Supplementary Material for Bayesian Nonparametric Multilevel Clustering with Contexts

Vu Nguyen[†], Dinh Phung[†], XuanLong Nguyen[‡], S. Venkatesh[†], and Hung Bui[*]
[†]*Centre for Pattern Recognition and Data Analytics (PRaDA),*
Deakin University, Australia.
{tvnguye,dinh.phung,svetha.venkatesh}@deakin.edu.au
[‡]*Department of Statistics, Dept of Electrical Engineering and Computer Science*
University of Michigan. xuanlong@umich.edu
[*]*Laboratory for Natural Language Understanding,*
Nuance Communications, Sunnyvale, CA 94085, USA. bui.h.hung@gmail.com

January 13, 2014

This note provides supplementary information for the main paper. It has three parts: a) the proof for the marginalization property of our proposed model, b) detailed derivations for our inference, and c) equations to show how the perplexity in the experiment was computed.

1 Proof for Marginalization Property (Theorem 4)

We start with a proposition on the marginalization result for DPM with the product measure then move on the final proof for our proposed model.

1.1 Marginalization of DPM with product measure

Let H be a measure over some measurable space (Θ, Σ). Let $\mathbb{P}$ be the set of all measures over (Θ, Σ), suitably endowed with some σ-algebra. Let $G \sim \mathrm{DP}(\alpha H)$ be a draw from a Dirichlet process.

Lemma 1. *Let $S_1 \dots S_n$ be n measurable sets in Σ. We form a measurable partition of Θ, a collection of disjoint measurable sets, that generate $S_1, \dots, S_n$ as follows. If S is a set, let $S^1 = S$ and $S^{-1} = \Theta \backslash S$. Then $S^* = \{\bigcap_{i=1}^n S_i^{c_i} | c_i \in \{1, -1\}\}$ is a partition of Θ into a finite collection of* disjoint *measurable sets with the property that any S_i can be written as a union of some sets in S^*. Let the element of S^* be $A_1 \dots A_{n^*}$ (note $n^* \leq 2^n$). Then the expectation*

$$\mathop{\mathbb{E}}_{G}\left[G\left(S_1\right), \dots, G\left(S_n\right)\right] = \int \prod_{i=1}^{n} G\left(S_i\right) DP\left(dG \mid \alpha H\right) \tag{1}$$

depends only on α and $H(A_i)$. In other words, the above expectation can be written as a function $E_n(\alpha, H(A_1), \dots H(A_{n^*}))$.

It is easy to see that since S_i can always be expressed as the sum of some disjoints A_i, $G\left(S_i\right)$ can respectively be written as the sum of some $G\left(A_i\right)$. Furthermore, by definition of a Dirichlet process, the vector $\left(G\left(A_1\right), \dots, G\left(A_{n^*}\right)\right)$ distributed according to a finite Dirichlet distribution $\left(\alpha H\left(A_1\right), \dots, \alpha H\left(A_{n^*}\right)\right)$, therefore the expectation $\mathop{\mathbb{E}}_{G}\left[G\left(S_i\right)\right]$ depends only on α and $H\left(A_i\right)$ (s).

Definition 2. (DPM) A DPM is a probability measure over $\Theta^n \ni (\theta_1, \dots, \theta_n)$ with the usual product sigma algebra Σ^n such that for every collection of measurable sets $\{(S_1, \dots, S_n) : S_i \in \Sigma, i = 1, \dots, n\}$:

$$\mathrm{DPM}(\theta_1 \in S_1, \dots, \theta_n \in S_n | \alpha, H) = \int_G \prod_{i=1}^{n} G\left(S_i\right) \mathrm{DP}\left(dG \mid \alpha H\right) \tag{2}$$

Consider two measurable spaces (Θ_1, Σ_1) and (Θ_2, Σ_2) and let (Θ, Σ) be their product space where $\Theta = \Theta_1 \times \Theta_2$ and $\Sigma = \Sigma_1 \times \Sigma_2$. We present the general theorem that states the marginal result from a product base measure.

Proposition 3. *Let H^* be a measure over the product space $\Theta = \Theta_1 \times \Theta_2$. Let H_1 be the marginal of H^* over Θ_1 in the sense that for any measurable set $A \in \Sigma_1$, $H_1(A) = H^*(A \times \Theta_2)$. Then:*

$$DPM\left(\theta_1^{(1)} \in S_1, \ldots, \theta_n^{(1)} \in S_n \mid \alpha, H_1\right) = DPM\left(\left(\theta_1^{(1)}, \theta_1^{(2)}\right) \in S_1 \times \Theta_2, \ldots, \left(\theta_n^{(1)}, \theta_n^{(2)}\right) \in S_n \times \Theta_2 \mid \alpha, H^*\right)$$

for every collection of measurable sets $\{(S_1, \ldots, S_n) : S_i \in \Sigma_1, i = 1, \ldots, n\}$.

Proof. Since $\{(S_1, \ldots, S_n) : S_i \in \Sigma_1, i = 1, \ldots, n\}$ are rectangles, expanding the RHS using Definition 2 gives:

$$RHS = \int G(S_1 \times \Theta_2) \ldots G(S_n \times \Theta_2)\, d\mathrm{DP}(dG|\alpha, H^*)$$

Let $T_i = S_i \times \Theta_2$, the above expression is the expectation of $\prod_i G(T_i)$ when $G \sim DP(\alpha H^*)$. Forming collection of the disjoint measurable sets $T^* = (B_1 \ldots B_{n^*})$ that generates T_i, then note that $B_i = A_i \times \Theta_2$, and $S^* = (A_1 \ldots A_{n^*})$ generates S_i. By definition of H_1, $H_1(A_i) = H^*(A_i \times \Theta_2) = H^*(B_i)$. Using the Lemma 1 above, $RHS = E_n(\alpha, H^*(B_1) \ldots H^*(B_{n^*}))$, while $LHS = E_n(\alpha, H_1(A_1) \ldots H_1(A_{n^*}))$ and they are indeed the same. □

We note that H^* can be any arbitrary measure on Θ and, in general, we do not require H^* to factorize as product measure.

