Variational inference for Dirichlet process mixtures

Mikolaj Kasprzak (mjkasprz@mit.edu)
Ali Ramadhan (alir@mit.edu)

6.435 Bayesian Modeling & Inference Lecture 14

April 7, 2022

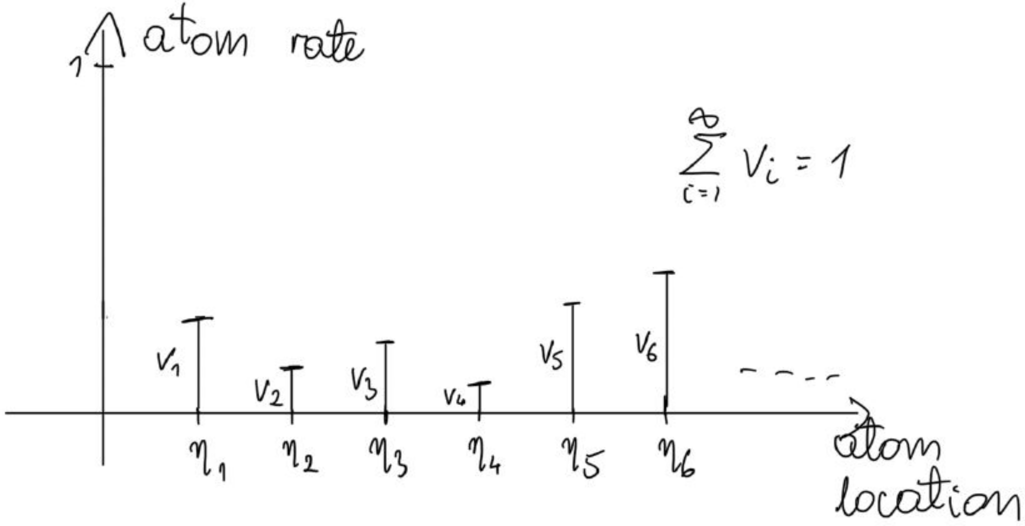
Blei & Jordan (2006), Bayesian Analysis **1**(1), pp. 121-144.

Plan for today

1. What is a Dirichlet Process and DP mixture and why do we care?
2. What is Variational Inference and why is it useful?
3. Why do we need Variational Inference for DP mixtures?
4. How can we go about constructing a VI algorithm for DP mixtures?
5. Demo: Dirichlet process mixture model for cluster assignment
6. Comparison between VI and Gibbs sampling for DP mixtures.
7. Empirical evaluation and example applications.

What is a Dirichlet Process?

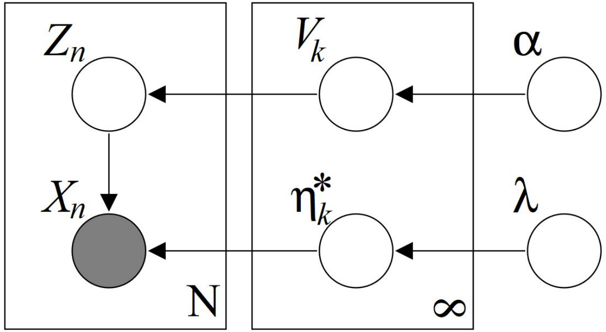
- Parameters: $\alpha > 0,\ G_0 -$ probability distribution
- Nature of the object:
  - $G \sim \mathrm{DP}(\alpha, G_0)$ is a random discrete probability distribution
  - $G : \{\text{sets}\} \to [0,1]$
- Stick-breaking construction:
  - $V_1, V_2, \cdots \overset{\text{i.i.d}}{\sim} \mathrm{Beta}(1, \alpha)$
  - $\eta_1^*, \eta_2^*, \cdots \overset{\text{i.i.d}}{\sim} G_0$
  - $G(\cdot) = \sum_{i=1}^{\infty} \pi(V) \mathbb{1}[\eta_i^* \in \cdot]$



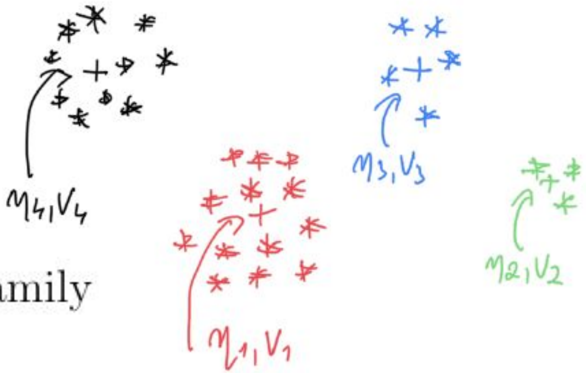
- Question: Is DP a useful object? What are its potential applications?

