Supplementary material

Anonymous CVPR supplementary material

Paper ID 1354

1. Gibbs sampling for the HTGMM model

This supplementary material presents the derivation for inferring the hidden variables in the HTGMM of the submitted paper. We follow the notation of the paper if we do not particularly mention about it. To derive the inference procedure in HTGMM, we need to compute the joint pdf of the HTGMM. By considering the dependency among the random variables in the model, the joint pdf can be derived as

$$p(\boldsymbol{\phi},\boldsymbol{q},\boldsymbol{T},\boldsymbol{z},\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{\pi}|\alpha,\beta,\gamma,\tau,\mu_o,\kappa_o,S_o) = \prod_{k=1}^{K} p(\phi_k \mid \beta)p(q_k \mid \gamma)* \\ [\prod_{d=1}^{D}\{\prod_{l=1}^{N} p(T_l^{(d)} \mid \phi_{z_l^{(d)}}, q_{z_l^{(d)}})p(z_l^{(d)} \mid \theta^{(d)})\}p(\theta^{(d)} \mid \mu_{c^{(d)}},\tau)p(\bar{\theta}^{(d)} \mid \mu_{c^{(d)}}, S_{c^{(d)}})p(c^{(d)} \mid \pi)]* \\ \prod_{m=1}^{M} p(\mu_m \mid S_m,\mu_o,\kappa_o)p(S_m,S_o) * p(\pi \mid \alpha), \tag{1}$$

where the bold character denotes the set of the corresponding elements indexed as in the right-hand side of the equation (1). We note that $p(\bar{\theta}^{(d)} \mid \mu_{c^{(d)}}, S_{c^{(d)}}, \theta^{(d)}) = p(\bar{\theta}^{(d)} \mid \mu_{c^{(d)}}, S_{c^{(d)}})$ when $p(\bar{\theta}^{(d)} \mid \theta^{(d)}) = 1$ as mentioned in the paper. To infer the posterior probability for each hidden variable, we should compute an integral to marginalize other variables. However, this equation is not tractable because c, z are natural numbers and the domain of this pdf is not Lebesgue Integrable [1].

Therefore, we use gibbs sampling approach [2] to infer the hidden variables in the proposed HTGMM. The problem is that our model has many random variables and hence has a large sample space. Accordingly, it is required to reduce the sample space for efficient solving of the problem. To reduce the sample space, we will pre-marginalize out some random variables before the sampling, which is referred to as collapsed gibbs sampling [5]. To utilize the collapsed gibbs sampling method in the proposed HTGMM, we first divide our models into two blocks by using blocked gibbs sampling approach [7]. This method can be applied to our model because the set of variables $\{\boldsymbol{\phi},\boldsymbol{q},\boldsymbol{z}\}$ and $\{\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{\pi}\}$ are conditionally independent given $\boldsymbol{\theta}$. This independency can be easily checked by applying Bayes ball algorithm [8] to the proposed HTGMM. By using the blocked gibbs sampler, we infer the random variables through iteration of the following two steps: *step 1*; update $\{\boldsymbol{\phi},\boldsymbol{q},\boldsymbol{z},\boldsymbol{\theta}\}$ given $\{\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{\pi}\}$ and *step 2*; update $\{\boldsymbol{c},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{\pi}\}$ given $\{\boldsymbol{\phi},\boldsymbol{q},\boldsymbol{z},\boldsymbol{\theta}\}$. For each update step, we marginalize all the random variables except the intractable variables $\boldsymbol{z}$ and $\boldsymbol{c}$. We can analytically compute the marginalizing calculation because the random variables are designed to satisfy the conjugate prior by introducing the augmented variable $\bar{\theta}$ as described in the paper. The detailed description of the update procedure is given in the following.

In *step 1*, we will sample only the random variable set $\boldsymbol{z}$. For simplicity, in the below, we will use the redefined notation of $\boldsymbol{T}$ and $\boldsymbol{z}$ by eliminating the chunk index d, that is, $\boldsymbol{z} = \{z_1, ..., z_i, ..., z_{N_o}\}$ and $\boldsymbol{T} = \{T_1, T_2, ..., T_i, ..., T_{N_o}\}$, where N_o indicates the number of all trajectories, i.e., $N_o = N * D$. z_i indicates the assignment variable to assign a pattern index to T_i, and T_i is defined by using the words as $T_i = \{w_{i1}, w_{i2}, ..., , w_{in}, ..., w_{iN_i}\}$, where N_i indicates the number of words in T_i. The chunk including T_i is indexed by d_i. Then, by the Bayes' rule, the conditional posterior distribution for z_i is given by

