

# These are notes for Buntine & Jakulin paper: Discrete Component Analysis [1]

Gamma-Poisson (GP) model [2]:

$$\mathbb{E}_{w \sim p(w|\mathbf{l}, \theta)} [w_j] = \sum_{k=1}^K \theta_{jk} l_k$$

- $w_j$  word count of  $j$ th word in a document

$$w_j \sim \text{Po}(w_j; (\theta \mathbf{l})_j) = \frac{(\theta \mathbf{l})_j^{w_j} \exp(-(\theta \mathbf{l})_j)}{w_j!}$$

- $l_k$  component scores (vector  $\mathbf{l}$ ) that indicate amount of the component in the document

$$l_k \sim \text{Gamma}(l_k; \alpha_k, \beta_k) = \frac{l_k^{\alpha_k-1} \beta_k^{\alpha_k} \exp(-\beta_k l_k)}{\Gamma(\alpha_k)}$$

- $\theta$  component loading matrix of size  $J \times K$ .  $\theta_{jk}$  controls partition of the  $k$ th component in the  $j$ th word

The log-likelihood of this model:

$$\begin{aligned} \log p(\mathbf{w}, \mathbf{l} | \theta, \text{GP}, K) &= \sum_{k=1}^K \left\{ \alpha_k \log(\beta_k) + (\alpha_k - 1) \log l_k - \beta_k l_k - \log \Gamma(\alpha_k) + \sum_{j=1}^J [w_j \log(\theta \mathbf{l})_j - (\theta \mathbf{l})_j - \log w_j!] \right\} \\ &= \sum_{k=1}^K \log \text{likelihood of } l_k + \sum_{j=1}^J \log \text{likelihood of } w_j \text{ given } \mathbf{l} \end{aligned} \quad (1)$$

## Section 6: Components assignment for words.

Introducing a discrete latent vector  $\mathbf{c}$  whose total count is  $\sum_j w_j$ . The count  $c_k$  gives the count of words in the document appearing in the  $k$ th component. It is derived from a latent matrix  $\mathbf{V}$  of size  $J \times K$  (entries  $v_{jk}$ ).

$$\begin{aligned} \sum_{j=1}^J v_{jk} &= c_k \\ \sum_{k=1}^K v_{jk} &= w_j \end{aligned}$$

The distribution underlying the GP model now becomes

$$\begin{aligned} l_k &\sim \text{Gamma}(l_k; \alpha_k, \beta_k) \\ c_k &\sim \text{Po}(c_k; l_k) \\ v_{j,k} &\sim \text{Multinom}(v_{jk}; \theta_{jk}, c_k) = c_k! \prod_j \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \end{aligned}$$

Proof:

We have  $p(c_k | l_k) = \text{Po}(c_k; l_k)$  and  $p(v_{jk} | c_k) = \text{Binom}(v_{jk}; \theta_{jk}, c_k) = \binom{c_k}{v_{jk}} \theta_{jk}^{v_{jk}} (1 - \theta_{jk})^{c_k - v_{jk}}$  (probability of having  $v_{jk}$  counts in  $c_k$  counts). Then:

$$\begin{aligned}
p(v_{jk}|l_k) &= \sum_{c_k} p(v_{jk}|c_k)p(c_k|l_k) \\
&= \sum_{c_k=v_{jk}}^{\infty} \frac{c_k!}{v_{jk}!(c_k - v_{jk})!} \theta_{jk}^{v_{jk}} (1 - \theta_{jk})^{c_k - v_{jk}} \times \frac{l_k^{c_k} \exp(-l_k)}{c_k!} \\
&= \frac{\exp(-l_k) \theta_{jk}^{v_{jk}}}{v_{jk}!} \sum_{c_k=v_{jk}}^{\infty} \frac{l_k^{c_k} (1 - \theta_{jk})^{c_k - v_{jk}}}{(c_k - v_{jk})!} \quad |\alpha_{jk} = c_k - v_{jk} \\
&= \frac{\exp(-l_k) (\theta_{jk} l_k)^{v_{jk}}}{v_{jk}!} \sum_{\alpha_{jk}=0}^{\infty} \frac{(l_k - \theta_{jk} l_k)^{\alpha_{jk}}}{(\alpha_{jk})!} \\
&= \frac{\exp(-l_k) (\theta_{jk} l_k)^{v_{jk}}}{v_{jk}!} \exp(l_k - \theta_{jk} l_k) \\
&= \frac{(\theta_{jk} l_k)^{v_{jk}} \exp(-\theta_{jk} l_k)}{v_{jk}!}
\end{aligned}$$