What is a DP mixture?

1. Parameters: $\alpha > 0,\ G_0 -$ probability distribution
2. The construction:



   1. Draw $V_1, V_2, \ldots | \alpha \sim \text{Beta}(1, \alpha)$ i.i.d
   2. Draw $\eta_1^*, \eta_2^*, \ldots | G_0 \sim G_0$ i.i.d.
   3. For the nth data point:
      - Draw $Z_n | \{v_1, v_2, \ldots\} \sim \text{Categorical}(\pi(v))$
      - Draw $X_n | z_n \sim p(x_n | \eta_{z_n}^*)$



3. Additional assumptions the paper makes:
   - $X_n | \{z_n, \eta_1^*, \eta_2^*, \ldots\}$ comes from an exponential family
   - G_0 is the corresponding conjugate prior.

Exponential families

1. Density: $p(x|\boldsymbol{\eta}) = h(x) \exp\left[\boldsymbol{\eta}^T \mathbf{T}(x) - \mathbf{a}(\boldsymbol{\eta})\right]$, for some functions $h, \mathbf{T}, \mathbf{a}$.
2. Conjugate prior: $p_{\text{prior}}(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) \propto \exp\left[\boldsymbol{\eta}^T \boldsymbol{\chi} - \nu \mathbf{a}(\boldsymbol{\eta})\right]$
3. Examples of exponential families: normal, exponential, gamma, beta, Dirichlet, chi-squared, Bernoulli, categorical, Poisson, geometric, binomial (with fixed number of trials), multinomial (with fixed number of trials).

4. In our case:

$$p(x_n|z_n, \eta_1^*, \eta_2^* \dots) = \prod_{i=1}^{\infty} \left[h(x_n) \exp\left[(\eta_i^*)^T x_n - a(\eta_i^*)\right]\right]^{\mathbb{1}[z_n=i]}$$

$$p_{\text{prior}}(\eta^*|\lambda) = h(\eta^*) \exp\left[\lambda_1^T \eta^* - \lambda_2 a(\eta^*) - a(\lambda)\right]$$

1. Draw $V_1, V_2, \dots |\alpha \sim \text{Beta}(1, \alpha)$ i.i.d
2. Draw $\eta_1^*, \eta_2^*, \dots |G_0 \sim G_0$ i.i.d.
3. For the nth data point:
   - Draw $Z_n|\{v_1, v_2, \dots\} \sim \text{Categorical}(\pi(v))$
   - Draw $X_n|z_n \sim p(x_n|\eta_{z_n}^*)$

Variational inference

1. Suppose you want to approximate the posterior distribution of latent variables:

$$p_{posterior}(\boldsymbol{w}|\boldsymbol{x},\theta) = \exp\left[\log p_{X,W}(\boldsymbol{x},\boldsymbol{w}|\theta) - \log p_X(\boldsymbol{x}|\theta)\right]$$

2. Choose a family of variational distributions: $\{q_\nu(w) : \nu \in \{\text{space of variational parameters}\}\}$
3. Minimize the KL divergence (with respect to the variational parameter)

$$D(q_\nu(\boldsymbol{w})\,||\,p(\boldsymbol{w}|\boldsymbol{x},\theta)) = \mathbb{E}_{\boldsymbol{W}\sim q_\nu} \log q_\nu(\boldsymbol{W}) - \mathbb{E}_{\boldsymbol{W}\sim q_\nu} \log p_{X,W}(\boldsymbol{x},\boldsymbol{W}|\theta) + \log p_X(\boldsymbol{x}|\theta)$$