1.2 Marginalization result for our proposed model

Recall that we are considering a product base-measure of the form $H^* = H \times \mathrm{DP}(vQ_0)$ for the group-level DP mixture. Drawing from a DP mixture with this base measure, each realization is a pair (θ_j, Q_j); θ_j is then used to generate the context x_j and Q_j is used to repeatedly generate the set of content observations w_{ji} within the group j. Specifically,

$$\begin{aligned} U &\sim \mathrm{DP}\left(\alpha(H \times \mathrm{DP}(vQ_0))\right) \quad \text{where } Q_0 \sim \mathrm{DP}(\eta S) \\ (\theta_j, Q_j) &\overset{\text{iid}}{\sim} U \text{ for } \ j = 1, \ldots, J \\ \varphi_{ji} &\overset{\text{iid}}{\sim} Q_j, \quad \text{for each } j \text{ and } i = 1, \ldots, N_j \end{aligned} \tag{3}$$

In the above, H and S are respectively base measures for context and content parameters θ_j and φ_{ji}. We start with a definition for nested Dirichlet Process Mixture (nDPM) to proceed further.

Definition 4. (nested DP mixture) A nDPM is a probability measure over $\Theta^{J \times \sum_{j=1}^{J} N_j}$ equipped with the usual product sigma algebra $\Sigma^{N_1} \times \ldots \times \Sigma^{N_J}$ such that for every collection of measurable sets $\{(S_{ji}) : S_{ji} \in \Sigma, j = 1, \ldots, J, i = 1 \ldots, N_j\}$:

$$\begin{aligned} &\mathrm{nDPM}(\varphi_{11} \in S_1^{(1)}, \ldots, \varphi_{1N_1} \in S_{n_1}^{(1)}, \ldots, \varphi_{J1} \in S_1^{(J)}, \ldots, \varphi_{JN_J} \in S_{N_J}^{(J)} | \alpha, v, \eta, S) \\ &= \int \left\{ \prod_{j=1}^{J} \int \prod_{i=1}^{N_j} Q_j\left(S_i^{(j)}\right) U(dQ_j) \right\} \mathrm{DP}\left(dU \mid \alpha \mathrm{DP}(vQ_0)\right) \mathrm{DP}\left(dQ_0 \mid \eta, S\right) \end{aligned}$$

We now state the main marginalization result for our proposed model.

Theorem 5. *Given α, H and α, v, η, S, let $\boldsymbol{\theta} = (\theta_j : \forall j)$ and $\boldsymbol{\varphi} = (\varphi_{ji} : \forall j, i)$ be generated as in Eq (3). Then, marginalizing out $\boldsymbol{\varphi}$ results in $DPM(\boldsymbol{\theta} \mid \alpha, H)$, whereas marginalizing out $\boldsymbol{\theta}$ results in $nDPM(\boldsymbol{\varphi}|\alpha, v, \eta, S)$.*

Proof. First we make observation that if we can show Proposition 3 still holds when H_1 is random with H_2 is fixed and vice versa, then the proof required is an immediate corollary of Proposition 3 by letting $H^* = H_1 \times H_2$ where we first let $H_1 = H$, $H_2 = \mathrm{DP}(vQ_0)$ to obtain the proof for the first result, and then swap the order $H_1 = \mathrm{DP}(vQ_0), H_2 = H$ to get the second result.

To see that Proposition 3 still holds when H_2 is a random measure and H_1 is fixed, we let the product base measure $H^* = H_1 \times H_2$ and further let μ be a prior probability measure for H_2, i.e, $H_2 \sim \mu(\cdot)$. Consider the marginalization over H_2:

$$\begin{aligned}
&\int_{H_2} \text{DPM}\left(\left(\theta_1^{(1)}, \theta_1^{(2)}\right) \in S_1 \times \Theta_2, \ldots, \left(\theta_n^{(1)}, \theta_n^{(2)}\right) \in S_n \times \Theta_2 \mid \alpha, H^*\right) \mu(H_2) \\
&\quad = \int_{\Sigma_2} \underbrace{\text{DPM}\left(\theta_1^{(1)} \in S_1, \ldots, \theta_n^{(1)} \in S_n \mid \alpha, H_1\right)}_{\text{constant w.r.t } H_2} \mu(H_2) \\
&\quad = \text{DPM}\left(\theta_1^{(1)} \in S_1, \ldots, \theta_n^{(1)} \in S_n \mid \alpha, H_1\right) \int_{\Sigma_2} \mu(H_2) \\
&\quad = \text{DPM}\left(\theta_1^{(1)} \in S_1, \ldots, \theta_n^{(1)} \in S_n \mid \alpha, H_1\right)
\end{aligned}$$

When H_1 is random and H_2 is fixed. Let $\lambda(\cdot)$ be a prior probability measure for H_1, ie., $H_1 \sim \lambda(\cdot)$. It is clear that Proposition 3 holds for each draw H_1 from $\lambda(\cdot)$. This complete our proof. □

1.3 Additional result for correlation analysis in nDPM

We now consider the correlation between φ_{ik} and $\varphi_{jk'}$ for arbitrary i, j, k and k', i.e., we need to evaluate:

$$P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid \alpha, \eta, v, S\right)$$

for two measurable sets $A_1, A_2 \in \Sigma$ by integrating out over all immediate random measures. We use an explicit stick-breaking representation for U where $U \sim \text{DP}(\alpha \text{DP}(vQ_0))$ as follows

$$U = \sum_{k=1}^{\infty} \pi_k \delta_{Q_k^*} \tag{4}$$

where $\pi \sim \text{GEM}(\alpha)$ and $Q_k^* \overset{\text{iid}}{\sim} \text{DP}(vQ_0)$. We use the notation $\delta_{Q_k^*}$ to denote the atomic measure on measure, placing its mass at measure Q_k^*.

For $i = j$, we have:

$$P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid Q_1, \ldots, Q_J\right) = Q_i(A_1) Q_i(A_2)$$

Sequentially take expectation over Q_i and U:

$$\begin{aligned}
\int_{Q_i} Q_i(A_1) Q_i(A_2) dU(Q_i) &= \int_{Q_i} Q_i(A_1) Q_i(A_2) d\left(\sum_{k=1}^{\infty} \pi_k \delta_{Q_k^*}\right) \\
&= \sum_k \pi_k \left[Q_k^*(A_1) Q_k^*(A_2)\right]
\end{aligned}$$

$$\begin{aligned}
\int_U \sum_{k=1}^{\infty} \pi_k \left[Q_k^*(A_1) Q_k^*(A_2)\right] d\text{DP}(U \mid \alpha \text{DP}(vQ_0)) &= \mathbb{E}\left\{\sum_k \pi_k \left[Q_k^*(A_1) Q_k^*(A_2)\right]\right\} \\
&= \sum_k \mathbb{E}[\pi_k] \mathbb{E}\left[Q_k^*(A_1) Q_k^*(A_2)\right] \\
&= \frac{Q_0(A_1 \cap A_2) + Q_0(A_1) Q_0(A_2)}{v(v+1)} \left(\sum_k \mathbb{E}[\pi_k]\right) \\
&= \frac{Q_0(A_1 \cap A_2) + Q_0(A_1) Q_0(A_2)}{v(v+1)}
\end{aligned}$$