$$P(z_i = j \mid \boldsymbol{z}_{-i}, \boldsymbol{T}) \propto p(T_i \mid z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{T}_{-i})P(z_i = j \mid \boldsymbol{z}_{-i}), \tag{2}$$

where $\boldsymbol{z}_{-i}$ is the set $\boldsymbol{z}$ excluding z_i, and this notation is also applied to the other variables in the same manner. The first term

in the right-hand side in (2) is a likelihood, and the second is a prior. For the first term, we have

$$p(T_i \mid z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{T}) = \int\int p(T_i|z_i = j, \phi_j, q_j) \cdot p(\phi_j, q_j \mid, \boldsymbol{z}_{-i}, \boldsymbol{T}_{-i}) d\phi_j dq_j \tag{3}$$

$$= \int\int \prod_{n=1}^{N_i} p(w_{in} \mid z_i = j, \phi_j) \prod_{m=1}^{N_i-1} p(w_{i(m+1)} \mid z_i = j, q_j(w_{im}, :)) \cdot p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{T}_{-i}) p(q_j|\boldsymbol{z}_{-i}, \boldsymbol{T}_{-i}) d\phi_j dq_j \tag{4}$$

$$= \int \prod_{n=1}^{N_i} p(w_{in} \mid z_i = j, \phi_j) p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{T}_{-i}) d\phi_j \int \prod_{m=1}^{N_i-1} p(w_{i(m+1)} \mid z_i = j, q_j(w_{im}, :)) p(q_j(w_{im}, :)|\boldsymbol{z}_{-i}, \boldsymbol{T}_{-i}) dq_j \tag{5}$$

$$= \int \prod_{n=1}^{N_i} p(w_{in} \mid z_i = j, \phi_j) p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in}) d\phi_j \int \prod_{m=1}^{N_i-1} p(w_{i(m+1)} \mid z_i = j, q_j(w_{im}, :)) p(q_j(w_{im}, :)|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-im}) dq_j \tag{6}$$

$$= \int \prod_{n=1}^{N_i} p(w_{in} \mid z_i = j, \phi_j) p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in}) d\phi_j * \tag{7}$$

$$[\prod_{m=1}^{N_i-1} \int p(w_{i(m+1)} \mid z_i = j, q_j(w_{im}, :)) p(q_j(w_{-im}, :)|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-im}) dq_j(w_{im}, :)]. \quad (\because \forall q_j(w_s, :) \perp q_j(w_l, :), s \neq l) \tag{8}$$

Note that ϕ and q are conditionally independent given T which has been applied to the procedure from (3) to (4). From Bayes' Rule, the second term in (7) becomes

$$p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in}) \propto p(\boldsymbol{w}_{-in} \mid \phi_j, \boldsymbol{z}_{-i}) p(\phi_j). \tag{9}$$

Since $p(\phi_j)$ is *Dirichlet*(β) and conjugate to $p(\boldsymbol{w}_{-in} \mid \phi_j, \boldsymbol{z}_{-i})$, the posterior $p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in})$ will be *Dirichlet*($\beta + n^{(w)}_{-in,j}$) as shown in the textbook [6], where $n^{(w)}_{-in,j}$ is the number of instances of word w assigned to pattern j, excluding w_{in}. The first term $p(w_{in} \mid z_i = j, \phi_j)$ in (7) is just $\phi^{(j)}_{w_{in}}$ according to the definition of HTGMM. Then, by following the multinomial-Dirichlet prior calculation given in the tutorial [3], we can easily complete the integral in (7) with

$$\int \prod_{n=1}^{N_i} p(w_{in} \mid z_i = j, \phi_j) p(\phi_j|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in}) d\phi_j = \prod_{n=1}^{N_i} \frac{n^{(w)}_{-in,j} + \beta}{n^{(\cdot)}_{-in,j} + W\beta}, \tag{10}$$

where W is the total number of words. $n^{(\cdot)}_{-in,j}$ is the total number of instances of all the words in $\boldsymbol{w}$ assigned to pattern j, excluding w_{in}. We can compute the integral in (8) using the similar derivation. From Bayes' Rule, the second term in (8) becomes

$$p(q_j(w_{in}, :)|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in}) \propto p(\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in} \mid q_j(w_{in}, :)) p(q_j(w_{in}, :)). \tag{11}$$

