Supplementary Materials

1 Proof of Proposition 1

Proof: We first consider the fixed-directional jump. In the following, we calculate the transition probability $\mathbb{P}\left((\boldsymbol{\theta}_j, \mathcal{M}_j) \mid (\boldsymbol{\theta}_i, \mathcal{M}_i)\right)$. Suppose the transition from $(\boldsymbol{\theta}_i, \mathcal{M}_i)$ to $(\boldsymbol{\theta}_j, \mathcal{M}_j)$ is achieved by some jumping distance r, then we have

$$\boldsymbol{\theta}_j = \boldsymbol{u} + r(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{u}}) \quad \text{and} \quad \boldsymbol{v} = \boldsymbol{\theta}_i + r(\widehat{\boldsymbol{v}} - \widehat{\boldsymbol{\theta}}_i).$$

It follows that

$$\begin{aligned}
\mathbb{P}\left((\boldsymbol{\theta}_j, \mathcal{M}_j) \mid (\boldsymbol{\theta}_i, \mathcal{M}_i)\right) &= \int q_i(\boldsymbol{u}) \times \frac{p_j\left(\boldsymbol{v}, \boldsymbol{\theta}_j, \mathcal{M}_j \mid \boldsymbol{y}\right)}{\sum_{k=1}^m p_j(\boldsymbol{v}^{(k)}, \boldsymbol{\theta}_j^{(k)}, \mathcal{M}_j \mid \boldsymbol{y})} \times \prod_{k=1}^{m-1} p(r^{(k)}) p(r) dr^{(1)} \cdots dr^{(m-1)} \\
&\quad \times \min\left\{1, \quad \frac{\sum_{k=1}^m p_j\left(\boldsymbol{v}^{(k)}, \boldsymbol{\theta}_j^{(k)}, \mathcal{M}_j \mid \boldsymbol{y}\right)}{\sum_{k=1}^m p_i\left(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{u}^{(k)}, \mathcal{M}_i \mid \boldsymbol{y}\right)}\right\} \\
&= \int \min\left\{ \left[\sum_{k=1}^m p_j\left(\boldsymbol{v}^{(k)}, \boldsymbol{\theta}_j^{(k)}, \mathcal{M}_j \mid \boldsymbol{y}\right)\right]^{-1}, \left[\sum_{k=1}^m p_i\left(\boldsymbol{\theta}_i^{(k)}, \boldsymbol{u}^{(k)}, \mathcal{M}_i \mid \boldsymbol{y}\right)\right]^{-1} \right\} \\
&\quad \times \prod_{k=1}^{m-1} p(r^{(k)}) dr^{(1)} \cdots dr^{(m-1)} \times p\left(\boldsymbol{\theta}_j, \mathcal{M}_j \mid \boldsymbol{y}\right) q_i(\boldsymbol{u}) q_j(\boldsymbol{v}) p(r),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\theta}_j^{(k)} &= \boldsymbol{u} + r^{(k)}(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{u}}), \quad \boldsymbol{v}^{(k)} = \boldsymbol{\theta}_i + r^{(k)}(\widehat{\boldsymbol{v}} - \widehat{\boldsymbol{\theta}}_i), \\
\boldsymbol{\theta}_i^{(k)} &= \boldsymbol{v} - r^{(k)}(\widehat{\boldsymbol{v}} - \widehat{\boldsymbol{\theta}}_i), \quad \boldsymbol{u}^{(k)} = \boldsymbol{\theta}_j - r^{(k)}(\widehat{\boldsymbol{\theta}}_j - \widehat{\boldsymbol{u}}).
\end{aligned}$$

Note that when we jump back from $(\boldsymbol{v}, \boldsymbol{\theta}_j)$ to $(\boldsymbol{\theta}_i, \boldsymbol{u})$, we flip the sign of the jumping direction, that is, changing the jumping direction to $(\widehat{\boldsymbol{u}} - \widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{v}})$, and keep the same jumping distance r. This is the reason why we do not require $p(r)$ to be symmetric and centered at 0. Since the Jacobian between $(\boldsymbol{v}, \boldsymbol{\theta}_j)$ and $(\boldsymbol{\theta}_i, \boldsymbol{u})$ is simply 1, and the multiple integral is symmetric with respect to the index i and j, the transition kernel satisfies the detailed balance condition, thus leaves $p\left(\boldsymbol{\theta}_k, \mathcal{M}_k \mid \boldsymbol{y}\right)$ invariant.

