Hierarchical Stick-breaking Feature Paintbox

Melanie F. Pradier
Harvard University
Cambridge, MA, USA

Weiwei Pan
Harvard University
Cambridge, MA, USA

Morris Yau
University of California
Berkeley, CA, USA

Rachit Singh
Harvard University
Cambridge, MA, USA

Finale Doshi-Velez
Harvard University
Cambridge, MA, USA

Abstract

Bayesian non-parametric latent feature models based on the Indian buffet process (IBP) provide interpretable representations of the data via binary-weighted matrix factorization. The IBP gives rise to a potentially infinite number of latent features, but has the disadvantage of assuming a priori independence between those features. In contrast, feature allocation models based on the paintbox representation allow for arbitrary complex correlations among the features. So far, there is no explicit generative construction nor effective inference algorithms for such representations. This paper presents the hierarchical stick-breaking paintbox process, a flexible tree-based prior that generalizes the IBP to capture correlation among the latent features. Theoretical proofs for asymptotic behavior of feature activation and exchangeability are given, as well as an efficient collapsed Gibbs sampler.

1 Introduction

Latent feature models have proved useful for revealing structures underlying complex data by decomposing observed attributes of the data into combinations of simple factors or features. However, since these features are generally unobserved and a priori unknown, assigning the appropriate number and combination of latent features to explain each observation is particularly challenging. In such scenarios, Bayesian non-parametrics provide a framework for building flexible latent feature models that adapt their capacity to explain available data [Gershman and Blei, 2012]. This flexibility is achieved by placing a distribution over an infinitely dimensional space of feature assignments, thus allowing the model to use as many latent features as would be needed to fit observations.

In particular, the Indian Buffet Process (IBP) is a feature-assignment prior that defines a probability distribution over classes of binary matrices with a potentially infinite number of columns Griffiths and Ghahramani [2011]. IBP priors have been used extensively across different fields for data exploration [Knowles and Ghahramani, 2011, Ruiz et al., 2014, Pradier et al., 2018]. However, the IBP prior assumes independence among the latent features, which might not be appropriate for many applications. In biology for instance, one expects many genetic traits to co-occur or to have complex inter-dependencies. In fact, there is a body of work attempting to extend the IBP prior to explicitly incorporate correlation into latent feature models models [Broderick et al., 2013, Doshi-Velez and Ghahramani, 2009, Chen et al., 2013, Ranganath and Blei, 2015].

Broderick et al. [2013] introduce feature allocation (FA) models, a general family of models for multi-label partition of data that uses latent features with potentially arbitrary correlations.[1] The authors define the *feature paintbox*, a graphical representation of the De Finetti mixing measure of this family. Despite its appealing flexibility, the feature paintbox lacks practical constructions and explicit inference algorithms in existing literature, because exchangeability—property under which indices to label the data points are irrelevant for inference—is hard to satisfy given the desiderata of arbitrary feature correlations.

[1] Using the IBP prior in latent feature models leads to a subset of FAs called "feature frequency models".

3rd Bayesian Nonparametrics Workshop at NeurIPS (BNP@NeurIPS 2018), Montréal, Canada.

This paper presents a novel flexible nonparametric prior for latent feature allocation models based on the feature paintbox representation. We treat the paintbox as a binary tree of conditional probabilities. Similar to the construction from Adams et al. [2010], we define a hierarchical stick-breaking process that allocates probability mass to different paths in the tree, each path corresponding to a binary vector of feature activation. We prove that the sequence of observations obtained from this prior is exchangeable, generalizing the IBP by accounting for both positive and negative correlations among features. We empirically show that, in cases where there are latent correlations among features, latent variable models incorporating our prior results in performance gains over those using IBP priors.

2 Background

Indian buffet process. The Indian buffet process (IBP) is a non-parametric prior over binary matrices with a finite number of rows and potentially unbounded number of columns [Griffiths and Ghahramani, 2011]. The IBP is often used in latent feature models as a prior over the feature-assignment matrix $\mathbf{Z}$ when the number of latent features K is unknown. In such models, $\mathbf{Z} \in \{0,1\}^{N \times K}$ is a binary matrix whose rows encode the active latent features for each data point $\mathbf{x}_n$, where $n = 1, \ldots, N$. Sampling $\mathbf{Z}$ from the IBP prior is denoted $\mathbf{Z} \sim \mathrm{IBP}(\alpha)$, where α is the mass parameter controlling the a priori activation probability of new features.

