Supplementary Materials for Bayesian Nonparametric Poisson Process Allocation for Time-Sequence Modeling

Hongyi Ding[*] **Mohammad Emtiyaz Khan**[†] **Issei Sato**[*†] **Masashi Sugiyama**[†*]
[*]The University of Tokyo, Japan [†]The RIKEN Center for AIP, Tokyo, Japan

A Evidence Lower Bound $\mathcal{L}_1(q)$

Using Jensen's inequality, we bound the marginal log likelihood of the observed sequence $\{\boldsymbol{y}_d\}$. Hereafter we omit hyper-parameters $a_0, b_0, \alpha, \boldsymbol{H}$ in $\ln p(Y; a_0, b_0, \alpha, \boldsymbol{H})$ for simplicity.

$$\begin{aligned}
\ln p(Y) &= \ln \Big[\int \Big(\prod_{d=1}^{D} p(\boldsymbol{y}_d|\theta_d, s_d, \boldsymbol{f}) p(s_d) p(\theta'_d) \Big) \\
&\quad \times \prod_{k=1}^{\infty} p(\boldsymbol{f}_{k,N}|\boldsymbol{f}_{k,M}) p(\boldsymbol{f}_{k,M}) d\boldsymbol{\theta'_d} d\boldsymbol{f} \Big] \\
&\geq \sum_{d=1}^{D} \mathbb{E} \ln p(\boldsymbol{y}_d|\theta_d, s_d, \boldsymbol{f}) + \sum_{d=1}^{D} \sum_{k=1}^{K-1} \mathbb{E} \ln p(\theta'_{dk}) \\
&\quad + \sum_{d=1}^{D} \mathbb{E} \ln p(s_d) + \sum_{k=1}^{K} \mathbb{E} \ln p(\boldsymbol{f}_{k,M}) \\
&\quad - \sum_{d=1}^{D} \sum_{k=1}^{K-1} \mathbb{E} \ln q(\theta'_{dk}) - \sum_{k=1}^{K} \mathbb{E} \ln q(\boldsymbol{f}_{k,M}) \triangleq \mathcal{L}_0(q).
\end{aligned} \tag{1}$$

First we introduce a lemma (Paisley, 2010).

Lemma 1. *(Paisley, 2010) Let $\{X_k\}_{k=1}^K$ be a set of positive random variables, then*

$$\mathbb{E} \ln \Big(\sum_{k=1}^{K} X_k \Big) \geq \ln \Big(\sum_{k=1}^{K} \exp(\mathbb{E} \ln X_k) \Big). \tag{2}$$

or equivalently if $X_k = \exp(Y_k)$ where Y_k is a random variable, then

$$\mathbb{E} \ln \Big(\sum_{k=1}^{K} \exp(Y_k) \Big) \geq \ln \Big(\sum_{k=1}^{K} \exp(\mathbb{E} Y_k) \Big). \tag{3}$$

Proof. The function $\ln(\cdot)$ is concave. Using an auxiliary probability vector, $(p_1, \ldots, p_K)$, where $p_k > 0$ and $\sum_{k=1}^K p_k = 1$, it follows from Jensens inequality that

$$\mathbb{E} \ln \Big(\sum_{k=1}^{K} X_k \Big) = \mathbb{E} \ln \Big(\sum_{k=1}^{K} p_k \frac{X_k}{p_k} \Big) \geq \sum_{k=1}^{K} p_k \mathbb{E} \ln \Big(\frac{X_k}{p_k} \Big) \tag{4}$$

Taking derivatives with respect to $\{p_k\}$, we have

$$p_k = \frac{\exp(\mathbb{E} \ln X_k)}{\sum_{v=1}^{K} \exp(\mathbb{E} \ln X_v)} \tag{5}$$

Inserting this back, we obtain the desired bound. □

Using Lemma 1, we could further bound the first term to allow for a practical variational inference. This result is the same as the one obtained by following the methodology in LPPA (Lloyd et al., 2016).

$$\begin{aligned}
&\mathbb{E} \ln p(\boldsymbol{y}_d|\theta_d, s_d, \boldsymbol{f}) \\
&= \sum_{n=1}^{N_d} \Big(\ln \eta_d + \mathbb{E} \ln \sum_{k=1}^{\infty} \exp(\ln \theta_{dk} + \ln f_k^2(t)) \Big) \\
&\quad - \eta_d \int_{\mathcal{T}} \mathbb{E} \sum_{k=1}^{\infty} \theta_{dk} f_k^2(s) ds \qquad (6) \\
&\geq \sum_{n=1}^{N_d} \Big(\ln \eta_d + \ln \sum_{k=1}^{\infty} \exp(\mathbb{E} \ln \theta_{dk} + \mathbb{E} \ln f_k^2(t)) \Big) \\
&\quad - \eta_d \int_{\mathcal{T}} \mathbb{E} \sum_{k=1}^{\infty} \theta_{dk} f_k^2(s) ds. \qquad (7)
\end{aligned}$$