For the adaptive-directional jump, we first standardize the jumping direction, that is, setting $\boldsymbol{e} = (\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})/||(\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})||$, so that the jumping distance is independent to the current

state $(\boldsymbol{\theta}_i, \boldsymbol{u})$. Besides, because the jumping direction always points to the mode of the augmented posterior distribution $p(\boldsymbol{v}, \boldsymbol{\theta}_j, M_j \mid \boldsymbol{y})$, when we jump back from $(\boldsymbol{v}, \boldsymbol{\theta}_j)$ to $(\boldsymbol{\theta}_i, \boldsymbol{u})$, we should flip the sign of the jumping distance as we won't flip the sign of the jumping direction. Consequently, $p(r)$ is required to be symmetric and centered at 0.

The proof of the reversibility of the transition kernel equipped with the adaptive-directional jump follows similarly (thus is omitted) as the case of the fixed-directional jump, but requires an additional calculation of the Jacobian, which is detailed as below. Suppose the transition from $(\boldsymbol{\theta}_i, \mathcal{M}_i)$ to $(\boldsymbol{\theta}_j, \mathcal{M}_j)$ is achieved by some jumping distance r, then we have

$$\boldsymbol{\theta}_j = \boldsymbol{u} + r\frac{\widehat{\boldsymbol{\theta}}_j - \boldsymbol{u}}{||(\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})||} \quad \text{and} \quad \boldsymbol{v} = \boldsymbol{\theta}_i + r\frac{\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i}{||(\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})||}.$$

For notational convenience, we define $\boldsymbol{x} \in \mathbb{R}^{d_i+d_j}$ as follows. For $k \in [d_j]$ and $l \in [d_i]$, let

$$x_k = \frac{(\widehat{\theta}_{jk} - u_k)}{||(\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})||^{3/2}}, \quad x_{d_j+l} = \frac{(\widehat{v}_l - \theta_{il})}{||(\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})||^{3/2}}.$$

Then we have

$$\left|\frac{\partial(\boldsymbol{\theta}_j, \boldsymbol{v})}{\partial(\boldsymbol{u}, \boldsymbol{\theta}_i)}\right| = \det\left((1 - r||\boldsymbol{x}||^2)I_{d_j} + r\boldsymbol{x}\boldsymbol{x}^\intercal\right) = \left[1 - \frac{r}{||(\widehat{\boldsymbol{v}} - \boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_j - \boldsymbol{u})||}\right]^{d_i+d_j-1}.$$

2 More details of the Log-Gaussian Cox process

We provide more details of the Log-Gaussian Cox process discussed in Section 5.1. We first detail the adaptive SMC algorithm in Algorithm 1. We then examine the effect of the threshold c used in the flat histogram criterion. Figure 1 shows that the WL mixture method is robust to the choice of c in the region $[0.1, 0.3]$.

3 Variational approximation for the Bayesian Lasso

Algorithm 2 details the coordinate ascent variational inference (CAVI) algorithm used in constructing the surrogate distribution for the Bayesian Lasso example (see Section 5.2). We use the mean-field variational family, in which we assume

$$q(\beta_j) \sim N(m_j, s_j^2), \quad q(\eta_j) \sim N(\phi_j, \zeta_j^2) \quad \text{for } j \in [p], \quad \text{and} \quad q(\xi) \sim N(u, v^2).$$

The definition of $\boldsymbol{\beta}, \boldsymbol{\eta}$ and ξ can be found in Section 5.2.

References

Zhou, Y., A. M. Johansen, and J. A. D. Aston (2016). Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics 25*(3), 701–726.

Algorithm 1 An adaptive sequential Monte Carlo (SMC) sampler (Zhou et al., 2016).

Input: proposal distribution $q(\boldsymbol{\theta})$, Markov kernels $\{K_t\}$.

1. Initialization.
   (a) Sample $\boldsymbol{\theta}_0^{(i)}$ from $q(\boldsymbol{\theta})$ for $i \in [n]$ independently.
   (b) Set $w_0^{(i)} = 1/n$ for $i \in [n]$. Set $\lambda_0 = 0$ and $t = 0$.
2. While $\lambda_t < 1$, iterate between the following steps.
   (a) Set $t \leftarrow t + 1$.
   (b) For some pre-specified $\kappa \in (0, 1)$, using binary search to find $\lambda_t \in (\lambda_{t-1}, 1]$ such that

$$n^{-1}\mathrm{CESS}_t(\lambda_t) = \frac{\left(\sum_{i=1}^n w_{t-1}^{(i)}(\gamma/q)(\boldsymbol{\theta}_{t-1}^{(i)})^{\lambda_t - \lambda_{t-1}}\right)^2}{\sum_{i=1}^n w_{t-1}^{(i)}(\gamma/q)(\boldsymbol{\theta}_{t-1}^{(i)})^{2(\lambda_t - \lambda_{t-1})}} = \kappa.$$

