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 nth 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 occurences 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

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

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 α_θ; 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 $H_\omega(d\omega_l) \sim \mathrm{NiG}(0, c_0, d_0)$ is drawn from the normal-inverse gamma distribution. The rest of the model is

$$p_{x_k}(t) = \Phi\big(\omega_{0k} + \textstyle\sum_{l=1}^{L}\omega_{lk}\exp(-\phi_k\,||t-t_l||_2^2)\big) \tag{2}$$
$$r_k^{n,t} \sim \mathrm{Ber}(p_{x_k}(t)) \tag{3}$$
$$G_{n,t} := \textstyle\sum_{k=1}^{\infty} r_k^{n,t}\pi_k\delta_{\theta_k} \tag{4}$$
$$\beta_k^{n,t} \sim \mathrm{Ga}(e,1), n=1,\ldots,N_t, k\in\mathbb{N} \tag{5}$$
$$\tilde{w}_{pntk} \sim \mathrm{Pois}(\theta_{kp} r_k^{n,t}\pi_k\beta_k^{n,t}) \tag{6}$$
$$w_{pnt} = \sum_{k=1}^{\infty}\tilde{w}_{pntk} \sim \mathrm{Pois}(\sum_{k=1}^{\infty}\theta_{kp} r_k^{n,t}\pi_k\beta_k^{n,t}) \tag{7}$$

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 $\pi_k \sim \mathrm{Ga}(1/K,1)$, $x_k := (\omega_{0k},\ldots,\omega_{Lk},\phi_k)$, $\omega_{lk} \sim \mathrm{NiG}(0,c_0,d_0)$, 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 straight-forward 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 ω_{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

$$r_k^{n,t} = \begin{cases} 1 & \text{if } \tilde{r}_k^{n,t} > 0 \\ 0 & \text{otherwise.}\end{cases}$$

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

- **Topics,** θ_k.
$$\theta_k|\ldots \sim \mathrm{Dir}(\alpha_\theta + \tilde{w}_{1\cdot\cdot k},\ldots,\alpha_\theta+\tilde{w}_{P\cdot\cdot k}) \tag{8}$$
where $\tilde{w}_{p\cdot\cdot k} = \sum_{t\in\mathcal{T}}\sum_{n=1}^{N_t}\tilde{w}_{pntk}$.

- **Global topic proportions,** π_k.
$$\pi_k|\ldots \sim \mathrm{Ga}(\tilde{w}_{\cdots k} + 1/K, \sum_{t\in\mathcal{T}}\sum_{n=1}^{N_t}\beta_k^{n,t}+1) \tag{9}$$
where $\tilde{w}_{\cdots k} = \sum_{p=1}^{P}\sum_{t\in\mathcal{T}}\sum_{n=1}^{N_t}\tilde{w}_{pntk}$.

- **Per-topic counts,** $\tilde{w}_{pntk}$.
$$(\tilde{w}_{pnt1},\ldots,\tilde{w}_{pntK})|\ldots \sim \mathrm{Mult}(w_{pnt};\xi_{pnt1},\ldots,\xi_{pntK}),$$
$$\text{where } \xi_{pntk} = \frac{\theta_{pk} r_k^{n,t}\pi_k\beta_k^{n,t}}{\sum_{j=1}^{K}\theta_{pj} r_j^{n,t}\pi_j\beta_j^{n,t}} \tag{10}$$

where we ensure that the denominator is greater than 0 by making sure that when sampling the $r_k^{n,t}$s, every document is not thinning at least one topic, i.e. $\forall t \forall n \exists j, r_j^{n,t} = 1$.

- **Per-document topic rate,** $\beta_k^{n,t}$.

$$\beta_k^{n,t} | \ldots \sim \mathrm{Ga}(\tilde{w}_{\cdot ntk} + a, r_k^{n,t} \pi_k + 1) \tag{11}$$

where $\tilde{w}_{\cdot ntk} = \sum_{p=1}^{P} \tilde{w}_{pntk}$.