The IBP prior on $\mathbf{Z}$ can be defined via a stick-breaking procedure. The idea is to start with a "stick" of unit length, then recursively break it at a point $\mathrm{Beta}(\alpha, 1)$ along its length, keeping the initial portion of the stick and discarding the excess [Teh et al., 2007]. The length of the stick after the k-th break represents the activation probability of the k-th latent feature. More formally, the process is defined as follows:

$$v_k \sim \text{Beta}(\alpha, 1), \quad \pi_k = \prod_{i=1}^{k} v_i, \quad z_{nk} \sim \text{Bernoulli}(\pi_k), \tag{1}$$

where π_k is the activation probability of feature k. We see that in this construction, latent features are generated independently, rendering the IBP prior potentially inappropriate for data sets where latent features have complex dependencies.

Feature allocation and feature paintbox. Feature allocation models (FAs) are multi-label partitions of data sets. In particular, FA models produce subgrouping of N data points, represented as a multi-set $\{A_1, A_2, \ldots\}$ of non-empty subsets of $\{1, \ldots, N\}$, wherein each data point belongs to an arbitrary large but finite number of groups, A_k. Each group A_k represents a unique latent feature. Thus, the multi-group membership or *feature assignment*, $\boldsymbol{z}_n$, encodes for the set of active latent features present in each data point $\mathbf{x}_n$. Furthermore, Broderick et al. [2013] prove that a large subclass of FAs, for which the sequence of feature assignment is *exchangeable* (the model is unaffected by re-indexing the data) and *regular* (no data point is assigned a unique feature in an infinite data collection), admitsa concrete construction via a *feature paintbox*, which represents the de Finetti mixing measure of the sequence of feature assignments, $\{z_n\}$.

The feature paintbox includes the well known IBP [Griffiths and Ghahramani, 2011] and, more generally, feature allocations with arbitrary inter-feature correlations [Broderick et al., 2013]. Unfortunately, inference algorithms for general feature paintboxes do not exist in current literature. In this work, we develop a novel construction and efficient inference for latent feature models with FA priors. We demonstrate empirically that models using paintbox priors consistently outperform models using IBP priors, in applications where underlying features are highly correlated or anti-correlated.

3 Hierarchical stick-breaking paintbox process

A feature paintbox can be thought of as an infinite collection of subsets $\{C_k\}_{k=1}^{\infty}$ of the unit interval. Sampling from a paintbox involves choosing a value in the unit interval and determining what subsets it intersects. We say a feature paintbox is in its *canonical form* if all mass probabilities are concentrated to the left (see Figure 1).

We now describe a construction for canonical paintboxes, and hence a generative model for infinite binary matrices, which we call *hierarchical stick-breaking paintbox process*. Let $\nu_{\boldsymbol{\epsilon}}$ be the conditional activation probability of feature k, given a binary string $\boldsymbol{\epsilon}$ of the previous $k-1$ activations for features 1 to $k-1$.Z For example, ν_{01} denotes the probability $p\left(\boldsymbol{z}_3 = 1 | \boldsymbol{z}_1 = 0, \boldsymbol{z}_2 = 1\right)$. We assume that $\nu_{\boldsymbol{\epsilon}} \sim \text{Beta}\left(\frac{\alpha}{K^p}, 1\right)$, where α is the concentration parameter of the process, p is a model parameter that controls the average rate at which the activation of subsequent features decays, and K refers to the truncation level, e.g., an upper bound for the number of latent features. Let $\pi_{\boldsymbol{\epsilon}} = p(\boldsymbol{\epsilon})$ refer to the joint probability distribution of the latent features represented by the binary string $\boldsymbol{\epsilon}$. We can recursively compute the probability of extended strings $\{\boldsymbol{\epsilon}1\}$ and $\{\boldsymbol{\epsilon}0\}$ as $\pi_{\{\boldsymbol{\epsilon}1\}} = \pi_{\boldsymbol{\epsilon}}\nu_{\boldsymbol{\epsilon}}$ and $\pi_{\{\boldsymbol{\epsilon}0\}} = \pi_{\boldsymbol{\epsilon}}(1-\nu_{\boldsymbol{\epsilon}})$. We define the hierarchical stick-breaking paintbox process (HSBP) as the following iterative process:

