Attraction Indian Buffet Distribution

David Dahl & Richard Warr
Brigham Young University

Abstract

We propose the attraction Indian buffet distribution (aIBD) as a latent feature model indexed by pairwise distances. Items that are "close" are more likely to share features. The aIBD extends the Indian buffet process (IBP) in the same way that the Ewens-Pitman attraction (EPA) distribution extends the Chinese restuarant process (CRP). We draw comparisons between the aIBD and the distance dependent Indian buffet process (ddIBP). We show that the aIBD maintains the expected number of columns of the IBP, has an explicit probability mass function, and reduces to the IBP in special cases.

Properties of IBP Maintained

Property	aIBD	ddIBP
Explicit pmf	Yes	No
Expected non-zero columns equal to that of IBP	Yes	No
Expected row sums equal to that of IBP	Yes	Yes
Expected column sums equal to that of IBP	No	No

Notation

Each row of the binary feature matrix $\mathbf{Z}$ denotes a "customer" and each column represents a "dish". $z_{i,k}$ is set to 1 if customer i takes dish k, and 0 otherwise.

- x_i is the number of *new* dishes that customer i takes and $y_{-i} = \sum_{j=1}^{i-1} x_j$ is the number of existing dishes before customer i.
- $m_{-i,k}$ is the number of customers that took dish k before customer customer i samples dishes.
- Let

$$h_{i,k} = \frac{\sum_{j=1}^{i-1} \lambda(\sigma_j, \sigma_i)\, z_{j,k}}{\sum_{j=1}^{i-1} \lambda(\sigma_j, \sigma_i)}$$

be the similarity component of the weight given to sampling dish k for customer i, where the permutation σ is the order in which customers are assigned dishes, and the similarity function $\lambda(i,j)$ maps the distance between customers i and j to a measure of how "close" the two customers are.

Demonstration

US Arrest Dataset in R

	Murder	Assault	Urban	Rape
NH	2.1	57	56	9.5
IA	2.2	56	57	11.3
WI	2.6	53	66	10.8
CA	9.0	276	91	40.6
NV	12.2	252	81	46.0

→

Euclidean Distance

NH	IA	WI	CA	NV
0.00	0.12	0.66	3.74	3.78
0.12	0.00	0.59	3.65	3.69
0.66	0.59	0.00	3.32	3.46
3.74	3.65	3.32	0.00	1.01
3.78	3.69	3.46	1.01	0.00

→

Similarity

NH	IA	WI	CA	NV
1.00	0.89	0.51	0.02	0.02
0.89	1.00	0.55	0.03	0.02
0.51	0.55	1.00	0.04	0.03
0.02	0.03	0.04	1.00	0.36
0.02	0.02	0.03	0.36	1.00

Expected Number of Shared Features

IBP

	NH	IA	WI	CA	NV
NH	1	0.5	0.5	0.5	0.5
IA	0.5	1	0.5	0.5	0.5
WI	0.5	0.5	1	0.5	0.5
CA	0.5	0.5	0.5	1	0.5
NV	0.5	0.5	0.5	0.5	1

aIBD

	NH	IA	WI	CA	NV
NH	1	0.59	0.57	0.44	0.44
IA	0.59	1	0.57	0.44	0.44
WI	0.57	0.57	1	0.45	0.44
CA	0.44	0.44	0.45	1	0.6
NV	0.44	0.44	0.44	0.6	1

ddIBP

	NH	IA	WI	CA	NV
NH	1	0.47	0.37	0.05	0.04
IA	0.47	1	0.38	0.05	0.05
WI	0.37	0.38	1	0.05	0.05
CA	0.05	0.05	0.05	1	0.25
NV	0.04	0.05	0.05	0.25	1

Probability of Feature Given First Two Rows

IBP

	1	2	3	4	5
NH	1	0	0	0	0
IA	0	1	1	0	0
WI	0.33	0.33	0.33	0.28	0.05
CA	0.33	0.33	0.33	0.23	0.08
NV	0.33	0.33	0.34	0.21	0.09

aIBD

	1	2	3	4	5
NH	1	0	0	0	0
IA	0	1	1	0	0
WI	0.29	0.34	0.34	0.24	0.04
CA	0.2	0.26	0.25	0.24	0.09
NV	0.2	0.27	0.26	0.23	0.1

ddIBP

	1	2	3	4	5
NH	1	0	0	0	0
IA	0	1	1	0	0
WI	0.25	0.27	0.27	0.38	0.08
CA	0.03	0.03	0.03	0.39	0.34
NV	0.03	0.03	0.03	0.23	0.29

Probability Mass Function

IBP:

