

A Supplementary Material

A.1 Summary of Truncated DCNT Generative Model

The DCNT model assumes the following generative process for a corpus of D documents, each containing N words, and conditioned on metadata encoded via F features:

1. Global parameters shared across all documents:
 - (a) For all features $f = 1, 2, \dots, F$:
 - i. Draw $\lambda_f \sim \text{Gam}(a_f, b_f)$
 - ii. Draw $\mu_f \sim N(0, \gamma_\mu)$
 - (b) Draw $\lambda_A \sim \text{Gam}(a_A, b_A)$
 - (c) Draw $\lambda_V \sim \text{Gam}(a_V, b_V)$
 - (d) Draw $\eta_{:k} \sim N(\mu, \Lambda^{-1})$
 - (e) Draw $A_{k\ell} \sim N(0, (k\lambda_A)^{-1})$
2. For each document $d = 1, 2, \dots, D$:
 - (a) Draw $u_{:d} \sim N(\eta^T \phi_{:d}, I_{\bar{K}})$
 - (b) Draw $v_{:d} \sim N(Au_{:d}, \lambda_v^{-1} I_{\bar{K}})$
 - (c) Let $\pi_{kd} = \psi(v_{kd}) \prod_{\ell=1}^{k-1} \psi(-v_{\ell d})$
 - (d) For each word $n = 1, 2, \dots, N$:
 - i. Draw $z_{dn} \sim \text{Mult}(\pi_{:d})$
 - ii. Draw $w_{dn} \sim \text{Mult}(\Omega_{z_{dn}})$

A.2 SCNT: A Singly Correlated Nonparametric Topic Model

To explore the benefits of our full square-root representation of topic correlations, we also consider a model where A is constrained to be a diagonal matrix. The posterior required by a Gibbs sampler then becomes

$$\begin{aligned} p(A_{kk} | v_{k:}, u_{k:}, \lambda_A, \lambda_v) &\propto N(A_{kk} | 0, \lambda_A^{-1}) N(v_{k:}^T | u_{k:}^T A_{kk}, \lambda_v^{-1} I_D) \\ &\propto N(A_{kk} | \lambda_A \lambda_v^{-1} + u_{k:} u_{k:}^T)^{-1} u_{k:} v_{k:}^T, (\lambda_A + \lambda_v u_{k:} u_{k:}^T)^{-1} \end{aligned} \quad (1)$$

Note that when A is constrained to be a diagonal matrix, all rows are assigned the same prior precision λ_A . The posterior for λ_A then equals

$$\begin{aligned} p(\lambda_A | A, a_A, b_A) &\propto \text{Gam}(\lambda_A | a_A, b_A) \prod_{k=1}^{\bar{K}} N(A_{kk} | 0, \lambda_A^{-1}) \\ &\propto \text{Gam}(\lambda_A | \frac{1}{2} \bar{K} + a_A, \frac{1}{2} \sum_{k=1}^{\bar{K}} A_{kk}^2 + b_A) \end{aligned} \quad (2)$$

A.3 Monte Carlo Estimation of DCNT Topic Covariances

The DCNT can model both positive and negative correlations among topic frequencies, but due to the nonlinearity associated with the logistic stick-breaking transformation, these covariances cannot be determined in closed form. We instead use a Monte Carlo estimate based on S samples from the covariance of each document, computed as follows:

$$\mathbb{E}[\pi_{:d}] = \frac{1}{S} \sum_{s=1}^S \pi_{:d}^s \quad (3)$$

$$\text{Cov}[\pi_{:d}] = \frac{1}{S} \sum_{s=1}^S (\pi_{:d}^s - \mathbb{E}[\pi_{:d}])(\pi_{:d}^s - \mathbb{E}[\pi_{:d}])^T \quad (4)$$

$$\hat{\Sigma} = \frac{1}{D} \sum_{d=1}^D \text{Cov}(\pi_{:d}) \quad (5)$$

Here, $\pi_{:d}^s$ is computed by mapping a single sample of $v_{:d}$, conditioned on the learned model parameters, through the logistic stick breaking transformation. For our visualizations, we set $S = 5000$ for each document d . We used a similar Monte Carlo estimator for the LDA model, conditioned on its Dirichlet topic weights α .