Subsequently, from the tutorial [3], the posterior $p(q_j(w_{in}, :)|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-in})$ is *Dirichlet*($\gamma + n^{(w)}_{-in}(w_{in})$). The term $n^{(w)}_{-in,j}(w_{in})$ is the number of instances of word w assigned to the transition probability starting from w_{in} for pattern j, excluding the current word w_{in}. By following the same procedure in (10), the integral in (8) is computed as

$$[\prod_{m=1}^{N_i-1} \int p(w_{i(m+1)} \mid z_i = j, q_j(w_{im}, :)) p(q_j(w_{im}, :)|\boldsymbol{z}_{-i}, \boldsymbol{w}_{-im}) dq_j(w_{im}, :)] = \prod_{m=1}^{N_i-1} \frac{n^{(w)}_{-im}(w_{im}) + \gamma}{n^{(\cdot)}_{-im}(w_{im}) + W\gamma}, \tag{12}$$

where $n^{(\cdot)}_{-im}(w_{im})$ is the total number of instances of all the words assigned to the transition probability starting from w_{in} for pattern j, excluding the current word w_{in}. Therefore, from the (10),(12), the probability $p(T_i \mid z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{T})$ in (3) is derived as

$$p(T_i \mid z_i = j, \boldsymbol{z}_{-i}, \boldsymbol{T}) \propto \prod_{n=1}^{N_i} \frac{n^{(w)}_{-in,j} + \beta}{n^{(\cdot)}_{-in,j} + W * \beta} \prod_{m=1}^{N_i-1} \frac{n^{(w)}_{-im}(w_{im}) + \gamma}{n^{(\cdot)}_{-in}(w_{im}) + W\gamma}. \tag{13}$$

In addition, we can find $p(z_i = j \mid \boldsymbol{z}_{-i})$ in (2) with the same procedure as in (10). We have

$$\begin{aligned} P(z_i = j \mid \boldsymbol{z}_{-i}) &= \int P(z_i = j \mid \theta^{(d_i)}) p(\theta^{(d_i)} \mid \boldsymbol{z}_{-i}) d\theta^{(d_i)} \\ &= \frac{n_{t_{-i,j}}^{(d_i)} + \tau \mu_c(j)}{n_{t_{-i,\cdot}}^{(d_i)} + K\tau \sum_{k=1}^{K} \mu_c(k)}, \end{aligned} \tag{14}$$

when $c^{(d_i)} = c$, because $p(\theta^{(d_i)})$ is defined as $Dirichlet(\tau\mu_c)$. The term $n_{t_{-i,j}}^{(d_i)}$ is the total number of trajectories in chunk d_i assigned to pattern j, excluding the current one. Therefore, from the (13),(14), the posterior (2) is solved as

$$P(z_i = j \mid \boldsymbol{z}_{-i}, \boldsymbol{T}) \propto \{\prod_{n=1}^{N_i} \frac{n_{-in,j}^{(w)} + \beta}{n_{-in,j}^{(\cdot)} + W * \beta} \prod_{m=1}^{N_i - 1} \frac{n_{-im}^{(w)}(w_{im}) + \gamma}{n_{-in}^{(\cdot)}(w_{im}) + W\gamma}\} \{\frac{n_{t_{-i,j}}^{(d_i)} + \tau \mu_c(j)}{n_{t_{-i,\cdot}}^{(d_i)} + K\tau \sum_{k=1}^{K} \mu_c(k)}\}. \tag{15}$$

We highlight that this derivation is possible by employing the augment variable $\bar{\theta}$ of which prior is the Gaussian distribution $\mathcal{N}(\mu_c, S_c)$. If we naively define the prior of $\theta^{(d)}$ as $\mathcal{N}(\mu_c, S_c)$, the integral in (14) is intractable because the Gaussian distribution is not a conjugate prior for the multinomial $\theta^{(d)}$. However, since we employ $\bar{\theta}^{(d)}$ which is given by deterministic mapping from $\theta^{(d)}$ and make $\bar{\theta}^{(d)}$ have the Gaussian prior, we can let $\theta^{(d)}$ has Dirichlet prior satisfying the conjugate prior. In *step 2*, we compute update equation considering both $\theta^{(d)}$ and $\bar{\theta}^{(d)}$.

