Geostatistical variable selection formulation

Let $Y_i$, $n_i$ and $p_i$ be the number of infected individuals, the number of individuals screened, and the prevalence of infection at location $i$ ($i = 1, \ldots, N$). We assume that $Y_i$ arises from a Binomial distribution, i.e., $Y_i \sim \text{Bin}(n_i, p_i)$. Potential predictors $X_{ij}^{(b)}$, $j_b = 1, \ldots, J_b$ are divided into $B$ groups $b$, ($b = 1, \ldots, B$), where $B - 1$ groups contain predictors which are considered highly correlated with a Pearson coefficient $> 0.9$, while the $B$th group includes predictors that exhibit only moderate correlation with other potential predictors. In addition, potential predictors presenting a non-linear association to the infection risk in explanatory variables. In particular, we model a categorical temporal trend $T_{il}$ ($l = 1, \ldots, L$ categories), the potential predictors $X_{ij}^{(b)}$ and a spatial random effect $\varphi_i$ on the logit scale, such as:

$$
\text{logit}(p_i) = \beta_0 + \sum_{l=1}^{L} T_{il} \beta_{1l} + \sum_{b=1}^{B} \sum_{j_b=1}^{J_b} \alpha_{j_b} \sum_{l=1}^{L} \xi_{lj_b} X_{ij}^{(b)} + \varphi_i,
$$

where regression coefficients of potential predictors $X_{j_b}$ are defined as the product of an overall contribution $\alpha_{j_b}$ and the effect $\xi_{lj_b}$ of each of its elements (i.e., categories).

Within a Bayesian framework of inference, we assign a spike and slab prior (Scheipl et al., 2012; Chammartin et al., 2013a,b) to $\alpha_{j_b}$, which is a scaled normal mixture of inverse-gamma, that is $\alpha_{j_b} \sim \text{N}(0, \tau_{\alpha}^2)$, where $\tau_{\alpha}^2 \sim \gamma_{2\alpha}^{(b)} \text{IG}(a_{\tau}, b_{\tau}) + (1 - \gamma_{2\alpha}^{(b)})v_0 \text{IG}(a_{\tau}, b_{\tau})$, where $a_{\tau}$ and $b_{\tau}$ are fixed parameters of non-informative inverse-gamma distribution set to 5 and 25, respectively, while $v_0$ is a small constant set to 0.00025, shrinking $\alpha_{j_b}$ to zero when the predictor is excluded. The product of the two indicators $\gamma_1^{(b)}$ and $\gamma_2^{(b)} = (\gamma_2^{(b)}; \ldots, \gamma_2^{(b)})^T$ indicates the presence or absence of the predictors in the model. In particular, $\gamma_1^{(b)}$ determines the presence or absence of the group $b$ in the model and $\gamma_2^{(b)}$, allows selection of a single predictor within the group. A Bernoulli and a categorical prior distribution are assigned to $\gamma_1^{(b)}$ and $\gamma_2^{(b)}$, respectively, such as $\gamma_1^{(b)} \sim \text{Bern}(\Omega_1^{(b)})$ and $\gamma_2^{(b)} \sim \text{Cat}(J_b, \Omega_1^{(b)}, \ldots, \Omega_2^{(b)})$ with inclusion probabilities $\Omega_1^{(b)}$ and $\Omega_2^{(b)}$. To allow greater flexibility in estimating model size, these probabilities are considered as hyper-parameters having non-informative beta and Dirichlet distributions: $\Omega_1^{(b)} \sim \text{Beta}(1, 1)$, $\Omega_2^{(b)} = (\Omega_{21}^{(b)}, \ldots, \Omega_{2J_b}^{(b)})^T \sim \text{Dirichlet}(1, \ldots, 1)$. A mixture of two Gaussian distributions is assumed for $\xi_{lj_b}$, $\xi_{lj_b} \sim \text{N}(m_{lj_b}, 1)$, $m_{lj_b} \sim 1/2\delta_{l}(m_{lj_b}+1/2)/\delta_{l}(m_{lj_b})$, which shrinks $\xi_{lj_b}$ towards $|1|$ (multiplicative identity). For predictors moderately correlated, $\gamma_2^{(b)}$ is fixed to 1, while the effect of linear predictors is only defined by an overall contribution of $\alpha$. In addition, non-informative normal priors have been assigned to the constant $\beta_0$ and the effects $\beta_{1l}$ of the temporal trend: $\beta_0, \beta_{1l} \sim \text{N}(0, 100)$.

