CHAPTER 6 - TIME AND SPACE

LECTURER: RICCARDO CORRADIN UNIVERSITY OF MILANO-BICOCCA

Several modern approaches in various fields, such as climatology, ecology, environmental health, real estate marketing, etc., face the task of analyzing structured data that are

- highly multivariate, with covariates and response variables;
- geographically referenced;
- temporally correlated.

Among these, time dependent models and spatial data analysis deals with observations that depend on a specific continuous domain, which either describe their observed time of realization or their dispersion over a coordinate set.

1 UNIVARIATE TIME-DEPENDENT DATA

Time-dependent real data appear quite often, since many quantities are observed together with time. In a Bayesian setting, we can also model time-dependent observations. We start considering data are real-valued. Hence, we consider a sequence of values $\{y_t\}_{t=1}^n$, which are ordered over time. Each datum lies in the same support, with $y_i \in \mathbb{Y} \subseteq \mathbb{R}$. Observations are not only time-dependent, but the sequence has also memory, since each value depends on previous observations.

More specifically, here we consider models belonging to the family of autoregressive model AR(p), of generic order p, where the current state is regressed on the past. Specifically, the model has memory until p time lags in the past. Hence, we consider model of the form

$$y_t = \phi_0 + \sum_{j=1}^{p} \phi_j y_{t-j} + \varepsilon_t, \qquad t \in \mathbb{Z},$$

where $\phi_0, \phi_1, \ldots, \phi_p$ are parameters inducing an autoregressive structure and $\varepsilon_t \sim N(0, \sigma^2)$ is an error term. Here, ϕ_0 plays the role of drift parameter, while the generic ϕ_j, $j = 1, \ldots, p$, measures the impact of the jth lagged element of the sequence on the current value. Once we observed a sample, the same model can be recasted in a matrix notation. Let us define the following quantities

$$\begin{aligned}
\boldsymbol{y}^\intercal &= (y_{p+1}, \ldots, y_t) \\
\mathrm{Z} &= \begin{bmatrix} \boldsymbol{z}_{p+1}^\intercal \\ \vdots \\ \boldsymbol{z}_t^\intercal \end{bmatrix}, \qquad \text{with} \qquad \boldsymbol{z}_i^\intercal = (1, y_{i-1}, \ldots, y_{i-p}), \\
\boldsymbol{\varepsilon}^\intercal &= (\varepsilon_{p+1}, \ldots, \varepsilon_t), \\
\boldsymbol{\phi}^\intercal &= (\phi_0, \phi_1, \ldots, \phi_p).
\end{aligned}$$

Hence, the same autoregressive model in a matrix form can be expressed as

$$\boldsymbol{y} = \mathrm{Z}\boldsymbol{\phi} + \boldsymbol{\varepsilon},$$

with $\boldsymbol{\phi} \in \mathbb{R}^{p+1}$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Remarkable cases can be obtained by suitably constraining the autoregressive parameters. Specifically, if we set $\phi = 0$, $\phi_j = 0$ for $j > 1$, we suppress the drift and we induce Markovian dependence, which results in having a random walk for $\phi_1 = 1$ and an oscillation around the mean for $0 < \phi_1 < 1$. The specific case for which also $\phi_1 = 0$ generate a white noise.

Remark 1.1. With higher autoregressive order, the model is stationary if the roots of the polynomial

$$p(x, \boldsymbol{\phi}) = 1 - \sum_{j=1}^{p} \phi_j x^j$$

lie outside the unit circle. However, stationarity is not fundamental for our inferential procedures. If a time series lead to non-stationary estimates, is good to know. Stationarity can be checked after the estimation. Further, priors more concentrated around the origin encourage stationarity.

A prior choice suitable for the previous model is simply coming from what we have done with linear regression models. More specifically, if we set

$$\begin{aligned} \boldsymbol{\phi} \mid \sigma^2 &\sim N(\boldsymbol{m}_0, \sigma^2 \Sigma_0), \\ \sigma^2 &\sim IG(a_0, b_0), \end{aligned}$$

a posteriori we obtain

$$\begin{aligned} \boldsymbol{\phi} \mid \{y_t\}_{t=1}^n, \sigma^2 &\sim N(\boldsymbol{m}_n, \sigma^2 \Sigma_n), \\ \sigma^2 \mid \{y_t\}_{t=1}^n &\sim IG(a_n, b_n), \end{aligned}$$

with

$$\begin{aligned} \Sigma_n &= \left(\Sigma_0^{-1} + \mathrm{Z}^\intercal \mathrm{Z}\right)^{-1}, \\ \boldsymbol{m}_n &= \Sigma_n \left(\Sigma_0^{-1} \boldsymbol{m}_0 + (\mathrm{Z}^\intercal \mathrm{Z}) \hat{\boldsymbol{\phi}}_{ML}\right) = \Sigma_n \left(\Sigma_0^{-1} \boldsymbol{m}_0 + \mathrm{Z}^\intercal \boldsymbol{y}\right), \\ a_n &= a_0 + \frac{n}{2}, \\ b_n &= b_0 + \frac{1}{2} \left(\boldsymbol{y}^\intercal \boldsymbol{y} - \boldsymbol{m}_n^\intercal \Sigma_n^{-1} \boldsymbol{m}_n + \boldsymbol{m}_0^\intercal \Sigma_0^{-1} \boldsymbol{m}_0\right). \end{aligned}$$

All the inferential procedure which we have seen in the third chapter are still valid. We can then obtain point estimate for the main parameter of the autoregressive model, construct credible intervals, perform tests, etc., as we have done in the past with the linear regression model.

Remark 1.2. Suppose, for example, that we want to test the presence of an autoregressive structure of order $p = 3$ versus the absence of an autoregressive structure. Hence, we can construct a Bayes factor as

$$\mathrm{BF}_{01} = \frac{m(y_{4:n} \mid M_0)}{m(y_{4:n} \mid M_1)}.$$

Note that we are ignoring the first 3 observations since we have to infer the autoregressive order $p = 3$ structure, and the marginals should be computed on the same data. The marginals are available in closed form (chapter 3), and the Bayes factor corresponds to

$$\mathrm{BF}_{01} = \sqrt{\frac{|\Sigma_n^{M_0}|}{|\Sigma_n^{M_1}|}} \left(\frac{b_n^{M_1}}{b_n^{M_0}}\right)^{a_n}.$$

where a_n is common to the two models, $\Sigma_n^{M_0}$, $b_N^{M_0}$ are the posterior parameter of the model M_0 and $\Sigma_n^{M_1}$, $b_N^{M_1}$ are the posterior parameter of the model M_1.

Once we estimate the model we can also perform predictive inference in the future. For the next observational dime, $n + 1$, we can directly resort to the predictive distribution which we derived in the third chapter, hence

$$\begin{aligned} f(y_{n+1} \mid y_{1:n}) &= \int f(y_{n+1} \mid \boldsymbol{\phi}, y_{(n-p):n}, \sigma^2)\pi(\boldsymbol{\phi}, \sigma^2 \mid y_{1:n})\mathrm{d}\boldsymbol{\phi}\mathrm{d}\sigma^2 \\ &\overset{d}{=} t_{2a_n}\left(\boldsymbol{z}_{n+1}^{\intercal}\boldsymbol{\phi}, \tfrac{b_n}{a_n}\big(1 + \boldsymbol{z}_{n+1}^{\intercal}\Sigma_n\boldsymbol{z}_{n+1}\big)\right) \end{aligned}$$

where $\boldsymbol{z}_{n+1}^{\intercal} = (1, y_{n-1}, \ldots, y_{n-p})$. Clearly, we can propagate our prediction in the future, conditioning on the upcoming predicted values, sampling a sequence $y_{n+1}, y_{n+1}, \ldots$ from the corresponding predictive distribution.