   If $n^{-1}\mathrm{CESS}_t(1) > \kappa$, set $\lambda_t = 1$ and $T = t$.
   (c) Compute the unnormalized weights $w_t^{(i)} = (\gamma_t/\gamma_{t-1})(\boldsymbol{\theta}_{t-1}^{(i)})$ for $i \in [n]$. The geometric sequence of the auxiliary distributions $\gamma_t(\boldsymbol{\theta})$ is defined as

$$\gamma_0(\boldsymbol{\theta}) = q(\boldsymbol{\theta}),\ \gamma_T(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}),\ \text{and}\ \gamma_t(\boldsymbol{\theta}) = q(\boldsymbol{\theta})^{\lambda_t}\gamma(\boldsymbol{\theta})^{1-\lambda_t}.$$

   (d) Compute $r_t = \sum_{i=1}^n w_{t-1}^{(i)} w_t^{(i)}$, which is an estimate of Z_t/Z_{t-1}. Z_t is the normalizing constant of γ_t.
   (e) Normalize the weights $\{w_t^{(i)}\}_{i\in[n]}$ to sum 1.
   (f) Compute the (normalized) effective sample size

$$\mathrm{ESS}_t = \frac{1}{n\sum_{i=1}^n \left(w_t^{(i)}\right)^2}.$$

   If $\mathrm{ESS}_t \le 0.5$, resample particles using systematic resampling, and set $w_t^{(i)} = 1/n$ for $i \in [n]$.
   (g) Move particles $\boldsymbol{\theta}_t^{(i)}$ according to the Markov kernel K_t for $i \in [n]$.

Output: normalizing constant estimate $\widehat{Z}_\gamma = \prod_{t=1}^T r_t$.

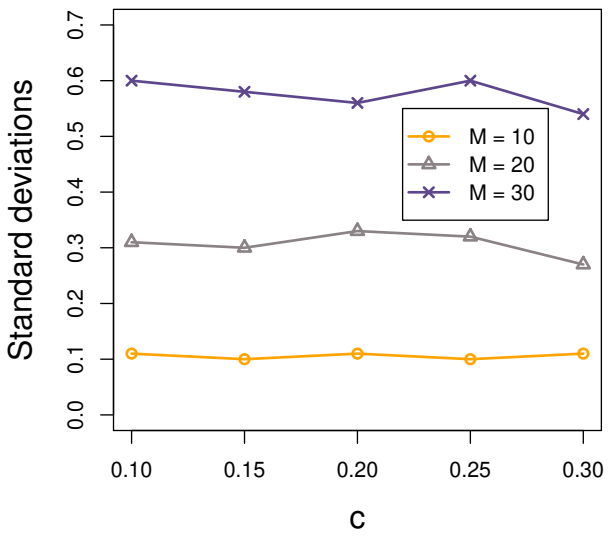

Figure 1: Demonstration of the effect of the threshold c used in the flat histogram criterion for the Log-Gaussian Cox process. The y-axis denotes the empirical standard deviation of the log normalizing constant estimates over 10 independent runs. The unit square is discretized into an $M \times M$ regular grid (see Section 5.1).

Algorithm 2 The CAVI updates for the Bayesian Lasso example.

1. Given $q(\beta_i)$ for $i \neq j$, $q(\boldsymbol{\eta})$ and $q(\xi)$, update $q(\beta_j)$.

$$(m_j, s_j) = \underset{(m,\ s>0)}{\arg\max} \left\{ e^{-u+v^2/2} \left[A_{-j} m - \frac{1}{2} \left((X^\intercal X)_{jj} + e^{-\phi_j + \zeta_j^2/2} \right) \left(m^2 + s^2 \right) \right] + \frac{1}{2} \log s^2 \right\},$$

   where $A_{-j} = \sum_{k=1}^{n} X_{kj} y_k - \sum_{k \neq j} (X^\intercal X)_{kj} m_k$.

2. Given $q(\eta_i)$ for $i \neq j$, $q(\boldsymbol{\beta})$ and $q(\xi)$, update $q(\eta_j)$.

$$(\phi_j, \zeta_j) = \underset{(\phi,\ \zeta>0)}{\arg\max} \left\{ \phi - \lambda^2 e^{\phi + \zeta^2/2} - (m_j^2 + s_j^2) e^{-u+v^2/2} e^{-\phi + \zeta^2/2} + \log \zeta^2 \right\}.$$

3. Given $q(\boldsymbol{\beta})$, $q(\boldsymbol{\eta})$, update $q(\xi)$.

$$(u, v) = \underset{(u,\ v>0)}{\arg\max} \left\{ -(n+p)u - B e^{-u+v^2/2} + \log v^2 \right\},$$

   where $B = \boldsymbol{y}^\intercal \boldsymbol{y} - 2\boldsymbol{y}^\intercal X m + \operatorname{tr}\left[(m m^\intercal + \operatorname{diag}(s_k^2)) X^\intercal X \right] + \sum_{k=1}^{p} (m_k^2 + s_k^2) e^{-\phi_k + \zeta_k^2/2}$.