Supplementary Material:
A unifying representation for a class of dependent random measures

1 Introduction

We present a complete description of the tGaP-PFA topic model, the associated Gibbs sampler and how to compute perplexity for unseen documents under the model with samples drawn from the Gibbs sampler.

2 Model

Recall that $w_{pnt}$ represents the number of occurrences of word $p$ in the $n$th document at time $t$, and that we decompose this as $w_{pnt} = \sum_{k=1}^{\infty} \tilde{w}_{pntk}$, where $\tilde{w}_{pntk}$ is the number of occurrences attributed to topic $k$. In the generative process presented below, $p$ indexes the vocabulary, $t$ indexes the observed times of documents, $n$ indexes the documents at a time $t$ and takes values in $\{1, \ldots, N_t\}$, and $k$ indexes the topics. Additionally, $l$ indexes the kernel functions of the RVM (Tipping, 2001) with centers $m_l$, which we take to be the locations of the observations (although this is not necessary).

The generative process is as follows

$$\Gamma := \sum_{k=1}^{\infty} \pi_k \delta(x_k, \theta_k) \sim \text{CRM}(\nu_{G0}(d\pi)H(dx)G_0(d\theta)),$$

where $x_k := (\omega_{0k}, \ldots, \omega_{Lk}, \phi_k)$; $\nu_{G0}(d\pi) = \pi^{-1}\exp(-\pi)d\pi$ is the Lévy measure of the gamma process with parameters $(1, 1)$; $B_0(d\theta)$ is the $P$-dimensional Dirichlet distribution with parameter $\alpha_0$; and $H(dx) = H_\phi(d\phi) \prod_{l=0}^{L} H_\omega(d\omega_l)$, where $H_\phi(d\phi)$ is the categorical distribution over the dictionary of kernel widths, and $H_\omega(d\omega_l) \sim \text{NiG}(0, c_0, d_0)$ is drawn from the normal-inverse gamma distri-
bution. The rest of the model is
\begin{equation}
p_{x_k}(t) = \Phi(\omega_{0k} + \sum_{l=1}^{L}\omega_{lk} \exp(-\phi_k \|t - t_l\|_2^2)) \tag{2}
\end{equation}
\begin{equation}
r_{k}^{n,t} \sim \text{Ber}(p_{x_k}(t)) \tag{3}
\end{equation}
\begin{equation}
G_{n,t} := \sum_{k=1}^{\infty} r_{k}^{n,t} \pi_k \delta_{\theta_k} \tag{4}
\end{equation}
\begin{equation}
\beta_{k}^{n,t} \sim \text{Ga}(\epsilon, 1), n = 1, \ldots, N_t, k \in \mathbb{N} \tag{5}
\end{equation}
\begin{equation}
\tilde{w}_{pntk} \sim \text{Pois}(\theta_{kp} r_{k}^{n,t} \pi_k \beta_{k}^{n,t}) \tag{6}
\end{equation}
\begin{equation}
 w_{pnt} = \sum_{k=1}^{\infty} \tilde{w}_{pntk} \sim \text{Pois}(\sum_{k=1}^{\infty} \theta_{kp} r_{k}^{n,t} \pi_k \beta_{k}^{n,t}) \tag{7}
\end{equation}

3 Gibbs sampler

We use a truncated version of the model by fixing the number of atoms we will represent to $K$ and forming the (finite) random measure, $\Gamma_K := \sum_{k=1}^{K} \pi_k \delta_{(x_k, \phi_k)}$, where $\pi_k \sim \text{Ga}(1/K, 1)$, $x_k := (\omega_{0k}, \ldots, \omega_{Lk}, \phi_k)$, $\omega_k \sim \text{NIG}(0, c_0, d_0)$, and $\phi_k \sim \{\phi_1^*, \ldots, \phi_d^*\}$. In the limit, $K \to \infty$, $\Gamma_K \to \Gamma$ in distribution. This truncation allows for the derivation of a straightforward Gibbs sampler. We assume $\mathcal{T}$ is the set of unique observed times.

We sample each of the variables in turn from their full conditional distributions. We use a standard data-augmentation technique for probit regression to sample the $\omega_{lk}$ variables by introducing an auxiliary variable $\tilde{r}_{k}^{n,t} \sim N(p_{x_k}(t), 1)$ for each topic $k$ at each document $n$ at time $t$, such that
\begin{equation}
r_{k}^{n,t} = \begin{cases} 
1 & \text{if } \tilde{r}_{k}^{n,t} > 0 \\
0 & \text{otherwise.}
\end{cases}
\end{equation}

See Albert & Chib (1993) for details of the data augmentation. The conditional distributions are as follows.

