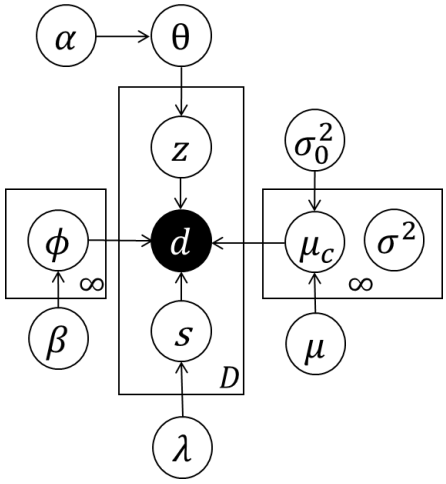

Figure 1: Graphical Representation of NMVM Model

Table 1: NOTATIONS

V	size of the vocabulary
D	number of documents in the corpus
L	average length of the documents
$\vec{d}$	documents in the corpus
$\vec{z}$	cluster assignment of the documents
z_d	cluster assignment of document d
e_d	embedding of document d
I	number of iterations
m_z	number of documents in cluster z
n_z	number of words in cluster z
e_z	sum of the embedding of documents in cluster z
n_z^w	number of occurrences of word w in cluster z
N_d	number of words in document d
N_d^w	number of occurrences of word w in document d

1 COLLAPSED GIBBS SAMPLING FOR NMVM

1.1 The Probability to Choose an Existed Cluster

In this part, we will give a detailed derivation of the probability to choose an existed cluster using collapsed Gibbs Sampling.

$$
\begin{aligned}
& p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta, \lambda, \sigma_0^2, \mu_0, \sigma^2) \\
=& p(z_d = z, s_d = 1|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta, \lambda, \sigma_0^2, \mu_0, \sigma^2) \\
& + p(z_d = z, s_d = 0|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta, \lambda, \sigma_0^2, \mu_0, \sigma^2) \qquad (1) \\
=& p(s_d = 1|\lambda) p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& + p(s_d = 0|\lambda) p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \sigma_0^2, \mu_0, \sigma^2) \qquad (2) \\
=& \lambda p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& + (1 - \lambda) p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \sigma_0^2, \mu_0, \sigma^2) \qquad (3)
\end{aligned}
$$

Equation (1) uses the Sum Rule of Probability and Equation (2) uses the property of D-Separation. Now, we derive the first term of Equation (3), the generation of bag-of-words part, as follows:

$$
\begin{aligned}
& p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& \propto p(z_d = z|\vec{z}_{\neg d}, \vec{d}_{\neg d}, \alpha, \beta) p(d|z_d = z, \vec{z}_{\neg d}, \vec{d}_{\neg d}, \alpha, \beta) && (4) \\
& = p(z_d = z|\vec{z}_{\neg d}, \alpha) p(d|z_d = z, \vec{d}_{z,\neg d}, \beta) && (5)
\end{aligned}
$$

Here, we use the Bayes Rule in Equation (4) and apply the property of D-Separation in Equation (5). The first term in Equation (5) is the probability of choosing cluster z for document d when the cluster assignments of the other documents have been known. The second term of Equation (5) indicates the predictive probability of document d when we know $\vec{d}_{z,\neg d}$, i.e., the other documents in cluster z currently.

Now, we will first derive the first term in Equation (5) as follows:

$$
\begin{aligned}
& p(z_d = z|\vec{z}_{\neg d}, \alpha) \\
& = \int p(z_d = z, \theta|\vec{z}_{\neg d}, \alpha) d\theta && (6) \\
& = \int p(\theta|\vec{z}_{\neg d}, \alpha) p(z_d = z|\vec{z}_{\neg d}, \theta, \alpha) d\theta && (7) \\
& = \int p(\theta|\vec{z}_{\neg d}, \alpha) p(z_d = z|\theta) d\theta && (8)
\end{aligned}
$$

Here, we use Sum Rule of Probability in Equation (6) and apply Product Rule of Probability in Equation (7). We finally employ the property of D-Separation in Equation (8). The first term of Equation (8) is the posterior of θ and the second term in Equation (8) is a *Categorical* distribution: $Mult(z_d = z|\theta) = \theta_z$

