Structured Variational Autoencoders for Beta-Bernoulli Processes

Rachit Singh* Jeffrey Ling* Finale Doshi-Velez

Harvard University

Summary

- Bayesian nonparametrics allows a model to adapt as data size grows, but inference can be difficult
- Beta-Bernoulli processes, a.k.a. Indian buffet processes, are priors over infinite dimensional binary feature matrices
- Variational autoencoders allow inference over extremely complicated likelihoods
- Structured VAE improves IBP inference on a mean field baseline

Beta-Bernoulli Process

- Defines a distribution on latent feature allocations $\mathbf{Z} \in \{0,1\}^{N \times K^+}$, where $z_{n,k}$ represents feature k for data point n
- Stick breaking process:

$$\nu_k \sim \text{Beta}(\alpha, 1); \quad \pi_k = \prod_{j=1}^{k} \nu_j; \quad z_{n,k} \sim \text{Bern}(\pi_k)$$

- Generative model for data $\mathbf{X} \in \mathbb{R}^{N \times D}$ with likelihood $p_\theta(\mathbf{X}|\mathbf{Z})$:

$$\mathbf{Z}, \boldsymbol{\nu} \sim \text{IBP}(\alpha); \quad \mathbf{A}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K^+}); \quad \mathbf{x}_n \sim p_\theta(\mathbf{x}_n | \mathbf{Z}_n \odot \mathbf{A}_n)$$

- Prior work in IBP VI focuses on exponential family likelihoods. We permit more flexible likelihoods (e.g. deep neural networks).

Variational Inference

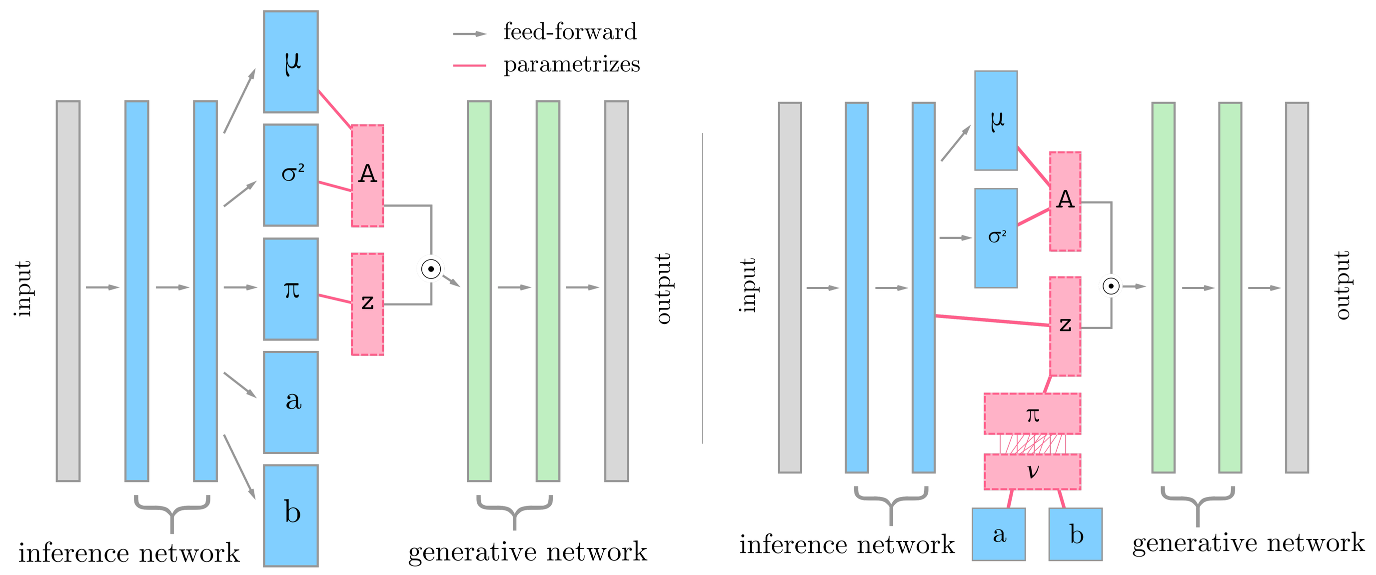


Figure: Mean field (left) and structured (right, ours) variational inferences for the Indian Buffet Process. The inference/generative networks can be arbitrary.