For *step 2*, we will sample only $c^{(d)}$, the assignment of the $\theta^{(d)}$, to infer the hidden variables $\{\boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{\pi}\}$. Similar to the equation (2), we compute the posterior distribution for $c^{(d)}$ as

$$\begin{aligned} P(c^{(d)} = c \mid \boldsymbol{c}_{-d}, \bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) &\propto P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) p(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta} \mid c^{(d)} = c, \boldsymbol{c}_{-d}) \\ &= P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) p(\bar{\theta}^{(d)}, \theta^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, \boldsymbol{\theta}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}) p(\bar{\boldsymbol{\theta}}_{-d}, \boldsymbol{\theta}_{-d} \mid c^{(d)} = c, \boldsymbol{c}_{-d}) \\ &\propto P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) (\bar{\theta}^{(d)}, \theta^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, \boldsymbol{\theta}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}). \end{aligned} \tag{16}$$

The equation (16) is further derived as

$$P(c^{(d)} = c \mid \boldsymbol{c}_{-d}, \bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) \propto P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) p(\bar{\theta}^{(d)}, \theta^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, \boldsymbol{\theta}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}) \tag{17}$$

$$\propto P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) p(\bar{\theta}^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, \boldsymbol{\theta}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}, \theta^{(d)}) p(\theta^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, \boldsymbol{\theta}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}) \tag{18}$$

$$\propto P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) p(\bar{\theta}^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}) p(\theta^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}), \quad (\because p(\bar{\theta}^{(d)} | \theta^{(d)}) = 1). \tag{19}$$

By using the same derivation step with (10), the first term in (19) is given by

$$P(c^{(d)} = c \mid \boldsymbol{c}_{-d}) = \frac{n_{m-d,c} + \alpha}{n_{m-d,(\cdot)} + M\alpha}, \tag{20}$$

Since $\bar{\theta}^{(d)}$ is drawn from Gaussian distribution, the second term in (19) is equivalent to Gaussian posterior distribution. Accordingly, by following the tutorial [4, 6], the second term is given as

$$p(\bar{\theta}^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}) = \zeta(\bar{\theta}^{(d)} \mid \mu_{-d,c}, \frac{\kappa_n + 1}{\kappa_n(\upsilon_n - D + 1)} S_{-d,c}, \upsilon_n - D + 1), \tag{21}$$

where $\zeta(\cdot)$ is standard-t distribution. The $\mu_{-d,c}, S_{-d,c}, \kappa_n$ and υ_n are given by

$$\begin{aligned} \mu_{-d,c} &= \frac{\kappa_o \mu_o + \sum_{d=1}^{D} \bar{\theta}^{(d)} I(c^{(d)} = c)}{\kappa_n}, \\ \kappa_n &= \kappa_o + \sum_{d=1}^{D} I(c^{(d)} = c), \\ \upsilon_n &= \upsilon_o + \sum_{d=1}^{D} I(c^{(d)} = c), \\ S_{-d,c} &= S_o + S_c + \kappa_o \mu_o \mu_o^T - \kappa_n \mu_{-d,c} \mu_{-d,c}^T, \\ S_c &= \sum_{d=1}^{D} \bar{\theta}^{(d)} \bar{\theta}^{(d)^T} I(c^{(d)} = c), \end{aligned} \tag{22}$$

where $I(\cdot)$ is an indicator function. As defined in our paper, the third term in (19) is Dirichlet distribution over $\tau\mu_{-d,c}$ and so given as

$$p(\theta^{(d)} \mid \bar{\boldsymbol{\theta}}_{-d}, c^{(d)} = c, \boldsymbol{c}_{-d}) = Dirichlet(\theta^{(d)} \mid \tau\mu_{-d,c}). \tag{23}$$

Therefore, from (20),(21),(23), the posterior equation (17) is solved as

$$P(c^{(d)} = c \,|\boldsymbol{c}_{-d}, \bar{\boldsymbol{\theta}}, \boldsymbol{\theta}) \propto \frac{n_{m-d,c} + \alpha}{n_{m-d,(\cdot)} + M\alpha} \cdot \zeta(\bar{\theta}^{(d)} \mid \mu_{-d,c}, \frac{\kappa_n + 1}{\kappa_n(\upsilon_n - D + 1)} S_{-d,c}, \upsilon_n - D + 1) \cdot Dirichlet(\theta^{(d)} \mid \tau\mu_{-d,c}). \tag{24}$$

By iteratively resampling $\boldsymbol{z}$ and $\boldsymbol{c}$ by the equations (15) and (24), we can infer the hidden variables of the proposed HT-GMM.

References

[1] R. G. Bartle. *The elements of integration and Lebesgue measure*. John Wiley & Sons, 2014. 1

[2] A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990. 1

[3] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. 2002. 2

[4] H. Kamper. Gibbs sampling for fitting finite and infinite gaussian mixture models. 2013. 3

[5] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994. 1

[6] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2, 3

[7] G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 291–317, 1997. 1

[8] D. Z. Wang, E. Michelakis, M. Garofalakis, and J. M. Hellerstein. Bayesstore: managing large, uncertain data repositories with probabilistic graphical models. *Proceedings of the VLDB Endowment*, 1(1):340–351, 2008. 1