Now, the following is the derivation of the first term in Equation (8)

$$
\begin{aligned}
& p(\theta|\vec{z}_{\neg d}, \alpha) \\
& = \frac{p(\theta|\alpha) p(\vec{z}_{\neg d}|\theta)}{\int p(\theta|\alpha) p(\vec{z}_{\neg d}|\theta) d\theta} && (9) \\
& = \frac{\frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha/K-1} \prod_{k=1}^{K} \theta_k^{m_{k,\neg d}}}{\int \frac{1}{\Delta(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha/K-1} \prod_{k=1}^{K} \theta_k^{m_{k,\neg d}} d\theta} && (10) \\
& = \frac{1}{\Delta(\vec{m}_{\neg d} + \alpha/K)} \prod_{k=1}^{K} \theta_k^{m_{k,\neg d}+\alpha/K-1} && (11) \\
& = Dir(\theta|\Delta(\vec{m}_{\neg d} + \alpha/K)) && (12)
\end{aligned}
$$

Here, we use the Bayes Rule in Equation (9). Then, we can derive the first term in Equation (5) as follows:

$$
\begin{aligned}
& p(z_d = z|\vec{z}_{\neg d}, \alpha) \\
& = \int Dir(\theta|\vec{m}_{\neg d} + \alpha/K) Mult(z_d = z|\theta) d\theta && (13) \\
& = \int \frac{1}{\Delta(\vec{m}_{\neg d} + \alpha/K)} \theta_z \prod_{k=1}^{K} \theta_k^{m_{k,\neg d}+\alpha/K-1} d\theta && (14) \\
& = \frac{\Delta(\vec{m} + \alpha/K)}{\Delta(\vec{m}_{\neg(d)} + \alpha/K)} && (15) \\
& = \frac{\prod_{k=1}^{K} \Gamma(m_k + \alpha/K)}{\Gamma(\sum_{k=1}^{K}(m_k + \alpha/K))} \frac{\Gamma(\sum_{k=1}^{K}(m_{k,\neg d} + \alpha/K))}{\prod_{k=1}^{K} \Gamma(m_{k,\neg d} + \alpha/K)} && (16) \\
& = \frac{\Gamma(m_{z,\neg d} + \alpha/K + 1)}{\Gamma(m_{z,\neg d} + \alpha/K)} \frac{\Gamma(D - 1 + \alpha)}{\Gamma(D + \alpha)} && (17) \\
& = \frac{m_{z,\neg d} + \alpha/K}{D - 1 + \alpha} && (18)
\end{aligned}
$$

In Equation (15), we employ the Δ function, which is defined as $\Delta(\vec{\alpha}) = \frac{\prod_{k=1}^{K}}{\Gamma(\sum_{k=1}^{K} \alpha)}$. Based on the property of Γ function:$\Gamma(x + 1) = x\Gamma(x)$, we can obtain Equation (18) from Equation (17). In Equation (18), $m_{z,\neg d}$ denotes the number of documents in cluster z with document d

removed, and D is the total number of documents in the dataset. Equation (18) means that, the document d is more likely to choose the cluster with more documents when we only consider the cluster assignments of other documents.

Now, we will derive the second term in Equation (5) as follows:

$$
\begin{aligned}
& p(d|z_d = z, \vec{d}_{z,\neg d}, \beta) \\
&= \int p(d, \Phi_z | z_d = z, \vec{d}_{z,\neg d}, \beta) d\Phi_z && (19) \\
&= \int p(\Phi_z | z_d = z, \vec{d}_{z,\neg d}, \beta) p(d|\Phi_z, z_d = z, \vec{d}_{z,\neg d}, \beta) d\Phi_z && (20) \\
&= \int p(\Phi_z | \vec{d}_{z,\neg d}, \beta) p(d|\Phi_z, z_d = z) d\Phi_z && (21)
\end{aligned}
$$

Here, we use the Sum Rule of Probability in Equation (19) and exploits the Product Rule of Probability in Equation (20). In Equation (21), we apply the property of D-Separation.

