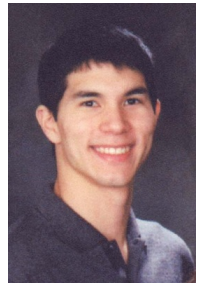
Variational Inference for the Indian Buffet Process

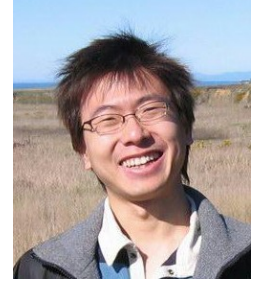
Finale Doshi-Velez[†]
Cambridge University

Kurt T. Miller[†]
UC Berkeley



Jurgen Van Gael[†]
Cambridge University

Yee Whye Teh
Gatsby Unit



[†] Authors contributed equally

Motivating example

We are interested in extracting unobserved features from observed data. For example:

Motivating example

We are interested in extracting unobserved features from observed data. For example:

- Latent classes $\Rightarrow$ Mixture models

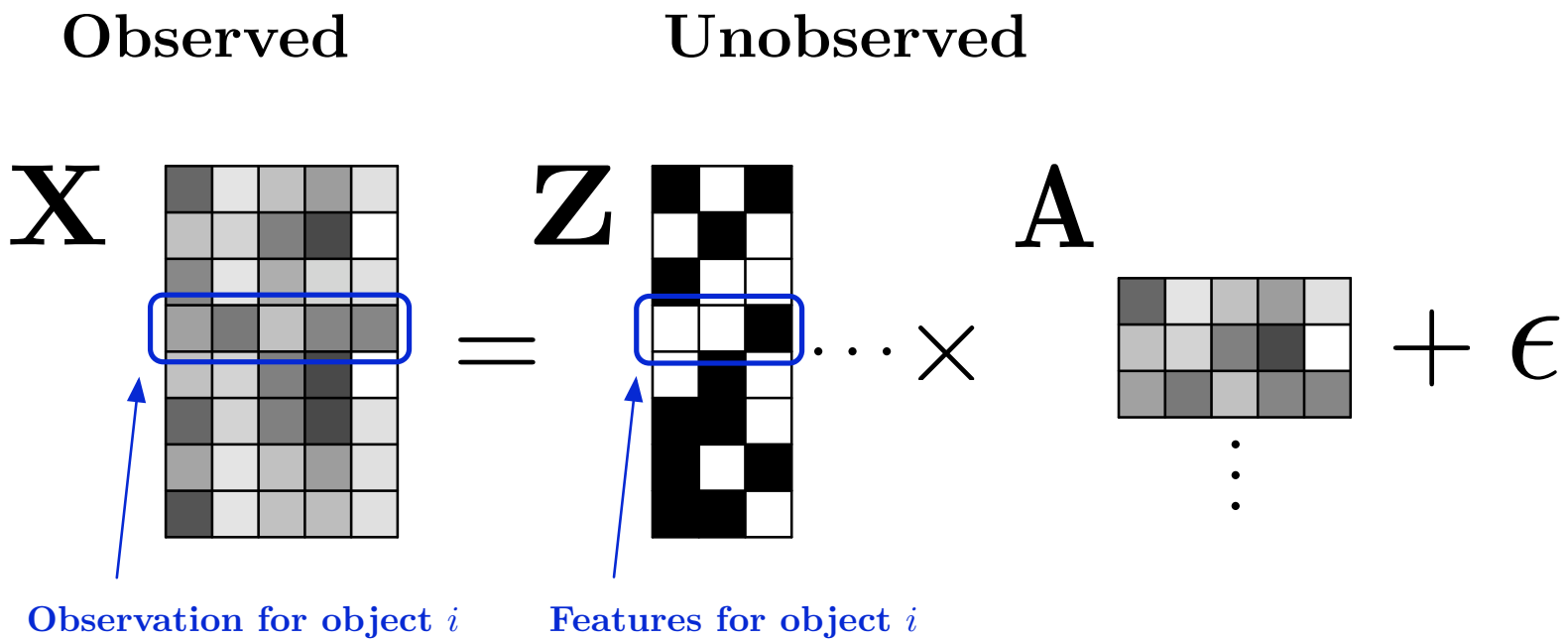
Motivating example

We are interested in extracting unobserved features from observed data. For example:

- Latent classes $\Rightarrow$ Mixture models
- Latent features $\Rightarrow$ Latent feature models

