Supplementary Material
Nonparametric Bayesian Topic Modelling with the Hierarchical Pitman-Yor Processes

Kar Wai Lim[a,b,*], Wray Buntine[c], Changyou Chen[d], Lan Du[c]

[a] *The Australian National University, Canberra ACT 0200, Australia*
[b] *Data61/NICTA, Locked Bag 8001, Canberra ACT 2601, Australia*
[c] *Monash University, Faculty of IT, Wellington Road, Clayton VIC 3800, Australia*
[d] *Duke University, Box 90291, Durham, NC 27708, United States*

Abstract

This supplementary material looks at performing Bayesian inference on the Twitter-Network Topic Model (TNTM) presented in the main article. In the TNTM, combining a GP with a HPYP makes its posterior inference non-trivial. Hence, we employ approximate inference by alternatively performing MCMC sampling on the HPYP topic model and the network model, conditioned on each other. For the HPYP topic model, we employ the flexible framework discussed in Section 3 to perform collapsed blocked Gibbs sampling. For the network model, we derive a Metropolis-Hastings (MH) algorithm based on the elliptical slice sampler (Murray et al., 2010). In addition, the author–topic distributions ν connecting the HPYP and the GP are sampled with an MH scheme since their posteriors do not follow a standard form. We note that the PYPs here can have multiple parents, so we extend the framework in Section 3 to allow for this.

Appendix A. Posterior Inference for TNTM

Appendix A.1. Decrementing the Counts Associated with a Word or Hashtag

When we remove a word or a hashtag during inference, we decrement by one the customer count from the PYP associated with the word or the hashtag, that is, $c_k^{\theta_d}$ for word w_{dn} ($z_{dn} = k$) and $c_k^{\theta'_d}$ for hashtag y_{dm} ($z'_{dm} = k$). Decrementing the customer count may or may not decrement the respective

[*] Corresponding author

Email addresses: karwai.lim@anu.edu.au (Kar Wai Lim), wray.buntine@monash.edu (Wray Buntine), cchangyou@gmail.com (Changyou Chen), lan.du@monash.edu (Lan Du)

table count. However, if the table count is decremented, then we would decrement the customer count of the parent PYP. This is relatively straight forward in Section 4.1 since the PYPs have only one parent. Here, when a PYP $\mathcal{N}$ has multiple parents, we would sample for one of its parent PYPs and decrement the table count corresponding to the parent PYP. Although not the same, the rationale of this procedure follows Section 4.1.

We explain in more details below. When the customer count $c_k^{\mathcal{N}}$ is decremented, we introduce an auxiliary variable $u_k^{\mathcal{N}}$ that indicates which parent of $\mathcal{N}$ to remove a table from, or none at all. The sample space for $u_k^{\mathcal{N}}$ is the P parent nodes $\mathcal{P}_1, \dots, \mathcal{P}_P$ of $\mathcal{N}$, plus $\emptyset$. When $u_k^{\mathcal{N}}$ is equal to $\mathcal{P}_i$, we decrement the table count $t_k^{\mathcal{N} \to \mathcal{P}_i}$ and subsequently decrement the customer count $c_k^{\mathcal{P}_i}$ in node $\mathcal{P}_i$. If $u_k^{\mathcal{N}}$ equals to $\emptyset$, we do not decrement any table count. The process is repeated recursively as long as a customer count is decremented, that is, we stop when $u_k^{\mathcal{N}} = \emptyset$.

The value of $u_k^{\mathcal{N}}$ is sampled as follows:

$$p(u_k^{\mathcal{N}}) = \begin{cases} t_k^{\mathcal{N} \to \mathcal{P}_i} / c_k^{\mathcal{N}} & \text{if} \quad u_k^{\mathcal{N}} = \mathcal{P}_i \\ 1 - \sum_{\mathcal{P}_i} p(u_k^{\mathcal{N}} = \mathcal{P}_i) & \text{if} \quad u_k^{\mathcal{N}} = \emptyset \quad . \end{cases} \tag{A.1}$$