Latent variables:

$\boldsymbol{\nu} = \{\nu_k\}_{k=1}^{\infty}$	global (in bijection to weights π_k)
$\boldsymbol{\psi}_n = \{\mathbf{Z}_n, \mathbf{A}_n\}$	local, with $\mathbf{Z}_n \in \{0,1\}^{K^+}, \mathbf{A}_n \in \mathbb{R}^{K^+}$
θ	likelihood parameters

- Learn the posterior $p(\boldsymbol{\nu}, \{\boldsymbol{\psi}_n\}_{n=1}^N | \mathbf{X})$ using **variational posterior** $q(\boldsymbol{\nu}, \{\boldsymbol{\psi}_n\}_{n=1}^N | \mathbf{X})$
- Assume a truncated posterior with support over finitely sized matrices (i.e. finite K^+)
- Parameters trained using minibatch stochastic gradient descent.

Inference

Mean Field (MF-IBP)

- Assume that q factors fully:

$$q_{\text{MF-IBP}}(\boldsymbol{\nu}, \{\boldsymbol{\psi}_n\}_{n=1}^N) = \prod_{k=1}^{K} q(\nu_k) \prod_{n=1}^{N} q(z_{n,k}) q(\mathbf{A}_n)$$

- Variational approximation:

$$q(\nu_k) = \text{Beta}(\nu_k | a_k(\mathbf{x}_n), b_k(\mathbf{x}_n))$$
$$q(z_{n,k}) = \text{Bern}(z_{n,k} | \pi_k(\mathbf{x}_n))$$
$$q(\mathbf{A}_n) = \mathcal{N}(\mathbf{A}_n | \mu(\mathbf{x}_n), \text{diag}(\sigma^2(\mathbf{x}_n)))$$

$a_k, b_k, \pi_k, \mu, \log\sigma$ are the outputs of neural networks which have input $\mathbf{x}_n$.

Extension: Structured (S-IBP)

- Factor q hierarchically over global and local variables:

$$q_{\text{S-IBP}}(\boldsymbol{\nu}, \{\boldsymbol{\psi}_n\}) = \prod_{k=1}^{K} q(\nu_k) \prod_{n=1}^{N} q(z_{n,k} \mid \nu_k) q(\mathbf{A}_n)$$

- Variational approximation:

$$q(\nu_k) = \text{Beta}(\nu_k | a_k, b_k)$$
$$q(z_{n,k} \mid \nu_k) = \text{Bern}(z_{n,k} | \pi_k)$$
$$q(\mathbf{A}_n) = \mathcal{N}(\mathbf{A}_n | \mu(\mathbf{x}_n), \text{diag}(\sigma^2(\mathbf{x}_n)))$$
$$\pi_k := \Pi_{j=1}^{k} \nu_j$$

a_k, b_k are global parameters to be learned.