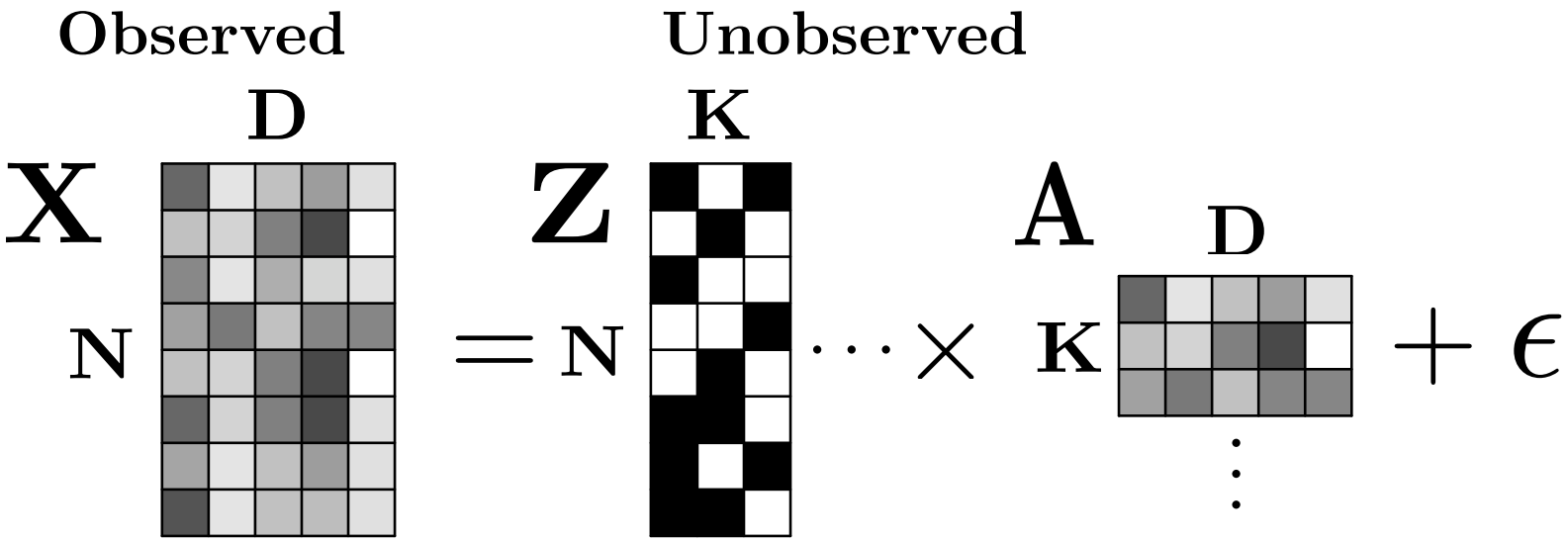
Linear Gaussian Latent Feature Model

We will focus on one example of a latent feature model:



Linear Gaussian Latent Feature Model

We will focus on one example of a latent feature model:



- N = Number of data points
- D = Dimension of observed data
- K = Number of latent features

Linear Gaussian Model Latent Feature Model

Goal: Infer Z and A given data X.

Linear Gaussian Model Latent Feature Model

Goal: Infer Z and A given data X.

Approach: Bayes' rule:

$$p(Z, A|X) \quad \propto \quad \underbrace{p(X|Z, A)p(A)}_{\text{Model specific}} \times \quad \underbrace{p(Z)}_{\text{Prior on binary matrices}}$$

Linear Gaussian Model Latent Feature Model

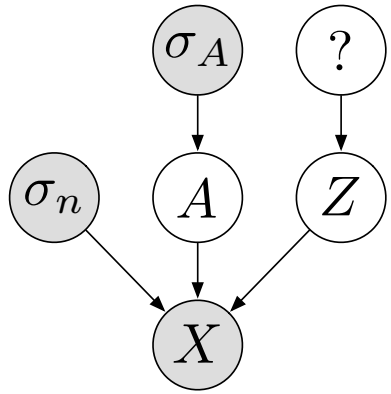
Goal: Infer Z and A given data X.

Approach: Bayes' rule:

$$p(Z, A|X) \quad \propto \quad \underbrace{p(X|Z, A)p(A)}_{\text{Model specific}} \times \quad \underbrace{p(Z)}_{\text{Prior on binary matrices}}$$

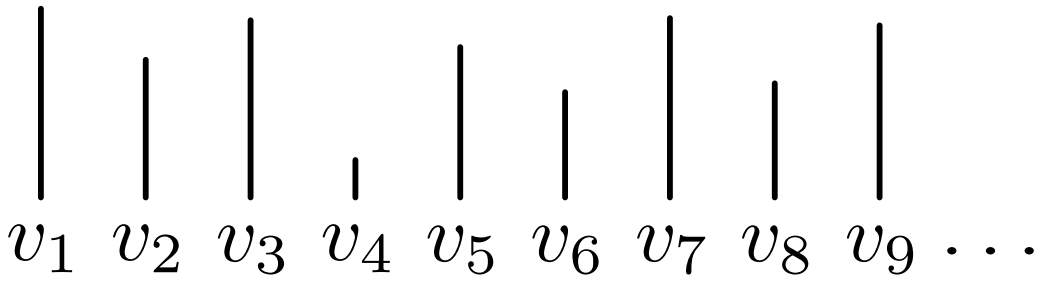
In the linear Gaussian model, we use

- $p(X|Z, A) \sim \mathcal{N}(ZA, \sigma_n^2 I)$
- $p(A) \sim \mathcal{N}(0, \sigma_A^2 I)$
- $p(Z) \sim ?$



