A unified view on Bayesian varying coefficient models

Maria Franco-Villoria[*]

Department of Economics "Marco Biagi", University of Modena and Reggio Emilia, Italy
e-mail: maria.francovilloria@unimore.it

Massimo Ventrucci

Department of Statistical Sciences, University of Bologna, Italy
e-mail: massimo.ventrucci@unibo.it

and

Håvard Rue

*CEMSE Division, King Abdullah University of Science and Technology, Thuwal,
Saudi Arabia*
e-mail: haavard.rue@kaust.edu.sa

Abstract: Varying coefficient models are useful in applications where the effect of the covariate might depend on some other covariate such as time or location. Various applications of these models often give rise to case-specific prior distributions for the parameter(s) describing how much the coefficients vary. In this work, we introduce a unified view of varying coefficients models, arguing for a way of specifying these prior distributions that are coherent across various applications, avoid overfitting and have a coherent interpretation. We do this by considering varying coefficients models as a flexible extension of the natural simpler model and capitalising on the recently proposed framework of penalized complexity (PC) priors. We illustrate our approach in two spatial examples where varying coefficient models are relevant.

Keywords and phrases: INLA, overfitting, penalized complexity prior, varying coefficient models.

Contents

[*]Corresponding author.

1. Introduction

Varying coefficient models (VCMs, [24]) can be seen as a general class of models that encompasses a large number of statistical models as special cases: the generalized linear model, generalized additive models, dynamic generalized linear models or even the more recent functional linear models. They can also be seen as a particular case of structured additive regression (STAR) models [11].

VCMs arise in a vast range of applications, ranging from economics to epidemiology [22, 25, 13, 15, 33, 12, 9, 43]. In practice, varying coefficient models are useful in presence of an *effect modifier*, a variable that "changes" the effect of a covariate of interest on the response. For the sake of a general notation that includes all cases discussed in this paper, consider the triplet (y_t, x_t, z_t), $t = 1, ..., n$, observed on n observational units, with z being the variable modifying the relationship between the covariate x and the response y. Following [24] who introduce VCMs as an extension of generalized linear models [34], we assume y belonging to the exponential family and model the effect of covariate x in the scale of the *linear predictor* $\eta = g(\mu)$, which is linked to the mean response μ via the *link function* g. The linear predictor of a generalized VCM is

$$\eta_t = \alpha + \beta(z_t)x_t \qquad t = 1, ..., n,$$

where $\beta(z_t)$, $t = 1, ..., n$, is the varying regression coefficient (VC), that can be regarded as a stochastic process on the effect modifier domain. For ease of notation we will use β_t to denote $\beta(z_t)$.

Depending on the nature of the effect modifier, that can be either a continuous variable (e.g. temperature) or a time/space index (e.g. day or municipality), we can envision several models for the varying coefficient $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$. For instance, we can assume exchangeability over $t = 1, \ldots, n$ with $cor(\beta_i, \beta_j) = \xi$ for $i \neq j$ if there is no natural ordering among the values of z. If the effect modifier is time, β_t might be a 1st order autoregressive (AR1) and ξ the lag-one correlation, or a spline if we want to ensure smoothness. The coefficients may also vary in space in a continuous or discrete way, in which case a Gaussian random field with a certain covariance function or a conditionally autoregressive (CAR) model can be assumed, respectively. These models have been treated separately in the literature, along with priors specifically chosen for each different model (see e.g. [6, 21, 22]). Further, the choice of the model describing the behaviour of the varying coefficient and the prior assigned to its parameter(s) (that controls the flexibility of the VC), are usually made at two different stages.

In this work we argue that 1) regardless of the model assumed for $\boldsymbol{\beta}$, all models can be seen in a unified way as a flexible extension of a simpler model where the varying coefficient is instead constant, and 2) the prior can be specified coherently with this model conception so that the issue of model and prior choice is tackled jointly. Following these two points, we propose a unified view on varying coefficient models where the prior is built under the recently proposed Penalized Complexity (PC) Prior framework [38].