4. Equivalently, maximize the ELBO (with respect to the variational parameter)

$$\text{ELBO} = \mathbb{E}_{\boldsymbol{W}\sim q_\nu} \log p_{X,W}(\boldsymbol{x},\boldsymbol{W}|\theta) - \mathbb{E}_{\boldsymbol{W}\sim q_\nu} \log q_\nu(\boldsymbol{W})$$

5. Mean-field VI: For an M-dimensional latent vector **W**, with conditionals following an exponential-family distribution $p(w_i|\boldsymbol{w}_{-i},\boldsymbol{x},\theta)$, choose

$$q_\nu(\boldsymbol{w}) = \prod_{i=1}^{M} \exp\left(v_i^T w_i - a(w_i)\right)$$

6. Question: Is Variational Inference useful? Do we need VI for DP mixtures? Can we do mean-field VI as described above, in this case?

Mean-field VI for DP mixtures

1. Latent variables: stick lengths, the atoms and cluster assignments: $\boldsymbol{W} = \{\boldsymbol{V}, \boldsymbol{\eta}^*, \boldsymbol{Z}\}$
2. Hyperparameters: scaling parameter and the parameter of G_0 : $\theta = \{\alpha, \lambda\}$
3. How to choose a variational family to approximate an infinite-dimensional random object depending on infinite sets $\boldsymbol{V} = \{V_1, V_2, \dots\}$ and $\boldsymbol{\eta}^* = \{\eta_1^*, \eta_2^*, \dots\}$?
4. Truncate! Stop breaking the stick! Fix some T and let the variational family satisfy $q(v_T = 1) = 1$. This means that the mixture proportions $\pi_t(\boldsymbol{v}) = 0, \text{ for all } t > T$.
5. The variational family is:

$$q_\nu(\boldsymbol{v}, \boldsymbol{v}^*, \boldsymbol{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{n=1}^{N} q_{\phi_n}(z_n)$$

where $q_{\gamma_t}(v_t)$ are beta, $q_{\tau_t}(\eta_t^*)$ are exponential family and $q_{\phi_n}(z_n)$ are categorical.

6. The free variational parameter: $\boldsymbol{\nu} = \{\gamma_1, \dots, \gamma_{T-1}, \tau_1, \dots, \tau_T, \phi_1, \dots, \phi_N\}$

A few questions

1. Why did we choose the stick-breaking representation of the Dirichlet Process to do VI?
2. Can we use the truncation trick and still call ourselves non-parametric Bayesians?
3. Does your answer to the previous question change if you a) truncate your generative model or b) truncate the variational approximation?
4. In general, what do we lose by truncating?
5. Is truncation the only reasonable approach? Can you think of other ways of bringing the DP mixture down to a finite-dimensional world?

Minimizing KL aka maximizing ELBO

$$\begin{aligned}\text{ELBO} =&\mathbb{E}_{\boldsymbol{W}\sim q_\nu} \log p_{X,W}(\boldsymbol{W}, \boldsymbol{x}|\theta) - \mathbb{E}_{\boldsymbol{W}\sim q_\nu} \log q_\nu(\boldsymbol{W}) \\ =&\mathbb{E}_q\left[\log p(\boldsymbol{V}|\alpha)\right] + \mathbb{E}_q\left[\log p(\boldsymbol{\eta}^*|\lambda)\right] + \sum_{n=1}^{N}\left(\mathbb{E}_q[\log p(Z_n|\boldsymbol{V})] + \mathbb{E}_q[\log p(x_n|Z_n)]\right) \\ &- \mathbb{E}_q[\log q(\boldsymbol{V}, \boldsymbol{\eta}^*, \boldsymbol{Z})]\end{aligned}$$