Integrating out $Q_0 \sim \text{DP}(vS)$ we get:

$$\begin{aligned} P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid \alpha, v, \eta, S\right) &= \underset{Q_0|\eta,S}{\mathbb{E}}\left[\frac{Q_0\left(A_1 \cap A_2\right) + Q_0\left(A_1\right) Q_0\left(A_2\right)}{v\left(v+1\right)}\right] \\ &= \frac{1}{v\left(v+1\right)}\left\{S\left(A_1 \cap A_2\right) + \frac{S\left(A_1 \cap A_2\right) + S\left(A_1\right) S\left(A_2\right)}{\eta\left(\eta+1\right)}\right\} \\ &= \frac{S\left(A_1 \cap A_2\right)}{v\left(v+1\right)} + \frac{S\left(A_1 \cap A_2\right) + S\left(A_1\right) S\left(A_2\right)}{v\left(v+1\right)\eta\left(\eta+1\right)} \end{aligned}$$

For $i \neq j$, since Q_i and Q_j are conditionally independent given U, we get:

$$P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid Q_1, \ldots, Q_J\right) = Q_i\left(A_1\right) Q_j\left(A_2\right)$$

Let $a_k = Q_k^*\left(A_1\right), b_k = Q_k^*\left(A_2\right)$ and using Definition (4), integrating out U conditional on Q_0 with the stick-breaking representation in Eq (4):

$$\begin{aligned} P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid vQ_0\right) &= \left(\int_U Q_i\left(A_1\right) dU\right)\left(\int_U Q_j\left(A_2\right) dU\right) \\ &= \mathbb{E}\left[\sum_k \pi_k Q_k^*\left(A_1\right)\right]\left[\sum_{k'} \pi_{k'} Q_{k'}^*\left(A_2\right)\right] \\ &= \mathbb{E}\left(\pi_1 a_1 + \pi_2 a_2 + \ldots\right)\left(\pi_1 b_1 + \pi_2 b_2 + \ldots\right) \\ &= \mathbb{E}\left(\sum_k \pi_k^2 a_k b_k\right) + \mathbb{E}\left(\sum_{k \neq k'} \pi_k \pi_{k'} a_k b_{k'}\right) \\ &= A\mathbb{E}\left(\sum_k \pi_k^2\right) + B\mathbb{E}\left(\sum_{k \neq k'} \pi_k \pi_{k'}\right) \\ &= A\sum_k \mathbb{E}\left[\pi_k^2\right] + B\left(1 - \sum_k \mathbb{E}\left[\pi_k^2\right]\right) \end{aligned}$$

where

$$A = \mathbb{E}\left[a_k b_k\right] = \mathbb{E}\left[Q_k^*\left(A_1\right) Q_k^*\left(A_2\right)\right] = \frac{Q_0\left(A_1 \cap A_2\right) + Q_0\left(A_1\right) Q_0\left(A_2\right)}{v\left(v+1\right)}$$

and since Q_k^* (s) are iid draw from DP $\left(vQ_0\right)$ we have:

$$\begin{aligned} B &= \mathbb{E}\left[a_k b_{k'}\right] = \mathbb{E}\left[Q_k^*\left(A_1\right) Q_{k'}^*\left(A_2\right)\right] = \mathbb{E}\left[Q_k^*\left(A_1\right)\right]\mathbb{E}\left[Q_{k'}^*\left(A_2\right)\right] \\ &= Q_0\left(A_1\right) Q_0\left(A_2\right) \end{aligned}$$

Lastly, since $\left(\pi_1, \pi_2, \ldots\right) \sim \text{GEM}\left(\alpha\right)$, using the property of its stick-breaking representation $\sum_k \mathbb{E}\left[\pi_k^2\right] = \frac{1}{1+\alpha}$. Put things together we obtain the expression for the correlation of φ_{ik} and $\varphi_{jk'}$ for $i \neq j$ conditional on Q_0 as:

$$\begin{aligned} P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid vQ_0\right) &= \frac{Q_0\left(A_1 \cap A_2\right) + Q_0\left(A_1\right) Q_0\left(A_2\right)}{\left(1+\alpha\right) v\left(v+1\right)} + \frac{\alpha}{1+\alpha} Q_0\left(A_1\right) Q_0\left(A_2\right) \\ &= \frac{Q_0\left(A_1 \cap A_2\right)}{\left(1+\alpha\right) v\left(v+1\right)} + \frac{\alpha v\left(v+1\right) + 1}{\left(1+\alpha\right) v\left(v+1\right)} Q_0\left(A_1\right) Q_0\left(A_2\right) \end{aligned}$$

Next, integrating out $Q_0 \sim \text{DP}\left(vS\right)$ we get:

$$\begin{aligned} P\left(\varphi_{ik} \in A_1, \varphi_{jk'} \in A_2 \mid \alpha, v, \eta, S\right) =& \frac{\alpha v\left(v+1\right) + 1}{\left(1+\alpha\right) v\left(v+1\right)}\mathbb{E}\left[Q_0\left(A_1\right) Q_0\left(A_2\right)\right] + \frac{\mathbb{E}\left[Q_0\left(A_1 \cap A_2\right)\right]}{\left(1+\alpha\right) v\left(v+1\right)} \\ =& \frac{\alpha v\left(v+1\right) + 1}{\left(1+\alpha\right) v\left(v+1\right)} \frac{S\left(A_1 \cap A_2\right) + S\left(A_1\right) S\left(A_2\right)}{\eta\left(\eta+1\right)} + \frac{S\left(A_1 \cap A_2\right)}{\left(1+\alpha\right) v\left(v+1\right)} \end{aligned}$$