Large matrix computation cost in estimating this latent spatial process $\varphi$ is overcome with the predictive process estimation (Banerjee et al., 2008). In more details, $\varphi$ is estimated from a subset of 200 locations (knots) $\{k, \ldots, K\}$ with latent observations $\varphi^* = (\varphi_1^*, \ldots, \varphi_K^*)^T$, $\varphi^* \sim \text{MVN}(0, V\Sigma^*)$. $V\Sigma^*$ is the $KK$ variance-covariance matrix modelled by an isotropic exponential correlation function of distance, i.e., $\Sigma_{cd} = \sigma_2^2 \exp(-d_{cd})$, where $d_{cd}$ is the Euclidean distance between locations $c$ and $d$, $\sigma_2^2$ is the geographical variability, and $\rho$ controls the rate of correlation decay. Inverse gamma distribution $\sigma_2^2 \sim \text{IG}(2.01, 1.01)$ is chosen for the variance $\sigma_2^2$ and a gamma distribution is assumed for the spatial decay $\rho$, $\rho \sim \text{G}(0.01, 0.01)$. Spatial random effect $\varphi$ at original set of locations are predicted via the conditional mean $Q^{-1}\Sigma^*^{-1}\varphi^*$, where $Q = \text{Cov}(\varphi^*, \varphi)$ is a $NK$ matrix of the covariance function between the $K$ knots and the $N$ observed locations. Minimax space filling sampling (Johnson et al., 1990; Diggle and Lophaven, 2006) is used to select the knots using the cover.design routine in R (The R Foundation for Statistical Computing R v.3.0.2).

Geostatistical variable selection was run in JAGS through the rjags library of R (The R Foundation for Statistical Computing v.3.0.2) in JAGS 3.4.0 with on chain sampler and 40,000 iterations (including a burn-in of 10,000 iterations). Final 10,000 iterations were used to calculate models posterior probabilities.
and the subset of variables included in the models with the highest posterior probabilities identified the final models.

**Bayesian spatio-temporal model formulation**

Our Bayesian spatio-temporal model formulation follows the approach introduced by Cameletti et al. (2013). In particular, we define \( Y_{it}, p_{it} \) and \( n_{it} \) as the number of infected individuals, the number of individuals screened, and the prevalence of infection at location \( i (i = 1, \ldots, N) \) for time \( t (t = 1, \ldots, T) \), and we assume \( Y_{it} \) to be generated from a binomial distribution, i.e., \( Y_{it} \sim \text{Bin}(p_{it}, n_{it}) \). Prevalence of infection is then linearly regressed on the logit scale as follows: 

\[
\text{logit}(p_{it}) = X_{it}^T \beta + \varphi_{it},
\]

where \( X \) is the matrix of explanatory variables (including an intercept, a temporal trend, and the predictors selected by the variable selection), \( \beta \) is the regression coefficient vector, and \( \varphi \) is a spatio-temporally-structured random effect. We allow the spatio-temporal process \( \varphi_{it} \) to change in time with a first order autoregressive process (AR1), such as:

\[
\varphi_{it} = \omega_{i1} \quad \text{if } t = 1
\]

\[
\varphi_{i,t} = a\varphi_{i,t-1} + \omega_{i,t} \quad \text{if } t = 2, \ldots, T,
\]

with a temporal autoregressive coefficient \( a, |a| < 1 \) and a temporally independent spatially-structured effect \( \omega \) which is assumed to be multivariate normal with zero mean and spatio-temporal covariance function of the Matérn family:

\[
\text{Cov}(\omega_{i,t}, \omega_{j,t'}) = \begin{cases} 
0 & \text{if } t \neq t' \\
\sigma^2_{\omega} C(d_{ij}) & \text{if } t = t' \text{ for } i \neq j,
\end{cases}
\]

where \( \sigma^2_{\omega} \) is the variance of the structured effect \( \omega \), \( (\sigma^2_{\omega} = \text{Var}(\omega_{i,t})) \). The spatial correlation function \( C(d_{ij}) \) is function of the Euclidean distance between locations \( i \) and \( j \) \( (d_{ij}) \) and is defined by the Matérn function given by:

\[
C(d_{ij}) = \frac{1}{\Gamma(v)2^{(v-1)/2}} (kd_{ij})^{v} K_v(kd_{ij}),
\]

where \( K_v \) is the modified Bessel function of second kind and order \( v \) \( (v > 0) \), \( v \) is a smoothing parameter controlling the rate of correlation decay fixed to 1, and \( \kappa \) \( (\kappa > 0) \) is a scaling parameter. The spatial range is defined as the minimum distance at which spatial correlation between locations is less than 10% and is given by \( \sqrt{8v/\kappa} \).

**References**

Banerjee S, Gelfand AE, Finley AO, and Sang H (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(4) :825–848.

Cameletti M, Lindgren F, Simpson D, and Rue H (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *ASA Advances in Statistical Analysis*, pp. 1–23.

Chammartin F, Hörlimann E, Raso G, N’Goran EK, Utzinger J, and Vounatsou P (2013a). Statistical methodological issues in mapping historical schistosomiasis survey data. *Acta Tropica*, 128(2) :345–352.

Chammartin F, Scholte RGC, Malone JB, Bavia ME, Nieto P, Utzinger J, and Vounatsou P (2013b). Modelling the geographical distribution of soil-transmitted helminth infections in Bolivia. *Parasites & Vectors*, 6(1) :1–14.

Diggle PJ and Lophaven SÅ (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, 33(1) :53–64.

Johnson ME, Moore LM, and Ylvisaker D (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2) :131–148.
Scheipl F, Fahrmeir L, and Kneib T (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500): 1518–1532.