

Dirichlet priors for IBM model 1

Wilker Aziz

April 26, 2017

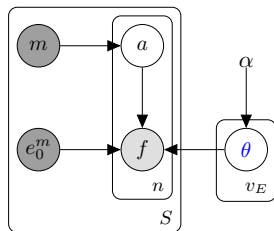
Content

IBM model 1

Bayesian IBM 1

Global assignments

- For each English type e ,
sample categorical parameters



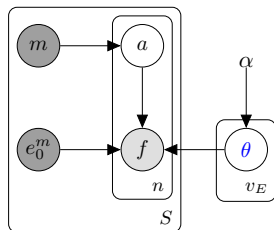
$$\theta_e \sim \text{Dir}(\alpha)$$

Bayesian IBM 1

Global assignments

- For each English type e ,
sample categorical parameters

$$\theta_e \sim \text{Dir}(\alpha)$$



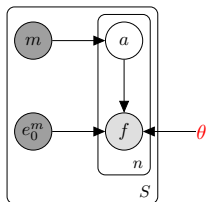
Local assignments

- For each French word position j ,

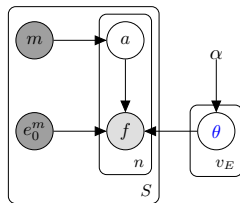
$$a_j \sim \mathcal{U}(0 \dots m)$$

$$f_j | e_{a_j} \sim \text{Cat}(\theta_{e_{a_j}})$$

MLE vs Bayesian IBM1



Incomplete data likelihood



$$P(f_1^n | e_1^m, \theta) = \prod_{j=1}^n \sum_{a_j=0}^m P(a_j | m) P(f_j | e_{a_j}, \theta) \quad (1)$$

Marginal likelihood (evidence)

$$P(f_1^n | e_1^m) = \int p(\theta | \alpha) P(f_1^n | e_1^m, \theta) d\theta \quad (2)$$

$$= \int p(\theta | \alpha) \prod_{j=1}^n \sum_{a_j=0}^m P(a_j | m) P(f_j | e_{a_j}, \theta) d\theta \quad (3)$$

Bayesian IBM 1: Joint Distribution

Sentence pair: (e_0^m, f_1^n)

$$p(f_1^n, a_1^n, \theta | e_0^m, \alpha) = P(a_1^n | m) \underbrace{\prod_e}_{\text{English types}} p(\theta_e | \alpha) \underbrace{\prod_f}_{\text{French types}} \overbrace{\theta_{f|e}^{\#(e \rightarrow f | a_1^n)}}^{\text{count } e \rightarrow f \text{ given } a_1^n} \quad (4)$$

$$= P(a_1^n | m) \underbrace{\prod_e \frac{\Gamma(\sum_f \alpha_f)}{\prod_f \Gamma(\alpha_f)}}_{\text{Dirichlet}} \underbrace{\prod_f \theta_{f|e}^{\alpha_f - 1} \prod_f \theta_{f|e}^{\#(e \rightarrow f | a_1^n)}}_{\text{Categorical}} \quad (5)$$

$$\propto P(a_1^n | m) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | a_1^n) + \alpha_f - 1} \quad (6)$$

Bayesian IBM 1: Joint Distribution (II)

Sentence pair: (e_0^m, f_1^n)

$$p(f_1^n, a_1^n, \theta | e_0^m, \alpha) \propto P(a_1^n | m) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | a_1^n) + \alpha_f - 1} \quad (7)$$

Corpus: (\mathbf{e}, \mathbf{f})

$$p(\mathbf{f}, \mathbf{a}, \theta | \mathbf{e}, \mathbf{m}, \alpha) \propto \prod_{(e_0^m, f_1^n, a_1^n)} P(a_1^n | m) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | a_1^n) + \alpha_f - 1} \quad (8)$$

$$= P(\mathbf{a} | \mathbf{m}) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f | \mathbf{a}) + \alpha_f - 1} \quad (9)$$