and so  $p(v_{jk}|l_k) \sim \text{Po}(v_{jk}; \theta_{jk} l_k)$ .

Now sum of two independent Poisson distributed variables  $Z = X_1 + X_2$  ( $X_i \sim \text{Po}(x; \lambda_i)$ ) is Poisson distributed:

$$\begin{aligned}
p(Z) &= \sum_{x_1=0}^z p(X_1) p(Z - X_1) \\
&= \sum_{x_1=0}^z \frac{\lambda_1^{x_1} e^{-\lambda_1}}{x_1!} \frac{\lambda_2^{z-x_1} e^{-\lambda_2}}{(z-x_1)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{z!} \sum_{x_1=0}^z \frac{z!}{x_1!(z-x_1)!} \lambda_1^{x_1} \lambda_2^{z-x_1} \\
&= \frac{(\lambda_1 + \lambda_2)^z e^{-(\lambda_1+\lambda_2)}}{z!}
\end{aligned}$$

for more by induction.

So  $w_j = \sum_{k=1}^K v_{jk}$  is Poisson distributed:

$$w_j \sim \text{Po}(w_j; \sum_{k=1}^K \theta_{jk} l_k)$$

The joint distribution for  $v_{jk}$  (each is Poisson):

$$\begin{aligned}
p(v_{1,k}, v_{2,k} \dots v_{J,k} | l_k, \theta_{jk}) &= \prod_{j=1}^J \frac{(\theta_{jk} l_k)^{v_{jk}} \exp(-\theta_{jk} l_k)}{v_{jk}!} \\
&= e^{-l_k \sum_j \theta_{jk}} l_k^{\sum_j v_{jk}} \prod_j \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \quad | \sum_j \theta_{jk} = 1, \sum_j v_{jk} = c_k \\
&= \frac{l_k^{c_k} e^{-l_k}}{c_k!} c_k! \prod_j \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \\
&= \text{Po}(c_k; l_k) \times \text{Multinom}(v_{jk}; \theta_{jk}, c_k)
\end{aligned}$$

The likelihood of GaP model with latent matrix  $V$  is then

$$\begin{aligned}
p(V, l | \alpha, \beta, \theta, K) &= \prod_k p(l_k | \alpha_k, \beta_k) \prod_{jk} p(v_{1k}, v_{2k} \dots v_{J,k} | l_k, \theta_{jk}) \\
&= \prod_k \text{Gamma}(l_k; \alpha_k, \beta_k) \prod_{jk} \text{Po}(c_k; l_k) \times \text{Multinom}(v_{jk}; \theta_{jk}, c_k)
\end{aligned}$$

explicitly:

$$p(V, l | \alpha, \beta, \theta, K) = \prod_k \frac{\beta_k^{\alpha_k} l_k^{c_k + \alpha_k - 1} \exp(-(\beta_k + 1)l_k)}{\Gamma(\alpha_k)} \prod_{jk} \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \quad (2)$$

and

$$\log p(V, l | \alpha, \beta, \theta, K) = \sum_k \left\{ (c_k + \alpha_k - 1) \log l_k - (\beta_k + 1)l_k + \alpha_k \log \beta_k - \log \Gamma(\alpha_k) + \sum_j [v_{jk} \log \theta_{jk} - \log v_{jk}!] \right\} \quad (3)$$

$w_j$  is derived from  $V$  so it is not represented.

