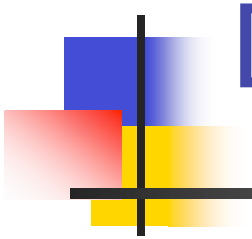
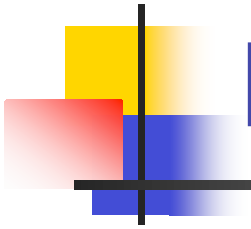


Variational Inference for Dirichlet Process Mixtures

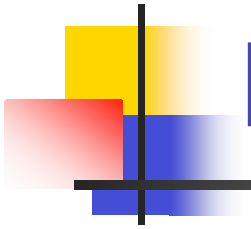
By David Blei and Michael Jordan

Presented by Daniel Acuna



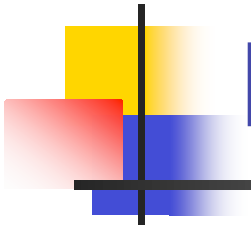
Motivation

- Non-parametric Bayesian models seem to be the right idea:
  - Do not fix the number of mixture components
- Dirichlet process is an elegant and principled way to “automatically” set the components
- Need to explore new methods that cope intractable nature of marginalization or conditional
- MCMC sampling methods widely used in this context, but there are other ideas



Motivation

- Variational inference have proved to be faster and more predictable (deterministic) than sampling
- The basic idea
  - Reformulate as an optimization problem
  - Relax the optimization problem
  - Optimize (find a bound of the original problem)



Background

- Dirichlet process mixture is a measure on measures
- Multiples representations and interpretations:
  - Ferguson Existent theorem
  - Blackwell-MacQueen urn scheme
  - Chinese restaurant process
  - Stick-breaking construction

Dirichlet process mixture model

$$
\begin{aligned}
G \mid \{\alpha, G_0\} &\sim \mathrm{DP}(\alpha, G_0) \\
\eta_n \mid G &\sim G \\
X_n \mid \eta_n &\sim p(x_n \mid \eta_n).
\end{aligned}
$$

- G_0 Base distribution
- α Positive scaling parameter

$\{\eta_1, \ldots, \eta_{n-1}\}$ exhibit a clustering effect

The DP mixture has a natural interpretation as a flexible mixture model in which the number of components is random and grows as new data are observed

Stick-breaking representation

- Two infinite collections of independent random variables

$$V_i \sim Beta(1,\alpha)$$

$$\eta_i^* \sim G_0$$

For i = {1,2,…}

- Stir-breaking representation of G

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j)$$

$$G = \sum_{i=1}^{\infty} \pi_i(\mathbf{v})\, \partial_{\eta_i^*}$$

- G is discrete!

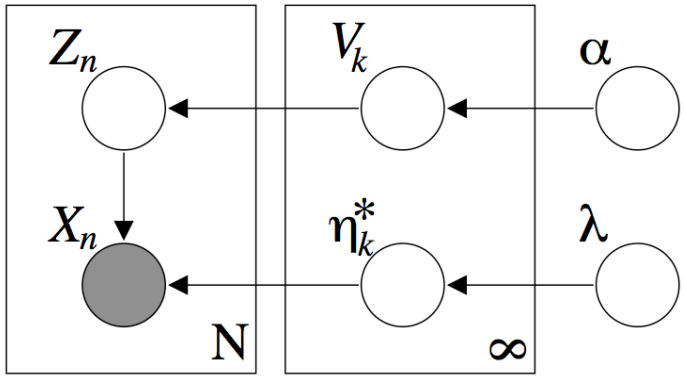
Sticking-breaking rep.

- The data can be described as arriving from

1) Draw $V_i \mid \alpha \sim Beta(1,\alpha), \qquad i = \{1,2,\ldots\}$
2) Draw $\eta_i^* \mid G_0 \sim G_0 \qquad i = \{1,2,\ldots\}$
3) For the n-th data point
   1) Draw $Z_n \mid \{v_1, v_2, \ldots\} \sim Mult(\pi(\mathbf{v}))$
   2) Draw $X_n \mid z_n \sim p(x_n \mid \eta_{z_n}^*)$

DP mixture for exponential families

- Observable data drawn from exponential family, the base distribution is the conjugate



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

$$p(\eta^* \mid \lambda) = h(\eta^*) \exp\{\lambda_1^T \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)\}$$

Variational inf. for DP mix.

- In DP, our goal

$$p(x \mid x_1, \ldots, x_N, \alpha, G_0) = \int p(x \mid \eta) p(\eta \mid x_1, \ldots, x_N, \alpha, G_0) d\eta$$

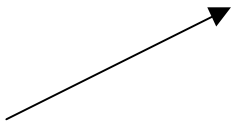
- But complex $p(\eta \mid x_1, \ldots, x_N, G_0, \alpha)$

- Variational inference uses a proposal distribution that breaks the dependency among latent variables

Variational inf. for DP mix.