- $\pi_{\emptyset} = 1$, $\nu_{\emptyset} \sim \mathrm{Beta}(\frac{\alpha}{K^p}, 1)$
- $\forall\, k = 1, \cdots, K$, and $j = 1, \cdots, 2^{k-1}$, draw $\nu_{\boldsymbol{\epsilon}_j} \sim \mathrm{Beta}(\frac{\alpha}{K^p}, 1)$, such that:

$$\begin{aligned}
\pi_1 &= \nu_{\emptyset} \\
\pi_0 &= (1-\nu_{\emptyset}) \\
\pi_{01} &= (1-\nu_{\emptyset})\nu_1 \\
\pi_{111} &= \nu_{\emptyset}\nu_1\nu_{11} \\
\pi_{010} &= (1-\nu_{\emptyset})\nu_1(1-\nu_{01}) \\
&\dots
\end{aligned}$$

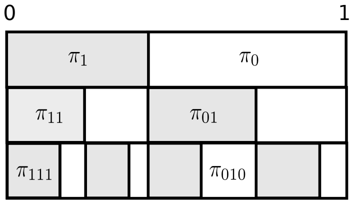


Figure 1: Example of canonical paintbox.

The feature paintbox induced by this iterative construction can be understood as a binary tree of conditional probabilities, which we denote $\boldsymbol{\nu} \sim \mathrm{HSBP}(\alpha, p)$. Given such paintbox, we can obtain a binary matrix $\mathbf{Z} \in \{0,1\}^{N\times K}$ by sampling each binary vector $\boldsymbol{z}_n$ of feature activations for observation n from a categorical distribution over the leaves of the tree, i.e., $\boldsymbol{z}_n \sim \mathrm{Multinomial}\left(1, \{\pi_{\boldsymbol{\epsilon}}\}_{\boldsymbol{\epsilon}\in\mathcal{S}_K}\right)$, where $\mathcal{S}_K$ is the set of all binary vectors of length K. This becomes intractable as K increases. Alternatively, we can sample each row $\boldsymbol{z}_n$ element-wise from each Bernoulli conditional probability distribution by traversing the tree top down:

$$p(\boldsymbol{z}_n) = \prod_{k=1}^{K} p\left(z_{nk}|\boldsymbol{z}_{n,1:(k-1)}\right). \tag{2}$$

Note that the rows of a binary matrix $\mathbf{Z}$ generated from the HSBP are exchangeable by construction, since the rows are conditionally i.i.d. given the sequence of Beta rv's $\boldsymbol{\nu}$ (see Appendix C).

Vanishing marginal feature probability as $K \to \infty$. Let π_k be the probability of activation for feature k, i.e., $\pi_k = \sum_{\boldsymbol{\epsilon}\in\mathcal{S}_{k=1}} \pi_{\boldsymbol{\epsilon}}$ where $\mathcal{S}_{k=1}$ refer to the set of binary strings whose k-th component is equal to one. The proposed iterative process gives rise to valid feature allocations if π_K vanishes as $K \to \infty$. This guarantees that there is no observation with an infinite number of latent features. The marginal probability of the last feature K can be written as

$$\pi_K = \sum_{\boldsymbol{\epsilon}\in\mathcal{S}_{K-1}} \pi_{\boldsymbol{\epsilon}1} = \sum_{\boldsymbol{\epsilon}\in\mathcal{S}_{K-1}} \prod_{\boldsymbol{\epsilon}'<\boldsymbol{\epsilon}} \nu_{\boldsymbol{\epsilon}'} \tag{3}$$

where $\mathcal{S}_{K-1}$ is the set of binary strings of length $(K-1)$, the notation $\boldsymbol{\epsilon}' < \boldsymbol{\epsilon}$ refers to the ancestors[2] of the binary string $\boldsymbol{\epsilon}$, and $\nu_{\boldsymbol{\epsilon}'}$ refer to Beta-distributed random variables, either $\nu_{\boldsymbol{\epsilon}'} \sim \mathrm{Beta}(\alpha/K^p, 1)$ or $\nu_{\boldsymbol{\epsilon}'} \sim \mathrm{Beta}(1, \alpha/K^p)$ depending on whether the last component of $\boldsymbol{\epsilon}'$ is one or zero. Since the random variables $\nu_{\boldsymbol{\epsilon}'}$ are independent, the expectation of its product is the product of the expectations. We can thus compute the expectation $\mathbb{E}\left[\pi_K\right]$ in closed-form and take the limit as $K \to \infty$ (details of the derivation can be found in Appendix A). The final result can be written as:

$$\lim_{K\to\infty} \mathbb{E}\left[\pi_K\right] = \lim_{K\to\infty} \frac{\alpha}{\alpha + K^p} = 0 \quad \forall p > 0. \tag{4}$$