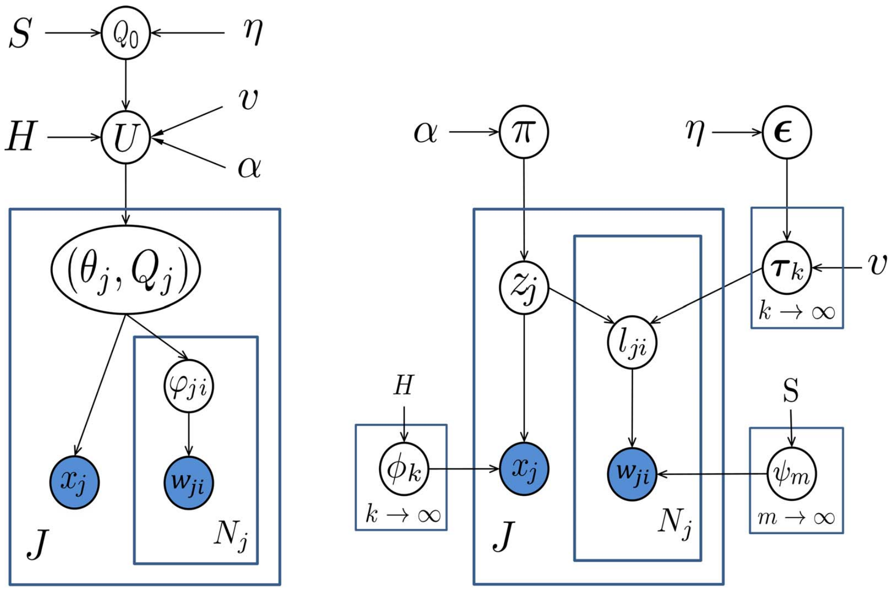


Figure 1: Generative (left) and stick-breaking (right) views of the proposed model.

2 Model Inference Derivations

We provide detailed derivations for model inference with the graphical model displayed in Fig 1. The variables ϕ_k, ψ_m, π, τ_k are integrated out due to conjugacy property. We need to sample these latent variables z, l, ϵ and hyper parameters α, v, η. For convenience of notation, we denote z_{-j} is a set of latent context variable z in all documents excluding document j, $\boldsymbol{l}_{j*}$ is all of hidden variables l_{ji} in document j, and $\boldsymbol{l}_{-j*}$ is all of l in other documents rather than document j-th.

Sampling z

Sampling context index z_j needs to take into account the influence of the corresponding context topics:

$$p(z_j = k \mid \mathbf{z}_{-j}, \boldsymbol{l}, \mathbf{x}, \alpha, H) \propto \underbrace{p\left(z_j = k \mid \mathbf{z}_{-j}, \alpha\right)}_{\text{CRP for context topic}} \underbrace{p\left(x_j \mid z_j = k, \mathbf{z}_{-j}, \mathbf{x}_{-j}, H\right)}_{\text{context predictive likelihood}} \tag{5}$$
$$\times \underbrace{p\left(\boldsymbol{l}_{j*} \mid z_j = k, \boldsymbol{l}_{-j*}, \mathbf{z}_{-j}, \epsilon, v\right)}_{\text{content latent marginal likelihood}}$$

The first term can easily be recognized as a form of Chinese Restaurant Process (CRP):

$$p\left(z_j = k \mid \mathbf{z}_{-j}, \alpha\right) = \begin{cases} \frac{n_{-j}^{k}}{n_{-j}^{*}+\alpha} & \text{if } k \text{is previously used} \\ \frac{\alpha}{n_{-j}^{*}+\alpha} & \text{if } k \text{is new} \end{cases}$$

where n_{-j}^{k} is the number of data $z_j = k$ excluding z_j, and n_{-j}^{*} is the count of all $\mathbf{z}$, except z_j.

The second expression is the predictive likelihood from the context observations under the context component ϕ_k. Specifically, let $f\left(\cdot \mid \phi\right)$ and $h\left(\cdot\right)$ be respectively the density function for $F\left(\phi\right)$ and H, the conjugacy between F and H allows us to integrate out the mixture component parameter ϕ_k , leaving us the conditional density of x_j under the mixture component k given all the context data items exclude x_j:

$$\begin{aligned} p\left(x_j \mid z_j = k, \mathbf{z}_{-j}, \mathbf{x}_{-j}, H\right) = & \frac{\int_{\phi_k} f\left(x_j \mid \phi_k\right) \prod_{j' \neq j, z_{j'} = k} f\left(x_{j'} \mid \phi_k\right) h\left(\phi_k\right) d\phi_k}{\int_{\phi_k} \prod_{j' \neq j, z_{j'} = k} f\left(x_{j'} \mid \phi_k\right) h\left(\phi_k\right) d\phi_k} \\ = & f_k^{-x_j}\left(x_j\right) \end{aligned}$$

Finally, the last term is the contribution from the multiple latent variables of corresponding topics to that context. Since $l_{ji} \mid z_j = k \overset{\text{iid}}{\sim} \text{Mult}\left(\boldsymbol{\tau}_k\right)$ where $\boldsymbol{\tau}_k \sim \text{Dir}\left(v\epsilon_1, \ldots, v\epsilon_M, \epsilon_{\text{new}}\right)$, we shall attempt to integrate out $\boldsymbol{\tau}_k$. Using the Multinomial-Dirichlet conjugacy property we proceed to compute the last term in Eq (5) as following:

$$p\left(\boldsymbol{l}_{j*} \mid z_j = k, \mathbf{z}_{-j}, \boldsymbol{l}_{-j*}, \epsilon, v\right) = \int_{\boldsymbol{\tau}_k} p\left(\boldsymbol{l}_{j*} \mid \boldsymbol{\tau}_k\right) \times p\left(\boldsymbol{\tau}_k \mid \left\{l_{j'*} \mid z_{j'} = k, j' \neq j\right\}, \epsilon, v\right) d\boldsymbol{\tau}_k \tag{6}$$

Recognizing the term $p\left(\boldsymbol{\tau}_k \mid \left\{\boldsymbol{l}_{j'*} \mid z_{j'} = k, j' \neq j\right\}, \boldsymbol{\epsilon}, v\right)$ is a posterior density, it is Dirichlet-distributed with the updated parameters

$$p\left(\boldsymbol{\tau}_k \mid \left\{\boldsymbol{l}_{j'*} \mid z_{j'} = k, j' \neq j\right\}\right) = \text{Dir}\left(v\epsilon_1 + c_{k,1}^{-j}, \ldots, v\epsilon_M + c_{k,M}^{-j}, v\epsilon_{\text{new}}\right) \tag{7}$$

where $c_{k,m}^{-j} = \sum_{j' \neq j} \sum_{i=1}^{N_{j'}} \mathbb{I}\left(l_{j'i} = m, z_{j'} = k\right)$ is the count of topic m being assigned to context k excluding document j. Using this result, $p\left(\boldsymbol{l}_{j*} \mid \boldsymbol{\tau}_k\right)$ is a predictive likelihood for $\boldsymbol{l}_{j*}$ under the posterior Dirichlet parameters $\boldsymbol{\tau}_k$ in Eq 7 and therefore can be evaluated to be:

$$\begin{aligned}
p\left(\boldsymbol{l}_{j*} \mid z_j = k, \mathbf{z}_{-j}, \boldsymbol{l}_{-j*}, \epsilon, v\right) &= \int_{\tau_k} p\left(\boldsymbol{l}_{j*} \mid \boldsymbol{\tau}_k\right) \times \text{Dir}\left(v\epsilon_1 + c_{k,1}^{-j}, \ldots, v\epsilon_M + c_{k,M}^{-j}, v\epsilon_{\text{new}}\right) d\boldsymbol{\tau}_k \\
&= \int_{\tau_k} \prod_{m=1}^{M} \tau_{k,m}^{c_{k,m}^{j}} \times \frac{\Gamma\left(\sum_{m=1}^{M}\left(v\epsilon_m + c_{k,m}^{-j}\right)\right)}{\prod_{m=1}^{M}\Gamma\left(v\epsilon_m + c_{k,m}^{-j}\right)} \times \prod_{m=1}^{M} \tau_{k,m}^{v\epsilon_m + c_{k,m}^{-j} - 1} d\boldsymbol{\tau}_k \\
&= \frac{\Gamma\left(\sum_{m=1}^{M}\left(v\epsilon_m + c_{k,m}^{-j}\right)\right)}{\prod_{m=1}^{M}\Gamma\left(v\epsilon_m + c_{k,m}^{-j}\right)} \times \int_{\tau_k} \prod_{m=1}^{M} \tau_{k,m}^{v\epsilon_m + c_{k,m}^{-j} + c_{k,m}^{j} - 1} d\boldsymbol{\tau}_k \\
&= \frac{\Gamma\left(\sum_{m=1}^{M}\left(v\epsilon_m + c_{k,m}^{-j}\right)\right)}{\prod_{m=1}^{M}\Gamma\left(v\epsilon_m + c_{k,m}^{-j}\right)} \times \frac{\prod_{m=1}^{M}\Gamma\left(v\epsilon_m + c_{k,m}^{-j} + c_{k,m}^{j}\right)}{\Gamma\left(\sum_{m=1}^{M}\left(v\epsilon_m + c_{k,m}^{-j} + c_{k,m}^{j}\right)\right)} \\
&= \frac{\Gamma\left(\sum_{m=1}^{M}\left(v\epsilon_m + c_{k,m}^{-j}\right)\right)}{\Gamma\left(\sum_{m=1}^{M}\left(v\epsilon_m + c_{k,m}^{-j}\right) + N_j\right)} \times \prod_{m=1}^{M} \frac{\Gamma\left(v\epsilon_m + c_{k,m}^{-j} + c_{k,m}^{j}\right)}{\Gamma\left(v\epsilon_m + c_{k,m}^{-j}\right)} \\
&= \begin{cases} A = \frac{\Gamma\left(\sum_m\left[v\epsilon_m + c_{k,m}^{-j}\right]\right)}{\Gamma\left(\sum_m\left[v\epsilon_m + c_{k,m}\right]\right)} \prod_m \frac{\Gamma\left(v\epsilon_m + c_{k,m}\right)}{\Gamma\left(v\epsilon_m + c_{k,m}^{-j}\right)} & \text{if } k \text{ previously used} \\ B = \frac{\Gamma\left(\sum_m v\epsilon_m\right)}{\Gamma\left(\sum_m v\epsilon_m + N_j\right)} \prod_m \frac{\Gamma\left(v\epsilon_m + c_{k,m}^{j}\right)}{\Gamma\left(v\epsilon_m\right)} & \text{if } k = k_{\text{new}} \end{cases}
\end{aligned}$$

note that $\epsilon = \left(\epsilon_1, \epsilon_2, \ldots \epsilon_M, \epsilon_{\text{new}}\right)$, here $\epsilon_{1:M} = \left(\epsilon_1, \epsilon_2, \ldots \epsilon_M\right)$, when sampling z_j we only use M active components from the previous iteration. In summary, the conditional distribution to sample z_j is given as:

$$p\left(z_j = k \mid \mathbf{z}_{-j}, \boldsymbol{l}, \mathbf{x}, \alpha, H\right) \propto \begin{cases} n_{-j}^{k} \times f_k^{-x_j}\left(x_j\right) \times A & \text{if } k \text{ previousely used} \\ \alpha \times f_{k_{\text{new}}}^{-x_{ji}}\left(x_{ji}\right) \times B & \text{if } k = k_{\text{new}} \end{cases}$$

Implementation note: to evaluate A and B, we make use of the marginal likelihood resulted from a Multinomial-Dirichlet conjugacy.

Sampling l

Let w_{-ji} be the same set as w excluding w_{ji}, i.e $w_{-ji} = \left\{w_{uv} : u \neq j \cap v \neq i\right\}$, then we can write

$$p\left(l_{ji} = m \mid l_{-ji}, z_j = k, v, w, S\right) \propto \underbrace{p\left(w_{ji} \mid w_{-ji}, l_{ji} = m, \rho\right)}_{\text{content predictive likelihood}} \times \underbrace{p\left(l_{ji} = m \mid \boldsymbol{l}_{-ji}, z_j = k, \epsilon_m, v\right)}_{\text{CRF for content topic}} \tag{8}$$

The first argument is computed as log likelihood predictive of the content with the component ψ_m

$$\begin{aligned} p\left(w_{ji} \mid w_{-ji}, l_{ji}=m, \rho\right) =& \frac{\int_{\lambda_m} s\left(w_{ji} \mid \lambda_m\right)\left[\prod_{u \in w_{-ji}(m)} y(u \mid \lambda_m)\right] s(\lambda_m) d\lambda_m}{\int_{\lambda_m}\left[\prod_{u \in w_{-ji}(m)} y\left(u \mid \lambda_m\right)\right] s\left(\lambda_m\right) d\lambda_m} \\ \triangleq & y_m^{-w_{ji}}\left(w_{ji}\right) \end{aligned} \tag{9}$$