Bayesian IBM 1: Posterior

Intractable marginalisation

$$p(\mathbf{a}, \theta | \mathbf{e}, \mathbf{m}, \mathbf{f}, \alpha) = \frac{p(\mathbf{f}, \mathbf{a}, \theta | \mathbf{e}, \mathbf{m}, \alpha)}{\int \sum_{\mathbf{a}'} p(\mathbf{f}, \mathbf{a}', \theta' | \mathbf{e}, \mathbf{m}, \alpha) d\theta'} \quad (10)$$

- ▶ θ is a global variable: posterior depends on the entire corpus

Variational inference

Optimise an approximation to true posterior $p(a_1^n, \theta | e_0^m, f_1^n)$

Mean field

$$q(a_1^n, \theta | \phi, \lambda) = q(\theta | \lambda) \times Q(a_1^n | \phi) \quad (11)$$

$$= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j | \phi_j) \quad (12)$$

Variational inference

Optimise an approximation to true posterior $p(a_1^n, \theta | e_0^m, f_1^n)$

Mean field

$$q(a_1^n, \theta | \phi, \lambda) = q(\theta | \lambda) \times Q(a_1^n | \phi) \quad (11)$$

$$= \prod_{\mathbf{e}} q(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}}) \times \prod_{j=1}^n Q(a_j | \phi_j) \quad (12)$$

Maximise ELBO

$$(\hat{\lambda}, \hat{\phi}) = \arg \max_{\lambda, \phi} \mathbb{E}_q[\log p(f_1^n, a_1^n, \theta | e_0^m, \alpha)] + \mathbb{H}(q) \quad (13)$$

$$= \arg \max_{\lambda, \phi} \sum_{j=1}^m \mathbb{E}_q[\log P(a_j | m) P(f_j | e_{a_j}, \theta) - Q(a_j | \phi_j)] \quad (14)$$

$$+ \sum_{\mathbf{e}} \mathbb{E}_q[\log p(\theta_{\mathbf{e}} | \alpha) - q(\theta_{\mathbf{e}} | \lambda_{\mathbf{e}})] \quad (15)$$

Bayesian IBM 1: Posterior (II)

Local variables

$$P(a_j|e_0^m, f_j, \theta, \alpha) = \frac{\overbrace{P(a_j|m)}^{\text{constant}} \overbrace{P(f_j, |e_{a_j}, \theta, \alpha)}^{\theta_{f_j|e_{a_j}}}}{\sum_{i=0}^m P(i|m) P(f_j, |e_i, \theta, \alpha)} \quad (16)$$

Bayesian IBM 1: Posterior (II)

Local variables

$$P(a_j | e_0^m, f_j, \theta, \alpha) = \frac{\overbrace{P(a_j | m)}^{\text{constant}} \overbrace{P(f_j, | e_{a_j}, \theta, \alpha)}^{\theta_{f_j | e_{a_j}}}}{\sum_{i=0}^m P(i | m) P(f_j, | e_i, \theta, \alpha)} \quad (16)$$

thus $Q(a_j | \phi_j) = \text{Cat}(\phi_j)$

Bayesian IBM 1: Posterior (II)

Local variables

$$P(a_j|e_0^m, f_j, \theta, \alpha) = \frac{\overbrace{P(a_j|m)}^{\text{constant}} \overbrace{P(f_j, |e_{a_j}, \theta, \alpha)}^{\theta_{f_j|e_{a_j}}}}{\sum_{i=0}^m P(i|m) P(f_j, |e_i, \theta, \alpha)} \quad (16)$$

thus $Q(a_j|\phi_j) = \text{Cat}(\phi_j)$

Global variables

$$p(\theta_e|\mathbf{e}, \mathbf{f}, \mathbf{a}, \alpha) \propto \prod_{(e_0^m, f_1^n, a_1^n)} p(f_1^n, a_1^n, \theta_e|e_0^m, \alpha) \quad (17)$$

$$= P(\mathbf{a}|\mathbf{m}) \prod_e \prod_f \theta_{f|e}^{\#(\mathbf{e} \rightarrow \mathbf{f}|\mathbf{a}) + \alpha_f - 1} \quad (18)$$