- **Topics, $\theta_k$.**

  \begin{equation}
  \theta_k|\ldots \sim \text{Dir}(\alpha_\theta + \tilde{w}_{1..k}, \ldots, \alpha_\theta + \tilde{w}_{P..k}) \tag{8}
  \end{equation}

  where $\tilde{w}_{P..k} = \sum_{t \in \mathcal{T}} \sum_{n=1}^{N_t} \tilde{w}_{pntk}$.

- **Global topic proportions, $\pi_k$.**

  \begin{equation}
  \pi_k|\ldots \sim \text{Ga}(\tilde{w}_{..k} + 1/K, \sum_{t \in \mathcal{T}} \sum_{n=1}^{N_t} \beta_{k}^{n,t} + 1) \tag{9}
  \end{equation}

  where $\tilde{w}_{..k} = \sum_{p=1}^{P} \sum_{t \in \mathcal{T}} \sum_{n=1}^{N_t} \tilde{w}_{pntk}$.

- **Per-topic counts, $\tilde{w}_{pntk}$.**

  \begin{equation}
  (\tilde{w}_{pnt1}, \ldots, \tilde{w}_{pntK})|\ldots \sim \text{Mult}(w_{pnt}; \xi_{pnt1}, \ldots, \xi_{pntK}), \tag{10}
  \end{equation}

  where $\xi_{pnt} = \frac{\theta_{pk} r_{k}^{n,t} \pi_k \beta_{k}^{n,t}}{\sum_{j=1}^{K} \theta_{pj} r_{j}^{n,t} \pi_j \beta_{j}^{n,t}}$.
where we ensure that the denominator is greater than 0 by making sure that when sampling the $r_{n,t}^{n,t}$s, every document is not thinning at least one topic, i.e. $\forall t \forall n \exists j, r_{n,t}^{n,t} = 1$.

- **Per-document topic rate**, $\beta_{n,t}^{n,t}$.
  
  $$ \beta_{n,t}^{n,t} | \ldots \sim \text{Ga}(\tilde{w}_{ntk} + a, r_{k}^{n,t} \pi_k + 1) $$  
  where $\tilde{w}_{ntk} = \sum_{p=1}^{P} \tilde{w}_{pntk}$.

- **Time-dependent indicators**, $r_{n,t}^{n,t}$: There are three cases:
  
  1. $\forall j, r_{n,t}^{n,t} = 0 \rightarrow r_{n,t}^{n,t} = 1$
  2. $\exists p, \tilde{w}_{pntk} > 0 \rightarrow r_{n,t}^{n,t} = 1$
  3. $\forall p, \tilde{w}_{pntk} = 0$

  Cases 1 and 2 are deterministic. For case 3 let $u_{pntk} \sim \text{Pois}(\rho_p)$ with $\rho_p = \theta_{pk} \pi_k r_{k}^{n,t}$ denote the fictitious count of word $p$ in the $n$th document at time $t$ assigned to topic $k$ disregarding $r_{n,t}^{n,t}$. The $u_{pntk}$ allow us to determine whether $\tilde{w}_{pntk} = 0$ because the topic has been thinned or because the topic is not popular (globally or for the individual document). Case 3 above then splits into the following cases:

  1. $\forall p, u_{pntk} = 0, r_{k}^{n,t} = 1$ with probability $\propto p(r_{k}^{n,t} = 1) \prod_{p=1}^{P} \text{Pois}(0; \rho_p)$
  2. $\exists p, u_{pntk} > 0, r_{k}^{n,t} = 0$ with probability $\propto p(r_{k}^{n,t} = 0) \left( 1 - \prod_{p=1}^{P} \text{Pois}(0; \rho_p) \right)$
  3. $\forall p, u_{pntk} = 0, r_{k}^{n,t} = 0$ with probability $\propto p(r_{k}^{n,t} = 0) \prod_{p=1}^{P} \text{Pois}(0; \rho_p)$

  We evaluate the three probabilities and sample from the resulting discrete distribution.