Using Equation (7), we implicitly collapse the indicator variables and obtain a lower bound of ELBO:

$$\begin{aligned}
\mathcal{L}_1(q) &\triangleq \sum_{n=1}^{N_d} \Big(\ln \eta_d + \ln \sum_{k=1}^{\infty} \exp(\mathbb{E} \ln \theta_{dk} + \mathbb{E} \ln f_k^2(t)) \Big) \\
&\quad - \eta_d \int_{\mathcal{T}} \mathbb{E} \sum_{k=1}^{\infty} \theta_{dk} f_k^2(s) ds + \sum_{d=1}^{D} \sum_{k=1}^{K-1} \mathbb{E} \ln \frac{p(\theta'_{dk})}{q(\theta'_{dk})} \\
&\quad + \sum_{d=1}^{D} \mathbb{E} \ln p(s_d) + \sum_{k=1}^{K} \mathbb{E} \ln \frac{p(\boldsymbol{f}_{k,M})}{q(\boldsymbol{f}_{k,M})}.
\end{aligned} \tag{8}$$

Now $q(f_{k,N}) = \mathcal{N}(\tilde{u}_k, \tilde{B}_k)$, where

$$\begin{aligned}
\tilde{u}_k &= \kappa_{k,NM} \kappa_{k,MM}^{-1} \mu_k, \\
\tilde{B}_k &= \kappa_{k,NN} - \kappa_{k,NM} \kappa_{k,MM}^{-1} \kappa_{k,MN} \\
&\quad + \kappa_{k,NM} \kappa_{k,MM}^{-1} \Sigma_k \kappa_{k,MM}^{-1} \kappa_{k,MN}
\end{aligned}$$

And the expectation parts in Equation (8) can be computed as:

$$\mathbb{E}\ln p(\theta'_{dk}) = \ln\alpha + (\alpha-1)\mathbb{E}[\ln(1-\theta'_{dk})], \tag{9}$$

$$\mathbb{E}\ln q(\theta'_{dk}) = \ln\frac{\Gamma(\tau_{dk,0}+\tau_{dk,1})}{\Gamma(\tau_{dk,0})\Gamma(\tau_{dk,1})} + (\tau_{dk,1}-1)\mathbb{E}[\ln(1-\theta'_{dk})] + (\tau_{dk,0}-1)\mathbb{E}[\ln\theta'_{dk}], \tag{10}$$

$$\mathbb{E}\ln p(s_d) = a_0\ln b_0 - \ln\Gamma(a_0) + (a_0-1)\ln\eta_d - b_0\eta_d, \tag{11}$$

$$\mathbb{E}\ln\frac{p(\boldsymbol{f}_{k,M})}{q(\boldsymbol{f}_{k,M})} = \frac{1}{2}\ln\frac{|\Sigma_k|}{|\kappa_{k,MM}|} + \frac{m}{2} - \frac{1}{2}tr\Big(\kappa^{-1}_{k,MM}(\Sigma_k + (\mu_k-g)(\mu_k-g)^T)\Big), \tag{12}$$

$$\mathbb{E}[\ln f_k^2(t_n^d)] = -G(-\frac{\tilde{u}^2_{k,n}}{2\tilde{B}_{k,nn}}) - C + \ln(\frac{\tilde{B}_{k,nn}}{2}), \tag{13}$$

$$\int_{\mathcal{T}}\mathbb{E}[f_k^2(s)]ds = \gamma|\mathcal{T}| - tr(\kappa^{-1}_{k,MM}\Psi_k) + tr(\kappa^{-1}_{k,MM}\Psi_k\kappa^{-1}_{k,MM}(\Sigma_k+\mu_k\mu_k^T)), \tag{14}$$

$G(x), x\le 0$ is calculated by a precomputed multi-resolution look-up table. C is a constant and $\Psi_k \in \mathbb{R}^{M\times M}, \Psi_{k,ij} = \int_{\mathcal{T}}\kappa_k(t_i,x)\kappa_k(x,t_j)dx$. Ψ_k is determined by the kernel hyper-parameter in κ_k and the region $\mathcal{T}$.