Derivation of the predictive distribution. Let $\mathbf{Z}_{-(nk)}$ refer to matrix $\mathbf{Z}$ except the element z_{nk}, and $\mathbf{Z}_{-n}$ to all elements from $\mathbf{Z}$ except row $\boldsymbol{z}_n$. We compute the conditional probability $p\left(z_{nk}|\mathbf{Z}_{-(nk)}\right)$ as follows:

$$\begin{aligned}
p\left(z_{nk}|\mathbf{Z}_{-(nk)}\right) &\propto \int_{\boldsymbol{\nu}} p\left(\boldsymbol{z}_n|\boldsymbol{\nu}\right) p\left(\boldsymbol{\nu}|\mathbf{Z}_{-n}\right) d\boldsymbol{\nu} && (5) \\
&\propto \prod_{\boldsymbol{\epsilon}\in\mathcal{S}_n} \frac{\left(\frac{\alpha}{K^p} + \phi_{\boldsymbol{\epsilon}1}^{-n}\right)^{z_{nk}} \left(1 + \phi_{\boldsymbol{\epsilon}0}^{-n}\right)^{(1-z_{nk})}}{\left(\frac{\alpha}{K^p} + 1 + \phi_{\boldsymbol{\epsilon}}^{-n}\right)}, && (6)
\end{aligned}$$

[2] We follow the same notation as in [Adams et al., 2010].

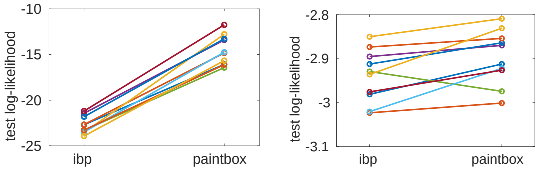
where $\phi_{\boldsymbol{\epsilon}'}^{-n}$ is a sufficient statistic accounting for the number of times that the binary vector $\boldsymbol{\epsilon}'$ appears in $\mathbf{Z}_{-n}$, and $\mathcal{S}_n$ is the set of subsequent partial binary vectors for observation n, i.e., $\mathcal{S}_n = \{z_{n1}, z_{n,(1:2)}, \ldots, z_{n,(1:K)}\}$. Eq. (5) can be integrated out analytically due to the conjugacy property between Bernoulli and Beta distributions (see complete derivation in Appendix B). Eq. (6) shows that the probability of each binary vector $\boldsymbol{z}_n$ only depends on the previous number of counts across the tree, i.e., this probability is independent of the sampling order of previous binary vectors, proving the exchangeability condition when $\boldsymbol{\nu}$ is integrated out. A collapsed Gibbs sampler can be derived straightforwardly by sampling from Eq. (6), see Appendix D for details.

4 Results

In this section, we compare an infinite latent feature model with Gaussian likelihood using either an IBP or HSBP prior. Inference is performed using a collapsed Gibbs sampler with Metropolis-Hasting proposals, as detailed in Appendix C. We consider two datasets: a synthetic one of correlated toy images ($N = 300$, $D = 36$), and the breast cancer dataset ($N = 500$, $D = 30$).[3] In the toy image dataset, the latent features correspond to four motifs in each of the four quadrants ("stair", "T", "square", and "cross"), exactly like Fig. 8 in [Griffiths and Ghahramani, 2011]. Yet, we construct the true matrix $\mathbf{Z}$ such that each latent feature has the same activation probability ($\forall k,\ \pi_k = 0.4$), but exhibit positive and negative correlations, i.e., 40% of the data exhibit one of the four latent features active alone (randomly), 30% of the data have ("T","square"), and the remaining 30% have ("stairs", ‘cross").

