Outline

- Poisson Matrix Factorization (PMF)

- Empirical Bayes Approach (EB-PMF)
  - sparsity assumption

- Background Model (EB-PMF-WBG)
  - new "sparsity" idea

  WBG: weighted background

- Results on simulated dataset

- Results on real dataset (text)

Poisson Matrix Factorization (PMF)

- Model $\underset{n\times p}{X_{ij}} \sim \text{Pois}(\sum_k l_{ik} f_{jk})$, $L \in R_+^{n\times K}$, $F \in R_+^{p\times K}$

- EM algorithm

$$\begin{cases} X_{ij} = \sum_k Z_{ijk} \\ Z_{ijk} \sim \text{Pois}(l_{ik} f_{jk}) \end{cases}$$

If we know Z, L $\longrightarrow$

$x_j = \sum_i Z_{ijk}$

$s_j = \sum_i l_{ik}$

Poisson Means problem

$X_j \sim \text{Pois}(s_j \lambda_j)$.

x_j, s_j known,

$\hat{\lambda}_j \leftarrow$?

MLE: $\frac{x_j}{s_j}$.

- $E_{Z|L,F,X} \log p(X,Z|L,F)$

$$= \sum_{i,j,k} \left(-l_{ik} f_{jk} + \bar{Z}_{ijk} \log(l_{ik} f_{jk})\right)$$

- E-step: compute $\bar{Z}_{ijk} \triangleq E_{Z|L,F,X}(Z_{ijk})$

- M-step: focus on f_{jk}, $j = 1, \cdots p$.

we have the Poisson Means problem,

and the MLE gives $f_{jk} = \dfrac{\sum_i \bar{Z}_{ijk}}{\sum_i l_{ik}}$

- want to impose some assumptions on L, F.

$\Rightarrow$ Empirical Bayes approach (EB PMF)

- Model :

$$\begin{cases} X_{ij} \sim \text{Pois}(\sum_k l_{ik} f_{jk}) \\ l_{ik} \sim g_k^{(L)}(\cdot), \quad f_{jk} \sim g_k^{(F)}(\cdot) \\ g_k^{(L)}, \ g_k^{(F)} \in \mathcal{G} \end{cases}$$

- impose assumption through the choice of $\mathcal{G}$
  (e.g. point gamma family,
  hope to impose sparsity assumption)

- $g_k^{(L)}, g_k^{(F)}$ is estimated from data
  (empirical Bayes)

- Model :
$$\begin{cases} X_{ij} \sim \mathrm{Pois}(\sum_k l_{ik} f_{jk}) \\ l_{ik} \sim g_k^{(L)}(\cdot), \quad f_{jk} \sim g_k^{(F)}(\cdot) \\ g_k^{(L)}, \ g_k^{(F)} \in \mathcal{G} \end{cases}$$

- Use Z - trick and Mean-field Variational Inference.

$$q(Z, L, F) = \prod_{ijk} q(Z_{ijk}) \prod_{ik} q(l_{ik}) \prod_{jk} q(f_{jk})$$

$$\mathrm{ELBO}(q, g) = E_q \log p(X, L, F, Z | g) - E_q[\log q(L, F, Z)]$$
$$= E_q[\log p(Z | L, F)] - KL(q_L \| g_L) - KL(q_F \| g_F) - \log q(Z)$$

look at parts relevant to f_{jk}, $j = 1, \cdots, p$

$$\mathrm{ELBO}(q_k^{(F)}, g_k^{(F)}) = \sum_{i,j} (-\bar{l}_{ik} \bar{f}_{jk} + \bar{Z}_{ijk} \overline{(\log f_{jk})}) - KL(q_k^F \| g_k^F)$$

$\Rightarrow$ EB Poisson Means : $X_j \sim \mathrm{Pois}(s_j \cdot \lambda_j)$
$\lambda_j \sim g(\cdot)$

