Copy Multiple Features in INLA / InlaBru

September 11, 2023

1 Background

Watson et al. (2019) proposed a framework that jointly modeling the distribution of an environmental process and a site-selection process, where the environmental process can be spatial, temporal, or spatio-temporal. By sharing the random effects between the two process, the joint model can detect the preferential sampling effects.

We consider a spatio-temporal environmental process Y_{st}, $s \in \mathcal{S}$, $t \in \mathcal{T}$. The space-time point is defined $(s,t) \in \mathcal{S} \times \mathcal{T}$, where $\mathcal{S}$ denoting the spatial domain of interest and $\mathcal{T}$ the temporal domain. Spatial network designer specifies a set of time points $T \subset \mathcal{T}$ at which to observe Y and at each time $t \in T$, a finite subset of sites $S_t \subset \mathcal{S}$ at which to do so.

$R_{st} \in \{0,1\}$ is a binary response for the site selection process. A Bayesian model is introduced for the joint distribution of the response vector (Y_{st}, R_{st}). By sharing random effects across the two processes, the stochastic dependence (if any) between $Y_{s,t}$ and $R_{s,t}$ and be quantified.

2 The joint model

We let $Y_i(t)$ denote the spatio-temporal observation process at site i, that is at locations $s_i \in \mathcal{P} \subset \mathcal{S}$, at time $t \in T$. We let $R_i(t)$ denote the random selection indicator for site $s_i \in \mathcal{P}$ at time t. We let $t_1, \ldots, t_N$ denote the N observation times, and let $r_{i,j} \in \{0,1\}$ denote the realization of $R_i(t_j)$, for $i \in \{1, \ldots, M\}$, $j \in \{1, \ldots, N\}$, where $M = |\mathcal{P}|$. The general model framework is

$$\begin{aligned}
Y_{i,j} \,|\, R_{i,j} = 1 &\sim f_Y(\mu_{i,j}, \theta_Y), \quad f_Y \sim \text{density}, \\
g(\mu_{i,j}) = \eta_{i,j} &= x_{i,j}^T \gamma + \sum_{k=1}^{q_1} u_{i,j,k} \beta_k(s_i, t_j), \qquad (1) \\
R_{i,j} &\sim \mathsf{Bern}(p_{i,j}), \\
h(p_{i,j}) = \nu_{i,j} &= v_{i,j}^T \alpha + \sum_{\ell=1}^{q_2} d_\ell \sum_{k=1}^{q_1} w_{i,j,\ell,k} \beta_k(s_i, \phi_{i,\ell,k}(t_j)) \qquad (2) \\
&+ \sum_{m=1}^{q_3} w^\star_{i,j,m} \beta^\star_m(s_i, t_j), \\
\beta_k(s_i, t_j) &\sim \text{(possibly shared) latent effect with parameters } \theta_k, \\
&k \in \{1, \ldots, q_1\}, \\
\beta^\star_m(s_i, t_j) &\sim \text{site selection only latent effect with parameters } \theta^\star_m, \\
&m \in \{1, \ldots, q_3\}, \\
\Theta &= (\theta_Y, \alpha, \gamma, d, \theta_1, \ldots, \theta_{q_1}, \theta^\star_1, \ldots, \theta^\star_{q_3}) \sim \text{Priors}, \\
x_{i,j} \in \mathbb{R}^{p_1}, u_{i,j} &\in \mathbb{R}^{q_1}, v_{i,j} \in \mathbb{R}^{p_2}, W_{i,j} \in \mathbb{R}^{q_2 \times q_1}, w^{\star T}_{i,j} \in \mathbb{R}^{q_3}
\end{aligned}$$

In the linear predictor $\eta_{i,j}$, we include a linear combination of fixed covariates $x_{i,j}$ with a linear combination of q_1 latent effects $\beta_k(s_i, t_j)$. These q_1 random effects can include any combinations of spatially-correlated processes, temporally correlated processes, spatial temporal processes and IID random effects. Note that we include the additional fixed covariates $u_{i,j}$ to allow for spatially-varying coefficient models, as well as both random slopes and/or scaled random effects.

