas pela introdução de nossa estrutura de agrupamento de entidades. Também aproveitamos esta oportunidade para realizar uma análise de sensibilidade prévia e considerar a sensibilidade da distribuição posterior à escolha da distribuição anterior no parâmetro de concentração α .

Antes de podermos realizar a inferência bayesiana, devemos primeiro descrever uma distribuição anterior adequada. Como em análises anteriores desses dados, optamos por deixar cada ordenação (das entidades) ser igualmente provável a priori; observe que o DP anterior também exige essa distribuição uniforme em ordenações. Como antes, deixamos $a = 1$ e, portanto, a distribuição a priori sobre os parâmetros de habilidade é $\hat{a}_k \text{ indep-} \sim Ga(1, 1)$ para $k = 1, \dots, K$. Conforme discutido ao considerar o agrupamento de classificadores, especificar um valor (anterior) para o parâmetro de concentração do processo de Dirichlet é um pouco difícil. Portanto, em vez disso, permitimos que α seja incerto e atribuímos uma distribuição a priori adequada. Avaliamos como a distribuição a posteriori é afetada pela escolha do α prior, considerando 4 análises separadas. Para as Análises 1 e 2, usamos os priores comumente usados na literatura (por exemplo, Rodriguez et al., 2008), ou seja, $\alpha = \beta = 1$ e $\alpha = \beta = 3$, respectivamente. Nas Análises 3 e 4, tomamos $\beta = 1$ e consideramos $\alpha = 3$ e $\alpha = 5$ para cada análise, respectivamente; A distribuição posterior nessas análises nos permitirá investigar o efeito do aumento da média anterior para α . A Tabela 3.3 (topo) mostra a distribuição anterior (induzida) para o número de clusters de entidades para cada uma das análises consideradas. Observe que, em cada caso, as probabilidades a priori foram obtidas por simulação, pois uma forma fechada de $\pi(N | \alpha, K)$ só existe quando α é uma constante fixa. As médias e desvios-padrão anteriores do número de agrupamentos de entidades, juntamente com o próprio parâmetro de concentração, também são apresentados na Tabela 3.3 (inferior) para cada análise. Finalmente, também precisamos especificar a probabilidade anterior de que cada classificador seja informativo. Embora, em geral, possa ser pragmático usar uma escolha conservadora do π_i , supomos que, pelo menos para esta análise, estamos bastante confiantes de que os classificadores são informativos - esses dados foram simulados sob o modelo padrão de Plackett-Luce (equivalente ao modelo WeightedPlackett-Luce com $w = 1$) - e, portanto, tome $\pi_i = 0,9$ para cada classificador.

j
AA
ba
1234567891011121314
≥ 15
11 0,24 0,21 0,17 0,13 0,10 0,06 0,04 0,03 0,01 0,01 0,00 0,00 0,00 0,00 0,0033 0,11 0,21 0,23 0,19
0,12 0,07 0,05 0,01 0,01 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00
0,00 0,00
31 0,02 0,05 0,09 0,12 0,14 0,15 0,13 0,11 0,08 0,05 0,03 0,02 0,01 0,00 0,0051 0,00 0,01 0,02
0,05 0,08 0,11 0,14 0,15 0,14 0,12 0,08 0,05 0,03 0,01 0,01

Aa	Ba	E(Ce)	SD (Ce)	E(a)	SD(a)
11		3.18	2.07	11	$1/\sqrt{3}$
33		3.45	1.76	1	$\sqrt{3}$
31		6.23	2.53	3	$\sqrt{5}$
51		8.04	2.70	5	

Quadro 3.3: Probabilidades prévias, $\Pr(N_e = j)$, do número de agrupamentos de entidades para cada análise (em cima) e as expectativas a priori e desvios-padrão do número de agrupamentos de entidades e do parâmetro de concentração (em baixo). Os valores modais são destacados em negrito.

Análise posterior

Geramos realizações a partir da distribuição posterior (para cada análise) usando o algoritmo de amostragem de Gibbs descrito na Seção 3.6.5. Depois de realizar algumas execuções piloto, parecia que escolher $m = 3$ deu uma boa mistura sobre os rótulos do cluster. Observe que, para este modelo, o aumento de m não afeta a carga computacional de forma tão significativa quanto quando consideramos os agrupamentos de classificadores, pois aqui só somos obrigados a desenhar m clusters de entidades auxiliares (escalares) e não clusters de classificadores (vetores de parâmetros). Dito isso, como ao considerar o agrupamento de classificadores, a distribuição discreta (condicional completa) sobre os rótulos de agrupamento também aumenta em dimensão. Cada cadeia de Markov foi inicializada em um sorteio aleatório da distribuição anterior. Executamos o esquema MCMC para iterações de 110K, descartando as primeiras amostras de 10K como burn-in e afinando as iterações restantes por um fator de 10. Isso deixou uma amostra posterior de 10K realizações (quase) não autocorrelacionadas da distribuição posterior. O tempo computacional necessário foi (aproximadamente) 215, 196, 223 e 249 segundos para as Análises 1–4, respectivamente. A Figura 3.6 mostra os gráficos de rastreamento da probabilidade de dados completos logarítmicos para todas as análises. As correntes parecem estar se misturando razoavelmente bem em cada caso. Novamente, avaliar a convergência e a mistura dessa maneira é conveniente não apenas porque o espaço de estados é vasto, mas também porque a dimensão da distribuição posterior muda a cada iteração (dependendo do número de clusters de entidades únicas). A convergência foi avaliada ainda mais inicializando várias cadeias em diferentes valores iniciais e verificando se a distribuição posterior obtida de cada cadeia é a mesma até o ruído estocástico.

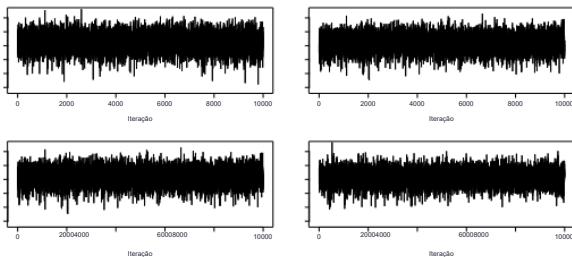


Figura 3.6: Gráficos de rastreamento da probabilidade de dados completos logarítmicos para as Análises 1 e 2 (canto superior esquerdo, direito) e Análises 3 e 4 (canto inferior esquerdo, direito).

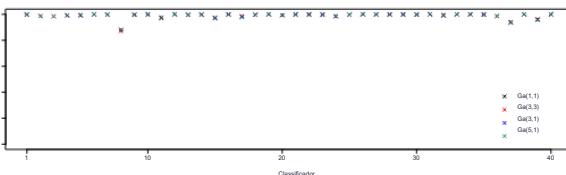


Figura 3.7: $\Pr(w_i = 1 | D)$ – Probabilidade posterior de que a classificação i seja informativa em cada análise (Análise 1: $a\alpha = b\alpha = 1$, Análise 2: $a\alpha = b\alpha = 3$). As cores distinguem entre os diferentes priors em α .

Em cada análise, observamos na Figura 3.7 que a probabilidade posterior de cada classificador ser informativo é grande com $\Pr(w_i = 1 | D) > 0,8$ para $i = 1, \dots, n$. Isso não é surpresa, pois esses dados foram simulados sob o modelo padrão de Plackett-Luce (que tem $w_i = 1$) e expressamos alta confiança em cada classificador ser informativo a priori. Observe que o ranker 8 tem uma probabilidade posterior menor do que a especificada no anterior ($\pi_1 = 0,9$). Uma inspeção mais detalhada dessa classificação revela que ela é um tanto atípica desse conjunto de dados: as entidades 2 e 4 aparecem nas 5 últimas posições e as entidades 13, 11 e 10 aparecem nas 5 primeiras posições. Essas características estão um pouco em desacordo com os verdadeiros valores de parâmetros a partir dos quais esses dados foram simulados e foram observadas quando analisamos esses dados sob o modelo Weighted Plackett-Luce sem agrupamento na Seção 2.7.

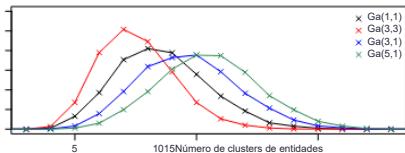


Figura 3.8: $\text{Pr}(N_e = ij|D)$ – Distribuição marginal posterior do número de agrupamentos de entidades para cada análise.

Agora voltamos nossa atenção para o agrupamento de entidades. A Figura 3.8 mostra a distribuição marginal posterior do número de clusters de entidades para cada análise (como polígonos de frequência). A primeira observação marcante que notamos é a diferença bastante grande nas posteriores marginais entre as análises. É claro que esse aspecto da distribuição posterior sempre foi bastante provável de ser afetado, dado que o parâmetro de concentração α controla o nível de agrupamento de entidades. Parece, pelo menos para esses dados, que as crenças anteriores para α desempenham um papel importante na análise. Isso pode ser devido à existência de poucas informações nesses dados sobre o número de clusters de entidades: lembre-se de que esses dados contêm apenas 40 classificações de 20 entidades.

Se compararmos as distribuições marginais posteriores do número de agrupamentos de entidades para as Análises 1 e 2, ou seja, aquelas que especificam uma média unitária no parâmetro de concentração a priori, notamos que há menos variação posterior para este último. Isso sugere que, para esses dados, as informações sobre a variação no número de clusters de entidades contidos no anterior também desempenham um papel importante na análise - lembre-se de que as distribuições anteriores para essas análises especificam desvios padrão de 1 e 1/3 para α respectivamente. Além disso, e talvez não surpreendentemente, observamos que, à medida que a média anterior para α aumenta, o mesmo acontece com a média posterior do número de agrupamentos de entidades; ver Análises 1, 3 e 4 na Figura 3.8.

Tal como acontece com o agrupamento de classificadores, o marginal posterior para o número de agrupamentos não conta a história completa. Novamente, usamos dendrogramas de ligação completa formados a partir da matriz de dissimilaridade com entradas Δ_{ij} , onde $\Delta_{ij} = \text{Pr}(\text{cei } 6 = \text{cej}|D)$ é a probabilidade posterior de que as entidades i e j não estejam agrupadas. A Figura 3.9 mostra os dendrogramas de agrupamento de entidades para cada análise. É claro que a estrutura de agrupamento mostrada nos dendrogramas é semelhante para cada análise e, portanto, esse aspecto da distribuição posterior é bastante robusto para a escolha de α anterior (ao contrário do posterior marginal sobre o número de agrupamentos de entidades). À medida que a média anterior para α aumenta, observamos que os valores de dissimilaridade nos quais os clusters se formam aumentam, ou seja, os rankers são agrupados com menos frequência. Isso é consistente com a observação de um número maior de clusters de entidades. A única mudança notável

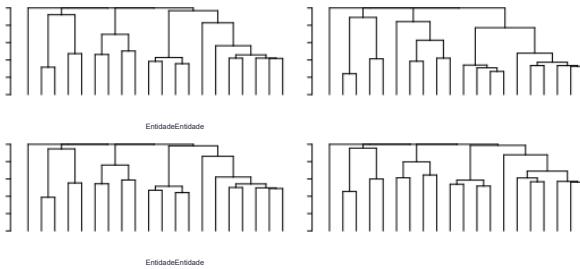


Figura 3.9: Dendogramas de agrupamento de entidades para as Análises 1, 2 (superior esquerdo e direito) e Análises 3,4 (inferior esquerdo e direito).

na estrutura de agrupamento dos dendrogramas é que a entidade 11 mudou de lealdade de entidades $\{4, 6, \dots, 10\}$ para entidades $\{12, 13, 14, 16\}$ na Análise 2 (canto superior direito). É agradável ver que, embora nosso modelo seja incapaz de detectar que cada entidade está dentro de seu próprio cluster (elas têm valores únicos com $\lambda = (20, 19, \dots, 1)$) as entidades que formam grupos são tipicamente aquelas com rótulos mais semelhantes.

É interessante ver como as inferências sobre os valores dos parâmetros de habilidade são afetadas pelo agrupamento de entidades incorporadas no modelo. A Tabela 3.4 mostra as médias marginais posteriores para cada uma das entidades K em todas as análises consideradas. Observe que, para facilitar a comparação, os parâmetros de habilidade foram redimensionados (offline) para que λ_{20} assuma seu valor verdadeiro, ou seja, deixamos $\lambda_k \rightarrow \lambda_k/\lambda_{20}$ para cada realização de nossa distribuição posterior. As classificações agregadas (formadas por entidades ordenadoras por sua média posterior) são as mesmas para cada análise e, portanto, isso é fornecido apenas uma vez na tabela. Segue-se que esse aspecto do posterior parece ser bastante robusto para a escolha do anterior para α . A(s) classificação(ões) agregada(s) (posterior) é(são) bastante semelhante(s) à classificação ótima, $\hat{x} = (1, 2, \dots, 20)$, formadas condicionalmente nos verdadeiros valores dos parâmetros. Lembre-se de que a classificação ótima é aquela que maximiza a probabilidade de Plackett-Luce e é formalmente definida em (2.7). Além disso, se compararmos as inferências poste-rior aqui com aquelas quando assumimos o modelo padrão de Plackett-Luce (sem agrupamento) na Seção 2.4.1, notamos semelhanças impressionantes. As classificações agregadas são as mesmas e há pouca discrepância entre as médias posteriores dos parâmetros de habilidade para cada modelo. Portanto, para esses dados, as inferências posteriores são robustas para incorporar a estrutura de agrupamento de entidades dentro do modelo. Observamos, no entanto, que a distribuição posterior formada sob o modelo que permite o agrupamento de entidades é muito mais rica em

Classificar Xagg	Análise			
	12	34		
13	24.76	23.83	25.39	25.88
21	24.03	23.34	24.57	24.85
35	23.10	22.56	23.44	23.71
42	22.56	22.04	22.75	22.95
54	19.66	19.68	19.61	19.61
66	19.50	19.45	19.39	19.34
78	19.04	19.09	18.91	18.84
87	18.24	18.36	18.04	17.95
99	18.08	18.19	17.85	17.69
1010	16.75	16.84	16.55	16.36
1111	13.02	12.97	12.98	13.04
1212	10.18	10.01	10.24	10.32
1313	9.21	9.10	9.28	9.33
1414	7.66	7.66	7.60	7.59
1516	7.61	7.58	7.56	7.55
1615	5.40	5.41	5.38	5.36
1717	4.95	4.92	4.92	4.91
1818	3.02	3.02	2.99	2.97
1919	2.95	2.96	2.92	2.89
2020	1.00	1.00	1.00	1.00

Tabela 3.4: Médias marginais posteriores dos parâmetros de habilidade para cada análise. A classificação agregada é a mesma em todas as análises.

informação. Por exemplo, podemos quantificar, de maneira baseada em princípios, o nível (posterior) de similaridade entre entidades – algo que exigiria uma abordagem ad hoc sob uma análise padrão (sem agrupamento).

3.7 Resumo

Neste capítulo, mostramos que é possível revelar a estrutura de grupo latente contida nos dados classificados, apelando para modelos de mistura de processos de Dirichlet. Na Seção 3.5, exploramos a área que recebeu muita atenção na literatura, ou seja, revelando diferenças entre as preferências dos classificadores. Utilizou-se uma mistura infinita de modelos ponderados de Plackett-Luce, que, por meio de estudos de simulação, mostrou-se um modelo apropriado para analisar tais dados. Na Seção 3.6, consideramos a noção de que os classificadores podem não ser capazes de distinguir entre certos grupos de entidades, ou seja, eles consideram algumas entidades indistinguíveis (empatadas em força). A análise usou uma nova distribuição prévia (processo de Dirichlet) sobre os próprios parâmetros de habilidade - algo que, na melhor das hipóteses,

nosso conhecimento, não foi considerado anteriormente na literatura. Isso permitiu a exploração da estrutura de agrupamento (potencialmente latente) dentro das entidades. Além disso, a riqueza de informações dentro da distribuição posterior nos permite quantificar o nível de similaridade entre as entidades - isso exigiria uma abordagem ad hoc se usasse técnicas padrão (sem agrupamento). Esquemas eficientes de amostragem posterior (marginal) foram discutidos e uma estratégia de amostragem de Gibbs (possibilitada pelo apelo a técnicas de aumento de dados) foi delineada para cada modelo.

Reconhecemos que os estudos de simulação neste capítulo consideraram dados que foram simulados a partir de modelos homogêneos. No entanto, acreditamos que nossos modelos tiveram um desempenho suficientemente bom e deram inferências razoáveis, mesmo no cenário em que nenhum ranker ou agrupamento de entidades estava presente. Foi particularmente interessante ver que a incorporação de nosso DPprior nos parâmetros de habilidade teve pouco efeito nas inferências posteriores.

No próximo capítulo, exploraremos técnicas de agrupamento bidirecional com o objetivo de construir um modelo único que possa explorar não apenas a heterogeneidade entre os classificadores, mas também a estrutura de agrupamento de entidades dentro dos grupos de classificadores. Como parte disso, consideraremos estudos de simulação sobre dados em que a estrutura de agrupamento (verdadeira) está presente e, portanto, a eficácia de nossos modelos (na recuperação da estrutura do classificador e do grupo de entidades) será examinada na última parte do Capítulo 4.

Capítulo 4

A VARINHA Bayesiana

4.1 Introdução

No Capítulo 3 (Seções 3.5 e 3.6), apresentamos duas distribuições prévias não paramétricas diferentes que permitiram o agrupamento de classificadores ou o agrupamento de entidades. Neste capítulo, desenvolvemos uma distribuição a priori não paramétrica que permite o agrupamento de classificadores e entidades. Fazemos isso apelando para outros priors bayesianos não paramétricos usados para agrupamento bidirecional. Começamos revisando brevemente alguns dos métodos existentes na literatura antes de descrever a distribuição a priori não paramétrica que usaremos para agrupar classificadores e entidades.

4.2 Clustering bidirecional

Existem alguns priors bayesianos não paramétricos que permitem várias camadas de agrupamento por meio de processos de Dirichlet. Dois deles são o Processo Hierárquico de Dirichlet (Teh et al., 2006) e o Processo de Dirichlet Aninhado (Rodriguez et al., 2008) que, para concisão, nos referimos como HDP e NDP, respectivamente.

O HDP tem a especificação de modelo típica

$$\begin{aligned} \lambda_i | \text{Você está} & \quad i = 1, \dots, n, \\ \text{inseparado}, & \\ G_i | \alpha, G_0 \text{ indep} & \sim DP(\alpha, \\ G_0), & i = 1, \dots, n, \\ G_0 | \gamma, H & \sim DP(\gamma, H). \end{aligned}$$

Sob este anterior, cada amostra é extraída de uma distribuição sobre um conjunto comum de átomos (que é uma realização de um processo de Dirichlet) cuja distribuição de base é, por sua vez, outra

Realização do processo de Dirichlet (com distribuição de base associada H). Em nossa configuração, uma realização típica deste anterior é uma matriz $n \times K$ Λ contendo os vetores de parâmetros de habilidade para cada um dos n classificadores. Os vetores de parâmetros de habilidade para cada classificador são extraídos do mesmo conjunto de átomos. No entanto, esses átomos têm pesos diferentes para cada um dos n rankers. Uma maneira de pensar sobre isso é que cada classificador recebe primeiro seu próprio DP exclusivo antes de extrair uma amostra desse DP para seus parâmetros de habilidade K .

O NDP é ligeiramente diferente e sob este NDP anterior há um único processo de Dirichlet cujos átomos são processos de Dirichlet únicos. Tal como acontece com o HDP, uma realização típica (em nosso contexto) deste anterior é a matriz de parâmetros Λ contendo os vetores skillparameter. No entanto, ao contrário do HDP, o NDP estipula que duas realizações do vetor de parâmetro de habilidade, digamos λ_1 e λ_2 , são extraídas de uma distribuição (realização de um DP) sobre os mesmos átomos com os mesmos pesos ou, alternativamente, uma distribuição sobre átomos diferentes com pesos diferentes. Formalmente, o NDP é definido por meio de sua representação de quebra de vara e deixamos $Q|a, \gamma, H \sim NDP(a, \gamma, H)$ denotar que Q segue a distribuição prévia do Processo de Dirichlet aninhado com representação de quebra de bastão

$$Q(L) = \sum_{s=1}^{\infty} \psi s \delta P_s(\lambda),$$

$$\text{vs indep} \sim \prod_{s=1}^{\infty} (1 - v^s), \quad s = 1, 2, \dots,$$

$$\text{vs indep} \sim \text{Beta}(1, \alpha), \quad s = 1, 2, \dots,$$

$$P_s(L) = \sum_{t=1}^{\infty} s t \delta P_t(\lambda), \quad s = 1, 2, \dots,$$

$$wst = ust \prod_{s=1}^{\infty} (1 - n_s), \quad s = 1, 2, \dots, \quad t = 1, 2, \dots,$$

$$ust \text{ indep} \sim \text{Beta}(1, \gamma), \quad s = 1, 2, \dots, \quad t = 1, 2, \dots,$$

$$\lambda^s t \text{ indep} \sim H, \quad s = 1, 2, \dots, \quad t = 1, 2, \dots.$$

Uma especificação de modelo típica para o NDP é, portanto,

$$\lambda_i | D \sim D, \quad i = 1, \dots, n,$$

$$G_i | Q \sim Q, \quad i = 1, \dots, n,$$

$$Q | a, c, H \sim NDP(a, c, H),$$

onde Q segue um processo de Dirichlet aninhado antes com parâmetros de concentração a, γ e distribuição de base H .

Infelizmente, nenhum desses antecedentes é apropriado para o nosso problema. Eles são projetados para situações em que x_{ij} é em si uma observação. No entanto, para dados classificados, esse não é o caso, pois uma observação é a classificação da entidade (vetor) x_i . Em nossa configuração, ambos os anteriores atribuem uma distribuição (realização de um DP) a cada classificador e, em seguida, extraem amostras para os parâmetros $Kskill$ (um para cada entidade) com base apenas nas informações contidas nessa classificação única. No entanto, para agrupar entidades (dentro de cada grupo de classificadores), exigimos informações de vários classificadores. As propriedades do NDP são um tanto desejáveis e, portanto, adaptamos essa distribuição prévia para que possa ser aplicada em um contexto de dados classificados. A adaptação necessária é bastante direta e é discutida em detalhes na próxima seção.

4.3 O Processo de Dirichlet Aninhado Adaptado (ANDP) antes

Como mencionado, precisamos adaptar o Processo de Dirichlet Aninhado antes para que ele possa ser usado em um contexto de dados classificados e permitir o agrupamento de classificadores e entidades. De acordo com o NDP padrão, os classificadores são primeiros atribuídos a uma distribuição (realização de um DP) antes que uma amostra dos parâmetros de habilidade K seja então desenhada (independentemente) para cada classificador. No entanto, adaptamos o anterior para que primeiro atribuirmos todos os classificadores a uma distribuição (realização de um DP) antes de prosseguir com a extração de uma única amostra (dos parâmetros de habilidade K) de cada uma das realizações de DP exclusivas às quais os classificadores são atribuídos. Essas amostras são extraídas com base nas informações de todos os classificadores atribuídos a cada realização de DP respectiva. Além disso, a amostra única (extraída de cada realização de DP respectiva) é compartilhada entre todos os classificadores dentro desse "cluster". Isso resulta em um anterior ligeiramente diferente do NDP, que chamamos de Processo de Dirichlet Aninhado Adaptado (ANDP) anterior, e isso determina que os classificadores atribuídos à mesma realização de DP (cluster) tenham exatamente o mesmo vetor de parâmetro de habilidade λ . Lembre-se de que o NDP requer apenas que os vetores de parâmetros para cada um dos classificadores (atribuídos ao mesmo cluster) sejam extraídos da mesma distribuição (realização de um DP).

A distribuição anterior do Processo de Dirichlet Aninhado Adaptado tem um processo de Dirichlet de "nível superior" cujos átomos são vetores de parâmetros λ . Cada um desses vetores de parâmetros é uma amostra de uma realização única de um processo de Dirichlet de "baixo nível". Deixamos $G|\alpha, \gamma, G_0 \sim \text{ANDP}(\alpha, \gamma, G_0)$

denotam que G segue a distribuição anterior ANDP com representação de quebra de bastão

$$G(L) = \sum_{s=1}^{\infty} \psi s \delta_{\lambda^s}(L^s), \quad (4.1)$$

ψ s indep~ $(1 - v')$, $s = 1, 2, \dots,$
 ψ s $\prod' < s$

$$P(\lambda^s) = \sum_{t=1}^{\infty} s t \delta_{\lambda^s}(t), \quad s = 1, 2, \dots, \quad (4.2)$$

ψ s indep~ Beta(1, α), $s = 1, 2, \dots,$
 ψ s $\prod' < s$

$$wst = ust \prod'_{< s} (1 - \text{nós}'), \quad s = 1, 2, \dots, \quad t = 1, 2, \dots,$$

ψ s indep~ Beta(1, α), $s = 1, 2, \dots, \quad t = 1, 2, \dots,$
 λ^s st indep~ G_0 , $s = 1, 2, \dots, \quad t = 1, 2, \dots.$

Vale a pena notar que os priores do NDP (e, portanto, do ANDP) são geralmente especificados usando dois parâmetros de concentração, um controla o agrupamento de nível superior (em nosso caso) e o segundo corresponde ao agrupamento de nível inferior (o agrupamento de entidades). No entanto, para o ANDP, optamos por introduzir um espaço dimensional infinito para nossos parâmetros de concentração de baixo nível, ou seja, introduzimos ψ s para $s \in N$ e $\lambda^s = (\psi_1, \psi_2, \dots)$ ser a coleção desses parâmetros de concentração. Embora essa alteração possa parecer um pouco incidental, significa que o ANDP anterior agora tem mais flexibilidade para lidar com diferentes níveis de agrupamento de entidades dentro de cada grupo de classificadores. Note-se que, uma vez que os processos de Dirichlet de baixo nível são, eles próprios, átomos do processo de Dirichlet de alto nível, os parâmetros de concentração associados ψ s devem ser intercambiáveis (em relação ao rótulo de agrupamento s). Isso tem como consequência que, se esses parâmetros forem escolhidos para serem constantes fixas, todos devem ser iguais, ou seja, $\psi = y$ com $y > 0$ e $s \in N$. Nesse cenário, a flexibilidade de modelagem adicional é perdida e o parâmetro de concentração (para o agrupamento de entidades) é o mesmo entre os grupos de classificadores. Alternativamente, ψ s pode receber uma distribuição prévia que não depende do rótulo de cluster s . Em outras palavras, podemos escolher ψ s indep~ $f(\cdot)$ a priori, mas não ψ s indep~ $f_s(\cdot)$ devido ao requisito de permutabilidade do processo de Dirichlet de nível superior. Notamos que se a densidade $f(\cdot)$ é uma mistura de distribuições Gama, então, como foi o caso quando consideramos o agrupamento unidirecional (de classificadores ou entidades), a distribuição condicional completa para cada ψ s é direta. É claro que outras especificações prévias podem ser escolhidas, mas vêm ao custo de uma perda de conjugação. Também observamos que o parâmetro de concentração do processo de Dirichlet de nível superior (α) continua sendo um escalar e controla o agrupamento dos classificadores.

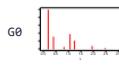
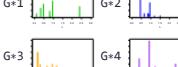
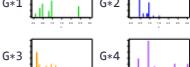
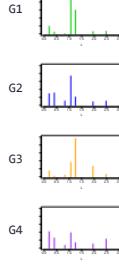
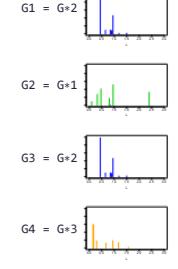
HDP	NDP	ANDP
$G_0 \sim DP(\gamma, H)G_i \sim DP(\alpha, G_0)\lambda_i \sim Gi$	$Q \sim NDP(\alpha, \gamma, H)Gi \sim Q\lambda_i \sim Gi$	$Q \alpha, \gamma, G_0 \sim ANDP(\alpha, \gamma, G_0)(\lambda_1, \dots, \lambda_n) Q \sim Q$
	 G+1 G+2 G+3 G+4 ... $\Delta=2\pi = 8\pi=11, \dots, \lambda=1K$	 G+1 G+2 G+3 G+4 ... $\lambda_1 = \lambda*2$ $\lambda_2 = \lambda*1$ $\lambda_3 = \lambda*2$ $\lambda_4 = \lambda*3$...
 G1 G2 G3 G4 ...	 G1 = G+2 G2 = G+1 G3 = G+2 G4 = G+3 ...	

Figura 4.1: Comparação de distribuições anteriores não paramétricas para agrupamento bidirecional

A Figura 4.1 fornece uma representação gráfica do HDP, NDP e ANDP (com base na Figura 1 em Rodriguez et al. (2008)) e, portanto, esclarece as diferenças entre essas distribuições prévias não paramétricas.

4.3.1 Geração de amostras anteriores (representação de quebra de vara)

Agora descrevemos como obter realizações da distribuição anterior ANDP quando os dados contêm n classificadores e K entidades. Lembre-se de que uma observação típica dessa distribuição anterior é a matriz Λ que contém os vetores de parâmetros para cada um dos n rankers.

Tal realização pode ser obtida de forma bastante trivial usando a representação de quebra de bastão delineada por (4.1) e (4.2). Um método marginal usando um esquema de urna Polya apropriado será considerado mais adiante neste capítulo. Ao implementar a abordagem de quebra de vara para um único processo de Dirichlet, observamos que devemos primeiro escolher um parâmetro de truncamento adequado que resulte em uma aproximação razoável da distribuição dimensional infinita definida pelo processo de Dirichlet. Talvez não seja surpreendente que a ANDP exija dois parâmetros de truncamento, N1 e N2, de modo que as distribuições G e P definidas em (4.1) e (4.2) sejam razoavelmente aproximadas. Esses parâmetros de truncamento devem ser escolhidos de modo que

$$\sum_{s=N1+1}^{\infty} \psi_s \cdot 0 \quad \text{e} \quad \sum_{s=1}^{N1} \sum_{t=N2+1}^{\infty} w_{st} \cdot 0, \quad (4.3)$$

ou, equivalentemente, de modo que

$$\sum_{s=1}^{N1} \psi_s \cdot 1 \quad \text{e} \quad \sum_{t=1}^{N2} w_{st} \cdot 1 \quad \text{para } s = 1, \dots, N1.$$

Podemos então gerar uma realização prévia usando o seguinte processo.

- Escolha N1 e N2 suficientemente grandes.
- Escolha $\alpha > 0$ ou amostra de uma distribuição anterior apropriada.
- Escolha $\gamma > 0$ ou amostra γ_s (independentemente) de uma distribuição prévia apropriada que seja trocável em s por $s = 1, \dots, N1$.
- Amostra λ_{st} indep~ G_0 para $s = 1, \dots, N1, t = 1, \dots, N2$.
- Amostra u_{st} indep~ $\text{Beta}(1, \gamma_s)$ para $s = 1, \dots, N1, t = 1, \dots, N2$.
- Definir $w_{st} = u_{st}(1 - \lambda_{st})$ para $s = 1, \dots, N1, t = 1, \dots, N2$.
Γ'<
- Amostra λ_{sk} da(s) distribuição(ões) discreta(s) com átomos λ_{ts} e pesos w_{st} .
garfo = 1, ..., K e $s = 1, \dots, N1$.
- Amostra v_s indep~ $\text{Beta}(1, \alpha)$ para $s = 1, \dots, N1$.
- Definir $\psi_s = v_s \cdot (1 - v')$ para $s = 1, \dots, N1$.
Γ'<
- Amostra λ_i da distribuição discreta com átomos $\{\lambda_{s1}, \dots, \lambda_{sN1}\}$ e pesos ψ_s para $i = 1, \dots, n$. Agora exploraremos essa distribuição anterior na próxima seção, investigando o efeito dos parâmetros de concentração nas realizações anteriores.

4.3.2 Explorando a ANDP antes

Agora investigamos como os parâmetros de concentração afetam as realizações da distribuição ANDP anterior. Suponha que temos $n = 50$ classificadores e $K = 20$ entidades. Segue-se que o número máximo de valores de parâmetros únicos que podem ser usados para resumir esses dados é $n \times K = 1000$; Esse cenário surgiria quando cada classificador fosse atribuído ao seu próprio cluster e cada entidade (dentro de cada grupo de classificadores) também estivesse dentro de seu próprio cluster. Além disso, o número máximo de vetores de parâmetros exclusivos (clusters de classificadores) é $n = 50$. A Figura 4.2 mostra o número de valores únicos de λ (número total de parâmetros de habilidade únicos em todos os clusters de classificadores) juntamente com o número de clusters de classificadores (vetores de parâmetros únicos) que existem para valores variados de α e γ . Aqui fixamos os parâmetros de concentração em rankergroups para serem constantes e, portanto, definimos $ys = \gamma$ para $s = 1, \dots, N1$. Note-se que a Figura 4.2 mostra as probabilidades empíricas a priori calculadas a partir de um milhão de realizações anteriores (independentes) elaboradas utilizando o método descrito na Secção 4.3.1.

Para α fixos, a distribuição anterior do número de clusters de classificadores (o número de vetores uniqueparameter λ) é a mesma para todos os valores γ ; veja a Figura 4.2 (canto superior esquerdo). Isso segue como o DP (4.1) cujos átomos são vetores de parâmetros de habilidade é condicionalmente independente de γ dado $\lambda * s$ para $s \in N$. A consequência disso é que γ não desempenha nenhum papel no nível de rankerclustering e esse aspecto da distribuição anterior é controlado apenas por α . Como quando consideramos agrupar classificadores (Seção 3.5), o número de agrupamentos de classificadores únicos aumenta à medida que α aumenta; veja a Figura 4.2 (canto superior direito). Segue-se de nossa primeira observação que, para qualquer escolha particular de α , a distribuição anterior do número de agrupamentos de classificadores é a mesma, independentemente de γ . Portanto, para este aspecto da distribuição anterior, permitir que α e γ variem é equivalente a considerar apenas a variação em α . Antes de considerarmos o nível de agrupamento de entidades, lembre-se de que aqui estamos considerando o número total de clusters de entidades em todos os clusters de classificadores, ou seja, o número total de parâmetros de habilidades exclusivas λ , e não o número de clusters de entidades dentro de clusters de classificadores individuais. Como esperado, para α fixos, o número de clusters de entidades aumenta à medida que γ aumenta; veja a Figura 4.2 (canto inferior esquerdo). Observe que aqui temos uma incerteza anterior maior para o número de clusters de entidades em comparação com quando tínhamos clustering apenas em entidades. Isso ocorre porque temos incertezas adicionais sobre o número de clusters de classificação com o ANDP anterior. Além disso, com γ fixo, o número de clusters de entidades únicas aumenta à medida que α aumenta; veja a Figura 4.2 (canto inferior direito). Isso talvez seja contra-intuitivo, pois, para α fixos, a distribuição anterior do número de clusters de entidades (dentro de um grupo de classificação) permanece inalterada. No entanto, à medida que α aumenta, o número de clusters de classificadores aumenta e cada cluster de classificadores exclusivo contém pelo menos um entitycluster. Portanto, à medida que o número de clusters de classificação aumenta (e, portanto, o número de vetores de parâmetro λ), também aumenta o número de valores λ únicos. Isso ocorre quando o vetor de parâmetro de habilidade para cada cluster de classificação diferente é extraído de uma distribuição exclusiva

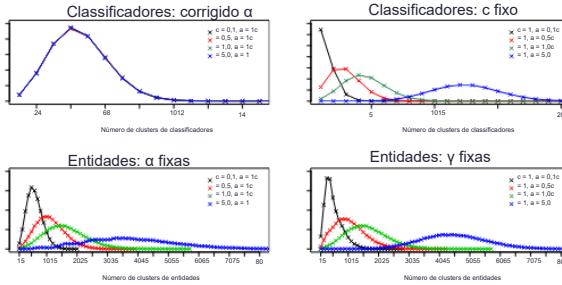


Figura 4.2: Número de agrupamentos de classificadores e entidades sob o Processo de Dirichlet Aninhado
Adaptado antes de vários valores de parâmetros de concentração α e γ .

(realização de um DP) e, portanto, $\Pr(\lambda c_1 i = \lambda c_2 j) = 0$ para dois rótulos de cluster únicos $c_1 \neq c_2$ e todos $i, j \in \{1, \dots, K\}$.

4.4 O modelo

Agora descrevemos o modelo completo que incorpora o parâmetro de confiabilidade do classificador e o agrupamento bidirecional de classificadores e entidades (dentro de grupos de classificadores). O modelo é uma mistura infinita de modelos de Plackett-Luce ponderado com o ANDP escolhido como a distribuição anterior. Referimo-nos a este modelo como a mistura de processo Weighted Adapted Nested Dirichlet (WAND) dos modelos Plackett-Luce. Os principais componentes deste modelo podem ser escritos como

$$\begin{aligned} X_i | \lambda_i, w_i &\sim PLW(\lambda_i, w_i), \quad i = 1, \dots, n, \\ (\lambda_1, \dots, \lambda_n) | Q &\sim Q \\ Q | \alpha, \gamma, G_0 &\sim ANDP(\alpha, \gamma, G_0), \end{aligned}$$

onde a representação de quebra de bastão do ANDP anterior é como na Seção 4.3.

4.5 Uma abordagem de amostragem condicional

Na Seção 3.4.1, discutimos diferentes métodos de implementação de algoritmos de amostragem MCMC para modelos da mistura do Processo de Dirichlet. Ao considerar o agrupamento unidirecional (Capítulo 3), foi bastante simples implementar um esquema de amostragem marginal baseado no Algoritmo 8 de Neal (Neal, 2000). No entanto, essa distribuição anterior não paramétrica é mais complexa devido ao agrupamento bidirecional que induz. Embora existam problemas inerentes à implementação de esquemas de amostragem condicional (como a escolha de parâmetros de truncamento e a exigência de mudanças de rótulo), esses métodos são frequentemente os mais intuitivos e, portanto, consideramos que são um bom ponto de partida neste caso.

4.5.1 Simulando dados do modelo WAND

Nesta seção, descrevemos como simular dados da mistura de processo Weighted Adapted NestedDirichlet (WAND) dos modelos Plackett-Luce. Mais uma vez, é útil introduzir primeiro indicadores de cluster latentes. Seja $c = (c_1, \dots, c_n)$ onde $c_i = j$ denota que a classificação está associada ao vetor de parâmetro λ_j . Também precisamos de indicadores para denotar o agrupamento de entidades dentro de cada grupo de classificação (vetor de parâmetro). Seja $d_{ij} = 1$ denote que o vetor de parâmetro j da entidade i está alocado ao cluster de entidades j e seja $D = (d_{ij})$ denote a matriz de indicadores de cluster de entidades latentes. Nessa notação, o valor do parâmetro de habilidade atribuído à entidade j da classificação i é $\lambda_{ci} d_{ci,j}$. Observe que, no NDP (baunilha), o parâmetro de habilidade correspondente a $\lambda_{ci} d_{ci,j}$ seria $\lambda_{ci, d_{ci,j}}$. Embora útil, a diferença é que, sob o NDP, cada ranker teria sua própria estrutura de agrupamento de entidades, ou seja, os indicadores de cluster (entidade) dentro de i não precisam ser os mesmos que os de j , mesmo que os rankers i e j estejam dentro do mesmo rankercluster ($c_i = c_j$). O ANDP, por outro lado, exige que, se dois classificadores estiverem no mesmo cluster, eles também tenham o agrupamento de entidades equivalente, de onde a estrutura de agrupamento para um classificador no cluster c_i é fornecida por d_{ci} . Lembre-se de que essa restrição é feita para que o agrupamento de entidades possa ser inferido, pois não há informações (sobre o agrupamento de entidades) contidas em uma única classificação i , e, portanto, devemos usar todas as informações desses classificadores dentro do cluster c para obter informações sobre d_{ci} .

Agora descrevemos como gerar dados classificados a partir desse modelo. Primeiro, precisamos especificar a estrutura de clustering do ranker e a estrutura de clustering de entidades dentro de cada cluster de ranker (ativo). Poderíamos, é claro, apenas escolher essas estruturas dando valores ao c and d_{ci} . Observe que, como estamos usando uma abordagem de amostragem condicional, não há mais um requisito para que os clusters ativos (entidade ou classificador) sejam rotulados incrementalmente a partir de 1, pois nosso espaço de estado (sobre indicadores de cluster e a coleção de parâmetros de habilidade exclusivos Λ) permanece de dimensão fixa, ou seja, $N_1 \times N_2$. Os rótulos de cluster são, portanto, necessários apenas

a ser escolhido de modo que $c \in \{1, \dots, N1\}$ e $dck \in \{1, \dots, N2\}$. Alternativamente, poderíamos desenhar estruturas de cluster para os classificadores e entidades da distribuição anterior do ANDP da seguinte forma.

- Escolha $N1$ e $N2$ suficientemente grandes.
- Escolha $\alpha > 0$ ou amostra de uma distribuição anterior apropriada.
- Amostra $v_s \sim \text{indep} \sim \text{Beta}(1, \alpha)$ para $s = 1, \dots, N1 - 1$ e deixe $v_{N1} = 1$.
- Definir $\psi_s = v_s(1 - v')$ para $s = 1, \dots, N1$.
 \prod^{N1-1}_s
- Amostra $c_i \sim \text{indep} \sim \text{Cat}(N1, \psi)$ para $i = 1, \dots, n$.
- Escolha $y_s = y > 0$ ou amostra y_s (independentemente) de uma distribuição prévia apropriada que seja trocável em relação a s por $s \in \{c\}$.
- Amostra $u_t \sim \text{indep} \sim \text{Beta}(1, y_s)$ para $s \in \{c\}, t = 1, \dots, N2 - 1$ e vamos $N2 = 1$.
- Definir $w_{st} = u_t(1 - \text{nós})$ para $s \in \{c\}, t = 1, \dots, N2$.
 \prod^{N2}_t
- Exemplo $dsk \sim \text{indep} \sim \text{Cat}(N2, w_s)$ para $s \in \{c\}, k = 1, \dots, K$. Uma vez que tenhamos uma estrutura de agrupamento de classificadores e entidades, precisamos escolher valores para os parâmetros de habilidade (específicos do cluster) λ_{sk} . Novamente, eles podem ser escolhidos explicitamente ou extraídos da distribuição anterior por amostragem $\lambda_{sk} \sim G_0$ para $s \in \{c\}, k \in \{ds\}$. Finalmente, precisamos especificar se os classificadores são informativos ou não e, portanto, devemos escolher um valor do binário w_i ou amostra-los independentemente das distribuições de Bern (p_i). Agora que todos os parâmetros do modelo são fornecidos, podemos usar a representação da variável latente exponencial do modelo de Plackett-Luce ponderado para gerar classificações. Uma coleção de n classificações completas $\{x_i\}_{i=1}^n$ é gerada através do seguinte processo. Para $i = 1, \dots, n$
- Amostra $v_{ij} \sim \text{indep} \sim \text{Exp}(\lambda w_i c_i d_{ci,j})$ para $j = 1, \dots, K$.
- Definir $x_{ij} = \frac{v_{ij}}{\sum_{q \in S_{ij}} v_{iq}}$ onde $S_{ij} = K \setminus \{x_{i1}, \dots, x_{ij-1}\}$ para $j = 1, \dots, K$.
 $\arg \min_{q \in S_{ij}} v_{iq}$

Tipos alternativos de classificações, como uma classificação top-5, podem ser obtidos a partir das classificações completas simuladas usando o mesmo processo discutido na Seção 2.2.5.

4.5.2Especificação prévia e variáveis latentes

Novamente, escolhemos uma distribuição prévia Gamma para os parâmetros de habilidade para que uma atualização conjugada possa ser executada (após o aumento dos dados). Como na Seção 3.6, nosso modelo permite o agrupamento de entidades e, portanto, precisamos escolher uma única distribuição de base G0 para todos os parâmetros de habilidade. Isso decorre do requisito de permutabilidade do PD anterior sobre os parâmetros da entidade, conforme discutido anteriormente. Aqui tomamos G0 = Ga (a, 1), dando λ_{st} indep~Ga (a, 1) a priori ; Lembre-se de que o parâmetro de taxa não é identificável por verossimilhança e, portanto, é fixado em um.

Observe que, como estamos usando uma abordagem de amostragem condicional, também devemos escolher parâmetros de truncamento para aproximar o aspecto dimensional infinito dos processos de Dirichlet. O espaço de estados dos parâmetros de habilidade, portanto, permanece de dimensão fixa ($N1 \times N2$). Seja $\Lambda = (\lambda_{st})$ a coleção de todos os parâmetros de habilidade únicos ($s = 1, \dots, N1, t = 1, \dots, N2$). A distribuição prévia sobre os parâmetros de habilidade é,

portanto,

$$p(L) = \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\lambda^{a-1} e^{-\lambda_{st}}}{C(a)}.$$

Escolhemos a distribuição a priori para que w_i seja como antes, com w_i indep~ $Bern(pi)$, $pi \in (0, 1)$ para $i = 1, \dots, n$. Lembre-se de que a escolha anterior $pi = 0$ não é permitida. Poderíamos especificar os parâmetros de concentração a e ys como constantes fixas, mas fazer escolhas específicas pode ser difícil. Além disso, o ANDP anterior exige que esses parâmetros de concentração de entidade sejam os mesmos para todos os clusters de classificadores. Em vez disso, optamos por colocar uma distribuição prévia sobre os parâmetros de concentração para agrupamento de classificadores e entidades e deixar $a \sim Ga(aa, ba)$ e ys indep~ $Ga(ay, by)$ para $s = 1, \dots, N1$ a priori. Também apresentamos os indicadores de cluster latente da seção anterior: lembre-se de que $ci = j$ denota que o ranker i está associado ao vetor de parâmetro λ_j e $dij = 'j'$ denota que a entidade j dentro do vetor de parâmetro i está alojada no cluster de entidade j . O valor do parâmetro de habilidade atribuído à entidade j a partir da classificação i é, portanto, dado por $\lambda_{ci, dci, j}$. Usando essas variáveis latentes, a probabilidade é

$$\pi(D|L, c, D, w) = \prod_{i=1}^n \prod_{j=1}^{N2} \frac{(\lambda_{ci, dci, xij})^{w_i}}{\sum_{m=j}^N (\lambda_{ci, dci, xm})^{w_i} + \sum_{m \in U_i} (\lambda_{ci, dci, m})^{w_i}} \quad (4.4)$$

Sem surpresa, dado o que vimos anteriormente, a forma da probabilidade não admite inferência bayesiana conjugada. A implementação de um amostrador de Gibbs para manter a eficiência computacional sem a necessidade de múltiplos parâmetros de ajuste é, no entanto, altamente desejável. Para facilitar isso, apelamos para a mesma técnica dos modelos anteriores, que

é, aumento de dados. Uma solução de amostragem de Gibbs pode ser obtida usando uma generalização direta das variáveis latentes em Caron e Doucet (2012), a saber:

$$\text{zij|D, } \Lambda, c, D, w \text{ indep} \sim \prod_{m=j}^{\hat{m}} (\lambda_{ci}, dci, xim) wi + \sum_{m \in U_i} (\lambda_{ci}, dci, m) w \quad (4.5)$$

pois $i = 1, \dots, n, j = 1, \dots, n_i$.