The expectation with regard to beta distribution is:

$$\mathbb{E}[\ln(1-\theta'_{dk})] = \psi(\tau_{dk,1}) - \psi(\tau_{dk,0}+\tau_{dk,1}),$$
$$\mathbb{E}[\ln(\theta'_{dk})] = \psi(\tau_{dk,0}) - \psi(\tau_{dk,0}+\tau_{dk,1}).$$

After adding augmented Lagrangian penalty function, the modified evidence lower bound is:

$$L_{\boldsymbol{v_i}}(\Phi,\boldsymbol{w_i}) \triangleq \mathcal{L}_1(q) - \sum_{k=1}^K w_{ik}\Big(\int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds - A\Big) - \sum_{k=1}^K\frac{v_{ik}}{2}\Big(\int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds - A\Big)^2. \tag{15}$$

A.1 Details of Derivatives

Based on the modified evidence lower bound in Equation (15), we could derive the parameter learning method.

- η_d. We list the term related to η_d in Equation (15) first.

$$L_{\eta_d} \triangleq N_d\ln\eta_d - \eta_d\int_{\mathcal{T}}\sum_{k=1}^K\mathbb{E}\Big(\theta_{dk}f_k^2(s)\Big)ds - \eta_d b_0 + (a_0-1)\ln\eta_d.$$

Obviously, there is a closed form update for η_d

$$\eta_d = \frac{N_d + a_0 - 1}{b_0 + \int_{\mathcal{T}}\sum_{k=1}^K\mathbb{E}\Big(\theta_{dk}f_k^2(s)\Big)ds}.$$

- $\tau_{dk,0}, \tau_{dk,1}$. We list the term related to these parameters in Equation (15) first

$$L_{\tau_{dk}} \triangleq \sum_{n=1}^{N_d}\Big[\ln\sum_{k=1}^{\infty}\exp\Big(\mathbb{E}_q[\ln\theta_{dk}] - \mathbb{E}_q[\ln f_k^2(t_n^d)]\Big)\Big] - \eta_d\int_{\mathcal{T}}\mathbb{E}\sum_{k=1}^{\infty}\theta_{dk}f_k^2(s)ds + \Big(\ln\frac{\Gamma(\tau_{dk,0})\Gamma(\tau_{dk,1})}{\Gamma(\tau_{dk,0}+\tau_{dk,1})} - (\tau_{dk,0}-1)\mathbb{E}\ln\theta'_{dk} + (\alpha-\tau_{dk,1})\mathbb{E}\ln(1-\theta'_{dk})\Big).$$

Let

$$L_{dnk} \triangleq \exp\Big(\mathbb{E}_q[\ln\theta_{dk}] + \mathbb{E}_q[\ln f_k^2(t_n^d)]\Big) = \exp\Big(\psi(\tau_{dk,0}) + \sum_{l=1}^{k-1}\psi(\tau_{dl,1}) - \sum_{l=1}^{k}\psi(\tau_{dl,0}+\tau_{dl,1}) + \mathbb{E}_q[\ln f_k^2(t_n^d)]\Big),$$
$$V_k \triangleq \int_{\mathcal{T}}\mathbb{E}f_k^2(s)ds$$

There is no closed form update for these variables, we use coordinate ascent method.

$$\frac{\partial L_{\tau_{dk}}}{\partial\tau_{dk,0}} = -\eta_d\Big(V_k\frac{\partial[\theta_{dk}]}{\partial\tau_{dk,0}} + \sum_{l=k+1}^K V_l\frac{\partial[\theta_{dl}]}{\partial\tau_{dk,0}}\Big) - \Big(\tau_{dk,0} - 1 - \sum_{n=1}^{N_d}\frac{L_{dnk}}{\sum_{v=1}^K L_{dnv}}\Big)\psi'(\tau_{dk,0}) + \Big(\tau_{dk,0} - 1 + \tau_{dk,1} - \alpha - \sum_{n=1}^{N_d}\frac{\sum_{v=k}^K L_{dnv}}{\sum_{v=1}^K L_{dnv}}\Big) \times\psi'(\tau_{dk,0}+\tau_{dk,1}),$$

$$\frac{\partial L_{\tau_{dk}}}{\partial\tau_{dk,1}} = -\eta_d\Big(V_k\frac{\partial[\theta_{dk}]}{\partial\tau_{dk,1}} + \sum_{l=k+1}^K V_l\frac{\partial[\theta_{dl}]}{\partial\tau_{dk,1}}\Big) - \Big(\tau_{dk,1} - \alpha - \sum_{n=1}^{N_d}\frac{\sum_{v=k+1}^K L_{dnv}}{\sum_{v=1}^K L_{dnv}}\Big)\psi'(\tau_{dk,1}) + \Big(\tau_{dk,0} - 1 + \tau_{dk,1} - \alpha - \sum_{n=1}^{N_d}\frac{\sum_{v=k}^K L_{dnv}}{\sum_{v=1}^K L_{dnv}}\Big) \times\psi'(\tau_{dk,0}+\tau_{dk,1}).$$