Bayesian IBM 1: Posterior (II)

Local variables

$$P(a_j|e_0^m, f_j, \theta, \alpha) = \frac{\overbrace{P(a_j|m)}^{\text{constant}} \overbrace{P(f_j, |e_{a_j}, \theta, \alpha)}^{\theta_{f_j|e_{a_j}}}}{\sum_{i=0}^m P(i|m) P(f_j, |e_i, \theta, \alpha)} \quad (16)$$

thus $Q(a_j|\phi_j) = \text{Cat}(\phi_j)$

Global variables

$$p(\theta_e|\mathbf{e}, \mathbf{f}, \mathbf{a}, \alpha) \propto \prod_{(e_0^m, f_1^n, a_1^n)} p(f_1^n, a_1^n, \theta_e|e_0^m, \alpha) \quad (17)$$

$$= P(\mathbf{a}|\mathbf{m}) \prod_e \prod_f \theta_{f|e}^{\#(e \rightarrow f|\mathbf{a}) + \alpha_f - 1} \quad (18)$$

thus $q(\theta_e|\lambda_e) = \text{Dir}(\lambda_e)$

VB for IBM1

Optimal $q(a_j|\phi_j)$

$$\phi_j = \frac{\exp \left(\Psi \left(\lambda_{f_j|e_{a_j}} \right) - \Psi \left(\sum_{\mathbf{f}} \lambda_{\mathbf{f}|e_{a_j}} \right) \right)}{\sum_{i=0}^m \exp \left(\Psi \left(\lambda_{f_j|e_i} \right) - \Psi \left(\sum_{\mathbf{f}} \lambda_{\mathbf{f}|e_i} \right) \right)} \quad (19)$$

VB for IBM1

Optimal $q(a_j|\phi_j)$

$$\phi_j = \frac{\exp\left(\Psi\left(\lambda_{f_j|e_{a_j}}\right) - \Psi\left(\sum_{\mathbf{f}} \lambda_{\mathbf{f}|e_{a_j}}\right)\right)}{\sum_{i=0}^m \exp\left(\Psi\left(\lambda_{f_j|e_i}\right) - \Psi\left(\sum_{\mathbf{f}} \lambda_{\mathbf{f}|e_i}\right)\right)} \quad (19)$$

Optimal $q(\theta_e|\lambda_e)$

$$\lambda_{\mathbf{f}|\mathbf{e}} = \alpha_{\mathbf{f}} + \sum_{(e_0^m, f_1^n)} \sum_{j=1}^n \mathbb{E}_{Q(A_j|\phi_j)}[\#(\mathbf{e} \rightarrow \mathbf{f}|A_j)] \quad (20)$$

Algorithmically

E-step as in MLE IBM1,

however, using $Q(a_j|\phi_j)$ instead of $P(a_j|e_0^m, f_j, \theta)$

- ▶ equivalent to using $\theta \approx \hat{\theta}$ where
- ▶ $\hat{\theta}_{f|e} = \exp(\Psi(\lambda_{f|e}) - \Psi(\sum_{f'} \lambda_{f'|e}))$

Algorithmically

E-step as in MLE IBM1,

however, using $Q(a_j|\phi_j)$ instead of $P(a_j|e_0^m, f_j, \theta)$

- ▶ equivalent to using $\theta \approx \hat{\theta}$ where
- ▶ $\hat{\theta}_{f|e} = \exp(\Psi(\lambda_{f|e}) - \Psi(\sum_{f'} \lambda_{f'|e}))$

M-step

- ▶ $\lambda_{f|e} = \alpha_f + \mathbb{E}[\#(e \rightarrow f)]$
where expected counts come from E-step

References I