Spatially varying coefficient models for areal data

Modelli a coefficienti variabili nello spazio per dati areali

Massimo Ventrucci, Maria Franco-Villoria and Haavard Rue

Abstract We discuss the use of penalized complexity priors for spatially varying coefficient models, introducing a natural base model choice that corresponds to a constant coefficient (no variation in space). We illustrate the use of these priors in a case study on air pollution and hospital admissions in Turin, Italy.
Abstract *In questo lavoro estendiamo la classe di distribuzioni a priori nota come Penalized Complexity Priors (PC priors) al caso di modelli a coefficienti variabili (VCM) per dati areali. Viene descritta una nuova parametrizzazione per il VCM, che facilit la definizione della PC prior. I metodi proposti sono illustrati su un caso di studio di epidemiologia spaziale, riguardante leffetto dellinquinamento atmosferico sul rischio di ospedalizzazione nella provincia di Torino.*

Key words: Varying Coefficient models, PC prior, INLA, RW1

1 Introduction

Varying coefficient models (VCMs) [1] are useful in the presence of an *effect modifier*, a variable that “changes” the effect of a covariate of interest on the response. Consider the simple case where there are n observational units indexed by $i = 1, \dots, n$ and one covariate x_i whose effect on the response y_i depends on another variable z_i ; the latter could be a continuous variable (e.g. temperature) or a

Massimo Ventrucci
University of Bologna, e-mail: massimo.ventrucci@unibo.it

Maria Franco-Villoria
University of Torino e-mail: maria.francovilloria@unito.it

Haavard Rue
King Abdullah University of Science and Technology e-mail: haavard.rue@kaust.edu.sa

time/space index (day, region, etc). Assuming y_i belonging to the exponential family, the linear predictor of a generalized VCM is

$$\eta_i = \alpha + f(z_i)x_i \qquad i = 1,\dots,n. \tag{1}$$

We follow a Bayesian hierarchical framework where the varying coefficient $f(z_i)$ in Eq. (1) is described by a vector of random effects $\theta = (\theta_1,\dots,\theta_n)^\top$ distributed at prior as a Gaussian Markov Random Field (GMRF) [7]. A GMRF is a multivariate normal distribution with mean vector μ and a sparse precision $Q(\tau)$ that depends on some hyper-parameters τ and whose non zero pattern specifies conditional dependencies among neighbouring random effects. For areal data, the index $i = 1,\dots,n$ indicates each of the non overlapping regions in a lattice. To capture correlation structure between the VCs in neighbouring regions, the standard approach is to use conditionally autoregressive (CAR) models proposed by Besag (1974) [3]. Models for areal data have been widely discussed in the literature and are useful, for example, in epidemiological studies [4], where data are not available at individual level but only at some aggregated level such as municipality, zip code, etc.

1.1 Penalized complexity (PC) priors and the base model

A useful parametrization for the varying coefficient in Model (1) is $\theta_i = \beta + \delta_i$, where δ_i indicates deviation from a constant slope β at value z_i. The model turns into a simple linear regression model when the varying coefficient $f(z)$ is constant over z, i.e. when $\delta_i = 0\ \forall i$. The linear case can be regarded as a *base model*, while the VCM can be seen as a flexible extension of it. If we consider $f(z)$ in terms of the vector of random effects $\theta = (\theta_1,\dots,\theta_n)^\top$ introduced in Section 1, the base model can be obtained setting the hyper-parameters τ to a particular value. Elicitation of priors for precision parameters is a long standing topic in the literature on hierarchical Bayesian models. Simpson et al. (2017) [2] recently introduced a new framework for building priors that avoid overfitting denoted as *Penalized Complexity (PC) priors*. PC priors are computed based on specific principles in which a model component is seen as a flexible parametrization of a base model. The idea is to penalize model complexity, defined in terms of distance from the base model, in such a way that the base model is favoured unless the available data support a more flexible one.