where we have

$$\frac{\partial[\theta_{dk}]}{\partial\tau_{dk,0}} = \frac{\tau_{dk,1}}{(\tau_{dk,0}+\tau_{dk,1})^2}\prod_{l=1}^{k-1}\frac{\tau_{dl,1}}{\tau_{dl,0}+\tau_{dl,1}},$$

$$\frac{\partial[\theta_{dk}]}{\partial\tau_{dk,1}} = -\frac{\tau_{dk,0}}{(\tau_{dk,0}+\tau_{dk,1})^2}\prod_{l=1}^{k-1}\frac{\tau_{dl,1}}{\tau_{dl,0}+\tau_{dl,1}},$$

$$\frac{\partial[\theta_{dl}]}{\partial\tau_{dk,0}} = -\frac{\tau_{dl,0}}{\tau_{dl,0}+\tau_{dl,1}}\frac{\tau_{dk,1}}{(\tau_{dk,0}+\tau_{dk,1})^2} \times\prod_{v=1,v\neq k}^{l-1}\frac{\tau_{dv,1}}{\tau_{dv,0}+\tau_{dv,1}},$$

$$\frac{\partial[\theta_{dl}]}{\partial\tau_{dk,1}} = \frac{\tau_{dl,0}}{\tau_{dl,0}+\tau_{dl,1}}\frac{\tau_{dk,0}}{(\tau_{dk,0}+\tau_{dk,1})^2} \times\prod_{v=1,v\neq k}^{l-1}\frac{\tau_{dv,1}}{\tau_{dv,0}+\tau_{dv,1}}.$$

- $\{\Sigma_k, \mu_k\}$. Take μ_k for an example.

$$\frac{\partial L_{\phi_k}}{\partial\mu_k} = \sum_{d=1}^{D}\Big(\sum_{n=1}^{N_d}\frac{1}{\sum_{v=1}^{K}L_{dnv}}\frac{\partial L_{dnk}}{\partial\mu_k}\Big) - \Big(w_{ik}+v_{ik}(V_k-A)+\sum_{d=1}^{D}\eta_d\mathbb{E}[\theta_{dk}]\Big)\frac{\partial V_k}{\partial\mu_k} + \frac{\partial}{\partial\mu_k}\Big[\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}\ln|\kappa_{k,MM}| - \frac{1}{2}tr\Big(\kappa_{k,MM}^{-1}(\Sigma_k+(\mu_k-g)(\mu_k-g)^T)\Big)\Big].$$

Hyper-parameter part: We could update the hyper-parameters in a similar way.

- Gaussian process hyper-parameters $\kappa_{k,MM}, \sigma$. Similar to that in $\{\Sigma_k, \mu_k\}$.
- Beta distribution prior α.

$$L_\alpha \triangleq D(K-1)\ln\alpha + (\alpha-1)\sum_{d=1}^{D}\sum_{k=1}^{K-1}(\psi(\tau_{dk,1}) - \psi(\tau_{dk,0}+\tau_{dk,1})).$$

Then we have a closed form update for α.

$$\alpha = \frac{D(K-1)}{\sum_{d=1}^{D}\sum_{k=1}^{K-1}\Big(\psi(\tau_{dk,1}+\tau_{dk,0})-\psi(\tau_{dk,1})\Big)}. \tag{16}$$

A.2 Proof of Upper Bound

Theorem 1. *Each optimization problem is upper bounded.*

$$L_{\boldsymbol{v_i}}(\Phi, \boldsymbol{w_i}) \leq \ln p(Y) + \sum_{k=1}^{K}\frac{w_{ik}^2}{2v_{ik}}, \quad i\in\mathbb{N}^+.$$

Proof. $\mathcal{L}_1(q)$ can be easily bounded by variational inference framework

$$\mathcal{L}_1(q) \leq \ln p(Y)$$

Let $h_{ik} = \int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds - A$, and then we have

$$\sum_{k=1}^{K}w_{ik}\Big(\int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds - A\Big) + \sum_{k=1}^{K}\frac{v_{ik}}{2}\Big(\int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds - A\Big)^2 = \sum_{k=1}^{K}(w_{ik}h_{ik}+\frac{v_{ik}}{2}h_{ik}^2) \geq \sum_{k=1}^{K}\frac{w_{ik}^2}{2v_{ik}}$$

Combining these two parts finishes the proof. □

A.3 A Bias When Using Lemma 1

Although the bound in Lemma 1 is rather tight, it can still add a bias which may lead to the over-shrinking phenomenon in the model. We illustrate the bias through the following simple model.