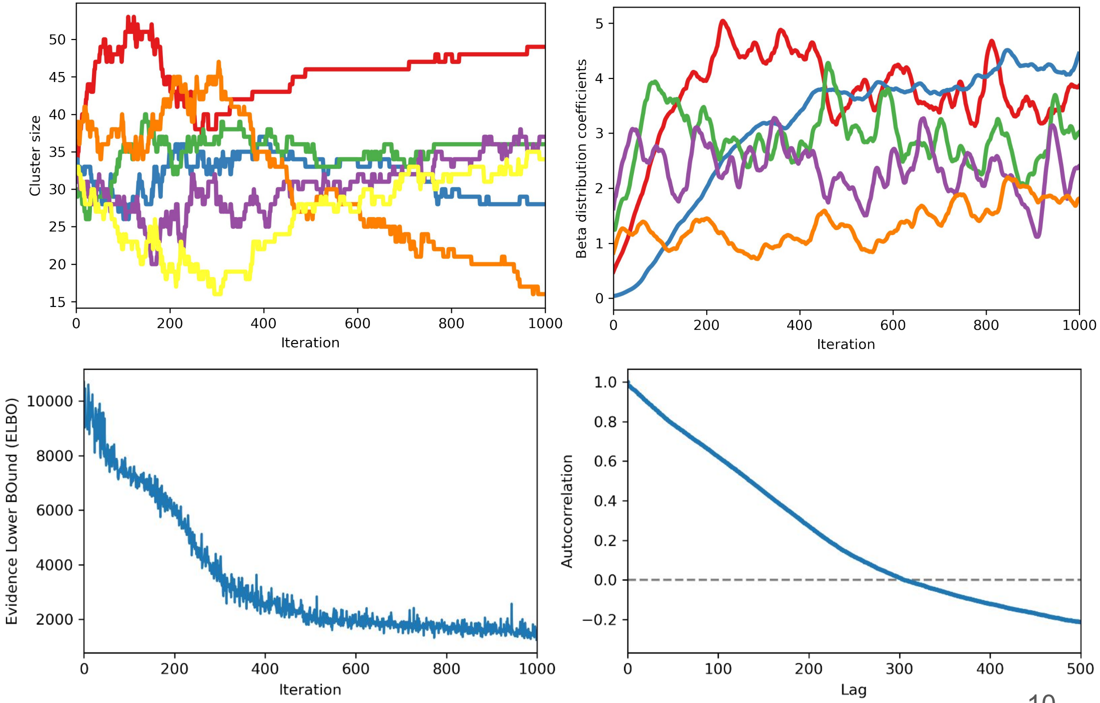
Use truncation to express the red term as a finite sum and apply coordinate ascent!

Calculating the predictive posterior:

$$\begin{aligned}p(X_{N+1}|\boldsymbol{X}, \alpha, \lambda) &= \int \left(\sum_{t=1}^{\infty} \pi_t(\boldsymbol{v}) p(x_{N+1}|\eta_t^*)\right) dP(\boldsymbol{v}, \boldsymbol{\eta}^*|\boldsymbol{x}, \lambda, \alpha) \\ &\approx \sum_{t=1}^{T} \mathbb{E}_q[\pi_t(\boldsymbol{V})]\mathbb{E}_q\left[p(x_{N+1}|\eta_t^*)\right]\end{aligned}$$

Question: Is the approximate predictive posterior easier to calculate than the true one? Why?

Demo: Let’s do some variational inference!



https://raw.githubusercontent.com/ali-ramadhan/random-jupyter-notebooks/master/Bayesian/clusters.mp4