The flexibility that VCMs offer can be desirable in certain applications and much work has been devoted to the development of flexible models. In our view, the VCM arises naturally from a simpler model; i.e. we can consider increasing the flexibility of the simple linear regression model $\eta_t = \alpha + \beta x_t$, $t = 1, \ldots, n$ by allowing the coefficient β to vary over t. Common choices of the prior might lead to overfitting, i.e. might *push* the model away from the simpler model even when a more flexible one is not appropriate [18, 19, 38]. The importance of using priors for VCMs that avoid overfitting is now beginning to be acknowledged in the literature by some authors such as [7] and [27], who consider shrinkage priors for the variance parameter in a time-varying setting, the latter allowing to adapt the shrinkage locally. We propose to use the more general PC prior approach to define priors that guarantee shrinkage to a simpler model for any kind of hyperparameter(s), including variance parameters. By treating varying coefficient models in a unified way we can use a single approach for doing so. The Penalized Complexity (PC) Prior framework considers a model component as a flexible extension of a simpler version of the model component, referred to as the base model. PC priors are defined on the scale of the distance from the base model and then transferred to the scale of the original parameter by a standard change of variable transformation. This strategy can be applied independently of the model choice for $\boldsymbol{\beta}$ describing the VC in a unique way, as the base model can always be easily identified in terms of a value for ξ. In this sense, we propose a coherent framework for building varying coefficient models.

The plan of the paper is as follows. Section 2 presents varying coefficient models in a unified way, while the general framework to construct PC priors is briefly reviewed in Section 3. In Section 4, several PC priors for ξ are derived under different model choices for $\boldsymbol{\beta}$, focusing first on the unstructured case (Section 4.1), where the realizations of the VC are assumed to be exchangeable. Structured cases, such as time and space are presented in Sections 4.2 and 4.3. Section 4.4 discusses the properties of PC priors and how these compare to other priors. Examples are illustrated in Section 5. The paper closes with a discussion in Section 6.

2. A unified view on varying coefficient models

Let us now specify a Bayesian generalized VCM, seeing it as a flexible extension of the simple generalized linear model $\eta_t = \alpha + \beta_0 x_t$, which will be denoted as *base model*; this can be thought of as the fit obtained if data do not show evidence for a varying coefficient but for a constant regression coefficient instead. Without loss of generality, we can assign the prior $\beta_0 \sim \mathcal{N}(0,1)$ to the base model:

$$
\begin{aligned}
\eta_t &= \alpha + \beta_0 x_t \qquad t = 1, ..., n, \\
\beta_0 &\sim N(0,1).
\end{aligned}
\tag{1}
$$

If we believe that the covariate effect is not constant in z, we can allow for deviation from β_0 in the form of a varying coefficient model,

$$
\begin{aligned}
\eta_t &= \alpha + \left(\beta_0 + \beta_t\right) x_t \qquad t = 1, ..., n, \\
\boldsymbol{\beta}|\xi &\sim \pi(\boldsymbol{\beta}|\xi),
\end{aligned}
\tag{2}
$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ is a vector of random effects defining a stochastic process over z, denoted as $\pi(\boldsymbol{\beta}|\xi)$ with ξ the associated hyperparameter(s).

In what follows we will assume the linear predictor $\eta_t = \alpha + \left(\beta_0 + \beta_t\right) x_t$ in Eq. (2) and consider different Gaussian models for $\pi(\boldsymbol{\beta}|\xi)$. We will focus on the VC models mostly used in applications, each of them representing a specific *extension* of the base model in Eq. (1), allowing us to view all the various cases in a unified manner.

