Particle EM in Latent Dirichlet Allocation model

Yubing Yao

December 20, 2016

Introduction-Particle EM

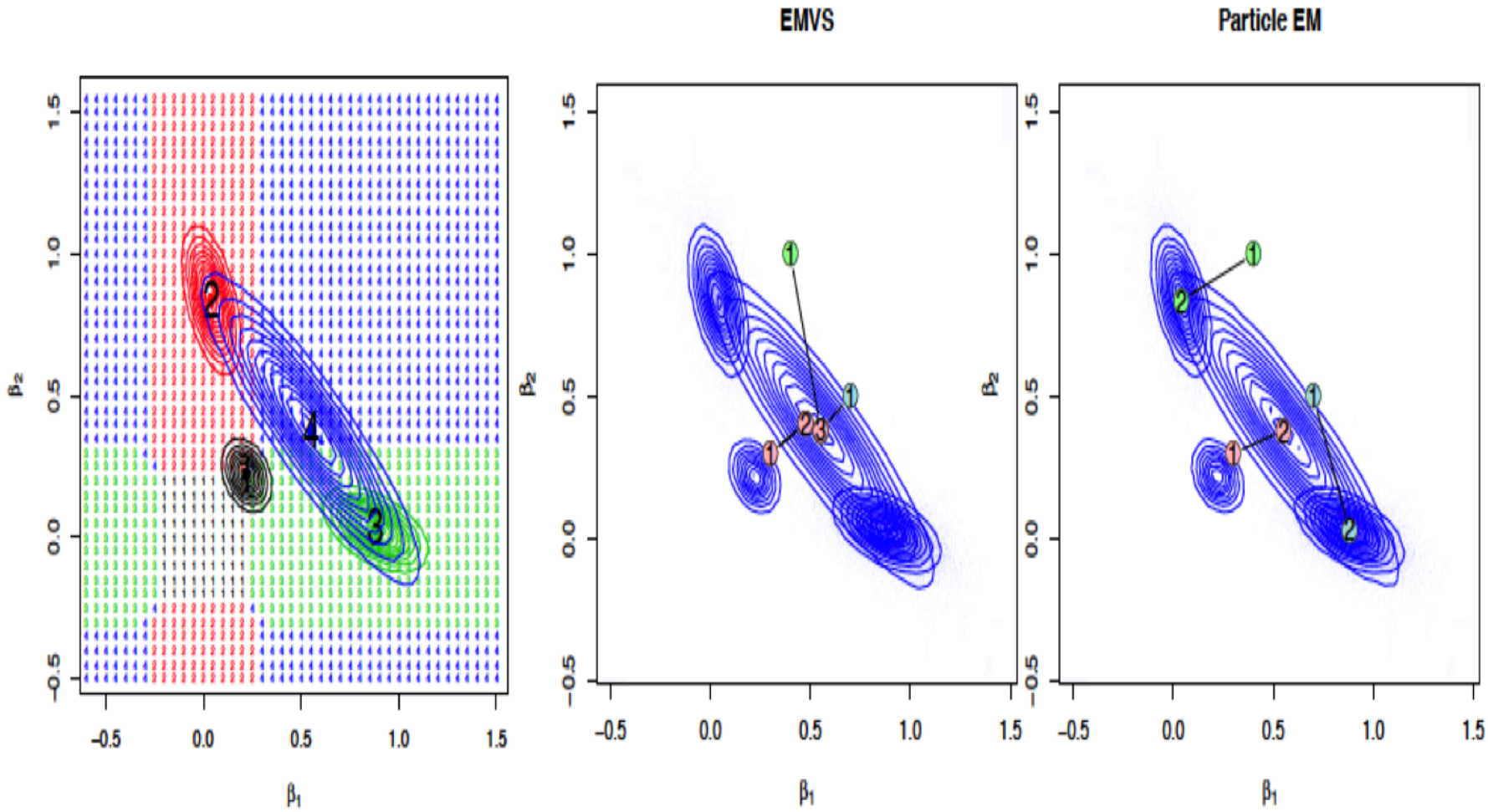
1. The Particle EM Algorithm was introduced by Veronika Rockova in Bayesian variable selection to find the best multiple point approximation of posterior marginal distribution in 2016.

2. The particle EM algorithm is a population based optimization method, and aims to overcome the vulnerability of local entrapment of the traditional EM algorithm when dealing with the multi-modal posterior/likelihood.