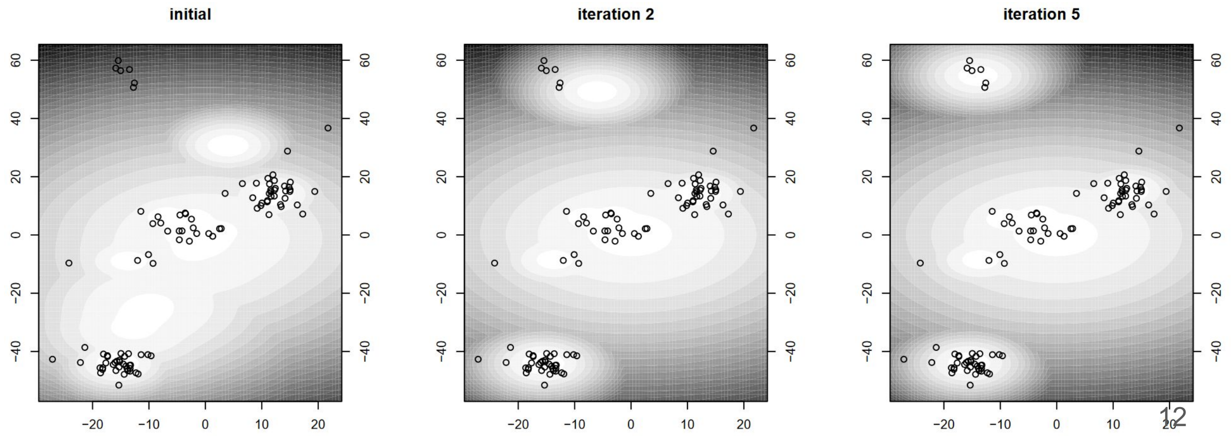
Using Dirichlet process mixture models

- Pyro: https://pyro.ai/examples/dirichlet_process_mixture.html
- PyMC3:https://docs.pymc.io/en/v3/pymc-examples/examples/mixture_models/dp_mix.html
- sklearn.mixture.DPGMM:https://scikit-learn.org/0.15/modules/generated/sklearn.mixture.DPGMM.html
- Turing.jl: https://turing.ml/dev/tutorials/06-infinite-mixture-model/
- Comparison: https://luiarthur.github.io/TuringBnpBenchmarks/dpsbgmm

Empirical comparison: Variational inference vs. Gibbs sampling

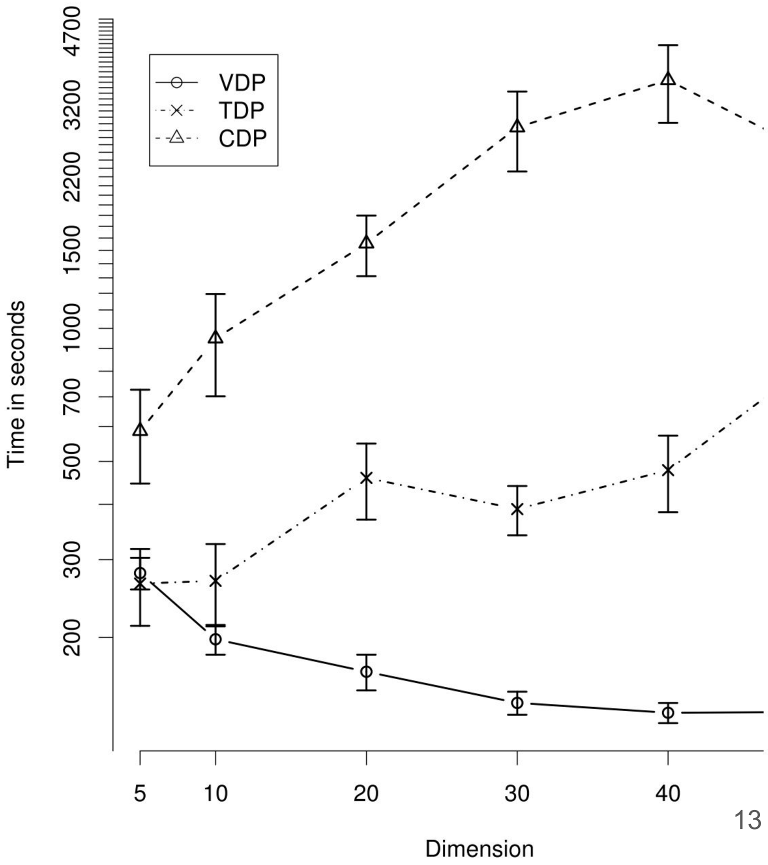
- VI can be faster.
- VI can use the Evidence lower bound (ELBO) can to assess convergence.
- VI optimization can fall into local maxima.
- VI only produces an approximation.
- No theory for evaluating VI disadvantages so the authors turn to empirical comparisons.

Problem: Dirichlet process mixture model for 100 data points sampled from a 2D DP mixture of Gaussians with diagonal covariance.



Empirical comparison: Variational inference vs. Gibbs sampling

- *Collapsed Gibbs*: marginalize over one or more variables when sampling for some other variable.
  - Integrate out G and η. Just sample cluster assignment c for each data point.
- *Blocked Gibbs*: group two or more variables together and samples from their joint distribution conditioned on all other variables.
  - Use a truncated Dirichlet process (TDP).