2.1. The unstructured case

The simplest correlation structure for random effects is to assume that they are exchangeable; this is commonly used to account for dependence among repeated measures in longitudinal models [29]. If $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)^{\mathsf{T}}$ are exchangeable over $t = 1, \ldots, n$, then $\boldsymbol{\beta} \sim \mathcal{N}(0, \tau^{-1} \boldsymbol{R}(\tilde{\rho}))$, where the correlation matrix is

$$
\boldsymbol{R}(\tilde{\rho}) =
\begin{bmatrix}
1 & \tilde{\rho} & \ldots & & \tilde{\rho} \\
\tilde{\rho} & 1 & \tilde{\rho} & \ldots & \tilde{\rho} \\
\cdot & & \cdot & & \cdot \\
\cdot & & & \cdot & \cdot \\
\tilde{\rho} & \tilde{\rho} & \ldots & \tilde{\rho} & 1
\end{bmatrix}
\tag{3}
$$

and τ is a precision parameter. For $\boldsymbol{R}(\tilde{\rho})$ to be positive definite, $-1/(n-1) < \tilde{\rho} < 1$ [38]. In the following, we consider $0 \le \tilde{\rho} < 1$.

In this case, Model (2) can be reparametrized as $\eta_t = \alpha + \beta_t x_t$, $t = 1, \ldots, n$ with

$$\boldsymbol{\beta} \sim N(0, \boldsymbol{R}(\rho)), \tag{4}$$

assuming unit marginal variance with no loss of generality (in practical applications τ can either be fixed to a single value if known or it could be considered as a parameter on which we impose a prior distribution). A sensible base model is $\rho = 1$, corresponding to $\beta_t = \beta \ \ \forall t$.

2.2. The structured case: temporal variation

In many real life applications the values of the effect modifier follow a natural ordering, e.g. time, so that it is not realistic to assume exchangeability of β_t. Instead, autoregressive (AR) models from time series analysis can be adopted [40]. An alternative is to consider the varying coefficient as a smooth function. A popular model in the context of smoothing with splines is the 2nd order random walk (RW2), that can be seen as a discrete representation of a continous (integrated) Wiener process that retains the Markov property and is computationally efficient [30]. It is also used in P-splines [32] where a RW2 is assigned to the coefficients of local B-spline basis functions. In the following we consider three cases: the 1st order autoregressive (AR1) and the 1st and 2nd order random walk (RW1, RW2). In all three cases, we always assume the linear predictor reported in Eq. (2), but consider different models for β_t.

2.2.1. The autoregressive model of first order

The most common model for dependence on time is the autoregressive process of first order (AR1), the discrete-time analogue of the Ornstein-Uhlenbeck process, characterized by a correlation function with exponential decay rate. A 1st order autoregressive prior on the varying coefficient is $\beta_t = \tilde{\rho}\beta_{t-1} + w_t$, where $|\tilde{\rho}| < 1$ represents the lag-one correlation, $w_t \sim \mathcal{N}(0, \tau^{-1}(1 - \tilde{\rho}^2))$, $t = 2, ..., n$, and $\beta_1 \sim \mathcal{N}(0, \tau^{-1})$. The varying coefficient has a joint distribution given by $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \tau^{-1}\boldsymbol{R}(\tilde{\rho})\right)$ with $\boldsymbol{R}(\tilde{\rho})_{ij} = (\tilde{\rho}^{|i-j|})$ and τ a precision parameter. Similarly to Section 2.1, we can reparametrize Model (2) as $\eta_t = \alpha + \beta_t x_t$, $t = 1, \ldots, n$, so that

$$\boldsymbol{\beta} \sim \mathcal{N}\left(0, \boldsymbol{R}(\rho)\right) \tag{5}$$

and $\beta_1 \sim N(0, 1)$. In this case the base model is $\rho = 1$, i.e. no change in time.

2.2.2. Random walk model of order one and two

We can consider the varying coefficient $\boldsymbol{\beta}$ in Eq. (2) as a smooth stochastic process on the effect modifier scale. The equivalence between smoothing splines

and Gaussian processes was shown in [26]. In a Bayesian framework, smoothing models are obtained using a random walk on the varying coefficients. A random walk is an intrinsic Gaussian Markov Random Field (IGMRF, [36] ch. 3), i.e. a process with the multivariate Gaussian density

$$\pi(\boldsymbol{\beta}|\tau) = (2\pi)^{-\texttt{rank}(\boldsymbol{K})/2}(|\tau\boldsymbol{K}|^*)^{1/2}\exp\left\{-\frac{\tau}{2}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{K}\boldsymbol{\beta}\right\} \tag{6}$$

where the structure matrix $\boldsymbol{K}$ is sparse and rank deficient ($\text{rank}(\boldsymbol{K}) = n - r$), τ is a scalar precision parameter and $|\tau\boldsymbol{K}|^*$ is the generalized determinant.