Usando essas variáveis latentes, a probabilidade completa dos dados é

$$\begin{aligned} \pi(D, Z|L, c, D, w) &= \pi(D|L, c, D, w)\pi(Z|D, L, c, D, w) \\ &= \prod_{i=1}^n \prod_{j=1}^{\hat{m}} \frac{(\lambda_{ci}, dci, xij) wi}{\exp \left(\sum_{m=j}^{\hat{m}} (\lambda_{ci}, dci, xim) wi + \sum_{m \in U_i} (\lambda_{ci}, dci, m) w \right)} \quad \text{eles } \square \\ &= \prod_{s=1}^{N_1} \prod_{t=1}^{N_2} \frac{\lambda_{bst}^s \prod_{i=1}^n \prod_{j=1}^{\hat{m}} (\lambda_{ci}, dci, xim) wi + \sum_{m \in U_i} (\lambda_{ci}, dci, m) w}{\exp \left(\sum_{m=j}^{\hat{m}} (\lambda_{ci}, dci, xim) wi + \sum_{m \in U_i} (\lambda_{ci}, dci, m) w \right)} \quad \text{eles } \square \\ &= \prod_{s=1}^{N_1} \prod_{t=1}^{N_2} \frac{\lambda_{bst}^{s-t} \prod_{i=1}^n W \sum_{j=1}^{\hat{m}} \zeta_{ij}(s, t) z_{ij}}{\exp \left(\sum_{m=j}^{\hat{m}} (\lambda_{ci}, dci, xim) wi + \sum_{m \in U_i} (\lambda_{ci}, dci, m) w \right)}, \quad (4.6) \end{aligned}$$

onde

$$\beta_{st} = \sum_{i=1}^n wi \sum_{j=1}^{\hat{m}} I(c_i = s) I(d_{ci}, x_{ij} = t) \quad (4.7)$$

é o número de vezes que a variável aleatória λ_{st} é atribuída a uma entidade dentro de uma classificação informativa, e

$$\zeta_{ij}(s, t) = I(c_i = s) \sum_{m=j}^{\hat{m}} I(d_{ci}, x_{im} = t) + \sum_{m \in U_i} I(d_{ci}, m = t) \quad (4.8)$$

é o número de vezes que a variável aleatória λ_{st} representa uma entidade dentro de uma classificação informativa e não é classificada acima de j th na i -ésima classificação.

4.5.3 Distribuições condicionais completas

Anteriormente, ao aplicar uma abordagem de amostragem marginal, usamos o Algoritmo 8 de Neal (Neal, 2000) para amostrar os indicadores latentes de suas distribuições condicionais completas. A distribuição condicional completa para o parâmetro de concentração DP também era conhecida; ver secção 3.4.2. Segue-se que poderíamos obter os FCDs para os parâmetros restantes considerando a densidade de todas as quantidades estocásticas (restantes) condicionada aos indicadores latentes e ao parâmetro de concentração. No entanto, ao implementar um

abordagem de amostragem, devemos derivar as distribuições condicionais completas para todas as quantidades desconhecidas (incluindo os indicadores latentes e o parâmetro de concentração). Aqui a distribuição posterior de interesse é $\pi(\Lambda, Z, v, c, u, D, \alpha, \gamma, w|D)$ e no restante desta seção derivamos as distribuições condicionais completas para cada uma das quantidades desconhecidas. Um esboço completo do esquema MCMC usado para gerar realizações posteriores é descrito na Seção 4.5.7.

Começamos construindo a densidade de todas as quantidades estocásticas como

$$\begin{aligned}
 p(L, D, Z, v, c, u, D, a, c, w) &= p(Z|D, L, c, D, w)\pi(D|L, c, D, w)\pi(D|u)\pi(u|\gamma)\pi(c|v)\pi(v|a)\pi(L)\pi(a)\pi(c)\pi(w) \\
 &= \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\lambda \beta stst}{\exp \left(\lambda st \sum_{i=1}^n W \sum_{j=1}^n \zeta_{ij}(s, t) z_{ij} \right)} \\
 &\quad \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{C(1 + ys)C(1)C(ys)}{(ust)ys - 1} \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{C(1 + a)C(1)C(a)}{(vs)a - 1} \times \\
 &\quad \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\lambda a - 1^{\circ} e^{-\lambda st}}{C(a)} \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{C(1 + a)C(1)C(a)}{(vs)a - 1} \times \prod_{i=1}^n \text{você} \\
 &\quad \times \frac{\text{baaa}\Gamma(aa)}{aaa - 1e - q\beta a} \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\Gamma(y\gamma(ay))}{y\gamma - 1s} \times \frac{E - ysby}{E - ysby} \prod_{i=1}^n \text{pwii} (1 - pi)1 - wi \\
 &= \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\lambda a + \beta st - 1st}{C(a) \exp \left(- \lambda st + \sum_{i=1}^n W \sum_{j=1}^n \zeta_{ij}(s, t) z_{ij} \right)} \\
 &\quad \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{ys(1 - ust)ys - 1}{\lambda st} \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\{ \prod_{i=1}^n (1 - n\delta s^*) \}}{\lambda st} \\
 &\quad \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{N1 - \alpha(1 - vs)a - 1 \prod_{i=1}^n \{ \text{contra} (1 - v^*) \}}{\alpha(1 - vs)a - 1} \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\{ \prod_{i=1}^n (1 - v^*) \} ms}{\alpha(1 - vs)a - 1} \times \frac{\text{baaa}\Gamma(aa)}{aaa - 1e - q\beta a} \\
 &\quad \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\Gamma(y\gamma(ay))}{y\gamma - 1s} \times \frac{E - ysby}{E - ysby} \prod_{i=1}^n \text{pwii} (1 - pi)1 - wi \\
 &= \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\lambda a + \beta st - 1st}{C(a) \exp \left(- \lambda st + \sum_{i=1}^n W \sum_{j=1}^n \zeta_{ij}(s, t) z_{ij} \right)} \\
 &\quad \times \prod_{s=1}^{N1-1} \prod_{t=s+1}^{N2} \frac{avms (1 - vs)^{a-1} \sum_{i=s+1}^{N1} m^i \prod_{i=s+1}^{N1} \prod_{t=i+1}^{N2-1} ysumstst (1 - ust)ys - 1 + \sum_{i=s+1}^{N2} ms^i}{\lambda st} \\
 &\quad \times \frac{\text{baaa}\Gamma(aa)}{aaa - 1e - q\beta a} \times \prod_{s=1}^{N1} \prod_{t=1}^{N2} \frac{\Gamma(y\gamma(ay))}{y\gamma - 1s} \times \frac{E - ysby}{E - ysby} \prod_{i=1}^n \text{pwii} (1 - pi)1 - wi, (4.9)
 \end{aligned}$$

onde

$$ms = \sum_{i=1}^n I(c_i = s) \quad e \quad mst = \sum_{j=1}^K I(d_{sj} = t)$$

são o número de classificadores atribuídos ao cluster de classificadores S e o número de entidades atribuídas ao cluster de entidades T dentro dos clusters de classificadores S, respectivamente.

A derivação da distribuição condicional completa (FCD) para cada quantidade aleatória dentro do modelo segue um procedimento semelhante ao observado anteriormente. Cada distribuição condicional completa é obtida tomando as partes apropriadas de (4.9). Os FCDs são os seguintes.

- Λ : Para $s = 1, \dots, N_1, t = 1, \dots,$

$$N_2, \lambda_{st}|D, Z, v, c, u, D, \alpha, \gamma, w \text{ indep} \sim Ga \square \quad \square a + \beta_{st}, 1 + \sum_{i=1}^n W \sum_{j=1}^E \zeta_{ij}(s, t) z_{ij} \square . \quad (4.10)$$

- Z : Conforme definido em (4.5) para $i = 1, \dots, n, j = 1, \dots, n_i$,

$$\begin{aligned} z_{ij} | \dots &\sim \text{Exp} \quad \square \sum_{m=j}^E (\lambda_{ci}, d_{ci}, x_{im}) w_i + \sum_{m \in i} (\lambda_{ci}, d_{ci}, m) w_i \square \\ \text{indep} \sim & \square m=j \quad \square \end{aligned} . \quad (4.11)$$

- em : Para $s = 1, \dots, N_1 - 1$,

$$vs | \dots \text{indep} \sim \left(\frac{1}{1 + ms}, \frac{1}{um + \sum_{i=s+1}^{N_1} m_i} \right) . \quad (4.12)$$

- U : Para $s = 1, \dots, N_1, t = 1, \dots, N_2 - 1$,

$$ust | \dots \text{indep} \sim \left(\frac{1}{1 + mst}, \frac{ys + \sum_{i=t+1}^{N_2} ms_i}{ms} \right) . \quad (4.13)$$

- c : Para $i = 1, \dots, n$, c_i tem uma distribuição discreta com probabilidades dadas por $Pr(c_i = s | \dots) = Pr(c_i = c_i | \dots)$

$$\begin{aligned} &\propto \prod_{s=1}^{N_1} \prod_{i=1}^n \left(1 - \prod_{s' \neq s} \lambda_{ci} \right)^{ms} \times \prod_{s=1}^{N_1} \prod_{t=1}^{N_2} \lambda_{st} \exp^{-\sum_{i=1}^n W \sum_{j=1}^E \zeta_{ij}(s, t) z_{ij}} \square \square \\ &= \prod_{i=1}^n \psi_{ci} \prod_{i=1}^n \prod_{j=1}^E (\lambda_{ci}, d_{ci}, x_{ij}) w_i \\ &\quad \times \prod_{i=1}^n \prod_{j=1}^E \sum_{m=j}^E (\lambda_{ci}, d_{ci}, m) w_i + \sum_{m \in i} (\lambda_{ci}, d_{ci}, m) w_i \square \square \end{aligned}$$

$$\propto \tilde{\psi}^{\sum_{j=1}^J (\lambda_{ci} - d_{ci})x_{ij}} w_i \quad \text{ado} \quad \sum_{m=j}^M (\lambda_{ci} - d_{ci})x_{im} w_i + \sum_{m \in U_i} (\lambda_{ci} - d_{ci,m}) w_i \quad (4.14)$$

onde $1 \leq s \leq N_1$, \tilde{c}_{is} tem $c_j = c_j$ para $j \neq i$ e $\tilde{c}_{is} = s$. Observe que esta é simplesmente a probabilidade completa de dados para o ranker i , dado que eles estão no cluster (ranker) s . Observe também que se $w_i = 0$, então

$$\begin{aligned} \Pr(c_i = s | w_i = 0, \dots) &= \Pr(c_i = c_j | w_i = 0, \dots) \\ &\propto \tilde{\psi}^{\sum_{j=1}^J (\lambda_{s,ds,0} - d_{s,ds})x_{ij}} \quad \text{ado} \quad \sum_{m=j}^M (\lambda_{s,ds,xim} - d_{s,ds,m})x_{im} \quad (4.14) \\ &= \tilde{\psi}^{\sum_{j=1}^J 1 \times X_P \sum_{m=j}^M 1} \quad \text{es} \quad \sum_{m \in U_i} 1 \\ &= \tilde{\psi}^{\sum_{j=1}^J \exp\{-\text{side}(K_i - j + 1)\}} \end{aligned}$$

$\propto \tilde{\psi}^s$ e, portanto, quando um classificador é considerado não informativo, o custo computacional é reduzido, pois não há avaliações de probabilidade necessárias.

- D: Para $s = 1, \dots, N_1, j = 1, \dots, K$, d_{sj} tem a distribuição discreta dada

$$\Pr(d_{sj} = t | \dots) = \Pr(d_{sj} = t | \dots)$$

$$\begin{aligned} &\propto \prod_{s=1}^{N_1} \prod_{t=1}^{N_2} \left\{ \prod_{i=1}^n \frac{(1 - \text{n}\delta_s)^{ms't}}{\lambda \beta s't} \right\} \exp \left(\lambda s't - \sum_{i=1}^n \sum_{j=1}^K \zeta_{ij}(s', t) z_{ij} \right) \\ &= \prod_{s=1}^{N_1} \prod_{j=1}^K \omega s' \tilde{d}_{sj} \prod_{i=1}^n \prod_{j=1}^K \left(\lambda \tilde{c}_{ij} - d_{ci,j} \right) w_i \\ &\quad \times X_P \lesssim \sum_{m=j'}^M \sum_{m \in U_i} (\lambda \tilde{c}_{ij} - d_{ci,m}) w_i \quad (4.15) \\ &\propto \omega s' \times \prod_{j=1}^K \left(\lambda \tilde{c}_{ij} - d_{ci,j} \right) w_i \\ &\quad \times X_P \lesssim \sum_{m=j'}^M \sum_{m \in U_i} (\lambda \tilde{c}_{ij} - d_{ci,m}) w_i \quad (4.15) \end{aligned}$$

onde $1 \leq t \leq N_2$, $F = \{i : ci = s\}$ é dst é dado por $ds' = ds'$ para $j \in F$ e $ds_j = t$.

Observe que para qualquer classificador $q \in F$, se $wq = 0$ então

$$\begin{aligned} & \prod_{j=1}^{\hat{E}} (\lambda_{ci, dci, xij})^{wi} \exp(-e_{les'}) \sum_{m=j'}^{\hat{E}} (\lambda_{ci, dci, xim})^{wi} + \sum_{m \in U_i} (\lambda_{ci, dci, m})^{wi} \\ &= \prod_{j=1}^{\hat{E}} \exp\{-z_{ij}(K_i - j + 1)\} \end{aligned}$$

e, portanto, é constante em relação às mudanças no valor do DSJ. Segue-se que podemos redefinir o conjunto como $F = \{i : (ci = s) \cap (wi = 1)\}$ o que ajudará a reduzir a carga computacional.

- w: Para $i = 1, \dots, n$, wi tem a distribuição discreta dada por $\Pr(wi = 1|w_{-i}, \dots) \propto \Pr(wi = 1)\pi(D|w_{-i}, wi = 1, \dots)\pi(Z|w_{-i}, wi = 1, \dots) \times \prod_{j=1}^{\hat{E}} \exp(-z_{ij}) \sum_{m=j}^{\hat{E}} (\lambda_{ci, dci, xim})^{wi} + \sum_{m \in U_i} (\lambda_{ci, dci, m})^{wi}$

$$\begin{aligned} & \propto \Pr(wi = 1) \prod_{j=1}^{\hat{E}} \lambda_{ci, dci, xij}^{-e_{les'}} \sum_{m=j}^{\hat{E}} \lambda_{ci, dci, xim} + \sum_{m \in U_i} \lambda_{ci, dci, m} \\ & \propto \Pr(wi = 0|w_{-i}, \dots) \\ & \propto \Pr(wi = 0)\pi(D|w_{-i}, ws = 0, \dots)\pi(Z|w_{-i}, wi = 0, \dots) \propto (1 - p_i) \prod_{j=1}^{\hat{E}} \exp(-z_{ij}) \sum_{m=j}^{\hat{E}} (\lambda_{ci, dci, xim})^{wi} \\ & \propto (1 - p_i) \prod_{j=1}^{\hat{E}} \exp\left(-\sum_{m=j}^{\hat{E}} z_{ij}(K_i - j + 1)\right). \end{aligned}$$

Portanto, para $i = 1, \dots, n$, a condicional completa é

$$wi | \dots \text{indep} \sim \text{Berma}(p_i), \quad (4.16)$$

onde

$$p_i = \frac{\Pr(wi = 1|w_{-i}, \dots) \Pr(wi = 1|w_{-i}, \dots)}{\Pr(wi = 0|w_{-i}, \dots)},$$

é a probabilidade de que a classificação i seja informativa (dadas as outras quantidades).

- α : De (4.9) temos

$$\begin{aligned} \pi(\alpha | \dots) &= \frac{\text{baadG}(aa)}{\text{aaa}-1e-\alpha\beta\alpha} \times \prod_{s=1}^{N1-1} \alpha(1-vs)^{\alpha-1} \\ &\propto \frac{\text{aaa}+N1-2e-\alpha\beta\alpha}{\text{aaa}+N1-2e-\alpha\beta\alpha} \prod_{s=1}^{N1-1} \exp \{ (\alpha-1) \log(1-vs) \} \\ &\propto \text{aaa}+n1-2 \exp \left\{ -\alpha \left(\frac{\text{ba}}{\text{ba}-1} \sum_{s=1}^{N1-1} \log(1-vs) \right) \right\}, \end{aligned}$$

de onde

$$a| \dots \sim \frac{\text{Ga}(\alpha + N1 - 1, \text{ba} - 1 \sum_{s=1}^{N1-1} \log(1-vs))}{\text{Ga}(\alpha)} \quad (4.17)$$

como $\text{ba} - \sum_{s=1}^{N1-1} \log(1-vs) > 0$ desde o $vs \in (0, 1)$.

- y : De (4.9) temos

$$\begin{aligned} p(c| \dots) &= \prod_{s=1}^{N1} \frac{\Gamma(y(s)y)}{\Gamma(y(s)-1)} \times \prod_{s=1}^{N1} \frac{y(s)(1-ust)^{y(s)-1}}{\Gamma(N2-y(s))} \\ &\propto \prod_{s=1}^{N1} \frac{y(s)^{y(s)+N2-2} (1-ust)^{y(s)-1}}{\Gamma(y(s))} \prod_{s=1}^{N1} \exp \{ (y(s)-1) \log(1-ust) \} \\ &\propto \prod_{s=1}^{N1} \frac{y(s)^{y(s)+N2-2} \exp \left\{ -y(s) \left(\frac{N2-1}{y(s)-1} \log(1-ust) \right) \right\}}{\Gamma(y(s))}, \end{aligned}$$

de onde para $s = 1, \dots, N1$,

$$y(s)| \dots \sim \frac{\text{indep} \sim \text{Ga}(\alpha y + N2 - 1, b y - 1 \sum_{t=1}^{N2-1} \log(1-para))}{\text{indep} \sim \text{Ga}(\alpha)} \quad (4.18)$$

como $b y - \sum_{t=1}^{N2-1} \log(1-ust) > 0$ desde o $ust \in (0, 1)$ para todos $s = 1, \dots, N1$. Notamos de passagem que, embora talvez contra-intuitivos, os FCDs computacionalmente mais caros para avaliar são as distribuições discretas sobre o ranker e os rótulos de cluster de entidade c e D . Isso ocorre porque vários cálculos de probabilidade são necessários (cada um envolvendo muitos termos) para amostrar cada alocação de cluster, principalmente para rótulos de entidade.

Para gerar realizações a partir da distribuição a posteriori $\pi(\Lambda, Z, v, c, u, D, \alpha, y, w|D)$ poderíamos empregar uma estratégia de amostragem de Gibbs e amostrar repetidamente dos FCDs dados em (4.10-4.18) por sua vez. No entanto, essa estratégia pode levar a cadeia de Markov a ficar presa em um modo local (da distribuição estacionária). Isso ocorre porque as distribuições condicionais completas em (4.14) e (4.15) permitem atualizar a alocação de cluster (classificador e entidade)

ções, c e D, usando atualizações uma de cada vez. Portanto, um grande cluster contendo vários rankers ou entidades terá dificuldade em alterar sua variável de alocação. Inicialmente, isso pode não parecer um problema significativo, pois a probabilidade (4.4) é invariante às permutações das variáveis indicadoras de cluster latentes. No entanto, se considerarmos a distribuição posterior completa $\pi(\Lambda, Z, v, c, u, D, \alpha, \gamma, w|D)$, torna-se claro que o valor da densidade será afetado por mudanças nos rótulos dos clusters por meio da construção dos pesos dos átomos (tanto para classificadores quanto para entidades dentro de cada cluster de classificadores). Os pesos são definidos como decrescentes estocasticamente e, portanto, os pesos dos clusters com rótulos maiores têm menor expectativa. Segue-se que, por exemplo, se tivéssemos dois clusters ocupados, preferiríamos que eles fossem rotulados como 1 e 2 em oposição a N1 – 2 e N1 – 1, pois a primeira rotulagem tem um valor aumentado de densidade posterior.

Felizmente, é possível superar esse problema introduzindo etapas apropriadas de troca de rótulos que melhoram a mistura da cadeia sobre os indicadores de cluster latentes. Descreveremos agora esses movimentos antes de delinear um esquema MCMC completo que pode ser usado para gerar realizações posteriores na Seção 4.5.7.

4.5.4 Movimentos de troca de rótulos

Papaspiliopoulos e Roberts (2008) e, mais recentemente, Hastie et al. (2015) observaram que, ao usar um esquema de amostragem condicional para modelos de mistura de processo de Dirichlet (derivado da construção de quebra de bastão), a cadeia de Markov pode sofrer com a má mistura das alocações de cluster. Inicialmente, isso não parece ser uma grande preocupação, pois a probabilidade (4.4) é invariante às permutações dos rótulos de cluster. No entanto, a distribuição full-posterior $\pi(\Lambda, Z, v, c, u, D, \alpha, \gamma, w|D)$ é afetado por alterações nos rótulos. Em particular, as inferências sobre os parâmetros de concentração α e γ podem ser significativamente afetadas se a cadeia de Markov não se misturar bem sobre os rótulos de cluster; ver Hastie et al. (2015) para obter detalhes. A má combinação de rótulos de cluster é um problema para as alocações de cluster de classificador c e as alocações de cluster de entidade em cada linha de D.

Como os pesos dos átomos dentro de cada processo de Dirichlet estão diminuindo estocasticamente, a alocação de aglomerados com maior suporte posterior é aquela em que o maior aglomerado tem rótulo 1, o segundo maior é rotulado como 2 e assim por diante. No entanto, nosso esquema de amostragem de Gibbs proposto executa atualizações individuais para as alocações de cluster. Isto resulta numa troca de atribuições de agrupamentos muito rara, podendo, por conseguinte, ser excessivamente influenciada pela inicialização (possivelmente aleatória) do amostrador. Para melhorar a mixagem, Papaspiliopoulos e Roberts (2008) propuseram dois movimentos de troca de rótulos, aqui chamados de Swap 1 e 2. Além disso, foi recentemente observado por Hastie et al. (2015) que, embora essas duas trocas encorajem o movimento para a alocação de cluster de maior suporte posterior, uma vez que esse estado é atingido, a mistura se torna pobre. Eles propõem um novo movimento de troca de rótulo, que

chamamos de Swap 3. Todas essas três trocas são aceitas ou rejeitadas pelas propostas de Metropolis-Hastings. Na seção a seguir, descrevemos como esses movimentos de proposta de troca de rótulo são implementados em nossa configuração de modelo. Como nosso modelo contém processos Dirichlet aninhados, prestamos atenção especial às trocas de espaço de estado necessárias dentro da cadeia de Markov quando um movimento de troca de rótulo é aceito.

4.5.5 Alocações de classificadores

Começamos considerando as trocas propostas para as alocações de cluster de classificadores latentes. Devido à natureza aninhada dos processos de Dirichlet, se alterarmos um rótulo de cluster de classificação, também precisaremos alterar os rótulos de todos os parâmetros correspondentes ao DP de baixo nível associado aos clusters de classificação. Agora descrevemos as trocas de espaço de estado necessárias junto com seus mecanismos de proposta.

Troca 1

Lembrando que $ms = \sum ni=1 I(ci = s)$ denota o número de classificadores atribuídos a rankerclusters s , deixamos $C(al) = \{s : ms > 0, 1 \leq s \leq N\}$ ser o conjunto de rótulos de alocação de cluster de ranker para os clusters que são preenchidos, ou seja, os rótulos de clusters aos quais um ou mais rankers são atribuídos. Esta etapa de troca de rótulos propõe trocar os rótulos de dois clusters "vivos" aleatórios $j, l \in C(al)$ e é aceita com probabilidade

$$\text{Min} \left\{ 1, \left(\frac{\psi_l}{\psi_j} \right)^{ml - mj} \right\}.$$

Se a troca for aceita, precisamos fazer uma série de alterações no espaço de estados da cadeia de Markov. Claro, devemos trocar os rótulos do cluster de classificação dentro do vetor de alocação c e também os parâmetros de habilidade associados a cada cluster de classificação, ou seja, trocar linhas apropriadas de Λ . No entanto, para preservar o agrupamento de entidades dentro de cada cluster ranker, também devemos trocar as linhas correspondentes de D , U e w , ou seja, trocar o processo Dirichlet de baixo nível associado a cada cluster ranker. Finalmente, devemos trocar ψ_l por ψ_j , pois o parâmetro de concentração único para cada um dos DPs de baixo nível também deve ser preservado. Os detalhes das alterações necessárias são fornecidos por (4.19)–(4.23) abaixo.

Em primeiro lugar, troque os rótulos no vetor de alocação do classificador c :

$$c'_l = \begin{matrix} \square & \square & \square \\ \square & \square \\ \square & \square & \square \end{matrix} \quad (4.19)$$

$\square \square \square l : ci = j, ji : ci = l,$ caso contrário.

Em segundo lugar, troque as linhas na matriz de parâmetros Λ :

$$\begin{array}{c} \square \square \square \\ \square \square \\ \square \square \end{array} = \begin{array}{c} \square \square \square \\ \square \square \\ \square \square \end{array} \quad (4.20)$$

$\square \square \square \square \lambda l \cdot i = j, \lambda j \cdot i = l, \lambda i \cdot \text{caso contrário.}$

Além disso, para manter a estrutura de clustering de entidades (dentro dos clusters de classificação), as linhas na matriz de alocação de clustering de entidades D são trocadas:

$$\begin{array}{c} \square \square \square \\ \square \square \\ \square \square \end{array} = \begin{array}{c} \square \square \square \\ \square \square \\ \square \square \end{array} \quad (4.21)$$

$\square \square \square \square d l \cdot i = j, d j \cdot i = l, d i \cdot \text{caso contrário.}$

Lembre-se de que também precisamos trocar o processo Dirichlet associado a cada um dos rankerclusters. Isso requer a troca das linhas dentro da matriz u e o recálculo dos pesos ω para cada cluster (classificador). No entanto, como os valores de u (nos quais os valores de ω são baseados) não mudam, é mais conveniente simplesmente renomeá-los também:

$$\begin{array}{c} \square \square \square \\ \square \square \\ \square \square \end{array} = \begin{array}{c} \square \square \square \\ \square \square \\ \square \square \end{array} \quad (4.22)$$

$\square \square \square \square u l \cdot i = j, u j \cdot i = l, u i \cdot \text{caso contrário.}$

Finalmente, também alteramos o valor dos parâmetros de concentração para os DPs correspondentes aos clusters j e l:

$$c'_{il} = \begin{cases} \gamma_l & i = j, \\ \gamma_j & i = l, \\ P & \text{caso contrário.} \end{cases} \quad (4.23)$$

O estado da cadeia de Markov é então atualizado deixando $c \rightarrow c'$, $L \rightarrow L'$, $D \rightarrow D'$, $u \rightarrow u'$, $\omega \rightarrow \omega'$ e $\gamma \rightarrow \gamma'$.

Troca 2

O segundo movimento de troca de rótulos que consideramos também foi derivado por Papaspiliopoulos e Roberts (2008) e propõe trocar os rótulos de dois grupos de classificação vizinhos, independentemente de estarem ocupados ou não. Primeiro, amostramos aleatoriamente um rótulo de cluster ranker j $\in \{1, \dots, N_1 - 1\}$ e então deixe l = j + 1. As variáveis de alocação de cluster j e l

são trocados com probabilidade

$$\text{Min} \left\{ \frac{1, (1 - VL)Mj(1)}{VJml} \right\}.$$

Se aceito, o espaço de estados da cadeia de Markov precisa ser alterado da mesma forma que para o Swap 1, ou seja, pelos swaps em (4.19)–(4.23). Além disso, para esta proposta, se troca por aceita, também trocamos os valores correspondentes das variáveis aleatórias beta associadas aos clusters de classificação:

$$v_l = \begin{cases} v_l & i = j, \\ v_j & i = l, \\ v_0 & \text{caso contrário.} \end{cases}$$

Agora também devemos recalcular os "pesos" para cada cluster de classificação recalculando ψ . Não é possível simplesmente renomear os pesos correspondentes devido à sua construção.

Troca 3

A proposta final de troca de rótulos que implementamos é a de Hastie et al. (2015); Isso também propõe trocar os rótulos de cluster de classificação de dois clusters vizinhos. Seja $C^* = \max_{1 \leq i \leq n} c_i$ o maior rótulo de um cluster ranker que está ocupado. Em seguida, amostramos um rótulo de cluster aleatório $j \in \{1, \dots, C^* - 1\}$ e seja $l = j + 1$.

Antes de indicar a probabilidade de aceitação, é útil definir os seguintes termos:

$$\begin{aligned} E1 &= E(\psi|l|c', \alpha)E(\psi|l|c, \alpha) = \frac{1 + \alpha + ml + \sum_{q>l} m^q + ml + \sum_{q>l} m^q}{\sum_{q>l} m^q}, \\ E2 &= E(\psi|l|c', \alpha) = \frac{\alpha + mj + \sum_{q>l} sqm + 1}{\alpha + mj + \sum_{q>l} sqm}, \\ \alpha E(\psi|l|c, \alpha) &= \alpha \psi_l + \psi_l, \\ \psi^+ &= \psi_l + \psi_l, \\ \hat{\psi} &= \psi_l E1 + \psi_l E2. \end{aligned}$$

Os rótulos de cluster de classificação j e l são então trocados com probabilidade

$$\text{Min} \left\{ \frac{1, (\psi^+ - \psi_l)ml + ml}{\hat{\psi}} \frac{Em|l|_1}{Em|l|_2} \right\}.$$

Novamente, se aceito, as trocas de espaço de estado fornecidas em (4.19) - (4.23) devem ser aplicadas. Além disso, também precisamos atualizar os pesos ψ (para os clusters de classificação) e os valores para o beta

variáveis aleatórias v:

$$\begin{aligned} \psi^i &= \begin{cases} \frac{\psi_i \psi + E1}{\psi} & i = j, \\ \frac{\psi_j \psi + E2}{\psi} & i = l, \\ \psi_i & \text{caso contrário.} \end{cases} & \begin{cases} \frac{\psi^j \prod_{k \neq i} q_k}{(1 - vq)} & i = j, \\ \frac{q^i_j (1 - v^j)}{\prod_{k \neq i} q_k} & i = l, \\ v^i & \text{caso contrário.} \end{cases} \end{aligned}$$

4.5.6 Dotações de entidades

Descobrimos que as trocas de rótulos de cluster propostas descritas na seção anterior são suficientes para garantir a mistura adequada nos rótulos de cluster do classificador. Para este modelo, no entanto, também precisamos garantir a mistura adequada das variáveis de cluster de entidade latente dentro de cada um dos grupos de classificação. Fazemos isso usando os mesmos três mecanismos de troca das alocações do classificador. No entanto, cada uma dessas trocas precisa ser realizada dentro de cada um dos processos de Dirichlet de baixo nível N1. As mudanças necessárias para o espaço de estados de nossa cadeia de Markov são um pouco mais diretas para os rótulos de entidade, pois estamos lidando com um processo típico de Dirichlet (cujos átomos são escalares) e, portanto, os detalhes (dados abaixo) são mais semelhantes aos descritos em Papaspiliopoulos e Roberts (2008) e Hastie et al. (2015).

Troca 1

Lembre-se primeiro de que $mst = \sum_{j=1}^K l(dsj = t)$ denota o número de entidades atribuídas ao entitycluster t dentro do cluster classificador s . Para $s = 1, \dots, N1$, seja $D(al)s = \{t : mst > 0, 1 \leq t \leq N2\}$ seja o conjunto de rótulos de cluster de entidades que são preenchidos dentro do processo de Dirichlet s . Observe que não há nenhum requisito para que um classificador seja atribuído a clusters de classificadores e consideraremos apenas os clusters de entidades aqui. Esta etapa de troca de rótulos propõe trocar os rótulos de dois clusters vivos aleatórios $j, l \in D(al)s$. Este swap é aceito com probabilidade

$$\min_{\text{Min}} \frac{\{1, (\omega_{sj})m_{sl} - m_{sj}\}}{\text{Max}}$$

Se a proposta for aceita, precisamos apenas fazer duas alterações no espaço de estados de nossa cadeia de Markov. Devemos, é claro, trocar os rótulos do cluster de entidades dentro da linha s da matriz de rótulos do entitycluster D , mas também devemos trocar os valores dos parâmetros de habilidade para as entidades j e l .

Formalmente, trocamos os rótulos nas linhas s de nossa matriz de alocação de entidades, D, deixando

$$d's_i = \begin{cases} \square & \square \\ \square & \square \\ \square & \square \end{cases} \quad (4.24)$$

$\square \square \square l_i : d_{si} = j, j_i : d_{si} = l, \text{ caso contrário.}$

Em seguida, trocamos os parâmetros de habilidade para os clusters de entidades dentro da linha s de Λ , permitindo que

$$\lambda' t_{si} = \begin{cases} \square & \square \\ \square & \square \\ \square & \square \end{cases} \quad (4.25)$$

$\square \square \square \lambda t_{si} = j, \lambda t_{sj} = l, \lambda t_{s} \text{ otherwise.}$

O estado da cadeia de Markov é então atualizado deixando $\Lambda \rightarrow \Lambda'$ e $D \rightarrow D'$. Todas as outras quantidades permanecem inalteradas.

Troca 2

Lembre-se de que, para essa troca, são feitas alterações nos rótulos dos clusters vizinhos, estejam eles ocupados ou não. Para $s = 1, \dots, N_1$, amostramos aleatoriamente um cluster de entidades rotulado $j \in \{1, \dots, N_2 - 1\}$ e deixe $l = j + 1$. Os rótulos de cluster de entidades j e l são trocados por probabilidade

$$\text{Min} \left\{ \frac{1, (1 - US_l) MS_j(1)}{US_j MS_l} \right\}.$$

Se aceito, atualizamos o espaço de estados de nossa cadeia de Markov como em (4.24) e (4.25). Lembre-se de que, ao aplicar esse movimento para os rótulos do classificador, também trocamos os valores correspondentes das variáveis aleatórias beta para esses dois clusters. A troca equivalente aqui é feita dentro da linha s da matriz que você e deixamos

$$u's_i = \begin{cases} \square & Usl & i = j, \\ \square & usj & i = l, \\ \square & Us & \text{caso contrário.} \end{cases}$$

Observe que os pesos do cluster de entidades dentro de DP s ($\omega s \cdot$) também devem ser recalculados devido à sua dependência (inerente) dos valores de u. Estes não podem ser simplesmente trocados devido à sua construção.

Troca 3

Finalmente, implementamos o movimento de troca de rótulos descrito por Hastie et al. (2015). Forn = 1, . . . , N1 deixe D*s = max1≤j≤Kdjs ser o maior rótulo de um cluster de entidade ativa dentro de DP s. Nós amostramos um rótulo de cluster aleatório j ∈ {1, . . . , D*s - 1}, defina l = j + 1 e calcule

$$E1 = \frac{E(wsj|d', ys)E(ws_l|d, ys) = 1 + ys + ms_l}{+ \sum_{q>l} ms_q ys + ms_l + \sum_{q>l} ms_q},$$

$$E2 = \frac{E(ws_j|d', ys + ms_j + \sum_{q>l} ms_q + 1)}{ys)E(ws_l|d, ys) = ys + ms_j + \sum_{q>l} ms_q, \\ w_+ = ws_j + ws_l,$$

$$\omega = ws_l E1 + ws_j E2.$$

A probabilidade de aceitação desta proposta é

$$\text{Min} \left\{ 1, \left(\frac{\omega_+ + ms_j + ms_l}{\omega} \right) \frac{E(ws_l|d, ys_+) E(ws_j|d', ys)}{E(ws_j|d, ys) E(ws_l|d, ys_+)} \right\}.$$

Novamente, se aceito, as trocas de espaço de estado fornecidas em (4.24) e (4.25) devem ser aplicadas. Além disso, devemos atualizar os pesos ω e os valores correspondentes para as variáveis aleatórias beta u:

$$\begin{aligned} & \boxed{\omega s_l \omega + E1} \quad i = j, \quad \boxed{\frac{\omega s_j \prod_{q < j} (1 - us_q)}{(1 - us_j)}} \quad i = j, \\ & \boxed{\omega s_i} \quad \boxed{\frac{\omega s_j \omega + E2}{\omega s_l \omega}} \quad i = l, \quad \boxed{\frac{\omega s_l (1 - us_l) \prod_{q < l} (1 - us_q)}{\omega s_j \prod_{q < j} (1 - us_q)}} \quad i = l, \\ & \boxed{\omega s_l} \quad \text{caso contrário.} \quad \boxed{\omega s_j} \quad \text{caso contrário.} \end{aligned}$$

4.5.7 MCMC – um amostrador condicional

Agora descrevemos um algoritmo para gerar amostras a partir da distribuição posterior $\pi(\Lambda, Z, v, c, u, D, \alpha, y, w|D)$. Na Seção 4.5.3, derivamos um conjunto completo de distribuições condicionais completas para cada quantidade aleatória de interesse, de modo que pudéssemos usar um esquema de amostragem de Gibbs por amostragem repetida dessas distribuições. No entanto, conforme discutido, há problemas com a mistura das variáveis indicadoras de cluster latentes c e D . Para remediar esse problema, empregamos movimentos adicionais de troca de rótulos com alterações propostas avaliadas nas etapas de Metropolis-Hastings, conforme descrito na seção anterior. O algoritmo que usamos é, portanto, um sampler Metropolis-within-Gibbs.

Dada uma escolha adequada de parâmetros de truncamento, N1 e N2, as amostras posteriores podem ser geradas executando repetidamente as etapas a seguir.

- Amostra λ_{st} de (4.10) para $s = 1, \dots, N1, t = 1, \dots, N2$.
- Amostra z_{ij} de (4.11) para $i = 1, \dots, n, j = 1, \dots, n_i$.
- Amostra w_i de (4.16) para $i = 1, \dots, n$.
- Amostra v_s de (4.12) para $s = 1, \dots, N1 - 1$, e deixe $v_{N1} = 1$.
- Calcule ψ , ou seja, seja $\psi_s = v_s / (1 - v')$ para $s = 1, \dots, N1$.
 $\prod_{s=1}^{N1} \psi_s < s$
- Amostra c_i de (4.14) para $i = 1, \dots, n$.
- Amostra u_{st} de (4.13) para $s = 1, \dots, N1, t = 1, \dots, N2 - 1$, e vamos $N2 = 1$.
- Calcule w , ou seja, seja $w_{st} = u_{st} / (1 - \psi_s)$ para $s = 1, \dots, N1, t = 1, \dots, N2$.
 $\prod_{t=1}^{N2} w_{st} < t$
- Amostra d_{sj} de (4.15) para $s = 1, \dots, N1, j = 1, \dots, K$.
- Exemplo α de (4.17).
- Amostra y_s de (4.18) para $s = 1, \dots, N1$.
- Propor trocas de rótulos de classificação como na Seção 4.5.5.
- Propor trocas de rótulos de entidade (para $s = 1, \dots, N1$) como na Seção 4.5.6.
- Redimensionar– Amostra $\Lambda \dagger \sim$
 $G_a(N1N2 a, 1)$.– Calcular $\Sigma = N2 \sum$

$$\sum_{s=1}^{N1} \sum_{t=1}^{N2} \Lambda_{st}$$

– Para $s = 1, \dots, N1, t = 1, \dots, N2$, seja $\lambda_{st} \rightarrow \lambda_{st} \Lambda \dagger / \Sigma$.

Uma observação sobre a escolha dos parâmetros de truncamento

Como o algoritmo emprega truncamento (usando N1 e N2), quaisquer amostras posteriores geradas são a partir de uma aproximação do verdadeiro posterior. O nível de aproximação depende da escolha dos parâmetros de truncamento N1 e N2, com as amostras sendo cada vez mais do verdadeiro posterior como $N1, N2 \rightarrow \infty$. A escolha de N1 e N2 a priori é um pouco difícil, uma vez que o nível de truncamento necessário depende fortemente do valor dos parâmetros de concentração, α e γ . A estratégia que defendemos é realizar algumas execuções piloto do

MCMC para medir valores plausíveis dos parâmetros de concentração. Os parâmetros de truncamento precisarão ser alterados para que as condições em (4.3) sejam mantidas. É claro que seria útil tornar os parâmetros de truncamento tão grandes quanto possível, mas o aumento de N1 e N2 tem um efeito importante na carga computacional e pode levar a uma amostragem prévia redundante considerável. A escolha do truncamento é, portanto, bastante específica da situação e talvez limitada pelos recursos computacionais disponíveis para o analista.

4.5.8 Um breve resumo

Até agora, este capítulo delineou a distribuição prévia do processo de Dirichlet Aninhado Adaptado (ANDP) e explorou algumas de suas características por meio de simulação. Em seguida, descrevemos a mistura do Processo de Dirichlet Aninhado Adaptado Ponderado dos modelos de Plackett-Luce (WAND). Isso pegou o ANDP como a distribuição anterior e o usou para misturar os modelos Weighted Plackett-Luce. Usando a representação de quebra de bastão do ANDP anterior e truncando o aspecto dimensional infinito, fomos capazes de derivar um conjunto completo de distribuições condicionais completas. Movimentos adicionais de troca de rótulos foram então introduzidos para melhorar a mixagem em relação aos rótulos de cluster. O algoritmo resultante foi um esquema de amostragem Metropolis-within-Gibbs. A adoção da abordagem de amostragem (condicional) descrita aqui tem, no entanto, algumas desvantagens, a mais notável das quais é que só podemos obter amostras de uma distribuição posterior aproximada. Outra questão é que é difícil determinar quais escolhas a priori dos parâmetros de truncamento levam a uma aproximação razoável do verdadeiro posterior. Idealmente, quanto ao agrupamento unidirecional no Capítulo 3, evitariamós métodos de truncamento e apelariamós para uma abordagem marginal de inferência. Na secção seguinte, discutimos como melhorar o regime MCMC para evitar tais aproximações através da implementação de um esquema de amostragem marginal bidirecional.

4.6 Uma abordagem de amostragem marginal

A implementação de métodos de amostragem marginal para modelos de agrupamento bidirecional que usam processos de Dirichlet pode ser um tanto desafiadora. Por exemplo, a construção do NDP padrão resulta em um esquema "totalmente" marginal sendo computacionalmente inviável (Rodriguez et al., 2008). O verdadeiro ponto crucial ao projetar esse esquema de amostragem marginal é obter realizações (posteriore) dos rótulos de cluster de nível superior, c . Para o processo padrão de Dirichlet aninhado, a amostragem c requer a avaliação de $\pi(x_i|G_s)$ (para todos s), ou seja, a probabilidade de observação x_i dado que está no cluster s . A questão aqui é que G_s é em si um processo de Dirichlet (lembre-se de que, sob o NDP, os átomos do processo de Dirichlet de nível superior são eles próprios processos de Dirichlet). Segue-se que a obtenção de um valor de $\pi(x_i|G_s)$ requer a avaliação de uma soma infinita – algo que é claramente problemático. Diante disso, talvez seja agora

claro que um esquema de amostragem de "truncamento único" poderia ser projetado para obter realizações posteriores sob o NDP anterior. Se os processos de Dirichlet de baixo nível G_s são truncados em um valor finito, digamos N_2 , então a avaliação de $\pi(x_i|G_s)$ é direta. O esquema de amostragem posterior resultante incluiria então um esquema marginal (baseado em um P'olyaurn) para amostrar os indicadores de cluster de nível superior c e uma abordagem de amostragem condicional para aproximar os indicadores posteriores sobre os indicadores de baixo nível D ; ver Rodriguez (2007) para detalhes completos. Tal esquema de amostragem não apenas aumentaria a eficiência computacional, mas também reduziria a aproximação à distribuição posterior em comparação com uma abordagem de "truncamento duplo" semelhante à discutida na Seção 4.5. Além disso, as movimentações adicionais de troca de rótulos não seriam mais necessárias para os rótulos do cluster de classificadores.

Dado que o processo de Dirichlet aninhado adaptado está inherentemente relacionado ao NDP, segue-se que também poderíamos implementar uma abordagem de truncamento único. Verifica-se, no entanto, que uma abordagem totalmente marginal é possível como resultado da adaptação feita à distribuição anterior. Lembre-se de que, para o NDP, a menos que os PDs de baixo nível sejam truncados, uma abordagem marginal para amostrar as variáveis indicadoras c_i é inviável devido à necessidade de avaliar $\pi(x_i|G_s)$. No entanto, para o ANDP, somos obrigados a avaliar $\pi(x_i|\lambda_s)$ onde $\lambda_s \sim G_s$, ou seja, λ_s é uma realização de um processo de Dirichlet e não um DP em si (como é para o NDP). Em outras palavras, a diferença sutil que permite esquemas marginais sob o ANDP é que o processo de Dirichlet de nível superior é bastante padrão, com os átomos sendo vetores de parâmetros que são realizações de DPs (independentes), enquanto, para o NDP, os átomos do DP de nível superior são eles próprios distribuições discretas (de dimensão infinita). A avaliação de $\pi(x_i|\lambda_s)$ é, portanto, trivial – é simplesmente a probabilidade (de Plackett-Luce) de classificação i condicionada aos parâmetros de habilidade para o cluster de classificação s .