3. The particle EM algorithm explores the whole search space by multiple repulsive particles and tries to capture the multiple modes of posterior distribution. Thus it can increase the possibility to identify a global mode by discovering a more comprehensive set of posterior modes.

A simple illustrative example comparing EM and particle EM

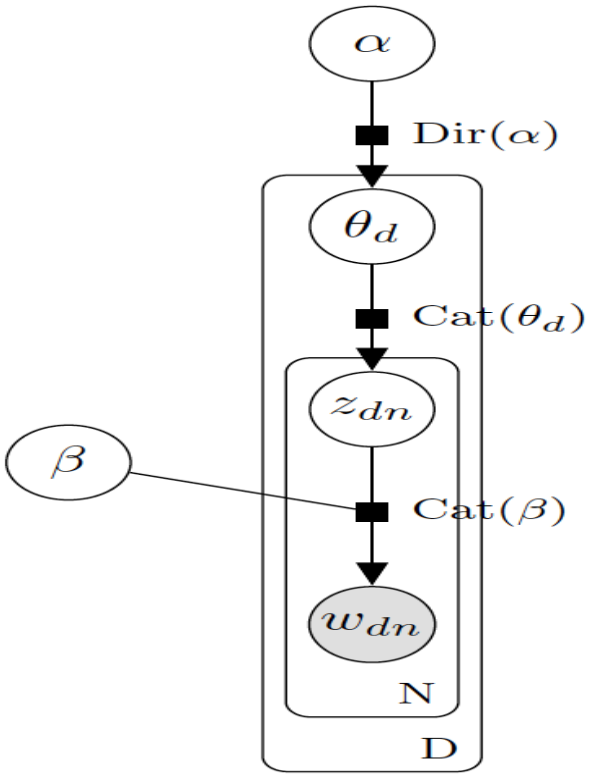
Two collinear predictors-$\boldsymbol{\beta} = (\beta_1, \beta_2)$, The posterior distribution-$\pi(\boldsymbol{\beta}|\boldsymbol{Y})$ have 4 modes with mode 3-the global mode.



Introduction- Latent Dirichlet Allocation(LDA) model

The Latent Dirichlet Allocation(LDA) model is a probabilistic model for collections of discrete data such as text corpora introduced by D. Blei et al. 2003 with model parameters-α, β.

$$
\begin{aligned}
p(\mathcal{D}|\alpha,\beta) &= \prod_{d=1}^{D} \int p(\theta_d, \boldsymbol{z}_d, \boldsymbol{w}_d|\alpha,\beta) d\theta_d \\
&= \prod_{d=1}^{D} \int p(\theta_d|\alpha) \prod_{n_d=1}^{N} p(\boldsymbol{z}_{dn_d}|\theta_d) \\
&\quad \times p(\boldsymbol{w}_{dn_d}|\boldsymbol{z}_{dn_d},\beta) d\theta_d
\end{aligned}
$$



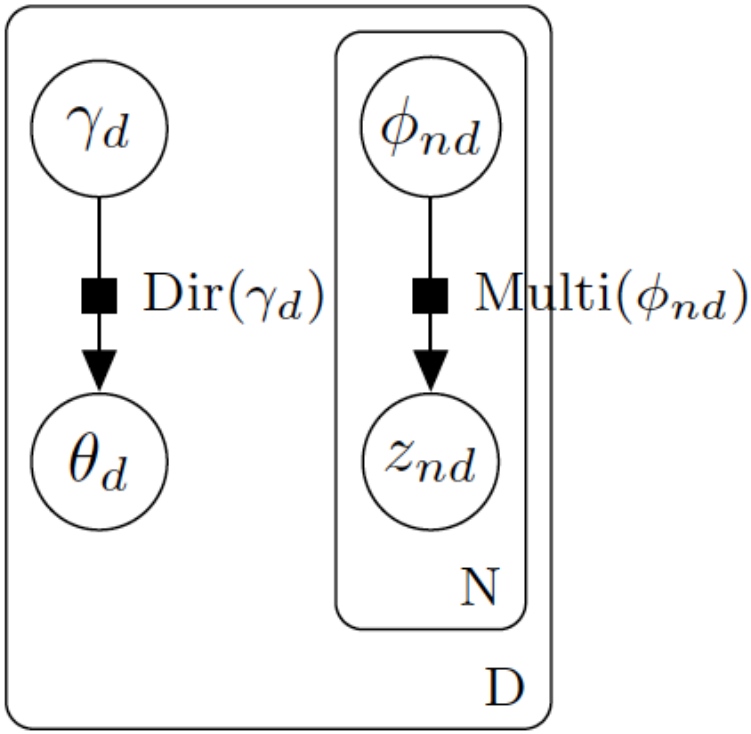
Variational Inference in LDA model

Two free variational parameters-γ, ϕ was introduced in D. Blei et al. 2003 to break the coupling between the parameters-θ and β in LDA such that:

$$q(\theta_d, \boldsymbol{z}_d|\gamma, \phi) = q(\theta_d|\gamma) \prod_{n=1}^{N_d} q(\boldsymbol{z}_{nd}|\phi_{dn})$$