The structure matrix encodes the conditional dependencies among the coefficients $\boldsymbol{\beta}$. To avoid scaling issues inherent in RW models, such as dependence on the graph, [39] propose to scale the matrix $\boldsymbol{K}$ by a factor equal to the geometric mean of the diagonal elements of the generalized inverse of $\boldsymbol{K}$, so that the marginal variance (subject to appropriate sum to zero constraints) is equal to 1. The rank deficiency of the structure matrix also identifies the order r of the IGMRF. Model (6) describes deviation from a polynomial model of degree $r-1$: e.g. a constant for RW1 ($r = 1$) and a linear trend for RW2 ($r = 2$). This means we need to impose a sum to zero constraint on $\boldsymbol{\beta}$ to avoid confounding with β_0 in Eq. (2), with the difference that using a RW2 will result in a smoother fit than if a RW1 is used. Without loss of generality, we assume equally spaced locations. The case of irregularly spaced locations differs only in the structure matrix $\boldsymbol{K}$ and the constraint, that has to be modified with the inclusion of appropriate integration weights [30].

The precision parameter τ regulates the amount of shrinkage towards the base model, that corresponds to $\tau = \infty$.

2.3. The structured case: spatial variation

Spatially structured models include the cases of continuous or discrete spatial variation. In the former case, the effect modifier is the pair of (scaled) latitude and longitude coordinates, $\boldsymbol{z}_t = \{\text{lat}_t, \text{lon}_t\}$ and β_t can be assumed as a realization from a spatial process. The class of Gaussian Random Field (GRF) models equipped with a *Matérn* covariance is the most popular model [42]. For areal data, the spatial units are identified by a one-dimensional region index, with no unique ordering among the regions. Neighbouring regions are assumed to be correlated, and the neighbourhood structure can be coded into a structure matrix. To model β_t, the standard approach is to use conditionally autoregressive (CAR) models proposed by [4]; see [44, 41] for applications.

2.3.1. Areal spatial variation

Models for areal data have been widely discussed in the literature and are useful, for example, in epidemiological studies [1], where data are not available at individual level but only at some aggregated level such as municipality or zip code (see Figure 6 for an example).

Assume the linear predictor in (2) where $t = 1, \ldots, n$ indicates each of the non overlapping regions in a lattice. Areas i and j are considered as neighbours, denoted as $i \sim j$, if they share a common border. The spatially varying coefficient $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathsf{T}}$ follows an Intrinsic Conditional Autoregressive (ICAR) model [4]:

$$\beta_t | \boldsymbol{\beta}_{-t}, \tau \sim \mathcal{N} \left(\frac{1}{n_t} \sum_{j:t \sim j} \beta_j, (n_t \tau)^{-1} \right)$$

with n_t the number of neighbours of region t and τ a precision parameter. The joint distribution for $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}|\tau) = (2\pi)^{-(n-1)/2} (|\tau \boldsymbol{K}|^*)^{1/2} \exp \left\{ -\frac{\tau}{2} \boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{\beta} \right\} \tag{7}$$

where the structure matrix $\boldsymbol{K}$ is singular with null space $\mathbf{1}$ and entries:

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

The base model, corresponding to no variation over area, is $\tau = \infty$.

2.3.2. Continuous spatial variation

In this case, $t = (\text{lat}_t, \text{lon}_t)$, properly scaled, represents location within a spatial region $D \subseteq \mathbb{R}^2$ and the spatially varying coefficient can be seen as a realization of a Gaussian random field (GRF) with a *Matérn* covariance function characterized by the marginal variance τ^{-1} and range parameter ϕ. These two parameters cannot be estimated consistently under infill asymptotics [45, 47], but only a function of those such as the product or the ratio, depending on the smoothness of the GRF.