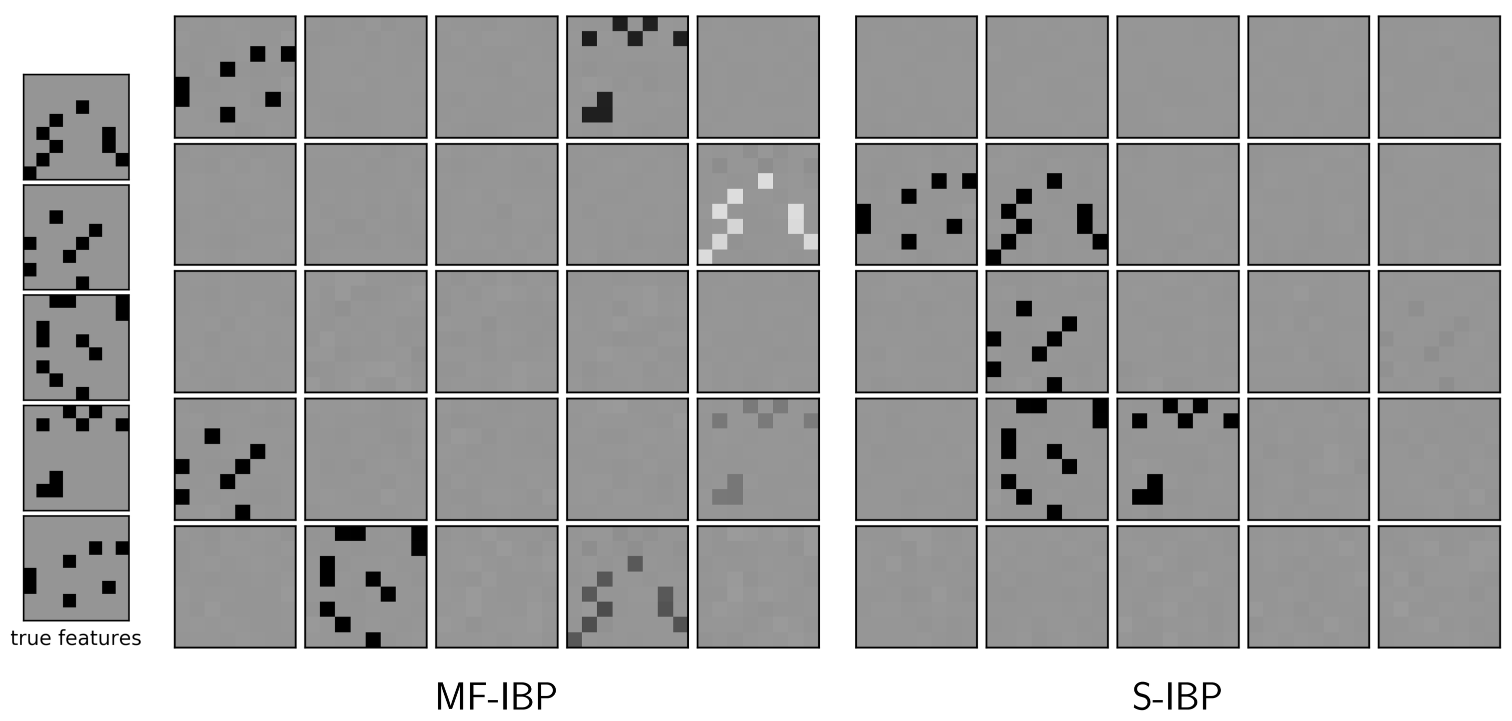
MF-IBP negative ELBO loss:

$$\mathcal{L}_{\text{MF-IBP}} = \sum_{n=1}^{N} -\mathbb{E}_q[\log p(\mathbf{x}_n | \boldsymbol{\psi}_n)] + \text{KL}(q(\mathbf{Z}_n | \boldsymbol{\nu}) \parallel p(\mathbf{Z}_n | \boldsymbol{\nu})) + \text{KL}(q(\mathbf{A}_n) \parallel p(\mathbf{A}_n)) + \text{KL}(q(\boldsymbol{\nu}) \parallel p(\boldsymbol{\nu}))$$

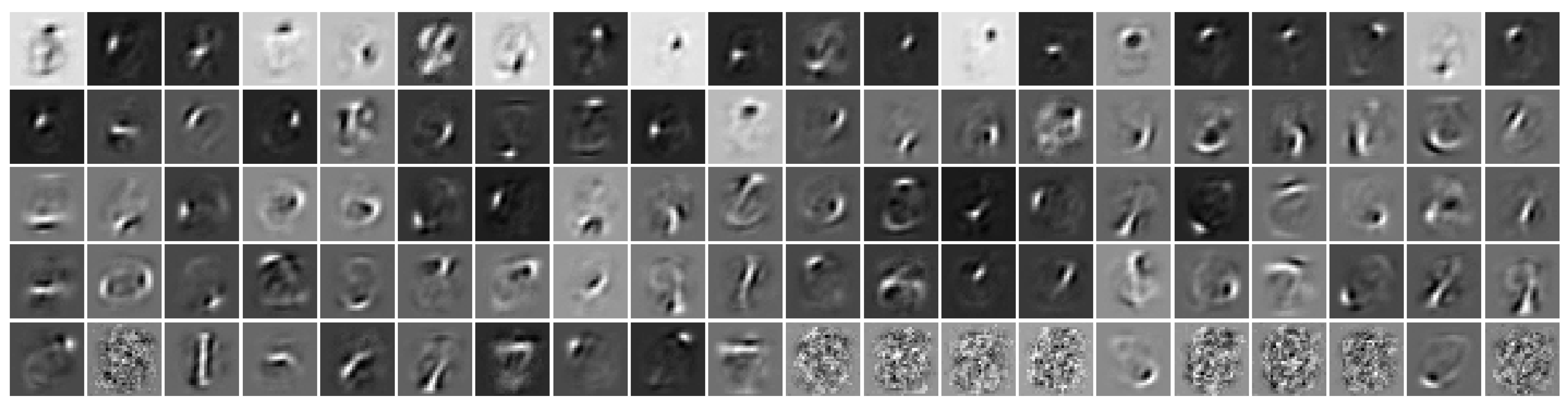
S-IBP loss derived from SSVI:

$$\mathcal{L}_{\text{S-IBP}} = \text{KL}(q(\boldsymbol{\nu}) \parallel p(\boldsymbol{\nu})) + \sum_{n=1}^{N} -\mathbb{E}_q[\log p(\mathbf{x}_n | \boldsymbol{\psi}_n)] + \text{KL}(q(\mathbf{Z}_n | \boldsymbol{\nu}) \parallel p(\mathbf{Z}_n | \boldsymbol{\nu})) + \text{KL}(q(\mathbf{A}_n) \parallel p(\mathbf{A}_n))$$

Visualizations



Features inferred on a synthetic dataset generated by a linear IBP model. Left is true features. Black is high, grey is 0, and white is low.



Inferred features (truncation 100) learned from MNIST in a linear IBP model. Note that many of the 'noise' features are never activated.

Training

Two methods to compute backpropagation gradients:

- Black box variational inference (BBVI): directly backpropagate with score estimator
- Replace non-reparameterizable variables in variational approximation:
  - Bernoulli → Gumbel softmax (Maddison et al. 2017, Jang et al. 2017)
  - Beta → Kumaraswamy (Nalisnick & Smyth 2017)

Evaluation

Evaluate using IWAE metric (Importance Weighted Auto Encoder):

$$\log p(\mathbf{x}) = \log \int_{\boldsymbol{\psi},\boldsymbol{\nu}} p(\mathbf{x}, \boldsymbol{\nu}, \boldsymbol{\psi}) \frac{q(\boldsymbol{\nu}, \boldsymbol{\psi})}{q(\boldsymbol{\nu}, \boldsymbol{\psi})} d\boldsymbol{\nu}\, d\boldsymbol{\psi}$$
$$\geq \mathbb{E}_q \left[\log \frac{1}{m} \sum_{j=1}^{m} \frac{p(\boldsymbol{\nu}_j)}{q(\boldsymbol{\nu}_j)} \prod_{n=1}^{N} \frac{p(\mathbf{x}_n, \boldsymbol{\psi}_{n,j} | \boldsymbol{\nu}_j)}{q(\boldsymbol{\psi}_{n,j} | \boldsymbol{\nu}_j)} \right]$$

	MNIST IWAE		Omniglot IWAE	
Model	Train	Test	Train	Test
MF-IBP BBVI	102.6	104.5	129.4	134.5
MF-IBP Gumbel	94.2	96.4	125.0	129.5
S-IBP BBVI	93.8	96.2	115.2	124.5
S-IBP Gumbel	81.7	**86.5**	101.4	**113.0**

Table: MNIST and Omniglot IWAE test results.

Open Source Code

Code: `https://github.com/rachtsingh/ibp_vae`.

- All models written in **Pytorch**
- Implements **GPU versions** of `lgamma`, `polygamma`, and sampling from Gamma/Beta distributions (for the MF-variants).

Runtime: (S-IBP Gumbel) 17.1 s/epoch on MNIST, (MF-IBP BBVI) 27.7 s/epoch on MNIST (38% speedup) - due mostly to speedups from Concrete/Kumaraswamy over BBVI.

Conclusion

- Demonstrates the utility of VAE inference for Beta-Bernoulli processes
- Structured variational approximation can improve on existing mean field methods
- Open source code will allow for future research

NIPS 2017 Advances in Approximate Bayesian Inference

{rachitsingh@college,jeffreyling@alumni,finale@seas}.harvard.edu