The Indian Buffet Process - Stick-breaking construction

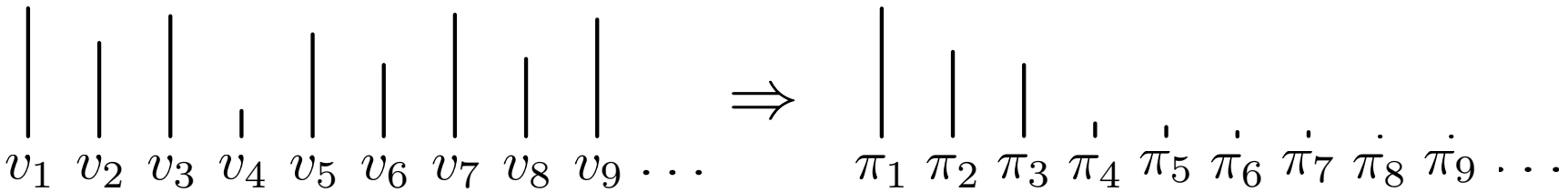
- First generate $v_1, v_2, \cdots \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1)$.



(Teh et al, 2007)

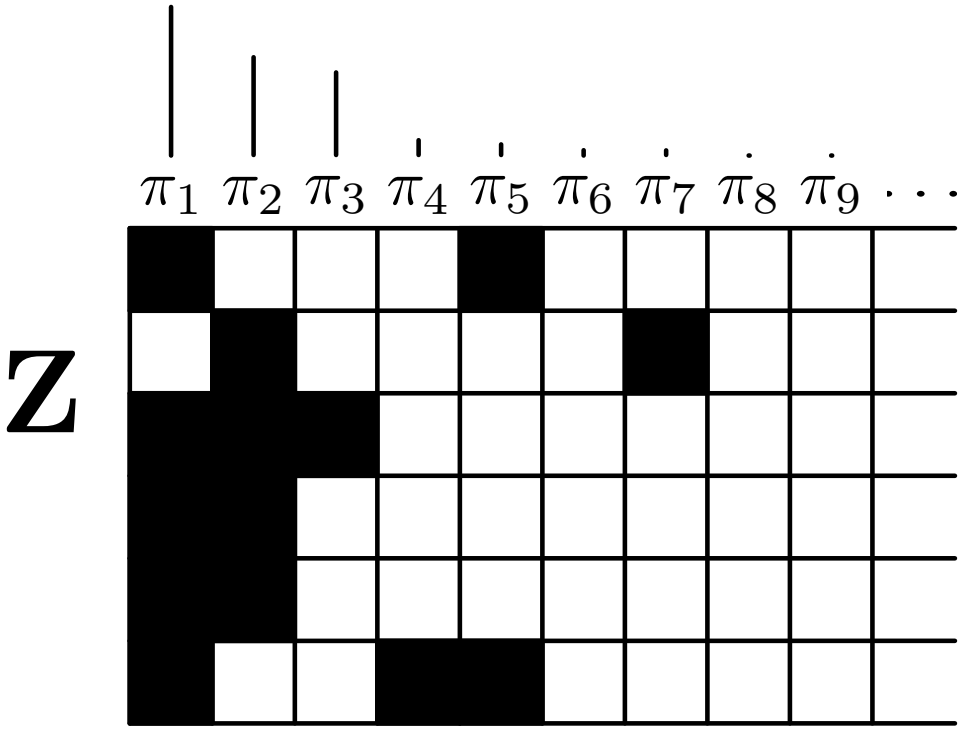
The Indian Buffet Process - Stick-breaking construction

- First generate $v_1, v_2, \cdots \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1)$.
- Let $\pi_i = \prod_{j=1}^{i} v_j$.



The Indian Buffet Process - Stick-breaking construction

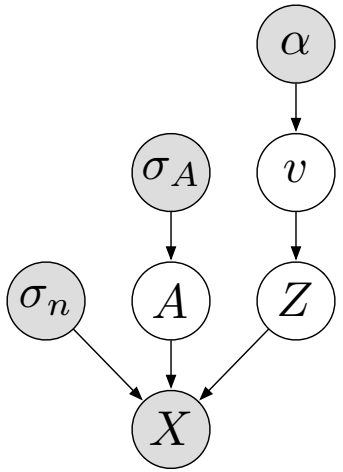
- First generate $v_1, v_2, \cdots \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, 1)$.
- Let $\pi_i = \prod_{j=1}^{i} v_j$.
- Sample $z_{nk} \sim \text{Bernoulli}(\pi_k)$.



Full Linear Gaussian Latent Feature Model

Model:

- $p(X|Z,A) \sim \mathcal{N}(ZA, \sigma_n^2 I)$
- $p(A) \sim \mathcal{N}(0, \sigma_A^2 I)$
- $p(Z) \sim \text{IBP}(\alpha)$

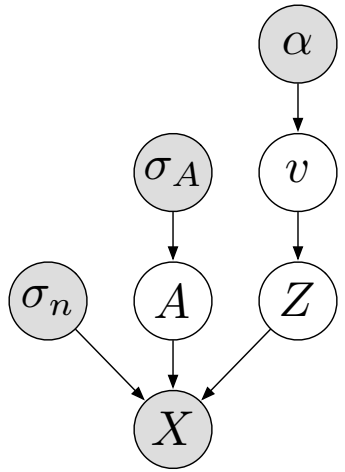


Given X, how do we do inference on Z and A?