Assuming the linear predictor in (2) the spatially varying coefficient

$$\boldsymbol{\beta} \sim \mathcal{N} \left(\mathbf{0}, \tau^{-1} \boldsymbol{R}(\phi) \right) \tag{8}$$

with $\boldsymbol{R}(\phi)_{ij} = (C(||i - j||))$, $C(\cdot)$ is a *Matérn* correlation function with fixed smoothness ν:

$$C(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{8\nu}h}{\phi} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{8\nu}h}{\phi} \right),$$

K_{ν} is the modified Bessel function of second kind and order ν and h represents the distance between any pair of locations. The base model in this case corresponds to $\tau = \infty$, $\phi = \infty$.

3. Review of Penalized Complexity (PC) Priors

The prior for all the hyperparameters in Section 2 can be built in a coherent way regardless of the assumed model for $\boldsymbol{\beta}$ using penalized complexity priors. In this section we summarize the four main principles underpinning the construction of PC priors, namely: support to Occam's razor (parsimony), penalisation of model complexity, constant rate penalisation and user-defined scaling. For a more detailed presentation of these principles the reader is referred to [38].

Let f_1 denote the density of a model component $\boldsymbol{w}$ where ξ is the parameter for which we need to specify a prior. The base model, corresponds to a fixed value of the parameter $\xi = \xi_0$ and is characterized by the density f_0.

1. The prior for ξ should give proper shrinkage to ξ_0 and decay with increasing complexity of f_1 in support of Occam's razor, ensuring parsimony; i.e. the simplest model is favoured unless there is evidence for a more flexible one.
2. The increased complexity of f_1 with respect to f_0 is measured using the Kullback-Leibler divergence [KLD, 28],

$$\mathrm{KLD}(f_1||f_0) = \int f_1(w) \log\left(\frac{f_1(w)}{f_0(w)}\right) dw.$$

   For ease of interpretation, the KLD is transformed to a unidirectional distance measure

$$d(\xi) = d(f_1||f_0) = \sqrt{2\mathrm{KLD}(f_1||f_0)} \tag{9}$$

   that can be interpreted as the distance from the flexible model f_1 to the base model f_0.
3. The PC prior is defined as an exponential distribution on the distance,

$$\pi(d(\xi)) = \lambda \exp(-\lambda d(\xi)), \tag{10}$$

   with rate $\lambda > 0$. The PC prior for ξ follows by a change of variable transformation.
4. The user must select λ based on his prior knowledge on the parameter of interest (or an interpretable transformation of it $Q(\xi)$). This knowledge can be expressed in terms of a probability statement, e.g. $\mathbb{P}(Q(\xi) > U) = a$, where U is an upper bound for $Q(\xi)$ and a is a (generally small) probability.

4. PC priors for varying coefficient models

The PC prior framework offers a unified approach for constructing priors for all the various models considered in Section 2 while guaranteeing proper shrinkage to the base model (1). Within this framework, we can always build the prior for

the corresponding flexibility parameter(s) as an exponential distribution on the distance from the base model, then transform it back to the original scale. Here we present the main results about the derivation of PC priors for the flexibility parameters of the models considered in Section 2, while technical details can be found in the appendix.

4.1. The unstructured case

As described in Section 2.1, the base model for Model (4) is $\rho = 1$. The PC prior for ρ:

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{2\sqrt{1-\rho}(1-\exp(-\theta))}, \quad 0 \leq \rho < 1, \quad \theta > 0. \tag{11}$$

The prior is scaled in terms of θ based on the prior belief that the user has about the parameter ρ in the form of (U, a) such that $\mathbb{P}(\rho > U) = a$. The corresponding value for θ is given by the solution of the equation

$$\frac{1 - \exp(-\theta\sqrt{1-U})}{1 - \exp(-\theta)} = a$$

that has to be solved numerically, provided that $a > \sqrt{1-U}$. The PC prior in (11) is illustrated in Figure 1 left panel.