Next, we will derive the first term of Equation (21) as follows:

$$
\begin{aligned}
& p(\Phi_z | \vec{d}_{z,\neg d}, \beta) \\
&= \frac{p(\Phi_z|\beta) p(\vec{d}_{z,\neg d}|\Phi_z)}{\int p(\Phi_z|\beta) p(\vec{d}_{z,\neg d}|\Phi_z) d\Phi_z} && (22) \\
&= \frac{\frac{1}{\Delta(\beta)} \prod_{t=1}^{V} \Phi_{z,t}^{\beta-1} \prod_{t=1}^{V} \Phi_{k,t}^{n^t_{z,\neg d}}}{\int \frac{1}{\Delta(\beta)} \prod_{t=1}^{V} \Phi_{z,t}^{\beta-1} \prod_{t=1}^{V} \Phi_{k,t}^{n^t_{z,\neg d}} d\Phi_z} && (23) \\
&= \frac{1}{\Delta(\vec{n}_z + \beta)} \prod_{t=1}^{V} \Phi_{z,t}^{n^t_{z,\neg d} + \beta - 1} && (24) \\
&= Dir(\Phi_z | \vec{n}_{z,\neg d} + \beta) && (25)
\end{aligned}
$$

Here, we use the Bayes Rule in Equation (22),

Next, we derive the second term of Equation (21) as follows:

$$
\begin{aligned}
& p(d|z_d = z, \vec{d}_{z,\neg d}, \beta) \\
&= \int Dir(\Phi_z | \vec{n}_{z,\neg d} + \beta) \prod_{w \in d} Mult(w|\Phi_z) d\Phi_z && (26) \\
&= \int \frac{1}{\Delta(\vec{n}_{z,\neg d})} \prod_{t=1}^{V} \Phi_{z,t}^{n^t_{z,\neg d} + \beta - 1} \prod_{w \in d} \Phi_{z,w}^{n^w_d} d\Phi_z && (27) \\
&= \frac{\Delta(\vec{n}_z + \beta)}{\Delta(\vec{n}_{z,\neg d} + \beta)} && (28) \\
&= \frac{\prod_{t=1}^{V} \Gamma(n^t_z + \beta)}{\Gamma(\sum_{t=1}^{V} (n^t_z + \beta))} \frac{\Gamma(\sum_{t=1}^{V} (n^t_{z,\neg d} + \beta))}{\prod_{t=1}^{V} \Gamma(n^t_{z,\neg d} + \beta)} && (29) \\
&= \frac{\prod_{w \in d} \prod_{j=1}^{N^w_d} (n^w_{z,\neg d} + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)} && (30)
\end{aligned}
$$

Here, we use the property of Γ function: $\frac{\Gamma(x+m)}{\Gamma(x)} = \prod_{i=1}^{m} (x + i - 1)$ when we obtain Equation (30) from Equation (29). In Equation (30), N^w_d and N_d denote the number of word w in document d and the total number of words in document d respectively, and $N_d = \sum_{w \in d} N^w_d$. In addition, $n^w_{z,\neg d}$ and $n_{z,\neg d}$ denotes the number of word w in cluster z and the total number of words in cluster z with document d removed, respectively, and $n_{z,\neg d} = \sum_{w=1}^{V} n^w_{z,\neg d}$. In fact, Equation (30) measures similarity between document d and cluster z to a certain extent. And, document d is more likely to choose a cluster whose documents share more words with it.

Eventually, we obtain the form of the first term of Equation (3) as follows:

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \propto \frac{m_{z,\neg d} + \alpha/K}{D - 1 + \alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)} \tag{31}$$

When $K \longrightarrow \infty$, the following is the final form of the first term of Equation (3)

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \propto \frac{m_{z,\neg d}}{D - 1 + \alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)} \tag{32}$$

Now, we will derive the second term of Equation (3), the generation of text embedding part, as follows:

$$\begin{aligned} & p(z_d = z|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \vec{d}, \alpha) \\ &= p(z_d = z|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \alpha, d, \vec{d}_{\neg d}) && (33) \\ &\propto p(z_d = z|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \alpha_0, \vec{d}_{\neg d}) \\ &\times p(d|z_d = z, \vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \alpha, \vec{d}_{\neg d}) && (34) \\ &= p(z_d = z|\vec{z}_{\neg d}, \alpha) p(d|\vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) && (35) \end{aligned}$$

Here, we use the Bayes Rule of Probability in Equation (34) and apply the property of D-Separation in Equation (35). We can notice that the first term of Equation (35) is as same as the first term of Equation (5), which we have derived already and the result is as follows:

$$p(z_d = z|\vec{z}_{\neg d}, \alpha) = \frac{m_{z,\neg d} + \alpha/K}{D - 1 + \alpha} \tag{36}$$