- In general, consider a model with hyperparameters θ , latent variables $\mathbf{W} = \{W_1, \ldots, W_M\}$ and observations x = $\{x_1, \ldots, x_N\}$
- The posterior distribution:

$$p(\mathbf{w} \mid \mathbf{x}, \theta) = \exp\{\log p(\mathbf{x}, \mathbf{w} \mid \theta) - \log p(\mathbf{x} \mid \theta)\}$$



Difficult!

Variational inf. for DP mix

- This is difficult

$$\log p(\mathbf{x} \mid \theta) = \log \int p(\mathbf{w}, \mathbf{x} \mid \theta) d\mathbf{w}$$

Because latent variables become dependent when conditioning on observed data

- We reformulate the problem using the mean-field method, which optimizes the KL divergence with respect to a *variational distribution*.

Variational inf. for DP mix

- This is, we aim to minimize the KL divergence between $q_\nu(\mathbf{w})$ and $p(\mathbf{w} \mid \mathbf{x}, \theta)$

$$\mathrm{D}(q_\nu(\mathbf{w}) || p(\mathbf{w} \mid \mathbf{x}, \theta)) = \mathrm{E}_q\left[\log q_\nu(\mathbf{W})\right] - \mathrm{E}_q\left[\log p(\mathbf{W}, \mathbf{x} \mid \theta)\right] + \log p(\mathbf{x} \mid \theta)$$

- Or equivalently, we try to maximize the lower bound

$$\log p(\mathbf{x} \mid \theta) \geq \mathrm{E}_q\left[\log p(\mathbf{W}, \mathbf{x} \mid \theta)\right] - \mathrm{E}_q\left[\log q_\nu(\mathbf{W})\right]$$

Mean field of exponential fam.

- For each latent variable, the conditional is a member of a exponential family:

$$p(w_i \mid \mathbf{w}_{-i}, \mathbf{x}, \theta) = h(w_i) \exp\{g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta)^T w_i - a(g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta))\}$$

- Where $g_i(\mathbf{w}_{-i}, \mathbf{x}, \theta)$ is the natural parameter of w_i when conditioned on the remaining latent variables
- Here the family of distributions is

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

$\boldsymbol{\nu} = \{\nu_1, \nu_2, \ldots, \nu_M\}$
Variational parameters

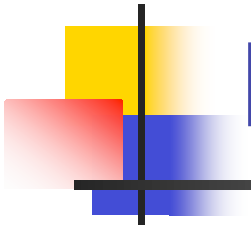
Mean-field of exponential family

- The optimization of KL divergence

$$\nu_i = \mathrm{E}_q \left[g_i(\mathbf{W}_{-i}, \mathbf{x}, \theta) \right]$$

after derivation (see Apendix)

- Notice:
  - Gibbs sampling, we draw w_i from $p(w_i|w_{-i},x,\theta)$
  - Here, we update v_i to set it equal E[$g_i(w_{-i},x,\theta)$]



DP mixtures

- The latent variables are stick lengths, atoms, and cluster assignment

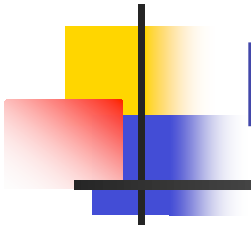
$$\mathbf{W} = \{\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z}\}$$

- The hyper parameters are the scaling and conjugate base distribution

$$\theta = \{\alpha, \lambda\}$$

- And the bound now is

$$\begin{aligned}\log p(\mathbf{x} \mid \alpha, \lambda) \geq\ & \mathrm{E}_q\left[\log p(\mathbf{V} \mid \alpha)\right] + \mathrm{E}_q\left[\log p(\boldsymbol{\eta}^* \mid \lambda)\right] \\ & + \sum_{n=1}^{N} \left(\mathrm{E}_q\left[\log p(Z_n \mid \mathbf{V})\right] + \mathrm{E}_q\left[\log p(x_n \mid Z_n)\right]\right) \\ & - \mathrm{E}_q\left[\log q(\mathbf{V}, \boldsymbol{\eta}^*, \mathbf{Z})\right].\end{aligned}$$



Relaxation of optimization

- To exploit this bound, with family q we need to approximate G
  - G is an *infinite-dimensional* random measure.
  - An approximation is to truncate the stick-breaking representation!

Relaxation of optimization

- Fix value T and q(v_T = 1)=1, then $\pi_t(\mathbf{v})$ are equal to zero for t>T
- (remember from $\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j)$
- Propose,

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

- $q_{\gamma_t}(v_t)$ Beta distributions
- $q_{\tau_t}(\eta_t^*)$ Exponential family distributions
- $q_{\phi_n}(z_n)$ Multinomial distributions

$$\boldsymbol{\nu} = \{\gamma_1, \ldots, \gamma_{T-1}, \tau_1, \ldots, \tau_T, \phi_1, \ldots, \phi_N\}$$

Optimization

- The optimization is performed by coordinate ascent algorithm
- From,

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

Infinite!

$$\begin{aligned} \mathrm{E}_q\left[\log p(Z_n \mid \mathbf{V})\right] &= \mathrm{E}_q\left[\log\left(\textstyle\prod_{i=1}^{\infty}(1-V_i)^{\mathbf{1}[Z_n>i]}V_i^{\mathbf{1}[Z_n=i]}\right)\right] \\ &= \textstyle\sum_{i=1}^{\infty} q(z_n > i)\mathrm{E}_q\left[\log(1-V_i)\right] + q(z_n = i)\mathrm{E}_q\left[\log V_i\right] \end{aligned}$$

Optimization

- But, $\mathrm{E}_q\left[\log(1 - V_T)\right] = 0$ and $q(z_n > T) = 0$.

