# Rethinking about ME algorithm

**Donglai Wei**

2011.4.23

## 1) Quick Test for Burstiness

We change the word counts into binary (1 for appear at least once, 0 for no appearance) and rerun the current ME code. But the result is still degenerate...
I guess it is the issue of the MODE estimation v.s. ENSEMBLE estimation instead of the model.

## 2) Rethinking about ME algorithm

(A) **Why $\alpha, \gamma$ doesn't help in ME**:
In Gibbs sampling and Meanfield, the objection function considers the whole assignment space and tries to find where the most mass acculmulates in the Likelihood space.
Considering AEP/WLLN, the statistics (number of topics/tables) of the result tends to converge to their expected value, which is exactly controlled by $\alpha, \gamma$.
In ME, however, we are caring about the mode of the likelihood space where $\alpha, \gamma$ may only be a small factor in the objective function.

(B) **Drawbacks of the ensemble solution**:

1 Though it is believed that we can get results with desired statistics by tuning $\alpha, \gamma$ with Gibbs/Meanfield, the result actually varies a lot with $\lambda$ for the likelihood term in the objective function.

2 The ensemble configuration space is complicated and the local move/gradient based method easily get stuck.

(C) **Justification for ME**:

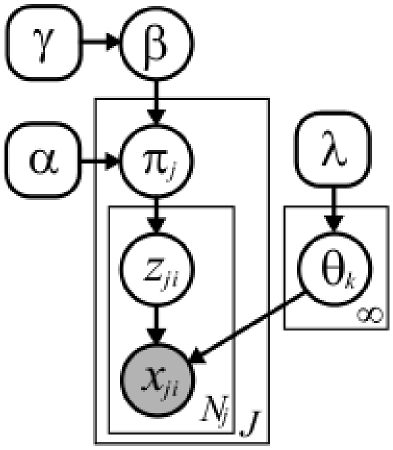
1 Though for less-structured data, the mode estimation that ME finds tends to be degenerate far away from the desired ensemble, we can reweight the Table-term and the Likelihood Term such that the mode estimation resides in the desired ensemble of the Likelihood space.

2 Also, ME can be used to reinitialize Gibbs/Meanfiled to help get out of stuck.

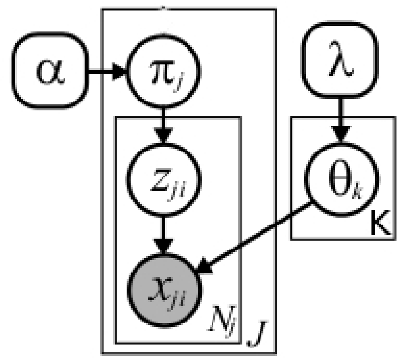
3 Theoretically, ME can easily clean up the topics, while Gibbs/Meanfield wanders around.

4 Experimentally, ME excels Gibb on bar data and real NIPS data in terms of predictive likelihood.

## 3) Formula for LDA evaluation




Stick-breaking representation for HDP




LDA

**HDP:**
$\beta \sim GEM(\gamma) \approx \mathcal{D}(\frac{\gamma}{K})$
$\pi_j \sim DP(\alpha, \beta) \approx \mathcal{D}(\alpha\beta)$
$\theta_k \sim H(\lambda) \approx \mathcal{D}(\lambda)$
$z_{ji} \sim \mathcal{M}(\pi_j)$
$x_{ji} \sim \mathcal{M}(\theta_{z_{ji}})$

**LDA:**
$\pi_j \sim \mathcal{D}(\alpha)$
$\theta_k \sim \mathcal{D}(\lambda)$
$z_{ji} \sim \mathcal{M}(\pi_j)$
$x_{ji} \sim \mathcal{M}(\theta_{z_{ji}})$

where
$\mathcal{D}(.)$: Dirichlet Distribution
$\mathcal{M}(.)$: Multinomial Distribution

(A) Get $t^*_{ji}, k^*_{jt}$ from ME algorithm in CRF representation and transform it into $\vec{z}^*$ in SB representation

(B) Given $\vec{z}^*$ and $\vec{x}$, approximate $\vec{\theta}^*$ and $\vec{\beta}^*$ in HDP for LDA evaluation

i. $\vec{\theta}^*$

$$\begin{aligned}
\vec{\theta}^* &= argmax_{\vec{\theta}}\ P(\vec{\theta}|\vec{z}^*, \vec{x}, \lambda) \\
&\sim argmax_{\vec{\theta}}\ P(\vec{\theta}, \vec{x}|\vec{z}^*, \lambda) \\
&= argmax_{\vec{\theta}}\ P(\vec{x}|\vec{z}^*, \vec{\theta})P(\vec{\theta}|\lambda) \\
&= argmax_{\vec{\theta}}\ [\Pi_{j,i}\mathcal{M}(x_{ji}; \theta_{z_{ji}})]\mathcal{D}(\vec{\theta}; \lambda) \\
&= argmax_{\vec{\theta}}\ [\Pi_{j,i}\mathcal{M}(x_{ji}; \theta_{z_{ji}})]\mathcal{D}(\vec{\theta}; \lambda) \\
&= argmax_{\vec{\theta}}\ \mathcal{D}(\vec{\theta}; \lambda + \vec{n}) \\
&= (\frac{\lambda + n_{kw}}{W\lambda + \sum_k n_{kw})})
\end{aligned}$$

where $n_{kw}$ is the count of number of words w in all restaurants that appear in topic k

ii. $\vec{\beta}^*$

$$\begin{aligned}
\vec{\beta}^* &= argmax_{\vec{\beta}}\ P(\vec{\beta}|\alpha,\gamma) \\
&= argmax_{\vec{\beta}}\ P(\vec{\beta},\vec{z}^*|\alpha,\gamma) \\
&= argmax_{\vec{\beta}}\ \int P(\vec{\beta},\vec{z}^*,\vec{\pi}|\alpha,\gamma)\ d\pi \\
&= argmax_{\vec{\beta}}\ [\int P(\vec{z}^*|\vec{\pi})P(\vec{\pi}|\alpha,\vec{\beta})\ d\pi]P(\vec{\beta}|\gamma) \\
&\approx argmax_{\vec{\beta}}\ [\int \mathcal{M}(\vec{z}^*;\vec{\pi})\mathcal{D}(\vec{\pi};\alpha\vec{\beta})\ d\pi]\mathcal{D}(\vec{\beta};\frac{\gamma}{K}) \\
&= argmax_{\vec{\beta}}\ [(\frac{\Pi_k\Gamma(\alpha\beta_k)}{\Gamma(\alpha\vec{\beta})})^J\Pi_j\frac{\Gamma(n_{j.}+\alpha\vec{\beta})}{\Pi_k\Gamma(n_{jk}+\alpha\beta_k)}][\frac{\Pi_k\Gamma(\frac{\gamma}{K})}{\Gamma(\gamma)}\Pi_k\beta_k^{\frac{\gamma}{K}}]
\end{aligned}$$

where $n_{jk}$ is the count of number of words in restaurant j that appear in topic k

(C) For LDA evaluation, use $\vec{\theta}^*$ for $\theta_{LDA}$ and and $\alpha\vec{\beta}^*$ for $\alpha_{LDA}$
(Note that $\vec{\beta}^*$ doesn't have a close form and need to be approximated from constraint nonlinear optimization.)