And the second term is inspired by Chinese Restaurant Franchise (CRF) as:

$$p\left(l_{ji}=m \mid \boldsymbol{l}_{-ji}, \epsilon_m, v\right)=\begin{cases} c_{k,m}+v\epsilon_m & \text{if } m \text{is used previously} \\ v\epsilon_{\text{new}} & \text{if } m=m_{\text{new}} \end{cases} \tag{10}$$

where $c_{k,m}$ is the number of data point $|\{l_{ji}|l_{ji}=m, z_j=k, 1\leq j\leq J, 1\leq i\leq N_j\}|$. The final form to sample l_{ji} is given as:

$$p\left(l_{ji}=m \mid \boldsymbol{l}_{-ji}, z_j=k, w, v, \epsilon\right) \propto \begin{cases} \left(c_{k,m}+v\epsilon_m\right)\times y_m^{-w_{ji}}\left(w_{ji}\right) & \text{if } m \text{is used previously} \\ v\epsilon_{\text{new}}\times y_m^{-w_{ji}}(w_{ji}) & \text{if } m=m_{\text{new}} \end{cases}$$

Sampling ϵ

Note that sampling ϵ require both z and l.

$$p\left(\epsilon \mid \boldsymbol{l}, \mathbf{z}, v, \eta\right) \propto p\left(\boldsymbol{l} \mid \epsilon, v, z, \eta\right) \times p\left(\epsilon \mid \eta\right) \tag{11}$$

Isolating the content variables l_{ji}^k generated by the same context $z_j=k$ into one group $l_j^k=\{l_{ji}: 1\leq i\leq N_j, z_j=k\}$ the first term of 11 can be expressed following:

$$\begin{aligned} p\left(l \mid \epsilon, v, z, \eta\right) =& \prod_{k=1}^{K}\int_{\tau_k} p\left(l_{**}^k \mid \tau_k\right) p\left(\tau_k \mid \epsilon\right) d\tau_k \\ =& \prod_{k=1}^{K}\frac{\Gamma(v)}{\Gamma\left(v+n_{k*}\right)}\prod_{m=1}^{M}\frac{\Gamma(v\epsilon_m+n_{km})}{\Gamma(v\epsilon_m)} \end{aligned}$$

where $n_{k*}=|\{w_{ji} \mid z_j=k, i=1,...N_j\}|$ and $n_{km}=|\{w_{ji} \mid z_j=k, l_{ji}=m, 1\leq j\leq J, 1\leq i\leq N_j,\}|$.

Let $\eta_r=\frac{\eta}{R}$, $\eta_{\text{new}}=\frac{R-M}{R}\eta$ and recall that $\epsilon\sim\text{Dir}\left(\eta_r,\ldots,\eta_r,\eta_{\text{new}}\right)$, the last term of Eq 11 is a Dirichlet density:

$$\begin{aligned} p\left(\boldsymbol{\epsilon} \mid \eta\right) =& \text{Dir}\left(\underbrace{\eta_1,\eta_2,...\eta_M}_{M},\eta_{\text{new}}\right) \\ =& \frac{\Gamma(M\times\eta_r+\eta_{\text{new}})}{[\Gamma(\eta_r)]^M\eta_{\text{new}}}\prod_{m=1}^{M}\epsilon_m^{\eta_r-1}\epsilon_{\text{new}}^{\eta_{\text{new}}-1} \end{aligned}$$

Using the result:

$$\frac{\Gamma(v\epsilon_m+n_{km})}{\Gamma(v\epsilon_m)}=\sum_{o_{km}=0}^{n_{km}}\text{Stirl}\left(o_{km},n_{km}\right)\left(v\epsilon_m\right)^{o_{km}}$$

Thus, Eq 11 becomes:

$$
\begin{aligned}
p\left(\boldsymbol{\epsilon} \mid \boldsymbol{l}, \mathbf{z}, v, \eta\right) =& \epsilon_{\text{new}}^{\eta_{\text{new}}-1} \prod_{k=1}^{K} \frac{\Gamma(v)}{\Gamma\left(v+n_{k*}\right)} \prod_{m=1}^{M} \epsilon_m^{\eta_m-1} \sum_{o_{km}=0}^{n_{km}} \text{Stirl}\left(o_{km}, n_{km}\right)\left(v\epsilon_m\right)^{o_{km}} \\
=& \epsilon_{\text{new}}^{\eta_{\text{new}}-1} \sum_{o_{km}=0}^{n_{km}} \prod_{k=1}^{K} \frac{\Gamma(v)}{\Gamma\left(v+n_{k*}\right)} \prod_{m=1}^{M} \epsilon_m^{\eta_m-1} \text{Stirl}\left(o_{km}, n_{km}\right)\left(v\epsilon_m\right)^{o_{km}} \\
p\left(\boldsymbol{\epsilon}, \boldsymbol{o} \mid \boldsymbol{l}, \mathbf{z}, v, \eta\right) =& \epsilon_{\text{new}}^{\eta_{\text{new}}-1} \prod_{k=1}^{K} \frac{\Gamma(v)}{\Gamma\left(v+n_{k*}\right)} \prod_{m=1}^{M} \epsilon_m^{\eta_m-1} \text{Stirl}\left(o_{km}, n_{km}\right)\left(v\epsilon_m\right)^{o_{km}}
\end{aligned}
$$

The probability of the auxiliary variable o_{km} is computed as:

$$
p(o_{km}) = \sum_{o_{km}=0}^{n_{km}} \text{Stirl}\left(o_{km}, n_{km}\right)\left(v\epsilon_m\right)^{o_{km}}
$$

Now let $o = (o_{km} : \forall k, m)$ we derive the following joint distribution:

$$
p\left(\boldsymbol{\epsilon} \mid o, \boldsymbol{l}, \mathbf{z}, v, \eta\right) = \epsilon_{\text{new}}^{\eta_{\text{new}}-1} \prod_{m=1}^{M} \epsilon_m^{\sum_K o_{km} + \eta_m - 1}
$$

As $R \to \infty$, we have

$$
p\left(\boldsymbol{\epsilon} \mid \boldsymbol{o}, \boldsymbol{l}, \mathbf{z}, v, \eta\right) \overset{\infty}{\cong} \epsilon_{\text{new}}^{\eta-1} \prod_{m=1}^{M} \epsilon_m^{\sum_K o_{km} - 1}
$$

Finally, we sample ϵ jointly with the auxiliary variable o_{km} by:

$$
p\left(o_{km} = h \mid \cdot\right) \propto \textbf{Stirl}\left(h, n_{km}\right)\left(v\epsilon_m\right)^h, \; h = 0, 1, \ldots, n_{km}
$$

$$
p(\epsilon) \propto \epsilon_{\textbf{new}}^{\eta-1} \prod_{m=1}^{M} \epsilon_m^{\sum_K o_{km} - 1}
$$

Sampling hyperparameters

In the proposed model, there are three hyper-parameters which need to be sampled : α, v and η.