1.1 Selecting the autoregressive order

Selecting the autoregressive order becomes quite natural, since we can resort to the strategies previously seen in the module. Ideally, we can produce several model estimates, and then look at which model has results to be the best. Hence, we can either look at information criteria, such as BIC/WAIC or similar, or we can perform Bayesian testing with an incremental complexity of the model. For example, we can start with the model without autoregressive order, i.e. AR(0) and test the AR(1) model against the simpler one. Then, we can proceed sequentially, testing AR(p+1) versus AR(p), until we reject the more complex model. Within the previous model assumptions, everything is in a closed form and such a procedure is quite simple to implement.

1.2 Improving predictions combining models

Ideally, we can also combine different prediction through Bayesian model averaging. Hence, instead of considering a single autoregressive order p, we can estimate several models with different autoregressive orders, $M_1, \ldots, M_p$, and produce an average prediction of the form

$$f(y_{n+1} \mid y_{1:n}) = \sum_{j=1}^{p} P(M_j \mid y_{1:n}) f(y_{n+1} \mid M_j, y_{1:n})$$

whereas $P(M_j \mid y_{1:n}) \propto m(y_{1:n} \mid M_J)$ is the posterior probability of the jth model, while $f(y_{n+1} \mid M_j, y_{1:n})$ is its predictive distribution.

1.3 INDUCING SPARSITY WITH SPIKE-AND-SLAB

It is quite common that, fixing an autoregressive order, not all the coefficients are significative but only a certain subset. Hence, instead of considering a diffuse multivariate Gaussian prior for ϕ, we can set a spike-and-slab prior for each element, having a specification like

$$\begin{aligned}\beta_j \mid \gamma_j &\sim (1-\gamma_j)N(0,\tau^2) + \gamma_j N(0, c^2\tau^2),\\ \gamma_j \mid \theta_j &\sim Be(\theta_j),\\ \theta_j &\sim Beta(a,b),\end{aligned}$$

for $j = 1, \dots, p$. Such a prior specification induce sparsity of the autoregressive coefficients, since it set close to zero some of the values. The same concepts and procedures we discussed for the regression case apply also here. Specifically, we consider an autoregressive coefficient close enough to zero if it falls in an interval of the form $(-\tau_j\epsilon_j, +\tau_j\epsilon_j)$, with

$$\epsilon_j = \sqrt{2\frac{\log(c_j)c_j^2}{c_j^2 - 1}}.$$

Hence, once we fix c_j and we want a specific value $\tau_j\epsilon_j$ for the interval, we can select a specific τ_j that guarantees the coverage. The autoregressive lags can be selected as usual, resorting for example to HPD, MFM or HS approaches.

1.4 DYNAMIC REGRESSION MODELS

If we have access to a set of time-dependent exogenous information, such as covariates, we can include them in our model specification. We assume then that alongside with the time-dependent sequence, for each observational time we also collect a sequence of covariates $\{\boldsymbol{x}_t\}_{t=1}^n$, with $\boldsymbol{x}_t \in \mathbb{X} \subseteq \mathbb{R}^p$. Hence, we can extend the model specification including a regression term, obtaining

$$y_t = \boldsymbol{x}_{t-q}^\intercal \boldsymbol{\beta} + \boldsymbol{z}_t^\intercal \boldsymbol{\phi} + \varepsilon_t, \qquad t = p+1, \dots, n,$$

where $\boldsymbol{x}_{t-q}$ is the vector of local covariates, possibly lagged by q, $\boldsymbol{z}_t^\intercal = (1, y_{t-1}, \dots, y_{t-p})$, $\boldsymbol{\phi}$ describes the autoregessive coefficients, and $\varepsilon_t \sim N(0, \sigma^2)$.

The previous model appears to remind us a mixed effect model. It has a similar structure, where the first term represent the fixed effect, while the second term describe the time-dependent dynamic. Hence, we can set independent priors on the regression coefficients, on the autoregressive term and on the variance, having

$$\begin{aligned}\boldsymbol{\beta} &\sim N(\boldsymbol{\eta}_0, \Sigma_0),\\ \boldsymbol{\phi} &\sim N(\boldsymbol{m}_0, \Lambda_0),\\ \sigma^2 &\sim IG(a_0, b_0).\end{aligned}$$

Hence, from the calculations we have done in chapter 3, a posteriori we have

$$\begin{aligned}\boldsymbol{\beta} \mid - &\sim N(\boldsymbol{\eta}_n, \Sigma_n),\\ \boldsymbol{\phi} \mid - &\sim N(\boldsymbol{m}_n, \Lambda_n),\\ \sigma^2 \mid - &\sim IG(a_n, b_n),\end{aligned}$$

where

$$\boldsymbol{\eta}_n = \Sigma_n \left(\Sigma_0^{-1}\boldsymbol{\eta}_0 + \frac{\mathrm{X}^\intercal \boldsymbol{y}^\phi}{\sigma^2} \right), \qquad \Sigma_n = \left(\Sigma_0^{-1} + \frac{\mathrm{X}^\intercal \mathrm{X}}{\sigma^2} \right)^{-1},$$
$$\boldsymbol{m}_n = \Lambda_0 \left(\Lambda_0^{-1}\boldsymbol{m}_0 + \frac{\mathrm{Z}\boldsymbol{y}^\beta}{\sigma^2} \right)^{-1}, \qquad \Lambda_n = \left(\Lambda_0^{-1} + \frac{\mathrm{Z}^\intercal \mathrm{Z}}{\sigma^2} \right)^{-1},$$
$$a_n = a_0 + \frac{n}{2}, \qquad b_n = b_0 + \frac{1}{2}\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta} - \boldsymbol{z}_i^\intercal \boldsymbol{\phi} \right)^2,$$

with $y_i^\gamma = y_i - \boldsymbol{z}_i^\intercal \boldsymbol{\phi}$ and $\boldsymbol{y}_j^{\boldsymbol{\beta}} = y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}$. The previous conditional distribution can be iterated in a Gibbs sampling scheme to produce a set of values from the distribution of interest. The model can be further extended by considering covariates with different lag times, following a similar strategies but with multiple fixed effects.