4.2. The structured case: temporal variation

4.2.1. The autoregressive model of first order

For Model (5), [40] derive the PC prior with base model at $\rho = 1$ as

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{2\sqrt{1-\rho}(1-\exp(-\sqrt{2}\theta))}, \quad |\rho| < 1, \quad \theta > 0. \tag{12}$$

The user can incorporate information on his/her prior belief about the size of the correlation parameter by setting U and a so that $\mathbb{P}(\rho > U) = a$. To work out θ the equation

$$\frac{1 - \exp(-\theta\sqrt{1-U})}{1 - \exp(-\sqrt{2}\theta)} = a, \quad a > \sqrt{(1-U)/2}$$

needs to be solved numerically for θ as in the unstructured case. The PC prior in (12) is illustrated in Figure 2 left panel.
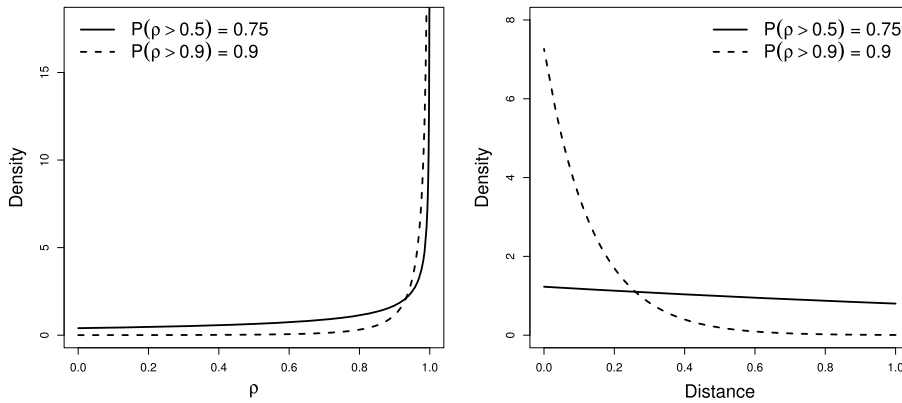
FIG 1. *Left panel: PC prior for ρ under the exchangeable model; the base model is $\rho = 1$. Right panel: the same PC prior plotted in the distance scale, $d(\rho)$; the base model is at $d(\rho) = 0$.*

FIG 2. *Left panel: PC prior for ρ under the AR(1) model; the base model is $\rho = 1$. Right panel: the same PC prior plotted in the distance scale, $d(\rho)$; the base model is at $d(\rho) = 0$.*
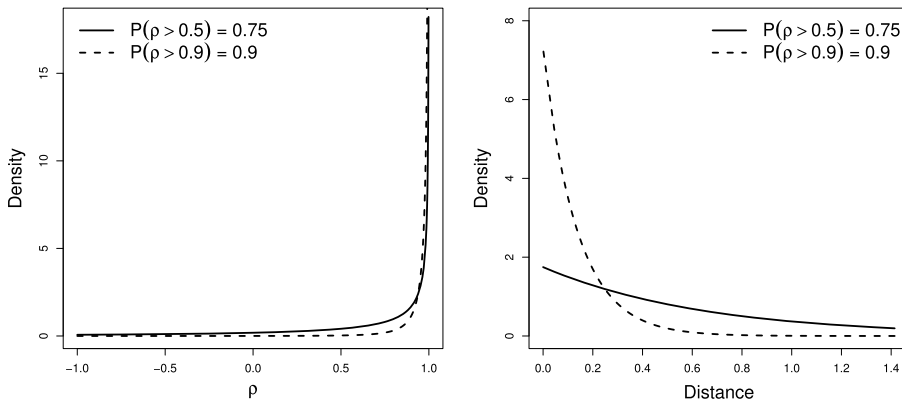
4.2.2. Random walk model of order one and two