To illustrate, when a word w_{dn} (with topic z_{dn}) is removed, we decrement $c_{z_{dn}}^{\theta_d}$, that is, $c_{z_{dn}}^{\theta_d}$ becomes $c_{z_{dn}}^{\theta_d} - 1$. We then determine if this word contributes to any table in node θ_d by sampling $u_{z_{dn}}^{\theta_d}$ from Equation (A.1). If $u_{z_{dn}}^{\theta_d} = \emptyset$, we do not decrement any table count and proceed with the next step in Gibbs sampling; otherwise, $u_{z_{dn}}^{\theta_d}$ can either be θ'_d or η_d, in these cases, we would decrement $t_{z_{dn}}^{\theta_d \to u_{z_{dn}}^{\theta_d}}$ and $c_{z_{dn}}^{u_{z_{dn}}^{\theta_d}}$, and continue the process recursively.

We present the decrementing process in Algorithm A. To remove a word w_{dn} during inference, we would need to decrement the counts contributed by w_{dn} (and z_{dn}). For the topic side, we decrement the counts associated with node $\mathcal{N} = \theta_d$ with group $k = z_{dn}$ using Algorithm A. While for the vocabulary side, we decrement the counts associated with the node $\mathcal{N} = \psi_{z_{dn}}$ with group $k = w_{dn}$. The effect of the word on the other PYP variables are implicitly considered through recursion.

Note that the procedure to decrementing a hashtag y_{dm} is similar, in this case, we decrement the counts for $\mathcal{N} = \theta'_d$ with $k = z'_{dm}$ (topic side), then decrement the counts for $\mathcal{N} = \psi'_{z'_{dm}}$ with $k = y_{dm}$ (vocabulary side).

Appendix A.2. Sampling a New Topic for a Word or a Hashtag

After decrementing, we sample a new topic for the word or the hashtag. The sampling process follows the procedure discussed in Section 4.2, but with different conditional posteriors (for both the word and the hashtag).

Algorithm A Decrementing counts associated with a PYP $\mathcal{N}$ and group k.

```
1. Decrement the customer count $c_k^{\mathcal{N}}$ by one.
2. Sample an auxiliary variable $u_k^{\mathcal{N}}$ with Equation (A.1).
3. For the sampled $u_k^{\mathcal{N}}$, perform the following:
   (a) If $u_k^{\mathcal{N}} = \emptyset$, exit the algorithm.
   (b) Otherwise, decrement the table count $t_k^{\mathcal{N}\to u_k^{\mathcal{N}}}$ by one and repeat Steps
       2 – 4 by replacing $\mathcal{N}$ with $u_k^{\mathcal{N}}$.
```

The full conditional posterior probability for the collapsed blocked Gibbs sampling can be derived easily. For instance, the conditional posterior for sampling the topic z_{dn} of word w_{dn} is

$$p(z_{dn}, \mathbf{T}, \mathbf{C} \,|\, \mathbf{Z}^{\circ -dn}, \mathbf{W}^{\circ}, \mathbf{T}^{-dn}, \mathbf{C}^{-dn}, \Xi) = \frac{p(\mathbf{Z}^{\circ}, \mathbf{T}, \mathbf{C} \,|\, \mathbf{W}^{\circ}, \Xi)}{p(\mathbf{Z}^{\circ -dn}, \mathbf{T}^{-dn}, \mathbf{C}^{-dn} \,|\, \mathbf{W}^{\circ}, \Xi)} \tag{A.2}$$

which can then be easily decomposed into simpler form (see discussion in Section 4.2) using Equation (63). Here, the superscript $\square^{-dn}$ indicates the word w_{dn} and the topic z_{dn} are removed from the respective sets. Similarly, the conditional posterior probability for sampling the topic z'_{dm} of hashtag y_{dm} can be derived as

$$p(z'_{dm}, \mathbf{T}, \mathbf{C} \,|\, \mathbf{Z}^{\circ -dm}, \mathbf{W}^{\circ}, \mathbf{T}^{-dm}, \mathbf{C}^{-dm}, \Xi) = \frac{p(\mathbf{Z}^{\circ}, \mathbf{T}, \mathbf{C} \,|\, \mathbf{W}^{\circ}, \Xi)}{p(\mathbf{Z}^{\circ -dm}, \mathbf{T}^{-dm}, \mathbf{C}^{-dm} \,|\, \mathbf{W}^{\circ}, \Xi)} \tag{A.3}$$

where the superscript $\square^{-dm}$ signals the removal of the hashtag y_{dm} and the topic z'_{dm}. As in Section 4.2, we compute the posterior for all possible changes to $\mathbf{T}$ and $\mathbf{C}$ corresponding to the new topic (for z_{dn} or z'_{dm}). We then sample the next state using a Gibbs sampler.