A seguir, desenvolvemos um algoritmo de amostragem marginal que pode ser usado para gerar amostras posteriores sob o modelo Bayesiano WAND. Embora o algoritmo de amostragem condicional descrito na Seção 4.5.7 seja capaz de gerar amostras posteriores aproximadas (sujeito à adição de movimentos de troca de rótulo), ainda é vantajoso apelar para uma abordagem marginal – não apenas para reduzir a carga computacional ao realizar inferência, mas também para que possamos gerar realizações a partir da verdadeira distribuição posterior. Conforme discutido na Seção 3.4.1, os algoritmos de amostragem marginal normalmente envolvem o uso de um esquema de urna P'olya que permite a marginalização da distribuição dimensional infinita (Escobar e West (1995), MacEachern e Müller (1998)) e, portanto, evita aproximações. Discutimos na Seção 4.5 como os amostradores condicionais dependem de parâmetros de truncamento para definir o tamanho máximo do espaço de estados (N_1, N_2 em nosso caso). Ao realizar a inferência, todo esse espaço de estado precisa ser atualizado a cada iteração e causa uma quantidade substancial de amostragem prévia redundante. Para produzir nosso amostrador marginal, recorremos novamente ao Algoritmo 8 de Neal (Neal, 2000). Lembre-se de que esse algoritmo só executa atualizações para os componentes exclusivos (clusters) que estão atualmente preenchidos e o re-

Os clusters não povoados restantes (teoricamente infinitos) não são considerados dentro do espaço amostral. Cada observação é então atribuída a um componente que está atualmente em uso ou a um dos m componentes auxiliares que são independentes da distribuição anterior. Nossa modelo tem agrupamento aninhado e, portanto, sob o método condicional, o espaço de estado para os parâmetros de habilidade Λ é a dimensão $N_1 \times N_2$. Se deixarmos N_r e N_s serem o número de clusters de ranker e o número de clusters de entidades dentro do cluster de ranker s respectivamente (para $s = 1, \dots, N_r$), então o tamanho do nosso espaço de estados é reduzido para $\sum N_{rs} = N_s N_1 \times N_2$ sob métodos marginais. É, portanto, evidente que os métodos marginais têm o potencial de reduzir substancialmente o número de operações necessárias por iteração. O Algoritmo 8 de Neal foi projetado para amostrar a partir de uma única mistura de DP e, portanto, projetamos uma versão aninhada que permitirá que a inferência seja realizada sob o modelo Bayesiano WAND.

Lembre-se de que o modelo WAND compreende uma mistura infinita de modelos de Plackett-Luce ponderado com o ANDP escolhido como a distribuição anterior. Tal como na secção 4.4, os principais componentes do modelo são

$$\begin{aligned} X_{ij}|\lambda_i, w_i &\sim PLW(\lambda_i, w_i), \quad i = 1, \dots, n, \\ (\lambda_1, \dots, \lambda_n)|Q &\sim Q \\ Q|\alpha, \gamma, G_0 &\sim ANDP(\alpha, \gamma, G_0), \end{aligned}$$

onde a representação de quebra de bastão do ANDP anterior é como na Seção 4.3.

4.6.1 Amostragem marginal da ANDP anterior

Começamos primeiro descrevendo como podemos (marginalmente) simular uma realização da distribuição ANDP prior por meio de um esquema de urna P'olya. Para delinear esse processo de forma concisa, usamos os indicadores de cluster latentes introduzidos na Seção 4.5.1. Lembre-se de que $c_i = j$ denota que a classificação i está associada ao vetor de parâmetro λ_j e $c = (c_1, \dots, c_n)$ é definido como a coleção desses rótulos de cluster de classificadores latentes. Para nosso agrupamento de entidades, deixamos $d_{ij} = 1$ para denotar que a entidade j dentro do vetor de parâmetro i está alocação ao cluster de entidades' e D ser a coleção de todos os rótulos de cluster de entidades latentes. Observe que, ao contrário da abordagem de amostragem condicional, sob a abordagem de amostragem marginal, os rótulos de cluster devem ser rotulados incrementalmente a partir de 1, ou seja, exigimos $c_i \in \{1, \dots, N_r\}$, (para $i = 1, \dots, n$) em que N_r denota o número de clusters de ranker e também $d_{sj} \in \{1, \dots, N_s\}$, (para $j = 1, \dots, K$) onde N_s denota o número de clusters de entidades dentro de clusters de ranker $s \in \{c\}$. Uma realização prioritária da ANDP é obtida usando o seguinte processo.

- Escolha $\alpha > 0$ ou amostra de uma distribuição anterior apropriada.
- Defina $c_1 = 1$ e o número (atual) de clusters de classificação $N_r = 1$.

- Para $i = 2, \dots, n$ simular a alocação do classificador i para um cluster de classificadores de acordo com

$$\Pr(c_i = j | c_1, \dots, c_{i-1}) = \frac{n_{rj|i} + i - 1}{a + i - 1} \quad \text{para } j = 1, \dots, N_r,$$

onde $n_{rj|i}$ denota o número de classificadores atualmente dentro do cluster de classificadores j (na iteração i) e $N_r \rightarrow N_r + 1$ se $c_i = N_r + 1$.

- Para cada cluster de classificação $s = 1, \dots, N_r$ Escolha $y = y_s > 0$ para todos os s ou amostra de uma distribuição anterior apropriada (observe que a escolha da distribuição anterior deve ser intercambiável em relação a s).— Defina $ds_1 = 1$ e o número atual de clusters de entidades dentro do cluster de classificação $sN_{es} = 1$.— Para $k = 2, \dots, K$ simule a alocação da entidade k a um cluster de entidades de acordo com

$$\Pr(ds_k = j | ds_1, \dots, ds_{k-1}) = \frac{ne_{kj,sys} + y_{s,j} + k - 1}{1},$$

onde $ne_{kj,s}$ denota o número de entidades atualmente dentro do cluster de entidades j (atualização k no cluster de classificação s) e $N_{es} \rightarrow N_{es} + 1$ se $ds_i = N_{es} + 1$.

- Simule λ_{sj} indep~ $\sim G_0$ para $s = 1, \dots, N_r, j = 1, \dots, N_{es}$.

4.6.2 Simulando dados do modelo WAND

Na Seção 4.5.1, descrevemos como os dados podem ser gerados a partir do modelo WAND usando uma abordagem de amostragem condicional. O mecanismo para a abordagem marginal é quase idêntico e difere apenas na forma como geramos uma realização (da estrutura de cluster) a partir do ANDP anterior. Segue-se que, se a estrutura de agrupamento for escolhida explicitamente, ou seja, não for extraída do ANDP anterior, o processo de geração de dados será o mesmo descrito na Seção 4.5.1. No entanto, se quisermos amostrar uma estrutura de agrupamento a partir da distribuição anterior, o uso do método marginal dado na Seção 4.6.1 é vantajoso, pois permite a simulação de realizações exatas a partir da distribuição anterior (em contraste com a aproximação fornecida pela abordagem de amostragem condicional). Condicional em uma amostra anterior (gerada usando o método da Seção 4.6.1), uma coleção de n classificações completas de K entidades é gerada usando o mesmo método da Seção 4.5.1:

Para $i = 1, \dots, n$

- Amostra $v_{ij} \sim \text{indep} \sim \text{Exp}(\lambda w_{ci}, d_{ci,j})$ para $j = 1, \dots, K$.

- Definir $x_{ij} = v_{ij}$ onde $S_{ij} = K \setminus \{x_1, \dots, x_{ij-1}\}$ para $j = 1, \dots, K$.
 $\arg\min_{q \in S_{ij}}$

Tipos alternativos de classificações, por exemplo, uma classificação entre os 5 primeiros, podem ser obtidos a partir das classificações completas simuladas usando o mesmo processo discutido na Seção 2.2.5.

4.6.3 Especificação prévia e variáveis latentes

Antes de realizarmos a inferência bayesiana, devemos primeiro escolher uma especificação prévia adequada para o nosso modelo. O Tribunal considera a mesma especificação prévia que quando consideramos a abordagem de amostragem condicional na seção 4.5.2. Deixamos $G_0 = Ga(a, 1)$ que dá $\lambda_{st} \sim \text{indep} \sim Ga(a, 1)$ a priori (lembre-se de que o parâmetro de taxa não é identificável e, portanto, escolhido como 1). Novamente, definimos Λ como a coleção de todos os parâmetros de habilidade exclusivos e observamos que a dimensão do espaço de parâmetros (sobre Λ) é $\sum N_r = 1 N$ es e, portanto, agora depende do número de clusters de classificadores e entidades. Segue-se que a distribuição a priori sobre Λ é

$$\pi(\Lambda | c, D) = \prod_{s=1}^S \prod_{t=1}^{N_s} \frac{\lambda^{a-1} e^{-\lambda s t}}{C(a)}.$$

Além disso, escolhemos o anterior nos indicadores de capacidade binária latente para ser $w_i \sim \text{indep} \sim \text{Bern}(p_i)$ com $p_i \in (0, 1]$ para $i = 1, \dots, n$. Lembre-se de que a escolha prévia de $p_i = 0$ não é permitida; Neste cenário, o classificador i tem probabilidade constante e, portanto, este classificador não tem informações sobre os parâmetros e, portanto, não deve ser considerado. As distribuições a priori nos parâmetros de concentração são $\alpha \sim Ga(ad, bd)$ e $\gamma \sim \text{indep} \sim Ga(ay, by)$ para $s = 1, \dots, N$.

Obviamente, a probabilidade no modelo Bayesiano WAND é a mesma, independentemente do método de amostragem usado. No entanto, observamos que, ao aumentar a probabilidade, é útil que a distribuição dos parâmetros de habilidade seja da dimensão $\sum N_r = 1 N$ es em vez da dimensão $N_1 \times N_2$ como na abordagem condicional. A probabilidade, condicionada aos indicadores de cluster latentes, é

$$\pi(D|L, c, D, w) = \prod_{i=1}^n \prod_{j=1}^{N_i} \frac{(\lambda_{ci}, d_{ci}, x_{ij})}{\sum_{m=j}^{N_i} (\lambda_{ci}, d_{ci}, x_{im}) w_i} \frac{w_i}{\sum_{m \in U_i} (\lambda_{ci}, d_{ci}, m) w_i} \quad (4.26)$$

Novamente, é útil usar variáveis latentes como em (4.5), pois elas fornecem atualizações semiconjugadas para os parâmetros de habilidade. As variáveis latentes são

$$\text{z}_{ij}|D, \Lambda, c, D, w \text{ indep} \sim \prod_{m=j}^n (\lambda_{ci}, d_{ci}, x_{im}) w_i + \sum_{m \in U_i} (\lambda_{ci}, d_{ci}, m) w_i \quad (4.27)$$

pois $i = 1, \dots, n$, $j = 1, \dots, n_i$.

Antes de construirmos a densidade de todas as quantidades estocásticas (e subsequentemente derivarmos as distribuições condicionais completas para cada variável aleatória dentro do modelo), é útil definir a verossimilhança completa dos dados como

$$\begin{aligned} p(D, Z|L, c, D, w, a, c) &= p(D|L, c, D, w)\pi(Z|D, L, c, D, w) \\ &= \prod_{i=1}^n \prod_{j=1}^{n_i} \frac{(\lambda_{ci}, d_{ci}, x_{ij}) w_i}{\exp \left(- \sum_{m=j}^n (\lambda_{ci}, d_{ci}, x_{im}) w_i + \sum_{m \in U_i} (\lambda_{ci}, d_{ci}, m) w_i \right)} \quad \text{eles } \square \\ &= \prod_{s=1}^n \prod_{t=1}^{tst} \frac{\lambda \beta_s}{\exp \left(- \sum_{j=1}^n (\lambda_{ci}, d_{ci}, x_{im}) w_i + \sum_{m \in U_i} (\lambda_{ci}, d_{ci}, m) w_i \right)} \quad \text{eles } \square \\ &= \prod_{s=1}^n \prod_{t=1}^{tst} \frac{\lambda \beta_s}{\exp \left(- \lambda \sum_{i=1}^n W \sum_{j=1}^{n_i} \zeta_{ij}(s, t) z_{ij} \right)}, \quad (4.28) \end{aligned}$$

onde

$$\beta_{st} = \sum_{i=1}^n w_i I(c_i = s) \sum_{j=1}^{n_i} I(d_{ci}, x_{ij} = t),$$

é o mesmo que na abordagem condicional (4.7) e dá o número de vezes que a variável aleatória λ_{st} é atribuída a uma entidade dentro de uma classificação informativa, e

$$\zeta_{ij}(s, t) = I(c_i = s) \sum_{m=j}^n I(d_{ci}, x_{im} = t) + \sum_{m \in U_i} I(d_{ci}, m = t),$$

é também definido na abordagem condicional em (4.8) e dá o número de vezes que a variável aleatória λ_{st} representa uma entidade que não é classificada acima de j th na i -ésima classificação.

4.6.4 Distribuições condicionais completas

A distribuição posterior é formada pela aplicação do Teorema de Bayes. A distribuição posterior $\pi(\Lambda, Z, c, D, w, a, y|D)$ é agora uma distribuição conjunta das variáveis aleatórias latentes Z , a coleção de parâmetros de habilidade únicos Λ , as variáveis indicadoras binárias w , as variáveis latentes

indicadores de cluster c, D e os parâmetros de concentração DP α, γ . As realizações posteriores dos indicadores latentes podem ser amostradas a partir de suas respectivas distribuições condicionais completas por meio de uma versão aninhada do Algoritmo 8 de Neal (Neal, 2000), que é descrito na próxima seção. Além disso, a distribuição condicional completa para os parâmetros de concentração de DP é semelhante à da Seção 3.4.2 e é dada na Seção 4.6.5 quando fornecemos um esboço completo do esquema MCMC usado para gerar amostras posteriores. Condicional às variáveis indicadoras de cluster latente e aos parâmetros de concentração de DP, a densidade de todas as quantidades estocásticas restantes é

$$\begin{aligned}
 p(L, D, Z, w|c, D, a, \gamma) &= p(L, D, Z, w|c, D) \\
 &= p(Z|D, L, c, D, w)\pi(D|L, c, D, w)\pi(L|c, D)\pi(w) \\
 &= \prod_{s=1}^S \prod_{t=1}^T \frac{\lambda \beta^{st}}{\exp^{-\lambda st}} \sum_{i=1}^n W \sum_{j=1}^{\bar{n}_i} \zeta_{ij}(s, t) z_{ij} \\
 &\quad \times \prod_{s=1}^S \prod_{t=1}^T \frac{\lambda a - \lambda st}{C(a)} \times \prod_{i=1}^n p_{wi} (1 - p_i)^{1-w_i} \\
 &= \prod_{s=1}^S \prod_{t=1}^T \frac{\lambda \beta^{st+a-1}}{C(a) \exp^{-\lambda st}} \sum_{i=1}^n W \sum_{j=1}^{\bar{n}_i} \zeta_{ij}(s, t) z_{ij} \\
 &\quad \times \prod_{i=1}^n p_{wi} (1 - p_i)^{1-w_i}.
 \end{aligned}$$

As distribuições condicionais completas são as seguintes.

- Λ : Para $s = 1, \dots, N$ $r, t = 1, \dots, N$ es

$$\begin{aligned}
 \lambda st|D, \Lambda-st, Z, c, D, w, \alpha, \gamma \text{ indep} \sim Ga & \\
 \lambda st|D, \Lambda-st, Z, c, D, w, \alpha, \gamma \text{ indep} \sim Ga & \\
 \lambda st|D, \Lambda-st, Z, c, D, w, \alpha, \gamma \text{ indep} \sim Ga &
 \end{aligned}$$

- Z : As variáveis latentes são definidas por sua distribuição condicional completa (4.27). Portanto, para $i = 1, \dots, n, j = 1, \dots, \bar{n}_i$,

$$\begin{aligned}
 z_{ij}|D, L, Z-i, c, D, w, \alpha, \gamma \text{ indep} \sim & \\
 \text{Exp} &
 \end{aligned}$$

- w : Tem o mesmo FCD da abordagem condicional. Portanto, para $i = 1, \dots, n, w_i$ segue a distribuição discreta dada por

$$\Pr(w_i = 1 | w_{-i}, \dots) \propto \Pr(w_i = 1) \pi(D|w_{-i}, w_i = 1, \dots) \prod_j \Pr(Z_j | w_{-i}, w_i = 1, \dots)$$

$$\Pr(w_i = 0 | w_{-i}, \dots) \propto \Pr(w_i = 0) \pi(D | w_{-i}, w_s = 0, \dots) \prod_j \pi(Z_j | w_{-i}, w_i = 0, \dots)$$

$$\propto (1 - \pi_i) \exp \left(- \sum_{j=1}^{K_i} z_{ij} (K_i - j + 1) \right).$$

Portanto, a condicional completa é

wi| . . . indep ~ Berna(pi),

para $i = 1, \dots, n$ onde

$$\rho_i = \frac{\Pr(w_i = 1 | w_{-i}, \dots) \Pr(w_i = 1 | w_{-i}, \dots)}{\Pr(w_i = 0 | w_{-i}, \dots)}$$

é a probabilidade de que a classificação i seja informativa (dadas as outras quantidades)

4.6.5 MCMC – um amostrador marginal

Estamos agora em posição de descrever o algoritmo usado para amostragem a partir da distribuição posterior $\pi(\Lambda, Z, C, D, w, \alpha, y|D)$ sob o modelo WAND. Lembre-se de que $N = |\{c_i\}| = 1, \dots, n$ é o número atual de clusters de classificação e N es = $|\{ds_j\}| = 1, \dots, K$ Denota o número de clusters das entidades dentro do cluster de classificação S . O estado da cadeia de Markov consiste então em $oc = (ci), D = (ds), \Lambda = (\lambda st), Z = (zij), w = (wi), y = (ys)$ e c para $s = 1, \dots, N$, $r = 1, \dots, N$ es, $i = 1, \dots, n$, $j = 1, \dots, n$, $n = 1, \dots, K$. Agora, se primeiramente definirmos a contribuição para os dados completos, a probabilidade do classificador i ser

$$f(x_i, z_i | L, c, D, w, \alpha, \gamma) = \prod_{j=1}^J \frac{\lambda w_{i,j} c_{i,j} x_{i,j}}{\exp(\sum_{m \in U_i} \lambda w_{i,m} c_{i,m})} - \sum_{m \in U_i} \lambda w_{i,m} c_{i,m} + \lambda w_{i,J+1} c_{i,J+1}$$

Em seguida, as atualizações procedem da seguinte forma.

- Para $i = 1, \dots, n$: Seja $qr - o$ o número de c_j distinto para $j \neq i$ e $hr = qr - mr$. Rotule esses valores c_j em $\{1, \dots, qr\}$. Se $c_i = c_j$ para algum $j \neq i$, desenhe λc_i indep~ $\sim DP(y_c, G_0)$ para $qr - c < hr$. Se $c_i = c_j \forall j \neq i$, seja c_i o rótulo $qr - 1$, e desenhe λc_i indep~ $\sim DP(y_c, G_0)$ para $qr - 1 < c < hr$.

Desenhe um novo valor para c_i de $\{1, \dots, h_r\}$ usando probabilidades

$$\begin{aligned} \Pr(c_i = c | D, \Lambda, Z, c_{-i}, D, w, \alpha, y) \\ &= \frac{\text{b } nr-i, c f(x_i, z_i | \Lambda, c_{-i}, c_i = c, D, w, \alpha, y), \quad 1 \leq c \leq qr-,}{\text{b } amr f(x_i, z_i | \Lambda, c_{-i}, c_i = c, D, w, \alpha, y), \quad qr- < c \leq hr,} \end{aligned}$$

onde $nr-i, c$ é o número de c_j para $j \neq i$ que são iguais a c , e b é uma constante normalizadora adequada. Altere o estado para conter apenas os λc que agora estão associados a uma ou mais observações, ou seja, seja $\Lambda = (\lambda c : c \in \{c_1, \dots, c_n\})$.

- Rotule novamente c para que $c_i \in \{1, \dots, N_r\}$ para $i = 1, \dots, n$.

- Para $s = 1, \dots, N_r$, $i = 1, \dots, K$: Seja $qe-s$ o número de ds_j distintos para $j \neq i$ and $hes = qe-s + m$. Rotule esses valores ds_j em $\{1, \dots, qe-s\}$. Se $ds_i = ds_j$ para algum $j \neq i$, desenhe λd indep~ G_0 para $qe-s < d \leq hes$. Se $ds_i \neq ds_j \forall j \neq i$, seja ds_i o rótulo $qe-s + 1$, e desenhe λd indep~ G_0 para $qe-s + 1 < d \leq hes$. Desenhe um novo valor para ds_i de $\{1, \dots, hes\}$ usando probabilidades $\Pr(ds_i = d | D, \Lambda, Z, c, D-s_i, w, \alpha, y)$

$$\begin{aligned} &\frac{\text{b } nes, -i, d f(x_i, z_i | L, c, D-s_i, Dsi = d, w, \alpha, y), \quad 1 \leq d \leq qe-s,}{\sum_{\substack{1 \leq j \leq R \\ j \neq i}} b_{j, s, m} f(x_i, z_i | L, c, D-s_i, Dsi = d, w, \alpha, y), \quad Q-s < d \leq ele,} \end{aligned}$$

onde $nes, -i, d$ é o número de ds_j para $j \neq i$ que são iguais a d , $R = \{i : c_i = s\}$ and b é a constante normalizadora adequada. Altere o estado para conter apenas aqueles que agora estão associados a uma ou mais observações, ou seja, seja $\Delta = (\lambda st : s = 1, \dots, N_r, t \in \{ds_1, \dots, ds_K\})$.

- Para $s = 1, \dots, N_r$ renomeie ds de modo que $ds_j \in \{1, \dots, N_{es}\}$ para $j = 1, \dots, K$.

- Para $s = 1, \dots, N_r$, $t = 1, \dots, N_{es}$

$$\begin{aligned} \text{sample} \lambda st | D, \Lambda-st, Z, c, D, w, \alpha, y \text{ indep} \sim & \frac{\square}{\square a + \beta st, 1 + \sum_{i=1}^n W \sum_{j=1}^E \zeta_{ij}(s, t) z_{ij} \square}, \\ Ga & \end{aligned}$$

- Para $i = 1, \dots, n, j = 1, \dots, n_i$

$$\text{sample} z_{ij} | D, L, Z-ij, c, D, w, \alpha, y \text{ indep} \sim \text{Exp} \frac{\square}{\square \sum_{m=j}^n \lambda w_{ici, dci, xim} + \frac{\lambda w_{ici, d}}{c_i, m} \square}.$$

- Para $i = 1, \dots, n$, amostra w_i da distribuição discreta dada por $\Pr(w_i = 1|D, \Lambda, Z, c, D, w-i, \alpha, \gamma) \propto p_i f(x_i, z_i|\Lambda, c, D, w-i, w_i = 1, \alpha, \gamma)$. $\Pr(w_i = 0|D, \Lambda, Z, c, D, w-i, \alpha, \gamma) \propto (1 - p_i) f(x_i, z_i|\Lambda, c, D, w-i, w_i = 0, \alpha, \gamma)$

$$\propto (1 - p_i) \exp \left[- \sum_{j=1}^K z_{ij} (K_j - j + 1) \right]. \quad \square$$

- Redimensionar Amostra $\lambda \dagger$
- $\sim \text{ga}(\alpha, \Sigma)$ (N é , 1).

– Calcule $\Sigma = \frac{\sum_{s=1}^m \sum_{t=1}^{n_s} z_{st} z_{st}^T}{\sum_{s=1}^m n_s}$.

– Para $s = 1, \dots, N$, $r, t = 1, \dots, N$ es , seja $\lambda_{st} \rightarrow \lambda_{st} \lambda \dagger / \Sigma$.

Em função da distribuição prévia descrita no ponto 4.6.3, os parâmetros de concentração podem ser amostrados a partir das seguintes misturas

- Amostra $\pi \sim \text{Ga}(\alpha_0 + N_r, \beta_0 - \log h) + (1 - \pi) \text{Ga}(\alpha_0 + N_r - 1, \beta_0 - \log h)$, onde $\pi(1 - \pi) = \alpha_0 + N_r - 1 \ln(\beta_0 - \log h)$, and $\pi \sim \text{Beta}(\alpha_0 + 1, n)$.

- Para $s = 1, \dots, N$ amostra $y_s | \cdot \cdot \cdot \sim \text{indep} \sim \text{Beta}(\alpha_s + N_{rs}, \beta_s - \log \eta_s) + (1 - \pi_s) \text{Beta}(\alpha_s + N_{rs} - 1, \beta_s - \log \eta_s)$, onde

$$\pi_s(1 - \pi_s) = \alpha_s + N_{rs} - 1 \quad \text{e} \quad \eta_s | \cdot \cdot \cdot \sim \text{indep} \sim \text{Beta}(\gamma_s + 1, K),$$

4.7 Estudos de simulação

Nesta seção, apresentamos dois estudos ilustrativos de simulação baseados em dados simulados que destacam a flexibilidade da WAND Bayesiana para a análise de dados classificados.

4.7.1 Estudo 1

Neste estudo, consideramos dois (novos) conjuntos de dados com classificações completas de $K = 20$ entidades. O primeiro conjunto de dados (Conjunto de Dados 3) contém classificações completas de $n = 40$ classificadores informativos ($w_i = 1$) em um único grupo de classificadores ($N r = 1$ e $c_i = 1$). Os parâmetros distintos de "habilidade" λ_k foram amostrados a partir da distribuição anterior com $\gamma_1 = 1$ e distribuição de bases $G_0 = \text{Ga}(1, 1)$. Esta simulação deu seis clusters de entidades únicas ($N e_1 = 6$). Os clusters de entidades são $\{1\}$ $\{2\}$ $\{3, 4\}$ $\{5, 6\}$ $\{7-16\}$ $\{17-20\}$ onde significa "é preferido para" e os parâmetros de habilidade distintos são $3,07, 1,83, 1,47, 0,85, 0,45, 0,04$ para cada cluster, respectivamente. Os dados simulados são apresentados no quadro B.3 nos apêndices. Observe que, para facilitar a interpretação, aplicamos uma permutação aos rótulos de entidade para que o λ_k diminuisse, com a entidade mais preferida sendo rotulada como 1 e a menos preferida rotulada como 20.

O segundo conjunto de dados (Conjunto de dados 4) contém classificações completas de $n = 50$ classificadores, consistindo naqueles no conjunto de dados 3 e um conjunto adicional de 10 classificadores não informativos, com $w_{41:50} = 0$. Essas classificações adicionais (aleatórias) são fornecidas nos apêndices da Tabela B.4.

Investigamos o efeito de classificações incompletas comparando a análise das classificações completas com as das classificações top-M para $M = 5, 10, 15$. Ou seja, analisamos os classificadores de dados assumindo que vemos $K_i = K = 20$ entidades e relatamos seu top $n_i = 5, 10, 15, 20$ entidades. Essas análises nos permitirão investigar o nível de incerteza introduzido apenas pela observação de truncamentos dos rankings. Também será interessante explorar se uma coleção de classificações completas é necessária se quisermos apenas inferir, digamos, as 5 principais entidades. Outro cenário que consideramos é o de uma análise chamada "restrita", na qual qualquer entidade não classificada por nenhum classificador é removida do conjunto de dados. A intuição por trás desse cenário é que nossas crenças sobre as ordenações de entidades que aparecem em pelo menos uma classificação não devem ser afetadas pelo fato de entidades não aparecidas serem incluídas ou não. Um exemplo disso está no Conjunto de Dados 3, onde, nos cenários 5 e 10 principais, as entidades 17, 18 e 20 não aparecem em nenhum dos rankings. Assim, também consideramos as análises restritas top-5 e top-10 do Conjunto de Dados 3 que usam $K_i = K = 17$.

Adotaremos a mesma distribuição de bases das análises anteriores e tomaremos $G_0 = \text{Ga}(1, 1)$. Além disso, escolhemos as distribuições anteriores para os parâmetros de concentração a serem $\alpha \sim$

$Ga(1, 1)$ e $ys \sim Ga(3, 3)$ para $s \in N$, observando que essas são escolhas comuns na literatura (Rodriguez et al., 2008). Aqui, consideramos dois cenários de um priori intercambiável para os parâmetros de confiabilidade do ranker w : um em que não temos certeza sobre sua capacidade (tomando $\pi = 0,5$) e outro em que estamos bastante confiantes de que eles são informativos (tomando $\pi = 0,9$).

Para gerar realizações a partir da distribuição posterior (para cada análise), implementamos o algoritmo de amostragem (marginal) descrito na Seção 4.6.5 com $mr = 2$ e $me = 3$ variáveis auxiliares (ranker e entidade). Cada cadeia de Markov foi inicializada em um sorteio aleatório da distribuição anterior. Para obter 10K realizações (quase) não autocorrelacionadas da distribuição posterior, permitimos a cada cadeia um período de burn-in de 10K iterações, em seguida, executamos o esquema por mais 1M iterações e reduzimos a saída por um fator de 100 em cada caso. As análises mais caras computacionalmente foram as do Conjunto de Dados 4 com $\pi = 0,9$ e o tempo computacional necessário para realizar a inferência foi (aproximadamente) 7, 16, 21 e 26 minutos para os casos top-5, top-10, top-15 e completos, respectivamente. A mistura das cadeias MCMC foi avaliada inspecionando gráficos de rastreamento e a convergência foi avaliada inicializando várias cadeias em diferentes valores iniciais e verificando se as distribuições posteriores resultantes eram equivalentes (até ruído estocástico).

A Figura 4.3 mostra a probabilidade posterior $Pr(w_i = 1|D)$ que o ranker i é informativo para cada cenário de informação e para ambas as opções de probabilidades anteriores $\pi = Pr(w_i = 1)$. Não é surpreendente que as análises restritas, devido à perda de informação, produzam resultados que são menos consistentes com as habilidades "conhecidas" dos classificadores ($w1:40 = 1$ e $w41:50 = 0$) do que as análises completas (irrestritas) que consideram todas as entidades. Este achado é claro para ambas as escolhas do π , mas é mais perceptível para o caso $\pi = 0,5$. Os resultados mostram corretamente que a maioria dos classificadores foi identificada como informativa (linha intermediária). Sem surpresa, essa identificação se torna mais clara à medida que classificações mais abrangentes são usadas (indo do top-5 até o completo). Também é interessante ver que a análise mais pobre em dados (top 5) se sai razoavelmente bem.

As análises para o conjunto de dados 4 mostram que os classificadores não informativos (aleatórios) foram identificados com bastante clareza, particularmente para o caso $\pi = 0,5$, onde a probabilidade posterior de que os classificadores 41 a 50 sejam informativos é muito próxima de zero (gráfico inferior esquerdo). As probabilidades para os classificadores de 1 a 40 permanecem semelhantes às encontradas ao analisar o Conjunto de Dados 3. A linha inferior de gráficos mostra a influência da escolha do π na distribuição anterior. Aqui vemos que ter um alto nível de confiança de que os classificadores não informativos são informativos pode potencialmente mascarar sua identificação. Fica claro aqui que as probabilidades posteriores $Pr(w_i = 1|D)$ para os classificadores não informativos estão bem separados daqueles para

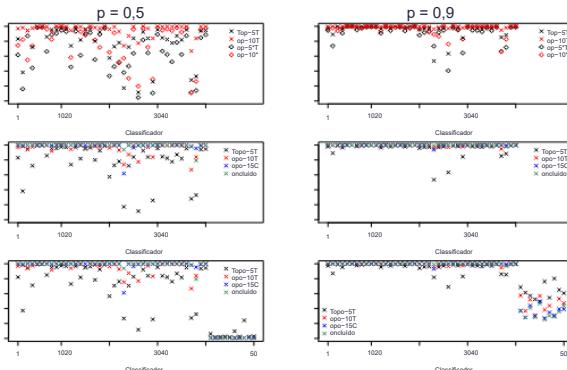


Figura 4.3: Gráficos da probabilidade posterior $\Pr(w_i = 1|D)$ que o ranker i é informativo para ambos os cenários de anterior em sua habilidade: $p_i = 0,5$ (coluna da esquerda) e $p_i = 0,9$ (coluna da direita). A linha superior de gráficos mostra a comparação entre as análises restritas (*) e completas (sem restrições) para o conjunto de dados 3. Os gráficos na linha do meio são aqueles para as análises completas usando o Conjunto de Dados 3, com os gráficos correspondentes usando o Conjunto de Dados 4 na linha inferior.

classificadores informativos, mas, no entanto, eles podem ser identificados como informativos se essa escolha for feita limitando essas probabilidades posteriores em, digamos, 0,5 ou até mais. Isso sugere que o analista deve usar uma escolha bastante conservadora do p_i e deve ter cuidado ao expressar confiança excessiva nas habilidades do ranker a priori. A Tabela 4.1 fornece a distribuição posterior do número de clusters de classificação N em cada análise. Para o conjunto de dados 3, vemos muito mais suporte posterior para um único grupo de classificadores nas análises completas (irrestritas) em comparação com seus equivalentes restritos, com pouca dependência da escolha do p_i . Além disso, para as análises do Conjunto de Dados 4, o suporte posterior para um único grupo de classificadores diminui, particularmente para o caso $p_i = 0,9$. De fato, para este caso, a alta confiança prévia de que todos os classificadores são informativos altera o número modal de agrupamentos de classificadores de um para dois, embora isso venha com um nível mais alto de incerteza posterior. Além disso, como antes, a probabilidade do número correto de rankerclusters aumenta à medida que classificações mais abrangentes são incluídas na análise.

No Capítulo 3, observamos que determinar a alocação de classificadores para grupos de classificadores pode ser problemático. Por exemplo, condicionar o número modal de clusters de ranker e alocar o ranker i para o cluster de ranker k usando apenas uma estimativa MCMC de $\Pr(c_i = k|N, r, D)$ pode

p = 0,5				p = 0,9						
Conjunto de	12	34	≥ 5	Conjunto de	12	34	≥ 5			
Topo-5*	0.73	0.20	0.05	0.02	0.00	0.65	0.23	0.08	0.02	0.01
Topo-10*	0.88	0.11	0.01	0.00	0.00	0.88	0.11	0.01	0.00	0.00
Topo-5 10 melhores	0.87	0.11	0.02	0.00	0.00	0.87	0.11	0.02	0.00	0.00
Topo-15	0.98	0.02	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.00
Completo	0.99	0.01	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Completo	1,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00
Conjunto de	12	34	≥ 5	Conjunto de	12	34	≥ 5			
Topo-5	0.74	0.20	0.05	0.01	0.00	0.06	0.29	0.29	0.19	0.10
10 melhores	0.88	0.10	0.02	0.00	0.00	0.09	0.38	0.30	0.16	0.06
Topo-15	0.89	0.10	0.01	0.00	0.00	0.15	0.36	0.28	0.14	0.05
Completo	0,89	0,10	0,01	0,00	0,00	0,16	0,35	0,28	0,14	0,06

Tabela 4.1: Distribuição posterior do número de clusters de classificação N r para análises restritas (*) e completas (irrestritas). Os números em negrito indicam valores modais.

não fornecer uma descrição adequada da distribuição posterior conjunta das alocações c, particularmente se não houver suporte posterior esmagador para a escolha de N r. Em vez disso, preferimos não condicionar um N r específico e usar a saída MCMC completa para examinar as probabilidades de co-agrupamento $\Pr(c_i = c_j | D)$. Definimos uma matriz de dissimilaridade $\Delta = [\Delta_{ij}]$, onde $\Delta_{ij} = \Pr(c_i \neq c_j | D)$ mede o quanto diferentes são os classificadores i e j e, em seguida, considera o dendrograma correspondente ao formar a alocação de classificadores para agrupamentos de classificadores, ao mesmo tempo em que considera a distribuição posterior de N r. Usamos o método de encadeamento completo, também conhecido como agrupamento de vizinhos mais distantes, pois isso tende a produzir aglomerados mais densamente compactados e não sofre de "encadeamento".

A distribuição posterior para o número de clusters de classificação na análise do Conjunto de Dados 3 dá suporte esmagador para o número verdadeiro N r = 1 para cada caso e, particularmente, para os casos completos, 15 e 10 principais. Portanto, a alocação de classificadores para rankerclusters é trivial. No entanto, a alocação não é tão direta na análise do conjunto de dados 4. Observe que aqui mantivemos o mesmo anterior para a concentração de nível superior a em vez de alterá-lo para refletir a heterogeneidade conhecida nas crenças do classificador no conjunto de dados 4.

A Figura 4.4 fornece os dendrogramas de ligação completos para os casos $p_1 = 0,5$ e $p_1 = 0,9$ na análise (classificação completa) do Conjunto de Dados 4. O método escolhe claramente os classificadores informativos e homogêneos (numerados de 1 a 40) e os coloca em um único grupo de classificadores. As probabilidades posteriores de que esses rankers sejam co-agrupados são de 0,997 e 0,979, respectivamente - essas probabilidades são fáceis de obter a partir das realizações (posteriores) de alocações de agrupamento e são dadas por $\Pr(c_1 = c_2 = \dots = c_{40} | D)$. Quando $p_1 = 0,5$, vemos que o classificador não informativo mais "semelhante" ao grupo informativo é o classificador 42. Este ranker

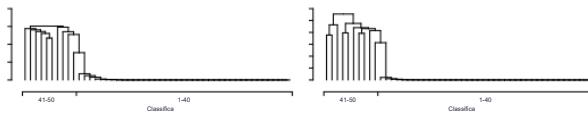


Figura 4.4: Dendrogramas para agrupamento de ranker no conjunto de dados 4 sob uma análise completa forpi = 0,5 (gráfico esquerdo) e pi = 0,9 (gráfico direito).

é co-agrupado com cada um dos classificadores informativos pelo menos 98,5% do tempo. Os outros classificadores não informativos também têm altas probabilidades de co-agrupamento e isso é consistente com o suporte posterior muito alto para um único grupo de classificadores, $\Pr(N = 1 | D, \pi = 0,5) = 0,89$. Este resultado ocorre como consequência do modelo reduzir a influência dos classificadores não informativos: $\Pr(w_i = 1 | D, \pi = 0,5) \approx 0,1$ para $i = 41, \dots, 50$. Por outro lado, quando estamos muito mais confiantes na capacidade dos classificadores (com $\pi = 0,9$), o classificador não informativo mais "semelhante" ao grupo informativo é co-agrupado com classificadores informativos pelo menos 68,6% do tempo, uma proporção muito menor do que no caso $\pi = 0,5$. Além disso, os classificadores não informativos não se separam em um único agrupamento distinto, com $0,311 \leq \alpha_{ij} \leq 0,555$ para qualquer $i, j \in \{41-50\}$, ou seja, qualquer par de classificadores não informativos são co-agrupados entre 44,5% e 68,9% do tempo. Talvez não seja surpreendente que o modelo não seja capaz de detectar semelhanças significativas entre qualquer par de classificadores não informativos, pois suas classificações são permutações aleatórias e são poucas em número.

A Tabela 4.2 fornece as distribuições marginais posteriores do número de clusters de entidades condicionais em um único cluster de classificação (para todas as análises de cada conjunto de dados). Observe que a análise $\pi = 0,9$ do Conjunto de Dados 4 fornece suporte posterior muito baixo para um único cluster ranker, com $\Pr(N = 1 | D, \pi = 0,9) = 0,16$ e, portanto, mais adiante na Seção 4.7.1, analisamos os resultados ao condicionar em dois clusters de classificadores (o número modal posterior). A tabela também mostra que o suporte posterior para o número correto de clusters de entidades ($N = 6$) aumenta à medida que as informações fornecidas em cada classificação aumentam. O custo de realizar uma análise restrita é especialmente visível no caso dos 5 principais. Como foi o caso do agrupamento de classificadores, a inclusão de classificações completas em comparação com as classificações dos 10 primeiros não tem um efeito significativo em nossas crenças posteriores.

A Figura 4.5 mostra dendrogramas da estrutura de agrupamento de entidades para cada uma das análises completas, condicionadas a um único cluster de classificadores. Observe que os clusters de entidades são semelhantes nas análises $\pi = 0,5$ e $\pi = 0,9$, particularmente para as entidades 1-6 e 15-20. As outras entidades (no lado direito de cada dendrograma) também têm um agrupamento semelhante

Conjunto de	$p = 0,5$								
	12	34	56	78	9	≥ 10			
Topo-5*	0.00	0.11	0.23	0.27	0.18	0.12	0.06	0.02	0.01
Topo-10*	0.00	0.00	0.05	0.19	0.28	0.23	0.15	0.06	0.02
Topo-5	0.00	0.00	0.15	0.27	0.24	0.17	0.09	0.04	0.02
10 melhores	0.00	0.00	0.03	0.13	0.22	0.25	0.19	0.11	0.05
Topo-15	0.00	0.00	0.00	0.07	0.21	0.29	0.22	0.13	0.05
Completar	0.00	0.00	0.00	0.09	0.24	0.29	0.21	0.11	0.04
Conjunto de	$p = 0,5$								
	12	34	56	78	9	≥ 10			
Topo-5	0.00	0.00	0.14	0.26	0.25	0.17	0.10	0.05	0.02
10 melhores	0.00	0.00	0.03	0.12	0.22	0.24	0.19	0.12	0.05
Topo-15	0.00	0.00	0.00	0.08	0.22	0.29	0.23	0.12	0.05
Completar	0.00	0.00	0.00	0.10	0.24	0.28	0.21	0.11	0.04
Conjunto de	$p = 0,9$								
	1	2	3	4	5	6	7	8	≥ 10
Topo-5*	0.00	0.10	0.28	0.29	0.20	0.08	0.03	0.02	0.00
Topo-10*	0.00	0.00	0.04	0.18	0.28	0.25	0.15	0.07	0.02
Topo-5	0.00	0.00	0.21	0.31	0.24	0.14	0.06	0.02	0.01
10 melhores	0.00	0.00	0.03	0.11	0.22	0.24	0.20	0.12	0.05
Topo-15	0.00	0.00	0.00	0.08	0.22	0.28	0.23	0.13	0.05
Completar	0.00	0.00	0.00	0.08	0.22	0.30	0.22	0.11	0.05
Conjunto de	1	2	3	4	5	6	7	8	≥ 10
	0.00	0.00	0.17	0.24	0.25	0.18	0.08	0.05	0.02
Topo-5	0.00	0.00	0.02	0.10	0.24	0.24	0.21	0.11	0.05
10 melhores	0.00	0.00	0.00	0.06	0.22	0.28	0.25	0.13	0.05
Topo-15	0.00	0.00	0.01	0.07	0.23	0.28	0.23	0.12	0.05
Completar	0.00	0.00	0.01	0.07	0.23	0.28	0.23	0.12	0.05

Tabela 4.2: Distribuição posterior do número de clusters de entidades, condicionada a um único rankercluster, para análises restritas (*) e completas (irrestritas). Os números em negrito indicam valores modais.

estrutura com apenas pequenas discrepâncias na ordem em que os pares de entidades se agrupam (e o agrupamento ocorre em níveis semelhantes da medida da dissimilaridade A_{ij}). Assim, os dendrogramas são bastante robustos para a adição de classificações não informativas. Isso pode ser parcialmente explicado pelos dendrogramas serem condicionais a um único cluster de classificadores e a análise do Conjunto de Dados 4 identificar corretamente os classificadores não informativos (com $w41: 50 = 0$).

Podemos explorar ainda mais nossa distribuição posterior, investigando onde entidades específicas provavelmente serão classificadas. Considere a probabilidade posterior de que uma entidade específica seja classificada no máximo, ou seja, $P_i = \Pr(\text{entidade } i \text{ no topo} | D)$. A Figura 4.6 exibe P_5 para todas as análises, P_{10} para todas, exceto o caso dos 5 primeiros e P_{15} para os 15 primeiros e análises completas. Curiosamente, as probabilidades posteriores nas análises restritas top-5 e top-10 são muito semelhantes àquelas sob a análise irrestrita (especialmente quando $p_i = 0,5$). Isso sugere que, para esses dados, esse aspecto da distribuição posterior é robusto para saber se as entidades não observadas (na análise restrita) estão ou não incluídas na análise. Os dois saíram

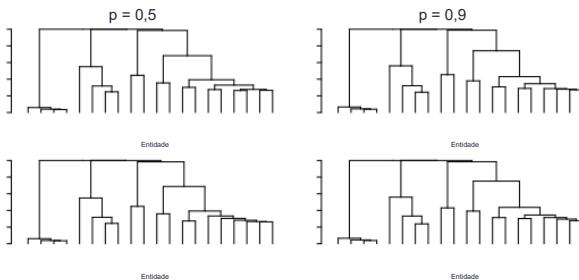


Figura 4.5: Dendrogramas para agrupamento de entidades para o conjunto de dados 3 (superior) e o conjunto de dados 4 (inferior) condicionais em um único cluster de classificação sob ambas as especificações anteriores para as análises completas.

gráficos manuais na Figura 4.6 mostram uma semelhança considerável entre as análises completas (irrestritas) dos dois conjuntos de dados (quando $\pi_i = 0,5$). Aqui, o modelo WAND foi capaz de identificar as chamadas classificações de spam no conjunto de dados 4 (ver Figura 4.3) e, portanto, essas classificações têm pouco efeito na análise. No entanto, este não é o caso quando tomamos $\pi_i = 0,9$. Nesse caso, o alto nível de confiança de que os classificadores são informativos resulta na relutância do modelo WAND em classificar qualquer classificador como não informativo; ver novamente a Figura 4.3. Isso leva os classificadores não informativos a contaminar a distribuição posterior dos parâmetros λ da entidade.

Anteriormente, questionamos se uma coleção de classificações completas é necessária se quisermos apenas inferir, digamos, as 5 principais entidades. Olhando para os resultados do Conjunto de Dados 3 (nos gráficos superiores da Figura 4.6), se considerarmos P5, podemos ver como a análise dos 5 primeiros é capaz de detectar que as entidades 1-4 são altamente prováveis de estar entre os 5 primeiros, no entanto, há alguma dúvida sobre qual é a 5^a entidade mais forte. Em contraste, as outras análises (top-10, top-15 e completa) indicam que a entidade 6 tem muito mais probabilidade de estar entre os top-5 em comparação com as entidades restantes. Além disso, observe como P5 diminui significativamente para entidades 7-20 no top-15 e complete as análises em comparação com o cenário top-10. Resultados semelhantes são obtidos para P10 e P15, ou seja, à medida que aumentamos as informações contidas nos rankings, ficamos mais certos sobre as entidades que são as mais preferidas. Em conclusão, embora seja possível identificar as entidades preferidas sem rankings completos, tem sido vantajoso incorporar o máximo de informações possível.