Sampling η

Using similar strategy and using technique from Escobar and West [3], we have

$$
p\left(M \mid \eta, u\right) = \text{Stirl}\left(M, u\right) \eta^M \frac{\Gamma\left(\eta\right)}{\Gamma\left(\eta + u\right)}
$$

where $u = \sum_m u_m$ with $u_m = \sum_K o_{km}$ is in the previous sampling ϵ and M is the number of active content atoms. Let $\eta \sim \text{Gamma}\left(\eta_1, \eta_2\right)$. Recall that:

$$
\frac{\Gamma\left(\eta\right)}{\Gamma\left(\eta + u\right)} = \int_0^1 t^{\eta}\left(1 - t\right)^{u-1}\left(1 + \frac{u}{\eta}\right) dt
$$

that we have just introduced an auxiliary variable t

$$p\left(t \mid \eta\right) \propto t^{\eta}\left(1-t\right)^{u-1} = \text{Beta}\left(\eta+1, u\right)$$

Therefore,

$$\begin{aligned}
p\left(\eta \mid t\right) &\propto \eta^{\eta_1-1+M} \exp\left\{-\eta\eta_2\right\} \times t^{\eta}\left(1-t\right)^{u-1}\left(1+\frac{u}{\eta}\right) \\
&= \eta^{\eta_1-1+M} \times \exp\left\{-\eta(\eta_2-\log t)\right\} \times \left(1-t\right)^{u-1} + \eta^{\eta_1-1+M-1} \exp\left\{-\eta(\eta_2-\log t)\right\} \times \left(1-t\right)^{u-1} u \\
&\propto \eta^{\eta_1-1+M} \exp\left\{-\eta(\eta_2-\log t)\right\} + u\eta^{\eta_1-1+M-1} \exp\left\{-\eta(\eta_2-\log t)\right\} \\
&= \pi_t \text{Gamma}\left(\eta_1+M, \eta_2-\log t\right) + \left(1-\pi_t\right)\text{Gamma}\left(\eta_1+M-1, \eta_2-\log t\right) \qquad (12)
\end{aligned}$$

where π_t satisfies this following equation to make the above expression a proper mixture density:

$$\frac{\pi_t}{1-\pi_t} = \frac{\eta_1+M-1}{u\left(\eta_2-\log t\right)} \qquad (13)$$

To re-sample η, we first sample $t \sim \text{Beta}\left(\eta+1, u\right)$, compute π_t as in equation 13, and then use π_t to select the correct Gamma distribution to sample η as in Eq. 12.

Sampling α

Again sampling α is similar to Escobar et al [3]. Assuming $\alpha \sim \text{Gamma}\left(\alpha_1, \alpha_2\right)$ with the auxiliary variable t:

$$\begin{aligned}
p\left(t \mid \alpha, K\right) \propto& t^{\alpha_1}\left(1-t\right)^{J-1} \\
p\left(t \mid \alpha, K\right) \propto& \text{Beta}\left(\alpha_1+1, J\right)
\end{aligned}$$

J: number of document

$$p\left(\eta \mid t, K\right) \sim \pi_t \text{Gamma}\left(\alpha_1+K, \alpha_2-\log(t)\right) + (1-\pi_t)\text{Gamma}\left(\alpha_1+K-1, \alpha_2-\log(t)\right)$$

where c, d are prior parameter for sampling η following Gamma distribution and $\frac{\pi_t}{1-\pi_t} = \frac{\alpha_1+K-1}{J(\alpha_2-\log t)}$

Sampling v

Sampling v is similar to sampling concentration parameter in HDP [6]. Denote $o_{k*} = \sum_m o_{km}$, where o_{km} is defined previously during the sampling step for $\boldsymbol{\epsilon}$, $n_{k*} = \sum_m n_{km}$, where n_{km} is the count of $\left|\left\{l_{ji} \mid z_{ji} = k, l_{ji} = m\right\}\right|$. Using similar technique in [6], we write:

$$p\left(o_{1*}, o_{2*}.., o_{K*} \mid v, n_{1*}, ...n_{K*}\right) = \prod_{k=1}^{K} \text{Stirl}(n_{k*}, o_{k*})\alpha_0^{o_{k*}} \frac{\Gamma(v)}{\Gamma\left(v+n_{k*}\right)}$$

where the last term can be expressed as

$$\frac{\Gamma(v)}{\Gamma\left(v+n_{k*}\right)} = \frac{1}{\Gamma(n_{k*})} \int_0^1 b_k^v \left(1-b_k\right)^{n_{k*}-1}\left(1+\frac{n_{k*}}{v}\right) db_k$$

Assuming $v \sim \text{Gamma}\left(v_1, v_2\right)$, define the auxiliary variables $b = \left(b_k \mid k = 1, \ldots, K\right), b_k \in [0,1]$ and $t = \left(t_k \mid k = 1, \ldots, K\right), t_k \in \{0,1\}$ we have

$$q\left(v, b, t\right) \propto v^{v_1-1+\sum_k M_k} \exp\left\{-vv_1\right\} \prod_{k=1}^{K} b_k^v \left(1-b_k\right)^{M_k-1}\left(\frac{M_k}{v}\right)^{t_k}$$

We will sample the auxiliary variables b_k, t_k in accordance with v that are defined below:

$$\begin{aligned} q(b_k \mid v) =& \mathrm{Beta}\left(v+1, o_{k*}\right) \\ q\left(t_k \mid .\right) =& \mathrm{Bernoulli}\left(\frac{o_{k*}/v}{1+o_{k*}/v}\right) \\ q(v \mid .) =& \mathrm{Gamma}\left(v_1 + \sum_k \left(o_{k*} - t_k\right), v_2 - \sum_k \log b_k\right) \end{aligned}$$

3 Relative Roles of Context and Content Data

Regarding the inference of the cluster index z_j (Eq. 5), to obtain the marginal likelihood (the third term in Eq. 5) one has to integrate out the words' topic labels l_{ji}. In doing so, it can be shown that the *sufficient* statistics coming from the content data toward the inference of the topic frequencies and the clustering labels will just be the empirical word frequency from each document. As each document becomes sufficiently long, the empirical word frequency quickly concentrates around its mean by the central limit theorem (CLT), so as soon as the effect of CLT kicks in, increasing document length further will do very little in improving this sufficient statistics.

Increasing the document length will probably not hurt, of course. But to what extent it contributes relative to the number of documents awaits a longer and richer story to be told.