Now,we will derive the second term of Equation (35) as follows:

$$\begin{aligned} & p(d|\vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) \\ &= \int_{\mu_z} p(d, \mu_z|\vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) d\mu_z && (37) \\ &= \int_{\mu_z} p(\mu_z|\vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) p(d|\mu_z, \vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) d\mu_z && (38) \\ &= \int_{\mu_z} p(\mu_z|\vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) p(d|\mu_z, \sigma^2) d\mu_z && (39) \end{aligned}$$

Here, we use the Sum Rule of Probability in Equation (37) and apply Product Rule of Probability in Equation (38). Equation (39) exploits the property of D-Separation. The first term of Equation (39) denotes the posterior distribution of μ_z and the second term of Equation (39) is a gaussian distribution as follows:

$$p(d|\mu_z, \sigma^2) = \mathcal{N}(d|\mu_z, \sigma^2) \tag{40}$$

Now, we will derive the first term of Equation (39) as follows:

$$\begin{aligned} & p(\mu_z|\vec{d}_{z,\neg d}, \sigma^2, \sigma_0^2, \mu_0) \\ &= \frac{p(\mu_z|\sigma^2, \sigma_0^2, \mu_0) p(\vec{d}_{z,\neg d}|\mu_z, \sigma^2, \sigma_0^2, \mu_0)}{\int_{\mu_z} p(\mu_z|\sigma^2, \sigma_0^2, \mu_0) p(\vec{d}_{z,\neg d}|\mu_z, \sigma^2, \sigma_0^2, \mu_0) d\mu_z} && (41) \\ &= \frac{p(\mu_z|\sigma_0^2, \mu_0) p(\vec{d}_{z,\neg d}|\mu_z, \sigma^2)}{\int_{\mu_z} p(\mu_z|\sigma_0^2, \mu_0) p(\vec{d}_{z,\neg d}|\mu_z, \sigma^2) d\mu_z} && (42) \\ &= \mathcal{N}(\mu_z|\mu_n, \sigma_n^2) && (43) \end{aligned}$$

Here, we use the Bayes Rule in Equation (41) and Equation (43) use the conjugacy for Gaussian distribution as follows:

$$p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2) \tag{44}$$

$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \tag{45}$$

$$p(\mu|x_1, x_2, \ldots, x_n) = \mathcal{N}(\mu|\frac{\sigma_0^2 \sum_i x_i + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}) \tag{46}$$

Here, Equation (44) and Equation (45) is the prior distribution of μ and likelihood of x, respectively. Then, we can obtain the posterior distribution of μ as Equation (46). In our method, $\vec{d}_{z,\neg d}$, the documents in cluster z with document d removed, is $x_1, x_2, \ldots, x_n$ in Equation (46) and therefore, $m_{z,\neg d}$, the number of documents in cluster z without considering document d, is n in Equation (46). Given above comparison, we can obtain Equation (43) and μ_n, σ_n^2 are as follows:

$$\mu_n = \frac{1}{m_{z,\neg d}\sigma_0^2 + \sigma^2}(\sigma^2\mu_0 + \sigma_0^2 \sum_{i \in z, i \neq d} \vec{e_i}) \tag{47}$$

$$\sigma_n^2 = (\frac{1}{\sigma_0^2} + \frac{m_{z,\neg d}}{\sigma^2})^{-1} \tag{48}$$

Here, e_i is text embedding representation of document i. Specially, in order to connect σ and σ_0, we make a trick: $\sigma_0^2 = \sigma^2/\kappa$. Then, the above two equations become the following form:

$$\mu_n = \frac{\kappa\mu_0 + \sum_{i \in z, i \neq d} \vec{e_i}}{\kappa + m_{z,\neg d}} \tag{49}$$

$$\sigma_n^2 = \frac{\sigma^2}{\kappa + m_{z,\neg d}} \tag{50}$$

Then, using the property of gaussian distribution, we can can obtain the result of Equation (39) as follows:

$$\int \mathcal{N}(d|\mu_z, \sigma^2)\mathcal{N}(\mu_z|\mu_n, \sigma_n^2)d\mu_z = \mathcal{N}(d|\mu_n, \sigma_n^2 + \sigma^2) \tag{51}$$