It is possible to integrate out  $l$  (not sure about discrete values...?):

$$\begin{aligned} p(V | \alpha, \beta, \theta, K) &= \int_0^\infty p(V, l | \alpha, \beta, \theta, K) dl \\ &= \prod_{jk} \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!} \prod_k \frac{\beta_k}{\Gamma(\alpha_k)} \int_0^\infty [l_k^{c_k + \alpha_k - 1} \exp(-(\beta_k + 1)l_k)] dl_k \end{aligned}$$

and

$$\begin{aligned} \int_0^\infty [l_k^{c_k + \alpha_k - 1} \exp(-(\beta_k + 1)l_k)] dl_k &= \int_0^\infty l_k^{z-1} \exp(-(\beta_k + 1)l_k) dl_k \quad | c_k + \alpha_k = z \\ &= \frac{1}{(\beta_k + 1)^z} \int_0^\infty t^{z-1} \exp(-t) dt \quad | (\beta_k + 1)l_k = t \\ &= \frac{1}{(\beta_k + 1)^z} \Gamma(z) \end{aligned}$$

so

$$p(V | \alpha, \beta, \theta, K) = \prod_k \frac{\beta_k}{(\beta_k + 1)^{c_k + \alpha_k}} \frac{\Gamma(c_k + \alpha_k)}{\Gamma(\alpha_k)} \prod_{jk} \frac{\theta_{jk}^{v_{jk}}}{v_{jk}!}$$

## EM algorithm

The term  $l_k^{(c_k + \alpha_k - 1)} = l_k^{(\sum_j v_{jk} + \alpha_k - 1)}$  in Eq.(2) links together  $l_k$  and  $V$  and prevents simple evaluation of  $\mathcal{Q}(\theta, \theta^{\text{old}}) = \mathbb{E}_{p(V, l | \theta^{\text{old}})} [\log p(V, l | \theta, \dots)]$  in the EM algorithm because of the term  $\mathbb{E}_{p(V, l | \theta^{\text{old}})} [v_{jk}]$ .

In the likelihood Eq.(1) is problematic the term  $w_k \log \sum_k \theta_{jk} l_k$ . (In [2] is the term  $\mathbb{E}_l [\log \sum_k \theta_{jk} l_k]$  approximated by  $\log \mathbb{E}_l [\sum_k \theta_{jk} l_k]$  which might be quite crude.)

## Section 7.1 Variational Approximation

Factorised approximate posterior distribution for latent variables:

$$p(l, V | w, \alpha, \beta, \theta, K) \approx q(l, V) = q_l(l) q_V(V)$$

Optimal solution [3]

$$\log q_l^*(l) = \mathbb{E}_{V \sim q_V} [\log p(V, l, w | \theta, \alpha, \beta)] + \text{const} \quad (4)$$

$$\log q_V^*(V) = \mathbb{E}_{l \sim q_l} [\log p(V, l, w | \theta, \alpha, \beta)] + \text{const} \quad (5)$$

The lower bound is given by [3]

$$\mathcal{L}(q, \theta) = \sum_z p(Z | X, \theta^{\text{old}}) \log p(X, Z | \theta) + C = \mathcal{Q}(\theta, \theta^{\text{old}}) + C$$

where

$$C = -\mathbb{E}_{l \sim q_l} [\log q_l] - \mathbb{E}_{V \sim q_V} [\log q_V] = H(q_l) + H(q_V)$$

are the entropy terms (independent on  $\theta$ ).

$$\log p(w|\theta, \alpha, \beta, K) \geq \mathbb{E}_{l, V \sim q(l, V)} [\log p(l, V, w|\theta, \alpha, \beta, K)] + C \quad (6)$$

