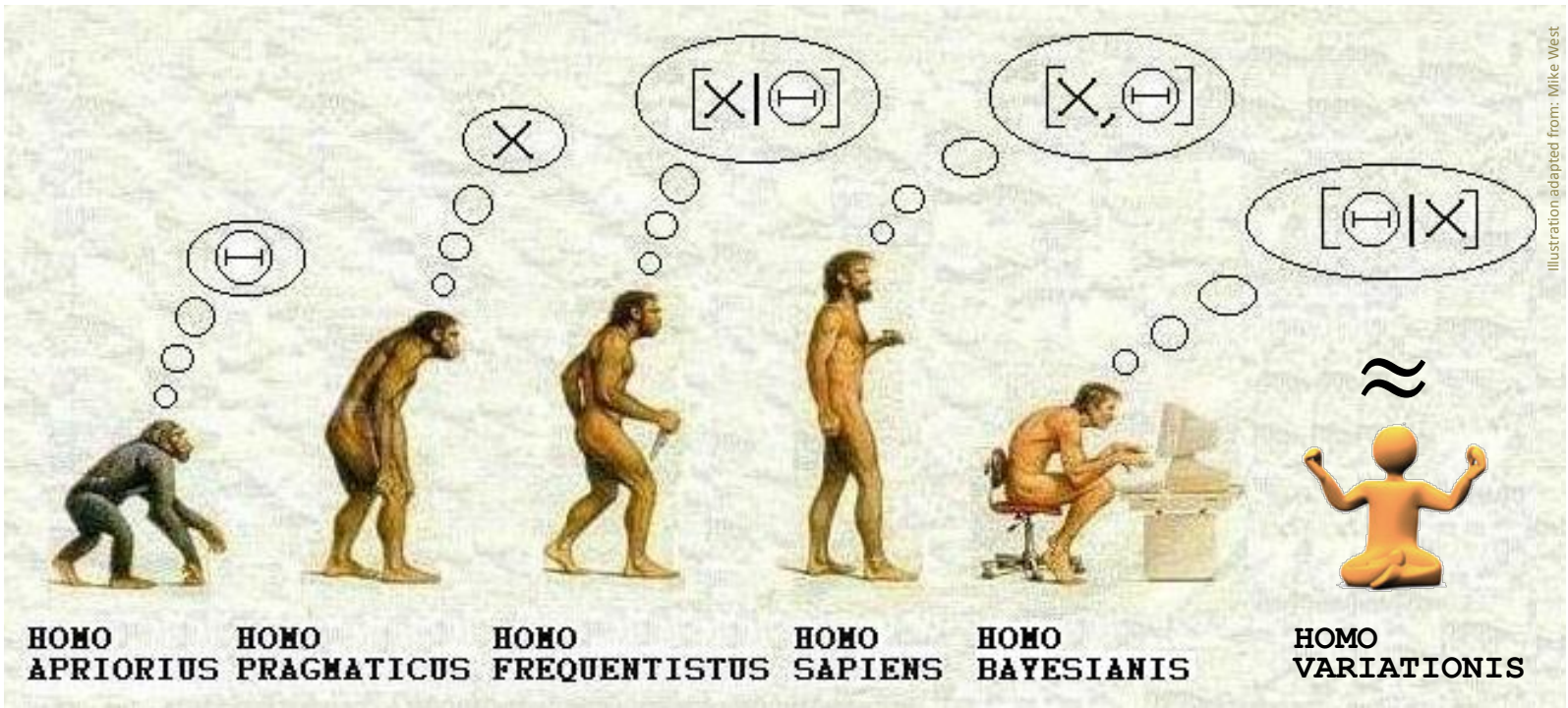
Variational Inference for DP Mixtures

Parameswaran Raman

December 4, 2015



Topics

- Intro
  - Why approximate inference?
  - Variational Inference vs MCMC
  - Variational Inference - Eg with a Generic Model
- Variational Inference for DP Mixtures
- Comparison Experiments
- References

Why Approximate Inference?

Bayesian Inference

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x,\theta) d\theta}$$

Key Challenge: Computing the log-partition function (or log marginal)

Solution: Approximate Inference techniques!