As for the site selection process $R_{i,j}$, the linear predictor $\nu_{i,j}$ may also include a linear combination of fixed covariates $v_{i,j}$ with a linear combination of latent effects. In particular, the latent effects appearing in the observation process $Y_{i,j}$ are allowed to exist in the linear predictor of the selection process $R_{i,j}$. The matrix $W_{i,j}$ is fixed beforehand, and allow for q_2 linear combinations of the latent effects from the $Y_{i,j}$ process to be copied across. The parameter vector d determines the degree to which each shared latent effect affects the R process and therefore measure the magnitude and direction of stochastic dependence between the two models term-by-term. These q_3 latent effects are independent from the $Y_{i,j}$ process to exist in the linear predictor.

2.1 A specific model for black smoke data in British

Watson et al. (2019) introduced a specific model in the family to model the black smoke data in British. Let $t_j^\star$ denote the jth time-scaled observations that lie in the interval $[0, 1]$.

The model for the observation process is

$$\begin{aligned}
Y_{i,j} \mid R_{i,j} &\sim \mathcal{N}(\mu_{i,j}, \sigma_\epsilon^2) \\
\mu_{i,j} &= \gamma_0 + \gamma_1 t_j^\star + \gamma_2 (t_j^\star)^2 + b_{0,i} + b_{1,i} t_j^\star + \beta_0(s_i) + \beta_1(s_i) t_j^\star + \beta_2(s_i)(t_j^\star)^2 \qquad (3) \\
[\beta_k(s_1), \dots, \beta_k(s_m)]^T &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma(\zeta_k)) \quad \text{for } k \in \{0, 1, 2\}, \quad \Sigma(\zeta_k) = \text{Matern}(\zeta_k) \\
[b_{0,i}, b_{1,i}] &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_b), \quad \Sigma_b = \begin{pmatrix} \sigma_{b,1}^2 & \rho_b \\ \rho_b & \sigma_{b,2}^2 \end{pmatrix}, \\
\theta &= (\sigma_\epsilon^2, \gamma, \zeta_k, \sigma_{b,1}^2, \rho_b) \sim \text{Priors}.
\end{aligned}$$

The model for site-selection process is

$$\begin{aligned}
R_{i,j} &\sim \mathsf{Bern}(p_{i,j}) \\
\operatorname{logit} p_{i,1} &= \alpha_{0,0} + \alpha_1 t_1^\star + \alpha_2(t_1^\star) + \beta_1^\star(t_1) \\
&\quad + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
&\quad + d_b[b_{0,i} + b_{1,i}(t_1^\star)] \\
&\quad + d_\beta[\beta_0(s_i) + \beta_1(s_i) t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2], \\
\text{for} j \neq 1 \quad \operatorname{logit} p_{i,j} &= \alpha_{0,1} + \alpha_1 t_j^\star + \alpha_2 (t_j^\star)^2 + \beta_1^\star t_j \\
&\quad + \alpha_{ret} r_{i,(j-1)} + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
&\quad + d_b[b_{0,i} + b_{1,i}(t_1^\star)] \\
&\quad + d_\beta[\beta_0(s_i) + \beta_1(s_i) t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2], \qquad (4) \\
I_{i,j} &= \mathbb{1}\left[\left(\sum_{\ell \neq i} r_{\ell,j-1} \mathbb{1}(\|s_i - s_\ell\| < c)\right) > 0\right], \\
[\beta_0^\star(s_1), \dots, \beta_0^\star(s_m)]^T &\sim \mathcal{N}(0, \Sigma(\zeta_R)), \Sigma(\zeta_R) = \text{Matern}(\zeta_R), \\
[\beta_1^\star(t_1), \dots, \beta_1^\star(t_T)]^T &\sim \text{AR1}(\rho_a, \sigma_a^2), \\
\theta_R = [\alpha, d_b, d_\beta, \rho_a, \sigma_a^2, \zeta_R] &\sim \text{Priors}
\end{aligned}$$