In all experiments, we run 10 different splits, 100 iterations of the Gibbs sampler, and initialize using non-negative matrix factorization [Lee and Seung, 2001]; 20% of the data is held-out for evaluation. Such test data is partially observed (only some dimensions are missing), the objective is to learn the latent features and reconstruct the missing observations. We fix $\sigma_x = 0.25$, $\alpha = 1$, and the truncation level $K = 6$.

Figure 1 shows the held-out log likelihood using an IBP or HSBP prior for the two datasets. The HSBP prior allows capturing implicit correlations among the latent features, resulting in better reconstructions of the missing observations. On top of better accuracy, the HSBP prior allows learning more interpretable dictionaries. Figure 2 compares the inferred dictionaries for the IBP and HSBP priors for the toy dataset. The IBP learns entangled representations corresponding to frequent co-occurrences of the features, and is not able to recover all the individual features due to the small dimensionality of the dataset. In contrast, the HSBP is able to recover the true four components.



(a) correlated toy images (b) breast cancer

Figure 1: Test log likelihood for a latent feature model with Gaussian likelihood and IBP or HSBP prior (each line corresponds to a different split of the data). a) correlated toy images, b) breast cancer dataset. **The HSBP prior improves performance significantly in the held-out data.**

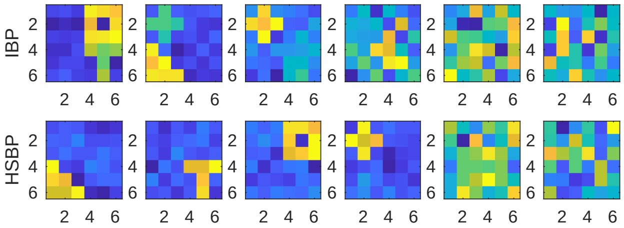


Figure 2: Dictionary learned for a latent feature model with Gaussian likelihood and IBP or HSBP prior over the latent feature activation matrix for the correlated toy images. **Using the HSBP prior, we are able to recover the true components.**

[3] `https://archive.ics.uci.edu/ml/datasets.html`

Acknowledgements

The authors would like to thank the Harvard Data Science Initiative (HDSI), the Center on Research for Computation and Society (CRCS), and the Institute for Applied Computational Sciences (IACS) at Harvard for financial support.

References

Ryan Prescott Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-Structured Stick Breaking Processes for Hierarchical Data. *arXiv:1006.1062 [stat]*, June 2010. URL `http://arxiv.org/abs/1006.1062`. arXiv: 1006.1062.

Tamara Broderick, Jim Pitman, and Michael I. Jordan. Feature Allocations, Probability Functions, and Paintboxes. *Bayesian Analysis*, 8(4):801–836, December 2013. ISSN 1936-0975, 1931-6690. doi: 10.1214/13-BA823. URL `https://projecteuclid.org/euclid.ba/1386166314`.

Mengjie Chen, Chao Gao, and Hongyu Zhao. Phylogenetic Indian Buffet Process: Theory and Applications in Integrative Analysis of Cancer Genomics. *stat*, 2013.

Finale Doshi-Velez and Zoubin Ghahramani. Correlated non-parametric latent feature models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 143–150. AUAI Press, 2009. URL `http://dl.acm.org/citation.cfm?id=1795132`.

S. J. Gershman and D. M. Blei. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012. URL `http://www.sciencedirect.com/science/article/pii/S002224961100071X`.

Thomas L. Griffiths and Zoubin Ghahramani. The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, 12:1185–1224, 2011. URL `http://mlg.eng.cam.ac.uk/pub/pdf/GriGha11.pdf`.

David Knowles and Zoubin Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, pages 1534–1552, 2011.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

M. F. Pradier, V. Stojkoski, Z. Utkovski, L. Kocarev, and F. Perez-Cruz. Sparse Three-parameter Restricted Indian Buffet Process for Understanding International Trade. In *International Conference on Acoustics, Speech and Signal Processing*, 2018.

R. Ranganath and D. Blei. Correlated Random Measures. *arXiv:1507.00720 [stat]*, July 2015. URL `http://arxiv.org/abs/1507.00720`. arXiv: 1507.00720.

Francisco JR Ruiz, Isabel Valera, Carlos Blanco, and Fernando Perez-Cruz. Bayesian nonparametric comorbidity analysis of psychiatric disorders. *Journal of Machine Learning Research*, 15(1): 1215–1247, 2014.

