BAYESIAN STATISTICAL MODELS - University of Milano-Bicocca

Case study - Time and space

Lecturer: Riccardo Corradin

We consider a set of time series of opening prices for the Google, Amazon, and Microsoft stocks, from 01 January 2025 to 31 January 2025, contained in the file *time_series.csv*. At first, we consider only the Amazon stock. We specify an univariate autoregressive model of the form

$$y_i = \phi_0 + \sum_{j=1}^{p} \phi_j y_{i-j} + \varepsilon_i, \qquad i = p+1, \ldots, t,$$

with $\varepsilon_i \sim N(0, \sigma^2)$, $\boldsymbol{\phi} \sim N(\mathbf{0}, \Sigma_0)$ and $\sigma^2 \sim IG(a_0, b_0)$, with $\Sigma_0 = \text{diag}_{p+1}(1)$, $a_0 = 2$ and $b_0 = 1$.

- Implement a Gibbs sampler for the posterior distributions of a generic AR(p) model.
- Consider an AR(1) model. Produce a sample of size $5\,000$, after $2\,000$ burnin iterations, from the posterior distribution of interest. Check the model fit.
- Reproduce the analysis considering an AR(7) model. Compare the models in terms of information criteria.

We now analyze all the series together. Let us consider now a VAR(1) model, assuming

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{y}_{i-1}^{\intercal} \mathrm{A} + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, t,$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Sigma)$. A priori we set $\boldsymbol{\mu} \sim N(\mathbf{0}, \Sigma/k_0)$, $\mathrm{A} \sim MN(0, \Sigma, \Psi_0)$ and $\Sigma \sim IW(\nu_0, \Lambda_0)$. We set $k_0 = 1$, $\nu_0 = 5$, $\Lambda_0 = \text{diag}_d(1)$ and $\Psi_0 = \text{diag}_d(1)$.

- Implement a Gibbs sampler for the posterior distributions of the VAR(1) model. Produce a sample of size $5\,000$, after $2\,000$ burnin iterations, from the posterior distribution of interest. Check the model fit.
- Obtain a posterior estimate for the autoregressive effects, emphasizing what effects are significatively different from 0 and how they affect the next observation.

We now consider first a set of data related to heavy metals measured in the top soil in a flood plain along the river Meuse, observed at some specific location, along with a handful of covariates such as the distance from the river and the altitude of each specific location. The data are available online in the *river.rda* file.

- y contains the logarithm of the zinc measurements.
- X contains the altitude (log-scaled) and the distance from the river of each point.
- S contains the coordinates of each observation.

We want to investigate if the spatial component is significant for the model specification. To this end, we consider an integrated model for point-referenced data of the form

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

with the following prior assumption

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

We assume a Gaussian covariance function, i.e. $\rho(d, \phi) = \mathrm{e}^{-\phi^2 d^2}$. The model specification is completed by setting $\boldsymbol{b}_0 = \mathbf{0}$, $\Lambda_0 = \mathrm{diag}(10^3)$, $a_\sigma = a_\tau = a_\phi = 2.1$ and $b_\sigma = b_\tau = b_\phi = 10$.

- Produce a code in STAN that produce a sample from the posterior distribution of interest, while performing prediction for a new single observation.
- Exclude the first observation from the dataset and use such observation to perform prediction. Sample $5\,000$ realizations from the posterior distribution of interest, after $2\,000$ burnin iterations. Check the convergence of the sampled chains.
- Write a function that estimate the model without the spatial component. Test if the first model is significantly different from this simpler model.

Lastly, we consider a set of data regarding 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. The data are observed for each specific county in North Carolina. The data are available online in the *sids.rda* file. For each observation, we then have

- y, number of sudden infant deaths specific for each county.
- X contains the number of births and non-white births in the time window of the study.
- W, the adjacency matrix of the counties.
- O, an offset quantity, namely the surface area of each county.

We consider a model of the form

$$\begin{aligned} Y_i \mid \mu_i &\overset{ind}{\sim} Poi(\mu_i), \qquad i = 1, \dots, n, \\ \log(\mu_i) &= \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \phi_i + \log(O_i), \qquad i = 1, \dots, n, \\ \boldsymbol{\beta} &\sim N(\boldsymbol{b}_0, \Lambda_0), \\ (\phi_1, \dots, \phi_n) &\sim CAR(W, \rho, \tau^2), \\ \rho &\sim Beta(a_\rho, b_\rho), \\ \tau^2 &\sim IG(a_\tau, b_\tau), \end{aligned}$$

where we consider the adjusted CAR specification for the spatial random effect, with

$$\phi_i \mid \boldsymbol{\phi}_{-i} \sim N\left(\rho \sum_{j=1}^{n} \frac{w_{ij}}{w_{i+}} \phi_j, \frac{\tau^2}{w_{i+}}\right), \qquad i = 1, \dots, n.$$

- Produce a code in STAN that produce a sample from the posterior distribution of interest.
- Sample $5\,000$ realizations from the posterior distribution of interest, after $2\,000$ burnin iterations. Check the convergence of the sampled chains.
- Plot the random effects associated at each spatial location.
- Write a function that estimate the model without the spatial component. Test if the first model is significantly different from this simpler model.