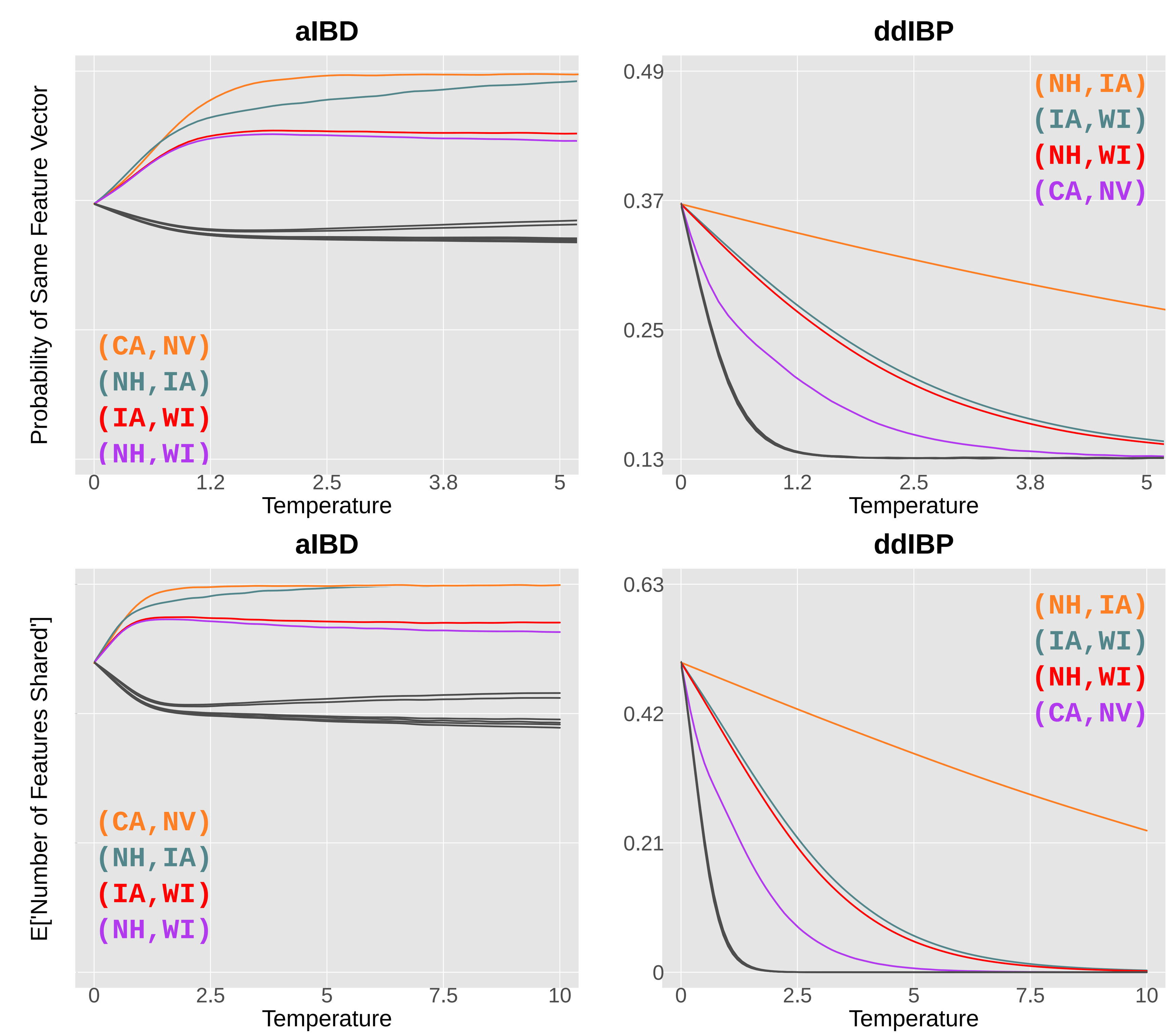
$$P(\mathbf{Z}|\alpha) = \frac{\alpha^K \exp\{-\alpha H_N\}}{\prod_{i=1}^{N} (i^{x_i}\, x_i!)} \prod_{i=2}^{N} \prod_{k=1}^{y_i} \left(\frac{m_{-i,k}}{i} \right)^{z_{i,k}} \left(1 - \frac{m_{-i,k}}{i} \right)^{1-z_{i,k}}$$

aIBD:

$$P(\mathbf{Z}|\mathbf{D}, \sigma, \alpha) = \frac{\alpha^K \exp\{-\alpha H_N\}}{\prod_{i=1}^{N} (i^{x_i}\, x_i!)} \prod_{i=2}^{N} \prod_{k=1}^{y_i} \left(\frac{h_{i,k}\,(i-1)}{i} \right)^{z_{i,k}} \left(1 - \frac{h_{i,k}\,(i-1)}{i} \right)^{1-z_{i,k}}$$

where $H_N = \sum_{i=1}^{N} \frac{1}{i}$ and N is the total number of "customers". Note: If $y_i = 0$ then the result of the product is 1.

Temperature Plots



Sampling Algorithm

To obtain a realization $\mathbf{Z}$ from an aIBD(α) with permutation σ and distance $\mathbf{D}$:

1. The first customer σ_1 takes a Poisson(α) number of dishes.
2. For customers $\sigma_i = 2$ to N,
   - For each previously sampled dish, customer σ_i takes dish k with probability $h_{i,k}\, m_{-i}/i$.
   - After sampling all previously sampled dishes, customer i samples Poisson(α/i) new dishes.

References

- Blei, D. M. and Frazier, P.I. (2011), "Distance Dependent Chinese Restaurant Process," *Journal of Machine Learning Research*, 12, 2383-2410.
- Dahl, D. B., Day, R., Tsai, J. W. (2017), "Random Partition Distribution Indexed by Pairwise Information," *Journal of the American Statistical Association*, 112, 721-732.
- Gershman, S. J., Frazier, P.I., and Blei, D. M. (2015), "Distance Dependent Infinite Latent Feature Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 334-345.
- Griffiths, T. L., and Ghahramani, Z. (2011), "The Indian Buffet Process: An Introduction and Review," *Journal of Machine Learning Research*, 12, 1185-1224.

http://dahl.byu.edu
warr@stat.byu.edu