2 MULTIVARIATE TIME-DEPENDENT DATA

We now consider the case of multivariate observed data, where at each time we observe a real-valued vector $\boldsymbol{y}_t \in \mathbb{Y} \subseteq \mathbb{R}^d$, hence producing a sequence of vectors $\{\boldsymbol{y}_t\}_{t=1}^n$ autocorrelated with each other. Ideally, each dimension of the vector can have possibly non-null correlation with the same dimension of others, with a lagged time, with

$$\mathrm{Corr}(y_{ti}, y_{t+\ell,j}) \in (-1, 1), \qquad \ell \geq 0,\ i, j \in \{1, \ldots, d\}.$$

We first consider the vector autoregressive model of order 1, here denoted with VAR(1). Under this model assumption we have

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \mathrm{A}\boldsymbol{y}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad t = 2, \ldots, n,$$

with $\boldsymbol{\mu}$ playing the role of drift parameter, while A models the autoregressive structure and $\boldsymbol{\varepsilon}_t$ is the local error term at time t. Specifically, we assume $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \Sigma)$. In a matrix representation, we have

$$\begin{pmatrix} y_{t1} \\ \vdots \\ y_{td} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} + \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \cdots & \vdots \\ a_{d1} & \cdots & a_{dd} \end{bmatrix} \begin{pmatrix} y_{t-1,1} \\ \vdots \\ y_{t-1,d} \end{pmatrix} + \begin{pmatrix} \varepsilon_{t1} \\ \vdots \\ \varepsilon_{td} \end{pmatrix}$$

whereas $\boldsymbol{a}_j \boldsymbol{y}_{t-1}$ is the autoregressive effect of the previous observed values $\boldsymbol{y}_{t-1}$ on the jth element of $\boldsymbol{y}_t$, where $\boldsymbol{a}_j$ is the jth row of A. Note that A is not a symmetric matrix, since we can have different effects on past dimension on the current ones rather than the opposite case.

Remark 2.1. Under the VAR(1) model assumption, we have (weak) stationarity if

$$\det(I_d - \mathrm{A}z) \neq 0, \qquad |z| \leq 1,$$

or equivalently if all the eigenvalues $\lambda_1, \ldots, \lambda_d$ of A are such that $0 < |\lambda_j| < 1$.

We need a suitable set of prior distributional assumptions for our model specification. At first, we can look at the expression of the likelihood function, to see if we could identify anything

conjugate. Hence, the likelihood for the previous model can be expressed as

$$\begin{aligned}\mathrm{L}(\mathrm{Y} \mid \boldsymbol{\mu}, \mathrm{A}, \Sigma) &= (2\pi)^{-nd/2}|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{t=1}^{n}(\boldsymbol{y}_t - \boldsymbol{\mu} - \mathrm{A}\boldsymbol{y}_{t-1})^{\intercal}\Sigma^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu} - \mathrm{A}\boldsymbol{y}_{t-1})\right\} \\ &= (2\pi)^{-nd/2}|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1}(\mathrm{Y} - \mathbf{1}_{n-1}\boldsymbol{\mu}^{\intercal} - \mathrm{ZA})^{\intercal}(\mathrm{Y} - \mathbf{1}_{n-1}\boldsymbol{\mu}^{\intercal} - \mathrm{ZA})\Big]\right\}\end{aligned}$$

where $\mathbf{1}_{n-1}$ is a $n-1$-dimensional vector of 1,

$$\mathrm{Y} = \begin{bmatrix} y_{21}, & \dots, & y_{2d} \\ \vdots & \vdots & \vdots \\ y_{n1}, & \dots, & y_{nd} \end{bmatrix}, \qquad \text{and} \qquad \mathrm{Z} = \begin{bmatrix} y_{11}, & \dots, & y_{1d} \\ \vdots & \vdots & \vdots \\ y_{n-1,1}, & \dots, & y_{n-1,d} \end{bmatrix}.$$

At first sight, we recognize that the σ^2 parameter enters in the distribution mimicking the structure of an inverse-gamma density function. Hence, we assume a priori $\sigma^2 \sim IG(a_0, b_0)$. Regarding the drift parameter, a priori we assume $\boldsymbol{\mu} \mid \sigma^2 \sim N(\boldsymbol{m}_0, \sigma^2\Sigma_0)$. For the autoregressive component, we need a distribution on the space of real-valued matrices of size $d \times d$. Specifically, a priori we assume a matrix-normal distribution.

Remark 2.2. We say that $\mathrm{X} \sim MN_{dq}(\mathrm{M}, \Psi, \Sigma)$ is a matrix-normal distributed random variable if its density function corresponds to

$$f(\mathrm{X} \mid \mathrm{M}, \Psi, \Sigma) = (2\pi)^{-dq/2}|\Psi|^{-d/2}|\Sigma|^{-q/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}(\mathrm{X} - \mathrm{M})^{\intercal}\Psi^{-1}(\mathrm{X} - \mathrm{M})\right]\right\},$$

where M is a location parameter and Σ, Ψ are scale parameters. Specifically, Ψ is a $d \times d$matrix, proportional to the correlation among rows, while Σ is a $q \times q$ matrix proportional to the correlation among columns.

Hence, the full model specification is given by

$$\begin{aligned}\mathrm{Y} \mid \boldsymbol{\mu}, \mathrm{A}, \Sigma &\sim MN_{nd}(\mathbf{1}\boldsymbol{\mu}^{\intercal}, \mathbf{I}_n, \Sigma), \\ \boldsymbol{\mu} \mid \Sigma &\sim N(\boldsymbol{m}_0, \Sigma/k_0), \\ \mathrm{A} \mid \Sigma &\sim MN_{dd}(\mathrm{M}_0, \Psi, \Sigma), \\ \Sigma &\sim IW(\nu_0, \Lambda_0).\end{aligned}$$

Proposition 2.3. *Under the previous model assumptions, a posteriori we have*

$$\begin{aligned}\boldsymbol{\mu}, \Sigma \mid \mathrm{A}, \mathrm{Y} &\sim NIW(\boldsymbol{m}_n, k_n, \nu_n, \Lambda_n), \\ \mathrm{A} \mid \boldsymbol{\mu}, \Sigma, \mathrm{A} &\sim MN_{dd}(\mathrm{M}_n, \Psi_n, \Sigma),\end{aligned}$$

with

$$\begin{aligned}k_n &= k_0 + n - 1, \qquad \boldsymbol{m}_n = \frac{1}{k_n}\left(k_0\boldsymbol{m}_0 + \mathrm{A}\sum_{t=1}^{n-1}\boldsymbol{y}_t\right), \\ \nu_n &= \nu_0 + n - 1, \qquad \boldsymbol{\Lambda}_n = \Lambda_0 + (\mathrm{A} - \mathrm{M}_0)^{\intercal}\Psi^{-1}(\mathrm{A} - \mathrm{M}_0) \\ &\qquad + \sum_{i=2}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\intercal} + \frac{k_0(n-1)}{k_n}(\boldsymbol{m}_0 - \bar{\boldsymbol{x}})(\boldsymbol{m}_0 - \bar{\boldsymbol{x}})^{\intercal}\end{aligned}$$

where $\boldsymbol{x}_i = \boldsymbol{y}_i - A\boldsymbol{y}_{i-1}$*, and*

$$\mathrm{M}_n = \Psi_n^{-1}(\Psi_0^{-1}\mathrm{M}_0 + \mathrm{Z}^\intercal U), \qquad \Psi_n = (\Psi_0^{-1} + \mathrm{Z}^\intercal\mathrm{Z})^{-1},$$

where $\mathrm{U} = \mathrm{Y} - \mathbf{1}\boldsymbol{\mu}^\intercal$.