A.4 Chib Style Estimation of Predictive Likelihoods

The Chib style estimator can be used to approximate the predictive likelihood of a held out document by marginalizing out the topic assignment variables z_d , and topic weights $v_{:d}$ and $u_{:d}$, to obtain $p(w_d | \zeta, \Gamma)$, where w_d refers to the set of N words in a held out document d , $\zeta = \{A, \Omega, \eta, \phi, \lambda_V\}$ are the parameters learned from training data, and Γ is the set of hyperparameters specified before training. The Chib-style estimator is based on a distinguished high-probability set of latent variables $(z_d^*, v_{:d}^*, u_{:d}^*)$, chosen so that:

$$p(w_d | \zeta, \Gamma) = \frac{p(w_d, z_d^*, v_{:d}^*, u_{:d}^* | \zeta, \Gamma)}{p(z_d^*, v_{:d}^*, u_{:d}^* | w_d, \zeta, \Gamma)} \quad (6)$$

$$p(w_d | \zeta, \Gamma) \approx \frac{p(w_{:d}, z_d^*, v_{:d}^*, u_{:d}^* | \zeta, \Gamma)}{\frac{1}{S} \sum_{s=1}^S T(z_d^*, v_{:d}^*, u_{:d}^* \leftarrow z_d^s, v_{:d}^s, u_{:d}^s)} \quad (7)$$

where $T(z_d^*, v_{:d}^*, u_{:d}^* \leftarrow z_d^s, v_{:d}^s, u_{:d}^s)$ is a reversible Markov chain operator used to numerically approximate the marginalization of z_d , $v_{:d}$, and $u_{:d}$ by calculating the transition probabilities from S samples from their respective posterior given w_d . These can be obtained via our standard Gibbs sampling updates for z_d and $u_{:d}$, and our Metropolis-Hastings independence sampler for $v_{:d}$ which we denote by $\text{MH}(\cdot)$. Depending on the direction of this chain, the respective posterior distributions used to evaluate the transition operators will be different. We denote the forward transition operator as $T(z^*, v^*, u^* \leftarrow z^s, v^s, u^s)$ and the reverse transition operator as $\tilde{T}(z^*, v^*, u^* \leftarrow z^s, v^s, u^s)$ which can be defined as follows:

$$T(z^*, v^*, u^* \leftarrow z^s, v^s, u^s) = p(z^* | v^s, z^s) q(v^* | v^s, z^*, u^s) p(u^* | v^*) \quad (8)$$

$$\tilde{T}(z^*, v^*, u^* \leftarrow z^s, v^s, u^s) = p(z^* | v^*, z^s) q(v^* | v^s, z^s, u^*) p(u^* | v^s) \quad (9)$$

The log posterior distributions have the following form for the forward transition $T(\cdot)$:

$$\log p(z^* | v^s, z^s, \Omega) = \sum_{n=1}^N \log \left[\frac{p(w_{dn} | z_{dn}^s = z_{dn}^*) p(z_{dn}^s = z_{dn}^* | \pi^s)}{\sum_{k=1}^K p(w_{dn} | z^s = k) p(z^s = k | \pi^s)} \right] \quad (10)$$

$$\log q(v^* | v^s, z^*, u^s) = \log N(v^* | A u^s, L) + \min \left[0, \sum_{k=1}^K n_k^* \log \pi_k^* - \sum_{k=1}^K n_k^* \log \pi_k^s \right] \quad (11)$$

$$\log p(u^* | v^*) = \log N(u^* | (I_{\bar{K}} + A^T L A)^{-1} (A^T L v^* + \eta^T \phi_d), (I_{\bar{K}} + A^T L A)^{-1}) \quad (12)$$

Our stick breaking weights for $v_{:d}$ are constructed as $\pi_{kd}^s = \psi(v_{kd}^s) \prod_{\ell=1}^{k-1} \psi(-v_{\ell d}^s)$, and the topic counts for document d are denoted by $n_k^s = \sum_{n=1}^N \delta(z_{dn}^s, k)$. The precision matrix for $u_{:d}$ under our prior is denoted by $L = \lambda_V I_{\bar{K}}$.

As in previous applications of similar Chib-style estimators, we set the length of the transition chain to be $S = 1000$, and run $T = 1000$ iterations to determine a high posterior probability state. Due to the use of a Metropolis-Hastings proposal, we need to reweight the final predictive likelihood by the probability of accepting the first sample of v^s from the high posterior state; runs where the proposal is rejected produce a likelihood estimate of zero. Since the entries of v_{kd} are independent in our posterior, we empirically estimate the probabilities of accepting each entry and use that estimate to determine the final predictive likelihood. For the NIPS corpus, we set the number of samples to estimate these rejection probabilities to $R = 25,000$.