3 The implementation in INLA / inlabru

To implement the preferential sampling model defined by Eq. (1) and Eq. (2) in INLA, or **inlabru**, we are supposed to specify two models. One for the observation process in the Gaussian family and one for the site selection process in the Bernoulli family. In particular, we want to share some linear combinations of latent factors between the observation model and the site selection model:

$$\sum_{\ell=1}^{q_2} d_\ell \sum_{k=1}^{q_1} w_{i,j,\ell,k} \beta_k(s_i, \phi_{i,\ell,k}(t_j)), \tag{5}$$

where we have q_2 set of factors to share, each being a linear combination of latent factors from the observation model. While both INLA and **inlabru** allow copying factors between models, each factor (component in **inlabru**) must be copied separately and therefore introduce one new scale parameter for each copied factor (by setting $fixed = FALSE$).

In our model, however, we only want one scale parameter for each combination of factors, i.e., we only want q_2 scale parameters when copying factors combinations in Eq. (5). For example in the specific model Eq. (3) and Eq. (4) for the black smoke data, the two copied factor combinations in the site selection process Eq. (4)are

$$d_b[b_{0,i} + b_{1,i}(t_1^\star)] \quad \text{and} \quad d_\beta[\beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2],$$

where d_b and d_β are two scale parameters. However, this is not directly achievable using the *copy* feature in INLA or **inlabru**. If we use the *copy* feature to copy each latent factor separately, there will be five (instead of two) new scale parameters introduced at each site and time point.

To copy the whole set of factors in implementing the model for black smoke data, Watson's code introduced two auxiliary factors and two auxiliary Gaussian models in addition to the original observation model and the site selection model:

$$\begin{aligned} 0 &= -C_b + [b_{0,i} + b_{1,i}(t_1^\star)] \\ 0 &= -C_\beta + [\beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2] \end{aligned}$$

where C_b and C_β are auxiliary factors. The latent factors C_b and C_β work as duplicates of the two factor combinations:

$$C_b = b_{0,i} + b_{1,i}(t_1^\star) \quad \text{and} \quad C_\beta = \beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2$$

To achieve this, the precision parameter of the two factors C_b and C_β are set to be ≈ 0 and the precision parameter of the two Gaussian auxiliary models are set to be $\approx \infty$. Also, these factors $b_{0,i}, b_{1,i}(t_1^\star), \beta_0(s_i), \beta_1(s_i)t_{j-1}^\star, \beta_2(s_i)(t_{j-1}^\star)^2$ are copied from the observation model Eq. (3) to the two auxiliary models with $fixed = TRUE$. The choice of the precision parameters make sure that C_b and C_β are precise copies of the factor combinations.

Now we need to fit four, instead of two models in INLA or **inlabru**. In general, the total number of models is $2 + q_2$. With the auxiliary factor C_b and C_β, instead of copying these two linear combinations of factors in the site selection model Eq. (4), now we copy C_b and C_β. By setting $fixed = FALSE$, two new scale parameters, d_b and d_β are added to C_b and C_β respectively.

4 The package for preferential sampling

We are currently working to develop an R package for the preferential sampling model defined by Eq. (1) and Eq. (2). The purpose of this package to facilitate the spatial prediction using the proposed preferential sampling model.

Given the specific form of the model (one observation model and one site-selection model), we would like to restrict the input of the user to simplify the API. In particular, we want the user to specify only formulas of the two models and provide the dataset. Given that the two models are both mixed effects models, we would like to use the syntax analogous to the that of the **lme4** package.

Internally, we want to convert the input of the user to proper models of **inlabru** and fit the model using **inlabru**. For example, one thing we might need to do is to introduce q_2 auxiliary factors and q_2 new models if the user-specified model wants to copy q_2 factor combinations from the observation model Eq. (1) to the site selection model Eq. (2).

References

Watson, J., Zidek, J. V., and Shaddick, G. (2019). A general theory for preferential sampling in environmental networks. *The Annals of Applied Statistics*, 13(4):2662 – 2700.