Step 1: $\hat{g} = \arg\max \log p(X | g, s) \triangleq l(g)$

Step 2 compute $p(\lambda | x, \hat{g}) = q$ in our model

https://github.com/stephenslab/ebpmf.alpha/blob/master/derivations/ebpmf.pdf

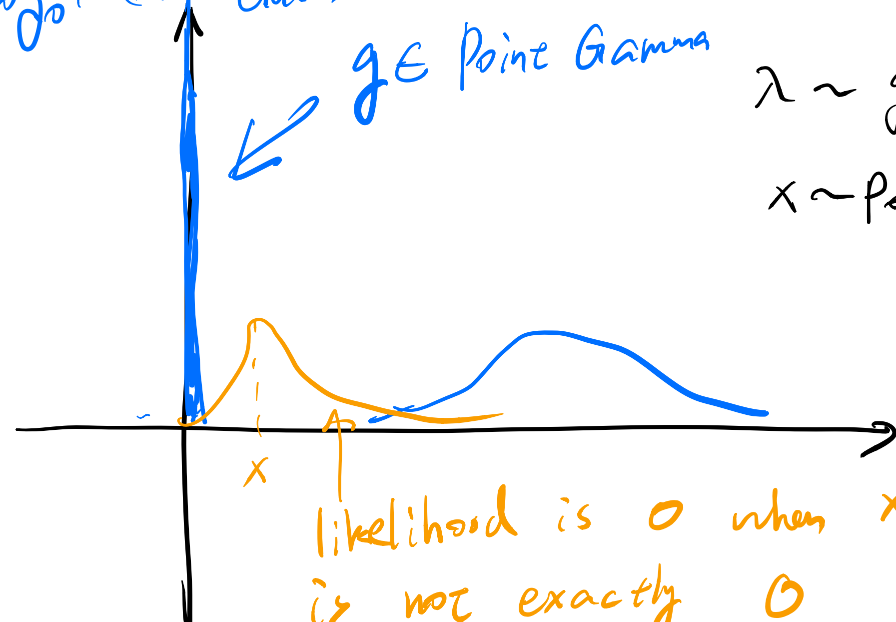
- Tried Sparsity Assumption. Not work

$g = \pi_0 \delta_0 + (1-\pi_0)\, Ga(\cdot\,; a, b)$

$g \in$ Point Gamma

$\lambda \sim g(\cdot)$

$x \sim Pois(\lambda)$



likelihood is 0 when x is not exactly 0

($Z_{ijk} \neq 0$ when $x_{ij} \neq 0$)

- Another view of "sparsity"

$$x_{ij} \sim \text{Pois}\left(\sum_k \tilde{l}_{ik} \tilde{f}_{jk}\right)$$

- Not assuming $\tilde{l}_{ik}$, $\tilde{F}_{jk}$ are sparse,

- $\tilde{f}_{jk} = f_{j0} f_{jk}$

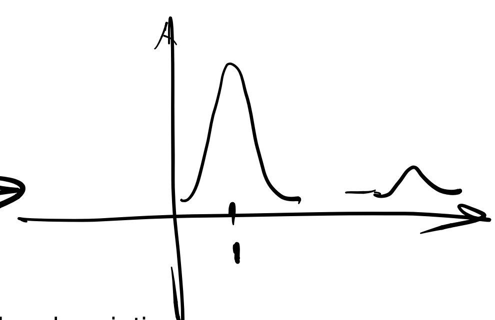
Background frequency for word j

in topic k, how much does word j deviate from its background.

- "sparse" means f_{jk} mostly around 1;

when $f_{jk} >> 1$, it's important for ~~topic~~ k.

- $f_{jk} \sim g_k^{(F)}$



https://zihao12.github.io/ebpmf_data_analysis/ebpmf_wbg_description

- New Model EB Poisson Matrix Factorization with weighted background (EBPMF-WBG)

$$\begin{cases} X_{ij} \sim \mathrm{Pois}\left(\sum_k w_k \, l_{i0} l_{ik} f_{j0} f_{jk}\right) \\ l_{ik} \sim \sum_{l=1}^{L} \pi_l^{(L,k)} \, \mathrm{Ga}(1/\phi_l, 1/\phi_l) \\ f_{jk} \sim \sum_{l=1}^{L} \pi_l^{(F,k)} \, \mathrm{Ga}(1/\phi_l, 1/\phi_l) \end{cases}$$

- Results on simulated dataset

https://zihao12.github.io/ebpmf_data_analysis/ebpmf_wbg_simulation_big2_2

- Results on real data

https://zihao12.shinyapps.io/topicview-app/