Algorithm: Chib Style Estimator of Marginal Likelihood for Document d

- 1: Fix (ζ, Γ) for each document d
- 2: **for** $t = 1 : T$ **do**
- 3: Sample $v_{:d}^t, z_d^t, u_{:d}^t$ from MCMC proposals to reach a high posterior probability state.
- 4: **end for**
- 5: Set $\mathbf{u}^* = u_{:d}^t, \mathbf{v}^* = v_{:d}^t$, and $\mathbf{z}^* = z_d^t$
- 6: Sample $s \sim \text{Unif}(1, 2, \dots, S)$
- 7: **for** $n = 1 : N$ **do**
- 8: Sample $z_{dn}^s \sim \text{Mult}(\theta_k^s)$ where $\theta_k^s = \frac{p(w_{dn} | z_{dn}^* = k)p(z_{dn}^* = k | \pi^s)}{\sum_{\ell=1}^K p(w_{dn} | z_{dn}^* = \ell)p(z_{dn}^* = \ell | \pi^s)}$
- 9: **end for**
- 10: **for** $1 : \bar{K}$ **do**
- 11: **for** $1 : R$ **do**
- 12: Sample $v_{kd}^r \sim \text{MH}(\mathbf{v}^*, \mathbf{z}^s, \mathbf{u}^*, \zeta)$
- 13: **if** $[1 - \sum_{k=1}^{\bar{K}} \delta(v_{kd}^r, v_{kd}^*)] > 0$ **then**
- 14: Set $\mathbf{v}^s = v_{:d}^r$
- 15: **end if**
- 16: **end for**
- 17: $\rho_k = \frac{1 - \sum_{r=1}^R \delta(v_{kd}^r, v_{kd}^*)}{R}$
- 18: **end for**
- 19: Sample $\mathbf{u}^s \sim N((I_{\bar{K}} + A^T L A)^{-1}(A^T L \mathbf{v}^s + \eta^T \phi_{:d}), (I_{\bar{K}} + A^T L A)^{-1})$
- 20: Begin Forward and Backward Chain
- 21: **for** $i = (s + 1) : S$ **do**
- 22: **for** $n = 1 : N$ **do**
- 23: Sample $z_{dn}^i \sim \text{Mult}(\theta_k^{i-1})$ where $\theta_k^{i-1} = \frac{p(w_{dn} | z_{dn}^{i-1})p(z_{dn}^{i-1} | \pi^{i-1})}{\sum_{\ell=1}^K p(w_{dn} | z_{dn}^{i-1} = \ell)p(z_{dn}^{i-1} = \ell | \pi^{i-1})}$
- 24: **end for**
- 25: **for** $k = 1 : \bar{K}$ **do**
- 26: Sample $v_{kd}^i \sim \text{MH}(\mathbf{v}^{i-1}, \mathbf{z}^i, \mathbf{u}^{i-1}, \zeta)$
- 27: **end for**
- 28: Sample $\mathbf{u}^i \sim N((I_{\bar{K}} + A^T L A)^{-1}(A^T L \mathbf{v}^i + \eta^T \phi_{:d}), (I_{\bar{K}} + A^T L A)^{-1})$
- 29: Calculate $T(\mathbf{z}^*, \mathbf{v}^*, \mathbf{u}^* \leftarrow \mathbf{z}^i, \mathbf{v}^i, \mathbf{u}^i)$
- 30: **end for**
- 31: **for** $j = (s - 1) : -1 : 1$ **do**
- 32: Sample $\mathbf{u}^j \sim N((I_{\bar{K}} + A^T L A)^{-1}(A^T L \mathbf{v}^{j+1} + \eta^T \phi_{:d}), (I_{\bar{K}} + A^T L A)^{-1})$
- 33: **for** $k = \bar{K} : -1 : 1$ **do**
- 34: Sample $v_{kd}^j \sim \text{MH}(\mathbf{v}^{j+1}, \mathbf{z}^{j+1}, \mathbf{u}^j, \zeta)$
- 35: **end for**
- 36: **for** $n = N : -1 : 1$ **do**
- 37: Sample $z_{dn}^j \sim \text{Mult}(\theta_k^{j+1})$ where $\theta_k^{j+1} = \frac{p(w_{dn} | z_{dn}^{j+1})p(z_{dn}^{j+1} | \pi^j)}{\sum_{\ell=1}^K p(w_{dn} | z_{dn}^{j+1} = \ell)p(z_{dn}^{j+1} = \ell | \pi^j)}$
- 38: **end for**
- 39: Calculate $\tilde{T}(\mathbf{z}^*, \mathbf{v}^*, \mathbf{u}^* \leftarrow \mathbf{z}^j, \mathbf{v}^j, \mathbf{u}^j)$
- 40: **end for**
- 41: $p(\mathbf{w} | \Omega, \eta, \phi, \lambda_V, \beta) \approx \frac{p(\mathbf{w}, \mathbf{z}^*, \mathbf{v}^*, \mathbf{u}^* | \Omega, \eta, \phi, \lambda_V, \beta) \prod_{k=1}^{\bar{K}} \rho_k}{\frac{1}{S} \sum_{s=1}^S T(\mathbf{z}^*, \mathbf{v}^*, \mathbf{u}^* \leftarrow \mathbf{z}^s, \mathbf{v}^s, \mathbf{u}^s)}$