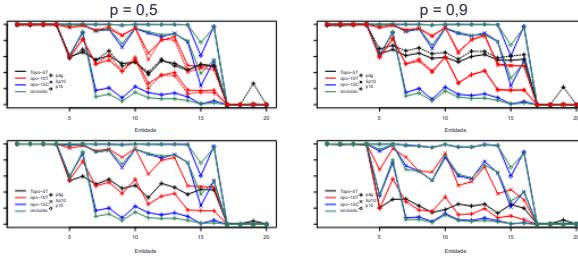


Figura 4.6: Probabilidades posteriores P5 para todas as análises, P10 para todas, exceto o caso dos 5 primeiros e P15 para os 15 primeiros e análises completas. As análises do Conjunto de Dados 3 e do Conjunto de Dados 4 são mostradas na linha superior e inferior, respectivamente, para cada escolha anterior de p .

Análise do conjunto de dados 4 condicional a dois clusters de classificação

Aqui, revisitamos a análise do Conjunto de Dados 4 sob a escolha de $p = 0,9$ a priori. Na seção anterior, examinamos a estrutura de agrupamento de entidades posterior condicionada a um único cluster de ranker. No entanto, houve pouco suporte posterior para um único cluster ranker, com $\Pr(N|D, p = 0,9) \leq 0,16$ em todas as análises consideradas, e o número posterior modal de clusters de ranker foi dois. Portanto, agora olhamos para a estrutura de agrupamento (posterior) condicionada a esse número modal de grupos de classificação (como seria feito se não tivéssemos conhecimento do mecanismo de geração desses dados).

A Tabela 4.3 fornece as distribuições marginais posteriores do número de clusters de entidades dentro de cada grupo de ranker (condicionada a dois clusters de ranker) para todas as análises. Observe que, dentro do cluster de ranker 1, o suporte posterior para o número correto de clusters de entidades ($N_{e1} = 6$) aumenta à medida que as informações fornecidas em cada ranking aumentam - isso também foi observado ao condicionar um único grupo de ranker. Para o cluster ranker 2, vemos maior incerteza nos posteriores marginais (em comparação com o cluster ranker 1), com apenas dois ou três clusters de entidades sendo mais prováveis em todas as análises. Claramente, os rankers no cluster 2 são menos capazes de distinguir entre as entidades - e isso talvez não seja surpreendente, pois esse cluster normalmente abriga os rankers não informativos.

A Figura 4.7 mostra dendrogramas da estrutura de agrupamento de entidades dentro dos clusters ranker 1 e 2 (esquerda e direita, respectivamente) para a análise completa do Conjunto de Dados 4 com $p = 0,9$, condicional a dois clusters ranker. Observe que os clusters de entidades no cluster de classificadores 1 são muito semelhantes aos do condicionamento em um único cluster de classificadores; ver Figura 4.5. Este

Conjunto de	Cluster	12	34	56	78	9	≥ 10
Topo-5	1	0.01	0.01	0.16	0.26	0.25	0.17
	2	0.14	0.23	0.22	0.18	0.12	0.06
10 melhores	1	0.00	0.00	0.02	0.11	0.21	0.23
	2	0.11	0.21	0.24	0.19	0.13	0.07
Topo-15	1	0.00	0.00	0.00	0.08	0.21	0.28
	2	0.18	0.25	0.21	0.16	0.10	0.05
Completar	1	0.00	0.00	0.00	0.08	0.25	0.28
	2	0.20	0.24	0.23	0.15	0.10	0.05

Tabela 4.3: Distribuição posterior do número de clusters de entidades, condicionada a dois clusters de classificação, para cada análise do Conjunto de Dados 4 com $\pi = 0.9$. Os números em negrito indicam valores modais.

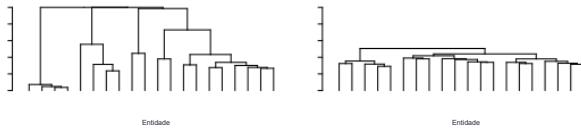


Figura 4.7: Dendrogramas de agrupamento de entidades (condicional a 2 clusters ranker) no cluster ranker 1 (esquerda) e cluster ranker 2 (direita) para a análise do conjunto de dados 4 com $\pi = 0.9$.

é provavelmente devido ao cluster de classificadores 1 contendo todos os classificadores informativos em ambos os casos (de um ou dois grupos de classificadores). Para o cluster de classificação 2, vemos que quaisquer duas entidades são agrupadas juntas pelo menos 49% do tempo ($\Delta_{ij} < 0,51$). Além disso, os agrupamentos de entidades são formados em níveis semelhantes de dissimilaridade, destacando novamente a incerteza sobre o agrupamento de entidades dentro desse grupo hierárquico.

4.7.2 Estudo 2

No estudo 2, analisamos um único conjunto de dados com $n = 40$ classificações completas de $K = 20$ entidades de classificadores informativos ($w_i = 1$). Também simulamos as alocações de cluster (para classificadores e entidades) marginalmente do anterior. Os parâmetros de habilidade distintos e alocações de agrupamento foram simulados usando $\alpha = ys = 1$ para $s \in N$, e $a = 1$ para que nossa distribuição de base seja $G_0 = Ga(1, 1)$. A simulação deu três clusters de classificadores ($N r = 3$) contendo 24, 12 e 4 classificadores, que rotulamos como classificadores 1–24, 25–36 e 37–40. Além disso, os rankerclusters continham 8, 6 e 3 clusters de entidades ($N e1 = 8$, $N e2 = 6$, $N e3 = 3$). A Tabela 4.4 mostra o agrupamento de entidades (dentro de cada cluster de classificadores) junto com os valores true associados dos parâmetros de habilidade na escala de log. Para facilitar a interpretação, as entidades são rotuladas de acordo com o tamanho de seu parâmetro de habilidade "agregado verdadeiro", maior para

Cluster de entidades							
CrRankers1234\$67811–241610,133,4,7,9,12,1525,11,14,17–20168		2.47	0.65	0.52	0.40	0.34	0.24
2	25–36	2–5,6	1,7,9,1	14,16	12	6,10,15,17	13,16–20
3	37–40	1,72	1,0,76	0,68	0,42	0,21	0,15
		2,6,10	7	1,3–5,8,9,11–20			
		1,54	1,16	0,64			

Tabela 4.4: Alocação real de entidades em clusters, juntamente com o valor do parâmetro true correspondente para cada um dos clusters de entidades.

menores, para que sejam rotulados com a entidade mais preferida em geral primeiro, até a entidade menos preferida em geral por último. Aqui, os valores agregados verdadeiros são uma média dos valores trueparameter dentro de cada cluster de classificadores, ponderados pelo tamanho dos clusters de classificadores. As classificações completas (simuladas) analisadas neste estudo podem ser encontradas na Tabela B.5 nos apêndices.

O objetivo deste estudo é investigar a capacidade do nosso modelo WAND de identificar (corretamente) diferentes grupos de classificadores e as preferências associadas a eles. A análise dada aqui usa a mesma distribuição de base e distribuição anterior para os parâmetros de concentração da entidade como na Seção 4.7.1, ou seja, $G_0 = \text{Ga}(1, 1)$ e $y_S \sim \text{Ga}(3, 3)$ para $s \in N$. Para refletir a heterogeneidade conhecida do ranker dentro desses dados, agora tomamos $a_\theta = b_\theta = 3$, ou seja, $\alpha \sim \text{Ga}(3, 3)$. Também consideramos o caso em que temos apenas uma confiança moderada em nossos classificadores sendo informativos, tomando $\pi_i = 0,5$.

As realizações da distribuição a posteriori foram obtidas usando o algoritmo de amostragem (marginal) descrito na Seção 4.6.5 com $m = 2$ e $m_e = 3$ variáveis auxiliares (ranker e entidade). A cadeia de Markov foi inicializada em um sorteio aleatório da distribuição anterior. Para obter 10K (quase) realizações não autocorrelacionadas da distribuição posterior, realizamos um período de queima de 10K iterações e, em seguida, executamos o esquema para mais 1M iterations e reduzimos a produção por um fator de 100. O tempo computacional necessário para realizar a inferência foi (aproximadamente) de 17 minutos. A mistura da cadeia MCMC foi avaliada inspecionando gráficos de rastreamento e a convergência foi avaliada inicializando várias cadeias em diferentes valores iniciais e verificando se as distribuições posteriores resultantes eram equivalentes (até ruído estocástico).

O gráfico da esquerda na Figura 4.8 mostra as probabilidades posteriores, $\Pr(w_i = 1|D)$, esse classificador é informativo. O gráfico mostra que, em geral, os classificadores nos (verdadeiros) agrupamentos de classificadores 1 e 2 (classificadores 1–36) são bem identificados como informativos. No entanto, os classificadores no (verdadeiro) cluster 3 são identificados como não informativos. A razão para essa identificação incorreta talvez se deva à (verdadeira) estrutura de agrupamento de entidades presente no cluster de classificação 3. A Tabela 4.4 mostra que esse cluster de classificação contém apenas 3 clusters de entidades, com um deles contendo 16 das 20 entidades e, portanto, é muito provável que as classificações neste cluster

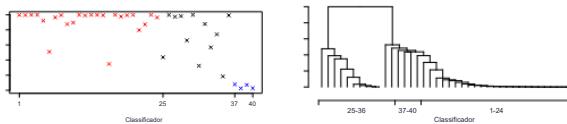


Figura 4.8: Gráfico da probabilidade posterior $\text{Pr}(w_i = 1|D)$ que o ranker i é informativo (à esquerda), as cores distinguem entre os clusters de ranker "verdadeiros". Dendrograma (ligação completa) calculado usando a dissimilaridade Δ_{ij} entre os rankers i e j (direita).

assemelham-se a uma permutação aleatória das entidades K .

O gráfico da direita na Figura 4.8 mostra o dendrograma de ligação completa determinado usando dissimilaridades Δ_{ij} . O dendrograma sugere que existem dois agrupamentos de classificadores (tomando a dissimilaridade $\epsilon \in (0,53, 1)$) que separa os classificadores numerados {25 – 30, 32, 33, 34, 36} dos classificadores restantes. O fato de existirem dois agrupamentos de classificadores é apoiado ainda mais pela distribuição marginal posterior do número de agrupamentos: $\text{Pr}(N = i|D) = 0,68, 0,25, 0,06, 0,01$ para $i = 2, 3, 4, 5$. Não é surpreendente que a análise não tenha identificado o terceiro grupo de classificadores, pois esse agrupamento contém apenas classificadores cujas classificações são virtualmente indistinguíveis de permutações aleatórias; em vez disso, o modelo preferiu considerar esses classificadores como não informativos e colocá-los dentro de grupos de classificadores informativos.

A Tabela 4.5 fornece a distribuição marginal posterior para o número de clusters de entidades dentro de cada cluster de ranker, condicionada ao número modal posterior de clusters de ranker. O número modal de clusters de entidades nos clusters de classificação 1 e 2 é seis e quatro, respectivamente (os valores verdadeiros correspondentes são oito e seis). Aqui, a análise identificou corretamente que o cluster ranker 1 é o cluster mais forte, na medida em que esses rankers são mais capazes de distinguir entre entidades. Os dendrogramas na Figura 4.9 sugerem que existem cinco clusters de entidades dentro do cluster ranker 1 (tomando ϵ de dissimilaridade (0,58, 0,83)) e três clusters de entidades no cluster ranker 2 (tomando ϵ de dissimilaridade (0,50, 0,69)). Observe que no classificador 1, a entidade preferida neste cluster (entidade 1) tem seu próprio cluster, e as entidades 16 e 8 (nos verdadeiros clusters de entidades 7 e 8) também formam um único cluster; Talvez isso não seja surpreendente, dados os valores "verdadeiros" dos parâmetros de habilidade para essas entidades dentro deste classificador

Cluster	1	2	3	4	5	6	7	8	9	≥ 10
1	0.00	0.00	0.00	0.07	0.20	0.25	0.21	0.14	0.08	0.05
2	0.00	0.11	0.25	0.26	0.18	0.11	0.05	0.02	0.01	0.01

Tabela 4.5: Distribuição posterior do número de clusters de entidades, condicionada a dois clusters de classificação.

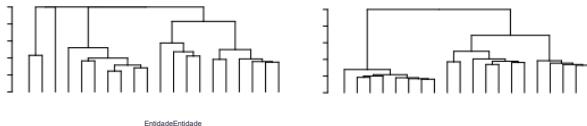


Figura 4.9: Dendogramas mostrando a dissimilaridade entre entidades dentro dos clusters ranker 1 (esquerda) e 2 (direita), condicionais a dois clusters ranker ($N = 2$).

cluster (ver Tabela 4.4). O cluster de entidade verdadeiro 6 é bastante bem identificado, com apenas a entidade 14 não sendo incluída e a entidade 3 (do cluster verdadeiro 4) se juntando ao cluster. Os dois clusters de entidades restantes identificados pelo dendrograma abrigam as outras entidades dos verdadeiros clusters de entidades 2–5. No cluster 2 do classificador, a estrutura de agrupamento de entidades "verdadeira" a partir da qual os dados foram simulados é amplamente preservada, mas os clusters inferidos são grupos de clusters "verdadeiros", com todas as entidades no cluster 1 "verdadeiro" sendo claramente identificadas em um cluster e aquelas nos clusters 2, 3 e 4 em outro cluster e aquelas nos clusters 5 e 6 em outro cluster. O facto de estes agrupamentos de entidades se terem fundido talvez não seja muito surpreendente, dados os valores reais (ver Quadro 4.4) e o número limitado de classificações observadas.

Agora investigamos a ordem de preferência das entidades dentro de cada grupo de classificação e uma ordem geral de preferência; ver Tabela 4.6. Aqui, a ordem de preferência dentro de cada grupo de classificação foi determinada pela média posterior dos parâmetros de "habilidade", calculada em média tanto no agrupamento de entidades quanto na alocação de classificadores para cada grupo de classificadores. A ordem geral de preferência foi calculada em todos os clusters de classificação. Comparando essas ordenações de preferência com a verdade (na Tabela 4.4), vemos que o modelo WAND teve um desempenho bastante bom na recuperação das verdadeiras preferências expressas nos clusters de classificação 1 e 2, especialmente para as entidades que são as mais e menos preferidas dentro de cada grupo de classificação. Não é de surpreender que haja um aumento do nível de identificação incorreta nas classificações intermediárias da ordem de preferência para ambos os clusters de ranker e, particularmente, para o cluster ranker 1. Isso talvez se deva, em parte, aos valores reais dos parâmetros de habilidade nos clusters de entidades 2–6 dentro de cada cluster ranker serem bastante semelhantes, com aqueles no cluster ranker 1 sendo os mais semelhantes; ver Tabela 4.4.

As entidades na Tabela 4.6 são listadas em ordem de seu parâmetro de habilidade "verdadeiro" geral. Embora o modelo WAND tenha permitido diferenças entre os classificadores, a ordenação geral inferida é muito diferente da ordem "verdadeira". Dito isso, as ordenações inferidas dentro dos clusters de classificação são muito semelhantes às ordenações "verdadeiras" e fornecem uma explicação muito melhor da heterogeneidade dentro do modelo que sustenta os dados. Isso ilustra como inferir ordenações de preferência usando resumos gerais (nível de população) de classificadores heterogêneos.

Classificador	Entidade	Cr1		Cr2		Agregado	
		Média (DP)	(DP)	Entidade	Média (DP)	Entidade	Média (DP)
11		3.54 (1.57)		2	1.91 (1.11)	1	2.93 (1.15)
2	7	1.02 (0.54)		8	1.77 (1.08)	7	1.13 (0.47)
3	13	1.02 (0.54)		5	1.73 (1.07)	4	1.02 (0.41)
4	10	0.91 (0.46)		4	1.71 (1.06)	2	0.97 (0.40)
5	14	0.89 (0.45)		3	1.65 (1.05)	14	0.96 (0.40)
66		0.87 (0.44)		16	1.51 (1.02)	12	0.95 (0.39)
7	12	0.83 (0.42)		1	1.44 (0.99)	9	0.89 (0.38)
8	15	0.76 (0.39)		7	1.41 (0.98)	3	0.82 (0.35)
94		0.75 (0.40)		9	1.36 (0.96)	13	0.80 (0.39)
10	9	0.70 (0.38)		11	1.31 (0.95)	5	0.78 (0.34)
11	2	0.60 (0.35)		12	1.23 (0.92)	10	0.73 (0.33)
12	3	0.49 (0.28)		14	1.13 (0.87)	6	0.69 (0.31)
13	20	0.44 (0.24)		10	0.25 (0.21)	11	0.64 (0.29)
14	18	0.44 (0.24)		17	0.22 (0.17)	15	0.62 (0.29)
15	17	0.41 (0.21)		15	0.22 (0.16)	8	0.52 (0.31)
16	5	0.41 (0.21)		20	0.22 (0.16)	16	0.47 (0.29)
17	11	0.37 (0.19)		13	0.21 (0.15)	20	0.39 (0.18)
18	19	0.37 (0.18)		19	0.21 (0.15)	18	0.38 (0.18)
1916		0.05 (0.07)		6	0.21 (0.15)	17	0.36 (0.16)
20	8	0.03 (0.08)		18	0.20 (0.14)	19	0.33 (0.14)

Tabela 4.6: Ordenações de preferência posteriores nos clusters de classificação 1 e 2 (condicionadas a dois clusters de classificação) e a classificação geral/agregada, com média (e desvio padrão) de seus parâmetros de habilidade.

pode ser muito enganoso mesmo quando se conhece os parâmetros de habilidade, muito menos ao tentar inferir seus valores.

4.8 Resumo

Neste capítulo, descrevemos o processo Adaptado, Nested Dirichlet anterior, que facilita o agrupamento bidirecional em classificadores e entidades (dentro de grupos de classificadores). Em seguida, usamos isso antes de formar o modelo WAND, tomando a distribuição de classificação subjacente como o modelo Weighted Plackett-Luce. Duas abordagens de inferência para o modelo WAND foram então consideradas. Na Seção 4.5, apelamos para uma abordagem de amostragem condicional. Embora intuitiva, essa abordagem de inferência para misturas de DP apresenta desvantagens quando comparada aos esquemas de amostragem marginal, conforme discutido no Capítulo 3. Na Seção 4.6, discutimos como um esquema marginal para amostragem posterior pode ser construído para essa adaptação do NDP. O esquema de amostragem marginal posterior que descrevemos na Seção 4.6.5 permite uma inferência rápida e eficiente em nosso modelo WAND.

Vimos através dos estudos de simulação na Seção 4.7 que inferências razoáveis podem ser feitas sob o modelo WAND, mesmo quando apenas informações limitadas (parciais) estão disponíveis.

A riqueza de informações na distribuição posterior nos permite inferir muitos detalhes da estrutura tanto entre grupos de rankers quanto entre grupos de entidades (dentro de rankergroups). A alta dimensão da distribuição a posteriori pode dificultar bastante a produção de resumos perspicazes, mas simples, e exploramos diferentes abordagens, que vão desde o condicionamento do número modal de grupos até a adoção de uma classificação baseada em cálculos de um resumo de matriz de dissimilaridade.

No próximo capítulo, consideraremos dois conjuntos de dados reais que foram analisados na literatura e comparamos suas conclusões com as obtidas com o ajuste do modelo WAND.

Capítulo 5

Análises de dados reais

5.1 Conjunto de dados da Roskam

Nesta seção, consideramos um conjunto de dados originalmente coletado em 1968 por Roskam, mais recentemente estudado por de Leeuw (2006). Os dados estão disponíveis no pacote R `homals` (de Leeuw and Mair, 2009) e também são fornecidos na Tabela B.6 nos apêndices. Os dados consistem em classificações obtidas de $n = 39$ psicólogos do Departamento de Psicologia da Universidade de Nijmegen (Holanda). Cada classificador dá uma classificação completa de $K = 9$ subáreas (entidades), listadas de acordo com a adequação das subáreas ao seu trabalho. As sub-áreas são: SOC - Psicologia Social, EDU - Psicologia da Educação e do Desenvolvimento, CLI - Psicologia Clínica, MAT - Psicologia Matemática e Estatística Psicológica, EXP - Psicologia Experimental, CUL - Psicologia Cultural e Psicologia da Religião, IND - Psicologia Industrial, TST - Construção e Validação de Testes, e PHY - Psicologia Fisiológica e Animal.

A heterogeneidade dentro desses dados foi analisada por de Leeuw (2006) usando uma análise de componentes principais não linear para detectar agrupamentos dentro dos rankings. Sua análise apoiou a ideia de que existem dois grupos de rankings: um grupo que favorece os campos qualitativos e o outro que favorece os campos quantitativos da psicologia. Uma análise de homogeneidade foi posteriormente realizada por de Leeuw e Mair (2009), que expôs agrupamentos de entidades dentro dos rankings. Mais recentemente, Choulakian (2016) realizou uma análise de correspondência de táxi para observar a estrutura entre os rankings e as entidades dentro dos grupos de classificação. Seus resultados apóiam as conclusões de de Leeuw (2006) e sugerem que os psicólogos compreendem dois grupos homogêneos com 23 e 16 membros, respectivamente. Dentro do grupo maior de classificadores, eles obtêm o agrupamento de entidades {MAT,EXP} {IND, TST} {PHY, SOC, EDU} CLI CUL, onde significa "é preferível", e áreas quantitativas da psicologia parecem ser preferidas. A pista correspondente

A classificação de entidades para o outro grupo de classificação é {EDU, CLI, SOC} {CUL, MAT, EXP} {TST, IND} PHY, e aqui as áreas qualitativas da psicologia parecem ser preferidas. Eles também concluem que o grupo maior de rankers é um pouco mais homogêneo do que o grupo menor.

Agora usamos nosso modelo WAND para investigar a estrutura de subgrupos nesses dados e consideramos nossa especificação prévia para a distribuição de base e parâmetros de concentração como $a = 1$ e $ad = ba = 1$, $ay = by = 3$. Esses dados contêm ordenações de preferências individuais que acreditamos ser informativas e, portanto, tomam $\pi = 0,75$. A distribuição posterior é bastante robusta para essa escolha; uma análise de sensibilidade segue na Seção 5.1.1. Relatamos os resultados de uma execução típica de nosso esquema MCMC inicializado a partir do anterior, com um burn-in de 10 mil iterações e, em seguida, executamos mais 1 milhão de iterações e diluídas em 100 para obter 10 mil realizações (quase) não autocorrelacionadas da distribuição posterior. A convergência foi avaliada usando vários valores iniciais, inspeção de traceplots de parâmetros e a probabilidade de dados log completos e estatísticas padrão disponíveis na coda do pacote R (Plummer et al., 2006). O esquema MCMC é executado rapidamente, com o código C em um singlthread de uma CPU Intel Core i7-4790S (velocidade de clock de 3,20 GHz) levando cerca de 5 minutos.

A Tabela 5.1 mostra a distribuição anterior e posterior para o número de clusters ranker. Os dados foram claramente informativos e sugerem que é provável que existam entre dois e quatro grupos de classificadores, sendo dois grupos os mais plausíveis. Observe que quase não há suporte posterior para sugerir que existe um único grupo de classificação (homogêneo) e, portanto, uma classificação agregada desse conjunto de dados pode ser enganosa. A Figura 5.1 mostra o dendrograma dos classificadores junto com a probabilidade posterior de que cada classificador seja informativo. O dendrograma sugere que existem dois grupos de classificação (considerando a dissimilaridade $> 0,60$), e isso é consistente com a distribuição posterior na Tabela 5.1 e as conclusões de análises anteriores. Observamos que os dados são consistentes com a maioria dos classificadores sendo informativos (com $\Pr(w_i = 1|D) \geq 0,8$), um aumento de suas probabilidades anteriores ($\pi = 0,75$). Além disso, os classificadores cujas probabilidades diminuíram (classificadores 1, 5, 8, 10, 13, 14, 15, 31) são aqueles com preferências (ligeiramente) diferentes e, portanto, atrasados para se juntar ao agrupamento da direita no dendrograma.

Agora nos voltamos para a estrutura de subgrupos de entidades dentro dos clusters ranker, e aqui condicionamos a existência de dois clusters ranker. A Figura 5.2 mostra a distribuição posterior (marginal) para o número de clusters de entidades dentro de cada cluster de ranker junto com o

1	2	3	4	5	6	7	8
Posterior	0,00	0,43	0,33	0,16	0,06	0,02	
Anterior	0,00	0,20	0,18	0,16	0,13	0,10	0,08
	0,10						

Tabela 5.1: Distribuição anterior e posterior do número de clusters de classificação (até 2 d.p.).

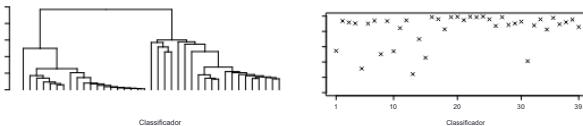


Figura 5.1: Conjunto de dados de Roskam: Dendrograma (à esquerda) mostrando a estrutura do cluster ranker junto com a probabilidade posterior, $\Pr(w_i = 1|D)$, para cada classificador i (direita).

distribuição prévia. Os dendrogramas na Figura 5.3 mostram a estrutura de agrupamento de entidades em cada cluster de classificadores. Definimos clusters de entidades em dissimilaridades nos intervalos $(0,45,0,95)$ e $(0,63,0,89)$ para os grupos de classificadores 1 e 2, respectivamente, e formamos uma ordenação de preferência desses agrupamentos de entidades examinando os posteriores marginais para os parâmetros de habilidade $\text{Acidcijd}^{\text{ent}}$ dentro de cada grupo de classificadores ci . Condicionando essas alocações para ranker e entitygroups e ordenando por média posterior, obtemos $\{\text{EXP}, \text{MAT}\}$ (TST, PHY, IND) $\{\text{EDU}, \text{SOC}, \text{CLI}\}$ {CUL} (com cluster de entidade significa $3,02, 0,72, 0,22, 0,06$) em rankercluster 1 e $\{\text{SOC}, \text{EDU}, \text{CLI}, \text{MAT}\}$ {CUL, IND, EXP, TST} {PHY} (com entitycluster significa $1,96, 0,82, 0,12$) no ranker cluster dois. Esses agrupamentos de entidades (dentro de grupos de classificação) são semelhantes aos fornecidos por Choulakian (2016). Além disso, se usarmos o valor médio de $\Pr(w_i = 1|D)$ como medida de homogeneidade dentro de um cluster ranker, obtemos $0,68$ e $0,56$ para os clusters 1 e 2, respectivamente, o que novamente concorda com a conclusão de Choulakian (2016) de que o cluster ranker 1 é mais homogêneo do que o cluster ranker 2. Observe que, para esta análise de dados, obtemos uma ordenação de entidade muito semelhante usando médias marginais posteriores dos parâmetros de habilidade dentro de cada grupo de classificadores (marginal sobre a distribuição de clusters de entidades); ver Tabela 5.2. De fato, a tabela sugere que os grupos de classificação quase têm preferências opostas (reversas) entre si.

Analisamos a sensibilidade da distribuição posterior (e inferências) a mudanças modestas na distribuição anterior. A distribuição posterior foi bastante insensível a

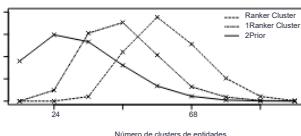


Figura 5.2: Densidades anteriores e posteriores marginais para o número de clusters de entidades dentro de cada cluster de ranker (condicional a dois clusters de ranker).

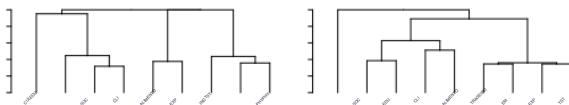


Figura 5.3: Conjunto de dados de Roskam: Dendrogramas mostrando a estrutura de agrupamento de entidades dentro do rankercluster 1 e 2 (esquerda e direita, respectivamente) condicionada a dois clusters ranker.

Classificador cluster1234567891EXPMATTSTPHYINDEDUSOCCLICUL	Classificar									
	3.13	2.68	0.76	0.70	0.63	0.27	0.22	0.20	0.07	
2	SOC	EDU	CLI	ALIMEN	TRASEI	EM	EXP	TST	PHY	
	1.95	1.75	1.49	1.32	0.94	0.90	0.87	0.87	0.10	

Tabela 5.2: Conjunto de dados de Roskam: classificações de entidades por média posterior dentro do cluster ranker (condicionalon dois clusters ranker). A classificação 1 corresponde à entidade mais preferida em cada cluster.

Alterações no índice (a) da distribuição de bases gama e alterações nos parâmetros (aa , ba , ay , by) das distribuições a priori gama para os parâmetros de concentração. A distribuição posterior foi mais sensível às mudanças nas probabilidades anteriores (pi) de os classificadores serem informativos. Não surpreendentemente, os mais afetados por tais mudanças foram seus equivalentes posteriores $Pr(wi = 1 | D)$, embora a conclusão de dois grupos de classificação e a participação nesses grupos tenham sido robustas. A alocação de entidades para grupos (dentro de cada cluster de ranker) também foi bastante robusta, com apenas uma pequena mudança na alocação no caso $p = 0,85$. A seção 5.1.1 contém os dendrogramas (classificador e entidade) e gráficos de $Pr(wi = 1 | D)$ para $pi = 0,65$ e $pi = 0,85$, além da escolha $pi = 0,75$ utilizada nesta análise.

5.1.1 Análise de sensibilidade prévia

Aqui olhamos para a sensibilidade da distribuição posterior a mudanças na probabilidade anterior de que um ranker seja informativo. Consideramos duas opções alternativas à utilizada em nossa análise anterior ($pi = 0,75$), a saber, $pi = 0,65$ e $pi = 0,85$. Para facilitar a referência, também incluímos os resultados para o caso $pi = 0,75$.

No geral, descobrimos que a distribuição posterior foi bastante robusta para a escolha da pi priori. Talvez sem surpresa, o aspecto de nossa distribuição posterior mais sensível a mudanças no pi foram seus equivalentes posteriores $Pr(wi = 1 | D)$; veja a Figura 5.4 (coluna da direita). No entanto, notamos que os rankers cuja probabilidade informativa diminui (anterior

→ posterior) permanecem os mesmos em cada caso: estes são os classificadores $\{1, 5, 8, 10, 13, 14, 15, 31\}$. Os dendrogramas da estrutura de agrupamento de classificadores são semelhantes para cada escolha anterior e indicam claramente que existem dois grupos de classificadores. Além disso, a alocação de classificadores para clusters é semelhante em cada caso; veja a Figura 5.4 (coluna da esquerda).

Curiosamente, observamos um aumento do suporte posterior para dois clusters ranker à medida que o π diminui - este também é o modo posterior em cada caso; veja a Figura 5.5. Além disso, condicionada à existência de clusters de dois rankers, o marginal posterior do número de clusters de entidades N es (dentro de cada cluster de ranker $s = 1, 2$) permanece bastante robusto para a escolha anterior; veja a Figura 5.5. Além disso, os dendrogramas da estrutura de agrupamento de entidades são semelhantes em cada caso; ver figura 5.6.

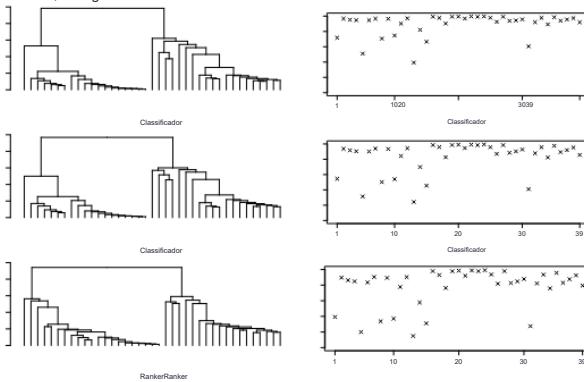


Figura 5.4: Conjunto de dados de Roskam: Dendrograma (à esquerda) mostrando a estrutura de agrupamento dos classificadores junto com a probabilidade posterior $\Pr(w_i = 1 | D)$ para cada classificação i (direita) para $\pi = 0,85, 0,75, 0,65$ (de cima para baixo, respectivamente).

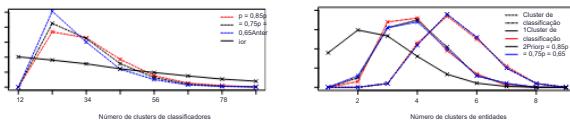


Figura 5.5: Conjunto de dados de Roskam: Densidades anteriores e marginais posteriores para o número de clusters rankers (gráfico esquerdo) e o número de clusters de entidades dentro de cada cluster ranker, condicional a clusters tworanker, (gráfico direito) para $\pi = 0,85, 0,75, 0,65$.

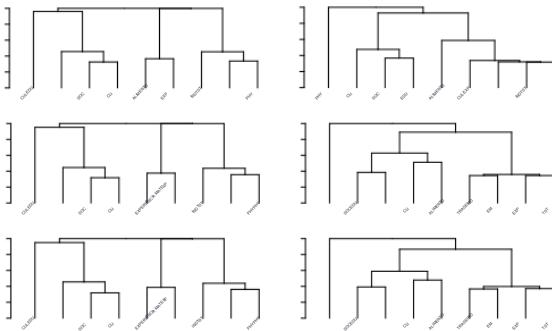


Figura 5.6: Conjunto de dados de Roskam: Dendrogramas da estrutura de agrupamento de entidades dentro do cluster ranker 1 (esquerda) e cluster ranker 2 (direita). Estes são mostrados para cada especificação anterior, $\pi = 0,85, 0,75, 0,65$, de cima para baixo, respectivamente.

5.2 Estudo da NBA

Consideramos agora outro conjunto de dados de classificações, estudado por Deng et al. (2014) e envolvendo classificações de equipes da NBA (National Basketball Association). Em seu artigo, Deng et al. propõem um modelo (denominado "Agregação Bayesiana de Dados Classificados", BARD) que visa agregar classificações e identificar as "entidades relevantes". Seu modelo também acomoda a possibilidade de que as classificações possam não ser igualmente confiáveis. Uma desvantagem do modelo BARD é que ele assume que todas as classificações vêm de um único grupo homogêneo. Agora investigamos essa suposição usando o modelo WAND e também produzimos uma classificação agregada para comparar com a classificação agregada do BARD.

Em 2011/12, a liga da NBA continha $K = 30$ equipes (entidades) e o conjunto de dados que consideramos tem uma classificação dessas equipes de cada um dos $n = 34$ classificadores. Os seis primeiros rankings completos foram obtidos a partir de probabilidades dadas em sites "profissionais" e os outros 8 primeiros rankings obtidos de amadores. Além disso, cada amador foi solicitado a se classificar em um dos seguintes grupos: "Fãs ávidos" (nunca perderam um jogo da NBA), "Fãs" (assistiram aos jogos da NBA com frequência), "Observadores pouco frequentes" (ocasionalmente assistiram aos jogos da NBA) e "Não interessados" (nunca assistiram a um jogo da NBA). Cada classificador considerou todas as equipes e, portanto, temos $K_i = K$ para $i = 1, \dots, n$. Os classificadores são numerados da seguinte forma: Profissionais (1-6), Fãs ávidos (7-12), Fãs (13-18), Observadores pouco frequentes (19-25) e Não interessados (26-34). Portanto, temos $n_i = K = 30$ para $i = 1, \dots, 6$ e $n_i = 8$ para $i = 7, \dots, n$. Os dados são apresentados no quadro B.7 nos apêndices. Mais detalhes sobre como esses dados foram coletados podem ser encontrados em Deng et al. (2014).

Agora analisamos esses dados usando nosso modelo WAND e vemos se é plausível que esses classificadores sejam homogêneos ou se os grupos autoavaliados se comportam de maneira diferente. Pegamos o mesmo prior para a distribuição base ($a = 1$) do exemplo anterior. No entanto, para refletir crenças fracas de que existem vários grupos de classificação, tomamos $ad = bc = 3$ além da escolha anterior para entidades, $ay = by = 3$. O antes que escolhemos a habilidade de cada ranker é baseado em quanta atenção eles supostamente prestam à NBA, com os rankers profissionais provavelmente sendo os mais informativos, seguidos pelos fãs ávidos, depois pelos fãs e assim por diante. Fazemos isso dando o mesmo valor π_i para cada ranker no mesmo grupo de "habilidade", com $\pi_i = 0,9$ para profissionais, $\pi_i = 0,7$ para fãs ávidos, $\pi_i = 0,5$ para fãs, $\pi_i = 0,3$ para observadores pouco frequentes e $\pi_i = 0,1$ para não interessado. Reconhecemos que, em geral, é improvável que esse tipo de informação esteja disponível para o analista e, portanto, na Seção 5.2.1, consideramos uma análise em que $\pi_i = 0,5$ para cada classificador (uma escolha sensata quando nenhuma informação está disponível). Geralmente, descobrimos que a distribuição posterior é bastante robusta para a escolha de π_i a priori, o que talvez não seja surpreendente, dado o que vimos nos estudos de simulação anteriores e (outras) análises de dados reais. Sem surpresa, o aspecto da distribuição posterior mais sensível a mudanças no π_i foi o seu posterior

equivalentes $\Pr(w_i = 1|D)$ embora as inferências em cada análise tenham sido robustas. Como na análise anterior, relatamos os resultados de uma execução típica de nosso esquema MCMC inicializado a partir do anterior, com um burn-in de 10K iterações e, em seguida, executado por mais 1M iterations e diluído em 100 para obter 10K (quase) realizações não autocorrelacionadas da distribuição posterior. A convergência foi avaliada usando vários valores iniciais, inspeção de traceplots de parâmetros e a probabilidade de dados completos de log e estatísticas padrão disponíveis na coda do pacote R. Novamente, o esquema MCMC é executado razoavelmente rápido, com o código C em um único thread de uma CPU Intel Core i7-4790S (velocidade de clock de 3,20 GHz) levando pouco menos de 18 minutos.

Nossa análise das realizações posteriores revela muito pouco suporte posterior para um único grupo homogêneo de classificadores, com maior suporte para dois grupos de classificadores ($\Pr(N_r = 1|D) = 0,00$, $\Pr(N_r = 2|D) = 0,80$ e $\Pr(N_r = 3|D) = 0,17$). A Figura 5.7 (à esquerda) mostra um dendrograma da estrutura de agrupamento posterior dos classificadores e confirma a conclusão de que existem dois grupos distintos de classificadores: um com classificadores 1-10, 12, 15 e o outro com classificadores 11, 14, 17-26, 28 e 32. Quase todos os outros classificadores são classificados como não informativos, com $\Pr(w_i = 1|D) < 0,25$, exceto o classificador informativo 16, que tem (aproximadamente) a mesma probabilidade de ser alocado a cada cluster; veja a Figura 5.7 (direita). Observe que obter um clustering usando a alocação MAP seria enganoso, pois a alocação MAP ocorre em apenas 60 das iterações de 10K na cadeia MCMC. Sem surpresa, os classificadores não informativos são tipicamente aqueles que prestam menos atenção à NBA, com valores médios de $\Pr(w_i = 1|D)$ para classificadores nos grupos autocertificados (de profissionais até indivíduos não interessados) de 1, 1, 0,87, 0,88, 0,34, respectivamente. Uma conclusão semelhante foi encontrada no BARB por meio de seus parâmetros de qualidade de classificação; ver Figura 8 em Deng et al. (2014).

A Figura 5.8 mostra a distribuição marginal posterior para o número de clusters de entidades dentro de cada cluster ranker (condicional à existência de dois clusters ranker) juntamente com a distribuição anterior. O número médio posterior de clusters de entidades para os clusters ranker 1 e 2 é 8,88 e 4,58, respectivamente, com desvios padrão correspondentes de 1,55 e 1,29.

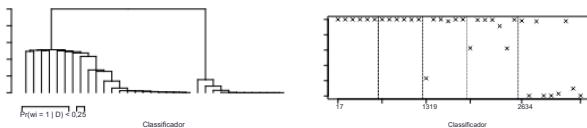


Figura 5.7: Conjunto de dados da NBA: Dendrograma (à esquerda) mostrando a estrutura de agrupamento de classificadores e destacando esses classificadores com $\Pr(w_i = 1|D) < 0,25$. Gráfico (à direita) das probabilidades posteriores $\Pr(w_i = 1|D)$ para cada ranker, com linhas verticais separando os grupos autocertificados.

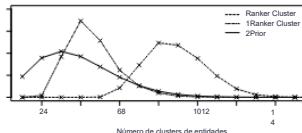


Figura 5.8: Densidades anteriores e posteriores marginais para o número de clusters de entidades dentro de cada cluster ranker (condicional a dois clusters ranker).

Essas distribuições sugerem que os classificadores dentro do cluster 1 são capazes de distinguir entre muito mais entidades do que aquelas no cluster 2. Novamente, isso não deve ser surpresa, já que o rankercluster 2 consiste principalmente em rankers que normalmente prestam pouca atenção à NBA. Os dendrogramas na Figura 5.9 mostram o agrupamento de entidades em cada cluster ranker e sugerem que existem seis clusters de entidades distintos dentro do cluster ranker 1 (considerando a dissimilaridade $> 0,81$) e três clusters de entidades no cluster ranker 2 (considerando a dissimilaridade $> 0,61$). Observamos que o agrupamento MAP fornece seis e dois clusters de entidades, respectivamente, embora haja relativamente poucas iterações MCMC contribuindo para a alocação MAP para qualquer cluster.

Também é interessante examinar as diferenças de preferências entre os dois clusters de classificação examinando as classificações agregadas dentro do cluster; ver Tabela 5.3. Como antes, eles são determinados pela média marginal posterior para cada entidade (dentro de cada cluster de classificadores). As linhas horizontais nesta tabela mostram o agrupamento de entidades MAP descrito acima e o número (muito pequeno) de ocorrências do MAP também é fornecido. Para que nossos resultados possam ser comparados aos de Deng et al. (2014), a tabela também inclui a classificação agregada geral, determinada ordenando a média da distribuição posterior (totalmente) marginal para cada entidade (marginalizada sobre clusters ranker).

As classificações de entidades no cluster de classificação 1 favorecem fortemente o Heat (entidade 1) e o Thunder (2), com os Bulls (10) como o 3º time mais preferido. O cluster 2 do Ranker também favorece o Heat, mas difere em suas preferências para a segunda e terceira posições - aqui sendo o

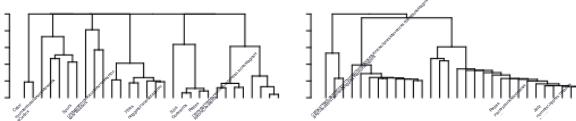


Figura 5.9: Conjunto de dados da NBA: Dendrogramas mostrando a estrutura do cluster de entidades dentro dos agrupamentos 1 e 2 (esquerda e direita, respectivamente) condicionada a dois agrupamentos de classificadores.

Classificador	Ranker cluster 1		Ranker cluster 2		Agregado		
	Entidade	Média (DP)	Entidade	Média (DP)	Entidade	Média (DP)	
11		5.63 (2.17)	1	3.18 (1.73)	1	Color	4.35 (1.38)
2	2	5.22 (2.38)	6	3.03 (1.77)	2	Trovão	2.61 (1.20)
3	10	1.48 (1.24)	4	2.23 (1.73)	6	Lakers	1.99 (0.95)
46		0.92 (0.54)	8	0.20 (0.16)	4	Celtas	1.52 (0.92)
5	9	0.86 (0.54)	10	0.20 (0.16)	10	Touros	0.81 (0.57)
64		0.75 (0.44)	9	0.19 (0.16)	9	Mavericks	0.52 (0.26)
73		0.74 (0.43)	3	0.19 (0.15)	3	Spurs	0.46 (0.22)
85		0.53 (0.36)	18	0.19 (0.15)	5	Clippers	0.28 (0.17)
9	11	0.32 (0.23)	11	0.18 (0.15)	11	Knicks	0.26 (0.13)
1012		0.20 (0.16)	20	0.18 (0.14)	8	76ers	0.13 (0.09)
11	7	0.05 (0.04)	2	0.17 (0.15)	18	Foguetes	0.12 (0.08)
12	13	0.05 (0.04)	26	0.15 (0.13)	12	Grizzlies	0.12 (0.08)
13	14	0.05 (0.04)	14	0.12 (0.12)	20	Sois	0.10 (0.08)
14	17	0.05 (0.03)	27	0.10 (0.10)	14	Magia	0.09 (0.07)
15	8	0.04 (0.03)	15	0.09 (0.10)	26	Reis	0.08 (0.07)
16	15	0.04 (0.03)	29	0.07 (0.08)	15	Hawks	0.07 (0.06)
17	18	0.03 (0.02)	23	0.07 (0.08)	13	Pepitas	0.07 (0.05)
18	19	0.02 (0.02)	13	0.07 (0.08)	7	Pacers	0.06 (0.04)
1920		0.00 (0.00)	22	0.06 (0.07)	27	Assistentes	0.05 (0.06)
20	22	0.00 (0.00)	25	0.06 (0.07)	17	TBlazers	0.05 (0.03)
21	21	0.00 (0.00)	21	0.05 (0.06)	23	Dois	0.04 (0.05)
22	23	0.00 (0.00)	7	0.05 (0.06)	29	Pilotos	0.04 (0.05)
23		0.00 (0.00)	19	0.05 (0.06)	19	Dólares	0.04 (0.04)
24	24	0.00 (0.00)	30	0.05 (0.06)	22	Guerreiros	0.04 (0.04)
2516		0.00 (0.00)	28	0.05 (0.05)	25	Pistões	0.04 (0.04)
26	26	0.00 (0.00)	16	0.04 (0.04)	21	Redes	0.03 (0.04)
27	27	0.00 (0.00)	24	0.04 (0.05)	30	Bobcats	0.03 (0.03)
28	28	0.00 (0.00)	5	0.04 (0.04)	28	Raptors	0.03 (0.03)
29	29	0.00 (0.00)	17	0.04 (0.04)	16	Jazz	0.03 (0.03)
3030		0.00 (0.00)	12	0.04 (0.04)	24	Hornets	0.03 (0.03)

MAPA: 24

MAPA: 67

Tabela 5.3: Análise da NBA: ordenações de preferência posterior dentro dos clusters de classificação 1 e 2 (condicional em dois clusters de classificação) e a classificação geral/agregada, com média (e desvio padrão) de seus parâmetros de habilidade. As linhas horizontais indicam o agrupamento de entidades MAP dentro de rankerclusters. Os números na parte inferior são o número de ocorrências em que o agrupamento MAP foi observado (de 8038 iterações com dois clusters de classificadores).

Lakers (6) e Celtics (4). Existem muitas diferenças nas ordenações de preferência entre os clusters de ranker, por exemplo, o Thunder e o Bulls aparecem nas posições 11 e 5 no cluster de ranker 2.

