Hierarchical Stick-breaking Feature Paintbox

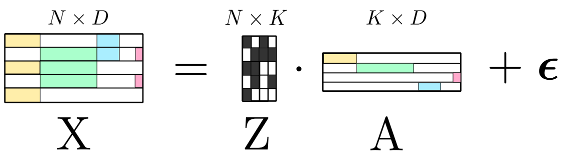
Melanie F. Pradier, Weiwei Pan, Morris Yau, Rachit Singh, and Finale Doshi-Velez

Motivation

Latent feature models decompose observed attributes of complex data into combinations of simple factors or features. We present:

- a novel feature model with a flexible nonparametric prior that allows for arbitrary correlations amongst the latent features
- tractable inference for our model via a collapsed Gibbs sampler

The HSBP Feature Model



$$\boldsymbol{\nu} \sim \text{HSBP}(\alpha, p) \qquad \boldsymbol{z}_n \sim \text{Mult}\left(1, \{\pi_{\boldsymbol{\epsilon}}\}_{\boldsymbol{\epsilon} \in \mathcal{S}_K}\right)$$

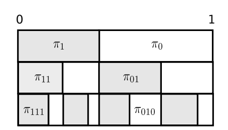
$$\mathbf{A} \sim \mathcal{N}(0, \sigma_0^2 \mathbf{I}) \qquad \mathbf{X}|\mathbf{Z}, \mathbf{A} \sim \mathcal{N}(\mathbf{ZA}, \sigma_x^2 \mathbf{I}),$$

The Feature Paintbox Prior

The hierarchical stick-breaking paintbox process (HSBP) has the following iterative construction:

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

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



Canonical paintbox example.

We can sample each row $\boldsymbol{z}_n$ element-wise from each Bernoulli conditional probability distribution by traversing the tree top down:

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

Properties of HSBP

Vanishing marginal feature probability. The proposed iterative process gives rise to valid feature allocations if π_K vanishes as $K \to \infty$. The marginal probability of feature K can be written as:

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

The expectation $\mathbb{E}[\pi_K]$ can be written in closed-form in the limit $K \to \infty$:

$$\begin{aligned} \lim_{K\to\infty} \mathbb{E}[\pi_K] &= \lim_{K\to\infty} \sum_{r=1}^{K} \binom{K-1}{r-1} \frac{(\alpha/K^P)^r}{(\alpha/K^P + 1)^K} \\ &= \lim_{K\to\infty} \frac{\alpha}{\alpha + K^p} = 0 \quad \forall p > 0 \end{aligned} \tag{2}$$

Exchangeability We can prove exchangeability if for any $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$, it holds that:

$$p(\boldsymbol{z}_2, \boldsymbol{z}_3 | \boldsymbol{z}_1) \stackrel{d}{=} p(\boldsymbol{z}_3, \boldsymbol{z}_2 | \boldsymbol{z}_1)$$

It is easy to show in Eq. (4) that the probability of a new vector $p\left(\boldsymbol{z}_n | \mathbf{Z}_{1:(n-1)}\right)$ only depends on the previous number of counts along the branch corresponding to $\boldsymbol{z}_n$, independently of the order of previous features.

Inference

We derive a collapsed Gibbs sampler:

$$p\left(z_{nk} | \mathbf{Z}_{-(nk)}\right) \propto \int_{\boldsymbol{\nu}} p\left(\boldsymbol{z}_n | \boldsymbol{\nu}\right) p\left(\boldsymbol{\nu} | \mathbf{Z}_{-n}\right) d\boldsymbol{\nu} \tag{3}$$

$$\propto \prod_{\boldsymbol{\epsilon} \in \mathcal{S}_n} \frac{\left(\frac{\alpha}{K^p} + \phi_{\boldsymbol{\epsilon}1}^{-n}\right)^{z_{nk}} \left(1 + \phi_{\boldsymbol{\epsilon}0}^{-n}\right)^{(1-z_{nk})}}{\left(\frac{\alpha}{K^p} + 1 + \phi_{\boldsymbol{\epsilon}}^{-n}\right)}, \tag{4}$$

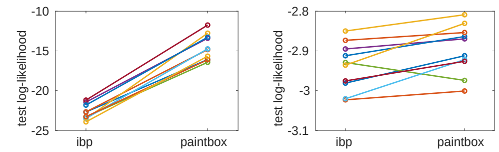
where $\phi_{\boldsymbol{\epsilon}'}^{-n}$ is a sufficient statistic accounting for the number of times that the binary vector $\boldsymbol{\epsilon}'$ appears in $\mathbf{Z}_{-n}$, and $\mathcal{S}_n$ is the set of subsequent partial binary vectors for observation n, i.e., $\mathcal{S}_n = \{z_{n1}, z_{n,(1:2)}, \ldots, z_{n,(1:K)}\}$.

More efficiently, we propose a Metropolis-Hasting within Gibbs with row-proposals according to Eq. (1).

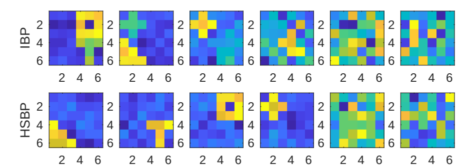
Results

We compare an infinite latent feature model with Gaussian likelihood using either an Indian Buffet Process or HSBP prior. Considered datasets: (left) correlated toy images ($N = 300$, $D = 36$), and (right) breast cancer dataset ($N = 500$, $D = 30$).

1. The HSBP prior improves performance substantially in the held-out data.



2. The HSBP prior improves recovery of the true components.



Discussion

- Paintbox as binary tree of conditional probabilities
- IBP generalization by accounting for both positive and negative correlations among features
- Better reconstruction + interpretable dictionaries
- Next: optimization, scalability, non-linear models

Acknowledgements

We thank the Harvard Data Science Initiative, the Center for Research on Computation and Society, and Institute of Applied Computational Sciences.

References

1. Thomas L. Griffiths and Zoubin Ghahramani. The Indian Buffet Process: An Introduction and Review. Journal of Machine Learning Research, 12:1185–1224, 2011.
2. Tamara Broderick, Jim Pitman, and Michael I. Jordan. Feature Allocations, Probability Functions, and Paintboxes. Bayesian Analysis, 8(4):801–836, December 2013.
3. Ryan Prescott Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-Structured Stick Breaking Processes for Hierarchical Data. arXiv:1006.1062 [stat], June 2010. arXiv: 1006.1062.