Introduction to STDGLM

Introduction

The `STDGLM` package provides a framework for fitting spatio-temporal dynamic generalized linear models. These models are useful for analyzing data that varies over both space and time, allowing for the incorporation of spatial and temporal dependencies in the modeling process. The package provides functions for fitting these models, as well as tools for visualizing and interpreting the results.

Installation

You can install the package from GitHub using the following command:

```
if (!requireNamespace("devtools", quietly = TRUE)) {
  install.packages("devtools")
}
devtools::install_github("czaccard/STDGLM")
```

Run the following command to load the package:

```
library(STDGLM)
```

Detailed Explanation on supported STDGLMs

Let p denote the number of *spatial units* (either georeferenced locations or areal units) where data is collected, and let T denote the number of *time points*. As for the current version of the package (`0.0.0.9000`), only Gaussian outcomes are supported. Specifically only dynamic linear models (DLMs) with the following **observation equation** can be handled:

$$y_{it} = x_{it}'\beta_{it} + z_{it}'\gamma + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2)$$

where:

- y_{it} is the response variable at spatial unit $i = 1, \dots, p$ and time $t = 1, \dots, T$,
- $x_{it} = (x_{1,it}, \dots, x_{J,it})'$ is a J-dimensional ($J \geq 1$) vector of covariates at spatial unit i at time t (an intercept may or may not be included here),
- $\beta_{it} = (\beta_{1,it}, \dots, \beta_{J,it})'$ is the state vector at time t at spatial unit i,
- z_{it} is a q-dimensional vector of covariates whose effects are constant (an intercept may or may not be included here),
- γ is a vector of non-varying coefficients,
- ϵ_{it} is the observation error at time t at spatial unit i.

The evolution of the state vector is described by the following **state equation**, which accounts for spatial correlations in the state vector:

$$\beta_{j,t} = F_{j,t}\beta_{j,t-1} + \eta_{j,t}, \quad \eta_{j,t} \sim N_p(0, \Sigma_{\eta,j}), \quad j = 1, \dots, J$$

where:

- $\beta_{j,t} = (\beta_{j,1t}, \dots, \beta_{j,pt})'$ is the state vector related to the j-th covariate $x_{j,t} = (x_{j,1t}, \dots, x_{j,pt})'$, at time t for all spatial units,
- $F_{j,t} = \phi_j^{(\mathsf{T})} I_p$ is a transition matrix,
- $\Sigma_{\eta,j}$ is the covariance matrix of the state evolution error $\eta_{j,t}$.

The transition matrix $F_{j,t}, j = 1, \dots, J$, is assumed to be a scalar multiple of the identity matrix. The parameter $\phi_j^{(\mathsf{T})}$ controls the temporal autocorrelation of the state vector.

The state evolution covariance matrix $\Sigma_{\eta,j}$ can be structured to reflect *spatial relationships*, e.g. by assuming an **exponential covariance function** if the data are point-referenced:

$$(\Sigma_{\eta,j})_{i\ell} = \rho_{j,2} \exp\left(-\frac{d_{i\ell}}{\rho_{j,2}}\right), \quad d_{i\ell} = \|s_i - s_\ell\|$$

where $\rho_{j,1}$ is the partial sill, $\rho_{j,2}$ is the range parameter, and $ d_{i }$ is the Euclidean distance between locations s_i and s_ℓ. At the moment, this is the only supported covariance structure for point-referenced data.

In this case, the evolution error $\eta_{j,t}$ is assumed to be a zero-mean Gaussian process with exponential covariance matrix parameteterized by $\rho_{j,1}$ and $\rho_{j,2}$, which we will denote as $\eta_{j,t} \sim GP(0, \rho_{j,1}, \rho_{j,2}; exp)$.

If the data are areal, a **proper conditional autoregressive (PCAR) covariance structure** is assumed:

$$\Sigma_{\eta,j} = \rho_{j,1} \left(D_w - \rho_{j,2} W\right)^{-1}$$

where W is a binary adjacency matrix, D_w is a diagonal matrix with row sums of W on the diagonal, and $\rho_{j,1}$ and $\rho_{j,2}$ are the conditional variance and autocorrelation parameters, respectively.

In this case, the evolution error $\eta_{j,t}$ follows a zero-mean PCAR process, and we will denote this as $\eta_{j,t} \sim PCAR(0, \rho_{j,1}, \rho_{j,2})$.

ANOVA Decomposition of the State Vector

The function `stdglm` allows for the decomposition of the state vector into components that can be interpreted as contributions from different sources of variability. Dropping the subscript j for the sake of simplicity, the state vector β_{it} is decomposed as follows:

$$\beta_{it} = \overline{\beta} + \beta_i^{(\mathsf{S})} + \beta_t^{(\mathsf{T})} + \beta_{it}^{(\mathsf{ST})}$$

where:

- $\overline{\beta}$ is the overall mean effect,
- $\beta_i^{(\mathsf{S})}$ is the spatial effect at spatial unit i,
- $\beta_t^{(\mathsf{T})}$ is the temporal effect at time t,
- $\beta_{it}^{(\mathsf{ST})}$ is the interaction effect between space and time at spatial unit i and time t.