$$Y_1 = X_1^2,\ Y_2 = X_2^2,\ X_1\sim\mathcal{N}(2,1),\ X_2\sim\mathcal{N}(2,4),$$

where $\mathcal{N}(\cdot)$ is the normal distribution. Using Lemma 1, we can arrive the following inequality:

$$\begin{aligned}\mathcal{L}_{left} &\triangleq \mathbb{E}_{p(Y_{1:2})}\ln\Big(wY_1+(1-w)Y_2\Big) \\ &\geq \ln\Big(w\exp(\mathbb{E}\ln Y_1)+(1-w)\exp(\mathbb{E}\ln Y_2)\Big) \\ &\triangleq \mathcal{L}_{right},\ w\in[0,1].\end{aligned} \tag{17}$$

We vary the value of w and plot $\mathcal{L}_{right}$ and $\mathcal{L}_{left}$. The result is given in Figure 1. We can see that the for $\mathcal{L}_{right}$ the optimal is $w=1$ while for $\mathcal{L}_{left}$ the optimal is obviously a mixture of two components. This is because the logarithm function will punish values which are closer to zero harder. Since Y_2 has a large variance, there will be a large proportion of samples near zero which makes the corresponding $\mathbb{E}\ln Y_2$ smaller and less favorable. This bias in the inequality may account for the shrinkage in LPPA and BaNPPA.

B Test Likelihood

In LPPA, the allocation matrix Θ is treated as hyper-parameters and all the parameters are $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{H}, \Theta\}$. Let $\Phi = \{\boldsymbol{H}, \Theta\}$. In variational inference we use the variational distribution $q(\boldsymbol{f};\Phi)$ to approximate the posterior $p(\boldsymbol{f}|Y_{train};\Phi)$. The test likelihood can be

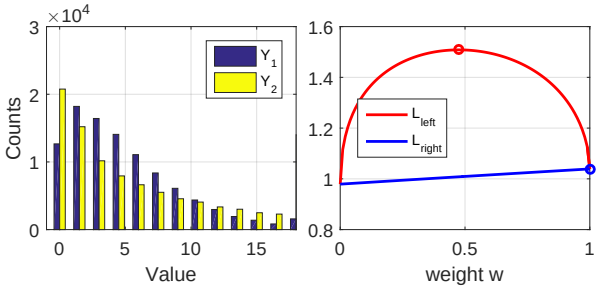

Figure 1: Bias in the inference with lower bound. Left: The histogram of Y_1 and Y_2. Right: $\mathcal{L}_{left}$ (Blue) versus $\mathcal{L}_{right}$ (Red) and the round marker indicates the maximum of the curve.

lower-bounded as follows.

$$\begin{aligned}
&\ln p(Y_{test}|Y_{train};\Phi) = \ln \int p(Y_{test}|\boldsymbol{f};\Phi)p(\boldsymbol{f}|Y_{train};\Phi)d\boldsymbol{f}\\
&\approx \ln \int p(Y_{test}|\boldsymbol{f};\Phi)q(\boldsymbol{f};\Phi)d\boldsymbol{f}\\
&\geq \int q(\boldsymbol{f};\Phi)\ln\frac{p(Y_{test}|\boldsymbol{f};\Phi)q(\boldsymbol{f};\Phi)}{q(\boldsymbol{f};\Phi)}d\boldsymbol{f}\\
&= \mathbb{E}_q \ln p(Y_{test}|\boldsymbol{f};\Phi)\\
&\geq \sum_{d=1}^{D}\sum_{n=1}^{N_d^{\text{test}}} \ln \sum_{k=1}^{K}\theta_{dk}\exp\left[\mathbb{E}_q(\ln f_k^2(t_n^d))\right]\\
&- \sum_{d=1}^{D}\sum_{k=1}^{K}\theta_{dk}\int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds \triangleq \mathcal{L}_{test}. \qquad (18)
\end{aligned}$$

In BaNPPA, all the parameters to be optimized are $\{\boldsymbol{\eta},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{H},a_0,b_0,\alpha\}$. Let $\Phi = \{\boldsymbol{H},a_0,b_0,\alpha\}$. However, if we follow the same deduction as LPPA, we will not arrive at a fair comparison since the inequality in Equation (18) is different in principle for LPPA and BaNPPA, and therefore, we draw L samples from variational distribution $q(\boldsymbol{s},\theta_d;a_0,b_0,\alpha)$ for $\boldsymbol{s},\theta_d$ and then follow the lower bound in Equation (18).