ELBO of variational inference in LDA model can be represented as:

$$\sum_{d=1}^{D} (\mathrm{E}_q[\log p(\theta_d|\alpha)] + \mathrm{E}_q[\log p(\boldsymbol{z}_d|\theta_d)] + \mathrm{E}_q[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)] - \mathrm{E}_q[\log q(\theta_d|\gamma_d)] - \mathrm{E}_q[\log q(\boldsymbol{z}_{n_d}|\phi_{dn})])$$



Particle approximation and Evidence Lower Bound

- We use a weighted mixture of atoms to approximate $\pi(\boldsymbol{z}_{dn}|\mathcal{D}), n = 1, 2, \ldots, N_d; d = 1, 2, \ldots, D$, and we denote:

$$q_{PEM}(\boldsymbol{z}_{dn}|\boldsymbol{\Gamma}_{dn}, \boldsymbol{\omega}) = \sum_{p=1}^{P} \omega_p \mathbb{I}\{\boldsymbol{z}_{dn} = \boldsymbol{z}_{pdn}\}$$

  where $\boldsymbol{\Gamma}_{dn} = [\boldsymbol{z}_{1dn}, \boldsymbol{z}_{2dn}, \ldots, \boldsymbol{z}_{Pdn}]$,corresponding importance weights-$\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_P)^T$, where $\sum_{p=1}^{P} \omega_p = 1, \forall d = 1, \ldots, D, n = 1, 2, \ldots, N_d; 0 \leq \omega_p \leq 1, \forall p = 1, 2, \ldots, P$.

- The corresponding Evidence Lower Bound (ELBO)-$\mathcal{L}_\lambda(\boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\omega}; \boldsymbol{\theta}, \alpha, \beta)$ is:

$$\begin{aligned}\sum_{d=1}^{D} (&\mathrm{E}_q[\log p(\theta_d|\alpha)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{z}_d|\theta_d)] + \mathrm{E}_{q_{PEM}}[\log p(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)] \\ &-\mathrm{E}_q[\log q(\theta_d|\boldsymbol{\gamma}_d)] - \lambda \mathrm{E}_{q_{PEM}}[\log q_{PEM}(\boldsymbol{z}_{n_d}|\boldsymbol{\Gamma}_d, \boldsymbol{\omega}_d)])\end{aligned} \tag{1}$$

  where $\lambda \geq 0$.

Particle EM-E step with single particle

- With single particle $P = 1$ and fix some document d, $\boldsymbol{w}_d$ then $\omega_1 = 1$, very similar to optimal estimator of variational parameters-ϕ_{dn}, γ_d in LDA model.
- The value of estimated $\hat{\phi}_{dni}$ maximizing the ELBO is:

$$\hat{\phi}_{dni} = \frac{\beta_{iv_d}^{1/\lambda} \exp((\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}))/\lambda)}{\sum_{i'=1}^{k} \beta_{i'v_d}^{1/\lambda} \exp((\Psi(\gamma_{di'}) - \Psi(\sum_{j=1}^{k} \gamma_{dj}))/\lambda)}$$

- The value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO is:

$$\hat{\gamma}_{di} = \alpha_i + \sum_{n=1}^{N_d} \phi_{dni}$$

Particle EM-M step with single particle

- In M step with single particle-$P = 1$, with all the documents-$\boldsymbol{w}_d, d = 1, \ldots, D$, maximize the ELBO with respect to model parameters-$\boldsymbol{\alpha}$,$\boldsymbol{\beta}$.
- The value of estimated $\hat{\beta}_{ij}$ maximizing the ELBO is:

$$\hat{\beta}_{ij} = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{j'=1}^{V} \phi_{dni} w_{dn}^{j'}}$$

- The estimation of $\alpha_i, i = 1, 2, \ldots, k$ by maximizing the ELBO can be solved numerically by Newton-Raphson method.