Full Linear Gaussian Latent Feature Model

Model:

- $p(X|Z,A) \sim \mathcal{N}(ZA, \sigma_n^2 I)$
- $p(A) \sim \mathcal{N}(0, \sigma_A^2 I)$
- $p(Z) \sim \text{IBP}(\alpha)$

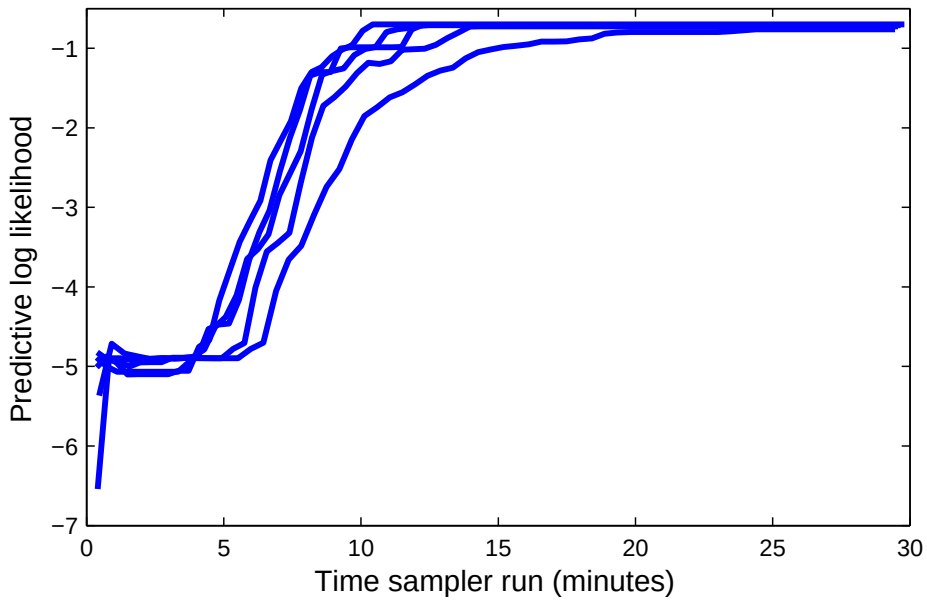


Given X, how do we do inference on Z and A?

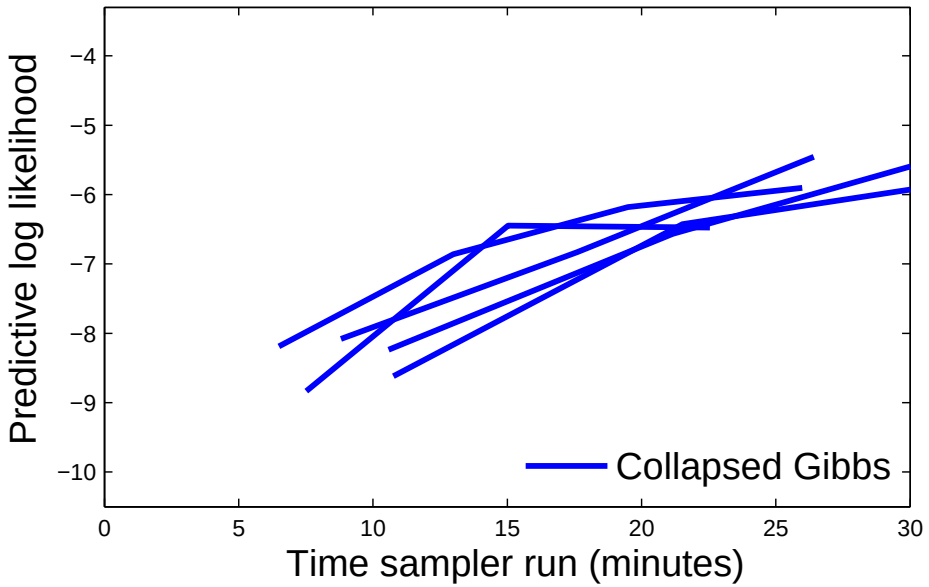
- Even for finite K, there are 2^{NK} possible Z.
- Many local optima.