- **Time-dependent indicators,** $r_k^{n,t}$**:** There are three cases:

1. $\forall j, r_j^{n,t} = 0 \rightarrow r_k^{n,t} = 1$
2. $\exists p, \tilde{w}_{pntk} > 0 \rightarrow r_k^{n,t} = 1$
3. $\forall p, \tilde{w}_{pntk} = 0$

Cases 1 and 2 are deterministic. For case 3 let $u_{pntk} \sim \mathrm{Pois}(\rho_p)$ with $\rho_p = \theta_{pk} \pi_k \beta_k^{n,t}$ denote the fictitious count of word p in the nth document at time t assigned to topic k disregarding $r_k^{n,t}$. 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 $\propto p(r_k^{n,t} = 1) \prod_{p=1}^{P} \mathrm{Pois}(0; \rho_p)$
2. $\exists p, u_{pntk} > 0,\ r_k^{n,t} = 0$ with probability $\propto p(r_k^{n,t} = 0) \left(1 - \prod_{p=1}^{P} \mathrm{Pois}(0; \rho_p)\right)$
3. $\forall p, u_{pntk} = 0,\ r_k^{n,t} = 0$ with probability $\propto p(r_k^{n,t} = 0) \prod_{p=1}^{P} \mathrm{Pois}(0; \rho_p)$

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

- **RVM weights,** ω_{lk}. We introduce the auxiliary variables λ_{lk} such that

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

Let $\boldsymbol{\omega}_k = (\omega_{0k}, \ldots, \omega_{Lk})^T$ be the vector of RVM weights and $\tilde{\mathbf{r}}_k$ be the vector of augmentation variables for all all time stamps, and

$$K_{tk} = (1, K(t, m_1, \phi_k), \ldots, K(t, m_L, \phi_k))^T \tag{12}$$

be the vector of the evaluation of the RVM kernels for time t. Then, the conditional of $\boldsymbol{\omega}_k$ is given by

$$\boldsymbol{\omega}_k | \tilde{\mathbf{r}}_k, \ldots \sim \mathrm{N}(\xi, B) \tag{13}$$

where $B = (\mathrm{diag}(\lambda_{0k}, \ldots, \lambda_{Lk}) + K_{tk}^T \tilde{\mathbf{r}}_k)^{-1}$ and $\xi = B K_{tk}^T \tilde{\mathbf{r}}_k$.

- **RVM auxiliary variables,** $\tilde{r}_k^{n,t}$.

$$p(\tilde{r}_k^{n,t}|\ldots) \propto \begin{cases} \mathrm{N}(K_{tk}^T \boldsymbol{\omega}_k, 1)\mathbf{1}(\tilde{r}_k^{n,t} > 0), & \text{if } r_k^{n,t} = 1 \\ \mathrm{N}(K_{tk}^T \boldsymbol{\omega}_k, 1)\mathbf{1}(\tilde{r}_k^{n,t} < 0), & \text{if } r_k^{n,t} = 0 \end{cases} \tag{14}$$

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

- **RVM precisions,** λ_{lk}.

$$\lambda_{lk}|\ldots \sim \mathrm{Ga}\left(c_0 + \frac{1}{2}, d_0 + \frac{1}{2}\omega_{lk}^2\right) \tag{15}$$

- **RVM kernel widths,** ϕ_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,

$$p(\phi_k = \phi_m^*|\ldots) \propto \frac{1}{M} \prod_{t\in\mathcal{T}} \prod_{n=1}^{N_t} \Phi(p_{\phi_m^*}(t))^{r_k^{n,t}} (1 - \Phi(p_{\phi_m^*}(t)))^{1-r_k^{n,t}} \tag{16}$$

where we have denoted the thinning function as a function of ϕ^* 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 $Y = [y_p^{n,t}]$ as

$$\exp\left(\frac{1}{y^{\cdot,\cdot}_{\cdot}} \sum_{p=1}^{P} \sum_{t\in\mathcal{T}} \sum_{n=1}^{N_t} y_p^{n,t} \log \frac{\sum_{b=1}^{B} \sum_{k=1}^{K} \theta_{pk}^{(b)} \pi_k^{(b)} r_{n,t,k}{}^{(b)} \beta_{n,t,k}^{(b)}}{\sum_{b=1}^{B} \sum_{p=1}^{P} \sum_{k=1}^{K} \theta_{pk}^{(b)} \pi_k^{(b)} r_{n,t,k}{}^{(b)} \beta_{n,t,k}^{(b)}}\right) \tag{17}$$

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

References

Albert, J.H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *JASA*, 88(422):669–679, 1993.

Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *JMLR*, 1:211–244, 2001.

Zhou, M., Hannah, L. A., Dunson, D. B., and Carin, L. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.

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