Then

$$\mathrm{E}_q\left[\log p(Z_n \mid \mathbf{V})\right] = \sum_{i=1}^{T} q(z_n > i)\mathrm{E}_q\left[\log(1 - V_i)\right] + q(z_n = i)\mathrm{E}_q\left[\log V_i\right]$$

Where

$$\begin{aligned} q(z_n = i) &= \phi_{n,i} \\ q(z_n > i) &= \textstyle\sum_{j=i+1}^{T} \phi_{n,j} \\ \mathrm{E}_q\left[\log V_i\right] &= \Psi(\gamma_{i,1}) - \Psi(\gamma_{i,1} + \gamma_{i,2}) \\ \mathrm{E}_q\left[\log(1 - V_i)\right] &= \Psi(\gamma_{i,2}) - \Psi(\gamma_{i,1} + \gamma_{i,2}). \end{aligned}$$

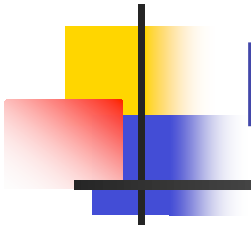
Optimization

- Finally, the mean-field coordinate ascent algorithm boils down to updates:

$$
\begin{aligned}
\gamma_{t,1} &= 1 + \textstyle\sum_n \phi_{n,t} \\
\gamma_{t,2} &= \alpha + \textstyle\sum_n \sum_{j=t+1}^{T} \phi_{n,j} \\
\tau_{t,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,t} x_n \\
\tau_{t,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,t}. \\
\phi_{n,t} &\propto \exp(S_t),
\end{aligned}
$$

for $t \in \{1, \ldots, T\}$ and $n \in \{1, \ldots, N\}$

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



Predictive distribution

$$p(x_{N+1} \mid \mathbf{x}, \alpha, \lambda) \approx \sum_{t=1}^{T} \mathrm{E}_q \left[\pi_t(\mathbf{V})\right] \mathrm{E}_q \left[p(x_{N+1} \mid \eta_t^*)\right]$$

where q depends implicitly on $\mathbf{x}$, α, and λ

Empirical comparison

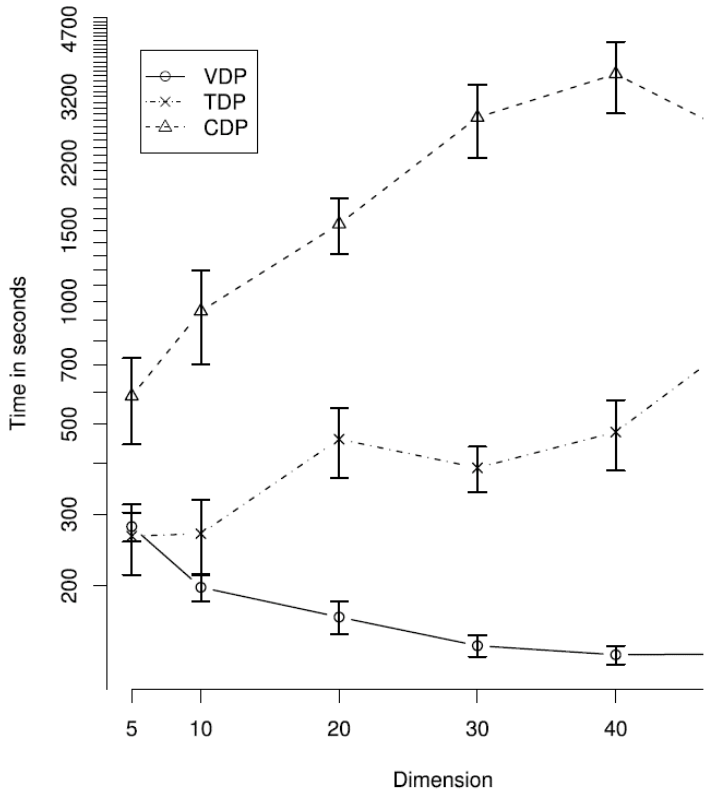
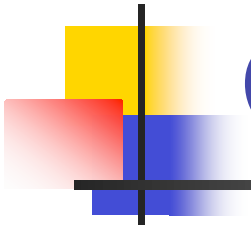


Figure 3: Mean convergence time and standard error across ten data sets per dimension for variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler.



Conclusion

- Faster than sampling for particular problems
- Unlikely, that one method will dominate another ➔ both have their pros and cons
- This is the simplest variational method (mean-field). Other methods are worth exploring.
- Check www.videolectures.net