Bayesian Hierarchical Structure

The Bayesian model is as follows, for $t = 1, \dots, T$ and $j = 1, \dots, J$:

$$\begin{aligned}
y_{it} &\sim N(x_{it}'\beta_{it} + z_{it}'\gamma, \sigma_\epsilon^2) \\
\beta_{j,t} &= 1_p\overline{\beta}_j + \beta_j^{(\mathsf{S})} + 1_p\beta_{j,t}^{(\mathsf{T})} + \beta_{j,t}^{(\mathsf{ST})} \\
\beta_j^{(\mathsf{S})} &= (\beta_{j,1}^{(\mathsf{S})}, \dots, \beta_{j,p}^{(\mathsf{S})})' \sim GP(0, \rho_{j,1}^{(\mathsf{S})}, \rho_{j,2}^{(\mathsf{S})}; exp) \text{ or } PCAR(0, \rho_{j,1}^{(\mathsf{S})}, \rho_{j,2}^{(\mathsf{S})}) \\
\beta_{j,t}^{(\mathsf{T})} &\sim N(\phi_j^{(\mathsf{T})}\beta_{j,t-1}^{(\mathsf{T})}, V_{\beta,j}^{(\mathsf{T})}) \\
\beta_{j,t}^{(\mathsf{ST})} &= (\beta_{j,1t}^{(\mathsf{ST})}, \dots, \beta_{j,pt}^{(\mathsf{ST})})' \sim GP(\phi_j^{(\mathsf{ST})}\beta_{j,t-1}^{(\mathsf{ST})}, \rho_{j,1}^{(\mathsf{ST})}, \rho_{j,2}^{(\mathsf{ST})}; exp) \text{ or } PCAR(\phi_j^{(\mathsf{ST})}\beta_{j,t-1}^{(\mathsf{ST})}, \rho_{j,1}^{(\mathsf{ST})}, \rho_{j,2}^{(\mathsf{ST})})
\end{aligned}$$

The model is completed with the following priors (again, dropping the subscript j, since these are common across j):

$$\begin{aligned}
\overline{\beta} &\sim N(0, V_\gamma) \\
\beta_0^{(\mathsf{T})} &\sim N(0, V_{\beta_0}) \\
\beta_0^{(\mathsf{ST})} &\sim N_p(0, V_{\beta_0} I_p) \\
\gamma &\sim N_q(0, V_\gamma) \\
\sigma_\epsilon^2 &\sim IG(a_\epsilon, b_\epsilon) \\
\rho_1^{(\mathsf{S})} &\sim IG(0.01, 0.01) \\
\rho_2^{(\mathsf{S})} &\sim U(a_\rho, b_\rho) \\
\rho_1^{(\mathsf{ST})} &\sim IG(0.01, 0.01) \\
\rho_2^{(\mathsf{ST})} &\sim U(a_\rho, b_\rho) \\
\phi^{(\mathsf{T})} &\sim TN_{(-1,1)}(0, 1) \\
\phi^{(\mathsf{ST})} &\sim TN_{(-1,1)}(0, 1) \\
V_\beta^{(\mathsf{T})} &\sim IG(a^{(\mathsf{T})}, b^{(\mathsf{T})})
\end{aligned}$$

where $TN_{(q,r)}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2 truncated to the interval (q, r). The hyperparameters a_ρ and b_ρ depend on the type of spatial data. If the data are point-referenced, they are set based on the minimum and maximum distances between points, respectively. If the data are areal, $a_\rho = 0.1$ and $b_\rho \to 1$.

Note that the spacetime-varying coefficients are assumed independent a priori across $j = 1, \dots, J$.

Efficient Inference and Identifiability

To build an efficient sampler, the algorithm proposed by Chan and Jeliazkov (2009) is used in conjuction with sparse matrix techniques.

To make the model identifiable, some **constraints** are imposed on the varying parameters at each MCMC iteration:

- Set $\sum_{t=1}^{T} \beta_t^{(\mathsf{T})} = 0$.
- Set $\sum_{i=1}^{p} \beta_i^{(\mathsf{S})} = 0$.
- Set $\sum_{i=1}^{p} \beta_{i,t}^{(\mathsf{ST})} = 0$ for each $t = 1, \dots, T$.
- Set $\sum_{t=1}^{T} \beta_{i,t}^{(\mathsf{ST})} = 0$ for each $i = 1, \dots, p$.

References

Chan, Joshua CC, and Ivan Jeliazkov. 2009. "Efficient Simulation and Integrated Likelihood Estimation in State Space Models." *International Journal of Mathematical Modelling and Numerical Optimisation* 1 (1-2): 101–20.