A análise dada em Deng et al. (2014) analisa as chamadas "entidades relevantes", definidas como aquelas entidades entre as 16 primeiras, e conclui que estas são {1, 2, ..., 15, 18}. A classificação agregada geral relatada em nosso modelo WAND fornece os 16 primeiros como {1, 2, ..., 6, 8, ..., 12, 14, 15, 18, 20, 26}; ver Tabela 5.3. Talvez surpreendentemente, apesar do

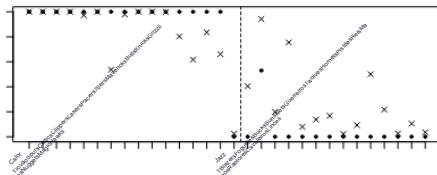


Figura 5.10: A probabilidade P16 de que cada entidade esteja entre os 16 primeiros sob o modelo WAND (\times) e as probabilidades de que cada entidade seja uma entidade relevante sob BARD ($-$). A linha vertical separa as equipes que realmente alcançaram os playoffs entre os 16 primeiros.

A análise do BARD assumindo apenas um único cluster de classificadores, há uma sobreposição considerável entre as listas dos 16 principais do WAND e do BARD - as diferenças são que as entidades 20 e 26 aparecem em nossa lista, enquanto as entidades 7 e 13 são omitidas, com as entidades 7 e 13 apenas faltando do nosso top 16 e aparecendo nas posições 18 e 17. No entanto, isso pode ser explicado pelo fato de a classificação agregada geral da WAND ser formada por um consenso entre o cluster de classificação muito discriminante 1 e o cluster de classificação muito menos discriminante 2.

Se agora compararmos o top-16 do BARD com os rankings nos clusters de ranker WAND, veremos que os rankings de entidades no cluster de ranker 1 são consistentes com os resultados do BARD, com as únicas diferenças sendo que a entidade 18 está classificada em 17º e a entidade 17 se move para o top-16. As classificações de entidades no cluster de classificação 2 são muito menos consistentes com os resultados do BARD, e isso é parcialmente explicado pela maior incerteza sobre as posições das entidades dentro desse cluster. A proximidade das médias posteriores da entidade (no cluster de classificação 2) ajuda a explicar esse nível de incerteza de classificação, pois esses classificadores claramente lutam para distinguir entre entidades.

A análise BARD também relata uma probabilidade de cada entidade ser uma entidade relevante, semelhante à probabilidade P16 de que cada entidade esteja entre as 16 primeiras no modelo WAND. Os valores para essas probabilidades nos modelos BARD e WAND são mostrados na Figura 5.10; A linha tracejada vertical separa as entidades de 1 a 16 do restante, ou seja, separa as equipes que realmente alcançaram os 16 primeiros playoffs naquela temporada das outras. É interessante ver que o modelo WAND coloca muito mais incerteza em muitas das 16 melhores equipes do que no BARD, no sentido de que seus valores de P16 são menores – os valores de BARD são essencialmente zero ou um.

5.2.1 Análise prévia de sensibilidade

Na Seção 5.2, usamos as habilidades de classificação (autodeclaradas) para formar um relativamente "informativo" antes dos pesos de classificação. Em geral, é improvável que essa informação esteja disponível para o analista e, portanto, aqui analisamos a sensibilidade da distribuição posterior a mudanças na probabilidade anterior de que um classificador seja informativo. A partir de nossa experiência usando WANDdescobrimos que, em um cenário onde há pouca informação sobre as classificabilidades, é melhor fazer escolhas conservadoras de π a priori. Consideramos uma escolha alternativa ao π "escalonado" usado anteriormente e deixamos $\pi_i = 0,5$ para todo i , de modo que cada ranker tenha a mesma probabilidade de ser informativo / não informativo e comparar o posterior em cada análise.

Geralmente, descobrimos que a distribuição a posteriori é bastante robusta para a escolha de π a priori, o que talvez não seja surpreendente, dado o que vimos nos estudos de simulação anteriores e (outras) análises de dados reais. Como antes, não foi surpresa que o aspecto da distribuição posterior mais sensível a mudanças no π fossem seus equivalentes posteriores $\Pr(w_i = 1|D)$; veja a Figura 5.11 (coluna da direita). Sem surpresa, as maiores discrepâncias são para os classificadores cujo π anterior foi mais aumentado (classificadores 19–25 e 26–34). Observamos, no entanto, que os classificadores que obtêm $\Pr(w_i = 1|D) < 0,25$ são semelhantes em ambas as análises, com apenas os classificadores 31 e 33 não mais abaixo desse limite para a análise $\pi_i = 0,5$ (observe $\Pr(w_{31} = 1|D) = 0,26$). Os dendrogramas da estrutura de agrupamento de classificadores são semelhantes para cada escolha anterior e indicam claramente que existem dois grupos de classificadores. Além disso, a alocação de classificadores para clusters é semelhante em cada caso; veja a Figura 5.11 (coluna da esquerda). Isso é ainda apoiado pela distribuição marginal posterior do número de grupos de classificadores, o que mostra que esses dados foram claramente informativos e sugerem 2 grupos de classificadores em ambas as escolhas anteriores; veja a Figura 5.12 (esquerda). Curiosamente, com a condição de haver dois clusters de ranker, o marginal posterior do número de clusters de entidades N_{es} (dentro de cada cluster de ranker $s = 1, 2$) sugere que há um pouco mais de clusters de entidades dentro do grupo ranker 2 sob a análise $\pi_i = 0,5$; veja a Figura 5.12 (à direita). Este é talvez um artefato de informações adicionais disponíveis neste cluster como $\Pr(w_i = 1|D)$ aumentou para alguns dos rankers dentro deste grupo. Os dendrogramas da estrutura de agrupamento de entidades no grupo ranker 1 são semelhantes em cada caso; veja a Figura 5.13 (coluna da esquerda). Embora os dendrogramas correspondentes do agrupamento de entidades para o grupo ranker 2 possam parecer ligeiramente diferentes para cada análise, após uma inspeção mais detalhada, fica claro que apenas a Magia e os Magos mudaram a associação do grupo da direita para o grupo central para a análise $\pi_i = 0,5$.

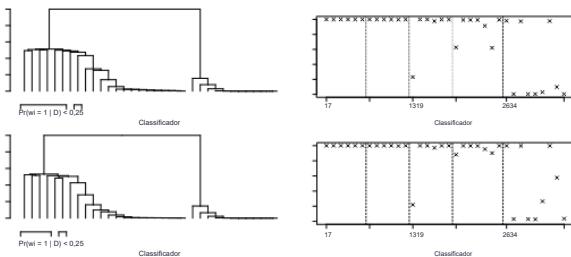


Figura 5.11: Conjunto de dados da NBA: Dendrograma (à esquerda) mostrando a estrutura de agrupamento de classificadores e destacando esses classificadores com $\text{Pr}(w_i = 1 | D) < 0,25$. Gráfico (à direita) das probabilidades posteriores $\text{Pr}(w_i = 1 | D)$ para cada ranker, com linhas verticais separando os grupos autocertificados. A linha superior mostra os resultados para a escolha "escalonada" de π e a linha inferior mostra os resultados correspondentes quando $\pi = 0,5$ para todos os classificadores.

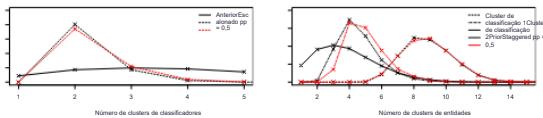


Figura 5.12: Conjunto de dados da NBA: Densidades posteriores anteriores e marginais para o número de clusters de rankers (gráfico à esquerda) e o número de clusters de entidades dentro de cada cluster de ranker, condicional a dois rankerclusters, (gráfico à direita) para escolha "escalonada" de π e $\pi = 0,5$.

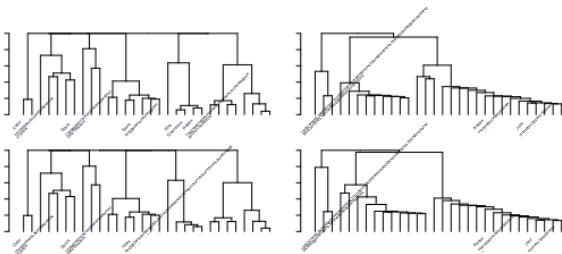


Figura 5.13: Conjunto de dados da NBA: Dendrogramas mostrando a estrutura do cluster de entidades dentro dos agrupadores 1 e 2 (esquerda e direita, respectivamente) condicionada a dois clusters de classificadores para escolha "escalonada" de π e $\pi = 0,5$, superior e inferior, respectivamente.

5.3 Resumo

Neste capítulo, ajustamos o modelo WAND a dois conjuntos de dados reais que foram previamente analisados na literatura. Em geral, verificamos que as inferências sob o modelo WAND foram semelhantes às obtidas sob outros modelos. No entanto, a riqueza das informações dentro da distribuição posterior (sob WAND) nos permite inferir informações adicionais sobre a estrutura entre os grupos de classificadores e entidades. Nossa análise dos dados da NBA também revelou fortes sinais de heterogeneidade entre as crenças dos rankers sobre as entidades. Segue-se que o modelo BARD pode não ser adequado a esses dados, dada a suposição de homogeneidade subjacente (ranker).

No próximo capítulo, consideraremos relaxar a suposição de um processo de classificação explícito, apelando para o modelo Extended Plackett-Luce (Mollica e Tardella, 2014).

Capítulo 6

O modelo Plackett-Luce estendido

6.1 Introdução

Neste capítulo, voltamos a considerar dados homogêneos classificados e consideraremos o modelo Ex-tended Plackett-Luce proposto por Mollica e Tardella (2014). Este modelo é uma extensão do modelo padrão de Plackett-Luce, que relaxa a suposição a priori de um processo de classificação explícito. Lembre-se de que o modelo padrão de Plackett-Luce (e também o modelo de Plackett-Luce ponderado) pressupõe que cada classificador forma sua classificação usando o processo de classificação direta; ver seção 2.2. Aqui, cada classificador forma sua classificação alocando primeiro sua entidade preferida, depois sua segunda entidade preferida e assim por diante até que sua entidade menos preferida seja alocada por último. Esta é uma suposição bastante forte. É fácil imaginar um cenário em que um classificador individual possa atribuir entidades a posições de maneira alternativa. Por exemplo, é bastante plausível que os classificadores possam achar mais fácil identificar suas entidades mais e menos preferidas primeiro, em vez daquelas entidades que colocam nas posições intermediárias de sua classificação. Nesse cenário, os classificadores podem formar sua classificação primeiro atribuindo suas entidades mais e menos preferidas a uma classificação antes de completar sua classificação (preenchendo as posições intermediárias) por meio de um processo de eliminação usando as entidades restantes (não alocadas), ou seja, eles usam um processo de classificação diferente. O efeito do (assumido) processo de classificação subjacente é um tanto desconhecido, com, até onde sabemos, apenas o modelo padrão (classificação direta) de Plackett-Luce e o modelo de Plackett-Luce reverso (classificação reversa) recebendo atenção significativa na literatura. O modelo de Plackett-Luce alargado permite que o processo de classificação subjacente seja mais explorado, uma vez que, em vez disso, permite todos os processos de classificação possíveis e permite que os dados nos informem qual é o mais plausível.

O restante deste capítulo é descrito a seguir. Começamos com uma discussão sobre o modelo de Plackett-Luce estendido e descrevemos o processo de geração de dados associado. Na Seção 6.2.2, consideramos a identificabilidade do processo de classificação e fornecemos algumas informações sobre onde as informações sobre o processo de classificação estão contidas nos dados. As seções restantes enfocam a inferência para o modelo de Plackett-Luce estendido e consideramos as abordagens de máxima verossimilhança e bayesiana com algoritmos de inferência eficientes apresentados em ambos os casos. Ao longo deste capítulo, realizamos vários estudos de simulação para demonstrar como inferências perspicazes podem ser obtidas usando o modelo ExtendedPlackett-Luce.

6.2 O modelo Extended Plackett-Luce

Mollica e Tardella (2014) propõem o modelo Extended Plackett-Luce (EPL) que permite que a suposição a priori de um processo de classificação implícita seja relaxada. Segue-se que, para este modelo, podemos aprender sobre o processo de classificação subjacente (possivelmente não observado) e os parâmetros das entidades. Antes de descrevermos o modelo de Plackett-Luce Estendido, é natural reformular o processo de classificação subjacente em termos de uma "ordem de escolha", onde a ordem de escolha é a ordem na qual os classificadores atribuem entidades a posições / classificações. Por exemplo, uma ordem de escolha de $(1, K, 2, 3, \dots, K - 1)$ corresponde ao processo de classificação em que um classificador primeiro atribui sua entidade preferida e , em seguida, sua entidade menor preferida antes de atribuir as entidades restantes em ordem de classificação da 2ª para baixo. Em outras palavras, os classificadores escolhem suas entidades mais e menos preferidas e, em seguida, atribuem as entidades restantes usando o processo de classificação direta. Observe que a ordem de escolha é apenas uma permutação das classificações 1 a K. O modelo EPL é definido através da introdução de um parâmetro adicional (livre) para representar a ordem de escolha dentro do modelo de Plackett-Luce. Suponha que existam K entidades e deixe σ denotar a ordem de escolha (possivelmente desconhecida), então a probabilidade de uma classificação particular sob o modelo de Plackett-Luce estendido é

$$\Pr(X_i = x_i | \lambda^*, \sigma) = \prod_{j=1}^K \frac{\lambda^{*\sigma_j} \sum K_m}{\sum_j \lambda^{*\sigma_j m}} \quad (6.1)$$

onde $\lambda^* \in R^K > 0$ são os parâmetros da entidade e $\sigma \in S^K$, o conjunto de todas as permutações possíveis das classificações 1 a K. Observe que, para manter a consistência notacional, adotamos uma representação alternativa à de Mollica e Tardella (2014): aqui $x_i = (x_{i1}, x_{i2}, \dots, x_{iK})$ representa a preferência das entidades relatadas pelo ranker i e assim, Como antes, a entidade X_{i1} é sua entidade preferida, X_{i2} é a segunda entidade mais preferida e assim por diante. A seguir, os rankings x_i são frequentemente chamados de ordenações de preferência para deixar claro que estamos considerando a preferência das entidades expressas pelo ranker i

independentemente da ordem de escolha. Além disso, sob o modelo de Plackett-Luce estendido, os parâmetros têm uma interpretação diferente daquelas sob o modelo de Plackett-Luce padrão (classificação direta) (discutido abaixo) e, portanto, deixamos λ^* ser os parâmetros das entidades para tornar essa distinção clara. Também observamos que o modelo Extended Plackett-Luce só é bem definido para classificações completas (Mollica e Tardella, 2014) e, portanto, cada ranker deve fornecer uma ordem de preferência de todas as entidades. Portanto, para o modelo ExtendedPlackett-Luce, exigimos $n = K$ para $i = 1, \dots, n$.

A forma da probabilidade estendida de Plackett-Luce é naturalmente bastante semelhante à da probabilidade padrão de Plackett-Luce. De fato, os modelos Plackett-Luce padrão e reverso são casos especiais do EPL. O modelo padrão (classificação direta) de Plackett-Luce é recuperado do modelo EPL quando σ é a permutação de identidade, ou seja, quando $\sigma(j) = j$ para $j = 1, \dots, K$. Além disso, o modelo de Plackett-Luce reverso (classificação inversa) é obtido quando $\sigma = (K, K-1, \dots, 1)$. Dado isso, talvez agora esteja claro que, embora σ seja nominalmente um parâmetro de modelo, cada $\sigma \in \text{SK}$ define um modelo de Plackett-Luce diferente.

Um aspecto fundamental da análise de dados classificados usando um modelo de Plackett-Luce é a interpretação dos parâmetros λ^* . Embora talvez não seja óbvio, para o modelo de Plackett-Luce, a interpretação dos parâmetros depende do processo de classificação subjacente. Segue-se que, para o modelo de Plackett-Luce estendido, a interpretação dos parâmetros depende do parâmetro de ordem de escolha σ . Isso fica claro se considerarmos um exemplo condicional em ordens de escolha conhecidas (fixas): suponha que temos duas entidades (rotuladas i, j) com parâmetros λ^*i e λ^*j onde $\lambda^*i > \lambda^*j$. Para o modelo padrão (classificação direta) de Plackett-Luce ($\sigma = (1, 2, \dots, K)$) a interpretação é que a entidade i é preferível à entidade j . No entanto, para o modelo de Plackett-Luce reverso (classificação inversa) ($\sigma = (K, K-1, \dots, 1)$) esses parâmetros são interpretados como a entidade j sendo preferida à entidade i . Segue-se que a ordem de preferência das entidades deve ser lida em relação à ordem de escolha. Em geral, a entidade com o maior parâmetro é a entidade com maior probabilidade de ser classificada na posição $\sigma(1)$. Além disso, condicional a uma entidade ser atribuída à classificação $\sigma(1)$, a entidade com o maior parâmetro das restantes é a que tem maior probabilidade de receber a classificação $\sigma(2)$. Embora para os processos de classificação para frente e para trás isso leve a uma interpretação natural dos parâmetros de habilidade, a interpretação para outros processos de classificação pode ser complicada. Por exemplo, suponha novamente que $\lambda^*i > \lambda^*j$, e agora considere a ordem de escolha como $\sigma = (5, 3, 2, 1, 4)$. Segue-se que, para esta ordem de escolha, a entidade i tem mais probabilidade de ser classificada em quinto lugar do que a entidade j . Além disso, se outra entidade $l \neq i, j$ recebe classificação 5, então a entidade i é preferida para a classificação 3 ($\sigma(2)$) em relação à entidade j . Essa interpretação não é exatamente intuitiva. Dado que o modelo Extended Plackett-Luce considera todos os $\sigma \in \text{SK}$, seria útil se pudéssemos conceber um método para interpretar consistentemente a preferência das entidades (por meio dos parâmetros), independentemente da ordem de escolha.

A seguir, discutimos como a probabilidade para o modelo EPL pode ser reescrita de forma que os parâmetros mantenham uma interpretação de ordem de preferência independentemente da ordem de escolha. Lembre-se de que a probabilidade de uma determinada preferência ordenar x_i para o modelo EPL é

$$\Pr(X_i = x_i | \lambda^*, \sigma) = \prod_{j=1}^K \frac{\lambda^{*x_{ij}} \sum K_m = j}{\lambda^{*x_i \sigma_m}}$$

Observe que o numerador do j -ésimo produto $\lambda^{*x_{ij}}$ é o parâmetro para a entidade na ordem de preferência o_j de ordem de preferência i . No entanto, para manter a interpretação da ordenação de preferência dos parâmetros, exigimos que o numerador do j -ésimo produto corresponda ao parâmetro para a entidade na classificação j em oposição ao da classificação o_j (como acima). Se deixarmos uma "classificação permutada" ser $x^* = x \circ \sigma$, ou seja, $x^*_{ij} = x_{i\sigma_j}$ para $j = 1, \dots, K$, então segue-se que a probabilidade de uma preferência ordenando x^* sob o modelo EPL pode ser escrita em termos da classificação permutada x^* como

$$\begin{aligned} \Pr(X_i = x_i | \lambda^*, \sigma) &= \prod_{j=1}^K \frac{\lambda^{*x_{i\sigma_j}} \sum K_m = j}{\lambda^{*x_{i\sigma_m}}} \\ &= \prod_{j=1}^K \frac{\lambda^{*x^*_{ij}} \sum K_m = j}{\lambda^{*x^*_{i\sigma_m}}}, \end{aligned} \quad (6.2)$$

e esta é simplesmente a probabilidade padrão (classificação futura) de Plackett-Luce definida sobre as classificações permutadas. De fato, já vimos um caso especial desse resultado no Capítulo 2, quando observamos que o modelo Reverse Plackett-Luce é equivalente ao modelo PL de classificação avançada aplicado a classificações que foram permutadas em ordem inversa. Note-se, no entanto, que, sob esta representação, os parâmetros λ^* ainda devem ser interpretados em relação à ordem de escolha, pois estamos analisando as classificações permutadas e, portanto, a entidade na classificação 1 das classificações permutadas (a entidade com maior probabilidade de ser escolhida primeiro) é aquela que tem preferência σ_1 . Segue-se que, embora a probabilidade tenha sido reescrita, o parâmetro skill, $\lambda^{*x^*_{ij}}$, ainda é o parâmetro para a entidade na posição o_j da ordem de preferência. A consequência disso é que, para este modelo, o maior parâmetro corresponderá à entidade que tem preferência σ_1 . Idealmente, a entidade com o maior parâmetro seria aquela que tem preferência 1 (como no modelo padrão de Plackett-Luce). Isso pode ser alcançado considerando a permutação inversa σ^{-1} que é definida de modo que

$$\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = \sigma I$$

onde σI é a permutação de identidade, ou seja, $\sigma I = (1, 2, \dots, K)$. Portanto, é claro que $\sigma^{-1}\sigma I = 1$ por definição. Agora seja $\lambda = (\lambda_1, \dots, \lambda_K)$ com $\lambda_k = \lambda^{*\sigma_k}$ para $k = 1, \dots, K$ seja uma coleção de parâmetros de habilidade então, por construção, temos $\lambda^{*k} = \lambda^{*\sigma^{-1}k}$ para $k = 1, \dots, K$.

Segue-se que $\lambda * x_{ij} = \lambda \sigma - 1 x_{ij}$ e, portanto, podemos escrever a probabilidade de uma ordem de preferência para o modelo de Plackett estendido

$$\begin{aligned}
 \Pr(X_i = x_i | \lambda, \sigma) &= \prod_{j=1}^K \frac{\lambda * x_{ij} / \sum K_m}{\lambda * x_{im}} \\
 &= \prod_{j=1}^K \frac{\lambda \sigma - 1 x_{ij} / \sum K}{\lambda \sigma - 1 x_{im}} \\
 &= \prod_{j=1}^K \frac{\lambda \sigma - 1 x_{ij} / \sum K}{m=j \lambda \sigma - 1 x_{im}}, \tag{6.3}
 \end{aligned}$$

para que o maior parâmetro de habilidade seja para a entidade que tem preferência 1, conforme desejado. Portanto, considerando os parâmetros de habilidade λ (em oposição a $\lambda*$) e usando a formulação (6.3), mantemos a interpretação da ordem de preferência dos parâmetros de habilidade, ou seja, $\lambda_i > \lambda_j$ indica que a entidade i é preferida à entidade j (independentemente da ordem de escolha). Dito isso, deixamos claro que, embora λ_k represente a preferência da entidade k, nossa intuição usual sobre as probabilidades que eles especificam sobre a classificação não se sustenta mais; λ_k não representa mais a probabilidade de que a entidade k receba classificação 1, a menos que $\sigma_1 = 1$.

6.2.1 Simulando dados do modelo Extended Plackett-Luce

O método para gerar dados do modelo de Plackett-Luce estendido é semelhante ao do modelo de Plackett-Luce padrão (classificação direta). No entanto, há uma diferença sutil, mas ainda importante. Para o modelo PL padrão, a entidade escolhida primeiro é também a mais preferida. No entanto, para o modelo de Plackett-Luce estendido, este não é mais o caso e a entidade escolhida primeiro é considerada como tendo classificação σ_1 . Lembre-se de que $x = x \circ \sigma$ denota uma classificação permutada onde x é a ordem de preferência correspondente e σ é a ordem de escolha. Agora, observando que, por construção, uma classificação permutada $x*$ é uma ordenação de entidades de acordo com a ordem em que foram escolhidas (e não a preferência das entidades), segue-se que podemos simular essas ordenações a partir do modelo PL padrão, pois elas seguem o processo de classificação direta por definição. Então, dada uma classificação permutada (simulada) $x*$ podemos obter a ordem de preferência correspondente trivialmente, lembrando que $x = x* \circ \sigma^{-1}$. Ao simular dados consistentes com o modelo de Plackett-Luce estendido, atenção adicional deve ser colocada na escolha dos parâmetros de habilidade. Embora λ_k ainda corresponda à força/habilidade da entidade k, nossa intuição usual sobre as probabilidades que eles especificam sobre a classificação não se sustenta mais. Lembre-se de que, para o modelo padrão de Plackett-Luce, a probabilidade de que a entidade k receba a classificação 1 é proporcional a λ_k . Agora, dado que o EPL

modelo é na verdade o modelo padrão de Plackett-Luce definido sobre as classificações permutadas com parâmetros λ^* segue-se que o parâmetro λ^*k é proporcional à probabilidade de que a entidade k seja classificada em primeiro lugar dentro de x^* , ou equivalenteamente, a probabilidade de que a entidade k seja classificada na posição $\sigma 1$ da ordem de preferência x é $Pr(x|\sigma = k) \propto \lambda^*k$ para $k = 1, \dots, n$, coleção K . A $X^* = \{x^i\}_{i=1}^n$ de n classificações permutadas usando permutação σ são geradas a partir do mecanismo de geração de dados de Plackett-Luce padrão (classificação direta) condicional aos parâmetros λ^* da seguinte forma.

Para $i = 1, \dots, n$,

1. Amostra $v_{ij} \sim \text{Exp}(\lambda^*ij)$ para $j = 1, \dots, K$.

2. Defina $x^*_{ij} = v_{ij}$ onde $S_{ij} = K \setminus \{x^*_{i1}, \dots, x^*_{ij-1}\}$ para $j = 1, \dots, K$.
 $\text{argmin}_q q \in S_{ij}$

As ordenações de preferência são então obtidas deixando $x_i = x^*_{i1} \circ \sigma^{-1}$ para $i = 1, \dots, n$.

Exemplo

Para fornecer clareza adicional, consideramos agora um breve exemplo que mostra como o mecanismo de geração de dados funciona na prática. Suponha que temos $K = 5$ entidades e deixe $K = \{a, b, c, d, e\}$ denotar a coleção de todas as entidades. Além disso, seja $\lambda = (\lambda_a, \lambda_b, \lambda_c, \lambda_d, \lambda_e) = (10, 8, 6, 4, 2)$ e, portanto, a entidade a é a mais preferida, a entidade b é a segunda mais preferida e em breve. Observe que, para o modelo padrão (classificação direta) de Plackett-Luce, $\sigma = (1, 2, 3, 4, 5)$, e com essa escolha de parâmetros de habilidade, a classificação ideal que maximiza a probabilidade padrão PL é $x^*_{11} = (a, b, c, d, e)$. Agora seja $\sigma = (5, 3, 2, 1, 4)$ para que os classificadores escolham primeiro sua entidade menos preferida e, em seguida, escolham sua 3^a, 2^a, 1^a e 4^a entidades mais preferidas respectivamente. Lembre-se de que $\lambda^*k = \lambda \sigma^{-1}k$ para $k = 1, \dots, K$ e assim $\lambda^* = (4, 6, 8, 2, 10)$ dado $\sigma^{-1} = (4, 3, 2, 5, 1)$. Segue-se que, no que diz respeito a λ^* , a entidade e é a entidade "mais forte" e, portanto, a mais provável de ser selecionada primeiro. Portanto, dado λ^* , a classificação permutada ótima é $x^*_{11} = (e, c, b, a, d)$ e, portanto, a ordem de preferência ótima correspondente é $x^*_{12} = x^*_{11} \circ \sigma^{-1} = (a, b, c, d, e)$, ou seja, a classificação ideal é a mesma que a do processo de classificação padrão dado λ , o que parece sensato. As classificações permutadas x^*_{11} são então geradas executando repetidamente as etapas 1 e 2 acima e as ordens de preferência correspondentes são dadas por $x_i = x^*_{i1} \circ \sigma^{-1}$.

6.2.2 Identificabilidade do processo de classificação

Nesta seção, discutimos a identificabilidade do processo de classificação. De fato, talvez não seja óbvio que os dados contenham qualquer informação para identificar o parâmetro de ordem de escolha σ . Novamente, suponha que temos $K = 5$ entidades com $K = \{a, b, c, d, e\}$ e parâmetro de habilidade

vetor $\lambda = (\lambda a, \lambda b, \lambda c, \lambda d, \lambda e)$. Se considerarmos agora uma única ordem de preferência $x1 = (1, 2, 3, 4, 5) \equiv (a, b, c, d, e)$ e duas ordenações de escolha diferentes $\sigma1 = (1, \dots, K)$ e $\sigma2 = (K, \dots, 1)$, então a probabilidade da ordem de preferência para cada ordem de escolha é

$$\Pr(x1|\lambda, \sigma1) = \frac{\lambda a \lambda a + \lambda b + \lambda c + \lambda d + \lambda e}{\lambda a \lambda a + \lambda b \lambda b + \lambda c + \lambda d + \lambda e} \frac{\lambda c \lambda c + \lambda d + \lambda e}{\lambda c \lambda c + \lambda d + \lambda e} \frac{\lambda d \lambda d + \lambda e}{\lambda d \lambda d + \lambda e}$$

$$\Pr(x1|\lambda, \sigma2) = \frac{\lambda a \lambda a + \lambda b + \lambda c + \lambda d + \lambda e}{\lambda a \lambda a + \lambda b \lambda b + \lambda c + \lambda d + \lambda e} \frac{\lambda c \lambda c + \lambda d + \lambda e}{\lambda c \lambda c + \lambda d + \lambda e} \frac{\lambda d \lambda d + \lambda e}{\lambda d \lambda d + \lambda e}$$

Portanto, neste cenário, claramente não há informações na ordem de preferência única $x1$ com a qual identificar a ordem de escolha, pois a probabilidade de $x1$ é a mesma para todos $\lambda \in \mathbb{R}^K > 0$ em ambas as ordenações de escolha. No entanto, consideremos a probabilidade de várias ordenações de preferência. Dado que as ordenações de preferência são condicionalmente independentes, a probabilidade sob o modelo de Plackett-Luce estendido é

$$\begin{aligned} \pi(D|\lambda, \sigma) &= \prod_{i=1}^n \Pr(x_i|\lambda, \sigma) \\ &= \prod_{i=1}^n \prod_{j=1}^{K-i} \frac{\lambda_{\sigma(i)-1} \lambda_{\sigma(j)}}{\lambda_{\sigma(i)-1} \lambda_{\sigma(j)}} \quad (6.4) \end{aligned}$$

onde D denota uma coleção de ordenações preferenciais $\{x_i\}_{i=1}^n$. Agora suponha que $n = 2$ e a ordem de preferência adicional seja $x2 = (3, 2, 1, 4, 5) \equiv (c, b, a, d, e)$ então de (6.4) segue-se que a probabilidade de $D = \{x1, x2\}$ sob a ordem de escolha $\sigma1$ é

$$\begin{aligned} \pi(D|\lambda, \sigma1) &= \frac{\lambda a \lambda a + \lambda b + \lambda c + \lambda d + \lambda e}{\lambda a \lambda a + \lambda b \lambda b + \lambda c + \lambda d + \lambda e} \frac{\lambda c \lambda c + \lambda d + \lambda e}{\lambda c \lambda c + \lambda d + \lambda e} \frac{\lambda d \lambda d + \lambda e}{\lambda d \lambda d + \lambda e} \\ &\times \frac{\lambda c \lambda c + \lambda b + \lambda a + \lambda d + \lambda e}{\lambda c \lambda c + \lambda b \lambda b + \lambda a + \lambda d + \lambda e} \frac{\lambda a \lambda a + \lambda b + \lambda d + \lambda e}{\lambda a \lambda a + \lambda b \lambda b + \lambda d + \lambda e} \end{aligned}$$

e sob a ordem de escolha $\sigma2$ é

$$\begin{aligned} \pi(D|\lambda, \sigma2) &= \frac{\lambda a \lambda a + \lambda b + \lambda c + \lambda d + \lambda e}{\lambda a \lambda a + \lambda b \lambda b + \lambda c + \lambda d + \lambda e} \frac{\lambda b \lambda b + \lambda e + \lambda d + \lambda c}{\lambda b \lambda b + \lambda e + \lambda d + \lambda c} \frac{\lambda c \lambda c + \lambda d + \lambda e}{\lambda c \lambda c + \lambda d + \lambda e} \frac{\lambda d \lambda d + \lambda e}{\lambda d \lambda d + \lambda e} \\ &\times \frac{\lambda a \lambda a + \lambda b + \lambda e + \lambda d + \lambda c}{\lambda a \lambda a + \lambda b \lambda b + \lambda e + \lambda d + \lambda c} \frac{\lambda b \lambda b + \lambda e + \lambda d + \lambda c}{\lambda b \lambda b + \lambda e + \lambda d + \lambda c} \frac{\lambda e \lambda e + \lambda d + \lambda c}{\lambda e \lambda e + \lambda d + \lambda c} \frac{\lambda d \lambda d + \lambda c}{\lambda d \lambda d + \lambda c} \end{aligned}$$

É claro que as duas probabilidades só são iguais quando $\lambda a = \lambda c = \lambda e$ e assim

$$\pi(D|\lambda, \sigma1) = \pi(D|\lambda, \sigma2) \iff \lambda a = \lambda c = \lambda e.$$

À medida que mais ordenações de preferência (únicas) são introduzidas na probabilidade, fica claro que restrições adicionais nos parâmetros de habilidade serão necessárias para que a probabilidade

Classificar um	Entidade de			
	a.C			
1	0.33	0.27	0.20	0.13
2	0.28	0.26	0.22	0.16
3	0.21	0.23	0.24	0.20
4	0.13	0.17	0.22	0.28
5	0.04	0.07	0.12	0.23

Tabela 6.1: Probabilidades de que cada entidade seja atribuída a uma classificação específica para o modelo padrão de Plackett-Luce com $\lambda = (5, 4, 3, 2, 1)$.

para ser o mesmo sob diferentes ordens de escolha. De fato, para n suficientemente grande, temos isso

$$\pi_i(D|\lambda, \sigma) = \pi_j(D|\lambda, \sigma) \iff \lambda_k = \lambda$$

para $i, j = 1, \dots, K$. Mollica e Tardella (2014) fornecem uma prova para esse efeito que mostra que cada ordem de escolha $\sigma \in S_K$ define uma distribuição única sobre as classificações. Segue-se que a ordem de escolha é identificável dado um número razoavelmente grande de classificações n . Infelizmente, é difícil fornecer uma justificação teórica para um valor suficiente de n , uma vez que isso dependerá não só do número de entidades K , mas também das posições das entidades dentro das ordens de preferência. Supomos, no entanto, que no cenário em que $n \geq K$ a ordem de escolha provavelmente será identificável.

Agora aproveitamos esta oportunidade para fornecer algumas dicas sobre onde as informações sobre a ordem de escolha estão contidas nos dados (ordens de preferência). Acontece que é a variação dentro das posições das ordens de preferência que nos permite determinar a ordem de escolha. Para ver isso, é útil primeiro considerar as probabilidades de cada entidade ser atribuída a uma classificação específica para o modelo padrão (classificação direta) de Plackett-Luce. Suponha que temos $K = 5$ entidades com parâmetros de habilidade $\lambda = (5, 4, 3, 2, 1)$. Mais uma vez, denotamos o conjunto completo de entidades como $K = \{a, b, c, d, e\}$ e, portanto, as entidades a e e são as entidades mais e menos preferidas, respectivamente. Podemos descrever a variação dentro das ordens de preferência calculando as probabilidades de que cada entidade seja atribuída a uma classificação específica; ver Tabela 6.1. Observe que a Tabela 6.1 não é simétrica e, portanto, a probabilidade de a entidade mais preferida ser classificada em último lugar não é a mesma que a probabilidade de a entidade menos preferida ser classificada em primeiro lugar. Curiosamente, também vemos que a entidade menos preferida (e) é classificada por último na maioria das vezes (probabilidade de 0,54). Em contraste, a entidade mais preferida (a) é classificada em primeiro lugar apenas um terço das vezes. Isso sugere que há muita incerteza sobre quais entidades são atribuídas às fileiras no início da ordenação de preferência em comparação com as últimas fileiras com essa escolha de λ . Em outras palavras, as entidades mais fracas são frequentemente claramente identificadas e aparecem perto da parte inferior das ordens de preferência, enquanto

Classificador	um	Entidade			Entidade		
		a.C	de	a.C	de		
≤ 1	0.33	0.27	0.20	0.13	0.07	≥ 1	1.00
≤ 2	0.61	0.53	0.42	0.29	0.15	≥ 2	0.66
≤ 3	0.82	0.76	0.46	0.49	0.27	≥ 3	0.38
≤ 4	0.95	0.93	0.88	0.77	0.47	≥ 4	0.17
≤ 5	1.00	1.00	1.00	1.00	1.00	≥ 5	0.04

Tabela 6.2: Probabilidades cumulativas de cada entidade ser classificada não inferior a (esquerda) e não superior a (direita) ou igual a cada posição. A entrada i, j corresponde a $\Pr(j \in x:i)$ (esquerda) e $\Pr(j \in x:i:K)$ (direita).

as entidades mais fortes são mais suscetíveis a parecer mais baixas do que talvez devesssem nas ordenações de preferência. O que estamos realmente vendo aqui é um artefato do processo de classificação avançada. Lembre-se de que o processo de classificação direta determina que um classificador primeiro atribui uma entidade à classificação 1 e, portanto, escolha sua entidade preferida do conjunto completo K , ou seja, das K entidades que estão disponíveis para seleção quando fazem essa escolha. O classificador então escolhe uma entidade para a classificação 2, onde essa escolha é condicionada ao fato de a entidade que ele colocou na classificação 1 não estar mais disponível para seleção e, portanto, nesta fase do processo de classificação, existem apenas $K - 1$ entidades possíveis para escolher. É claro que em cada etapa do processo de classificação há uma entidade a menos para escolher do que na etapa anterior. Segue-se que, quando um classificador está alocando suas entidades menos preferidas, a maioria das outras entidades já terá sido alocada e, portanto, há apenas algumas para escolher e isso resulta na variação reduzida de quais entidades são atribuídas a essas classificações. O nível de variação dentro de cada classificação é ainda destacado na Tabela 6.2, que mostra as probabilidades cumulativas de cada entidade ser classificada não inferior a (esquerda) e não superior a (direita) cada classificação.

Agora examinamos as probabilidades de que as entidades recebam uma classificação específica para o modelo ExtendedPlackett-Luce. Suponha que a ordem de escolha seja $\sigma = (5, 3, 2, 1, 4)$ e, como antes, deixe os parâmetros de habilidade serem $\lambda = (5, 4, 3, 2, 1)$, ou seja, mantermos a preferência das entidades como antes, com a entidade a sendo a mais preferida e a entidade e a menos preferida. A Tabela 6.3 (canto superior esquerdo) mostra as probabilidades de que cada entidade seja atribuída a uma classificação específica (dentro da ordem de preferência). Ao comparar essas probabilidades com as do modelo padrão (forward ranking) de Plackett-Luce na Tabela 6.1, fica claro que a distribuição (de entidades) nas fileiras não é a mesma para o modelo Extended Plackett-Luce, ou seja, o modelo EPL define uma distribuição de classificação diferente. No entanto, como veremos agora, essas distribuições estão inherentemente relacionadas. Lembre-se de que, ao delinear o mecanismo de geração de dados para o modelo EPL, observamos que os parâmetros de habilidade λ não são mais proporcionais à probabilidade de uma entidade ser classificada em primeiro lugar na ordem de preferência e, em vez disso, o

		Entidade de					Entidade de				
		Classificar um a.C			Classificar um a.C		Classificar um a.C			Classificar um a.C	
		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
1	0.28	0.22	0.17	0.20	0.13		0.13	0.20	0.27	0.07	0.33
2	0.20	0.24	0.23	0.12	0.21		0.16	0.22	0.26	0.08	0.28
3	0.16	0.22	0.26	0.08	0.28		0.20	0.24	0.23	0.12	0.21
4	0.23	0.12	0.07	0.54	0.04		0.28	0.22	0.17	0.20	0.13
5	0.13	0.20	0.27	0.07	0.33		0.23	0.12	0.07	0.54	0.04

		Entidade					Entidade				
		Classificar Ce b			Para		Classificar Ce b			Para	
		S1	S2	S3	S4	S5	S1	S2	S3	S4	S5
		0.33	0.27	0.20	0.13	0.07					
		0.28	0.26	0.22	0.16	0.08					
		0.21	0.23	0.24	0.20	0.12					
		0.13	0.17	0.22	0.28	0.20					
		0.04	0.07	0.12	0.23	0.54					

Tabela 6.3: Probabilidades de cada entidade ser classificada em cada posição. A entrada i, j corresponde a $\Pr(x_i = j)$, ou seja, a entidade de probabilidade j recebe a classificação i .

Os parâmetros "originais" $\lambda \cdot k$ são proporcionais à probabilidade de que a entidade k seja classificada na posição $\sigma 1$. Diante disso, parece sensato considerar as probabilidades de que cada entidade receba a classificação σi . A Tabela 6.3 (canto superior direito) mostra essas probabilidades e notamos que elas são obtidas simplesmente aplicando a σ de permutação às linhas da tabela na Tabela 6.3 (canto superior esquerdo). Observe que essas probabilidades são, na verdade, as probabilidades de que cada entidade seja classificada em primeiro lugar nas classificações permutadas x^* . Agora, lembrando que a ordem de escolha é $\sigma = (5, 3, 2, 1, 4)$, segue-se que a entidade com maior probabilidade de ser a primeira em x^* é aquela que é a quinta ($\sigma 1^a$) entidade preferida. Se, portanto, permutarmos as colunas da tabela de modo que, em relação a x^* , as entidades sejam as mais para menos preferidas da esquerda para a direita, então as probabilidades para o modelo de Plackett-Luce estendido tornam-se as mesmas que as do modelo de Plackett-Luce padrão (classificação direta); ver Tabela 6.3 (abaixo) e Tabela 6.1. σ Na seção seguinte, examinamos o quanto informativa é a função de verossimilhança sobre a ordem de escolha σ antes de considerarmos a análise bayesiana na Seção 6.4.

6.3 Informações de probabilidade sobre a ordem de escolha

Antes de considerarmos uma análise totalmente bayesiana do modelo de Plackett-Luce estendido, examinamos o quanto informativa é a função de verossimilhança sobre a ordem de escolha σ . Nesta seção, verificamos que a ordem de escolha não é apenas identificável, mas a função de verossimilhança também pode ser bastante informativa. Examinamos essa questão observando a probabilidade maximizada

função em função da ordem de escolha σ . A probabilidade é maximizada na estimativa de máxima verossimilhança $\hat{\lambda}(\sigma)$ e essa estimativa dependerá da ordem de escolha.

Lembre-se da Seção 2.2.2 que a probabilidade padrão de Plackett-Luce é invariante à multiplicação escalar dos parâmetros de habilidade. Além disso, vimos na Seção 6.2 que o modelo ExtendedPlackett-Luce é simplesmente o modelo padrão de Plackett-Luce avaliado sobre as classificações por silenciamento $x^* = x \circ \sigma$ (com parâmetros λ^*). Assim, a probabilidade EPL também é invariante à multiplicação escalar dos parâmetros λ^* (e, portanto, também dos parâmetros de habilidade λ), ou seja, $\Pr(D | \lambda, \sigma) = \Pr(D | C\lambda, \sigma)$ para qualquer $C \in R^{>0}$ e, portanto, $\hat{\lambda}$ não pode ser identificado. No entanto, esse problema é resolvido restringindo os parâmetros de habilidade para que $\sum_k \lambda_k = 1$, ou seja, colocando o vetor do parâmetro de habilidade λ no simplex ($K - 1$) dimensional. Agora examinamos as informações na função de verossimilhança observando a verossimilhança logarítmica máxima para uma determinada ordem de escolha σ . Hunter (2004) propôs um algoritmo de Minorização/Maximização (MM) que permite obter a estimativa de máxima verossimilhança dos parâmetros do modelo padrão de Plackett-Luce. Aqui, ao notar novamente que o modelo de Plackett-Luce estendido é simplesmente o modelo padrão de Plackett-Luce, dadas as classificações permutadas $x^* = x \circ \sigma$ e com os parâmetros λ^* , segue-se que podemos obter o MLE $\hat{\lambda}^*$ dada uma ordem de escolha fixa σ aplicando o algoritmo MM padrão às ordenações de preferência permutadas. Observe que, se desejado, o MLE $\hat{\lambda}$ dos parâmetros de habilidade é fácil de obter com $\hat{\lambda}_k = \hat{\lambda}^* \circ \sigma_k$ para $k = 1, \dots, K$. Segue-se que, para qualquer $\sigma \in S_K$, podemos deixar $D = \{x^*_i\}_{i=1}^n$ ser a coleção de classificações permutadas e usar o algoritmo MM para obter $\hat{\lambda}^*$ de modo que a probabilidade padrão de Plackett-Luce

$$\pi_{PL}(D|\lambda^*) = \prod_{i=1}^n \prod_{j=1}^K \frac{\lambda^*_{i \circ j} \sum K_m}{\sum_{l=j}^K \lambda^*_{i \circ l}} \quad (6.5)$$

é maximizado. Agora descrevemos o algoritmo MM que nos permite obter o MLE $\hat{\lambda}^*$ dos parâmetros dada uma ordem de escolha fixa.

6.3.1 Algoritmo MM

A coleção de parâmetros de habilidade $\hat{\lambda}$ que maximizam a probabilidade de Plackett-Luce estendida para uma determinada ordem de escolha σ pode ser obtida a partir dos parâmetros $\hat{\lambda}^*$ que maximizam o modelo padrão (classificação direta) de Plackett-Luce dada uma coleção de classificações permutadas $X^* = \{x_i \circ \sigma\}_{i=1}^n$. Especificamente $\hat{\lambda}_k = \hat{\lambda}^* \circ \sigma_k$ para $k = 1, \dots, K$ onde $\hat{\lambda}^*$ é obtido da seguinte forma.

1. Inicializar: Seja $t = 0$ e $\lambda^*(0) = (\lambda^*(0)_1, \lambda^*(0)_2, \dots, \lambda^*(0)_K)$

$$\frac{2. \text{ Letra}}{\lambda^*(t)k} = \frac{w_k}{\sum_{i=1}^n \sum_{j=1}^K \delta_{ij}(k)} \rightarrow \text{para } k = 1, \dots, K.$$

3. Redimensionar:

- calcular $\Sigma(t) = \sum_{k=1}^K \lambda^*(t)k$.
- seja $\lambda^*(t)k / \Sigma(t)$ para $k = 1, \dots, K$.