Inference in the Linear Gaussian Model

$N = 100,\ D = 500, K = 25$

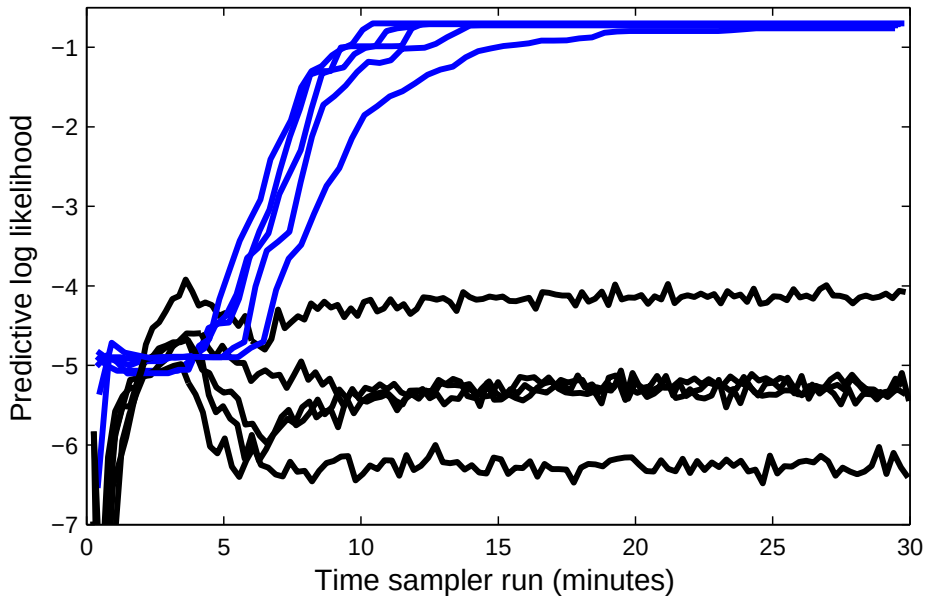


$N = 500,\ D = 500, K = 25$

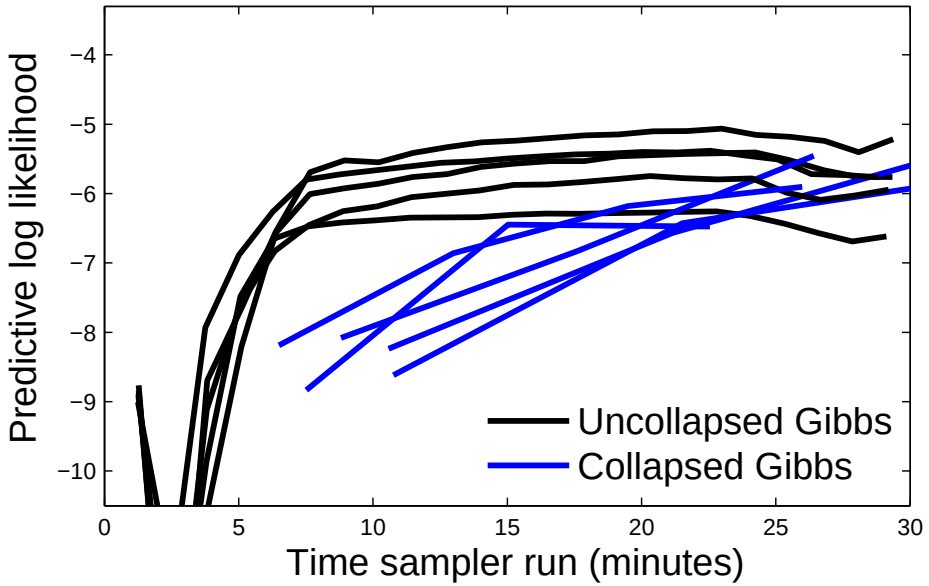


Inference in the Linear Gaussian Model

$N = 100,\ D = 500, K = 25$

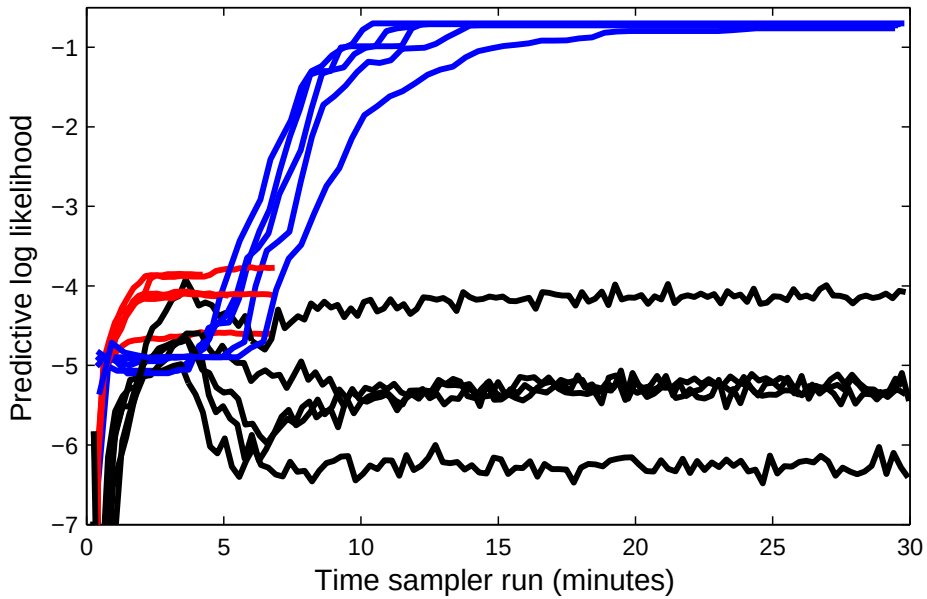


$N = 500,\ D = 500, K = 25$

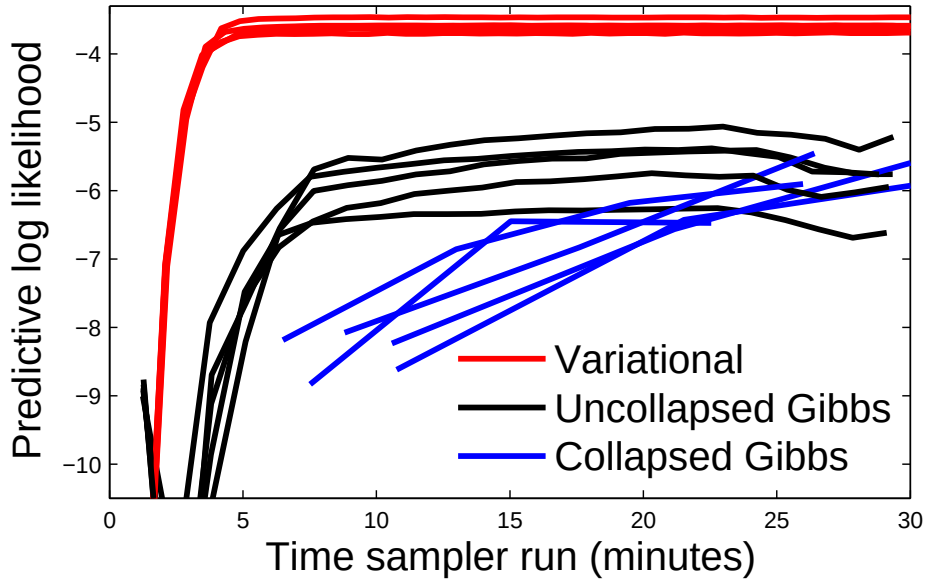


Inference in the Linear Gaussian Model

$N = 100,\ D = 500, K = 25$



$N = 500,\ D = 500, K = 25$



Mean Field Variational Inference

Approximate $p(Z, A|X)$ with distribution $q(Z, A)$ from a family Q that is "close" to $p(Z, A|X)$.

Mean Field Variational Inference

Approximate $p(Z, A|X)$ with distribution $q(Z, A)$ from a family Q that is "close" to $p(Z, A|X)$.

How do we define "close"? We will attempt to find

$$q(Z, A) \quad = \quad \arg\min_{q \in Q} D(q(Z, A)||p(Z, A|X)).$$

How do we choose Q?

$p(Z, A|X)$ is a distribution over infinitely many features.

Trick (Blei and Jordan, 2004): Let Q be a truncated family where we assume that Z is nonzero in at most the first K columns.

Why can we do this? Intuitively, the probability π_k that z_{nk} is one decreases exponentially quickly.

Truncation bound

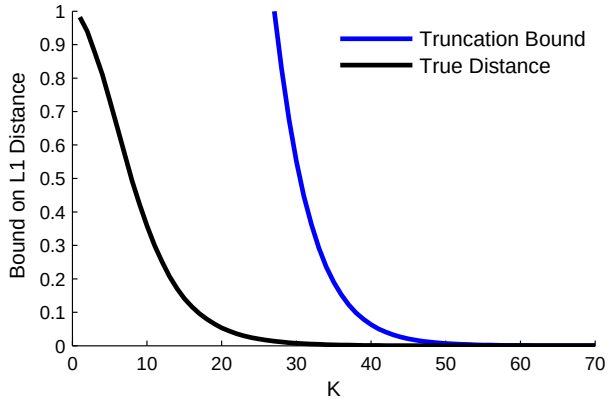