2 PC prior for spatially varying coefficient in areal data

The spatially varying coefficient $\theta = (\theta_1,...,\theta_n)^\top$ follows an Intrinsic Conditional Autoregressive (ICAR) model [3]:

$$\theta_i|\theta_{-i},\tau \sim N\left(\sum_{j:i\sim j}\frac{\theta_j}{n_i},(n_i\tau)^{-1}\right)$$

where $i \sim j$ denotes neighbouring regions (sharing a common border) and n_i denotes the number of neighbours of region i. The joint distribution for θ is

$$\pi(\theta|\tau) = (2\pi)^{-(n-1)/2}(|\tau R|^*)^{1/2}\exp\left\{-\frac{\tau}{2}\theta^\top R\theta\right\} \quad (2)$$

where the structure matrix R is a singular matrix (whose null space is 1) with entries:

$$R_{i,j} = \begin{cases} n_i & i=j \\ -1 & i\sim j \\ 0 & otherwise, \end{cases}$$

which is equivalent to assuming a random walk of order 1 (RW1) prior on $\theta = (\theta_1,\dots,\theta_n)^\top$. The RW1 is a first-order intrinsic Gaussian Markov Random Field ([7] ch. 3) and describes deviations from an arbitrary overall level. In the context of a VCM it is natural to interpret the latter as the constant slope β. In this sense, the RW1 characterizes a varying coefficient in a very intuitive way: it is a prior that shrinks towards a natural base model given by $f(z_i) = \beta$, $\forall i = 1,...,n$, with τ controlling the amount of shrinkage. When $\tau^{-1} = 0$ we have $f(z_i) = \beta$, which implies the linear regression model, $\eta_i = \alpha + \beta x_i$. For $\tau^{-1} > 0$, $f(z_i)$ incorporates higher degree of complexity w.r.t. the constant slope, leading to the flexible VCM. The PC prior is an exponential distribution on the distance scale (measured using the Kullback-Leibler divergence [8]). A change of variable gives the PC prior in the scale of the precision. For a generic Gaussian Random effect conditional on τ, the PC prior for τ is the Gumbel type 2 with density,

$$\pi(\tau) = 0.5\lambda\tau^{-1.5}\exp(-\lambda\tau^{-0.5}); \quad (3)$$

for more details see [2]. The parameter λ in Eq. (3) can be selected through a user-defined scaling approach. The user can encode the information available (at prior) on the degree of flexibility of the VCM model with respect to the base model based. [2] suggest eliciting the probability of a tail event regarding the marginal standard deviation, i.e. $Pr(1/\sqrt{\tau} > U) = a$, which yields $\lambda = -\log(a)/U$.

3 Application: PM$_{10}$ and hospital admissions in Torino, Italy

Data on daily hospital admission are available from hospital discharge registers for the 315 municipalities in the province of Torino, Italy in 2004. In total, there are 12743 residents hospitalized for respiratory causes, aggregated by municipality and day. On the other hand, daily particular matter PM$_{10}$ ($\mu g/m^3$) data are available at municipality level, as estimates based on daily average PM$_{10}$ concentration. Average temperature (Kelvin degrees) is also available at each municipality and day. The application goal is to estimate the effect of PM_{10} on the risk of hospitalization for respiratory causes. We consider the following model:

$$\begin{aligned} y_{it} &\sim Poisson(E_{it}\exp(\eta_{i,t})) \\ \eta_{i,t} &= \alpha_t + u_i + \gamma temp_{i,t} + \beta_i PM_{10,i,t} \quad (4) \\ (\alpha_1,...,\alpha_{366})^\top &\sim RW2(\tau_{rw2}) \quad (5) \\ (u_1,...,u_{315})^\top &\sim BYM(\tau_{bym},\rho_{bym}) \quad (6) \\ (\beta_1,...,\beta_{315})^\top &\sim ICAR(\tau_{ICAR}) \quad (7) \end{aligned}$$

where y_{it} and E_{it} are the observed and expected number of hospitalizations in municipality $i = 1,\dots,315$ and day $t = 1,\dots,366$ respectively and $\exp(\eta_{i,t})$ is the relative risk of hospitalization in municipality i and time t. The pollution covariate $PM_{10,i,t}$ is taken as the sum of estimated daily average concentrations in the three days before t, in region i. The pollution effect, β_i, is allowed to vary from municipality to municipality.

