Dirichlet priors for IBM model 1

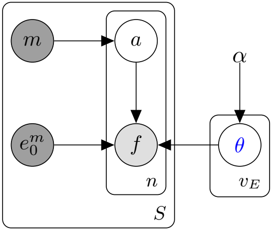
Wilker Aziz

April 28, 2017

Content

IBM model 1

Bayesian IBM 1

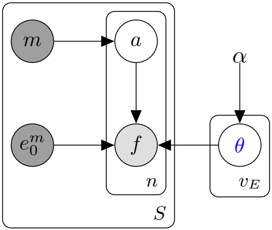


Global assignments

- For each English type e, sample categorical parameters

$$\theta_{\mathsf{e}} \sim \mathrm{Dir}(\alpha)$$

Bayesian IBM 1



Global assignments

- For each English type e, sample categorical parameters

$$\theta_{\mathsf{e}} \sim \mathrm{Dir}(\alpha)$$

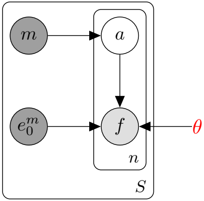
Local assignments

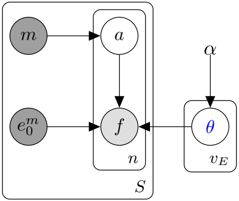
- For each French word position j,

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

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

MLE vs Bayesian IBM1





Incomplete data likelihood

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

Marginal likelihood (evidence)

$$P(f_1^n|e_1^m, \alpha) = \int p(\theta|\alpha)P(f_1^n|e_1^m, \theta)\mathrm{d}\theta \tag{2}$$

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

Bayesian IBM 1: Joint Distribution

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

$$p(f_1^n, a_1^n, \theta | e_0^m, \alpha) = P(a_1^n|m) \underbrace{\prod_{\mathsf{e}}}_{\text{English types}} p(\theta_{\mathsf{e}}|\alpha) \underbrace{\prod_{\mathsf{f}}}_{\text{French types}} \theta_{\mathsf{f}|\mathsf{e}}^{\overbrace{\#(\mathsf{e} \to \mathsf{f}|a_1^n)}^{\text{count e}\to\text{f given } a_1^n}} \tag{4}$$

$$= P(a_1^n|m) \prod_{\mathsf{e}} \underbrace{\frac{\Gamma(\sum_{\mathsf{f}} \alpha_{\mathsf{f}})}{\prod_{\mathsf{f}} \Gamma(\alpha_{\mathsf{f}})} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\alpha_{\mathsf{f}}-1}}_{\text{Dirichlet}} \underbrace{\prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e}\to\mathsf{f}|a_1^n)}}_{\text{Categorical}} \tag{5}$$

$$\propto P(a_1^n|m) \prod_{\mathsf{e}} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e}\to\mathsf{f}|a_1^n)+\alpha_{\mathsf{f}}-1} \tag{6}$$

Bayesian IBM 1: Joint Distribution (II)

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

$$p(f_1^n, a_1^n, \theta | e_0^m, \alpha) \propto P(a_1^n | m) \prod_{\mathsf{e}} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e} \to \mathsf{f} | a_1^n) + \alpha_{\mathsf{f}} - 1} \tag{7}$$

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

$$p(\mathbf{f}, \mathbf{a}, \theta | \mathbf{e}, \mathbf{m}, \alpha) \propto \prod_{(e_0^m, f_1^n, a_1^n)} P(a_1^n | m) \prod_{\mathsf{e}} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e} \to \mathsf{f} | a_1^n) + \alpha_{\mathsf{f}} - 1} \tag{8}$$

$$= P(\mathbf{a} | \mathbf{m}) \prod_{\mathsf{e}} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e} \to \mathsf{f} | \mathbf{a}) + \alpha_{\mathsf{f}} - 1} \tag{9}$$

Bayesian IBM 1: Posterior

Intractable marginalisation

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

- 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

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

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

Variational inference

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

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

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

Maximise ELBO

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

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

$$+ \sum_{\mathsf{e}} \underbrace{\mathbb{E}_q[\log p(\theta_e|\alpha) - \log q(\theta_e|\lambda_e)]}_{-\,\mathrm{KL}(q(\theta_e|\lambda_e)||p(\theta_e|\alpha))} \tag{15}$$

Bayesian IBM 1: Posterior (II)

Local variables

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

Bayesian IBM 1: Posterior (II)

Local variables

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

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

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

Global variables

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

$$= P(\mathbf{a}|\mathbf{m}) \prod_{\mathsf{e}} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e} \to \mathsf{f}|\mathbf{a}) + \alpha_{\mathsf{f}} - 1} \quad (18)$$

Bayesian IBM 1: Posterior (II)

Local variables

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

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

Global variables

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

$$= P(\mathbf{a}|\mathbf{m}) \prod_{\mathsf{e}} \prod_{\mathsf{f}} \theta_{\mathsf{f}|\mathsf{e}}^{\#(\mathsf{e} \to \mathsf{f}|\mathbf{a}) + \alpha_{\mathsf{f}} - 1} \tag{18}$$

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

VB for IBM1

Optimal $q(a_j|\phi_j)$

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

VB for IBM1

Optimal $q(a_j|\phi_j)$

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

Optimal $q(\theta_e|\lambda_e)$

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

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
- $\hat{\theta}_{\mathsf{f}|\mathsf{e}} = \exp\left(\Psi\left(\lambda_{\mathsf{f}|\mathsf{e}}\right) - \Psi\left(\sum_{\mathsf{f}'} \lambda_{\mathsf{f}'|\mathsf{e}}\right)\right)$

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
- $\hat{\theta}_{\mathsf{f}|\mathsf{e}} = \exp\left(\Psi\left(\lambda_{\mathsf{f}|\mathsf{e}}\right) - \Psi\left(\sum_{\mathsf{f}'} \lambda_{\mathsf{f}'|\mathsf{e}}\right)\right)$

M-step

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

References I