Proof. We separate the proof in two parts. First, we look at the term regarding drift and dispersion. Hence, we have

$$\begin{aligned}
\pi(\boldsymbol{\mu},\Sigma \mid \mathrm{Y},\mathrm{A}) &\propto L(\mathrm{Y} \mid \boldsymbol{\mu},\mathrm{A},\Sigma)\pi(\boldsymbol{\mu} \mid \Sigma)\pi(\boldsymbol{A} \mid \Sigma)\pi(\Sigma) \\
&\propto |\Sigma|^{-(n-1)/2}|\Sigma|^{-1/2}|\Sigma|^{-d/2}|\Sigma|^{-(\nu_0+p+1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\Big(\Sigma^{-1}(\mathrm{Y}-\mathbf{1}_{n-1}\boldsymbol{\mu}^\intercal-\mathrm{ZA})^\intercal(\mathrm{Y}-\mathbf{1}_{n-1}\boldsymbol{\mu}^\intercal-\mathrm{ZA})\Big)\right.\right. \\
&\quad +(\boldsymbol{\mu}-\boldsymbol{m}_0)^\intercal\left(\frac{\Sigma}{k_0}\right)^{-1}(\boldsymbol{\mu}-\boldsymbol{m}_0)+\mathrm{tr}\Big(\Lambda_0\Sigma^{-1}\Big) \\
&\quad \left.\left.+\mathrm{tr}\Big(\Sigma^{-1}(\mathrm{A}-\mathrm{M}_0)^\intercal\Psi^{-1}(\mathrm{A}-\mathrm{M}_0)\Big)\right]\right\} \\
&= |\Sigma|^{-1/2}|\Sigma|^{-(\nu_0+n-1+d+p+1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\Big(\Sigma^{-1}(\mathrm{U}-\mathbf{1}_{n-1}\boldsymbol{\mu}^\intercal)^\intercal(\mathrm{U}-\mathbf{1}_{n-1}\boldsymbol{\mu}^\intercal)\Big)\right.\right. \\
&\quad +(\boldsymbol{\mu}-\boldsymbol{m}_0)^\intercal\left(\frac{\Sigma}{k_0}\right)^{-1}(\boldsymbol{\mu}-\boldsymbol{m}_0) \\
&\quad \left.\left.+\mathrm{tr}\Big(\big(\Lambda_0+(\mathrm{A}-\mathrm{M}_0)^\intercal\Psi^{-1}(\mathrm{A}-\mathrm{M}_0)\big)\Sigma^{-1}\Big)\right]\right\}
\end{aligned}$$

with $\mathrm{U} = \mathrm{Y} - \mathrm{ZA}$. The previous is nothing but the posterior distribution of a multivariate Gaussian likelihood with a NIW prior, which we already saw in the previous chapter. Similarly, we set $\mathrm{V} = \mathrm{Y} - \mathbf{1}\boldsymbol{\mu}^\intercal$. Hence, for the autoregressive term, we have

$$\begin{aligned}
\pi(A \mid \mathrm{Y},\boldsymbol{\mu},\Sigma) &\propto L(\mathrm{Y} \mid \boldsymbol{\mu},\mathrm{A},\Sigma)\pi(\boldsymbol{A} \mid \Sigma) \\
&\propto \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\Big(\Sigma^{-1}(\mathrm{Y}-\mathbf{1}_{n-1}\boldsymbol{\mu}^\intercal-\mathrm{ZA})^\intercal(\mathrm{Y}-\mathbf{1}_{n-1}\boldsymbol{\mu}^\intercal-\mathrm{ZA})\Big)\right.\right. \\
&\quad \left.\left.+\mathrm{tr}\Big(\Sigma^{-1}(\mathrm{A}-\mathrm{M}_0)^\intercal\Psi^{-1}(\mathrm{A}-\mathrm{M}_0)\Big)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\Big((\mathrm{V}-\mathrm{ZA})^\intercal(\mathrm{V}-\mathrm{ZA})+(\mathrm{A}-\mathrm{M}_0)^\intercal\Psi^{-1}(\mathrm{A}-\mathrm{M}_0)\Big)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\Big(\mathrm{A}^\intercal\mathrm{Z}^\intercal\mathrm{ZA}+\mathrm{A}^\intercal\Psi^{-1}\mathrm{A}-2\mathrm{A}^\intercal\mathrm{Z}^\intercal\mathrm{V}-2\mathrm{A}^\intercal\Psi^{-1}\mathrm{M}_0+C_{M_0,Z,V,\Psi}\Big)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\Big((\mathrm{A}-\mathrm{M}_n)^\intercal\Psi_n(\mathrm{A}-\mathrm{M}_n)\Big)\right]\right\},
\end{aligned}$$

which identifies a matrix Normal distribution.

□

The previous posterior characterization can be used, similarly to the univariate case, to perform posterior inference, such as point estimation, credible intervals, Bayesian testing and prediction. Generalization to higher cases can be done. A similar approach holds for the posterior distribution, working for each specific autoregression matrix, given the rest.

3 BAYESIAN MODEL OF SPATIAL DATA

Spatial observations are sample data which depend on a specific coordinate of measurement. Those coordinates are providing informations and insights that can be helpful to better understand and interpret what we are studying. Spatial data are usually classified into 3 types, depending on the specific structure of the data.

- Point-referenced data (geostatistical data): $y(\boldsymbol{s})$ random vector at location $\boldsymbol{s} \in \mathbb{R}^r$. The coordinate $\boldsymbol{s}$ varies continuously over D, a subset of $\mathbb{R}^r$ that contains a r-dim rectangle of positive volume.
- Areal data: $y(\boldsymbol{s})$, $\boldsymbol{s} \in D$, and D is partitioned into a finite number of areal units with well-defined boundaries.
- Point pattern data: D is random, and the index set of D gives the locations of random events that are the spatial point pattern. For example, $y(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in D$.

As statistician, we want to investigate if the spatial domain has an impact on the data structure. Specifically, we want to study if there is any spatial pattern in data $y(\boldsymbol{s}_1), y(\boldsymbol{s}_2), \ldots, y(\boldsymbol{s}_n)$. Spatial pattern suggests that measurements near to each other will tend to take more similar values than those for units far from each other. Independent measurements for the units means there is no pattern.

3.1 POINT-REFERENCED DATA

In this framework, data $y(\boldsymbol{s})$ are given at specific locations $\boldsymbol{s} \in D \subseteq \mathbb{R}^r$. For example, $y(\boldsymbol{s})$ are level of a pollutant at site $\boldsymbol{s}$. While we can assume the existence of a pollutant level at all possible sites, in practice the data are a partial realization of a spatial process at specific locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. A fundamental object to deal with this type of data is the underlying stochastic process $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$. We observe the process at fixed locations. Hence, the data we observe are

$$y_{1:n} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\intercal.$$

The process is centered in $\mu(\boldsymbol{s}) = \mathbb{E}[y(\boldsymbol{s})]$, which plays the role of mean parameter. We also assume that the variance of $y(\boldsymbol{s})$ exists at each $\boldsymbol{s} \in D$. Regularity and tractability come from assuming $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\intercal$ distributed according to a multivariate Gaussian distribution, which we set as our distributional assumption.