In the case of Model (6), the amount of deviation from the base model depends on τ, with base model at $\tau = \infty$. [38] derive the PC prior for τ as a Gumbel$(1/2, \theta)$ type 2 distribution

$$\pi(\tau) = \frac{\theta}{2} \tau^{-3/2} \exp\left(-\theta/\sqrt{\tau}\right), \quad \tau > 0, \theta > 0. \tag{13}$$

To derive the scaling parameter θ, [38] suggest to bound the marginal standard deviation, $1/\sqrt{\tau}$. This way it is sufficient to specify (U, a) and solve $\mathbb{P}(1/\sqrt{\tau} > U) = a$ for θ, which gives $\theta = -\log(a)/U$. To aid the user in specifying parameters (U, a), [38] provide a general rule of thumb: "the marginal
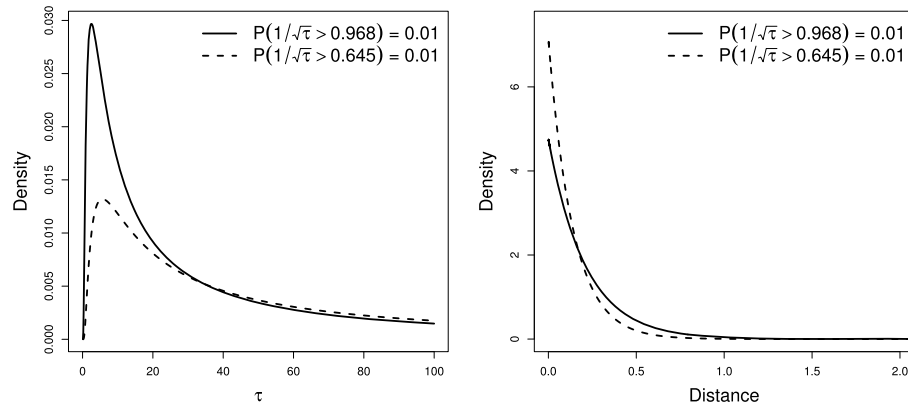
FIG 3. *Left panel: PC prior for τ under the RW model; the base model is $\tau = \infty$. Right panel: the same PC prior plotted in the distance scale, $d(\tau)$; the base model is at $d(\tau) = 0$.*

standard deviation of $\boldsymbol{\beta}$ with $\boldsymbol{K} = \boldsymbol{I}$, after the type-2 Gumbel distribution for τ is integrated out, is about $0.31U$ when $\alpha = 0.01$"; e.g. if we think a standard deviation of approximately 0.3 is a reasonable upper bound for the random effect (i.e. the varying coefficient), we need to set $U = 0.3/0.31 = 0.968$. The PC prior in (13) is illustrated in Figure 3 left panel.

4.3. The structured case: spatial variation

4.3.1. Areal spatial variation

It is clear from Eq. (7) that the ICAR model can be seen as a RW1 model (Eq. (6) with $\text{rank}(\boldsymbol{K}) = n - 1$), and hence the PC prior for τ is (13) as in the previous section.

4.3.2. Continuous spatial variation

PC priors for the range and marginal variance parameters of a GRF with *Matérn* covariance function have been derived by [20]. The joint PC prior for (τ, ϕ) with base model at $\tau = \infty$, $\phi = \infty$:

$$\pi(\tau, \phi) = \lambda_\phi \phi^{-2} \exp\left(-\lambda_\phi \phi^{-1}\right) \frac{\lambda_\tau}{2} \tau^{-3/2} \exp\left(-\frac{\lambda_\tau}{\sqrt{\tau}}\right), \qquad \tau > 0, \phi > 0 \quad (14)$$

where, once the user fixes U_ϕ, a_ϕ, U_τ, a_τ such that $\mathbb{P}(\phi < U_\phi) = a_\phi$, $\mathbb{P}(1/\sqrt{\tau} > U_\tau) = a_\tau$ the parameters λ_ϕ, λ_τ are calculated as