More formally, let $m_K(X)$ be the marginal of X when Z and A are integrated out when we truncate the stick-breaking construction at column K.

Truncation bound

More formally, let $m_K(X)$ be the marginal of X when Z and A are integrated out when we truncate the stick-breaking construction at column K.

Then we can show

$$\frac{1}{4}\int |m_K(X) - m_\infty(X)| dX \leq 1 - \exp\left(-N\alpha\left(\frac{\alpha}{\alpha+1}\right)^K\right).$$



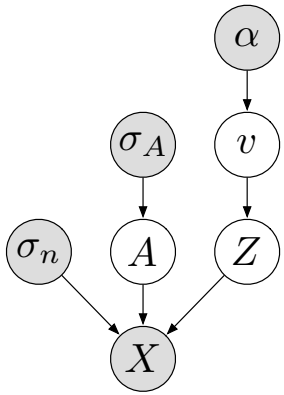
This is the first such bound for the IBP and can serve as a guideline for how to choose K for the family Q.

How do we choose Q?

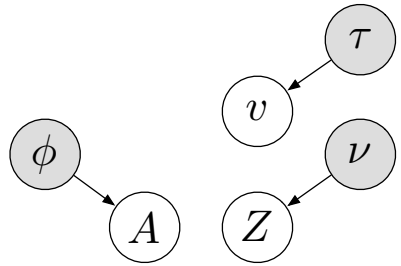
We let our family Q be the parameterized family (introducing the stick-breaking variables v)

$$q(Z, A, v) = q_\nu(Z) q_\phi(A) q_\tau(v)$$

True distribution:



Variational distribution:



- $q_{\nu_{nk}}(z_{nk}) = \text{Bernoulli}(z_{nk}; \nu_{nk})$
- $q_{\phi_k}(A_{k\cdot}) = \mathcal{N}(A_{k\cdot}; \bar{\phi}_k, \Phi_k)$
- $q_{\tau_k}(v_k) = \text{Beta}(v_k; \tau_{k1}, \tau_{k2})$

Inference

Inference now reduces to finding *variational parameters* (τ, ϕ, Φ, ν) such that $q \in Q$ is "close" to p.

$$(\tau, \phi, \Phi, \nu) \quad = \quad \arg\min_{q \in Q} D(q(Z, A)||p(Z, A|X)).$$

This is not a convex optimization, so we can only hope to find a local optimum.