Appendix A.3. Estimating the Probability Vectors of the PYPs with Multiple Parents

Following Section 4.4, we estimate the various probability distributions of the PYPs by their posterior means. For a PYP $\mathcal{N}$ with a single PYP parent $\mathcal{P}_1$, as discussed in Section 4.4, we can estimate its probability vector $\hat{\mathcal{N}} = (\hat{\mathcal{N}}_1, \dots, \hat{\mathcal{N}}_K)$ as

$$\begin{aligned} \hat{\mathcal{N}}_k &= \mathbb{E}[\mathcal{N}_k \,|\, \mathbf{Z}^{\circ}, \mathbf{W}^{\circ}, \mathbf{T}, \mathbf{C}, \Xi] \\ &= \frac{\big(\alpha^{\mathcal{N}} T^{\mathcal{N}} + \beta^{\mathcal{N}}\big)\mathbb{E}[\mathcal{P}_{1k} \,|\, \mathbf{Z}^{\circ}, \mathbf{W}^{\circ}, \mathbf{T}, \mathbf{C}, \Xi] + c_k^{\mathcal{N}} - \alpha^{\mathcal{N}} T_k^{\mathcal{N}}}{\beta^{\mathcal{N}} + C^{\mathcal{N}}} \,, \end{aligned} \tag{A.4}$$

which lets one analyse the probability vectors in a topic model using recursion.

Unlike the above, the posterior mean is slightly more complicated for a PYP $\mathcal{N}$ that has multiple PYP parents $\mathcal{P}_1, \ldots, \mathcal{P}_P$. Formally, we define the PYP $\mathcal{N}$ as

$$\mathcal{N} \,|\, \mathcal{P}_1, \ldots, \mathcal{P}_P \sim \mathrm{PYP}\Big(\alpha^{\mathcal{N}}, \beta^{\mathcal{N}}, \rho_1^{\mathcal{N}} \mathcal{P}_1 + \cdots + \rho_P^{\mathcal{N}} \mathcal{P}_P\Big) , \tag{A.5}$$

where the mixing proportion $\rho^{\mathcal{N}} = (\rho_1^{\mathcal{N}}, \ldots, \rho_P^{\mathcal{N}})$ follows a Dirichlet distribution with parameter $\lambda^{\mathcal{N}} = (\lambda_1^{\mathcal{N}}, \ldots, \lambda_P^{\mathcal{N}})$:

$$\rho^{\mathcal{N}} \sim \mathrm{Dirichlet}\big(\lambda^{\mathcal{N}}\big) . \tag{A.6}$$

Before we can estimate the probability vector, we first estimate the mixing proportion with its posterior mean given the customer counts and table counts:

$$\hat{\rho}_i^{\mathcal{N}} = \mathbb{E}[\rho_i^{\mathcal{N}} \,|\, \mathbf{Z}^\circ, \mathbf{W}^\circ, \mathbf{T}, \mathbf{C}, \Xi] = \frac{T^{\mathcal{N} \to \mathcal{P}_i} + \lambda_i^{\mathcal{N}}}{T^{\mathcal{N}} + \sum_i \lambda_i^{\mathcal{N}}} \,. \tag{A.7}$$

Then, we can estimate the probability vector $\hat{\mathcal{N}} = (\hat{\mathcal{N}}_1, \ldots, \hat{\mathcal{N}}_K)$ by

$$\hat{\mathcal{N}}_k = \frac{\big(\alpha^{\mathcal{N}} T^{\mathcal{N}} + \beta^{\mathcal{N}}\big) \hat{H}_k^{\mathcal{N}} + c_k^{\mathcal{N}} - \alpha^{\mathcal{N}} T_k^{\mathcal{N}}}{\beta^{\mathcal{N}} + C^{\mathcal{N}}} \,, \tag{A.8}$$

where $\hat{H}^{\mathcal{N}} = (\hat{H}_1^{\mathcal{N}}, \ldots, \hat{H}_K^{\mathcal{N}})$ is the expected base distribution:

$$\hat{H}_k^{\mathcal{N}} = \sum_{i=1}^{P} \hat{\rho}_i^{\mathcal{N}} \mathbb{E}[\mathcal{P}_{ik} \,|\, \mathbf{Z}^\circ, \mathbf{W}^\circ, \mathbf{T}, \mathbf{C}, \Xi] \,. \tag{A.9}$$