The functional form of the complete likelihood suggests

$$q_l(l) = \prod_k \text{Gamma}(l_l; \alpha_k, \beta_k) = \prod_k \frac{l_k^{\alpha_k-1} b_k^{\alpha_k} \exp(-b_k l_k)}{\Gamma(\alpha_k)} \quad (7)$$

$$q_V(V) = \prod_{jk} \text{Mutlinom}(v_{jk}; n_{jk}, w_j) = \prod_{jk} \frac{w_j!}{v_{jk}!} n_{jk}^{v_{jk}} \quad (8)$$

with  $\sum_k n_{jk} = 1$ .

Then from Eq.(4), (7) and (3) keeping terms dependent on  $l$

$$(a_k - 1) \log l_k - b_k l_k + \text{const} = \left( \sum_j \mathbb{E}_V [v_{jk}] + \alpha_k - 1 \right) \log l_k - (\beta_k + 1) l_k + \text{const}$$

where  $c_k = \sum_j v_{jk}$ . Form Eq.(5), (8) and (3) keeping terms dependent on  $V$

$$v_{jk} \log n_{jk} - \log v_{jk}! + \text{const} = v_{jk} \mathbb{E}_l [\log l_k] + v_{jk} \log \theta_{jk} - \log v_{jk}! + \text{const}$$

so the rewrite rules for the parameters:

$$\begin{aligned} n_{jk} &= \frac{1}{z_j} \theta_{jk} \exp(\mathbb{E}_l [\log l_k]) \\ a_k &= \sum_j n_{jk} w_j + \alpha_k \\ b_k &= 1 + \beta_k \end{aligned} \quad (9)$$

where  $z_j$  is the normalisation constant ( $\sum_k n_{jk} = 1$ ) so  $z_j = \sum_k \theta_{jk} \exp(\mathbb{E}_l [\log l_k])$  and  $\sum_j \mathbb{E}_V [v_{jk}] = \sum_j n_{jk} w_j$  (Eq.(5)).  $\mathbb{E}_{l \sim q_l} [\log l_k] = \psi_0(a_k) - \log b_k$  where  $\psi_0$  is digamma function (logarithmic derivation of the gamma function...)

Now recompute model parameter  $\theta$  by maximising lower bound Eq.(6) (keeping constraints  $\sum_j \theta_{jk} = 1$ ). Keeping only term dependent on  $\theta_{jk}$ :

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{j,k} \mathbb{E}_{q_V(V)} [v_{jk}] \log \theta_{jk} + \text{const} \\ &= \sum_{j,k} n_{jk} w_j \log \theta_{jk} + \text{const} \end{aligned}$$

(from Eq.(8)  $\mathbb{E}_{q_V(V)} [v_{jk}] = w_j n_{jk}$ )

$$0 = \frac{\partial}{\partial \theta_{mn}} \left[ \sum_{j,k} n_{jk} w_j \log \theta_{jk} + \lambda_k (1 - \sum_p \theta_{pk}) \right]$$

we get

$$\theta_{mn} = \frac{n_{mn} w_m}{\lambda_n}$$

and from normalisation constraints  $\lambda_n = \sum_m n_{mn} w_m$ .

If we take likelihood function over all documents ( $i = 1 : L$ ) each  $w_j \rightarrow w_{j(i)}$  and  $n_{jk} \rightarrow n_{jk(i)}$  then we get

$$\theta_{mn} = \frac{\sum_i n_{mn(i)} w_{m(i)}}{\lambda_n} \quad (10)$$

Buntine [1] even introduce prior on  $\theta_{jk} \sim \text{Dirichlet}(\theta_{jk}; \gamma, J) = C(\gamma_j) \prod_{j=1}^J \theta_{jk}^{\gamma_j - 1}$ . This is incorporated into the complete log-likelihood function  $p(V, l, w, \theta | \alpha, \beta, K)$  so that lower bound  $\mathbb{E}_{l, V \sim q(l, V)} [\log p(l, V, w, \theta | \alpha, \beta, K)]$  and terms dependent on  $\theta$ :