Inference

Inference now reduces to finding *variational parameters* (τ, ϕ, Φ, ν) such that $q \in Q$ is "close" to p.

$$(\tau, \phi, \Phi, \nu) \quad = \quad \arg\min_{q \in Q} D(q(Z, A)||p(Z, A|X)).$$

This is not a convex optimization, so we can only hope to find a local optimum.

$\Rightarrow$ Parameter updates done iteratively.

Parameter updates

Many calculations are straightforward exponential family calculations.

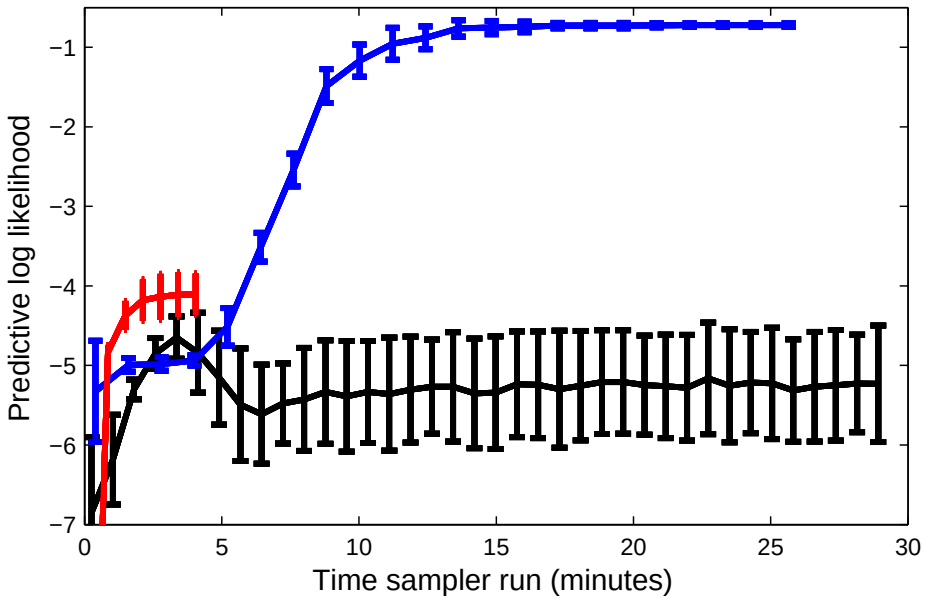
The only nontrivial calculation is $\mathbb{E}_{\boldsymbol{v},\boldsymbol{Z}}\left[\log p(Z_{nk}|\boldsymbol{v})\right]$ which requires evaluating

$$\mathbb{E}_{\boldsymbol{v}}\left[\log\left(1-\prod_{m=1}^{k} v_m\right)\right]$$

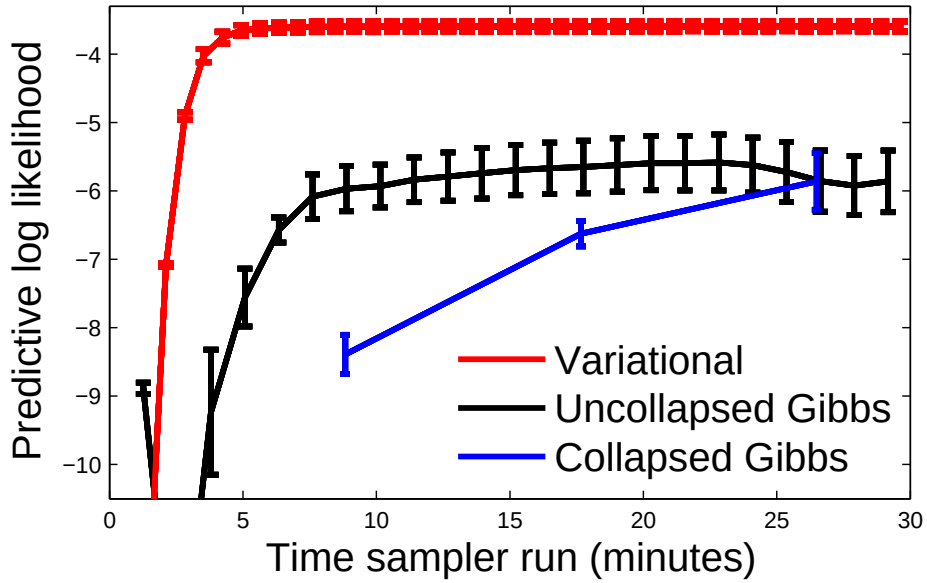
We provide an efficient way to lower bound this term.