With these formulations, all the topic distributions and the word distributions in the TNTM can be reconstructed from the customer counts and table counts. For instance, the author–topic distribution ν_i of each author i can be determined recursively by first estimating the topic distribution μ_0 . The word distributions for each topic are similarly estimated.

Appendix A.4. MH Algorithm for the Random Function Network Model

Here, we discuss how we learn the topic distributions μ_0 and ν from the random function network model. We configure the MH algorithm to start after running one thousand iterations of the collapsed blocked Gibbs sampler, this is to we can quickly initialise the TNTM with the HPYP topic model before running the full algorithm. In addition, this allows us to demonstrate the improvement to the TNTM due to the random function network model.

To facilitate the MH algorithm, we have to represent the topic distributions μ_0 and ν explicitly as probability vectors, that is, we do not store the customer counts and table counts for μ_0 and ν after starting the MH algorithm. In the MH algorithm, we propose new samples for μ_0 and ν, and then accept or reject the samples. The details for the MH algorithm is as follow.

In each iteration of the MH algorithm, we use the Dirichlet distributions as proposal distributions for μ_0 and ν:

$$q(\mu_0^{\text{new}} \,|\, \mu_0) = \text{Dirichlet}(\beta^{\mu_0}\mu_0)\,, \tag{A.10}$$

$$q(\nu_i^{\text{new}} \,|\, \nu_i) = \text{Dirichlet}(\beta^{\nu_i}\nu_i)\,. \tag{A.11}$$

These proposed μ_0 and ν are sampled given the their previous values, and we note that the first μ_0 and ν are computed using the technique discussed in Appendix A.3. These proposed samples are subsequently used to sample $\mathbf{Q}^{\text{new}}$. We first compute the quantities ς^{new} and κ^{new} using the proposed μ_0^{new} and ν^{new} with Equation (61) and Equation (62). Then we sample $\mathbf{Q}^{\text{new}}$ given ς^{new} and κ^{new} using the elliptical slice sampler (see Murray et al., 2010):

$$\mathbf{Q}^{\text{new}} \sim \text{GP}(\varsigma^{\text{new}}, \kappa^{\text{new}})\,. \tag{A.12}$$

Finally, we compute the acceptance probability $A' = \min(A, 1)$, where

$$\begin{aligned} A = {} & \frac{p(\mathbf{Q}^{\text{new}} \,|\, \mathbf{X}, \nu^{\text{new}}, \mathbf{\Xi})}{p(\mathbf{Q}^{\text{old}} \,|\, \mathbf{X}, \nu^{\text{old}}, \mathbf{\Xi})} \, \frac{f^*(\mu_0^{\text{new}} \,|\, \nu^{\text{new}}, \mathbf{T}) \, \prod_{i=1}^{A} f^*(\nu_i^{\text{new}} \,|\, \mathbf{T})}{f^*(\mu_0^{\text{old}} \,|\, \nu^{\text{old}}, \mathbf{T}) \, \prod_{i=1}^{A} f^*(\nu_i^{\text{old}} \,|\, \mathbf{T})} \\ & \times \frac{q(\mu_0^{\text{old}} \,|\, \mu_0^{\text{new}}) \, \prod_{i=1}^{A} q(\nu_i^{\text{old}} \,|\, \nu_i^{\text{new}})}{q(\mu_0^{\text{new}} \,|\, \mu_0^{\text{old}}) \, \prod_{i=1}^{A} q(\nu_i^{\text{new}} \,|\, \nu_i^{\text{old}})}\,, \end{aligned} \tag{A.13}$$

and we define $f^*(\mu_0 \,|\, \nu, \mathbf{T})$ and $f^*(\nu \,|\, \mathbf{T})$ as

$$f^*(\mu_0 \,|\, \nu, \mathbf{T}) = \prod_{k=1}^{K} (\mu_{0k})^{t_k^{\mu_1} + \sum_{i=1}^{A} \nu_i}\,, \tag{A.14}$$

$$f^*(\nu_i \,|\, \mathbf{T}) = \prod_{k=1}^{K} (\nu_{ik})^{\sum_{d=1}^{D} t_k^{\eta_d} I(a_d = i)}\,. \tag{A.15}$$