Approaches: Type I (sampling based), Type II (variational approximation based), or mix of both

Approximate Inference - Type I

Sampling based methods

1. Design an algorithm that draws samples $\theta_1, \ldots, \theta_k$ from $p(\theta|x)$
2. Use the samples for further computations (say: approximating an expectation originally intractable)

Pros and Cons:

- asymptotically exact
- slow and computationally expensive

Examples: Gibbs Sampling, Importance Sampling, Rejection Sampling

Approximate Inference - Type II

Variational methods

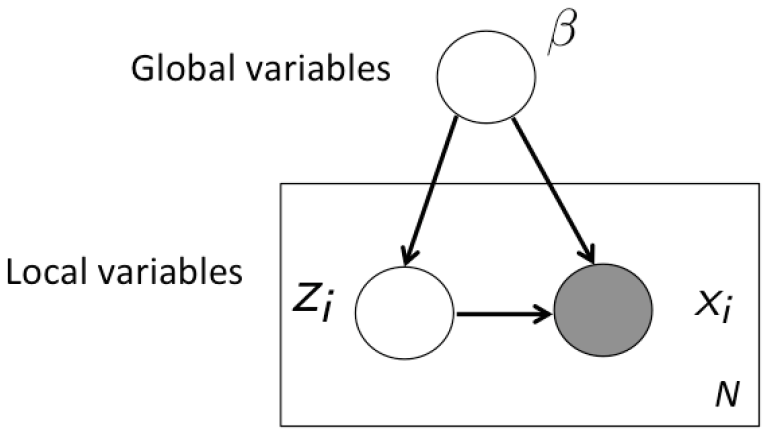
1. Find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta|x)$
2. Turn inference into an optimization problem involving $q(\theta)$

Pros and Cons:

- can be much faster
- deterministic algorithm
- hard to derive variational updates
- prone to local minima

Examples: Variational Bayes, Expectation Propagation

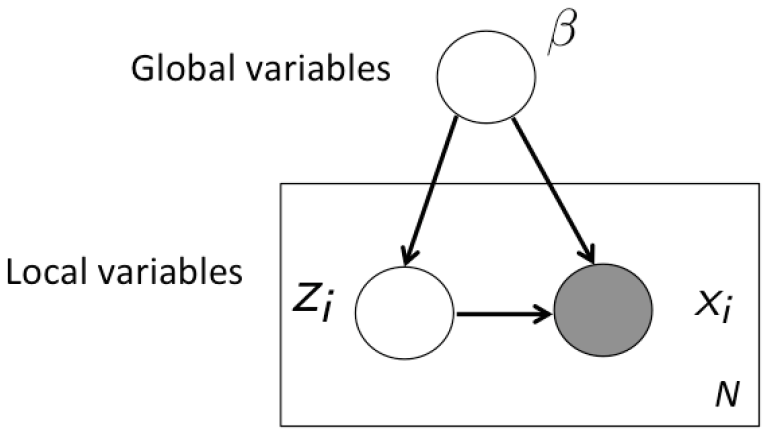
Generic Model



$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- Observations are $x = x_{1:N}$
- Local variables are $z = z_{1:N}$
- Global variables are β
- The ith data point x_i only depends on z_i and β
- **Goal**: To compute $p(\beta, z|x)$

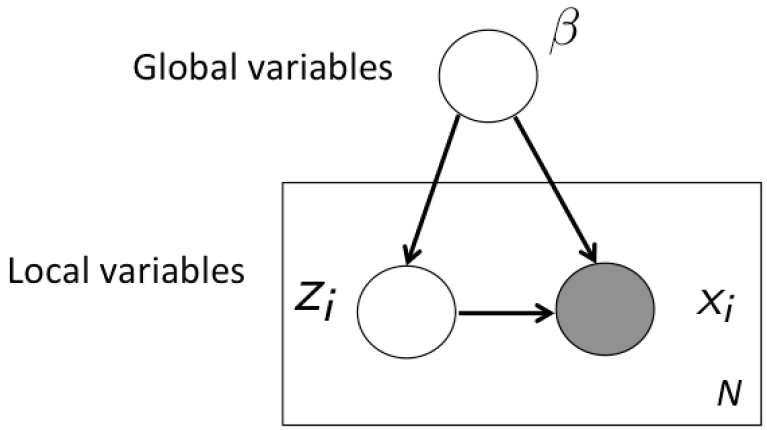
Generic Model



$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- Bayesian Mixture Models
- Time series models (HMMs, Kalman Filters)
- Factorial Models
- Matrix Factorization (PCA, etc)
- Dirichlet Process Mixtures, HDPs
- Multilevel regression
- Mixed-membership models (LDA and variants)