$$\begin{aligned}
&\mathbb{E}_q \ln p(Y_{test}|\boldsymbol{s},\boldsymbol{\Theta},\boldsymbol{f};\Phi)\\
&= \int q(\boldsymbol{s},\boldsymbol{\Theta},\boldsymbol{f};\Phi)\ln p(Y_{test}|\boldsymbol{s},\boldsymbol{\Theta},\boldsymbol{f};\Phi)d\boldsymbol{s}d\boldsymbol{\Theta}d\boldsymbol{f}\\
&\approx \frac{1}{\tilde{L}}\sum_{l=1}^{\tilde{L}}\int q(\boldsymbol{f};H)\ln p(Y_{test}|\boldsymbol{s}_l,\boldsymbol{\Theta}_l,\boldsymbol{f};H)d\boldsymbol{f}\\
&\geq \frac{1}{\tilde{L}}\sum_{l=1}^{\tilde{L}}\Big(\sum_{d=1}^{D}\sum_{n=1}^{N_d^{\text{test}}}\ln\Big(s_{l,d}\sum_{k=1}^{K}\theta_{l,dk}\exp\left[\mathbb{E}_q(\ln f_k^2(t_n^d))\right]\Big)\\
&- \sum_{d=1}^{D}s_{l,d}\sum_{k=1}^{K}\theta_{l,dk}\int_{\mathcal{T}}\mathbb{E}_q[f_k^2(s)]ds\Big). \qquad (19)
\end{aligned}$$

C Additional Experiment Results

C.1 Details of the Data Sets

- **Synthetic dataset**.
  A) In $\lambda_d(t) = s_d\sum_{k=1}^{4}\theta_{dk}\tilde{f}(t;\psi_k),\ t\in[0,60]$.

$$\begin{aligned}
s_d &\sim \text{Gamma}(2,3),\\
\theta_d &\sim \text{Dirichlet}(1.2,1,0.8,0.6),\\
\tilde{f}(t;\psi_k) &\propto \exp\Big(-\frac{(t-15+10k)^2}{10}\Big)\\
&+\exp\Big(-\frac{(t-55+10k)^2}{10}\Big).
\end{aligned}$$

  Each $\tilde{f}(t;\psi_k)$ is either a Gaussian distribution or a mixture of two Gaussian distributions normalized by its integral.
  B) In $\lambda_d(t) = s_d\sum_{k=1}^{6}\theta_{dk}f_k(t)$.

$$\begin{aligned}
s_d &\sim \text{Gamma}(2,3),\\
\theta_d &\sim \text{Dirichlet}(1.2,1,0.8,0.6,0.5,0.5),\\
\tilde{f}(t;\psi_k) &\propto \exp\Big(-\frac{(t-15+10k)^2}{10}\Big)\\
&+\exp\Big(-\frac{(t-75+10k)^2}{10}\Big).
\end{aligned}$$

  Each $\tilde{f}(t;\psi_k)$ is either a Gaussian distribution or a mixture of two Gaussian distributions normalized by its integral. We use the rejection sampling method for the inhomogeneous Poisson process to generate the time sequences.

- **citation dataset**. Two examples with different citation patterns are given in Figure 3.

C.2 The Comparison of the Train Likelihood

The comparison of the train likelihood $\mathcal{L}_{train}$ is given in Figure 2. We can notice that for LPPA, the train likelihood keeps increasing when we increase K. This is also a sign of over-fitting.

C.3 Computation Time

We plot the change of the training likelihood in one trial in Figure 4. For total computational complexity, both BaNPPA-NC and BaNPPA take more computation time but are still comparable to LPPA. Two reasons account for this fact. One is that there are more parameters to be optimized in BaNPPA and BaNPPA-NC and the other is that BaNPPA potentially has an infinite number of problems to be solved. In Figure 4, we can notice that the training likelihood for BaNPPA and the training likelihood for BaNPPA-NC stabilize rather quickly. This is because we use Equation (19) to calculate the likelihood and there are no divergence terms in it.

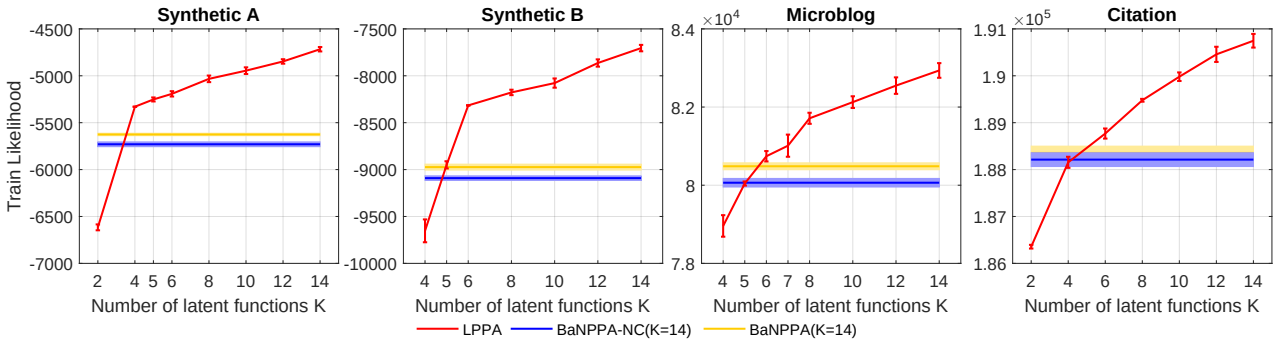

Figure 2: The comparison of the train likelihood for three algorithms. For LPPA, we change the number of latent functions K. For BaNPPA/BaNPPA-NC, we fix $K = 14$ and optimize the hyper-parameter α using the variational expectation-maximization. Error bars and shaded area represent the 95% confidence intervals.