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i, j, k} \mathbb{E}_{q_V(V)} [v_{jk(i)}] \log \theta_{jk} + (\gamma_j - 1) \log \theta_{jk} + \text{const} \\ &= \left( \sum_{i, j, k} n_{jk(i)} w_{j(i)} + \gamma_j - 1 \right) \log \theta_{jk} + \text{const} \end{aligned}$$

and by maximising with normalisation constraints:

$$\theta_{mn} \propto \sum_i n_{mn(i)} w_{m(i)} + \gamma_j \quad (11)$$

The lower bound Eq.(6)

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}_{l, V \sim q(l, V)} \left[ \sum_k (c_k + \alpha_k - 1) \log l_k - (\beta_k + 1) l_k + \log \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} + \sum_j [v_{jk} \log \theta_{jk} - \log v_{jk!}] \right] + C \\ &= \sum_k \mathbb{E}_l [\log l_k] \left( \sum_j \mathbb{E}_V [v_{jk}] + \alpha_k - 1 \right) - (\beta_k + 1) l_k + \log \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \\ &\quad + \sum_j [\mathbb{E}_V [v_{jk}] (\log n_{jk} + \log z_j - \mathbb{E}_l [\log l_k] - \mathbb{E}_V [\log v_{jk!}])] + C \\ &= \sum_k \mathbb{E}_l [\log l_k] (\alpha_k - 1) - (\beta_k + 1) l_k + \log \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} + \sum_j [\mathbb{E}_V [v_{jk}] (\log n_{jk} + \log z_j - \log v_{jk!})] + C \end{aligned}$$

where Eq.(9) for  $\theta$  was used and  $c_k = \sum_j v_{jk}$ .

Including entropy terms  $C = H(q_l) + H(q_V)$  from Eq.(6)

$$\begin{aligned} H(q_l) &= - \sum_k \left\{ (a_k - 1) \mathbb{E}_l [\log l_k] - b_k \mathbb{E}_l [l_k] - \log \frac{b_k^{a_k}}{\Gamma(a_k)} \right\} \\ H(q_V) &= - \sum_{jk} \{ -\mathbb{E}_V [\log v_{jk!}] + \mathbb{E}_V [v_{jk}] \log n_{jk} + \log w_j! \} \end{aligned}$$

we get

$$\mathcal{L} = \sum_k \mathbb{E}_l [\log l_k] (\alpha_k - a_k) + \sum_j w_j \log z_j + \sum_k \log \frac{\Gamma(a_k) \beta_k^{\alpha_k}}{\Gamma(\alpha_k) b_k^{a_k}} - \log \prod_j w_j! \quad (12)$$

where Eq.(9) for  $b_k$  and  $\sum_k n_{jk} = 1$  was used.

After initialisation the algorithm then repeats until convergence:

1. For each document: update  $n_{jk}$  and  $a_k$  according to Eq.(9) (variational E step).
2. Update  $\theta$  according to Eq.(10) or (11) (variational M step).
3. Compute lower bound on log-probability Eq.(12) and check for convergence.

## NMF

add correspondance to NMF...

## References

- [1] Buntine W, Jakulin A. Discrete component analysis. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, eds. Subspace, Latent Structure and Feature Selection. Springer; 2006:1â33. Available at: <http://www.springerlink.com/index/d53027666542q3v7.pdf> [Accessed January 26, 2011].

- [2] Canny, J. (2004). GaP: a factor model for discrete data. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (p. 122–129). ACM. Retrieved January 25, 2011, from <http://portal.acm.org/citation.cfm?id=1009016>.
- [3] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. (M. Jordan, J. Kleinberg, & B. Scholkopf, Eds.)Pattern Recognition (p. 738). Springer. doi: 10.1117/1.2819119.