Next, we get the probability choosing an existed cluster for text embedding part as follows:

$$\begin{aligned} &p(z_d = z|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \vec{d}, \alpha) \\ &\propto \frac{m_{z,\neg d} + \alpha/K}{D - 1 + \alpha}\mathcal{N}(d|\mu_n, \sigma_n^2 + \sigma^2) \end{aligned} \tag{52}$$

when $K \longrightarrow \infty$, we obtain the second term of (3) as follows:

$$\begin{aligned} &p(z_d = z|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \vec{d}, \alpha) \\ &\propto \frac{m_{z,\neg d}}{D - 1 + \alpha}\mathcal{N}(d|\mu_n, \sigma_n^2 + \sigma^2) \end{aligned} \tag{53}$$

given above deduction, we can obtain the probability choosing an existed cluster as follows:

$$\begin{aligned} &\lambda\frac{m_{z,\neg d}}{D - 1 + \alpha}\frac{\prod_{w \in d}\prod_{j=1}^{N_d^w}(n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d}(n_{z,\neg d} + V\beta + i - 1)} \\ &+(1 - \lambda)\frac{m_{z,\neg d}}{D - 1 + \alpha}\mathcal{N}(d|\mu_n, \sigma_n^2 + \sigma^2) \end{aligned} \tag{54}$$

1.2 the probability to choose a new cluster

In this part, we will give a detailed derivation of the probability to choose a new cluster using collapsed Gibbs Sampling.

$$
\begin{aligned}
& p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}, \alpha, \beta, \lambda, \sigma_0^2, \mu_0, \sigma^2) \\
=& p(z_d = K+1, s_d = 1 | \vec{z}_{\neg d}, \vec{d}, \alpha, \beta, \lambda, \sigma_0^2, \mu_0, \sigma^2) \\
& + p(z_d = K+1, s_d = 0 | \vec{z}_{\neg d}, \vec{d}, \alpha, \beta, \lambda, \sigma_0^2, \mu_0, \sigma^2) && (55) \\
=& p(s_d = 1|\lambda) p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& + p(s_d = 0|\lambda) p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}, \sigma_0^2, \mu_0, \sigma^2) && (56) \\
=& \lambda p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& + (1-\lambda) p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}, \sigma_0^2, \mu_0, \sigma^2) && (57)
\end{aligned}
$$

We use the Sum Rule of Probability in Equation (55) and apply the property of *Bernoulli* distribution in Equation (57).The two parts of Equation (57) are the probability choosing a new cluster for the BoW part and text embedding part, respectively. We will first derive the first term of (57) as follows:

$$
\begin{aligned}
& p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& \propto p(z_d = K+1 | \vec{z}_{\neg d}, \vec{d}_{\neg d}, \alpha, \beta) p(d | z_d = K+1, \vec{z}_{\neg d}, \vec{d}_{\neg d}, \alpha, \beta) && (58) \\
& = p(z_d = K+1 | \vec{z}_{\neg d}, \alpha) p(d | z_d = K+1, \beta) && (59)
\end{aligned}
$$

Here, we use the Bayes Rule in Equation (58) and apply the property of D-Separation in Equation (59).We can derive the second term of Equation (59) as follows:

$$
\begin{aligned}
& p(z_d = K+1 | \vec{z}_{\neg d}, \alpha) \\
& = 1 - \sum_{k=1}^{K} p(z_d = k | \vec{z}_{\neg d}, \alpha) && (60) \\
& = 1 - \frac{\sum_{k=1}^{K} m_{k,\neg d}}{D - 1 + \alpha} && (61) \\
& = 1 - \frac{D-1}{D-1+\alpha} && (62) \\
& = \frac{\alpha}{D-1+\alpha} && (63)
\end{aligned}
$$

Then, we will derive the first term of Equation (59) as follows:

$$
\begin{aligned}
& p(d|z_d = K+1, \beta) \\
&= \int p(d, \Phi_{K+1}|z_d = K+1, \beta) d\Phi_{K+1} && (64)\\
&= \int p(\Phi_{K+1}|z_d = K+1, \beta) p(d|\Phi_{K+1}, z_d = K+1, \beta) d\Phi_{K+1} && (65)\\
&= \int p(\Phi_{K+1}|\beta) p(d|\Phi_{K+1}, z_d = K+1) d\Phi_{K+1} && (66)\\
&= \int Dir(\Phi_{K+1}|\beta) \prod_{w \in d} Mult(w|\Phi_{K+1}) d\Phi_{K+1} && (67)\\
&= \int \frac{1}{\Delta(\beta)} \prod_{t=1}^{V} \Phi_{K+1,t}^{\beta-1} \prod_{w \in d} \Phi_{K+1,w}^{N_d^w} d\Phi_{K+1} && (68)\\
&= \frac{\Delta(\vec{n}_{K+1} + \beta)}{\Delta(\beta)} && (69)\\
&= \frac{\sum_{t=1}^{V} \Gamma(n_{K+1}^t + \beta)}{\Gamma(\sum_{t=1}^{V}(n_{K+1}^t + \beta))} \frac{\Gamma(\sum_{t=1}^{V}(\beta))}{\sum_{t=1}^{V} \Gamma(\beta)} && (70)\\
&= \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (\beta + j - 1)}{\prod_{i=1}^{N_d} (V\beta + i - 1)} && (71)
\end{aligned}
$$

Here, we use the Sum Rule in Equation (64) and apply the Product Rule of probability in Equation (65). Equation (67) exploits the property of D-Separation. Then, we can obtain Equation (71) from Equation (70) using the following property of Δ function: $\frac{\Delta(x+m)}{\Delta(x)} = \prod_{i=1}^{m}(x + i - 1)$. Then, we obtain the final form in Equation (57) as follows:

$$
\begin{aligned}
& p(z_d = K+1|\vec{z}_{\neg d}, \vec{d}, \alpha, \beta) \\
& \propto \frac{\alpha}{D - 1 + \alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (\beta + j - 1)}{\prod_{i=1}^{N_d} (V\beta + i - 1)} && (72)
\end{aligned}
$$

now, we, will derive the second term of Equation (57) as follows:

$$
\begin{aligned}
& p(z_d = K+1|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \vec{d}, \alpha) \\
&= p(z_d = K+1|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \alpha, d, \vec{d}_{\neg d}) && (73)\\
&\propto p(z_d = K+1|\vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \alpha_0, \vec{d}_{\neg d}) \\
&\times p(d|z_d = K+1, \vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \alpha, \vec{d}_{\neg d}) && (74)\\
&= p(z_d = K+1|\vec{z}_{\neg d}, \alpha) p(d|\vec{d}_{z=K+1,\neg d}, \sigma^2, \sigma_0^2, \mu_0) && (75)
\end{aligned}
$$

Here, we use the Bayes Rule in Equation (74) and apply the Property of D-Separation in Equation (75). We can notice that the first term of Equation (75) is as same as the first term in Equation (59) and therefore it equals $\frac{\alpha}{D-1+\alpha}$. Besides, there is no document in the new cluster, so $m_{z,\neg d} = 0, \sum_{i \in z, i \neq d} \vec{e}_i = \vec{0}$. Therefor, given the property of Gaussian distribution in Equation (46) and Equation (51), we can obtain the

second term in Equation (57) as follows:

$$\mu_n = \frac{1}{m_{z,\neg d}\sigma_0^2 + \sigma^2}(\sigma^2\mu_0 + \sigma_0^2 \sum_{i \in z, i \neq d} \vec{e}_i) \tag{76}$$

$$= \frac{1}{0\sigma_0^2 + \sigma^2}(\sigma^2\mu_0 + 0\sigma_0^2) \tag{77}$$

$$= \mu_0 \tag{78}$$

$$\sigma_n^2 = (\frac{1}{\sigma_0^2} + \frac{m_{z,\neg d}}{\sigma^2})^{-1}$$

$$= (\frac{1}{\sigma_0^2} + \frac{0}{\sigma^2})^{-1} \tag{79}$$

$$= \sigma_0^2 \tag{80}$$

the following is the final form of the second term of Equation (57)

$$p(z_d = K+1 | \vec{z}_{\neg d}, \sigma^2, \sigma_0^2, \mu_0, \vec{d}, \alpha)$$

$$\propto \frac{\alpha}{D-1+\alpha}\mathcal{N}(d|\mu_0, \sigma_0^2 + \sigma^2) \tag{81}$$

we finally obtain the probability choosing a new cluster for a document as follows

$$\lambda \frac{\alpha}{D-1+\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (\beta + j - 1)}{\prod_{i=1}^{N_d} (V\beta + i - 1)}$$

$$+(1-\lambda)\frac{\alpha}{D-1+\alpha}\mathcal{N}(d|\mu_0, \sigma_0^2 + \sigma^2) \tag{82}$$