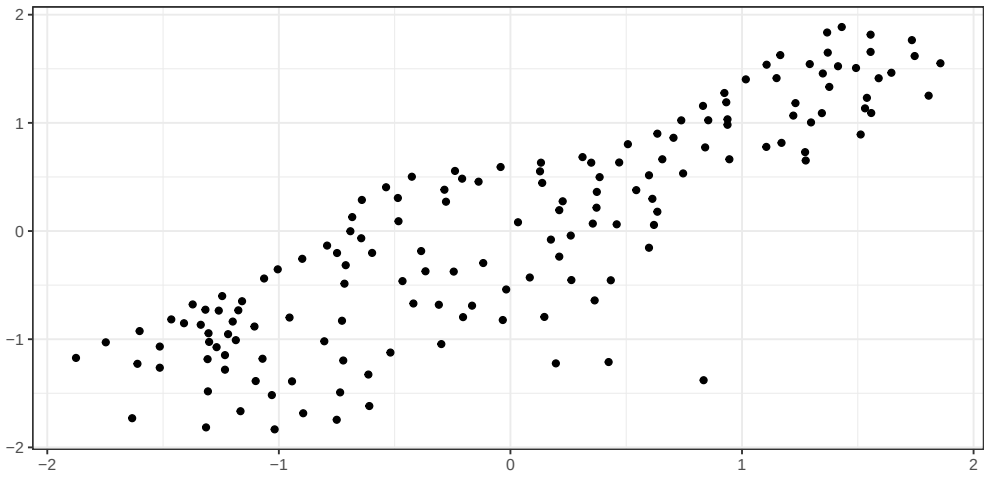

Figure 1: Locations (we also have topsoil heavy metal concentrations along with a number of soil and landscape variables at the observation locations) collected in a flood plain of the river Meuse, near the village of Stein (NL).

A fundamental concept is the stationarity of the underlying process. Under stationarity, the characteristic of such a processes such as mean, variance and covariance do not change upon shifting the support.

- The process is said to be strictly (strong) stationary if

  S1) for any $n \geq 1$, any set of sites $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, and $\boldsymbol{h} \in \mathbb{R}^r$, we have

$$(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)) \stackrel{d}{=} (y(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, y(\boldsymbol{s}_n + \boldsymbol{h})), \qquad \text{with } \boldsymbol{s}_j + \boldsymbol{h} \in D.$$

- The process is said to be weakly stationary if

  W1) $\mu(\boldsymbol{s}) = \mu$, i.e. constant mean over the spatial domain.

  W2) For the covariance term, we have

$$\text{cov}(y(\boldsymbol{s}), y(\boldsymbol{s} + \boldsymbol{h})) = C(\boldsymbol{h}), \qquad \text{with } \boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h} \in D.$$

  In practice, the covariance can be summarized in a covariance function.

- The process is said to be intrinsic stationary if

  I1) For the lagged first moment we have $\mathbb{E}[y(\boldsymbol{s} + \boldsymbol{h}) - y(\boldsymbol{s})] = 0$.

  I2) For the lagged second moment, we have

$$\mathbb{E}[(y(\boldsymbol{s} + \boldsymbol{h}) - y(\boldsymbol{s}))^2] = \text{var}(y(\boldsymbol{s} + \boldsymbol{h}) - y(\boldsymbol{s})) = 2\gamma(\boldsymbol{h}),$$

  depending solely on $\boldsymbol{h}$, where $2\gamma(\boldsymbol{h})$ is called variogram and $\gamma(\boldsymbol{h})$ semivariogram.

Note that: strong stationarity $\Rightarrow$ weak stationarity $\Rightarrow$ intrinsic stationarity.

The two fundamental tools we use to characterize the distributions of a point-referenced spatial model are the semivariogram or the covariance function. The former models the average

dissimilarity between two points on the support, while the latter models their average similarity. We recall the following relation between the semivariogram $\gamma(\boldsymbol{h})$ and the covariance function $C(\boldsymbol{h})$

$$\gamma(\boldsymbol{h}) = C(0) - C(\boldsymbol{h}) \quad \Leftrightarrow \quad C(\boldsymbol{h}) = C(0) - \gamma(\boldsymbol{h}) = \lim_{||\boldsymbol{u}||\to\infty} \gamma(\boldsymbol{u}) - \gamma(\boldsymbol{h}).$$

Further, we can break down the variability of our process into three sources, having

Nuggets: $\tau^2 = \lim_{d\to 0^+} \gamma(d)$, represent the non-spatial variability

Range: $R = 1/\phi$, where ϕ is the decay parameter

Sill: $\tau^2 + \sigma^2 = \lim_{d\to+\infty} \gamma(d)$

The first represents the initial value that the semivariogram is taking, measuring the local variability of the distribution at each observed point. The last term, the sill, represent the limit value of the semivariogram. Since the semivariogram is usually upperbounded, the sill represent the maximum value its taking in the limit case. Finally, the range corresponds to the distance between two points after which the semivariogram reaches the sill value. Finally, we remark that the model is said to be isotropic if $\gamma(\boldsymbol{h}) = \gamma(||\boldsymbol{h}||)$. In isotropic models, the average similarity among two points, separated by $\boldsymbol{h}$, does not depends on specific coordinates of $\boldsymbol{h}$ but on the magnitude of their distance, quantified as $||\boldsymbol{h}||$. A model that is both isotropic and stationary is called homogeneous.

Example 3.1. Exponential semivariogram/covariance.

$$\gamma(d) = \gamma(||\boldsymbol{h}||) = \begin{cases} \tau^2 + \sigma^2(1 - \mathrm{e}^{-\phi d}) & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases}$$

$$C(d) = C(||\boldsymbol{h}||) = \begin{cases} \sigma^2 \mathrm{e}^{-\phi d} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

Note that, for $d > 0$, we use the notation $C(d) = \sigma^2\rho(d, \phi)$.

Example 3.2. Powered exponential semivariogram/covariance.

$$\gamma(d) = \gamma(||\boldsymbol{h}||) = \begin{cases} \tau^2 + \sigma^2(1 - \mathrm{e}^{-|\phi d|^p}) & \text{if } d > 0 \\ 0 & \text{if } d = 0 \end{cases}$$

$$C(d) = C(||\boldsymbol{h}||) = \begin{cases} \sigma^2 \mathrm{e}^{-|\phi d|^p} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

with $0 < p \leq 2$ and $\rho(d, \phi) = e^{-|\phi d|^p}$.

Example 3.3. Gaussian semivariogram/covariance.

$$C(d) = C(||\boldsymbol{h}||) = \begin{cases} \sigma^2 e^{-\phi^2 d^2} & \text{if } d > 0 \\ \tau^2 + \sigma^2 & \text{if } d = 0 \end{cases}$$

and $\rho(d, \phi) = e^{-\phi^2 d^2}$.