Results: Synthetic data

$N = 100,\ D = 500, K = 25$



$N = 500,\ D = 500, K = 25$

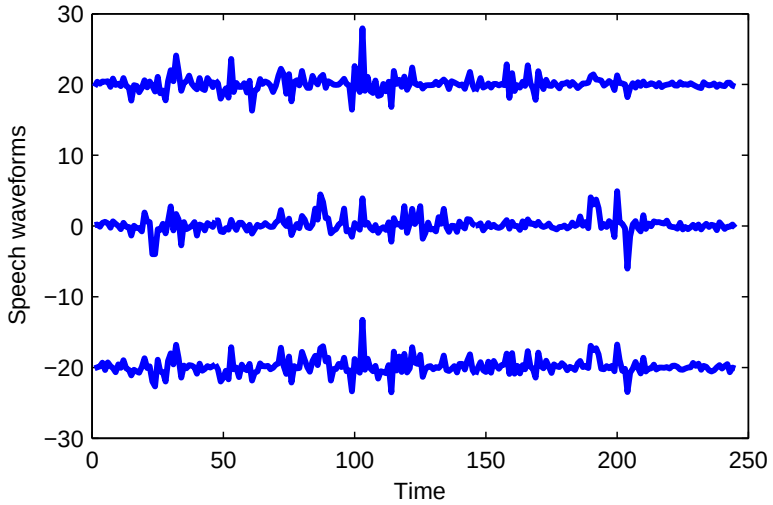


Results: Real data

2 data sets:

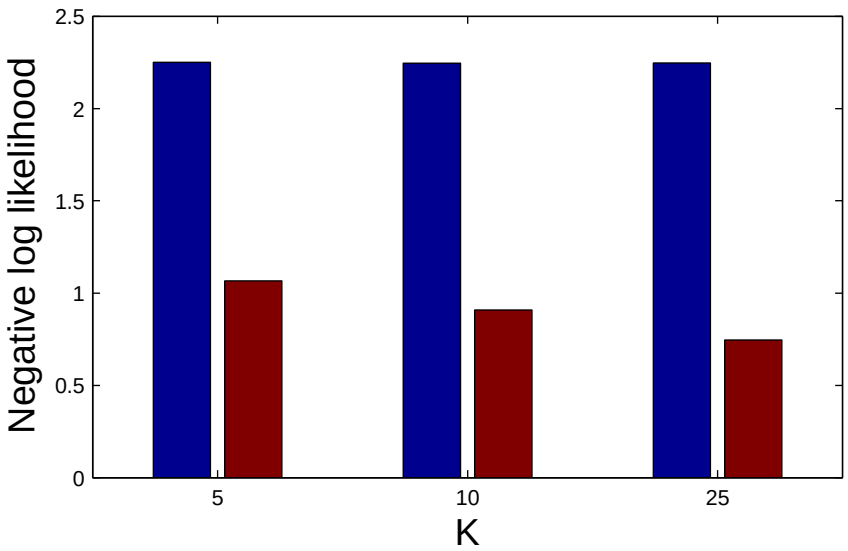
- Yale faces data set: linear Gaussian model, $N = 721, D = 1024$ (32$\times$ 32 images)

- Speech data set: iICA model, $N = 245, D = 10$



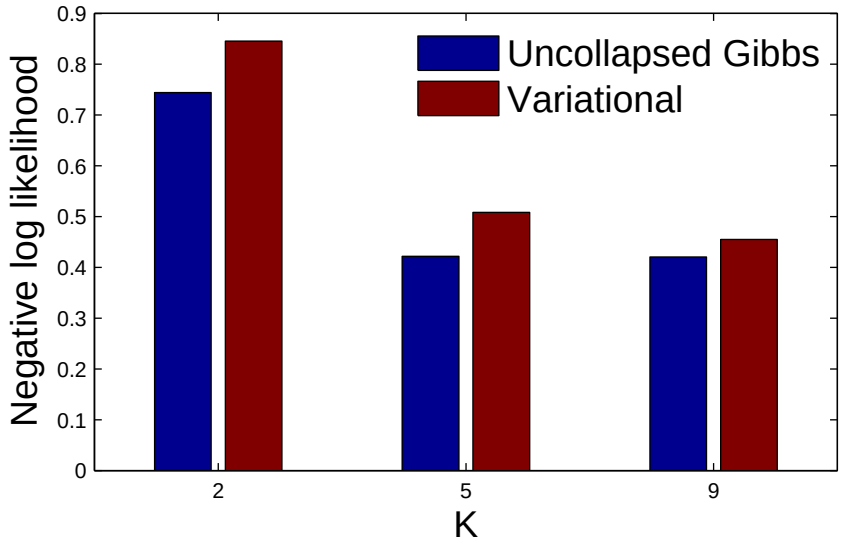
Results: Real data

Faces data set: $N = 721, D = 1024$



Large D, N - Variational helps

Speech data set: $N = 245, D = 10$



Small N, D - Variational does not help

Summary

- We present the first variational inference algorithm for the IBP.
- For large N and D, it finds better local optima than the samplers.
- We also present the first truncation bound for the IBP.

Code will be available soon from our websites.

Questions?