Generic Model



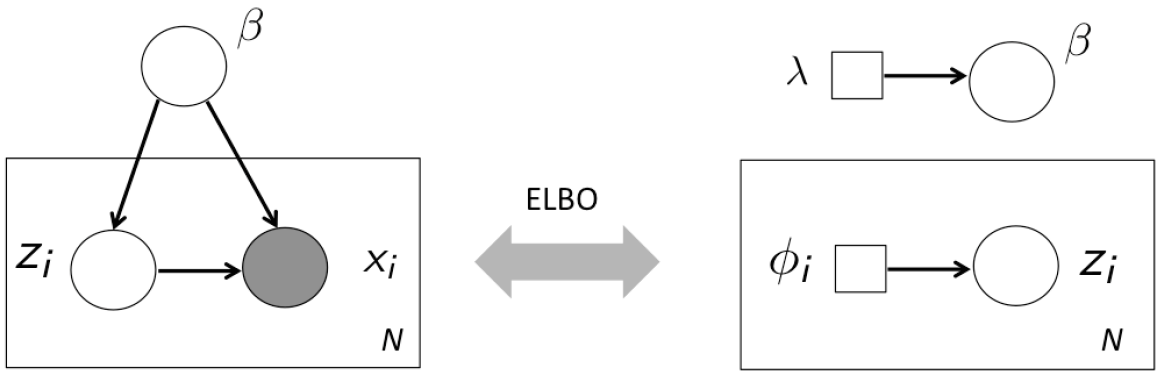
$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variables
- Assume each complete conditional $\sim$ exp family,

$$p(z_i|\beta, x_i) \propto \exp\{\langle z_i, \eta_l(\beta, x_i)\rangle - g(\eta_l(\beta, x_i))\}$$

$$p(\beta|z, x) \propto \exp\{\langle \beta, \eta_g(z, x)\rangle - g(\eta_g(z, x))\}$$

Evidence Lower Bound (ELBO)



- Introduce a **variational distribution** over the latent variables $q(\beta, z)$
- We optimize the **evidence lower bound** (ELBO) with respect to q

$$\log p(x) \geq \mathbb{E}_q \Big[\log p(\beta, Z, x) \Big] - \mathbb{E}_q \Big[\log q(\beta, Z) \Big]$$

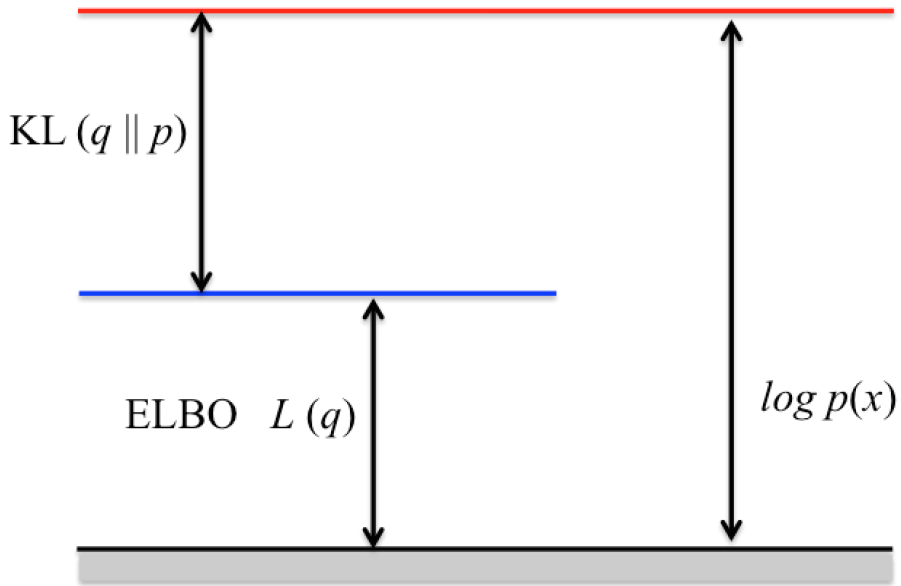
- Up to a constant, this is the negative KL divergence between q and the posterior
- *The ELBO links the observations/model to the variational distribution*