We confirm this argument by varying the document length and the number of documents in the synthetic document and see how they affect the *posterior* of the clustering labels. Each experiment is repeated 20 times. We record the mean and standard deviation of the clustering performance by NMI score. As can be seen from Fig 2, using context observation makes the model more robust in recovering the true document clusters.

4 Perplexity Evaluation

The standard perplexity proposed by Blei et al [2], used to evaluate the proposed model as following:

$$\text{perplexity}\left(w^{\text{Test}}\right) = \exp\left\{-\frac{\sum_{j=1}^{J_{\text{Test}}} \log p\left(w_j^{\text{Test}}\right)}{\sum_{j=1}^{J_{\text{Test}}} N_j^{\text{Test}}}\right\}$$

During individual sampling iteration t, we utilize the important sampling approach [5] to compute $p\left(w_{\text{Test}}\right)$. The posterior estimation of ψ_m in a Multinomial-Dirichlet case is defined below, note that it can be in other types of conjugacies [4] (e.g. Gaussian-Wishart, Binomial-Poisson):

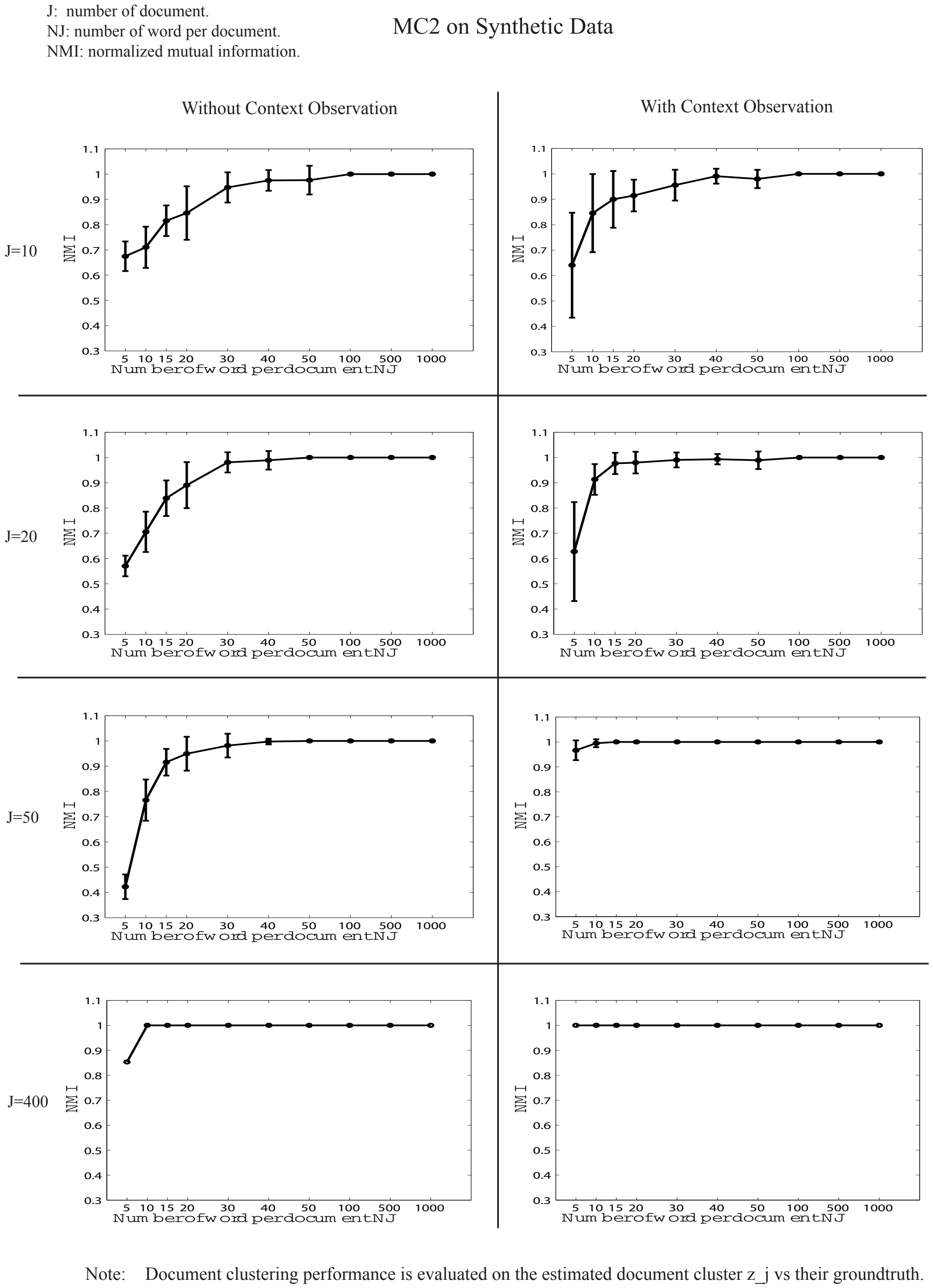
$$\psi_{m,v}^t = \frac{n_{m,v}^t + \text{smooth}}{\sum_{u=1}^V n_{m,v}^t + V \times \text{smooth}}$$

$$\tau_{k,m}^t = \frac{c_{k,m} + vv \times \epsilon_m}{\sum_{m=1}^M \left(c_{k,m} + vv \times \epsilon_m\right)}$$

where $n_{m,v}^t$ is number of times a word v, $v \in \{1, ..., V\}$ is assigned to context topic ψ_m in iteration t, and $c_{k,m}$ is the count of the set $\{w_{ji} \mid z_j = k, l_{ji} = m, 0 \leq j \leq J, 0 \leq i \leq N_j\}$. There is a constant smooth parameter [1] that influence on the count, roughly set as 0.1. Supposed that we estimate $z_j^{\text{Test}} = k$ and $l_{ji}^{\text{Test}} = m$, then the probability $p\left(w_j^{\text{Test}}\right)$ is computed as:

$$p\left(w_j^{\text{Test}}\right) = \prod_{i=1}^{N_j^{\text{Test}}} \frac{1}{T} \sum_{t=1}^T \tau_{k,m}^t \psi_{m,w_{ji}^{\text{Test}}}^t$$

where T is the number of collected Gibbs samples.



J: number of document.
NJ: number of word per document.
NMI: normalized mutual information.

MC2 on Synthetic Data

Note: Document clustering performance is evaluated on the estimated document cluster z_j vs their groundtruth.

Figure 2: Document clustering performance with different numbers of observed words and documents.

References

[1] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430), 1995.

[4] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2003.

[5] Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2006.

[6] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.