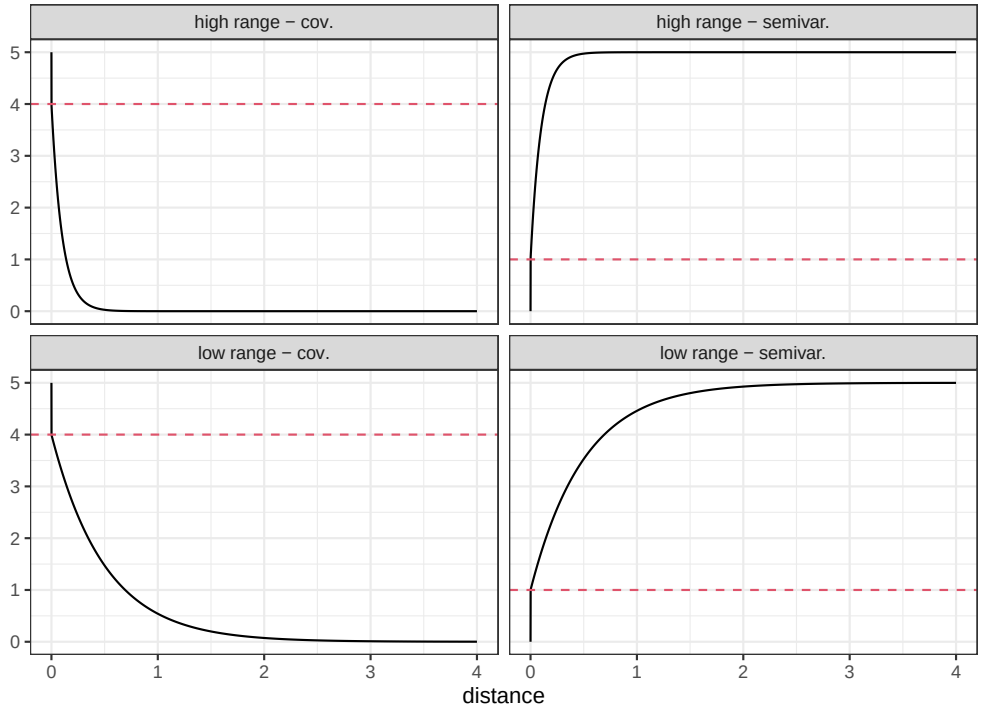

Figure 2: Graphical illustration of covariance function and semivariogram for different values of the range parameter $\phi \in \{2, 10\}$, top and bottom rows respectively.

We introduce a first Bayesian spatial regression model for point-referenced data. We consider something that reminds a mixed model, having random effects, discussed in the third chapter. We consider a model of the form

$$y(\boldsymbol{s}) = \boldsymbol{x}^{\intercal}(\boldsymbol{s})\boldsymbol{\beta} + \omega(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),$$

where the residual term of the model is partitioned in two parts: $\omega(\boldsymbol{s})$ is the spatial residual term, where $\{\omega(\boldsymbol{s})\}_{s\in D}$ is a spatial Gaussian process, capturing the residual spatial association. Its distribution is indexed by the dispersion parameters σ^2 and ϕ. $\{\epsilon(\boldsymbol{s})\}_{s\in D}$ is a sequence of uncorrelated pure error terms, variability at distances smaller than the smallest interlocation distance, with a distribution indexed by τ^2.

Let X be a $n \times p$ matrix with $\boldsymbol{x}^{\intercal}(\boldsymbol{s}_i)$ being its ith row, and $\boldsymbol{\omega} = (\omega(\boldsymbol{s}_1), \ldots, \omega(\boldsymbol{s}_n))^{\intercal}$. The model specification is completed by setting

$$\begin{aligned}
\mathrm{y} \mid \boldsymbol{\omega}, \boldsymbol{\beta}, \tau^2 &\sim N(\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 \mathrm{I}_n), \\
\boldsymbol{\omega} \mid \boldsymbol{\theta} &\sim N(\mathbf{0}, \Sigma(\boldsymbol{\theta})), \qquad \text{where } [\Sigma(\boldsymbol{\theta})]_{ij} = \sigma^2 \rho(||\boldsymbol{s}_i - \boldsymbol{s}_j||, \boldsymbol{\theta}), \\
\boldsymbol{\beta} &\sim N(\boldsymbol{b}_0, \Lambda_0), \\
\boldsymbol{\theta} = (\sigma^2, \phi, \tau^2) &\sim \pi(\sigma^2, \phi, \tau^2).
\end{aligned}$$

where $\pi(\sigma^2, \phi, \tau^2)$ usually is assumed to factorize in independent components. The crucial part here is how we model dependencies across distinct locations, specifically through the

distribution of $\boldsymbol{\omega}$. Finally, considering the support of the parameters in $\boldsymbol{\theta}$, a suitable prior choice is given by the following assumptions

$$\begin{aligned}\sigma^2 &\sim IG(a_\sigma, b_\sigma),\\ \tau^2 &\sim IG(a_\tau, b_\tau),\\ \phi &\sim IG(a_\phi, b_\phi).\end{aligned}$$

Proposition 3.4. *We can integrate out the random effect $\boldsymbol{\omega}$, obtaining as marginal model*

$$\begin{aligned}\mathrm{y} \mid \boldsymbol{\beta}, \tau^2 &\sim N(\mathrm{X}\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 \mathrm{I}_n), \qquad \textit{where } [H(\phi)]_{ij} = \rho(||\boldsymbol{s}_i - \boldsymbol{s}_j||, \phi),\\ \boldsymbol{\beta} &\sim N(\boldsymbol{b}_0, \Lambda_0),\\ \boldsymbol{\theta} = (\sigma^2, \phi, \tau^2) &\sim \pi(\sigma^2, \phi, \tau^2).\end{aligned}$$

The model can be implemented in STAN, by suitably constructing the correlation/covariance matrices needed in the spatial residual term (first specification) or in the marginal distribution of the data (second specification).

As usual in spatial analysis, one of our main scopes is to perform kriging. In a Bayesian framework, kriging is nothing but Bayesian prediction. We want to predict the response Y_0 at a new location $\boldsymbol{s}_0$, given a vector of predictors $\boldsymbol{x}_0 = \boldsymbol{x}(\boldsymbol{s}_0)$, by computing the predictive distribution

$$\begin{aligned}f(y_0 \mid \boldsymbol{x}_0, \boldsymbol{y}, \mathrm{X}) &= \int_{\mathbb{R}^p \times \mathbb{R}^3_+} \mathcal{L}(y_0, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{x}_0, \boldsymbol{y}, \mathrm{X}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{\theta}\\ &= \int_{\mathbb{R}^p \times \mathbb{R}^3_+} \mathcal{L}(y_0 \mid \boldsymbol{x}_0, \boldsymbol{\beta}, \boldsymbol{\theta}) \pi(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}, \mathrm{X}) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{\theta},\end{aligned}$$

where $\pi(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}, \mathrm{X})$ denotes the posterior distribution of interest. In general, the previous integral can be solved numerically, starting from an MCMC output. In practice, we can compute directly the quantity we need in STAN. Under the Gaussian model, we can write explicitly $\mathcal{L}(y_0 \mid \boldsymbol{x}_0, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}, \mathrm{X})$.