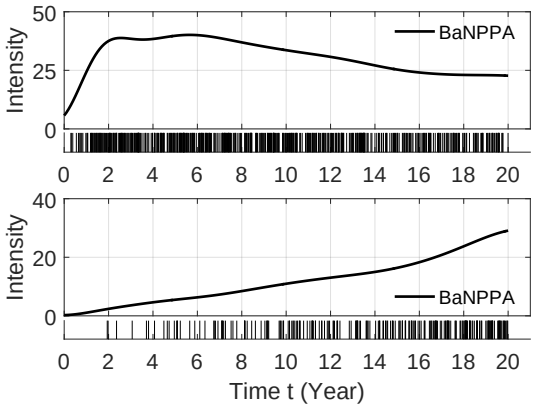

Figure 3: **Citation data set**. Top: A paper which slowly gets citation and becomes popular many years later. Bottom: A paper which quickly gets citation after being published. Smooth lines are the mean intensity function inferred from LPPA and BaNPPA. Small bars is the time of each citation. The x-axis indicates the time in year after publication.

C.4 Synthetic Data Sets with a Relatively Large K

We add three more synthetic data set with a larger K.

C) We sample 200 sequences from $\lambda_d(t) = s_d \sum_{k=1}^{6} \theta_{dk} \tilde{f}(t;\psi_k)$, where s_d, θ_d are drawn from Dirichlet distribution and Gamma distribution.

$$s_d \sim \text{Gamma}(2,3),$$
$$\boldsymbol{\theta}_d \sim \text{Dir}(0.8, 0.4, 0.2, 0.2, 0.2, 0.2).$$

We use $\tilde{f}(t;\psi_k) = \exp(-(t-15+10k)^2/10)$, $k = 1,\ldots,6$, $t \in [0,60]$ as basis intensity functions.

D) We sample 200 sequences from $\lambda_d(t) = s_d \sum_{k=1}^{8} \tilde{f}(t;\psi_k)$, where s_d, θ_d are drawn from Dirichlet distribution and Gamma distribution.

$$s_d \sim \text{Gamma}(2,3),$$
$$\boldsymbol{\theta}_d \sim \text{Dir}(0.8, 0.4, 0.4, 0.2, 0.2, 0.2, 0.1, 0.1).$$

We use $\tilde{f}(t;\psi_k) \propto \exp(-(t-15+10k)^2/10)$, $k = 1,\ldots,8$, $t \in [0,80]$ as basis intensity functions.

E) We sample 200 sequences from $\lambda_d(t) = s_d \sum_{k=1}^{10} \tilde{f}(t;\psi_k)$, where s_d, θ_d are drawn from Dirichlet distribution and Gamma distribution.

$$s_d \sim \text{Gamma}(2,3),$$
$$\boldsymbol{\theta}_d \sim \text{Dir}(0.8, 0.6, 0.4, 0.4, 0.4, 0.2, 0.2, 0.2, 0.1, 0.1).$$

We use $\tilde{f}(t;\psi_k) \propto \exp(-(t-15+10k)^2/10)$, $k = 1,\ldots,10$, $t \in [0,100]$ as basis intensity functions.

In the experiment, we fix the hyper-parameter a_0 and b_0 and the length-scale hyper-parameters in all $\kappa_{k,MM}$ to be 4.3081 (Close to the half of the span of $\tilde{f}(t;\psi_k)$). This means we only optimize the mixture weights and the variational distribution $q(m,S)$ for Gaussian processes.

We vary the hyper-parameter $\alpha = [1.1, 2, 3, 4, 5]$. The result is given in Figure. We can see that BaNPPA-NC tends to over-shrink the components even when $\alpha = 5$ and gets a worse result.

References

Lloyd, C., Gunter, T., Osborne, M., Roberts, S., and Nickson, T. (2016). Latent point process allocation. In *Artificial Intelligence and Statistics*, pages 389–397.

Paisley, J. (2010). Two useful bounds for variational inference. Technical report, Technical report, Department of Computer Science, Princeton University, Princeton, NJ.