The $f^*(\cdot)$ corresponds to the topic model posterior of the variables μ_0 and ν after we represent them as probability vectors explicitly. Note that we treat the acceptance probability A as 1 when the expression in Equation (A.13) evaluates to more than 1. We then accept the proposed samples with probability A, if the sample are not accepted, we keep the respective old values. This completes one iteration of the MH scheme. We summarise the MH algorithm in Algorithm B.

Algorithm B Performing the MH algorithm for one iteration.

1. Propose a new μ_0^{new} with Equation (A.10).
2. For each author i, propose a new ν_i^{new} with Equation (A.11).
3. Compute the mean function ς^{new} and the covariance matrix κ^{new} with Equation (61) and Equation (62).
4. Sample $\mathbf{Q}^{\text{new}}$ from Equation (A.12) using the elliptical slice sampler from Murray et al. (2010).
5. Accept or reject the samples with acceptance probability from Equation (A.13).

Appendix A.5. Hyperparameter Sampling

We sample the hyperparameters β using an auxiliary variable sampler while leaving α fixed. We note that the auxiliary variable sampler for PYPs that have multiple parents are identical to that of PYPs with single parent, since the sampler only used the total customer counts $C^{\mathcal{N}}$ and the total table counts $T^{\mathcal{N}}$ for a PYP $\mathcal{N}$. We refer the readers to Section 4.3 for details.

We would like to point out that hyperparameter sampling is performed for all PYPs in TNTM for the first one thousand iterations. After that, as μ_0 and ν are represented as probability vectors explicitly, we only sample the hyperparameters for the other PYPs (except μ_0 and ν). We note that sampling the concentration parameters allows the topic distributions of each author to vary, that is, some authors have few very specific topics and some other authors can have a wider range of topics. For simplicity, we fix the kernel hyperparameters s, l and σ to 1. Additionally, we also make the priors for the mixing proportions uninformative by setting the λ to 1. We summarise the full inference algorithm for the TNTM in Algorithm C.

Bibliography

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, AISTATS 2010, pages 541–548, Brookline, MA, USA. Microtome Publishing.

Algorithm C Full inference algorithm for the TNTM.

```
1. Initialise the HPYP topic model by assigning random topic to the latent topic $z_{dn}$ associated with each word $w_{dn}$, and to the latent topic $z'_{dm}$ associated with each hashtag $y_{dm}$. Then update all the relevant customer counts $\mathbf{C}$ and table counts $\mathbf{T}$.
2. For each word $w_{dn}$ in each document $d$, perform the following:
    (a) Decrement the counts associated with $w_{dn}$ (see Appendix A.1).
    (b) Blocked sample a new topic for $z_{dn}$ and corresponding customer counts $\mathbf{C}$ and table counts $\mathbf{T}$ (with Equation (A.2)).
    (c) Update (increment counts) the topic model based on the sample.
3. For each hashtag $y_{dm}$ in each document $d$, perform the following:
    (a) Decrement the counts associated with $y_{dm}$ (see Appendix A.1).
    (b) Blocked sample a new topic for $z'_{dn}$ and corresponding customer counts $\mathbf{C}$ and table counts $\mathbf{T}$ (with Equation (A.3)).
    (c) Update (increment counts) the topic model based on the sample.
4. Sample the hyperparameter $\beta^{\mathcal{N}}$ for each PYP $\mathcal{N}$ (see Appendix A.5).
5. Repeat Steps $2-4$ for 1,000 iterations.
6. Alternatingly perform the MH algorithm (Algorithm B) and the collapsed blocked Gibbs sampler conditioned on $\mu_0$ and $\nu$.
7. Sample the hyperparameter $\beta^{\mathcal{N}}$ for each PYP $\mathcal{N}$ except for $\mu_0$ and $\nu$.
8. Repeat Steps $6-7$ until the model converges or when a fix number of iterations is reached.
```