Evidence Lower Bound (ELBO)

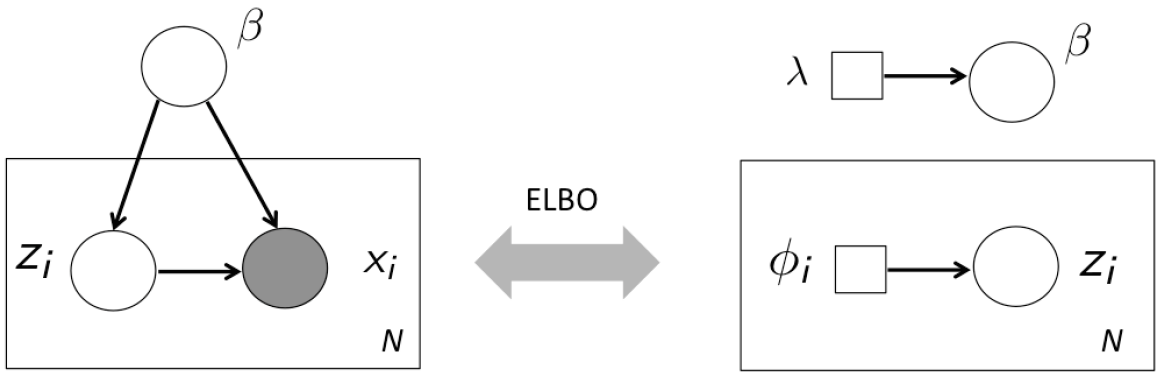
Jensen's Inequality

When f is concave, $f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x,z)dz \\
&= \log \int_z p(x,z)\frac{q(z)}{q(z)}dz \\
&= \log \int_z \frac{p(x,z)}{q(z)}q(z)dz \\
&= \log(\mathbb{E}_q[\frac{p(x,z)}{q(z)}]) \\
&\geq \mathbb{E}_q[\log \frac{p(x,z)}{q(z)}]
\end{aligned}
$$



Mean Field Approximation



- We specify $q(\beta, z)$ to be a fully factored variational distribution,

$$q(\beta, z) = q(\beta|\lambda) \prod_{i=1}^{N} q(z_i|\phi_i)$$

- Each instance of each variable has its own distribution
- Each component (factor) is in the same family as the model conditional,

$$p(\beta|z, x) \propto \exp\{\langle \beta, \eta_g(z, x) \rangle - g(\eta_g(z, x))\}$$

$$q(\beta|\lambda) \propto \exp\{\langle \beta, \lambda \rangle - g(\lambda)\}$$

(And same for the local variational parameters)

Optimization of ELBO

- **Objective function**: We optimize the ELBO wrt variational parameters

$$\mathcal{L}(\lambda, \phi_{1:N}) = \mathbb{E}_q[\log p(\beta, x, z)] - \mathbb{E}_q[\log q(\beta, z)]$$

- Same as finding the $q(\beta, z)$ that is closest in KL Divergence to $p(\beta, z|x)$
- Coordinate Ascent: Iteratively update each variational parameter, holding others fixed
  - **Local Step** (*Var E-Step*):

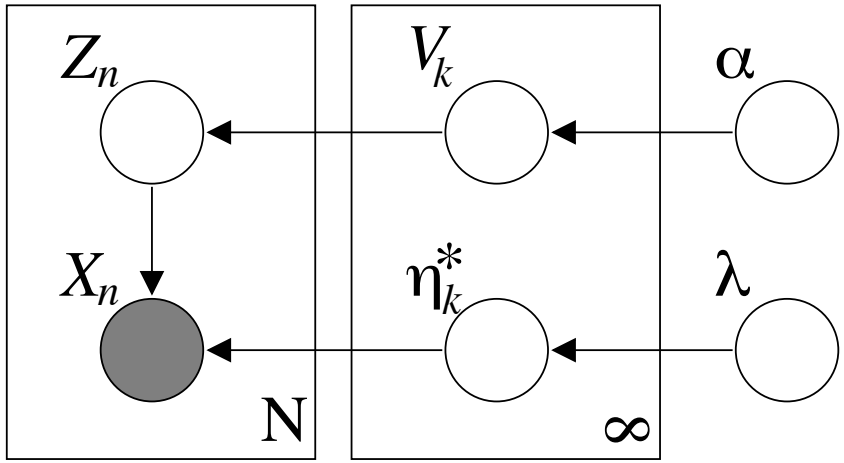
$$\phi_i = \mathbb{E}_q\left[\eta_l(x_i, \beta)\right]$$