Remark 3.5. Recall that, from standard multivariate Gaussian proprieties, if

$$\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}\right)$$

with $\Omega_{12} = \Omega_{21}^\intercal$. Then, the conditional distribution of $\boldsymbol{y}_1 \mid \boldsymbol{y}_2$ is still a Gaussian distribution, with mean and covariance matrix

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_1 \mid \boldsymbol{y}_2] &= \boldsymbol{\mu}_1 + \Omega_{12}\Omega_{22}^{-1}(\boldsymbol{Y}_2 - \boldsymbol{\mu}_2),\\ \mathrm{var}(\boldsymbol{y}_1 \mid \boldsymbol{y}_2) &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}.\end{aligned}$$

In our framework, we have $\boldsymbol{y}_1 = y(\boldsymbol{s}_0)$, $\boldsymbol{y}_2 = \boldsymbol{y}$. Then, we have

$$\Omega_{11} = \sigma^2 + \tau^2, \quad \Omega_{12} = \boldsymbol{\gamma}^\intercal, \quad \Omega_{22} = \sigma^2 H(\phi) + \tau^2 \mathrm{I}_n,$$

where $\boldsymbol{\gamma}^\intercal = (\sigma^2\rho(d_{01}, \phi), \ldots, \sigma^2\rho(d_{0n}, \phi))$. Hence,

$$\begin{aligned}\mathbb{E}[y(\boldsymbol{s}_0) \mid \boldsymbol{y}] &= \boldsymbol{x}_0^\intercal\boldsymbol{\beta} + \boldsymbol{\gamma}^\intercal(\sigma^2 H(\phi) + \tau^2 \mathrm{I}_n)^{-1}(\boldsymbol{y} - \mathrm{X}\boldsymbol{\beta}),\\ \mathrm{var}(y(\boldsymbol{s}_0) \mid \boldsymbol{y}) &= \sigma^2 + \tau^2 - \boldsymbol{\gamma}^\intercal(\sigma^2 H(\phi) + \tau^2 \mathrm{I}_n)^{-1}\boldsymbol{\gamma}.\end{aligned}$$

3.2 AREAL DATA

The second type of spatial data we are considering consists of areal data. We recall that areal data $y(\boldsymbol{s})$, $\boldsymbol{s} \in D$, consists of a spatial data where the domain D is partitioned into a finite number of areal units. Hence, our realizations are $\boldsymbol{y} = (y_1, \dots, y_n)$, continuous, binary, count, etc., associated to n distinct areal units $\boldsymbol{s} = \{s_1, \dots, s_n\}$. We also have a $n \times n$ matrix, here called W, that describes how different areas are, in some way, connected. Typically, we set $w_{ii} = 0$, $i = 1, \dots, n$, i.e. an observation is not connected with itself. Further, we have

- $w_{ij} = 1$ if the ith and the jth area share at least a common boundary.
- w_{ij} could reflect the distance among units, e.g. a decreasing function of intercentroidal distance.

W is usually a symmetric matrix, and it can be marginally standardized by defining

$$\tilde{w}_{ij} = \frac{w_{ij}}{w_{i+}}, \qquad w_{i+} = \sum_{j=1}^{n} w_{ij}$$

Is not symmetric anymore, but is a stochastic matrix with row summing up to 1.

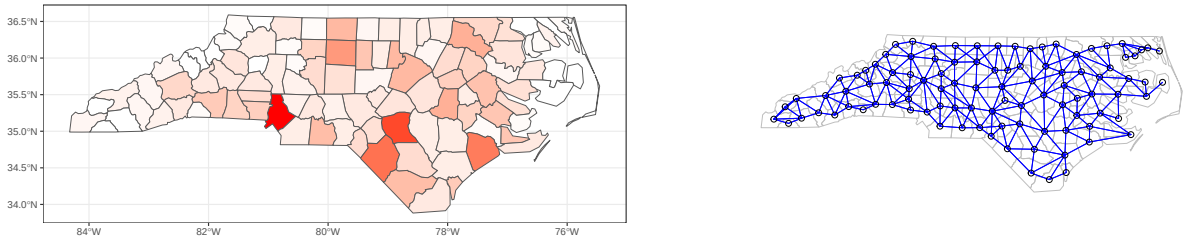

Figure 3: Left: sudden infant deaths in North Carolina for 1974-78. We also have access to other information for each specific areas, such as number of births and number of non-white birth. Right: the corresponding graph connecting different areas, and defining the adjacency matrix.

Typical quantities to measure the strength of spatial association among different areal units are the following. Moran's I, which is defined as

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i \neq j} w_{ij}) \sum_{i=n}^{n} (y_j - \bar{y})^2}$$

which is the analogue of lagged autocorrelation for time series. By construction, is not constricted in $[-1, 1]$. Geary's C, which is defined as

$$C = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})^2}{2(\sum_{i \neq j} w_{ij}) \sum_{i=n}^{n} (y_j - \bar{y})^2}$$

The C index is never negative. Small values (between 0 and 1) indicate positive spatial association.

A relevant issue here is how we can specify a joint distribution, playing the role of the likelihood term, for $\boldsymbol{y} = (y_1, \dots, y_n)^\intercal$, that incorporates also the spatial dependence we across

areas. A possible strategy that we can explore is to build up the joint distribution starting from the full conditional distributions of each observation, giving all the others, i.e. $\mathcal{L}(y_i \mid \boldsymbol{y}_{-i})$, where $\boldsymbol{y}_{-i}$ denotes all the observations discarding the ith term. The joint distribution can be determined by the product of the full conditionals. However, the joint distribution can be improper.

Instead of considering the whole support, we denote by ∂_i a generic neighborhood of i. Suppose we specify the full conditionals in a local fashion, by considering

$$\mathcal{L}(y_i \mid \boldsymbol{y}_{-i}) = \mathcal{L}(y_i \mid y_j \in \partial_i), \quad i = 1, \dots, n.$$

By specify the full conditionals in the previous way, we identify a unique joint distribution, while we are inside the Markov random field domain.

Remark 3.6. A Markov random field is nothing but a mathematical model used to describe a set of random observations having local dependencies. It is specified through an undirected graph representation $\mathcal{G} = (V, E)$. Hence, all the possible areas are the nodes of our graph V and the connections among areas are described by the edges E. Specifically, the graph structure drives all the dependencies, having that two non-adjacent observations, say y_j, y_ℓ are independent conditionally on the remaining ones, in formula

$$y_j \perp\!\!\!\perp y_\ell \mid \{y_i\}_{i \neq j,l}, \qquad \text{if} \qquad E_{j\ell} = 0.$$

A remarkable case is the AR(1) model without a drift, whereas an observation are defined as

$$y_t = \phi y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, 1), \qquad |\phi| < 1.$$

Assuming $y_1 \sim N(0, \frac{1}{1-\phi^2})$, we have

$$\boldsymbol{y} \sim N(\boldsymbol{0}, \Phi), \qquad \Phi = \begin{bmatrix} 1 & -\phi & 0 & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & 0 & \cdots & 0 \\ 0 & -\phi & 1+\phi^2 & -\phi & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & -\phi & 1 \end{bmatrix}.$$

Conditionally autoregressive (CAR) models are an example of Marov random fields, where the joint distribution is a Gibbs distribution, i.e. it exists but it can be improper. We introduce the CAR case with continuous Y_is, Gaussian distributed. The same framework can be extended more in general to exponential family models. We then set, for each single observation, a model of the form

$$y_i \mid \boldsymbol{y}_{-i} \sim N\left(\sum_{j \neq i} b_{ij} y_j, \tau_i^2\right), \quad i = 1, \dots, n.$$

These full conditional are compatible, and we obtain as joint distribution

$$f(y_1, \dots, y_n) \propto \exp\left\{-\frac{1}{2}\boldsymbol{y}^\intercal D^{-1}(\mathrm{I}_n - B)\boldsymbol{y}\right\},$$

where $[B]_{ij} = b_{ij}$ and $D = \mathrm{diag}(\tau_1^2, \dots, \tau_n^2)$.