Particle EM-E step with multiple particles

- With $P > 1$, we need alternatively update the particle location-$[\phi_{1dn}, \phi_{2dn}, \ldots, \phi_{Pdn}]$ and their corresponding importance weights-$(\omega_1, \omega_2, \ldots, \omega_P)^T, \forall n = 1, 2, \ldots, N_d; d = 1, \ldots, D$.
- Apply the Newton-Raphson method, for each fixed $n = 1, 2, \ldots, N_d, d = 1, 2, \ldots, D, i = 1, 2, \ldots, k$, denote $\phi_{dni} = [\phi_{1dni}, \phi_{2dni}, \ldots, \phi_{Pdni}]^T$, then by iterating the equation below we can find the maximal-ϕ_{dni}:

$$\phi_{dni(new)} = \phi_{dni(old)} - H(\phi_{dni(old)})^{-1} g(\phi_{dni(old)})$$

- The value of estimated importance weight-$\hat{\omega}_p$ is:

$$\hat{\omega}_p = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \left(\phi_{pdni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{pdni} \log \beta_{iv_d} \right)}{\sum_{p'=1}^{P} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \left(\phi_{p'dni} (\Psi(\gamma_{di}) - \Psi(\sum_{j=1}^{k} \gamma_{dj})) + \phi_{p'dni} \log \beta_{iv_d} \right)} \tag{2}$$

- The value of estimated $\hat{\gamma}_{di}$ maximizing the ELBO is:

$$\hat{\gamma}_{di} = \alpha_i + \sum_{n=1}^{N_d} \sum_{p=1}^{P} \omega_p \phi_{pdni}$$

Particle EM-M step with multiple particles

- In M step with multiple particles-$P > 1$, with all the documents- $\boldsymbol{w}_d, d = 1, \ldots, D$, maximize the ELBO with respect to model parameters-$\boldsymbol{\alpha}$,$\boldsymbol{\beta}$.
- The value of estimated $\hat{\beta}_{ij}$ maximizing the ELBO is:

$$\hat{\beta}_{ij} = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{p=1}^{P} \omega_p \phi_{pdni} w_{dn}^j}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{j'=1}^{V} \sum_{p=1}^{P} \omega_p \phi_{pdni} w_{dn}^{j'}}$$

- The estimation of $\alpha_i, i = 1, 2, \ldots, k$ by maximizing the ELBO can be solved numerically by Newton-Raphson method with a Hessian matrix with special structure:

$$\boldsymbol{\alpha}_{(new)} = \boldsymbol{\alpha}_{(old)} - H'(\boldsymbol{\alpha}_{(old)})^{-1} g'(\boldsymbol{\alpha}_{(old)})$$

where $H'(\boldsymbol{\alpha}), g'(\boldsymbol{\alpha})$ are the Hessian matrix and gradient respectively at the point-$\boldsymbol{\alpha}$ defined above.
And the Hessian matrix $H'(\boldsymbol{\alpha})$ with the special form:

$$H'(\boldsymbol{\alpha}) = diag(\boldsymbol{h}) - \Psi'(\sum_{j=1}^{k} \alpha_j)\mathbf{1}\mathbf{1}^T$$

where $\boldsymbol{h} = [D\Psi'(\alpha_1), D\Psi'(\alpha_2), \ldots, D\Psi'(\alpha_k)]^T$

Stochastic Variational Inference in LDA

Stochastic variational inference is stochastic optimization with **noisy natural gradients** to optimize the variational objective function.
Which are local and global in LDA?

- **Local:** Document d, word $w_{d,1:N}$, topic parameters θ_d, topic assignment $z_{d,1:N}$.
- **Global:** hidden variable $\beta_{1:K}$, govern parameter $\lambda_{k,1:V}$

$\Rightarrow$ The $V-$dimensional variational distribution for β_k:

$$q(\beta_k) = \text{Dirichlet}(\lambda_k)$$

SVI Algorithm in LDA

1. Initialize $\lambda^{(0)}$ randomly.
2. Set the step-size schedule ρ_t appropriately
3. **repeat**
4. Sample a document w_d uniformly from the data set.
5. Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
6. **repeat**
7. For $n \in \{1, \ldots, N\}$ set

$$\phi_{dn}^k \propto \exp\{\mathbb{E}[\log \theta_{dk}] + [\log \beta_{k,w_{dn}}]\}, k \in \{1, \ldots, K\}.$$

8. Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
9. **until** local parameters ϕ_{dn} and γ_d converge.
10. $k \in \{1, \ldots, K\}$ set intermediate topics

$$\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{dn}^k w_{dn}.$$

11. $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$
12. **until** forever