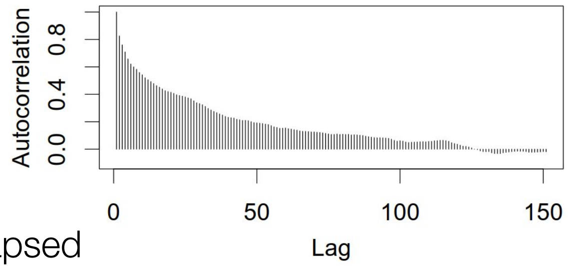


Empirical comparison: Variational inference vs. Gibbs sampling

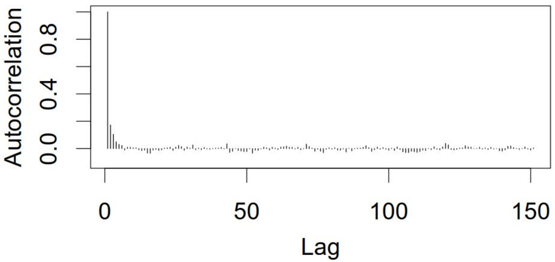
- Testing on held-out data: treat each data point as the 101st data point and compute its conditional probability.

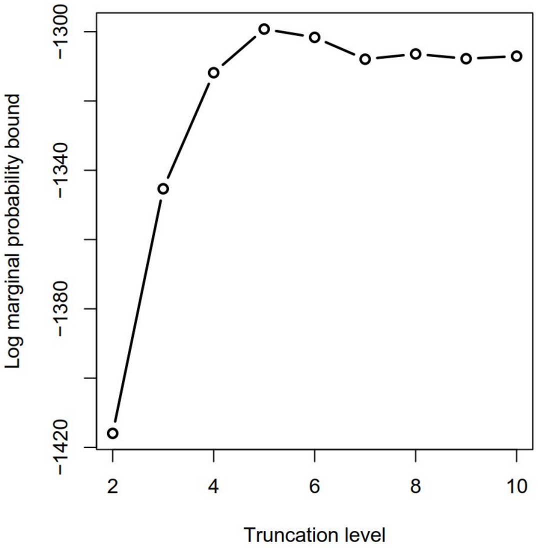
Dim	Mean held out log probability (Std err)		
	Variational	Collapsed Gibbs	Truncated Gibbs
5	-147.96 (4.12)	-148.08 (3.93)	-147.93 (3.88)
10	-266.59 (7.69)	-266.29 (7.64)	-265.89 (7.66)
20	-494.12 (7.31)	-492.32 (7.54)	-491.96 (7.59)
30	-721.55 (8.18)	-720.05 (7.92)	-720.02 (7.96)
40	-943.39 (10.65)	-941.04 (10.15)	-940.71 (10.23)
50	-1151.01 (15.23)	-1148.51 (14.78)	-1147.48 (14.55)

Blocked (truncated DP)