- **RVM weights**, $\omega_{lk}$. We introduce the auxiliary variables $\lambda_{lk}$ such that

  $$ \lambda_{lk} \sim \text{Ga}(c_0, d_0) $$
  $$ \omega_{lk} \sim N(0, \lambda_{lk}^{-1}) $$

  Let $\omega_k = (\omega_{l0}, \ldots, \omega_{Lk})^T$ be the vector of RVM weights and $\tilde{\mathbf{r}}_k$ be the vector of augmentation variables for all all time stamps, and

  $$ K_{lk} = \left( 1, K(t, m_1, \phi_k), \ldots, K(t, m_L, \phi_k) \right)^T $$  \hspace{1cm}  (12)

  be the vector of the evaluation of the RVM kernels for time $t$. Then, the conditional of $\omega_k$ is given by

  $$ \omega_k | \tilde{\mathbf{r}}_k, \ldots \sim N(\xi, B) $$  \hspace{1cm}  (13)

  where $B = (\text{diag}(\lambda_{l0}, \ldots, \lambda_{Lk}) + K_{lk}^T \tilde{\mathbf{r}}_k)^{-1}$ and $\xi = BK_{lk}^T \tilde{\mathbf{r}}_k$. 


• RVM auxiliary variables, $\tilde{r}_{k,n,t}^{n,t}$.

$$p(\tilde{r}_{k,n,t}^{n,t} | \ldots) \propto \begin{cases} 
N(K^T_k \omega_k, 1) \mathbf{1}(\tilde{r}_{k}^{n,t} > 0), & \text{if } r_{k}^{n,t} = 1 \\
N(K^T_k \omega_k, 1) \mathbf{1}(\tilde{r}_{k}^{n,t} < 0), & \text{if } r_{k}^{n,t} = 0
\end{cases}$$ (14)

which is a truncated normal distribution that we sample using the inversion method described in Albert & Chib (1993).

• RVM precisions, $\lambda_{lk}$.

$$\lambda_{lk} | \ldots \sim \text{Ga}(c_0 + \frac{1}{2}, d_0 + \frac{1}{2} \omega^2_{lk})$$ (15)

• RVM kernel widths, $\phi_k$. We assume a finite dictionary $\{\phi^*_1, \ldots, \phi^*_M\}$ of possible values for the RVM kernel widths, and a uniform prior on these values,

$$p(\phi_k = \phi^*_m | \ldots) \propto \frac{1}{M} \prod_{t \in T} \prod_{n=1}^{N_t} \Phi(p_{\phi^*_m}^{(t)}(\tilde{r}_{n,t,k}^{n,t}))^{r_{k,n,t}^{n,t}} (1 - \Phi(p_{\phi^*_m}^{(t)}(\tilde{r}_{n,t,k}^{n,t})))^{1-r_{k,n,t}^{n,t}}$$ (16)

where we have denoted the thinning function as a function of $\phi^*$ as the other variables are held fixed.

4 Perplexity

Similarly to Zhou et al. (2012), given $B$ samples of the model parameters and latent variables we compute a Monte Carlo estimate of the held-out perplexity for unobserved counts $\mathbf{Y} = [y_{n,t}^{p}]$ as

$$\exp \left( \frac{1}{y^*} \sum_{p=1}^{P} \sum_{t \in T} \sum_{n=1}^{N_t} y_{n,t}^{p,\hat{t}} \log \frac{\sum_{b=1}^{B} \sum_{k=1}^{K} \phi_{pk}^{(b)} \pi_{k}^{(b)} r_{n,t,k}^{(b)} \beta_{n,t,k}^{(b)}}{\sum_{b=1}^{B} \sum_{p=1}^{P} \sum_{k=1}^{K} \pi_{k}^{(b)} r_{n,t,k}^{(b)} \beta_{n,t,k}^{(b)}} \right)$$ (17)

where we have used a superscript $b$ to denote the $b$th sample of the parameters and latent variables$^1$ and $y^* = \sum_{p=1}^{P} \sum_{t \in T} \sum_{n=1}^{N_t} y_{n,t}^{p}$ denotes the held-out number of occurrences of word $p$ in the $n$th document at time $t$.

References


$^1$We have denoted the $b$th samples of $r_{k,n,t}^{n,t}$ and $\beta_{n,t,k}^{(b)}$ as $r_{k,n,t}^{(b)}$ and $\beta_{n,t,k}^{(b)}$ for readability.