4. Defina $t = t + 1$ e retorne à Etapa 2.

Aqui, a quantidade $w_k = \sum_{i=1}^n \sum_{j=1}^K I(x_{ij} = k)$ é o número de vezes que o parâmetro λ^*k representa uma entidade na coleção de classificações permutadas e $\delta_{ij}(k) = I(k$

$\in \{x_{1j}, \dots, x_{nj}\}$) é uma variável indicadora sobre o evento em que a entidade k não aparece acima da posição j na classificação permutada i . Observe que $w_k = n$ dado que consideramos apenas classificações completas para o modelo de Plackett estendido

O número de iterações a serem executadas é uma questão importante a ser considerada ao implementar o algoritmo MM acima. Hunter (2004) mostra que no limite como $t \rightarrow \infty$ o (verdadeiro) MLE λ^* é obtido quase certamente. É claro que, na prática, devemos considerar apenas um número finito de iterações para que $\lambda^*(t)$ seja uma aproximação razoável do (verdadeiro) MLE λ^* . Escolher um número finito de iterações para executar a priori é difícil, pois não há garantia de que o algoritmo terá收敛ido neste ponto. Para evitar fazer essa escolha, é sensato implementar uma regra de parada. Uma regra de parada é uma instrução lógica (matemática) que é avaliada a cada iteração e o algoritmo prossegue até que essa instrução lógica seja satisfeita. Embora haja uma riqueza de regras de parada possíveis que poderíamos considerar, para o propósito desta tese, implementamos a regra de parada direta usada por Hunter (2004) e tomamos $\lambda^*(t)$ como o MLE quando a norma L2 da mudança no valor do vetor parâmetro é menor que 10^{-9} , ou seja, quando

$$\|\lambda^*(t) - \lambda^*(t-1)\|_2 = \sqrt{\sum_{k=1}^K (\lambda^*(t)_k - \lambda^*(t-1)_k)^2} < 10^{-9} \quad (6.6)$$

está satisfeita.

6.3.2 Estudo de simulação

Neste estudo, analisamos um único conjunto de dados com $n = 5000$ classificações completas (preferências) de $K = 5$ entidades. Consideramos um grande número de ordenações de preferência para que o MLE λ^* para cada um dos $K! = 120$ diferentes ordens de escolha possíveis são obtidas após relativamente

poucas iterações do algoritmo MM. As ordenações de preferência foram simuladas a partir do mecanismo de geração de dados descrito na Seção 6.2.1 com os parâmetros de habilidade $\lambda = (5, 4, 3, 2, 1)$ e ordem de escolha $\sigma = (5, 3, 2, 1, 4)$. O foco principal deste estudo é verificar se a ordem de escolha é identificável e investigar valores maximizados da probabilidade logarítmica para diferentes ordens de escolha. Aqui, os valores reais dos MLEs ' λ^* ' (sob as diferentes ordens de escolha) não são uma preocupação primária, mas, no entanto, esses parâmetros devem ser interpretados em relação à ordem de escolha σ pois estamos trabalhando com o modelo padrão (classificação direta) de Plackett-Luce com classificações permutadas. Lembre-se da Seção 6.3 que podemos usar o algoritmo MM para obter o MLE ' λ^* ' que maximiza $\pi_{PL}(D = \{x_{ni}\}_{ni=1}^N | \lambda^*)$ para qualquer permutação σj , e assim obter os MLEs ' λ^* ' dado σj para todas as permutações possíveis $j = 1, \dots, K!$. Cada algoritmo MM é inicializado com $\lambda^*k = \lambda^* = 1/K$ e prossegue até que a regra de parada (6.6) seja satisfeita. A Figura 6.1 mostra $\log \pi(D | \lambda^*, \sigma j)$, ou seja, o log-verossimilhança do modelo Extended Plackett-Luce avaliado no MLE dos parâmetros de habilidade ' λ^* ' para cada ordem de escolha σj . Claramente, a probabilidade logarítmica não é constante e, de fato, a ordem de escolha assumida pode ter um grande efeito na probabilidade logarítmica geral. O quadro 6.4 mostra (um subconjunto de) a classificação das ordens de escolha (permutações) com base no valor do log-verossimilhança no MLE correspondente para os parâmetros de competência. A distância de Kendall-tau entre cada ordem de escolha e a permutação verdadeira $\sigma = (5, 3, 2, 1, 4)$ também é fornecida. Curiosamente, cada uma das 5 principais ordens de escolha tem $\sigma 4 = 1$ e $\sigma 5 = 4$ e, portanto, identificamos claramente que a entidade mais preferida é escolhida em 4º e a quarta entidade preferida é escolhida por último. Além disso, talvez surpreendentemente, a ordem de escolha inversa ($\sigma = (4, 1, 2, 3, 5)$) para aquela a partir da qual esses dados foram simulados resulta na nona maior probabilidade logarítmica. Por definição, essa permutação é a distância mais distante (Kendall-tau) da ordem de escolha "verdadeira" e, portanto, é claro que os modos locais podem ser separados por grandes distâncias dentro do espaço de permutação. Esta é uma observação fundamental e

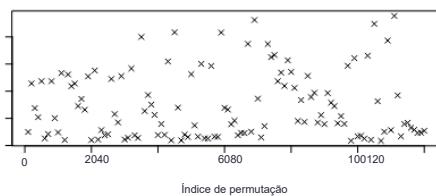


Figura 6.1: $\pi(D | \lambda^*, \sigma j)$: log-verossimilhança maximizada dada cada ordem de escolha σj e o respectivo MLE ' λ^* ' para $j = 1, \dots, K!$.

Classificação	Log-verossimilhança	Ordem de escolha	Distância Kendall-tau
1-21613.95(5,3,2,1,4)02-21689.21(3,5,2,1,4)13-			
21780.05(5,2,3,1,4)14-21914.95(2,5,3,1,4)25-			
21920.27(3,2,5,1,4)2			
...	
9	-22131.35	(4,1,2,3,5)	10
...	
84	-23750.71	(1,2,3,4,5)	7
...	
116	-23900.67	(1,5,2,4,3)	5
117	-23904.19	(2,5,1,4,3)	4
118	-23905.65	(2,5,4,1,3)	5
119	-23912.46	(5,1,2,4,3)	4
120	-23914.28	(5,2,4,1,3)	4

Tabela 6.4: Um subconjunto da classificação das ordens de escolha (permutações) com base no valor da verossimilhança logarítmica avaliada no MLE correspondente para os parâmetros de habilidade ($\pi(D|\lambda_j, \sigma_j)$).

desempenhará um papel importante quando considerarmos um esquema de inferência bayesiana para o modelo EPL na seção seguinte.

Notamos de passagem que a ordem de escolha $\sigma = (5, 3, 2, 1, 4)$ a partir da qual esses dados foram simulados é a que fornece a maior probabilidade logarítmica maximizada, ou seja, é também o MLE para σ para esses dados. No entanto, em geral, é problemático determinar o MLE para σ , pois isso requer uma pesquisa em todas as ordens de escolha possíveis ($K!$), e isso pode ser proibitivo quando K não é pequeno.

6.4 Inferência – uma abordagem bayesiana

Nesta seção, consideramos uma abordagem bayesiana para inferência em oposição à abordagem de máxima verossimilhança adotada anteriormente. Adotando uma abordagem bayesiana, podemos obter a distribuição posterior $\pi(\sigma|D)$ e, como consequência, quantificar a incerteza sobre o parâmetro de ordem de escolha de maneira baseada em princípios. Até onde sabemos, a única solução bayesiana atual é fornecida por Mollica e Tardella (2018), mas isso depende de um espaço amostral restrito para σ . Aqui pretendemos desenvolver métodos MCMC capazes de explorar todo o espaço amostral. Isso não é simples, dado que $\sigma \in SK$ e, portanto, atenção especial deve ser dada a como podemos construir um esquema de amostragem de MonteCarlo em cadeia de Markov capaz de explorar efetivamente esse espaço e direcionar a distribuição posterior correta.

6.4.1 Especificação prévia e variáveis latentes

Antes de realizarmos a inferência bayesiana, devemos primeiro escolher uma especificação prévia adequada para o nosso modelo. Consideraremos a mesma especificação prévia para os parâmetros de habilidade que quando consideramos o modelo padrão de Plackett-Luce no Capítulo 2, ou seja, deixamos λ_k indep~ $\text{Ga}(\alpha_k, 1)$ a priori (lembre-se de que o parâmetro de taxa não é identificável por verossimilhança e, portanto, escolhido como 1). Segue-se que a distribuição a priori sobre λ é

$$p(\lambda) = \prod_{k=1}^K \frac{\lambda_k^{\alpha_k - 1} e^{-\lambda_k}}{\lambda_k \Gamma(\alpha_k)}.$$

Observe que aqui somos livres para escolher um parâmetro de forma exclusivo α_k para cada entidade, pois não consideramos o agrupamento de entidades (e, portanto, não há restrições de permutabilidade). No entanto, para o modelo de Plackett-Luce estendido, atenção adicional deve ser dada à escolha de λ_k , pois a distribuição de classificação não é especificada apenas por λ , mas também pela ordem de escolha. Lembre-se da Seção 6.2.1 que a probabilidade de a entidade k receber a classificação σ_1 na ordenação de preferência é proporcional a $\lambda_k^{\alpha_k}$, ou seja, $\Pr(x|1 = k) \propto \lambda_k^{\alpha_k}$ onde $\lambda_k = \lambda \sigma_1^{-1} k$ para $k = 1, \dots, K$. Segue-se que especificar um informativo anterior para os parâmetros de habilidade é complicado, a menos que a ordem de escolha σ seja fixa e, portanto, na prática, é frequentemente útil deixar $\lambda_k = \lambda$ para que todas as ordens de preferência sejam igualmente prováveis a priori (independentemente da ordem de escolha σ).

Devemos também escolher uma distribuição prévia adequada sobre as possíveis ordenações de escolha. Cada ordenação de escolha σ é um elemento de S^K e, portanto, há $K!$ possíveis ordenações de escolhadas K entidades. Segue-se que a distribuição a priori sobre as ordenações de escolha é uma distribuição discreta com $K!$ valores possíveis. Assumimos que cada ordem de escolha (permutação) é igualmente provável a priori e, portanto,

$$\Pr(\sigma = \sigma_i) = 1/K!$$

para $i = 1, \dots, K!$. Observe que, se desejado, é possível considerar um subconjunto de todas as ordenações de escolha possíveis fazendo uma escolha apropriada de probabilidades anteriores.

Lembre-se da Seção 6.2.2 que a probabilidade sob o modelo de Plackett-Luce estendido é

$$\pi(D|\lambda, \sigma) = \prod_{i=1}^n \prod_{j=1}^{K_i} \frac{\lambda_{\sigma(i)-1}^{x_{ij}}}{\lambda_{\sigma(i)-1}^{x_{ij}} \sum_{m=j}^K \lambda_{\sigma(i)-1}^{x_{im}}}.$$

e, portanto, é de forma semelhante à do modelo padrão de Plackett-Luce. Segue-se do que vimos anteriormente que a forma da verossimilhança não admite inferência bayesiana conjugada. A implementação de um amostrador de Gibbs para manter a eficiência computacional sem a necessidade de vários parâmetros de ajuste é, no entanto, altamente desejável.

Para facilitar isso, apelamos para a mesma técnica dos modelos anteriores, ou seja, usoamento de dados. Uma solução de amostragem de Gibbs pode ser obtida usando uma generalização direta das variáveis latentes em Caron e Doucet (2012), a saber:

$$z_{ij|D, \lambda, \sigma} \text{ indep} \sim \frac{\sum_{m=j}^K L_{S-1}}{\sum_{m=1}^K L_{I \Sigma M}}$$

pois $i = 1, \dots, n$, $j = 1, \dots, K$.

Usando essas variáveis latentes, a probabilidade completa dos dados é

$$\pi(D, Z|\lambda, \sigma) = \pi(D|\lambda, \sigma)\pi(Z|D, I, S) \\ = \prod_{i=1}^n \prod_{j=1}^K \frac{\exp^{-\lambda_{S-1}x_{ij}}}{\sum_{m=j}^K \lambda_{I \Sigma M}} \quad (6.7)$$

Além disso, dada a nossa escolha de distribuição anterior, a densidade de todas as quantidades estocásticas é

$$p(\lambda, D, Z, \sigma) = \pi(D|\lambda, \sigma)\pi(Z|D, I, S)p(I)p(S) \\ = \prod_{i=1}^n \prod_{j=1}^K \frac{\exp^{-\lambda_{S-1}x_{ij}}}{\sum_{m=j}^K \lambda_{I \Sigma M}} \times \prod_{k=1}^K \frac{\lambda_{ak-1}e^{-\lambda_k} \times 1K!}{\lambda_k \Gamma(\lambda_k)} \quad (6.8)$$

6.4.2 Distribuições condicionais completas para λ, Z

A partir de (6.8) podemos obter as distribuições condicionais completas dos parâmetros de habilidade λ e as variáveis latentes Z como

- λ : Para $k = 1, \dots, K$,

$$\lambda_k | \dots \sim \frac{\delta_{ij}(k)z_{ij}}{1 + \sum_{i=1}^n \sum_{j=1}^K \delta_{ij}(k)z_{ij}}$$

onde

$$\beta_k = \sum_{i=1}^n \sum_{j=1}^K I(\sigma_{-1}x_{ij} = k) \text{ e } \delta_{ij}(k) = \sum_{m=j}^K I(\sigma_{-1}x_{ij} = k)$$

são o número de vezes que λ_k representa uma entidade dentro dos dados e uma variável indicadora sobre o evento em que λ_k representa uma entidade que aparece em uma posição não superior a j na classificação permutada i , respectivamente. Observe que $\beta_k = n$ (para $k = 1, \dots, K$), dado que consideramos apenas as classificações completas para o modelo Extended Plackett-Luce.

- Z: Para $i = 1, \dots, n, j = 1, \dots, K$,

$$z_{ij} | \cdot \sim \text{Exp}_{\text{indep} \sim} \sum_{m=j}^K \frac{\lambda_m}{\sum_{m=1}^K \lambda_m}.$$

6.4.3 Distribuição condicional completa para σ

A distribuição condicional completa para a ordem de escolha σ também é direta e é a distribuição discreta com probabilidades

$$\Pr(\sigma = \sigma| \cdot \cdot \cdot) \propto \pi(D, Z|\lambda, \sigma = \sigma) \Pr(\sigma = \sigma)$$

para $i = 1, \dots, K$. Claramente, a amostragem desta condicional completa exigirá $K!$ avaliações completas de probabilidade de dados e, portanto, uma atualização de Gibbs para σ provavelmente só é plausível se K for suficientemente pequeno; talvez não muito maior que 5. É claro que as probabilidades $\Pr(\sigma = \sigma| \cdot \cdot \cdot)$ e $\Pr(\sigma = \sigma| \cdot \cdot \cdot)$ são condicionalmente independentes para $i = j$ e, portanto, podem ser computadas em paralelo, o que pode facilitar essa abordagem para valores ligeiramente maiores de K . A carga computacional também aumentará com n devido à avaliação de $\pi(D, Z|\lambda, \sigma = \sigma)$. No entanto, a probabilidade total dos dados também pode ser calculada em paralelo (para cada ordem de escolha) como consequência de as ordenações de preferência serem condicionalmente independentes. Embora a computação paralela possa ser benéfica, essa abordagem provavelmente permanece inviável até mesmo para um número modesto de entidades e, portanto, na próxima seção, consideraremos uma abordagem alternativa capaz de gerar realizações posteriores quando a amostragem de Gibbs é inviável.

6.4.4 Propostas de Metropolis-Hastings para σ

Suponha que estejamos em um cenário em que a execução de uma atualização de Gibbs para σ não seja computacionalmente viável. Claro, ainda desejamos obter a distribuição posterior $\pi(\sigma, \lambda, Z|D)$ e, em vez disso, consideraremos uma atualização de Metropolis-Hastings para σ que nos permitirá direcionar a distribuição posterior (além das condicionais completas para λ e Z da Seção 6.4.2). Dado $\sigma \in SK$, segue-se que devemos construir um mecanismo de proposta adequado que permita que a cadeia de Markov explore eficientemente o espaço (discreto) de todos os $K!$ possíveis permutações. A partir de nossa investigação sobre a probabilidade do modelo de Plackett-Luce estendido dadas diferentes ordens de escolha na Seção 6.3.2, fica claro que $\pi(D|\lambda, \sigma)$ é multimodal. Além disso, localmodes podem estar a grandes distâncias um do outro dentro do espaço de permutação. Nosso mecanismo de proposta deve, portanto, ser capaz de fazer grandes saltos dentro do espaço de permutação, pois apenas propor pequenos movimentos (em termos de distância) pode resultar em nossa cadeia ficar presa em um modo local e não explorar todo o espaço. Com isso em mente, nossa distribuição de propostas $q(\sigma'|\sigma)$ compreenderá 5 mecanismos alternativos de propostas que ocorrem com

probabilidades $p_{prop} = (p_{prop1}, \dots, p_{prop5})$, que devem ser especificados a priori. Mais formalmente Nossa distribuição proposta é a distribuição de mistura de 5 componentes

$$q(s|\sigma) = \sum_{i=1}^5 p_{prop} q_i(s|\sigma) \quad (6.9)$$

onde $q_i(\sigma|s)$ denota o i -ésimo componente. Agora descrevemos cada componente da distribuição da mistura com mais detalhes.

Proposta 1 – a troca aleatória

Para nosso primeiro mecanismo de proposta, consideraremos uma "troca aleatória". Amostramos (uniformemente) aleatoriamente e com substituição duas posições $p_1, p_2 \in \{1, \dots, K\}$ e deixamos a ordem de escolha proposta σ ser a ordem de escolha atual σ onde os elementos nas posições p_1 e p_2 foram trocados. Mais formalmente, amostramos duas posições p_1 e p_2 a partir das distribuições discretas

$$\begin{aligned} \Pr(p_1 = p_1) &= \frac{1}{K}, & \text{duran } 1 \leq p_1 \leq K \\ \Pr(p_2 = p_2) &= \frac{1}{K}, & \text{duran } 1 \leq p_2 \leq K \end{aligned}$$

e deixe

$$\sigma_{tr1} = \sigma p_2, \sigma_{tr2} = \sigma p_1, \text{ e } \sigma_{ti} = \sigma i \text{ para } i \in \{1, \dots, K\} \setminus \{p_1, p_2\}.$$

O fato de as duas posições serem amostradas com substituição pode parecer incidental, no entanto, há boas razões para isso. Lembre-se de que exigimos que nosso mecanismo de proposta seja capaz de propor grandes saltos em torno do espaço de permutação. É evidente que, se considerarmos apenas a troca de duas posições dentro de σ o número de propostas únicas será reduzido; especificamente $(K C_2 + 1) K!$ (incluindo o "swap nulo" $p_1 = p_2$). Por conseguinte, para aumentar o número de propostas possíveis, é sensato considerar a possibilidade de trocar duas posições > 1 vezes. No entanto, ao não permitir a troca nula, limitamos substancialmente as possíveis propostas, dependendo se s é escolhido para ser par ou ímpar. Por exemplo, suponha que $\sigma = (1, 2, 3, 4, 5)$ e a troca de posições $p_1 = p_2$ não seja permitida. Neste cenário, é impossível obter $\sigma = (2, 1, 3, 4, 5)$ se $s > 1$ for par e, em contraste, se $s > 1$ for ímpar, então $\sigma = (2, 1, 3, 5, 4)$ não pode ser obtido. No entanto, ao permitir a troca nula, esse problema é evitado, pois as posições de troca $p_1 = p_2$ efetivamente mudam s de pares ou ímpares sem alterar o estado atual de σ ; esta noção é discutida mais adiante na Seção 6.4.5.

Uma característica fundamental da implementação de uma proposta Metropolis-Hastings é a chamada proposta $q(\sigma|\sigma')/q(\sigma'|\sigma)$. Claro, a proporção da proposta vai envolver a proporção do

distribuição de mistura (6.9), no entanto, é útil se considerarmos primeiro cada proposta por vez, ou seja, $q_1(\sigma | \sigma\ddagger) / q_1(\sigma\ddagger | \sigma)$. Considere primeiro o caso quando $s = 1$. Para este mecanismo de proposta, é claro que existem duas maneiras possíveis de formar o $\sigma\ddagger$ da proposta; quer trocando de posições (P_1, P_2) quer trocando de posições (P_2, P_1). A razão proposta é formada considerando a probabilidade de obter σ dado o valor proposto $\sigma\ddagger$ juntamente com a probabilidade do "movimento reverso", ou seja, a probabilidade de obter $\sigma\ddagger$ dado o estado atual σ . Daqui resulta que, tendo em conta

$$\Pr(p_1 = p_1, p_2 = p_2) = \frac{1}{K} \cdot 2 = \Pr(p_1 = p_2, p_2 = p_1),$$

A proporção da proposta é

$$\frac{q_1(\sigma | \sigma\ddagger) q_1(\sigma\ddagger | \sigma)}{q_2(\sigma | \sigma\ddagger) q_2(\sigma\ddagger | \sigma)} = \frac{\Pr(p_1 = p_2, p_2 = p_1) + \Pr(p_1 = p_1, p_2 = p_2)}{2 \Pr(p_1 = p_1, p_2 = p_2) + \Pr(p_1 = p_2, p_2 = p_1)} =$$

$$= 1$$

e, portanto, esse mecanismo de proposta é claramente simétrico. Uma generalização direta deste resultado mostra que esta proposta permanece simétrica quando se considera a troca de duas posições > 1 vezes; isso é discutido mais adiante na Seção 6.4.5.

Proposta 2 – a troca de Poisson

O segundo mecanismo de proposta é o que chamamos de "swap de Poisson". Quanto à troca aleatória (Proposta 1), formamos a ordem de escolha proposta trocando duas posições (p_1, p_2) da σ de permutação atual, no entanto, as posições que trocamos não são mais escolhidas uniformemente ao acaso. Aqui, em vez disso, consideramos a troca de posições p_1 e $p_2 = p_1 + m$ onde m segue uma distribuição de Poisson (mistura). A ideia aqui é que trocar de posição mais próximas umas das outras (m minúsculo) levará a maiores taxas de aceitação e, portanto, podemos ajustar esse mecanismo de proposta por meio da escolha da distribuição para m .

Formalmente, amostramos a posição p_1 uniformemente ao acaso, ou seja, a partir da distribuição discreta

$$\Pr(p_1 = p_1) = \frac{1}{K}, \quad \text{duran} \frac{1}{te} \leq p_1 \leq K.$$

Então, em contraste com a Proposta 1, trocamos a posição p_1 por uma posição que está a um certo número de posições de distância, ou seja, $m = (-1)^p f$ onde $p \sim \text{Bern}(0,5)$, $f \sim \text{Po}(\tau)$ e assim $p_2 = p_1 + m$. Segue-se que o parâmetro τ é considerado um parâmetro de afinação

o que nos permite alterar a distribuição da distância entre os swaps propostos para obter taxas de aceitação razoáveis. Claro, exigimos que $p_2 \in \{1, \dots, K\}$ e, portanto, definimos a ordem de escolha como cíclica, ou seja, supomos que σ_1 está próximo a σ_K . Segue-se que podemos deixar $p_2 \rightarrow p_2 \text{ Mod } K$ onde Mod é o módulo superior dado por

$$x \text{ Contra } K \equiv \{(x - 1) \text{ contra } K\} + 1.$$

A ordem de escolha proposta σ^\dagger é então formada como para a Proposta 1, ou seja,

$$\sigma_1^\dagger = \sigma p_2, \sigma_2^\dagger = \sigma p_1, \text{ e } \sigma_i^\dagger = \sigma i \text{ para } i \in \{1, \dots, K\} \setminus \{p_1, p_2\}.$$

Antes de considerarmos a razão proposta para esta troca, é útil notar que a distribuição de m é simétrica em torno de 0, ou seja, $\Pr(m = c) = \Pr(m = -c)$ para todo $c \in N$. A proporção proposta para esta troca é, portanto,

$$\begin{aligned} q_2(\sigma|\sigma^\dagger)q_2(\sigma^\dagger|\sigma) &= \Pr(p_1 = p_2) \Pr(p_2 = p_1) \\ p_1 = p_2) \Pr(p_1 = p_1) \Pr(p_2 = p_2|p_1 = p_1) &= \\ K \Pr(p_1 + m = p_1|p_1 = p_2) \Pr(p_1 + m = p_2|p_1 = p_1) &= \Pr(m = p_1 - p_1|p_1 = p_2) \Pr(m = p_2 - p_1|p_1 = p_1) = \Pr(m = p_1 - p_2) \Pr(m = p_2 - p_2) = \Pr(m = p_1 - p_2) \end{aligned}$$

$$= 1,$$

E então essa troca é claramente simétrica. Mais uma vez, uma generalização direta desse resultado mostra que esse mecanismo de proposta permanece simétrico ao aplicar os swaps $s > 1$.

Proposta 3 – inserção aleatória

A terceira proposta que consideramos é uma proposta de "inserção aleatória" (Bez'akov'a et al., 2006). Em contraste com as propostas anteriores (baseadas em swap) aqui, a ordem de escolha proposta σ^\dagger é formada tomando o valor na posição p_1 e inserindo-o de volta na permutação, de modo que fique na posição p_2 . Segue-se que esse mecanismo move várias posições dentro da permutação, ao contrário dos movimentos de swap considerados anteriormente. Para ver isso, talvez seja útil considerar permutar ("embaralhar") um baralho de cartas. De acordo com esta proposta, o novo baralho permitido é formado removendo a carta que está atualmente na posição p_1 e, em seguida, reinserindo-a no baralho para que tenha a posição p_2 . Claramente, o

As cartas nas posições entre p_1 e p_2 também devem se mover para acomodar isso. Essas cartas se moverão para cima ou para baixo em uma posição, dependendo do valor do par (p_1, p_2) .

Formalmente, amostramos duas posições (p_1, p_2) das distribuições discretas

$$\Pr(p_1 = p_1) = \frac{1}{K}, \quad \text{durante } 1 \leq p_1 \leq K$$

$$\Pr(p_2 = p_2 | p_1 = p_1) = \frac{1}{K-1}, \quad \text{durante } 1 \leq p_2 \neq p_1 \leq K$$

e deixe

$$s^\dagger = \begin{cases} \square & (s_1, \dots, s_{p_1-1}, s_{p_1+1}, \dots, s_{p_2}, s_{p_1}, s_{p_2+1}, \dots, s_K), & \text{se } p_1 < p_2 \\ \square & (s_1, \dots, s_{p_2-1}, s_{p_1}, s_{p_2}, \dots, s_{p_1-1}, s_{p_1+1}, \dots, s_K), & \text{caso contrário.} \end{cases}$$

Por exemplo, suponha que a ordem de escolha atual seja $\sigma = (1, 2, 3, 4, 5)$ e tenhamos $(p_1, p_2) = (2, 4)$, então a ordem de escolha proposta é $\sigma^\dagger = (1, 3, 4, 2, 5)$, ou, se em vez disso $(p_1, p_2) = (4, 2)$, então a ordem de escolha proposta seria $\sigma^\dagger = (1, 4, 2, 3, 5)$.

Quanto aos mecanismos de proposta anteriores, a proporção da proposta para esta mudança é direta. Suponha que a permutação proposta seja formada pela aplicação dos movimentos acima dados ($p_1 = p_1, p_2 = p_2$), então fica claro que a ordem de escolha atual pode ser recuperada da σ^\dagger proposta considerando $(p_1 = p_2, p_2 = p_1)$ e assim

$$\frac{q_3(\sigma|\sigma^\dagger)q_3(\sigma^\dagger|\sigma)}{p_1 = p_2} = \Pr(p_1 = p_2) \Pr(p_2 = p_1) \frac{p_1 = p_2}{p_1 = p_2} \Pr(p_1 = p_1) \Pr(p_2 = p_2 | p_1 = p_1) =$$

$$\frac{K(K-1)K(K-1)}{K(K-1)K(K-1)} = 1.$$

Proposta 4 – proposta prévia

O quarto mecanismo de proposta que consideramos é uma "proposta prévia". Aqui σ^\dagger é simplesmente um sorteio independente da distribuição anterior $\pi(\sigma)$. Formalmente, esta é uma proposta de independência, pois a distribuição da proposta é independente do estado atual da cadeia de Markov, ou seja, $q(\sigma^\dagger|\sigma) = q(\sigma^\dagger)$. Daqui resulta que o rácio proposto é simples e é dado por

$$\frac{q_4(s|\sigma^\dagger)q_4(s^*|s)}{p(s)p(\sigma^\dagger)} = \frac{\Pr(s = s^*)}{\Pr(s = s^\dagger)}$$

$$= 1$$

como nossa distribuição anterior é uniforme em todas as permutações. É claro que esta proposta pode não ser simétrica se for especificada uma opção prévia alternativa.

Proposta 5 – proposta inversa

Nosso mecanismo de proposta final é a "proposta reversa". Aqui deixamos σ^\dagger ser a ordem inversa da σ de permutação atual, ou seja, se $\sigma = (1, 2, 3, 4, 5)$, então a ordem proposta é $\sigma^\dagger = (5, 4, 3, 2, 1)$. Observe que, em geral, a permutação reversa não é a permutação inversa. Este mecanismo de proposta é motivado por nossa observação da investigação sobre a probabilidade do modelo EPL (sob todas as ordens de escolha possíveis) na Seção 6.3.2, onde vimos que o inverso da ordem de escolha "verdadeira" apareceu bastante na classificação das ordens de escolha; ver Tabela 6.4. Além disso, é improvável que o inverso da ordem de escolha atual seja proposto por qualquer um dos mecanismos de proposta anteriores, pois, por construção, a permutação reversa está a uma grande distância do estado atual. Formalmente, deixamos

$$\sigma^\dagger = \sigma K:1 = (\sigma K, \dots, \sigma 1).$$

com probabilidade 1. Segue-se que a distribuição da proposta é $q5(\sigma^\dagger|\sigma) = \delta\sigma K:1$ onde δx denota a medida de probabilidade de Dirac concentrada em x e, portanto, a razão da proposta é

$$\frac{q5(s|\sigma^\dagger)q5}{(s|\sigma^\dagger|s)} \frac{\delta\sigma^\dagger K:1}{K:1} = 1. \quad (6.10)$$

A probabilidade de aceitação

Em geral, a probabilidade de aceitação de σ^\dagger é $\min(1, A)$ onde

$$A = \frac{\pi(\sigma^\dagger | \dots) p(s | \dots)}{q(s | \sigma^\dagger) q(s | s)} \times$$

A partir da densidade de todas as grandezas estocásticas (6.8), fica claro que

$$\begin{aligned} p(s | \dots) &\propto p(D | I, s) p(Z | D, I, s) p(s) \\ &= \pi(D, Z | \lambda, \sigma) \pi(\sigma), \end{aligned}$$

e assim, sem surpresa, a distribuição posterior de σ é proporcional aos dados completos vezes a anterior.

Lembre-se de que nossa distribuição proposta é a distribuição de mistura de 5 componentes (6,9) e, observando que $q_i(\sigma \uparrow | \sigma) = q_i(\sigma | \sigma \uparrow)$ para $i = 1, \dots, 5$ segue-se que a proporção da proposta é

$$\frac{q(s|s \uparrow)q(s \uparrow|s)}{\sum_{i=1}^5 \frac{\prod_{j \neq i} p_j}{\prod_{j \neq i} p_j} q_i(s \uparrow|s)} = \frac{\sum_{i=1}^5 \frac{\prod_{j \neq i} p_j}{\prod_{j \neq i} p_j} q_i(s \uparrow|s)}{\sum_{i=1}^5 \frac{\prod_{j \neq i} p_j}{\prod_{j \neq i} p_j} q_i(s \uparrow|s)} = 1.$$

Claramente, a probabilidade de aceitação simplifica substancialmente e pode ser escrita como $\min(1, A)$ onde $A = \pi(D, Z|\lambda, \sigma \uparrow) \Pr(\sigma = \sigma \uparrow) \pi(D, Z|\lambda, \sigma) \Pr(\sigma = \sigma) = \pi(D, Z|\lambda, \sigma \uparrow) \pi(D, Z|\lambda, \sigma)$, que é simplesmente a razão da probabilidade completa dos dados dada a permutação proposta e a atual (assumindo que cada ordem de escolha é escolhida para ser igualmente provável a priori).

6.4.5 Considerações adicionais

Explorar grandes espaços discretos, como o conjunto de todas as permutações S_5 com um algoritmo Metropolis-Hastings, é uma tarefa não trivial. A distribuição de propostas discutida na Seção 6.4.4 compreende uma mistura de mecanismos de proposta "locais" (Propostas 1-3) e "globais" (Propostas 4 e 5) para o parâmetro de ordem de escolha em uma tentativa de facilitar a exploração efetiva desse grande espaço discreto. Embora seja razoavelmente simples de implementar, é útil começar por considerar a distribuição da proposta com mais pormenor, uma vez que uma implementação ingênua pode resultar numa distribuição ineficiente da proposta devido às subtilezas envolvidas em alguns dos mecanismos da proposta; particularmente os movimentos de troca.

A primeira coisa a notar é que, por construção, tanto a Proposta 4 quanto a 5 têm o potencial de propor grandes saltos (em termos de distância) da permutação atual. Embora útil para escapar de modos locais, grandes saltos normalmente resultam em menores probabilidades de aceitação. Diante disso, talvez seja sensato considerar apenas a construção de $\sigma \uparrow$ a partir da Proposta 4 ou 5 com relativa pouca frequência em comparação com as propostas mais locais. Isto pode ser conseguido através de uma escolha adequada dos pesos dos componentes da mistura na distribuição proposta (6.9). À primeira vista, também podemos pensar que a Proposta 3 pode sofrer de baixas taxas de aceitação, uma vez que a permutação proposta é formada pela movimentação de várias posições dentro da permutação atual. No entanto, a construção deste mecanismo de proposta determina que grande parte da estrutura que está presente na permutação atual também está presente na permutação proposta. Segue-se que, em termos de distância, a permutação proposta não estará tão longe da permutação atual, pois grande parte da ordem é preservada.

Talvez surpreendentemente, são as Propostas 1 e 2 que têm potencial para causar problemas na distribuição desta proposta. Observe que, trocando apenas duas posições dentro σ o número de propostas exclusivas que podem ser geradas é $(K C_2 + 1) K!$ (incluindo o "swap nulo" $p_1 = p_2$) e, portanto, poderíamos facilmente ficar presos em um modo local. Portanto, como aludido na Seção 6.4.4, talvez seja sensato considerar a troca de duas posições $s > 1$ vezes para permitir que nosso mecanismo de proposta proponha saltos maiores em torno do espaço de permutação. Cada uma das trocas s deve ser executada iterativamente. Por exemplo, suponha que temos s pares de posições para trocar $(p_1, p_2)_1, \dots, (p_1, p_2)_s$ que são amostras das distribuições discretas apropriadas. A ordem de escolha proposta $\sigma \dagger$ é obtida primeiro formando uma "proposta temporária" $\sigma \dagger 1$ trocando posições $(p_1, p_2)_1$ dentro do estado atual σ , em seguida, considerando $\sigma \dagger 1$ como o estado atual da cadeia, obtemos outra proposta temporária $\sigma \dagger 2$ trocando posições $(p_1, p_2)_2$ dentro de $\sigma \dagger 1$. Esse processo continua e a ordem de escolha proposta é $\sigma \dagger = \sigma \dagger s$. Naturalmente, à medida que s aumenta, também aumenta o número de propostas possíveis, N_p . A distância entre as ordens de escolha atuais e propostas também aumentará normalmente com s e, portanto, o número de swaps é nominalmente um parâmetro de ajuste e pode ser ajustado para aumentar as taxas de aceitação. Claro, devemos considerar como o desempenho de $s > 1$ swaps afeta a proporção da proposta. Consideramos isso para a Proposta 1 e observamos que argumentos semelhantes se aplicam à Proposta 2. Lembre-se de que cada uma das trocas s é realizada iterativamente e, portanto, podemos escrever a distribuição da proposta como

$$\begin{aligned} q_1(\sigma \dagger | \sigma) &= q_1(\sigma \dagger s | \sigma \dagger s-1, \dots, \sigma \dagger 1) \times \dots \times \\ q_1(s \dagger 1 | s) &= q_1(s \dagger s | s \dagger s-1) \times \dots \times Q_1(S \dagger 1 | S) \end{aligned}$$

Dado que a distribuição da proposta para cada swap s é condicionalmente independente, dada a proposta temporária $\sigma \dagger s-1$. Além disso, também é claro que

$$q_1(\sigma | \sigma \dagger) = q_1(\sigma \dagger s-1 | \sigma \dagger s) \times \dots \times 1^{\text{o trimestre}}(s \dagger 1)$$

e assim a razão proposta $q_1(\sigma | \sigma \dagger) / q_1(\sigma \dagger | \sigma) = 1$ para todo $s \geq 1$ que se segue do resultado que $q_1(\sigma_i | \sigma_j) = q_1(\sigma_j | \sigma_i)$ para qualquer $\sigma_i, \sigma_j \in SK$; consulte a Seção 6.4.4. Voltamos agora brevemente nossa atenção para o motivo pelo qual é sensato permitir a troca nula ao gerar permutações propostas usando um mecanismo de troca (Propostas 1 e 2), conforme mencionado na Seção 6.4.4. Naturalmente, esperamos que o número de propostas possíveis (N_p) aumente à medida que consideramos mais swaps, mas especificamente, podemos imaginar que $N_p \rightarrow K! \text{as } s \rightarrow \infty$. No entanto, se supormos que a troca de posições $p_1 = p_2$ não é permitida, então $N_p = K!/2$ para s grande. Além disso, o conjunto de possíveis permutações que podem ser propostas depende se s é escolhido para ser par ou ímpar. Se deixarmos $\Sigma_{\text{ímpar}}$ e Σ_{par} denotarem os conjuntos de permutações possíveis que podem ser geradas quando s é ímpar e par, respectivamente, então pode ser mostrado que $SK = \{\Sigma_{\text{ímpar}} \cup \Sigma_{\text{par}}\}$ e $\Sigma_{\text{ímpar}} \cap \Sigma_{\text{par}} = \emptyset$. Segue-se que se $s > 1$ é

fixado a priori e depois trocando repetidamente duas posições $p_1 \leftrightarrow p_2$ só nos permite explorar metade das permutações possíveis - aquelas com paridade par ou ímpar (dependendo da paridade da permutação atual). Embora aqui tenhamos considerado o resultado assintótico, fica claro a partir do nosso exemplo na Seção 6.4.4 que não permitir a troca nula resulta no mecanismo de proposta ser incapaz de propor algumas permutações que são próximas à distância, dependendo da escolha de s . Uma estratégia para evitar esse problema é considerar s como uma variável aleatória em oposição a uma constante fixa. Escolher $s \sim \text{Geom}(ps)$ talvez seja sensato, no entanto, optamos por considerar s fixo e permitir a troca nula.

6.4.6 Um algoritmo Metropolis-within-Gibbs para o modelo EPL

Um algoritmo de Metropolis-within-Gibbs pode ser construído para direcionar a distribuição posterior conjunta dos parâmetros de habilidade λ , o parâmetro de ordem de escolha σ e as variáveis latentes Z . As amostras posteriores são geradas da seguinte forma.

1. Inicializar: escolha $\sigma \in SK$, $\lambda \in RK > 0$ e $z_{ij} > 0$ para $i = 1, \dots, n$, $j = 1, \dots, K$.
2. Execute repetidamente as seguintes etapas:
 - Amostra $\lambda_k | \dots, e + N, 1 + \sum_{i=1}^n \sum_{j=1}^K \delta_{ij}(k) z_{ij}$ para $k = 1, \dots, K$, onde $\delta_{ij}(k)$ é como na Seção 6.4.2.
 - Amostra $z_{ij} | \dots, \text{Exp} \left(\frac{K \sum_{m=j}^M L_{S-1} X_i \Sigma}{M} \right)$ pois $i = 1, \dots, n$, $j = 1, \dots, K$.
 - Amostra σ' da distribuição discreta com probabilidades $\Pr(\sigma' = i) = p_{\text{prop}}^{i-1}$ durante $i = 1, \dots, S$
 - propor σ' usando o mecanismo de proposta – seja $\sigma \rightarrow \sigma'$ com probabilidade $\min(1, A)$ onde $A = \frac{\pi(D, Z|\lambda, \sigma)}{\pi(D, Z|\lambda, \sigma')} \frac{\Pr(\sigma' = \sigma)}{\Pr(\sigma = \sigma')}$
 - Reescala – amostra $\lambda'_k \sim Ga(K \sum k = 1, 1)$.

- calcular $\Sigma = \sum_{k=1}^K \lambda_k$.
- seja $\lambda_k \rightarrow \lambda_k \Lambda'_k / \Sigma$ para $k = 1, \dots, K$.

Se K for suficientemente pequeno, ou tivermos o poder computacional para fazê-lo, poderíamos amostrar σ de sua distribuição condicional completa, conforme discutido na Seção 6.4.3, caso em que teríamos um amostrador de Gibbs direto.

6.4.7 Estudo de simulação – Metrópole dentro de Gibbs

Neste estudo, consideraremos um novo conjunto de dados (Conjunto de dados 6) com $n = 100$ classificações completas de $K = 10$ entidades. Os parâmetros de habilidade usados para simular esses dados são $\lambda = (10, 9, \dots, 1)$ e a ordem de escolha (processo de classificação) é escolhida para ser $\sigma = (3, 10, 9, 1, 7, 4, 5, 6, 8, 2)$. Os dados simulados são apresentados nos quadros B.8 e B.9 dos apêndices.

Adotaremos a mesma distribuição a priori para os parâmetros de habilidade usada em análises anteriores, ou seja, pegue $\lambda_k \sim \text{Indep} \sim \text{Ga}(1, 1)$ e então $a_k = a = 1$. Além disso, assumimos que cada ordem de escolha é igualmente provável a priori e, portanto, $\Pr(\sigma = \sigma_i) = 1 / K!$ para todos $\sigma_i \in SK$. Dada essa distribuição anterior, usamos o algoritmo Metropolis-within-Gibbs da Seção 6.4.6 na tentativa de obter realizações posteriores de $\pi(\lambda, \sigma, Z | D)$. Claro, também devemos escolher parâmetros de ajuste adequados para a distribuição da proposta (6.9) do parâmetro choiceorder σ . A partir de nossa discussão na Seção 6.4.5, acreditamos que é sensato letpprop = (0,3, 0,3, 0,3, 0,05, 0,05) para que as propostas "locais" (swaps e inserção) sejam favorecidas em relação às propostas globais (anteriores e reversas), uma vez que as primeiras são talvez mais prováveis de serem aceitas. Também optamos por realizar trocas $s = 2$ ao usar as propostas 1 e 2 e tomar $\tau = 3$ dentro da última, de modo que a distância esperada entre as posições trocadas seja 3.

Para investigar as propriedades de convergência e mistura de nosso sampler Metropolis-within-Gibbs, consideramos 5 cadeias; cada um dos quais é inicializado em um parâmetro de ordem de escolha diferente o como mostrado na Tabela 6.5. Naturalmente, esperamos que o esquema MCMC tenha sido construído de forma que seja capaz de explorar (eficientemente) o conjunto de todas as permutações e (rapidamente) convergir para sua distribuição estacionária, independentemente da ordem de escolha inicial. Relatamos os resultados de uma execução típica de nosso esquema MCMC inicializado conforme descrito, com um burn-in de 10K iterações e, em seguida, executado por mais 1M iterações e diluído em 100 para obter 10K ("posterior") amostras. O esquema MCMC é executado razoavelmente rápido, com código C em um único thread de uma CPU Intel Core i7-4790S (velocidade de clock de 3,20 GHz) levando cerca de 1 minuto e 35 segundos (para cada cadeia).

Cadeia	Permutação inicial
1	$\sigma = (1, 2, \dots, 10)$
2	$\sigma = (10, 9, \dots, 1)$
3	$\sigma = (3, 10, 9, 1, 7, 4, 5, 6, 8, 2)$
4	$\sigma = (2, 8, 6, 5, 4, 7, 1, 9, 10, 3)$
5	$\sigma = (8, 10, 7, 4, 3, 1, 2, 6, 5, 9)$

Tabela 6.5: Permutações (ordens de escolha) usadas para a inicialização de cada cadeia

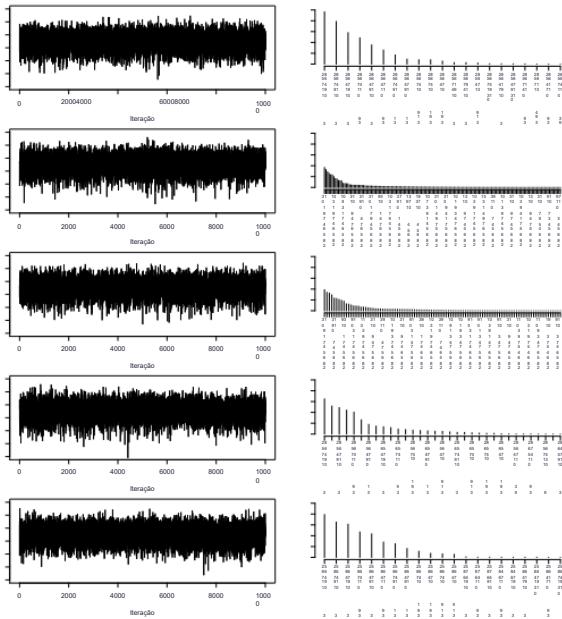


Figura 6.2: Gráficos de rastreamento da verossimilhança dos dados logarítmicos completos (esquerda) e do $\pi(\sigma|D)$ (direita) para cadeias 1 a 5 de cima para baixo, respectivamente (abordagem Metropolis-within-Gibbs).