  - **Global Step** (*Var M-Step*):

$$\lambda = \mathbb{E}_q\left[\eta_g(x, z)\right]$$

Exponential Family DP Mixture

$$G|\{\alpha, G_0\} \sim DP(\alpha, G_0)$$
$$\eta_n|G \sim G$$
$$X_n|\eta_n \sim p(x_n|\eta_n)$$



where,

- $\eta_1, \ldots, \eta_n$ are the variables,
- $\eta_1^*, \cdots, \eta_{|c|}^*$ denote distinct values of $\eta_1, \ldots, \eta_n$ (i.e.. atoms)
- V_k denote stick lengths
- Z_n denote assignments of observations to atoms

We are interested in: $p(V, \eta^*, Z)$

Deriving Variational Updates

Step 1: Write the evidence lower bound (ELBO):

$$\log p(x|\alpha,\lambda) \geq \mathbb{E}_q[\log p(V|\alpha)] + \mathbb{E}_q[\log p(\eta^*|\alpha)] + \sum_{n=1}^{N}\left(\mathbb{E}_q[\log p(Z_n|V)] + \mathbb{E}_q[\log p(x_n|Z_n)]\right) - \mathbb{E}_q[\log q(V,\eta^*,Z)]$$

Step 2: Specify the (factorizable) mean field variational family:

$$q(v,\eta^*,z) = \prod_{t=1}^{T-1}\underbrace{q(v_t|\gamma_t)}_{beta}\quad\prod_{t=1}^{T}\underbrace{q(\eta_t^*|\tau_t)}_{expfamily}\quad\prod_{n=1}^{N}\underbrace{q(z_n|\phi_n)}_{mult}$$

where new variational parameters $\{\gamma_1,\ldots,\gamma_{T-1},\tau_1,\ldots,\tau_T,\phi_1,\ldots,\phi_N\}$ have been introduced.

Also, to handle the infinite set $V=\{V_1,V_2,\ldots\}$, *truncated stick-breaking representation* is used in [1].

Deriving Variational Updates - (Contd)

Step 3: Coord Ascent Updates (closed form):

