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

Gamma-Poisson (GP) model [?]:

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

- w_j word count of jth word in a document

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

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

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

- θ component loading matrix of size $J \times K$. θ_{jk} controles partition of te kth component in the jth word

The log-likelihood of this model:

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

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 kth component. It is derived from a latent matrix $\mathbf{V}$ of size $J \times K$ (entries v_{jk}).

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

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:

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

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

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

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

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

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

The likelihood of GaP model with latent matrix V is then

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

explicitly:

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

and

$$\log p(V,l|\alpha,\beta,\theta,K)=\sum_k\left\{(c_k+\alpha_k-1)\log l_k-(\beta_k+1)l_k+\alpha_k\log\beta_k-\log\Gamma(\alpha_k)+\sum_j\left[v_{jk}\log\theta_{jk}-\log v_{jk}!\right]\right\} \tag{2}$$

w_j is derived from V so it is not represented... The term $l_k^{(c_k+\alpha_k-1)}=l_k^{(\sum_j v_{jk}+\alpha_k-1)}$ links together l_k and V and prevents simple evaluation of $\mathcal{Q}(\theta,\theta^{\mathrm{old}})=\mathbb{E}_{p(V,l|\theta^{\mathrm{old}})}\left[\log p(V,l|\theta,...)\right]$ in EM algorithm as because of the term $\mathbb{E}_{p(V,l|\theta^{\mathrm{old}})}\left[v_{jk}\right]$

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

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

and

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

so

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

Section 7.1 Variational Approximation

Factirised approximative 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]

$$\log q_l^*(l)=\mathbb{E}_{V\sim q_V}\left[\log p(V,l,w|\theta,\alpha,\beta)\right]+\mathrm{const} \tag{3}$$
$$\log q_V^*(V)=\mathbb{E}_{l\sim q_l}\left[\log p(V,l,w|\theta,\alpha,\beta)\right]+\mathrm{const} \tag{4}$$

The lower bound is given by [3]

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

(const is an entropy term independent on q, θ). So

$$\log p(w|\theta,\alpha,\beta,K)\geq\mathbb{E}_{l,V\sim q(l,V)}\left[\log p(l,V,w|\theta,\alpha,\beta,K)\right]+\mathrm{const} \tag{5}$$

The functional form of the complete likelihood suggests

$$q_l(l) = \prod_k \text{Gamma}(l_l; \alpha_k, \beta_k) = \prod_k \frac{l_k^{a_k-1} b_k^{a_k} \exp(-b_k l_k)}{\Gamma(a_k)} \tag{6}$$

$$q_V(V) = \prod_{jk} \text{Mutlinom}(v_{jk}; n_{jk}, w_j) = \prod_{jk} \frac{w_j!}{v_{jk}!} n_{jk}^{v_{jk}} \tag{7}$$

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

Then from Eq.(3), (6) and (2) keeping terms dependent on l

$$(a_k - 1)\log l_k - b_k l_k + \text{const} = (c_k + \alpha_k - 1)\log l_k - (\beta_k + 1) l_k + \text{const}$$

and form Eq.(4), (7) and (2) keeping terms dependent on V

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

so the rewrite rules for the parameters:

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

where z_{jk} is the normalisation constat ($\sum_k n_{jk} = 1$) so $z_{jk} = \sum_k \theta_{jk} \exp(\mathbb{E}_l \left[\log l_k\right])$ and $\sum_j n_{jk} w_j = c_k$. (n_{jk} is the proportion of the w_j it kth component). $\mathbb{E}_{l \sim q_l} \left[\log l_k\right] = \psi_0(a_k) - \log b_k$ where ψ_0 is digamma fucntion (logarithmic derivation of the gamma function...)

Now recompute model parameter $\theta$by maximizing lower bound Eq.(5) (keeping constraints $\sum_j \theta_{jk} = 1$). Keeping only term dependent on θ_{jk}:

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

(from Eq.(7) $\mathbb{E}_{q_V(V)} \left[v_{jk}\right] = w_j n_{jk}$)

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

we get

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

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

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

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

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

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

and by maximizing with normalization constraints:

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

References

[1] Buntine, W., & Jakulin, A. (2006). Discrete Componenet Analysis. In C. Saunders, M. Grobelnik, S. Gunn, & J. Shawe-Taylor (Eds.), Subspace, Latent Structure and Feature Selection (pp. 1-33). Springer.

[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.