Random effects (5) and (6) capture residual temporal and spatial structure, respectively. The temporal random effects are assigned a RW2. The spatial random effect u_i is the sum of two spatially structured and unstructured random effects associated to municipality i, as defined by the popular BYM (Besag, York and Mollié) model [5]. We follow the BYM parametrization introduced by [6] that has two hyperparameters: a marginal precision τ_{bym}, that allows shrinkage of the risk surface to a flat field, and a mixing parameter $\rho_{bym} \in (0,1)$, that handles the contribution from the structured and unstructured components. We use the PC priors derived in [6] for (τ_{bym},ρ_{bym}), while we use the PC prior in Eq. 3 for τ_{rw2} and τ_{ICAR}.

Figure 1 displays the municipality specific posterior relative risk for a $10\mu g/m^3$ increase in PM_{10} (panel a). The values obtained range from 1.06 to 1.2 and are largely comparable with estimates reported in literature. Panel (b) in Figure 1 shows the posterior probability of an increased risk associated to pollution, demonstrating that changes in the VCs across municipalities are substantial, with Turin (the *hotspot* in the south-east area) showing the largest increased risk.

Results from a small sensitivity analysis on the choice of U and a indicate that posterior relative risks remain basically unchanged across the different prior scenarios, unless a prior for the precision that puts a lot of probability mass around the base model is used, in which case the risk pattern is more shrunk towards no variation.

4 Discussion and Future Work

Elicitation of a prior for the hyper-parameters τ is a crucial aspect for practitioners who wish to specify a Bayesian VCM model, as τ regulates the degree of flexibility of the VCM w.r.t. the base model. Regardless the chosen model for the varying coefficient, a suitable PC prior shrinking to a sensible base model can always be defined

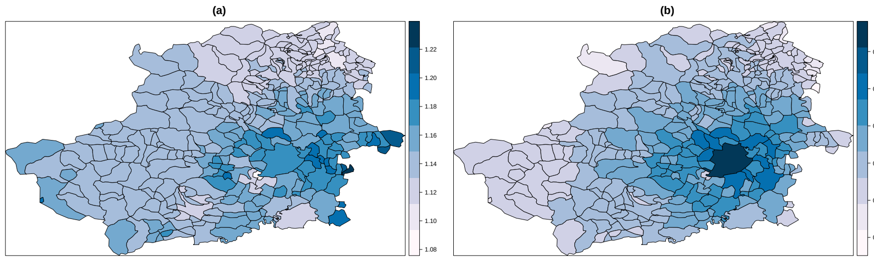

Fig. 1 (a) Posterior relative risks for $10\mu g/m^3$ increase in PM_{10} and (b) posterior probability for an increased risk, $P(\beta_i > 0|y)$.

through the application of predefined principles. In our case study, there seems to be evidence in the data for a varying effect of pollution. From an epidemiological point of view, there seems to be two possible explanations for a spatially-varying pollution effect. First, the result might be the effect of an unobserved confounding variable which is not captured by the random effects in the model. Second, the PM_{10} chemical composition might change substantially over space, so that the PM_{10} may be more or less dangerous for people, according to where they live. An alternative parametrization for the ICAR model is currently being developed.

References

1. Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.
2. Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Srbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28.
3. Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). series b. *Journal of the Royal Statistical society*, 36(2):192–225.
4. Banerjee, S., Carlin, B., and Gelfand, A. (2015). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press/Chapman & Hall. Monographs on Statistics and Applied Probability.
5. Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–21.
6. Riebler, A., Srbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165. PMID: 27566770.
7. Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.
8. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.