Collapsed





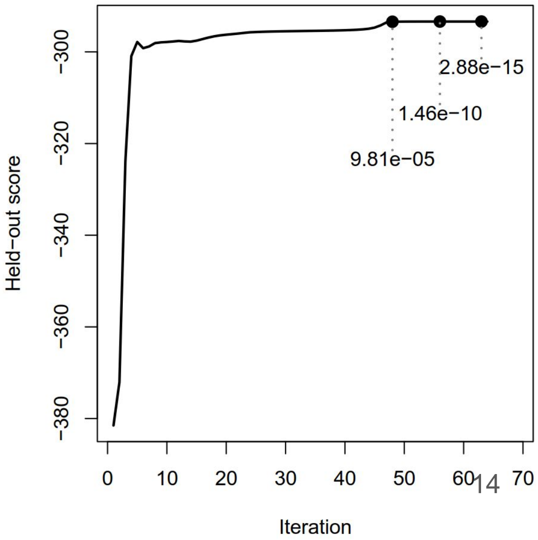
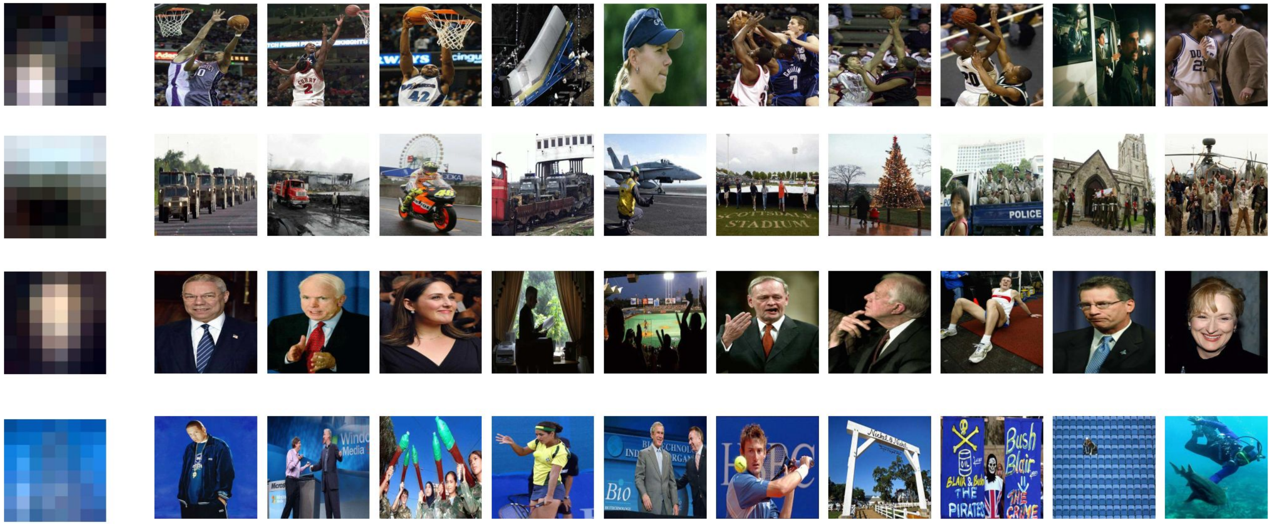
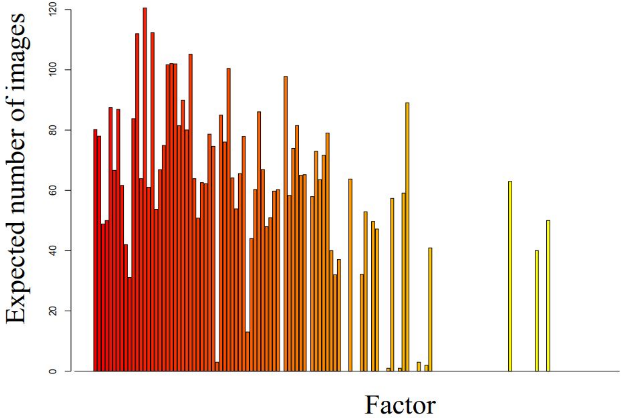
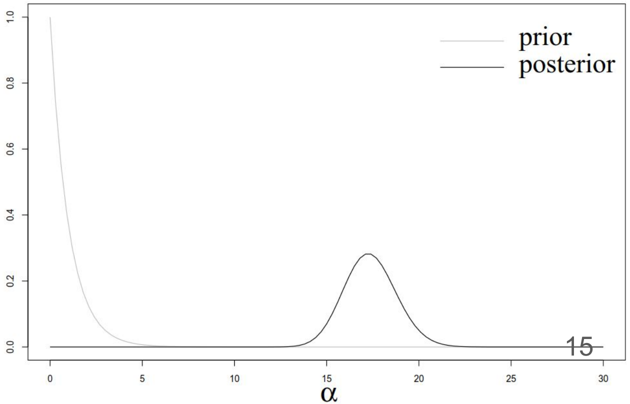


Image analysis

- Analyze 5,000 images reduced to an 8x8 grid of average RGB values.







Conclusions

1. DP mixture models are useful for applications where the number of clusters (categories, topics, ...) may potentially grow without a bound.
2. In order to do inference with DP mixture models in practice we need to approximate the posterior with some tractable, finite-dimensional distribution.
3. The authors describe a computationally efficient mean-field variational inference algorithm for DP mixture models, which outputs such an approximation.
4. They use the stick-breaking construction of DP and truncate this procedure at some fixed point in order to be in a finite-dimensional world.
5. Variational inference for DP mixture models can be faster and scale better than MCMC methods (Gibbs sampling).