Copy Multiple Features in INLA / inlabru

September 17, 2023

1 The package for preferential sampling

We are currently working to develop an R package for the preferential sampling model proposed by Watson et al. (2019) which fits an observational model and a site selection model that shares latent factors with the observation model. The purpose of this package is to facilitate spatial prediction using the proposed preferential sampling model.

Since the joint model is restricted to two mixed effects models (one for the observation process and one for the site-selection process), 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 in addition to 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**.

2 A preferential sampling model for black smoke data in British

We consider a spatio-temporal environmental process Y_{st}, $s \in \mathcal{S}$, $t \in \mathcal{T}$, 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. Watson et al. (2019) proposed one such preferential sampling model to analyze 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

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

The model for site-selection process is

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

The latent effects appearing in the observation process $Y_{i,j}$ are allowed to exist in the linear predictor of the selection process $R_{i,j}$. In particular, the two linear combinations of the latent effects, $b_{0,i} + b_{1,i}(t_1^\star)$ and $\beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2$, from the $Y_{i,j}$ process are copied across. The parameters d_b and d_β determine 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.

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. Also, we want to share two linear combinations of latent factors between the observation model and the site selection model:

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

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 two scale parameters d_b and d_β for these two linear combinations of factors:

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

where d_b and d_β are two scale parameters. This is not directly achievable using the *copy* feature in INLA or **inlabru**, and 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.

3.1 An alternative approach using auxiliary models

To copy the linear combinations of factors in implementing the model for black smoke data, Watson et al. (2019) introduced two auxiliary factors and two auxiliary Gaussian models in addition to the original joint model:

$$0 = -C_b + [b_{0,i} + b_{1,i}(t_1^\star)] \qquad (3)$$

$$0 = -C_\beta + [\beta_0(s_i) + \beta_1(s_i)t_{j-1}^\star + \beta_2(s_i)(t_{j-1}^\star)^2] \qquad (4)$$

where C_b and C_β are auxiliary latent factors. These individual 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 separately from the observation model Eq. (1) to the two auxiliary models, Eq. (3) and Eq. (4), with the argument $fixed = TRUE$.

By setting the precision parameter of the two factors C_b and C_β to be ≈ 0 and setting the precision parameter of the two Gaussian auxiliary models to be $\approx \infty$, the latent factors C_b and C_β duplicate of the two factor combinations:

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

Given the two auxiliary models, the new model for site-selection process copies C_b and C_β from Eq. (3) and Eq. (4) instead with the argument $fixed = FALSE$:

$$\begin{aligned}
& \operatorname{logit} p_{i,1} = \alpha_{0,0} + \alpha_1 t_1^\star + \alpha_2(t_1^\star) + \beta_1^\star(t_1) \\
& \qquad + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
& \qquad + d_b C_b + d_\beta C_\beta, \\
\text{for} j \neq 1 \quad & \operatorname{logit} p_{i,j} = \alpha_{0,1} + \alpha_1 t_j^\star + \alpha_2(t_j^\star)^2 + \beta_1^\star t_j \\
& \qquad + \alpha_{ret} r_{i,(j-1)} + \alpha_{rep} I_{i,2} + \beta_0^\star(s_i) \\
& \qquad + d_b C_b + d_\beta C_\beta.
\end{aligned}$$

With the auxiliary models and factors, it is possible to copy the linear combination of factors without introducing too many scale parameters. However, this approach requires us to fit four, instead of two models in INLA(or **inlabru**), and in general, more auxiliary models and factors will be required if more linear combinations of factors need to be shared between the observation process and the site selection process.

4 Question

To simplify the API of our package so that users can use the preferential sampling model to make spatial predictions easily, we want to follow the syntax of the package **lme4** and ask users to only provide formulas of two mixed effects models. Inside the package, we need to convert the joint model to **inlabru**(or INLA) models. Since the approach used by Watson et al. (2019) requires one more additional model for each linear combination of factors to be copied, this increases the complexity of the code. So we wonder if there is more straightforward way to copy multiple / linear combination of factors across models in INLA or **inlabru**.

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.