Ideally, the previous expression suggests us a joint multivariate normal distribution for $\boldsymbol{y}$, with $\boldsymbol{0}$ mean and covariance matrix $\Sigma = (\mathrm{I}_n - B)^{-1}D$. But we should be careful, as Σ^{-1} and hence $D^{-1}(\mathrm{I}_n - B)$ must be symmetric and nonsingular.

Enforcing symmetry

First, to enforce symmetry in $D^{-1}(\mathrm{I}_n - B)$, we should satisfy the following conditions

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}, \qquad i, j = 1, \ldots, n.$$

From the previous, is apparent that B does not need to be symmetric. From the definition of the Gaussian CAR model, b_{ij} relates observation i and j. Hence, we can return back to our proximity matrix W, which is assumed to be symmetric. Suppose now that we set $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$. Then the previous is satisfied since W is symmetric and

$$\frac{w_{ij}w_{i+}}{\tau_2 w_{i+}} = \frac{w_{ji}w_{j+}}{\tau_2 w_{j+}}, \qquad i, j = 1, \ldots, n,$$

leading to full conditionals of the form

$$y_i \mid \boldsymbol{y}_{-i} \sim N\left(\sum_{j=1}^{n} \frac{w_{ij}}{w_{i+}} y_j, \frac{\tau^2}{w_{i+}}\right).$$

From the previous full conditionals, the joint distribution we obtain takes form

$$f(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2}\boldsymbol{y}^\intercal (D_w - W)\boldsymbol{y}\right),$$

where D_w is a diagonal matrix with $[D_w]_{ii} = w_{i+}$.

Enforcing non-singularity

Secondly, we can note a second aspect. Unfortunately, within the previous construction we have

$$(D_w - W)\mathbf{1} = \mathbf{0},$$

hence $(D_w - W) = \Sigma^{-1}$ is singular, so that Σ does not exist. Note that, while for Σ singular we do not have a density function, but a distribution that lives in a lower dimensional space, when Σ^{-1} is singular we do have a density function, but not integrable, hence improper. With some algebra, the previous density function can be rewritten as

$$f(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2}\sum_{i \neq j} w_{ij}(y_i - y_j)^2\right),$$

The improperty is still apparent from the previous, we can add any constant to all the Y_is, and the previous is unaffected. However, it can still be used as improper model. The previous is usually referred to as intrinsically autoregressive (IAR) model.

A slight variation of the previous model gives us a proper distribution. We redefine $\Sigma^{-1} = (D_w - \rho W)$, by suitably choosing ρ such that Σ^{-1} is nonsingular. The nonsingularity is guaranteed by setting $\rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$, where $\lambda_{(1)} < \lambda_{(2)} \cdots < \lambda_{(n)}$ are the ordered eigenvalues of $D_w^{1/2} W D_w^{-1/2}$. Moreover, since

$$\mathrm{tr}(D_w^{1/2} W D_w^{-1/2}) = 0 = \sum_{i=1}^{n} \lambda_{(i)},$$

then we have $\lambda_{(1)} < 0$, $\lambda_{(n)} > 0$ and 0 being in the set of interest $(1/\lambda_{(1)}, 1/\lambda_{(n)})$.

Simpler bounds can be alternatively obtained by looking at the scaled matrix we defined above, $\tilde{W} = \text{diag}(1/w_{1+}, \ldots, 1/w_{n+})W$. Such a matrix is not symmetric, but is row stochastic (i.e. all of its rows sum to 1). Then, Σ^{-1} can be written as $M^{-1}(\text{I}_n - \alpha\tilde{W})$, where M is diagonal. Further, if $|\alpha| < 1$, then $(\text{I}_n - \alpha\tilde{W})$ is nonsingular.

Full model specification

Under the first constrain, with $\Sigma^{-1} = (D_w - \rho W)$, with W symmetric matrix, we have

$$y_i \mid \boldsymbol{y}_{-i} \sim N\left(\rho \sum_{j=1}^{n} \frac{w_{ij}}{w_{i+}} y_j, \frac{\tau^2}{w_{i+}}\right), \qquad i = 1, \ldots, n.$$

Typically, we set $\rho \in (0, 1)$. For the boundary values, if $\rho = 0$, then

$$y_i \overset{iid}{\sim} N(0, \tau^2/w_{i+}), \qquad i = 1, \ldots, n.$$

Otherwise, if $\rho = 1$, then we are back to the improper intrinsic CAR model. Usually we set a prior on ρ. A prior with mass near to 1 encourages spatial association among different areas. The model can be used directly as distribution for the data, or can be combined in more complex models where the spatial association is dictated by a latent parameter level. In a general setting, we consider $\boldsymbol{s} = \{s_1, \ldots, s_n\}$ areal units; $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ response variables; $\boldsymbol{o} = \{o_1, \ldots, o_n\}$ offsets, additional information that we have on the areal units. The spatial pattern in the response is modelled by X a matrix of covariates, where $\boldsymbol{x}_i^\intercal = (x_{i1}, \ldots, x_{ip})$ is the covariate vector associated to the ith area, and a set of random effects $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$. We consider a generic case arising from GLMM. The model specification is given by

$$\begin{aligned}
y_i \mid \mu_i &\overset{ind}{\sim} f(y_i \mid \mu_i, \nu^2), \qquad i = 1, \ldots, n, \\
g(\mu_i) &= \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i + o_i, \qquad i = 1, \ldots, n, \\
\boldsymbol{\beta} &\sim N(\boldsymbol{b}_0, \Lambda_0), \\
\nu^2 &\sim IG(a_\nu, b_\nu), \\
(\phi_1, \ldots, \phi_n) &\sim CAR(W, \rho, \tau^2), \\
\rho &\sim Beta(a_\rho, b_\rho), \\
\tau^2 &\sim IG(a_\tau, b_\tau).
\end{aligned}$$

Regarding specific distributional assumption for $f(y_i \mid \mu_i, \nu^2)$, we can assume the usual ones, depending on the type of response variable, such as

- Gaussian, $y_i \sim N(\mu_i, \nu^2)$ and $\mu_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i + o_i$.
- Bernoulli, $y_i \sim Be(\theta_i)$ and

$$\mu_i = \log(\theta_i/(1 - \theta_i)) = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i + o_i.$$

- Poisson, $y_i \sim Poi(\mu_i)$ and $\log(\mu_i) = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i + o_i$.

Example 3.7. The offset is needed in case we want to adjust what we observe on some feature of the area, such as the dimension. Suppose for example that we are interested into model rates, but we observe counts. For instance, suppose we observe a count Y in an area with surface o. Our model is a Poisson with $\mathbb{E}[Y \mid x] = \mu_x$, but we are interested in

$$\log \frac{\mu_x}{o} = \beta_0 + \beta_1 x \qquad \Longrightarrow \qquad \log \mu_x = \log(o) + \beta_0 + \beta_1 x.$$

Then $\log(o)$ is the corresponding offset.