A inspeção dos gráficos de rastreamento mostrando a probabilidade dos dados log completos [$\log \pi(D, Z|\lambda, \sigma)$] para cada cadeia mostra que o esquema MCMC parece estar se misturando razoavelmente bem e, de fato, mostra sinais de convergência; veja a Figura 6.2 à esquerda. No entanto, se olharmos para a distribuição posterior marginal de σ (definida pelas probabilidades $\Pr(\sigma = \sigma_i | \dots)$), então fica claro que as cadeias não atingiram sua distribuição estacionária; veja a Figura 6.2 à direita. É claro que, dado que o esquema MCMC é executado rapidamente, poderíamos realizar mais iterações na esperança de que as cadeias atinjam a convergência. No entanto, em nossa experiência, isso geralmente é difícil de alcançar. A partir da Figura 6.2 (à direita), fica claro que as cadeias 1, 4 e 5 ficaram presas no modo local em torno da ordem de escolha "verdadeira" reversa. Embora tenhamos introduzido a Proposta 5 (o

Mecanismo de proposta1234				
	5			
# proposto	302995302982	303379	5024650399	
# aceito	18671921	1383	00	
Pr (aceitar)	0.006	0.006	0.005	00

Tabela 6.6: Taxas de aceitação para cada um dos 5 mecanismos de proposta para a ordem de escolha σ .

proposta reversa) na esperança de evitar isso, parece que, neste cenário, este *ProposalMove* não permite que a cadeia se move para o outro modo (em torno da ordem de escolha "verdadeira"). Esta observação é ainda mais apoiada se considerarmos as taxas de aceitação da distribuição de propostas para σ . A taxa de aceitação global de σ é inferior a 1% e o Quadro 6.6 mostra as taxas de aceitação para cada um dos 5 mecanismos propostos (componentes da proposta de mistura). Observe que os resultados mostrados são os obtidos da cadeia 1, no entanto, esses resultados são indicativos das outras cadeias. A partir da Tabela 6.6, vemos que nem o inverso nem a proposta anterior (*Propostas 4 ou 5*) são aceitos e, portanto, nossas cadeias lutam para escapar dos modos locais. É evidente que esta é uma questão que tem de ser resolvida. A partir de uma investigação mais aprofundada da distribuição posterior (e também dos MLEs λ_j da Seção 6.3 que maximizam a probabilidade de Plackett-Luce estendida para cada ordem de escolha), fica claro que existe uma grande correlação entre λ e σ , ou seja, entre os parâmetros de habilidade e a ordem de escolha. Além disso, ao longo desta tese, vimos que, para uma ordem de escolha fixa, a amostragem de Gibbs por meio de aumento de dados leva à rápida convergência da cadeia de Markov para a distribuição estacionária $\pi(\lambda, Z|D, \sigma)$. Segue-se que, dada uma ordem de escolha fixa σ , se os parâmetros de habilidade λ (e os parâmetros latentes Z) forem suficientemente otimizados, então será difícil propor uma ordem de escolha alternativa razoável (dados esses λ, Z) e, portanto, nossa cadeia ficará presa neste modo. Uma estratégia típica para melhorar a mistura em tal cenário é considerar uma atualização conjunta dos parâmetros altamente correlacionados (Gemanian e Lopes, 2006). Infelizmente, realizar uma atualização conjunta de (λ, Z, σ) não é simples - a dificuldade é como construir um mecanismo de proposta adequado. Claro, poderíamos usar $q(\sigma|\lambda)$ para gerar uma ordem de escolha proposta como antes, então, dada a ordem de escolha proposta σ^* , extrair os parâmetros de habilidade e variáveis latentes de suas distribuições condicionais completas. Ao usar essa estratégia, temos as seguintes opções

- | | |
|---|--|
| 1. desenhar
$s \dagger s$
Sorteio $L \dagger Z, p^*$
desenhar $Z \dagger l, p^*$ | 2. desenhe $s \dagger s$
desenhar $Z \dagger l,$
p^*
Sorteio $L \dagger Z^*,$
págt |
|---|--|

no entanto, nenhuma das opções provavelmente gerará propostas razoáveis ($\lambda \dagger$, $Z \dagger$, $\sigma \dagger$), dado que, claramente, λ e Z serão altamente correlacionados.

Diante disso, devemos, portanto, considerar uma estratégia alternativa na tentativa de quebrar as correlações e obter um esquema MCMC que exiba uma boa mistura em relação à ordem de escolha σ . Propomos remover as variáveis latentes Z do espaço de parâmetros e, em vez disso, considerar as atualizações de Metropolis-Hastings para os parâmetros de habilidade λ_k e a ordem de escolha σ ; Este é o tópico da próxima seção.

6.4.8 Metropolis-Hastings propostas para a

A partir da seção anterior, fica claro que nosso esquema de amostragem Metropolis-within-Gibbs é inadequado para gerar realizações a partir da distribuição posterior para o modelo ExtendedPlackett-Luce. Agora, em vez disso, consideramos uma atualização de Metropolis-Hastings para os parâmetros de habilidade λ , portanto, não precisamos introduzir as variáveis latentes Z no espaço de amostra. Em primeiro lugar, observe que, dado que não estamos considerando as variáveis latentes Z , a densidade de todas as quantidades estocásticas é agora

$$p(l, D, s) = p(D|l, s)p(l)p(s) \\ = \prod_{i=1}^n \prod_{j=1}^K \frac{\lambda^{\sigma-1} x_{ij}^{\sigma} \sum_{m=j}^K x_{im}}{\lambda^{\sigma-1} x_{im}^{\sigma}} \times \prod_{k=1}^K \frac{\lambda^{ak-1} k^{\sigma-1}}{\lambda^k \Gamma(ak)} \frac{1}{1^K} \quad (6.11)$$

e assim

$$p(\lambda_k | \dots) \propto \prod_{i=1}^n \prod_{j=1}^K \frac{\lambda^{\sigma-1} x_{ij}^{\sigma} \sum_{m=j}^K x_{im}}{\lambda^{\sigma-1} x_{im}^{\sigma}} \frac{\lambda^{ak-1} k^{\sigma-1}}{\lambda^k \Gamma(ak)} \\ = \pi(D|\lambda, \sigma) \pi(\lambda_k) \quad (6.12)$$

para $k = 1, \dots, K$.

Claramente, esta é uma distribuição não padrão e, portanto, usamos uma etapa Metropolis-Hastings para direcionar essa distribuição. Dado $\lambda_k > 0$, é sensato considerar uma distribuição de proposta log-normal centrada no valor atual para os parâmetros de habilidade, ou seja, pegue $\lambda \dagger | \lambda_k$ indep~LN($\log \lambda_k, \sigma^2 \lambda_k$). Para esta escolha de distribuição de proposta, a probabilidade de aceitação é $\min(1, A)$ onde

$$A = \frac{\pi(\lambda \dagger | \dots) q(\lambda_k)}{\pi(\lambda_k | \dots) q(\lambda \dagger | \lambda_k)} \\ \lambda \dagger | \lambda_k = \frac{\pi(D | \lambda - k, \lambda_k = \lambda \dagger | \lambda_k)}{\pi(D | \lambda, \sigma)} \times \frac{(L+K)^{ak}}{\lambda^k} e^{(\lambda_k - \lambda \dagger | \lambda_k)} \\ \sigma) \pi(D | \lambda, \sigma)$$

Além disso, como a densidade de todas as quantidades estocásticas mudou agora (dado que não consideramos mais as variáveis latentes Z), é claro que a probabilidade de aceitação de σ também deve mudar. A partir da forma da densidade de todas as grandezas estocásticas (6.11), fica claro que $\pi(\sigma | \dots) \propto \pi(D|\lambda, \sigma)\pi(\sigma)$ e, portanto, dado o mecanismo de proposta descrito na Seção 6.4.4, a probabilidade de aceitação para $\sigma \uparrow$ é $\min(1, A)$ onde

$$A = \frac{\pi(D|\lambda, \sigma \uparrow)}{\sigma \uparrow \pi(D|\lambda, \sigma)} \Pr(\sigma = \sigma \uparrow)$$

que é simplesmente a razão da probabilidade EPL dada a permutação proposta e a atual (assumindo que cada ordem de escolha é escolhida para ser igualmente provável a priori).

6.4.9 Um algoritmo Metropolis-Hastings para o modelo EPL

Usando os resultados da Seção 6.4.8, podemos construir um algoritmo de Metropolis-Hastings para direcionar a distribuição posterior conjunta dos parâmetros de habilidade λ e o parâmetro de ordem de escolha σ da seguinte forma.

1. Inicializar: escolha $\sigma \in SK$ e $\lambda \in RK > 0$

2. Execute repetidamente as seguintes etapas:

- Para $k = 1, \dots, K$ — desenhe $\lambda \uparrow k | \lambda_k \sim LN(\log \lambda_k, \sigma^2 \lambda_k)$ — seja $\lambda_k \rightarrow \lambda \uparrow k$ com probabilidade $\min(1, A)$
- onde

$$A = \frac{\pi(D|\lambda - k, \lambda_k = \lambda \uparrow k)}{\sigma \uparrow \pi(D|\lambda, \sigma)} \times \frac{(\lambda \uparrow k)^k}{\lambda_k} e^{(\lambda_k - \lambda \uparrow k)}$$

- Amostra ' da distribuição discreta com probabilidades $\Pr(' = i) = p_{\text{prop}} \quad \text{durante } i = 1, \dots, 5$
 - propor $\sigma \uparrow$ usando o mecanismo de proposta '— seja $\sigma \rightarrow \sigma \uparrow$ com probabilidade $\min(1, A)$ onde $A = \frac{\pi(D|\lambda, \sigma \uparrow)}{\sigma \uparrow \pi(D|\lambda, \sigma)} \Pr(\sigma = \sigma \uparrow) \cdot$
- Reescalonar – amostra $\Lambda \uparrow \sim Ga(K \sum k=1 \lambda_k, 1)$.

- calcular $\Sigma = \sum_{k=1}^K \lambda_k$.
- seja $\lambda_k \rightarrow \lambda_k / \Sigma$ para $k = 1, \dots, K$.

6.4.10 Estudo de simulação – Metropolis-Hastings

Neste estudo, revisitamos o Conjunto de Dados 6, que foi introduzido pela primeira vez na Seção 6.4.7. No entanto, aqui realizamos inferência bayesiana usando o algoritmo Metropolis-Hastings da Seção 6.4.9 em oposição ao amostrador Metropolis-within-Gibbs considerado anteriormente. Lembre-se de que o conjunto de dados 6 compreende $n = 100$ classificações completas de $K = 10$ entidades. Os parâmetros de habilidade e a ordem de escolha (processo de classificação) usados para simular esses dados são $\lambda = (10, 9, \dots, 1)$ e $\sigma = (3, 10, 9, 1, 7, 4, 5, 6, 8, 2)$, respectivamente.

Escolhemos a mesma distribuição a priori da Seção 6.4.7 e, portanto, $\alpha_k = \alpha = 1$ e cada ordem de escolha é igualmente provável a priori, ou seja, $\Pr(\sigma = \sigma_i) = 1 / K!$ para todos $\sigma_i \in SK$. Além disso, tomamos os mesmos parâmetros de ajuste para a distribuição da proposta da ordem de escolha, ou seja, optamos por realizar $s = 2$ trocas ao usar as propostas 1 e 2 e tomar $r = 3$ na última. Novamente, as probabilidades de cada mecanismo de proposta ser usado são $p_{prop} = (0,3, 0,3, 0,3, 0,05, 0,05)$. Quanto à análise anterior, consideramos 5 cadeias, cada uma das quais é inicializada em um parâmetro de ordem de escolha diferente σ conforme mostrado na Tabela 6.5. Aqui, inicializamos fazendo sorteios independentes do anterior ($\alpha_k \text{ Indep} \sim \text{Ga}(1, 1)$) e deixamos o parâmetro de ajuste ser $\sigma_{\lambda k} = \sigma_{\lambda} = 0,75$, que determinamos ser sensato após realizar várias execuções piloto. Os resultados relatados são de uma execução típica de nosso esquema MCMC inicializado conforme descrito, com um burn-in de 20 mil iterações e, em seguida, executado por mais 2 milhões de iterações e diluído em 200 para obter amostras de 10 mil ("posteriore"). O MCMCscheme é executado razoavelmente rápido, com código C em um único thread de um Intel Core i7-4790SCPU (velocidade de clock de 3,20 GHz) levando cerca de 24 minutos e 30 segundos (para cada cadeia). Observe que este algoritmo requer mais tempo computacional do que o algoritmo Metropolis-within-Gibbs correspondente da Seção 6.4.6 devido às avaliações de verossimilhança adicionais necessárias para calcular as taxas de aceitação de MH para os parâmetros de habilidade λk . A Figura 6.3 (à esquerda) mostra os gráficos de rastreamento da probabilidade logarítmica ($\log \pi(D|\lambda, \sigma)$) para cada cadeia, a partir da qual vemos que algumas das cadeias exibem sinais de má mistura. Apesar dos possíveis problemas de mistura, será interessante ver como é a distribuição posterior em cada cadeia. Lembre-se de que, quando consideramos um amostrador Metropolis-within-Gibbs, ficou claro a partir do marginal posterior sobre a ordem de escolha σ que as cadeias não haviam atingido sua distribuição estacionária. No entanto, em contraste com nossa análise anterior, nosso esquema MCMC parece estar se misturando muito bem ao longo do σ com nenhuma das cadeias ficando presa no modo local em torno da permutação "verdadeira" reversa; veja a Figura 6.3 (direita). Com efeito, embora apenas um pequeno número de ordens de escolha seja rotulado nas parcelas, verifica-se um aumento significativo do número de ordens de escolha que têm apoio posterior no âmbito desta análise; ver Figuras 6.2 e 6.3 (à direita).

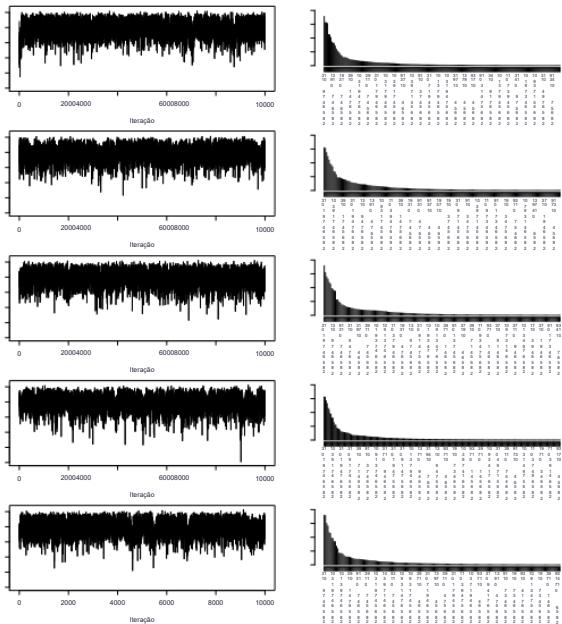


Figura 6.3: Gráficos de rastreamento da probabilidade de dados logarítmicos completos (esquerda) e do π posterior marginal ($\sigma|D$) (direita) para cadeias 1 a 5 de cima para baixo, respectivamente (abordagem Metropolis-Hastings).

Para investigar mais a fundo a distribuição posterior marginal de σ é útil observar quais ordens de escolha recebem suporte posterior alto. A Figura 6.4 mostra as 25 ordens de escolha com maior suporte posterior (e suas probabilidades posteriores correspondentes) das cadeias 1 a 5, ou seja, um subconjunto do posterior marginal. Embora existam discrepâncias entre os posteriores produzidos a partir de cada cadeia, é claro que a abordagem Metropolis-Hastings está tendo um desempenho substancialmente melhor do que o algoritmo Metropolis-within-Gibbs. É claro, no entanto, que a mistura sobre as ordens de escolha deve ser melhorada ainda mais, pois, embora razoável aqui, esse problema só piorará à medida que K aumentar. Para nos ajudar a explorar melhor SK , o conjunto

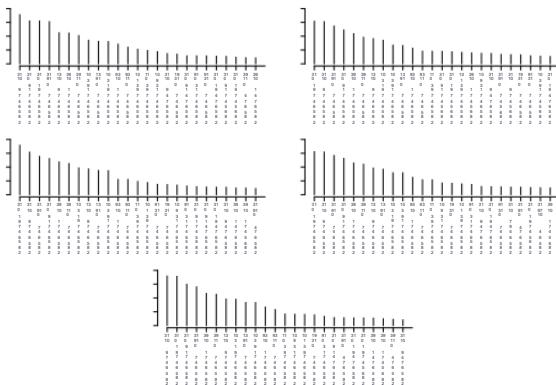


Figura 6.4: Subconjunto do π marginal posterior ($\pi|D$) mostrando as 25 ordens de escolha com maior suporte posterior das cadeias 1 a 5 (lidas da esquerda para a direita)

de todas as permutações, apelamos para a cadeia de Markov acoplada a Metrópole Monte Carlo, que é o tópico da próxima seção.

6.5 Metrópole acoplada cadeia de Markov Monte Carlo

Cadeia de Markov acoplada à metrópole Monte Carlo (Geyer, 1991), ou têmpera paralela, como também é conhecida na literatura bayesiana, é uma técnica de amostragem que visa melhorar a mistura de cadeias de Markov em comparação com os métodos MCMC padrão (Gilks e Roberts, 1996). Embora os métodos de Monte Carlo (MC3) da cadeia de Markov acoplada à metrópole possam ser aplicados em muitos cenários, esta técnica de amostragem é particularmente útil quando a distribuição-alvo é multimodal (Brooks, 1998). A ideia geral por trás da cadeia de Markov acoplada à Metrópole Monte Carlo é considerar várias cadeias "Markov" - uma das quais visa a densidade de interesse e as cadeias restantes visam densidades "temperadas" que são construídas para que sejam mais fáceis de explorar. As múltiplas cadeias são então acopladas ao Metropolis por meio da proposta de trocas de espaço de estado entre as cadeias. Segue-se que as amostras das melhores cadeias de mistura (aqueles que visam as densidades temperadas) podem ser filtradas para a cadeia que visa a densidade de interesse e, assim, melhorar a exploração dessa densidade. Notamos que, estritamente falando, cada uma das cadeias não é Markoviana, pois, devido à

(Metrópole) acoplamento das cadeias, cada cadeia não depende mais apenas do valor anterior dentro dessa cadeia, mas também dos valores anteriores de todas as outras cadeias. Na próxima seção, discutimos como é mais fácil construir uma cadeia de Markov para atingir eficientemente uma densidade temperada em oposição à densidade não temperada. Concluímos esta seção descrevendo a metodologia subjacente ao revenimento paralelo e fornecemos um esboço genérico do algoritmo.

6.5.1A vantagem de visar densidades temperadas

Aqui discutimos como a mistura (e, portanto, a eficiência de amostragem) de cadeias de Markov cujas densidades alvo são multimodais é melhorada quando elas visam uma densidade temperada adequada. Suponha que estamos interessados em projetar um esquema padrão de cadeia de MarkovMonte Carlo para atingir a densidade $\pi(\theta)$ onde

$$\theta \sim 12 N(-2, 0,52) + 12 N(2, 0,52),$$

isto é, a densidade de interesse é uma mistura normal de dois componentes igualmente ponderada com médias de componentes de ± 2 e desvios padrão comuns de 0,5. Segue-se que a densidade alvo $\pi(\theta)$ é multimodal por construção; veja a Figura 6.5. É claro que, dado que conhecemos a densidade alvo, é simples gerar realizações de θ , no entanto, para fins de argumentação, suponhamos que a distribuição de θ seja desconhecida. Nesse cenário, uma estratégia sensata seria usar um algoritmo de Metropolis-Hastings para direcionar a densidade. Dado que $\theta \in \mathbb{R}$, um passeio aleatório normal é uma distribuição de proposta plausível, ou seja, $\theta^* | \theta \sim N(\theta, \sigma^2)$. No entanto, deve ficar claro que, se tal algoritmo for inicializado em um dos modos da distribuição alvo, a cadeia de Markov terá dificuldade para "colmatar" a área de baixa probabilidade e explorar o outro modo. É claro que, para este exemplo, poderíamos inventar uma distribuição de proposta que encorajasse o algoritmo a saltar entre o

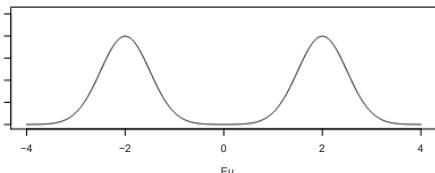


Figura 6.5: $\pi(\theta)$: gráfico de densidade de uma mistura normal de dois componentes igualmente ponderada com médias compostas de ± 2 e desvios-padrão de 0,05.

modos (por exemplo, $q(\theta^*|\theta) = \delta - \theta$) como conhecemos a forma do alvo. No entanto, em geral, não será esse o caso e, portanto, a construção de uma proposta adequada não será trivial; especialmente se o destino contiver um grande número de modos. Vamos agora considerar a densidade temperada $\pi(\theta) 1/T$ onde $T \geq 1$ é conhecido como temperatura. Claramente, a densidade alvo de interesse é recuperada quando $T = 1$. A Figura 6.6 mostra as densidades temperadas para as temperaturas $T \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. A partir disso, observamos que, à medida que aumentamos a temperatura ("calor"), a densidade revenida se torna "mais plana" e segue-se que $\pi(\theta) 1/T$ é completamente plana (uniforme) quando $T = \infty$. Torna-se, portanto, cada vez mais simples construir uma cadeia de Markov capaz de direcionar efetivamente $\pi(\theta) 1/T$ como $T \rightarrow \infty$. Claro, só estamos interessados em direcionar a densidade $\pi(\theta) 1/T$ quando $T = 1$. No entanto, como todas as densidades temperadas estão inherentemente relacionadas, podemos usar as amostras aceitas de uma cadeia visando uma densidade temperada para melhorar o

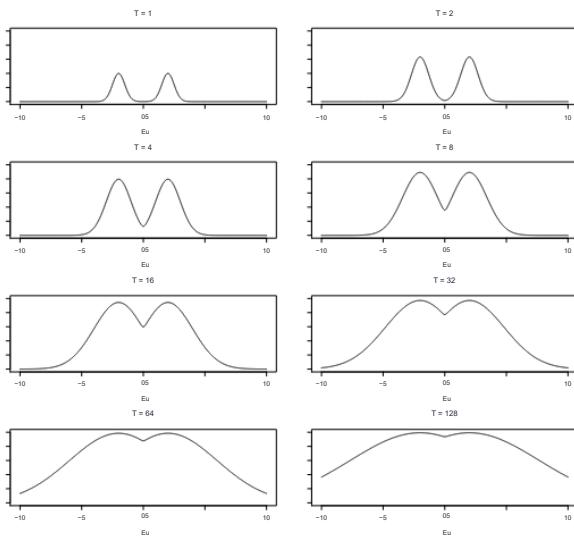


Figura 6.6: Densidades temperadas $\pi(\theta) 1/T$ para $T \in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

mistura dentro da cadeia visando a densidade de interesse. Na próxima seção, discutiremos como podemos tirar proveito dessa técnica em um cenário de inferência bayesiana.

6.5.2 Témpora paralela

Témpora paralela (ou MC3) é a noção de usar densidades temperadas dentro de um contexto de inferência bayesiana. Suponha que estejamos interessados em direcionar uma distribuição posterior (possivelmente) multimodal $\pi(\theta|x)$. Agora, em contraste com os esquemas MCMC padrão, construímos várias cadeias "Markov". Uma cadeia tem como alvo o verdadeiro $\pi(\theta|x)$ e as outras cadeias têm como alvo os posteriores temperados $\tilde{\pi}(\theta|x)$. As múltiplas cadeias são então acopladas a Metropolis por meio de uma proposta (Metropolis-Hastings) envolvendo trocas de espaço de estado entre as cadeias e para que amostras de cadeias de mistura melhores (visando os posteriores temperados) possam ser filtradas para a cadeia visando o verdadeiro posterior. Agora definimos formalmente os posteriores temperados e, em seguida, consideraremos o mecanismo proposto para as trocas de espaço de estados entre cadeias.

Lembre-se do Teorema de Bayes que o verdadeiro posterior é

$$w(\theta|x) \propto w(x|\theta)p(\theta)$$

e deixe o c^{th} temperado posterior ser

$$\tilde{\pi}_c(\theta_c|x) \propto \pi(x|\theta_c)1/T_c\pi(\theta)$$

onde $T_c > 1$ é a temperatura da cadeia c . Notamos que os posteriores temperados são formados apenas pela moderação do componente de verossimilhança do verdadeiro posterior, pois, dado que estamos trabalhando dentro do paradigma bayesiano, nossas crenças a priori devem ser consistentes, irrespectivas do posterior que estamos almejando.

A questão agora é como trocar os estados das cadeias sem afetar suas respectivas distribuições de destino. Suponha que temos cadeias C , isto é, uma cadeia para atingir o posterior verdadeiro e outras cadeias $C - 1$ visando densidades temperadas. Observando que $\tilde{\pi}_c(\theta_c|x) = \pi(\theta|x)$ quando $T_c = 1$ é útil deixar $T = (1, T_2, T_3, \dots, T_C)$ de modo que as cadeias que estão sendo consideradas sejam dadas por $\tilde{\pi}_c(\theta_c|x)$ para $c = 1, \dots, C$. Agora, se considerarmos que todas as cadeias C estão evoluindo juntas, segue-se que, como os posteriores $\tilde{\pi}_c(\theta_c|x)$ são condicionalmente independentes dado x , juntas as cadeias estão visando a articulação posterior

$$\pi(\theta_1, \dots, \theta_C | x) \prod_{c=1}^C \tilde{\pi}_c(\theta_c|x). \quad (6.13)$$

Como um breve aparte, notamos que o alvo conjunto (6.13) é preservado independentemente das atualizações dentro da cadeia dos parâmetros como consequência de π_i e π_j serem condicionalmente independentes dado x .

Pendentes. De fato, usar diferentes atualizações dentro da cadeia é uma estratégia sensata, dada a forma da densidade alvo é diferente para cada cadeia. Voltando à proposta de trocas de espaço de estado entre cadeias, deve ficar claro que estamos simplesmente propondo trocar θ_i e θ_j por alguns $i \neq j$ dentro do alvo conjunto (6.13). Seja $\theta = (\theta_1, \dots, \theta_C)$ denotam o estado atual da cadeia conjunta e, portanto, o estado proposto da cadeia é $\theta^* = (\theta^*_1, \dots, \theta^*_C)$ onde $\theta^*_i = \theta_j$, $\theta^*_j = \theta_i$ e $\theta^* = \theta$ para $i \neq j$. Assumindo um mecanismo de proposta simétrico, ou seja, que a probabilidade de propor a troca dos estados das cadeias (i, j) é a mesma que a probabilidade de propor a troca dos estados das cadeias (j, i) , a probabilidade de aceitação da troca do espaço de estados é $\min(1, A)$ onde

$$\begin{aligned}
 A &= \frac{\pi(\theta^*|x)\pi(\theta|x)}{\prod_{c=1}^C \frac{\pi(\theta+c|x)}{\pi(\theta|x)}} \\
 &= \frac{\pi(\theta^*_1|x)\pi(\theta_1|x)}{\theta_1\pi(\theta_1|x)} \cdot \frac{\pi(\theta^*_2|x)\pi(\theta_2|x)}{\theta_2\pi(\theta_2|x)} \cdots \frac{\pi(\theta^*_C|x)\pi(\theta_C|x)}{\theta_C\pi(\theta_C|x)} \\
 &= \frac{\pi(x|\theta^*)}{\prod_{i=1}^C \pi(x|\theta_i)} \cdot \frac{\pi(\theta^*)}{\prod_{i=1}^C \pi(\theta_i)} \quad (6.14)
 \end{aligned}$$

É claro que, se o mecanismo de proposta não for simétrico, a probabilidade A deve ser multiplicada pela razão de proposta $q(\theta | \theta^*) / q(\theta^* | \theta)$. Além disso, é simples generalizar o mecanismo de proposta acima para permitir cenários em que a proposta envolve a troca de parâmetros de mais de 2 cadeias. No entanto, tal deve ser evitado, uma vez que tal proposta pode ter baixas taxas de aceitação; isso será discutido mais adiante na Seção 6.5.4, onde consideraremos o ajuste das distribuições de propostas dentro de um esquema de amostragem MC3. Na próxima seção, consideraremos um esboço geral do algoritmo para um esquema de amostragem MC3 antes de discutir o ajuste e uma generalização adicional desse método.

6.5.3 Esboço geral do algoritmo

Um esboço geral do algoritmo de um esquema de amostragem de Monte Carlo da cadeia de Markov acoplada à metrópole usando cadeias C é o seguinte.

- Inicializar:— seja $T1 = 1$ e escolha $Tc > 1$ para $c = 2, \dots, C$ — seja $\theta = (\theta_1, \dots, \theta_C)$
- Execute repetidamente as seguintes etapas:

1. Para $c = 1, \dots, C$,

– extraia θ_c de $\pi_c(\theta_c|x) \propto \pi(x|\theta_c)1/T_c\pi(\theta)$ usando técnicas MCMC padrão

2. Faça uma amostra de um par de rótulos de cadeia (i, j) onde $1 \leq$

$i = j \leq C$ – seja $\theta_i \rightarrow \theta_j$ e $\theta_j \rightarrow \theta_i$ com probabilidade $\min(1, A)$ onde

$$A = \frac{\pi(x|\theta_j)1/T_{j|j}}{\theta_j1/T_j\pi(x|\theta_i)1/T_{i|i}\pi(\theta_i)1/T_j}$$

As realizações posteriores de interesse são θ_1 , ou seja, os sorteios aceitos da cadeia 1. Dado que cada uma das cadeias restantes tem como alvo uma densidade posterior temperada, os extrações dessas cadeias não são de interesse e, portanto, a coleta de amostras $(\theta_2, \dots, \theta_C)$ pode ser descartada.

6.5.4 Ajustando um esquema de amostragem MC3

Lembre-se de que os esquemas de Monte Carlo da cadeia de Markov acoplados à Metropolis consideram cadeias C ; cada um dos quais visa uma densidade alternativa (temperada). Além disso, como essas cadeias são condicionalmente independentes dado x , cada cadeia pode ser considerada um esquema MCMC padrão independente visando sua respectiva densidade. Segue-se que cada cadeia deve, portanto, ser ajustada de uma maneira típica, ou seja, conforme discutido na Seção 1.3.2. Note-se, no entanto, que como cada cadeia tem como alvo uma densidade alternativa, é sensato considerar diferentes distribuições de propostas para cada cadeia. Intuitivamente, esperamos que saltos maiores ao redor do espaço amostral sejam mais plausíveis para T grande, isso ocorre à medida que a densidade alvo se torna "mais plana" à medida que a temperatura da cadeia aumenta; veja a Seção 6.5.1. Ajustar a proposta entre cadeias (Etapa 2 do algoritmo MC3) pode ser complicado em geral. A taxa de aceitação das trocas entre cadeias não é afetada apenas pela escolha das temperaturas T_c , mas também pelo mecanismo que determina as cadeias entre as quais propor uma troca. Atchad'e et al. (2011) mostram que, ao visar uma densidade normal, a taxa de aceitação ideal (aquele que maximiza a distância de salto quadrada esperada) entre as cadeias é de 0,234. É claro que, em qualquer cenário realista, é improvável que tenhamos como alvo uma densidade normal usando MC3, pois técnicas mais padrão seriam suficientes para gerar amostras de um alvo normal. É geralmente aceito que entre as taxas de aceitação em cadeia de cerca de 20% a 60% fornecem uma mistura razoável (com relação à densidade conjunta de $\theta_1, \dots, \theta_C$); ver, por exemplo, Geyer e Thompson (1995) ou, mais recentemente, Altekaret et al. (2004). A questão é, portanto, como escolher as temperaturas e o mecanismo de troca para atingir essas taxas de aceitação. A estratégia que sugerimos, também defendida por Wilkinson (2013), é escolher as temperaturas de modo que $T_i < T_j$ para $i < j$ e considerar trocas entre cadeias adjacentes. A intuição por trás dessa estratégia é que as densidades-alvo se tornam cada vez mais diferentes à medida que $|T_i - T_j|$ aumenta e, portanto, é menos provável que os swaps

ser aceito. Além disso, as temperaturas devem ser escolhidas de modo que exibam espaçamento geométrico, ou seja, $T_c + 1 / T_c = r$ para algum $r > 1$. Observe que isso reduz o problema de escolher $C - 1$ temperaturas para o de escolher T_2 (ou equivalente mente r) como dado $T_1 = 1$ e T_2 as temperaturas restantes são completamente determinadas. Normalmente, é sensato realizar algumas execuções piloto do esquema MC3 para determinar uma proporção de temperatura adequada. Claro, esta é apenas uma estratégia geral e cada temperatura pode precisar de ajustes dependendo da situação.

6.5.5 Metrópole paralela acoplada cadeia de Markov Monte Carlo

Uma desvantagem notável do uso de esquemas de amostragem de Monte Carlo da cadeia de Markov acoplada ao Metropolis é o aumento significativo na quantidade de computação necessária em comparação com os amostradores MCMC padrão. Lembre-se de que um esquema de amostragem MC3 considera cadeias C ; cada um dos quais precisa ser atualizado para gerar uma única amostra a partir da densidade desejada (observe que, para esquemas MCMC padrão, apenas uma única cadeia precisa ser atualizada para gerar o mesmo número de amostras do alvo). É, portanto, claro que, para os esquemas MC3, a quantidade de computação necessária para gerar realizações (posteriore) aumenta linearmente com C . No entanto, embora não seja possível reduzir a quantidade de computação por se, é possível reduzir o tempo necessário para realizar esses cálculos apelando para a cadeia de Markov acoplada à metrópole paralela Monte Carlo (pMC3) – este método também é conhecido como cadeia de Markov acoplada a metrópole multi-core Monte Carlo no MC4 na literatura. A partir da Etapa 1 no algoritmo geral fornecido na Seção 6.5.3, deve ficar claro que o MC3 padrão considera a atualização de cada cadeia em série, ou seja, desenhamos sequencialmente θ_c para $c = 1, \dots, C$, de seus respectivos alvos π_c . Normalmente, esta etapa é a parte mais cara computacionalmente do algoritmo, pois envolve várias avaliações de verossimilhança (moderadas) por construção. No entanto, explorando a independência condicional das cadeias, podemos reduzir a quantidade de tempo necessária para executar esta etapa; esta é a ideia por trás da cadeia de Markov acoplada à Parallel Metropolis Monte Carlo (pMC3). Lembre-se de que π_i e π_j são condicionalmente independentes dado x e, portanto, segue-se que podemos realizar atualizações dentro da cadeia das cadeias direcionadas a essas densidades de forma independente, sem afetar a distribuição conjunta do alvo $\pi(\theta_1, \dots, \theta_C | x)$. Claramente, essas atualizações podem ser executadas em paralelo em vários núcleos. É importante notar, no entanto, que a Etapa 2 do algoritmo MC3 só pode ser executada depois que todas as cadeias individuais forem atualizadas, ou seja, devemos esperar até que a última cadeia seja atualizada antes de prosseguirmos. O número de núcleos a serem usados (`ncores`) é um fator importante ao usar o pMC3; especialmente de que todas as cadeias C devem ser atualizadas antes de prosseguirmos para a Etapa 2 do algoritmo MC3. Naturalmente, podemos pensar que aumentar o número de núcleos reduzirá a quantidade de tempo que leva para atualizar as cadeias C e, embora isso seja verdade em geral, não é bem assim

ncore	Núcleo			Hora	
	12	34	56	Total	Relativo
1	i1 —	— —	— —	-6 -	1
	i2 —	— —	— —	— -	
	i3 —	— —	— —	— -	
	i4 —	— —	— —	— -	
	i5 —	— —	— —	— -	
06	20102 —			3 1/2	
	i3 —	i4 —	— —		
	i5 —	i6 —	— —		
3	i1 040506	i2 401026394	i3 —	-2 —	1/3
	i5 —	i6 —	— —	2 1/3	
5	i1 06	i2 6010203640506	i3 —	i4 —	i5 -2
				1 1/3	
06	6010203640506 —				
				1 1/6	

Tabela 6.7: Alocação de trabalho ideal teórica entre núcleos com tempo de execução total e relativo (assumindo que o trabalho leva uma unidade de tempo).

tão simples. Para ver isso, consideraremos um exemplo: suponha que escolhamos usar $C = 6$ cadeias dentro de um esquema de amostragem pMC3 e, portanto, devemos atualizar 01, ..., 06 em cada iteração. Se assumirmos que a atualização de 0c leva tempo unitário, o tempo total necessário para atualizar todas as cadeias de Markov $C = 6$ para diferentes valores de ncore pode ser visto na Tabela 6.7. Claro, este é um cronograma de trabalho teoricamente ideal que não é alcançável na prática (devido a sobrecargas adicionais, como alocação de memória e inicialização de núcleos / trabalhos), no entanto, ilustra a ideia geral. Observe que quando ncore = 1 pMC3 é o mesmo que o padrão MC3, ou seja, cada cadeia é atualizada em série. A partir da Tabela 6.7, fica claro que podemos reduzir pela metade o tempo necessário para atualizar as cadeias C simplesmente considerando dois núcleos. Além disso, para este exemplo, tomando ncore = 3, reduzimos a etapa de atualização para um terço do tempo em comparação com o MC3 padrão. Observe, no entanto, que não melhoramos isso se ncore = 4, 5, pois embora possamos atualizar mais cadeias na primeira unidade de tempo, isso é de pouco benefício, pois toda a etapa de atualização só é concluída quando todas as cadeias C são atualizadas. Claramente, tomar ncore > C não proporcionará ganhos adicionais. Segue-se que, em geral, a melhor estratégia é escolher ncore = C, no entanto, isso pode não ser possível quando se considera um grande número de cadeias. Neste cenário é sensato escolher o maior número possível decores ncore que também é um divisor de C, ou seja, escolher o maior ncore possível para que $C \bmod ncore = 0$. Altekar et al. (2004) fornecem mais detalhes computacionais e fornecem ganhos de desempenho relativos mais realistas de pMC3 (contra MC3) em oposição aos ganhos teóricos considerados aqui.

6.6 Inferência – uma abordagem bayesiana (revisitada)

Voltamos agora à construção de um esquema de inferência bayesiana que é capaz de gerar realizações posteriores dos parâmetros de ordem de habilidade e escolha, implementando um algoritmo de Monte Carlo de cadeia de Markov acoplado a metrópole par-allel para direcionar a distribuição posterior conjunta. Esta seção será concluída com um estudo de simulação onde mostramos que a mistura de nossas cadeias de Markov é melhorada consideravelmente usando MC3 em oposição às técnicas MCMC padrão desenvolvidas na Seção 6.4.

6.6.1 Um algoritmo pMC3 para o modelo EPL

Um algoritmo de Monte Carlo de cadeia de Markov acoplado a Metropolis paralelo para direcionar a distribuição conjunta-posterior dos parâmetros de habilidade λ e o parâmetro de ordem de escolha σ é o seguinte.

1. Afinação:

- escolha o número de cadeias (C) e o número de núcleos ($ncores$)
- seja $T1 = 1$ e escolha $Tc > 1$ para $c = 1, \dots, C$

2. Inicializar: escolha $\sigma_0 \in SK$ e $\lambda_0 \in RK > 0$ para $c = 1, \dots, C$

3. Para $c = 1, \dots, C$ executa, paralelamente, as seguintes etapas:

- Para $k = 1, \dots, K$ – desenhe $\lambda_{tck} | \lambda_{ck} \sim LN(\log \lambda_{ck}, \sigma^2 \lambda_{ck})$ – seja $\lambda_{ck} \rightarrow \lambda_{tck}$ com probabilidade $\min(1, A)$ onde
$$A = \frac{\{ \frac{\pi(D|\lambda_c, -k, \lambda_{ck} = \lambda_{tck}, \sigma_{ck})}{\sigma_{ck} \pi(D|\lambda_c, \sigma_c)} \}^{1/Tc}}{\lambda_{ck}^{(Tc)}} \times e^{(\lambda_{tck} - \lambda_{ck})}$$
- Amostra ' da distribuição discreta com probabilidades $Pr(' = i) = p_{prop,i,c}$ durante $i = 1, \dots, 5$
 - propor σ_{tck} usando o mecanismo de proposta ' –
 - seja $\sigma_c \rightarrow \sigma_{tck}$ com probabilidade $\min(1, A)$ onde
$$A = \frac{\{ \frac{\pi(D|\lambda_c, \sigma_{tck})}{\sigma_{tck} \pi(D|\lambda_c, \sigma_c)} \}^{1/Tc}}{\sigma_{tck}^{(Tc)}} \times \frac{Pr(\sigma = \sigma_{tck})}{Pr(\sigma = \sigma_c)}$$

- Redimensionar – amostra
 $\lambda_{ik} \sim Ga(\frac{1}{K}, \sum_{k=1}^K \lambda_{ik})$.

– calcular $\Sigma c = \sum_{k=1}^K \lambda_{ik}$.

– seja $\lambda_{ik} \rightarrow \lambda_{ik} \lambda_{ik}/\Sigma c$ para $k = 1, \dots, K$.

4. Amostra de um par de rótulos de cadeia (i, j) onde $1 \leq i \neq j \leq C$

- seja $(\lambda_i, \sigma_i) \rightarrow (\lambda_j, \sigma_j) \rightarrow (\lambda_i, \sigma_i)$ com probabilidade $\min(1, A)$ onde

$$A = \frac{\pi(D|\lambda_j, \sigma_j)1/T_{jpo}(D|\lambda_j, \sigma_j)}{\pi(D|\lambda_i, \sigma_i)1/T_j\pi(D|\lambda_i, \sigma_i)1/T_{jpo}(D|\lambda_j, \sigma_j)1/T_j}$$

5. Retorne à Etapa 3.

6.6.2 Estudo de simulação

Para este estudo, revisitamos novamente o conjunto de dados 6, que analisamos anteriormente usando técnicas MCMC padrão na Seção 6.4. Lembre-se de que esses dados compreendem $n = 100$ classificações completas de $K = 10$ entidades. Os parâmetros de habilidade e a ordem de escolha (processo de classificação) usados para simular esses dados são $\lambda = (10, 9, \dots, 1)$ e $\sigma = (3, 10, 9, 1, 7, 4, 5, 6, 8, 2)$, respectivamente. Aqui, realizamos inferência bayesiana usando o algoritmo pMC3 descrito na Seção 6.6.1 com $C = 5$ cadeias rodando em ncores = $C = 5$ núcleos. Adotamos a mesma distribuição a priori das análises anteriores desses dados, ou seja, $a_k = a = 1$ e cada ordem de escolha é escolhida para ser igualmente provável a priori. A motivação subjacente a esta escolha de distribuição a priori é dada nas seções 6.4.7 e 6.4.10. Além disso, a distribuição da proposta para o parâmetro de ordem de escolha é escolhida para ser a mesma de antes; ou seja, $s = 2$, $\tau = 3$ e $pprop = (0.3, 0.3, 0.3, 0.05, 0.05)$ dentro de cada cadeia. Antes de podermos realizar a inferência bayesiana usando o algoritmo pMC3 da Seção 6.6.1, também devemos escolher as temperaturas apropriadas para cada cadeia, juntamente com um mecanismo adequado para propor as trocas entre cadeias. Escolhemos as temperaturas como $T = (1, 0.75, 1, 0.6, 1, 0.5, 1, 0.4, 1)$ que foram determinadas após fazer alguns ajustes manuais (guiados por execuções piloto) para as temperaturas obtidas usando espaçamento geométrico com razão 4/3, ou seja, $T_{c+1}/T_c = 4/3$. Além disso, conforme discutido na Seção 6.5.4, consideramos sensato considerar apenas as trocas entre duas cadeias adjacentes e amostramos os rótulos das cadeias $(i, i + 1)$ uniformemente ao acaso. As taxas de aceitação entre cadeias resultantes são de cerca de 50% a 60%, o que consideramos aceitável e deve nos permitir nos beneficiar do uso de um esquema MC3. Claro, também devemos escolher um parâmetro de ajuste para a proposta Log-normalrandom walk que gera os parâmetros de habilidade propostos. Como no anterior

Análises: escolhemos $\sigma\lambda = \sigma\lambda = 0,75$, o que dá taxas de aceitação em torno de 19% a 25% (dentro de cada cadeia).

Neste estudo, inicializamos cada uma das cadeias C (dentro de nosso esquema de inferência única) a partir da distribuição anterior, ou seja, extraímos λ_{CK} indep~ $\text{Ga}(1, 1)$ e amostramos uniformemente c_k do conjunto de todas as permutações SK para $c = 1, \dots, C$ e $k = 1, \dots, K$. Os resultados relatados são de uma execução típica de nosso esquema pMC3 inicializado conforme descrito, com um burn-in de 20K iterações e, em seguida, executado por mais 2M iterações e diluído em 200 para obter 10K (quase) amostras posteriores não autocorrelacionadas. O esquema pMC3 é executado razoavelmente rápido, com código C em ncores = C = 5 núcleos de uma CPU Intel Core i7-4790S (velocidade de clock de 3,20 GHz) levando cerca de 35 minutos e 20 segundos. Observe que a análise equivalente leva cerca de 123 minutos e 45 segundos em uma implementação MC3 padrão (serial) (ncores = 1) e, portanto, claramente o uso da computação paralela é vantajoso. Além disso, devido às avaliações adicionais de probabilidade necessárias para propor as trocas entre cadeias, o esquema pMC3 requer mais tempo de CPU do que o esquema MCMC Metropolis-Hastings padrão da Seção 6.4.10, onde a análise equivalente levou aproximadamente 24 minutos e 30 segundos.

A Figura 6.7 mostra um gráfico de rastreamento da log-verossimilhança (esquerda) e também a distribuição posterior marginal da ordem de escolha σ (direita). Aqui, em contraste com as análises bayesianas anteriores, a distribuição posterior marginal para σ obtida de vários esquemas de pMC3 (inicializados em valores diferentes) são consistentes até o ruído estocástico (não relatado aqui) e, portanto, para este exemplo, MC3 nos permitiu gerar realizações posteriores. A Figura 6.8 mostra as 25 ordens de escolha com maior suporte posterior e suas probabilidades posteriores correspondentes. Observe que a ordem de escolha usada para simular esses dados é mostrada em vermelho e tem probabilidade posterior $\Pr(\sigma = (3, 10, 9, 1, 7, 4, 5, 6, 8, 2)|D) = 0,011$. Talvez não seja surpreendente que não vejamos um grande suporte posterior para a ordem de escolha usada para simular esses dados (ou mesmo qualquer ordem de escolha), uma vez que estamos considerando apenas $n = 100$ observações e há $10!$ possíveis pedidos de escolhas. Dito isso, é agradável ver que as ordens de escolha com maior suporte posterior são bastante semelhantes àquela que foi usada para simular esses dados. Outra observação interessante da Figura 6.8 é que parece que podemos identificar claramente as posições mais baixas dentro da ordem de escolha e grande parte da incerteza reside nas primeiras 4 entradas de σ .

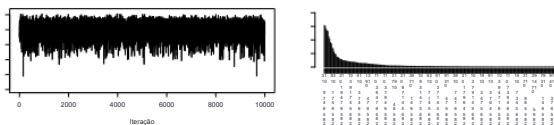


Figura 6.7: Traçado da log-verossimilhança (esquerda) e do $\pi(\sigma|D)$ (direita).