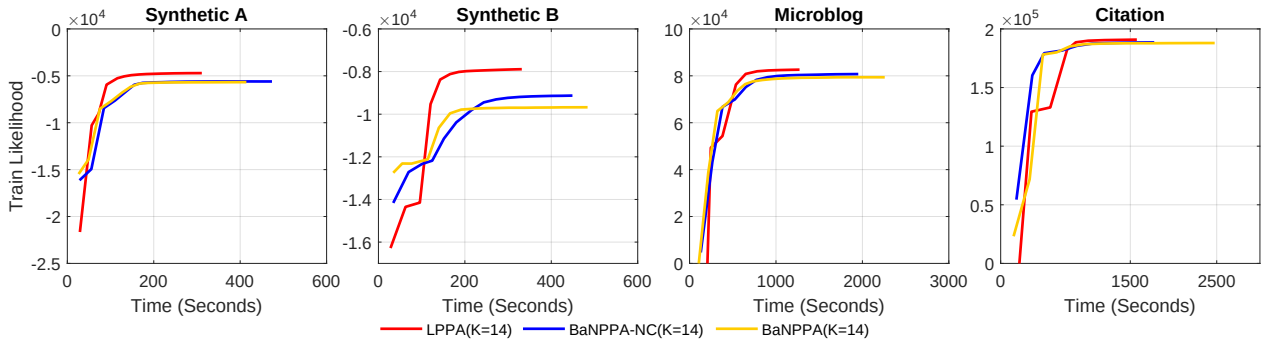

Figure 4: The comparison of the training likelihood versus time for four data sets (K=14) when optimizing the hyper-parameter α. The result of one trial is shown.

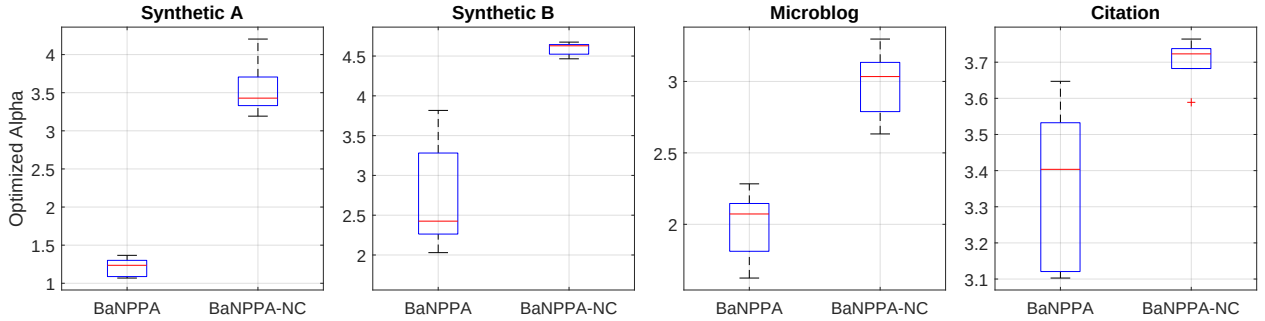

Figure 5: The comparison of the optimized α for four data sets (K=14) when optimizing the hyper-parameter α .

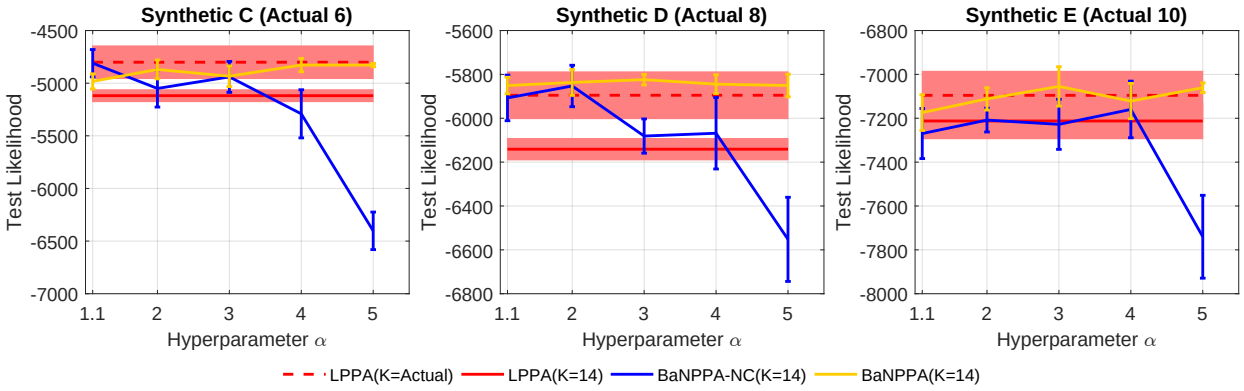

Figure 6: The comparison of the test likelihood for three additional data sets (K=14) when fixing the hyper-parameter $\alpha = [1.1, 2, 4, 6, 8]$. Error bars and shaded area represent the 95% confidence intervals.