$$\gamma_{t,1} = 1 + \sum_{N} \phi_{n,t}$$

$$\gamma_{t,2} = \alpha + \sum_{N} \sum_{j=t+1}^{T} \phi_{n,j}$$

$$\tau_{t,1} = \lambda_1 + \sum_{N} \phi_{n,t} x_n$$

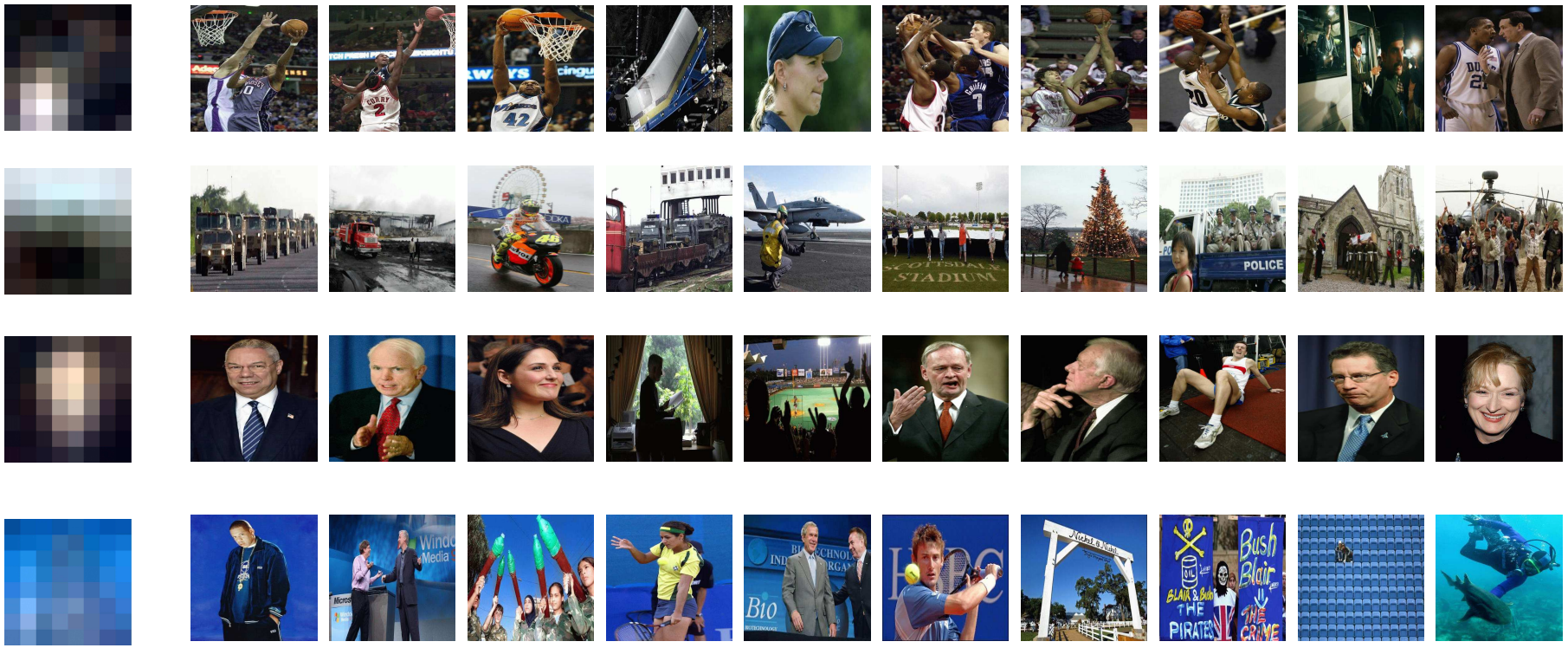
$$\tau_{t,2} = \lambda_2 + \sum_{N} \phi_{n,t}$$

$$\phi_{n,t} = \exp(S_t)$$

where,

$S_t = \mathbb{E}_q[\log V_t] + \sum_{i=1}^{t-1} \mathbb{E}_q[\log(1 - V_i)] + \mathbb{E}_q[\eta_t^*]^T X_n - \mathbb{E}_q[a(\eta_t^*)]$

Comparison Experiments - Image Analysis



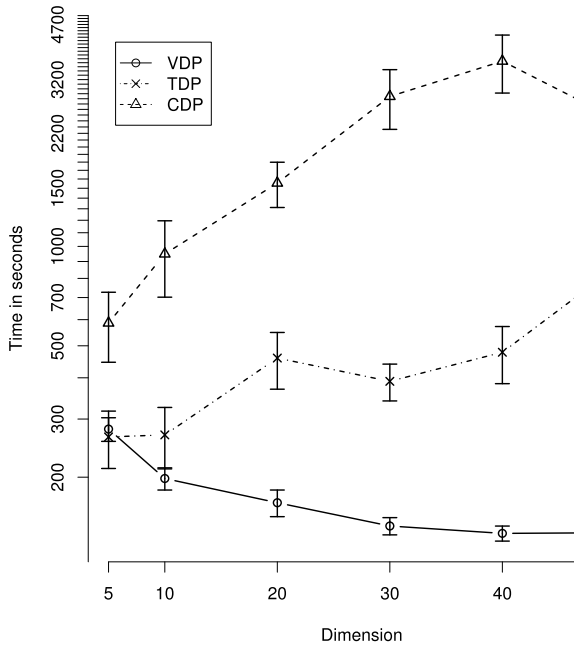
Problem: Identify # of mixture components in the collection of images

Comparison Experiments - Image Analysis

Convergence time:

- Variational algorithm took approximately **four hours** to converge
- Collapsed Gibbs Sampling was drastically slower (one iteration took **15 mins**)

Interesting Observation:



References

[1] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[2] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[3] Chong Wang and David M Blei. Variational inference in nonconjugate models. *The Journal of Machine Learning Research*, 14(1):1005–1031, 2013.