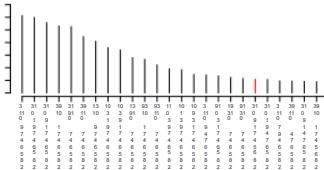


Figura 6.8: Subconjunto do π marginal posterior ($\sigma|D$) mostrando as 25 ordens de escolha com maior suporte posterior (vermelho denota ordem de escolha usada para simular esses dados)

A Figura 6.9 mostra os boxplots da distribuição marginal posterior para cada condição logarítmica λ_k tanto na ordem de escolha modal posterior ($\sigma = (3, 1, 10, 9, 7, 4, 6, 5, 8, 2)$) quanto na ordem de escolha "verdadeira" ($\sigma = (3, 10, 9, 1, 7, 4, 5, 6, 8, 2)$) em branco e vermelho, respectivamente. As cruzes azuis denotam os valores verdadeiros a partir dos quais esses dados foram simulados e notamos que redimensionamos os valores para que $\lambda_{10} = 1$ seja constante e, portanto, omitido do gráfico junto com quaisquer outliers. Claramente, há uma grande incerteza posterior sobre os valores dos parâmetros de habilidade, no entanto, isso talvez não seja surpreendente, dada a incerteza associada à ordem de escolha (e o pequeno número de observações). Além disso, esses boxplots são construídos com base em um número relativamente pequeno de realizações (posteriore) devido à condição de uma determinada ordem de escolha. No entanto, é agradável ver que o marginalposterior para os parâmetros de habilidade são coerentes sob ambas as ordens de escolha selecionadas e as classificações agregadas (formadas pela ordenação dos parâmetros de habilidade em sua média posterior) são $xagg = (4, 3, 1, 2, 5, 6, 8, 7, 9, 10)$ sob a permutação modal posterior e $xagg = (2, 1, 3, 4, 5, 6, 7, 8, 9, 10)$ sob a verdadeira permutação - lembre-se de que a verdadeira preferência das entidades é $(1, \dots, 10)$.

Também é interessante ver se há algum benefício em usar o complicado modelo EPL e se, em vez disso, podemos fazer inferências razoáveis sobre esses dados usando o modelo padrão (classificação futura) de Plackett-Luce. Para responder a esta pergunta, analisamos esses dados usando o algoritmo (amostragem de Gibbs) para o modelo PL padrão do Capítulo 2. A Figura 6.10 mostra boxplots da distribuição posterior marginal para cada $\log \lambda_k$ sob a ordem de escolha superior do modelo EPL e aqueles obtidos sob o PL padrão em branco e verde, respectivamente. A partir disso, fica claro que o modelo padrão de Plackett-Luce não é adequado para esses dados e, sob tal análise, concluiríamos que a classificação agregada é $xagg = (2, 8, 5, 7, 4, 9, 6, 1, 3, 10)$ e há pouca discrepância entre a preferência das entidades - o que sabemos não ser verdade. Segue-se que o modelo padrão de Plackett-Luce não é um bom modelo para esses dados e devemos esperar obter inferências "melhores" ao usar o EPL se os dados não forem gerados a partir do processo de classificação futura.

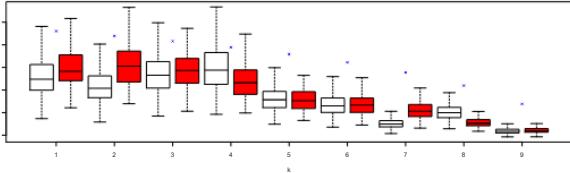


Figura 6.9: Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{10} = 1$. As densidades em cada caso são mostradas em branco e vermelho para aquelas obtidas sob a ordem de escolha com o maior suporte posterior e a ordem de escolha verdadeira, respectivamente. As cruzes azuis representam os valores reais a partir dos quais esses dados foram simulados (escala logarítmica).

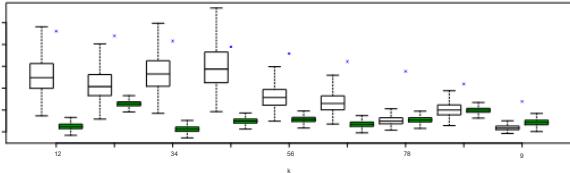


Figura 6.10: Boxplots resumindo as densidades marginais posteriores para cada $\log \lambda_k$, dado que $\lambda_{10} = 1$. As densidades em cada caso são mostradas em branco e verde para aquelas obtidas sob o modelo EPL (para a ordem de escolha com o maior suporte posterior) e aquelas obtidas sob o modelo padrão de Plackett-Luce ($\sigma = (1, \dots, K)$), respectivamente. As cruzes azuis representam os valores verdadeiros a partir dos quais esses dados foram simulados (escala logarítmica).

6.7 Resumo

Neste capítulo, descrevemos o modelo de Plackett-Luce estendido, que permite que a suposição a priori de uma ordem de escolha explícita (processo de classificação) seja relaxada. A ordem de escolha, que é um elemento do conjunto de todas as permutações SK , é representada por um parâmetro livre adicional dentro do modelo. Descrevemos o modelo proposto em detalhes e consideramos se a ordem de escolha era identificável dada uma coleção de classificações. Na Seção 6.3, verificamos que cada ordem de escolha alternativa resulta em alterações na verossimilhança (máxima) por meio de um estudo de simulação usando estimativa de máxima verossimilhança.

O restante deste capítulo se concentrou no problema desafiador de realizar uma análise totalmente bayesiana do modelo de Plackett-Luce estendido. Na secção 6.4.4, o Tribunal considerou

numerosos movimentos de "troca" que esperávamos que nos permitissem explorar efetivamente o conjunto de todas as permutações quando o número de entidades é grande. Infelizmente, a abordagem de aumento de dados (para dar atualizações a Gibbs dos parâmetros de habilidade) foi ineficaz, o que acreditamos ser causado pela grande correlação posterior entre os parâmetros dentro desse modelo. Para superar isso, consideramos uma abordagem de amostragem Metropolis-Hastings e vimos por meio de estudos de simulação na Seção 6.4.10 que, embora essa abordagem tenha se mostrado promissora, a mistura das cadeias de Markov permaneceu pobre. O passo natural foi então considerar a cadeia de Markov acoplada a Metropolis Monte Carlo (MC3) e na Seção 6.5 passamos um tempo explorando essa ideia e demos um esboço genérico do algoritmo e discutimos como a computação paralela pode ser usada para aumentar a velocidade dos esquemas de inferência. Na Seção 6.6.2, consideramos um estudo de simulação usando um esquema de amostragem pMC3 que deu resultados promissores e a cadeia de Markov pareceu explorar efetivamente a parte posterior. Dito isso, trabalhos futuros são necessários para verificar se essa abordagem também funcionará para conjuntos de dados maiores - nossa intuição nos leva a acreditar que o problema de inferência para o EPL pode se tornar muito desafiador para mais do que, digamos, $K = 20$ entidades. Seria igualmente bom aumentar a flexibilidade da modelização apelando para as misturas de processos de Dirichlet de modelos EPL, mas isso também aumentará a dificuldade da amostragem posterior.

No próximo capítulo, concluiremos esta tese e fornecemos uma visão geral dos principais resultados, juntamente com algumas direções potenciais para trabalhos futuros.

Capítulo 7

Conclusões

A intenção desta tese foi explorar modelos flexíveis que permitam a identificação de (possível) estrutura de subgrupos dentro de dados ranqueados. Além disso, queríamos que nossa estrutura de modelagem fosse capaz de capturar a heterogeneidade potencial entre as habilidades dos classificadores. Também investigamos o efeito do processo de classificação e desenvolvemos métodos para aumentar ainda mais a flexibilidade da modelagem, relaxando a suposição de um processo de classificação explícito.

Consideramos primeiro o modelo (padrão) de Plackett-Luce (Luce, 1959; Plackett, 1975). Por meio de estudos de simulação, foi mostrado como as inferências desse modelo podem ser afetadas até mesmo por uma quantidade modesta de classificações espúrias. Segue-se que este modelo não é adequado para lidar com um cenário em que alguns dos classificadores não estão bem informados sobre as entidades que estão sendo classificadas. Estendemos o modelo padrão de Plackett-Luce para o novo modelo WeightedPlackett-Luce (WPL), onde o modelo WPL permite confiabilidade (potencial) diferente por meio de um modelo de mistura de dois componentes. A inferência bayesiana para este modelo é simplificada por uma ligeira extensão das abordagens de aumento de dados existentes (Caron e Doucet, 2012) que produz um esquema de amostragem de Gibbs eficiente. Estudos de simulação mostraram que o modelo Weighted Plackett-Luce é capaz de identificar corretamente classificadores espúrios (quando presentes) dentro de uma coleção de dados. Além disso, em contraste com o modelo padrão de Plackett-Luce, as inferências do modelo WPL mostraram-se bastante robustas para a adição de classificações não informativas.

No Capítulo 3, apresentamos modelos capazes de explorar a (possível) estrutura de subgrupos dentro de dados classificados. Mais especificamente, nosso objetivo foi identificar grupos homogêneos de indivíduos que compartilham crenças semelhantes e também analisamos como um único grupo de classificadores pode ter dificuldade em distinguir entre certas entidades. Para implementar essa estrutura, apelamos para métodos de agrupamento não paramétricos bayesianos, especificamente usando modelos de mistura de processos de Dirichlet (Ferguson, 1973; Antoniak, 1974). Apresentamos dois modelos. O primeiro al-

lowed para a noção de que os rankers podem ser heterogêneos em suas crenças sobre as entidades. As misturas de processos finitos e de Dirichlet dos modelos de Plackett-Luce (padrão) foram exploradas na literatura e estendemos ligeiramente essa abordagem construindo um modelo que compreende uma mistura de processos de Dirichlet de modelos de Plackett-Luce ponderados. O segundo modelo que apresentamos permite a noção de que um grupo homogêneo de classificadores pode não ser capaz de distinguir entre certos grupos de entidades, ou seja, eles podem considerar algumas entidades indistinguíveis (amarradas em força). Permitimos essa estrutura considerando uma mistura de processos de Dirichlet sobre os parâmetros de habilidade no modelo Weighted Plackett-Luce. Até onde sabemos, essa abordagem não foi considerada anteriormente na literatura. Estudos de simulação mostraram que este modelo provou ser eficaz na detecção de grupos de entidades e também teve um desempenho razoavelmente bom quando não estavam presentes grupos de entidades, o que foi tranquilizador. Além disso, esse modelo nos permitiu quantificar o nível de similaridade (posterior) entre as entidades de maneira baseada em princípios; Isso exigiria uma abordagem ad hoc se usasse técnicas padrão (sem agrupamento). A inferência bayesiana para cada modelo prossegue por meio de esquemas eficientes de amostragem marginal (Neal, 2000).

O Capítulo 4 apresentou a mistura do processo Weighted Adapted Nested Dirichlet (WAND) dos modelos de Plackett-Luce. Este modelo combina os aspectos de cada modelo apresentados nos capítulos anteriores. Mais especificamente, esse modelo permite o agrupamento de classificadores e entidades, juntamente com a possibilidade de (potencial) confiabilidade de classificadores diferentes. A permissão para o agrupamento de classificadores e entidades foi alcançada apelando para técnicas de agrupamento bidirecional; especificamente adaptando o processo Nested Dirichlet anterior de Rodriguez et al. (2008) para que pudesse lidar com dados classificados. O modelo (WAND) foi então formado tomando o processo Nested Dirichlet adaptado como a distribuição anterior sobre os parâmetros de habilidade e o modelo Plackett-Luce ponderado como a distribuição de classificação. Abordagens condicionais e marginais para inferência posterior foram apresentadas para este modelo. A estrutura de modelagem descrita também permite que inferências sejam feitas usando apenas classificações incompletas, como classificações top-M ou parciais. Vimos através dos estudos de simulação que inferências razoáveis podem ser feitas sob o modelo WAND, mesmo quando apenas informações limitadas (parciais) estão disponíveis. Embora não seja considerado aqui, na Seção 2.2.4 descrevemos como os laços dentro dos rankings podem ser facilmente explicados em nossa abordagem de inferência baseada em simulação. A riqueza de informações na distribuição posterior nos permite inferir muitos detalhes da estrutura tanto entre grupos de rankers quanto entre grupos de entidades (dentro de grupos de rankers), em contraste com muitas outras análises (bayesianas). A alta dimensionalidade da distribuição posterior pode dificultar bastante a produção de resumos perspicazes, mas simples, e exploramos diferentes abordagens, que vão desde o condicionamento do número modal de grupos até a adoção de uma classificação baseada em cálculos de um resumo de matriz de dissimilaridade. O Capítulo 5 continha análises de vários conjuntos de dados reais que foram analisados na literatura, e compararamos suas conclusões com as obtidas do ajuste do

Modelo de WAND. Em geral, descobrimos que o modelo WAND é adequado para modelar dados classificados e fornece informações valiosas sobre a estrutura de subgrupos dentro de dados classificados que não seriam possíveis em outros modelos.

Também consideramos relaxar a suposição de um processo de classificação conhecido, observando o modelo Extended Plackett-Luce recentemente desenvolvido (Mollica e Tardella, 2014). Neste modelo, o processo de classificação ("ordem de escolha") é representado por um parâmetro additionalfree que é um elemento do conjunto de todas as permutações SK . Algumas informações sobre o modelo foram fornecidas e também discutimos a identificabilidade do processo de classificação. Para motivar ainda mais a identificabilidade do processo de classificação, consideramos um estudo de simulação que mostrou que ordens de escolha alternativas resultam em mudanças na probabilidade (maximizada). Em seguida, consideramos o problema desafiador de realizar inferência bayesiana para o modelo de Plackett-Luce estendido. Nosso objetivo foi estender a solução de Mollica e Tardella (2018), considerando um espaço amostral irrestrito para o parâmetro de ordem de escolha. Apresentamos vários movimentos de "troca" que esperávamos que nos permitissem explorar efetivamente o conjunto de todas as permutações quando o número de entidades é grande. Infelizmente, a abordagem de aumento de dados (para fornecer atualizações de Gibbs dos parâmetros de habilidade) foi ineficaz e acreditamos que isso foi causado pela grande correlação posterior entre os parâmetros dentro deste modelo. Na tentativa de superar esse problema, removemos os parâmetros latentes do modelo e, em vez disso, consideramos uma abordagem de amostragem de Metropolis-Hastings e, embora essa abordagem tenha se mostrado promissora, ficou evidente (por meio de estudos de simulação) que a mistura das cadeias de Markov permaneceu pobre. Para melhorar a mistura, recorremos à cadeia de Markov acoplada Metropolis Monte Carlo (MC3). Um estudo de simulação usando um esquema de amostragem MC3 (paralelo) deu resultados promissores e parecia que somos capazes de explorar a distribuição posterior de forma eficaz. Observamos que deve-se ter cuidado ao realizar a inferência bayesiana para a EPL, pois é improvável que nossa solução seja bem dimensionada para cenários em que o número de entidades é grande.

7.1 Trabalho futuro

Acreditamos que esta pesquisa oferece muitas oportunidades para extensão e trabalho futuro dentro da análise bayesiana de dados classificados. Por exemplo, pode ser possível remover pesos de classificador binário w_i do modelo WAND e, em vez disso, introduzir um "cluster de spam" para abrigar os classificadores não informativos. Na configuração de mistura finita, isso seria razoavelmente simples de alcançar. No entanto, para modelos de mistura infinita, é um pouco mais complicado. Para ser equivalente a uma especificação de $w_i = 0$, o cluster de spam precisaria conter apenas um único cluster de entidade e é improvável que isso ocorra, a menos que o parâmetro concentration da distribuição de mistura seja escolhido para ser pequeno. Claro, não queremos

O parâmetro de concentração (agrupamento de entidades) deve ser pequeno em todos os grupos de classificadores, pois ainda gostaríamos de aprender sobre a estrutura do subgrupo de entidades para grupos de classificadores que são informativos. Uma estratégia poderia ser escolher a distribuição prévia dos parâmetros de concentração como uma mistura em que um dos componentes está fortemente concentrado em pequenos valores. No entanto, nossa intuição nos leva a acreditar que essa abordagem pode resultar em vários clusters de "spam", tornando a identificação de classificadores não informativos um pouco mais complicada. Portanto, pode ser mais vantajoso lidar com esse aspecto com opesos de classificação binária. No entanto, isso merece mais pesquisas.

Embora nosso trabalho tenha se concentrado principalmente no modelo (ponderado) de Plackett-Luce, o ANDPprior também é adequado para outros modelos de classificação paramétrica. Um modelo natural a considerar é o modelo de Benter (Benter, 1994), uma vez que é uma extensão direta do modelo PL (padrão). O modelo de Benter possui parâmetros adicionais que representam a "importância" de cada etapa do processo de classificação. Em teoria, também seria possível introduzir nossos indicadores de classificação binária e, portanto, considerar um modelo de Benter ponderado. No entanto, isso também pode dar origem a problemas identificáveis. Além disso, dado que o processo de classificação futura implica implicitamente que há mais incerteza nas primeiras fileiras de uma observação, é improvável que o parâmetro "importância da posição" dentro do modelo de Benter seja capaz de lidar adequadamente com um cenário em que esse não seja o caso, ou seja, quando há mais certeza sobre as entidades nas fileiras mais altas do que aquelas nas últimas fileiras. No entanto, seria interessante ver se o modelo Benter é capaz de mitigar esse artefato do processo de classificação futura.

Nossa exploração do modelo Extended Plackett-Luce abriu muitos caminhos potenciais para pesquisas futuras. O problema de inferência associado é desafiador e, embora nosso esquema de inferência pareça ser adequado para um número modesto de entidades, é improvável que tenha um bom desempenho em cenários em que o número de entidades é grande. Pode ser possível evitar considerar explicitamente o conjunto de todas as permutações e, em vez disso, considerar um parâmetro contínuo (multivariado) que se encontra no simplex ($K - 1$) dimensional (com a permutação implícita sendo dada ordenando a realização). A construção de uma distribuição de propostas adequada provavelmente será mais direta neste cenário e, embora a multimodalidade ainda possa ser um problema, pode ser possível superá-la apelando para o Monte Carlo hamiltoniano (Duane et al., 1987; Neal, 2011). Uma vez que o problema de inferência tenha sido explorado mais a fundo, seria interessante estender a flexibilidade de modelagem considerando misturas finitas ou infinitas de modelos de Plackett estendido. Observamos, no entanto, que o ANDP anterior provavelmente não é adequado para esse cenário, pois o agrupamento de entidades pode tornar o processo de classificação não identificável; Mais trabalho analítico / empírico seria necessário para verificar isso. Finalmente, um tópico instigante de pesquisa potencial é a noção de um modelo de "Plackett-Luce hierárquico". Em vez de considerar o parâmetro do processo de classificação no modelo de Plackett-Luce estendido como uma permutação, poderíamos considerá-lo

para ser um ranking em si. Esta classificação (o processo de classificação) poderia então ser modelizada utilizando um modelo de Plackett-Luce e, portanto, cada posição no processo de classificação teria um parâmetro de competência correspondente. Se desejado, o processo de classificação também poderia ser modelado por outro modelo de Plackett-Luce estendido e, portanto, as camadas aninhadas dos modelos de Plackett-Luce poderiam continuar indefinidamente. . .

Apêndice A

Variado

A.1 Derivação do FCD para o parâmetro de concentração DP α

Central para a implementação de um modelo de mistura de processo de Dirichlet é a escolha do parâmetro de concentração α . Muitas vezes, desejamos inferir esse parâmetro a partir dos dados, o que é possível se incorporarmos α em nossa análise. Supondo que temos n amostras e uma densidade contínua $\pi(\alpha)$, foi demonstrado por Antoniak (1974) que o prévio implícito no número de clusters N_c é

$$\frac{\pi(N_c|\alpha, n)}{a^{N_c}} = c_n(N_c)! \frac{C(\alpha)^{N_c}}{(n + N_c)^{N_c}} \quad (A.1)$$

para $N_c = 1, \dots, n$. A função $c_n(N_c)$ é definida como a densidade do número de clusters condicionais $\text{em } \alpha = 1$, ou seja, $c_n(N_c) = \pi(N_c|\alpha = 1, n)$ e, portanto, não envolve α . Se supormos que amostramos nossos parâmetros Λ , então deve ficar claro que também obtivemos uma amostra de N_c (o número de clusters) que é dada pelo número de entradas únicas dentro de Λ . Além disso, também assumimos que temos uma amostra de nossos indicadores de cluster latentes, c , e, portanto, a configuração desses dados nos grupos N_c também é conhecida. Agora, dado que N_c , Λ e c são conhecidos, pode-se mostrar que os dados D são condicionalmente independentes de α . Daqui resulta que, para $\alpha > 0$

$$p(a|N_c, \Lambda, D) = \pi(\alpha|N_c) \propto p(N_c|\alpha)p(\alpha).$$

e, substituindo em (A.1) dá

$$\begin{aligned}
 p(a|N c, \Lambda, D) &\propto cn(N c)n! \frac{C(a)C(a+n)}{p(a)} \\
 &\propto \alpha N c C(a)C(a+n) p(a) = \\
 &\propto \alpha N c (\alpha + n) B(\alpha + 1, \\
 &\quad n) \alpha \Gamma(n) \frac{\int_0^1 (1-x)^{n-1} dx}{p(a)} \pi(\alpha). \tag{A.2}
 \end{aligned}$$

Felizmente, podemos construir uma distribuição conjunta de α e uma quantidade $\eta \in (0, 1)$ que tem (A.2) como sua densidade marginal. Essa densidade de junta assume a forma, para $\alpha > 0$ e $\eta \in (0, 1)$

$$p(a, h|N c) \propto \alpha N c - 1 (\alpha + n) \eta \alpha (1 - h)^{n-1} \pi(\alpha). \tag{A.3}$$

Componente único anterior

Para manter a conjugação, atribuímos α uma mistura de distribuições Gama a priori. Se considerarmos primeiro o caso simples em que há apenas um único componente de mistura, ou seja, $\alpha \sim Ga(a_0, b_0)$, segue-se que a densidade conjunta de α e η é, para $\alpha > 0$ e $\eta \in (0, 1)$

$$\begin{aligned}
 p(a, h|N c) &\propto \alpha N c - 1 (\alpha + n) \eta \alpha (1 - h)^{n-1} ba00C(a_0) \\
 &\propto \alpha a_0 - 1 e^{-\alpha b_0} \\
 &\quad = ba00C(a_0) \alpha a_0 + N c - 2 e^{-\alpha b_0} (\alpha + n) \eta \alpha (1 - h)^{n-1}. \tag{A.4}
 \end{aligned}$$

De (A.4) a distribuição marginal para α é, para $\alpha > 0$,

$$\begin{aligned}
 \pi(\alpha|\eta, N c) &\propto ba00C(a_0) \\
 &\propto \alpha a_0 + N c - 2 e^{-\alpha b_0} (\alpha + n) \eta \alpha \\
 &\propto ba00C(a_0) \alpha a_0 + N c - 2 e^{-\alpha(b_0 - \log h)} (\alpha + n) = ba00C(a_0) \\
 &\quad \alpha a_0 + N c - 1 e^{-\alpha(b_0 - \log h)} + n ba00C(a_0) \\
 &\quad \alpha a_0 + N c - 2 e^{-\alpha(b_0 - \log h)},
 \end{aligned}$$

e assim temos

$$\alpha | \dots \sim \pi 0 \text{ Ga}(a_0 + N c, b_0 - \log h) + (1 - \pi 0) \text{ Ga}(a_0 + N c - 1, b_0 - \log h).$$

O peso da mistura π_0 é dado por

$$\begin{aligned}\pi_0 &= \frac{\int_{\infty}^{\infty} \alpha a_0 + Nc - 1 - e^{-\alpha(b_0 - \log h)}}{ba_0 C(a_0)} \\ \Rightarrow \pi_0 &= \frac{C(a_0 + Nc)(b_0 - \log h)a_0 + Nc - 1}{ba_0 C(a_0)}\end{aligned}\quad (\text{A.5})$$

e

$$\begin{aligned}(1 - \pi_0) &= \frac{\int_0^{\infty} \alpha a_0 + Nc - 2 - e^{-\alpha(b_0 - \log h)}}{nba_0 C(a_0)} \\ \Rightarrow (1 - \pi_0) &= \frac{C(a_0 + Nc - 1)(b_0 - \log h)a_0 + Nc - 1}{nba_0 C(a_0)}\end{aligned}\quad (\text{A.6})$$

de onde

$$\frac{\pi_0(1 - \pi_0)}{1n(b_0 - \log h)} = \frac{a_0 + Nc - 1}{1n(b_0 - \log h)}\quad (\text{A.7})$$

De (A.4) também deduzimos a distribuição condicional para h ser, para $n \in (0, 1)$

$$p(h|\cdot) \propto h^{a_0}(1 - h)^{Nc - 1},$$

Isto é

$$h|\cdot \sim \text{Beta}(a_0 + 1, Nc)\quad (\text{A.8})$$

Apêndice B

Datasets

Classificação 41	Classificador									
	50
1	16	5	8	10	20	9	19	8	2	15
223	11	139	10	14	16	2093188	12	15		
19252	14341	126	118	1199	10256	1013				
10 29	1364	18642	194	1378	20	13	207849			
142	91	11	19789	15	1864	16	15	177		
14919	20	1395	13	12	19611017741	154				
2043	1611139	20	20	18	18	11	12	1581215		
19 18	1861	17	155	1713317	1738	18	18			
17 18	141163	12	17	12	107	12	111510	14		
175734	1494167	115	16	12	177	108				
13175	16	1571	15611	121814	1822	11523				
1661920	17	14	19	14	1435	11	102012	13		
108	66	16	13	185						

Tabela B.2: 10 classificações não informativas adicionais usadas para formar o conjunto de dados 2.

Classificação 41	Classificador	50
1	11 7 20 13 17	8 15 19 18 18
21	15 12 137 17 10	206318 20 162
1015	41 19415 19 17 176	1173 111512 151
19	186 127 16613 14 14	2096 1892
1072	20 1276 159 147	10 1481029 101 169
156	2096 18 134 14 141	18 1581073 10
16	20 17485 13111995	18 165
1363	212148 1935 13 19	1743134569 19
122	3 13 1514643824	12 149
1115	26273 20 16 20 17162	10 1858
1835	125171714 14 11	10 11 167183
16	12 11 2282 149199	178 1543 11 16
17	12 205 13 1117	15 20184

Tabela B.4: 10 classificações não informativas adicionais usadas para formar o conjunto de dados 4.

Dataset
S

Classificar	
	Classificador 1911 5 4 6 2 7 3 9 8
21	3 2 6 7 5 4 8 9 3 1 6 4 7 3 2 8 5
941	6 7 3 2 5 4 8 9 5 2 7 4 3 9 5 1
6 662	3 5 8 4 1 6 7 9 7 2 1 5 3 4 7
8 6 98	2 3 5 1 6 8 9 4 7 9 2 8 3 1 4
6 5 7	9 10 2 3 1 4 9 5 8 6 7 1 1 2 7 4
1 8 6	5 3 9 1 2 3 2 1 8 4 5 7 6 9 1 3 3
1 6 7	3 8 4 9 5 2 1 4 3 1 5 4 7 8 2 9
6 1 53	2 4 8 5 9 1 6 7 1 6 4 8 5 7 1 9
2 3 6	7 4 5 8 9 1 7 6 3 2 1 8 4 8 5 9
3 2 6	7 1 1 9 5 4 7 8 9 2 3 6 1 2 0 5 4
2 7 8	9 3 1 6 2 1 5 2 9 8 4 7 1 3
6 2 25	7 4 9 8 3 2 1 6 2 3 5 4 9 7 3 2
8 1 6	2 4 5 4 9 8 3 7 2 1 6 2 5 5 4 2 8
9 3 1	7 6 2 6 5 7 8 1 4 9 2 6 3 2 7 5 4
9 7 1	2 8 3 6 2 8 1 4 5 8 9 7 6 2
3 2 96	1 8 4 2 3 7 5 9 3 0 6 1 4 3 2 5
8 7 9	3 1 6 3 2 5 1 7 8 9 4 3 2 7 4 8 1
2 3 6	5 9 3 3 7 5 4 8 1 9 2 6 3 3 4 7 4
2 8 6	1 3 5 9 3 5 7 5 8 4 2 9 3 1
6 3 67	1 5 4 3 2 6 8 9 3 7 7 4 5 8 1 9
6 2 3 3	8 9 5 3 4 7 8 2 1 6 3 9 9 4 8 5
1 2 3	7 6
<hr/>	
Nº da entidade	12
Área	SOC EDU CLI MAT EXP CUL IND TST PHY
	34
	56
	78
	9

Tabela B.6: Dados psicológicos de Roskam

Dataset
S

225

Apêndice B.

Dataset
s

Classificação	Classificador									
	81	...	90	...	100					
19	699	10	49	3	7	10	10	59	57	3
28	88	22	22	22	58	62	89	22	22	10
3	75	14	96	16	48	91	771	10	45	89
4101	5653546469643514465532387693218	10	1088871762							
10314871194716277864732	1051	103	10512359	10699						
10281477758887328254665594748694796544349	1039110696	1031	105	1037						
1031613133										

Tabela B.9: 20 classificações adicionais usadas para formar o conjunto de dados 6

Bibliografia

- Aldous, D. J. (1985), Permutabilidade e tópicos relacionados, em "Ecole d'Eté de Probabilités de Saint-Flour XIII-1983", Springer, pp. 1–198.
- Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. e Ronquist, F. (2004), 'Metropo- Paralelo cadeia de Markov acoplada a lis Monte Carlo para inferência filogenética bayesiana', Bioinformatics 20(3), 407–415.
- Antoniak, C. E. (1974), 'Misturas de processos de Dirichlet com aplicações a Bayesian problemas não paramétricos', *The Annals of Statistics* 2, 1152–1174.
- Atchad'e, Y. F., Roberts, G. O. e Rosenthal, J. S. (2011), 'Rumo ao dimensionamento ideal de Cadeia de Markov acoplada à metrópole Monte Carlo ', *Estatística e Computação* 21 (4), 555–568.
- Baker, R. D. e McHale, I. G. (2015), 'Evolução determinística da força em múltiplos modelos de comparação: Quem é o maior jogador de golfe?', *Scandinavian Journal of Statistics* 42(1), 180–196.
- Benter, W. (1994), Sistemas de handicap e apostas de corridas de cavalos baseados em computador: A relatório, em 'Eficiência dos mercados de apostas em pistas de corrida', Academic Press, pp. 183–198.
- Bernardo, J. M. e Smith, A. F. M. (1994), Teoria Bayesiana, Wiley, Chichester, Reino Unido
- Bez'akov'a, I., Kalai, A. e Santhanam, R. (2006), Seleção de modelo de grafo usando máximo probabilidade, em 'Anais da 23ª conferência internacional sobre aprendizado de máquina', ACM, pp. 105–112.
- Blackwell, D. e MacQueen, J. B. (1973), 'Distribuições de Ferguson por meio de esquemas de urnas Pólya', Os Anais de Estatística 1, 353–355.
- Bradley, R. A. e Terry, M. E. (1952), 'Análise de classificação de projetos de blocos incompletos: I. O método de comparações emparelhadas', *Biometrika* 39 (3-4), 324–345.
- Breslow, N., Crowley, J. et al. (1974), 'Um grande estudo de amostra da tábua de vida e do produto limitar estimativas sob censura aleatória', *The Annals of Statistics* 2(3), 437–453.

- Brooks, S. (1998), 'Método de Monte Carlo da cadeia de Markov e sua aplicação', *Journal of the Royal Statistical Society: Série D (o Estatístico)* 47(1), 69–100.
- Caron, F. e Doucet, A. (2012), 'Inferência bayesiana eficiente para Bradley generalizado-Modelos de Terry', *Jornal de Estatística Computacional e Gráfica* 21(1), 174–196.
- Caron, F., Teh, Y. W. e Murphy, T. B. (2014), 'Plackett-Luce não paramétrico bayesiano modelos para a análise de preferências por programas de graduação universitária', *The Annals of Applied Statistics* 8(2), 1145–1181.
- Casella, G. e Robert, C. P. (1996), 'Rao-Blackwellisation of sampling schemes', *Biometrika* 83(1), 81–94.
- Chib, S. e Greenberg, E. (1995), 'Entendendo o algoritmo Metropolis-Hastings', *O Estatístico Americano* 49(4), 327–335.
- Choulakian, V. (2016), 'Componentes de mistura globalmente homogêneos e heterogeneidade local de dados de classificação', pré-impressão arXiv arXiv:1608.05058
- Daíhl, D. B. (2006), Clustering baseado em modelo para dados de expressão por meio de um processo de Dirichlet modelo de mistura, em K.-A. Do, P. Müller e M. Vannucci, eds, 'Bayesian Inference for Gene Expression and Proteomics', Cambridge University Press, pp. 201–218.
- de Leeuw, J. (2006), Análise de componentes principais não lineares, em M. Greenacre e J. Bla-sius, eds, 'Análise de correspondência múltipla e métodos relacionados', CRC Press, capítulo 4, pp. 107–134.
- de Leeuw, J. e Mair, P. (2009), 'Métodos Gifi para escalonamento ideal em R: O pacote homals', *Jornal de Software Estatístico* 31(4), 1–20. URL: <http://www.jstatsoft.org/v31/i04/>
- Deng, K., Han, S., Li, K. J. e Liu, J. S. (2014), 'Agregação bayesiana de pedidos baseados em dados de classificação', *Journal of the American Statistical Association* 109(507), 1023–1039.
- Diaconis, P. (1988), 'Representações de grupo em probabilidade e estatística', Notas de aula-Série de Monografias 11, 1–192.
- Duane, S., Kennedy, A. D., Pendleton, B. J. e Roweth, D. (1987), 'Monte Carlo Híbrido', *Cartas de Física B* 195(2), 216–222.
- Escobar, M. D. e West, M. (1995), 'Estimativa de densidade bayesiana e inferência usando misturas', *Jornal da Associação Americana de Estatística* 90 (430), 577–588.
- Everitt, B. e Hand, D. (1981), 'Distribuições de misturas finitas', Monografias sobre Aplicado Probabilidade e Estatística, Londres: Chapman e Hall, 1981.

- Everitt, B. S., Landau, S., Leese, M. e Stahl, D. (2011), 'Agrupamento hierárquico', Cluster Análise, 5^a Edição, pp. 71–110.
- Ferguson, T. S. (1973), 'Uma análise bayesiana de alguns problemas não paramétricos', The Annals de Estatística 1, 209–230.
- Gamerman, D. e Lopes, H. F. (2006), Monte Carlo da cadeia de Markov: simulação estocástica para inferência bayesiana, Chapman e Hall/CRC.
- Gelfand, A. E. e Smith, A. F. (1990), 'Abordagens baseadas em amostragem para calcular densidades marginais', Jornal da Associação Americana de Estatística 85 (410), 398–409.
- Gelman, A. (1996), Inferência e convergência de monitoramento, em W. Gilks, S. Richardson e D. Spiegelhalter, eds, 'Markov Chain Monte Carlo na Prática', Chapman & Hall / CRCInterdisciplinary Statistics, Taylor & Francis.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. e Rubin, D. B. (2014), Análise de Dados Bayesianos, 3^a edn, CRC press Boca Raton, FL.
- Gelman, A., Roberts, G. O., Gilks, W. R. et al. (1996), Regras de salto de metrópole eficiente, em J. M. Bernardo, J. Berger, A. P. Dawid e J. F. M. Smith, eds, 'Bayesian Statistics5', Oxford University Press, pp. 599–608.
- Geman, S. e Geman, D. (1984), 'Relaxamento estocástico, distribuições de Gibbs e o Restauração bayesiana de imagens', IEEE Transactions on Pattern Analysis and MachineIntelligence 6, 721–741.
- Geweke, J. (1992), Avaliando a precisão das abordagens baseadas em amostragem para calcular momentos posteriores, em J. M. Bernardo, J. Berger, A. P. Dawid e J. F. M. Smith, eds, 'Bayesian Statistics 4', Oxford University Press, pp. 169–193.
- Geyer, C. J. (1991), Probabilidade máxima de Monte Carlo da cadeia de Markov, em 'Proceedings do 23º Simpósio sobre a Interface', Ciência da Computação e Estatística, InterfaceFoundation of North America, pp. 156–163.
- Geyer, C. J. e Thompson, E. A. (1995), 'Recozimento da cadeia de Markov Monte Carlo com aplicações à inferência ancestral', Journal of the American Statistical Association90(431), 909–920.
- Gilks, W. R. e Roberts, G. O. (1996), Estratégias para melhorar o MCMC, em W. Gilks, S. Richardson e D. Spiegelhalter, eds, 'Markov chain Monte Carlo in Practice', Chapman & Hall / CRC Interdisciplinary Statistics, Taylor & Francis.

- Glickman, M. E. e Hennessy, J. (2015), 'Um modelo logit ordenado por classificação estocástica para classificação de jogos e esportes multicompetitivos', *Journal of Quantitative Analysis in Sports*11(3), 131–144.
- Gormley, I. C. e Murphy, T. B. (2006), 'Análise da aplicação da faculdade irlandesa de terceiro nível dados de cônices', *Journal of the Royal Statistical Society: Série A (Estatísticas na Sociedade)* 169 (2), 361–379.
- Gormley, I. C. e Murphy, T. B. (2008a), 'Explorando blocos de votação dentro da eleição irlandesa torate: uma abordagem de modelagem de mistura', *Journal of the American Statistical Association*103(483), 1014–1027.
- Gormley, I. C. e Murphy, T. B. (2008b), 'Uma mistura de modelos de especialistas para dados de classificação com aplicações em estudos eleitorais', *Anais de Estatística Aplicada* 2(4), 1452–1477.
- Gormley, I. C. e Murphy, T. B. (2009), 'Um modelo de grau de associação para dados de classificação', *Análise Bayesiana* 4(2), 265–295.
- Graves, T., Reese, C. S. e Fitzgerald, M. (2003), 'Modelos hierárquicos para permutações: Análise dos resultados do automobilismo', *Journal of the American Statistical Association*98 (462), 282–291.
- Hastie, D. I., Liverani, S. e Richardson, S. (2015), 'Amostragem do processo de Dirichlet modelos de mistura com parâmetro de concentração desconhecido: problemas de mistura em grandes implementações de dados', *Statistics and Computing* 25(5), 1023–1037.
- Hastings, W. K. (1970), 'Métodos de amostragem de Monte Carlo usando cadeias de Markov e seus Aplicações', *Biometrika* 57(1), 97–109.
- Henderson, D. A. e Kirrane, L. J. (2018), 'Uma comparação de truncado e ponderado no tempo Modelos de Plackett-Luce para previsão probabilística dos resultados da Fórmula Um', *BayesianAnalysis* 13(2), 335–358.
- Henery, R. (1981), 'Probabilidades de permutação como modelos para corridas de cavalos', *Journal of the Sociedade Real de Estatística. Série B (Metodologia Estatística)*, pp. 86–91.
- Hjort, N. L., Holmes, C. C., Müller, P. e Walker, S. G., eds (2010), *Bayesian Nonparamétricas*, Cambridge University Press, Cambridge UK.
- Hunter, D. R. (2004), 'Algoritmos MM para modelos generalizados de Bradley-Terry', *The Annals de Estatística* 32(1), 384–406.
- Ishwaran, H. e James, L. F. (2001), 'Métodos de amostragem de Gibbs para priores de quebra de bastão', *Jornal da Associação Americana de Estatística* 96(453), 161–173.

- Ishwaran, H. e Zarepour, M. (2002), 'Representações de soma exatas e aproximadas para o processo de Dirichlet', *Canadian Journal of Statistics* 30(2), 269–283.
- Lau, J. W. e Green, P. J. (2007), 'Procedimentos de agrupamento baseados em modelo bayesiano', *Journal de Estatística Computacional e Gráfica* 16(3), 526–558.
- Lindsay, B. G. (1995), Modelos de mistura: teoria, geometria e aplicações, em 'NSF-CBMS série de conferências regionais em probabilidade e estatística', JSTOR, pp. 1–163.
- Luce, R. (1959), *Comportamento de Escolha Individual: Uma Análise Teórica*, Wiley.
- MacEachern, S. N. e Müller, P. (1998), 'Estimando a mistura de modelos de processo de Dirichlet', *Jornal de Estatística Computacional e Gráfica* 7(2), 223–238.
- Mallows, C. L. (1957), 'Modelos de classificação não nulos. I', *Biometrika* 44(1–2), 114–130.
- Marden, J. I. (1995), *Analisando e Modelando Dados de Classificação*, Chapman e Hall, Londres.
- Medvedovic, M. e Sivaganesan, S. (2002), 'Modelo de mistura infinita bayesiano baseado em clustering de perfis de expressão gênica', *Bioinformática* 18(9), 1194–1206.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. e Teller, E. (1953), 'Equação de cálculos de estado por máquinas de computação rápida', *The Journal of Chemical Physics* 21(6), 1087–1092.
- Mollica, C. e Tardella, L. (2014), 'Perfil de epitopos por meio de modelagem de mistura de classificados dados', *Estatísticas em Medicina* 33(21), 3738–3758.
- Mollica, C. e Tardella, L. (2016), 'Modelos de mistura Bayesian Plackett-Luce para dados classificados', *Psychometrika* 82(2), 1–17.
- Mollica, C. e Tardella, L. (2018), 'Algoritmos e diagnósticos para a análise de classificações de preferência com o modelo Extended Plackett-Luce', arXiv preprint arXiv:1803.02881 .
- Neal, R. M. (2000), «Métodos de amostragem em cadeia de Markov para modelos de mistura de processos de Dirichlet», *Jornal de Estatística Computacional e Gráfica* 9(2), 249–265.
- Neal, R. M. (2011), MCMC usando dinâmica hamiltoniana, em S. Brooks, A. Gelman, G. Jones e X.-L. Meng, eds, 'Manual da Cadeia Markov Monte Carlo', CRC Press, pp. 93–162.
- Papaspiliopoulos, O. e Roberts, G. O. (2008), 'Cadeia retrospectiva de Markov Monte Carlo métodos para modelos hierárquicos do processo de Dirichlet', *Biometrika* 95(1), 169–186.

- Pitman, J. e Yor, M. (1997), 'A distribuição de Poisson-Dirichlet de dois parâmetros derivada de um subordinador estável', *The Annals of Probability* 25(2), 855–900.
- Plackett, R. L. (1975), 'A análise de permutações', *Applied Statistics* 24, 193–202.
- Plummer, M., Best, N., Cowles, K. e Vines, K. (2006), 'CODA: Diagnóstico de Convergência e Análise de Saída para MCMC', *R News* 6(1), 7–11. URL: <https://journal.r-project.org/archive/>
- Raftery, A. E. e Lewis, S. (1992), Quantas iterações no sampler de Gibbs?, em J. M. Bernardo, J. Berger, A. P. Dawid e J. F. M. Smith, eds, 'Bayesian Statistics 4', Oxford University Press, pp. 763–773.
- Raftery, A. E. e Lewis, S. (1996), Implementando MCMC, em 'Markov Chain Monte Carlo na prática', W. R. Gilks, S. Richardson e D. J. Spiegelhalter, eds, Chapman & Hall, Londres, pp. 115–130.
- Rastelli, R. e Friel, N. (2017), 'Estimadores bayesianos ideais para cluster de variáveis latentes modelos', *Estatística e Computação* pp. 1–18.
- Richardson, S. e Green, P. J. (1997), 'Sobre a análise bayesiana de misturas com um desconhecido número de componentes (com discussão)', *Journal of the Royal Statistical Society: SeriesB (Metodologia Estatística)* 59(4), 731–792.
- Roberts, G. O. e Rosenthal (2001), 'Dimensionamento ideal para várias metrópoles-Hastings algoritmos', *Ciência Estatística* 16(4), 351–367.
- Rodriguez, A. (2007), Alguns avanços na modelagem não paramétrica bayesiana, tese de doutorado, Universidade Duke.
- Rodriguez, A., Dunson, D. B. e Gelfand, A. E. (2008), 'O processo Nested Dirichlet', *Jornal da Associação Americana de Estatística* 103(483), 1131–1154.
- Sethuraman, J. (1994), 'Uma definição construtiva de prioris de Dirichlet', *Statistica Sinica* 4, 639–650.
- Sherlock, C. (2013), 'Dimensionamento ideal da metrópole de passeio aleatório: critérios gerais para a regra de aceitação de 0,234', *Journal of Applied Probability* 50(1), 1–15.
- Sherlock, C. e Roberts, G. (2009), 'Escala ideal da caminhada aleatória Metropolis on alvos unimodais elípticos simétricos', *Bernoulli* 15(3), 774–798.
- Stephens, M. (2000), 'Lidando com a troca de rótulos em modelos de mistura', *Jornal do Royal Statistical Society: Série B (Metodologia Estatística)* 62(4), 795–809.

- Stern, H. (1990), 'Modelos para distribuições em permutações', *Journal of the American Association Estatística* 85(410), 558–564.
- Tanner, M. A. e Wong, W. H. (1987), 'O cálculo de distribuições posteriores por aumento de dados', *Journal of the American Statistical Association* 82(398), 528–540.
- Teh, Y. W., Jordan, M. I., Beal, M. J. e Blei, D. M. (2006), 'Dirichlet Hierárquico processos', *Journal of the American Statistical Association* 101(476), 1566–1581.
- Vigneau, E., Courcoux, P. e Semenou, M. (1999), 'Análise de dados de preferência classificada: usando modelos de classe latente', *Qualidade e Preferência Alimentar* 10(3), 201–207.
- Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A. e Arjas, E. (2018), 'Probabilístico aprendizagem de preferência com o modelo de classificação de Mallows', *Journal of Machine Learning Re-search* 18(158), 1–49.
- Walker, S. G. (2007), 'Amostragem do modelo de mistura de Dirichlet com fatias', *Comunicações em Estatística - Simulação e Computação* 36(1), 45–54.
- West, M. (1992), Estimativa de hiperparâmetros em modelos de mistura de processos de Dirichlet, Duke Documento de discussão ISDS da universidade # 92-A03.
- Wilkinson D. (2013), «Paralelo Têmpera e Metrópole Acoplado MCMC», <https://darrenjw.wordpress.com/tag/mc3/>. [Online; acessado em 4 de maio de 2018].
- Yu, P. L., Lam, K. e Lo, S. (2005), 'Análise fatorial para dados classificados com aplicação para uma pesquisa de atitude de seleção de emprego', *Journal of the Royal Statistical Society: Série A (Estatísticas na Sociedade)* 168(3), 583–597.