Yee W. Teh, Dilan Gorur, and Zoubin Ghahramani. Stick-breaking construction for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 556–563, 2007. URL `http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS07_TehGG.pdf`.

Appendix A: Derivation to prove vanishing marginal feature probability

As described in the main text, the marginal probability of feature K can be written as the sum of terms, where each term corresponds to the product of Beta-distributed random variables:

$$\pi_K = \sum_{\boldsymbol{\epsilon} \in \mathcal{S}_{K-1}} \pi_{\boldsymbol{\epsilon}1} = \sum_{\boldsymbol{\epsilon} \in \mathcal{S}_{K-1}} \prod_{\boldsymbol{\epsilon}' < \boldsymbol{\epsilon}} \nu_{\boldsymbol{\epsilon}'} \tag{7}$$

where $\mathcal{S}_{K-1}$ is the set of binary strings of length $(K-1)$, the notation $\boldsymbol{\epsilon}' < \boldsymbol{\epsilon}$ refers to the ancestors[4] of the binary string $\boldsymbol{\epsilon}$, and $\nu_{\boldsymbol{\epsilon}'}$ refer to Beta-distributed random variables, either $\nu_{\boldsymbol{\epsilon}'} \sim \mathrm{Beta}(\alpha/K^p, 1)$ or $\nu_{\boldsymbol{\epsilon}'} \sim \mathrm{Beta}(1, \alpha/K^p)$ depending on whether the last component of $\boldsymbol{\epsilon}'$ is one or zero. Since the random variables $\nu_{\boldsymbol{\epsilon}'}$ are independent, the expectation of its product is the product of the individual expectations. We can thus compute its expectation in close-form and take the limit as $K \to \infty$. In particular:

$$\begin{aligned}
\mathbb{E}\left[\pi_K\right] &= \sum_{\boldsymbol{\epsilon} \in \mathcal{S}_{K-1}} \mathbb{E}\left[\pi_{\boldsymbol{\epsilon}1}\right] && (8)\\
&= \sum_{r=1}^{K} \left(\begin{array}{c} K-1 \\ r-1 \end{array} \right) \left(\frac{\alpha/K^P}{\alpha/K^P + 1} \right)^r \left(\frac{1}{\alpha/K^P + 1} \right)^{K-r}, && (9)
\end{aligned}$$

where r is the number of ones (active features) in the binary string $\{\boldsymbol{\epsilon}1\}$. Equation (9) arises from partitioning the random variables $\nu_{\boldsymbol{\epsilon}'}$ in two sets: those distributed according to $\mathrm{Beta}(\alpha/K^p, 1)$ and $\mathrm{Beta}(\alpha/K^p, 1)$.

$$\begin{aligned}
\lim_{K \to \infty} \mathbb{E}\left[\pi_K\right] &= \lim_{K \to \infty} \sum_{r=1}^{K} \left(\begin{array}{c} K-1 \\ r-1 \end{array} \right) \frac{\left(\alpha/K^P\right)^r}{\left(\alpha/K^P + 1\right)^K} && (10)\\
&= \lim_{K \to \infty} \frac{\alpha}{\alpha + K^p} = 0 \quad \forall p > 0 && (11)
\end{aligned}$$

We have demonstrated that $\lim\limits_{K \to \infty} \mathbb{E}\left[\pi_K\right] \to 0$, making it a valid Bayesian non-parametric prior.

Appendix B: Derivation of the predictive distribution

Let $\mathbf{Z}_{-(nk)}$ refer to matrix $\mathbf{Z}$ except the element z_{nk}, and $\mathbf{Z}_{-n}$ to all elements from $\mathbf{Z}$ except row $\boldsymbol{z}_n$. We compute the conditional probability $p\left(z_{nk} | \mathbf{Z}_{-(nk)}\right)$ as follows:

$$\begin{aligned}
p\left(z_{nk} | \mathbf{Z}_{-(nk)}\right) &\propto p\left(z_{nk}, \mathbf{Z}_{\neg n}, \boldsymbol{z}_{n,\neg k}\right) && (12)\\
&= \int_{\nu_{\boldsymbol{\epsilon}}} \int_{\boldsymbol{\nu}_{\neg \boldsymbol{\epsilon}}} p\left(z_{nk}, \mathbf{Z}_{\neg n}, \boldsymbol{z}_{n,\neg k}, \nu_{\boldsymbol{\epsilon}}, \boldsymbol{\nu}_{\neg \boldsymbol{\epsilon}}\right) d\boldsymbol{\nu}_{\neg \boldsymbol{\epsilon}} d\nu_{\boldsymbol{\epsilon}} && (13)\\
&\propto \int_{\boldsymbol{\nu}} p\left(z_{nk}, \boldsymbol{z}_{n,-k} | \boldsymbol{\nu}\right) p\left(\mathbf{Z}_{-n} | \boldsymbol{\nu}\right) p\left(\boldsymbol{\nu}\right) d\boldsymbol{\nu} && (14)\\
&\propto \int_{\boldsymbol{\nu}} p\left(\boldsymbol{z}_n | \boldsymbol{\nu}\right) p\left(\boldsymbol{\nu} | \mathbf{Z}_{-n}\right) d\boldsymbol{\nu} && (15)
\end{aligned}$$

This expression can be integrated out analytically due to the conjugacy property between Bernoulli and Beta distributions, such that:

$$\begin{aligned}
p\left(z_{nk} | \mathbf{Z}_{-(nk)}\right) &\propto \prod_{\boldsymbol{\epsilon} \in \mathcal{S}_n} \frac{\Gamma\left(\frac{\alpha}{K^p} + \phi_{\boldsymbol{\epsilon}1}^{-n} + z_{nk}\right) \Gamma\left(1 + \phi_{\boldsymbol{\epsilon}0}^{-n} + (1 - z_{nk})\right)}{\Gamma\left(\frac{\alpha}{K^p} + 1 + \phi_{\boldsymbol{\epsilon}}^{-n} + 1\right)} \frac{\Gamma\left(\frac{\alpha}{K^p} + 1 + \phi_{\boldsymbol{\epsilon}}^{-n}\right)}{\Gamma\left(\frac{\alpha}{K^p} + \phi_{\boldsymbol{\epsilon}1}^{-n}\right) \Gamma\left(1 + \phi_{\boldsymbol{\epsilon}0}^{-n}\right)}\\
&\propto \prod_{\boldsymbol{\epsilon} \in \mathcal{S}_n} \frac{\left(\frac{\alpha}{K^p} + \phi_{\boldsymbol{\epsilon}1}^{-n}\right)^{z_{nk}} \left(1 + \phi_{\boldsymbol{\epsilon}0}^{-n}\right)^{(1 - z_{nk})}}{\left(\frac{\alpha}{K^p} + 1 + \phi_{\boldsymbol{\epsilon}}^{-n}\right)} && (16)
\end{aligned}$$

where $\phi_{\boldsymbol{\epsilon}'}^{-n}$ is a sufficient statistic accounting for the number of times the binary vector $\boldsymbol{\epsilon}'$ appears in $\mathbf{Z}_{-n}$, and $\mathcal{S}_n$ is the set of subsequent partial binary vectors for observation n, i.e.,

[4] We follow the same notation as in [Adams et al., 2010].

$\mathcal{S}_n = \{z_{n1}, z_{n,(1:2)}, \ldots, z_{n,(1:K)}\}$. We denote $\Phi^{-n} = \{\phi_{\boldsymbol{\epsilon}'}\}_{\boldsymbol{\epsilon}' \in \mathcal{S}}$ as the tree of sufficient statistics given $\mathbf{Z}_{-n}$, i.e., the number of counts for each binary string $\boldsymbol{\epsilon}'$ of length $k = 1, \ldots K$ given $\mathbf{Z}_{-n}$. Equation (16) shows that the probability of each binary vector $\boldsymbol{z}_n$ only depends on the number of counts across the tree, i.e., this probability is independent of the sampling order of previous binary vectors.

Appendix C: Exchangeability

Algorithms for posterior inference are often greatly simplified by the assumption of exchangeability. A natural question to ask is whether the sequence of observations generated by the HSBP prior is exchangeable. According to Theorem 10 in [Broderick et al., 2013], any *regular*[5] exchangeable feature allocation (FA) admits a feature paintbox representation. Given any feature paintbox, we can build a random feature allocation for each observation independently, e.g., by sampling from the appropriate subset of conditional probability distributions as we go down the tree. Thus, the distribution $p(\boldsymbol{z}|\boldsymbol{\nu})$ is exchangeable, i.e., any permutation $\rho(\cdot)$ of the observation indices do not change the probability:

$$p\left(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3, \ldots |\boldsymbol{\nu}\right) = p\left(\boldsymbol{z}_{\rho(1)}, \boldsymbol{z}_{\rho(2)}, \boldsymbol{z}_{\rho(3)} \ldots |\boldsymbol{\nu}\right). \tag{17}$$

The exchangeability condition when $\boldsymbol{\nu}$ is integrated out also holds. Indeed, Eq. (6) shows that the probability of each binary vector z_n given all other rows $\mathbf{Z}_{-n}$ only depends on the previous number of counts across the tree, i.e., this probability is independent of the sampling order of previous binary vectors, proving exchangeability.

Appendix D: Linear feature paintbox model

Using the HSBP as a prior, we can build an infinite latent feature model with *a priori arbitrary correlations* between the latent features. Let $\mathbf{X} \in \mathbb{R}^{N \times D}$ denote the observation matrix. The complete generative process can be written as:

$$\begin{aligned} \boldsymbol{\nu} &\sim \text{HSBP}(\alpha, p) & \boldsymbol{z}_n &\sim \text{Multinomial}\left(1, \{\pi_{\boldsymbol{\epsilon}}\}_{\boldsymbol{\epsilon} \in \mathcal{S}_K}\right) \\ \mathbf{A} &\sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}) & \mathbf{X}|\mathbf{Z}, \mathbf{A} &\sim \mathcal{N}(\mathbf{Z}\mathbf{A}, \sigma_x^2 \mathbf{I}), \end{aligned} \tag{18}$$

where $\mathbf{A} \in \mathbb{R}^{K \times D}$ is the dictionary matrix, $\mathbf{Z} \in \{0,1\}^{N \times K}$ is the feature allocation matrix, $\boldsymbol{z}_n \in \{0,1\}^{1 \times K}$ is the feature allocation vector for observation n, $\pi_{\boldsymbol{\epsilon}}$ refer to the joint probability distribution of the latent features represented by the binary string $\boldsymbol{\epsilon}$, and $\mathcal{S}_K$ is the set of all binary vectors of length K.

Inference. We perform inference using a collapsed Gibbs sampler with row-wise Metropolis-Hasting (MH) proposals $p(\boldsymbol{z}_n|\mathbf{Z}_{-n})$ stated in Equation 6. In particular, we compute:

$$p(\boldsymbol{z}_n|\mathbf{X}, \mathbf{Z}_{-n}) \propto p(\mathbf{X}|\boldsymbol{z}_n) p(\boldsymbol{z}_n|\mathbf{Z}_{-n}). \tag{19}$$

For the row-wise sampler, we propose an entire row $\boldsymbol{z}_n^*$ for each observation n iteratively, and accept with probability:

$$a = \min\left(1, \frac{p(\mathbf{X}|\mathbf{Z}^*) p(\mathbf{Z}^*)}{p(\mathbf{X}|\mathbf{Z}) p(\mathbf{Z})} \frac{g(\mathbf{Z}|\mathbf{Z}^*)}{g(\mathbf{Z}^*|\mathbf{Z})}\right), \tag{20}$$

where g is the proposal distribution. Let us propose a new vector $\boldsymbol{z}_n^* \sim p(\boldsymbol{z}|\mathbf{Z}_{-n})$ according to Eq. (16). In that case, Eq. (20) simplifies to a ratio of likelihoods:

$$a = \min\left(1, \frac{p(\mathbf{X}|\mathbf{Z}^*)}{p(\mathbf{X}|\mathbf{Z})} \frac{p(\boldsymbol{z}_n^*|\mathbf{Z}_{-n}) p(\mathbf{Z}_{-n})}{p(\boldsymbol{z}_n|\mathbf{Z}_{-n}) p(\mathbf{Z}_{-n})} \frac{p(\boldsymbol{z}_n|\mathbf{Z}_{-n})}{p(\boldsymbol{z}_n^*|\mathbf{Z}_{-n})}\right) = \min\left(1, \frac{p(\mathbf{X}|\mathbf{Z}^*)}{p(\mathbf{X}|\mathbf{Z})}\right) \tag{21}$$

[5] Regular means that with probability one, there is no index (row) with unique feature collection.