$$\lambda_\phi = -\log(a_\phi) U_\phi, \qquad \lambda_\tau = -\frac{\log(a_\tau)}{U_\tau}.$$
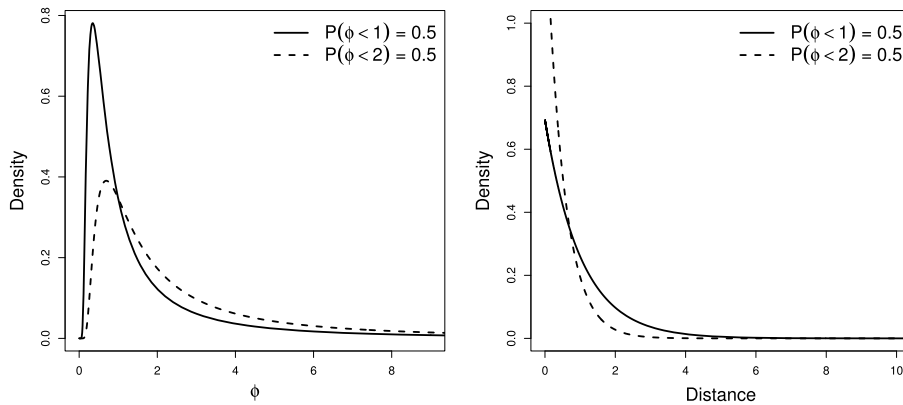
FIG 4. *Left panel: PC prior for range parameter ϕ of the* Matérn *covariance function; the base model is $\phi = \infty$. Right panel: the same PC prior plotted in the distance scale, $d(\phi)$; the base model is at $d(\phi) = 0$.*

Note that, as the result of a convenient reparametrization (see appendix A.4) the joint PC prior for (τ, ϕ) factorizes as the product of the marginal densities. The PC prior for ϕ is illustrated in Figure 4 left panel (while the PC prior for τ is the same as in Figure 3).

4.4. Properties of PC priors in the context of VCMs

In Eq. (10) the PC prior is defined as an exponential distribution on the distance scale $d = \sqrt{2\mathrm{KLD}(f_1||f_0)}$, which implies two important properties. First the exponential ensures constant rate penalization,

$$\frac{\pi(d + \delta)}{\pi(d)} = \frac{\lambda \exp(-\lambda(d + \delta))}{\lambda \exp(-\lambda d)} = r^\delta,$$

where $r = \exp(-\lambda)$ is the constant decay rate. The relative change in the density for adding an extra δ does only depend on δ, not on d. In many cases d will not be an easy-to-interpret measure of complexity, thus the memory-less property becomes a practical device to penalize increasingly flexible models. (An example where the distance is well interpretable is for the case of independent Gaussian random effects, where d corresponds to the marginal standard deviation of such random effects [38]).

A second important property is that the mode of the PC prior is at distance 0, meaning that PC priors naturally contract to the base model and prevent overfitting by construction. This is illustrated in the right panels of Figures 1, 2, 3, 4, where each of the presented PC priors is displayed in their distance scale $d = \sqrt{2\mathrm{KLD}(f_1||f_0)}$. [38] describe an *overfitting prior* as a prior that places insufficient mass at $d = 0$, suggesting that "a prior overfits if its density in a sensible parametrization is zero at the base model" (for justification of

this choice see [38] section 2.4). The idea is the following: while priors that contract to $d = 0$ avoid overfitting because they always give a chance for the base model to arise in the posterior, priors that go to 0 at $d = 0$ may incur in overfitting issues, because they may drag the posterior away from the base model, even when the latter is the true one. As an example, the conjugate Gamma for τ is an overfitting prior [38]. We believe this property is very important, as in the context of VCMs it is advisable to use priors that allow the *constant coefficient* to arise in the posterior, unless data show evidence for a varying coefficient.

4.4.1. Comparison with other priors

Plotting priors on the distance scale is a useful tool to judge the behaviour near the base model. Figure 5 displays three different priors for the lag-one correlation ρ of an AR1: the PC prior in Eq. (12), the reference prior [2, 3] and the uniform on $(-1, 1)$. All priors are plotted in the distance scale. The behaviour near the base model attained by the three priors is very different. The PC prior contracts to the base model as it peaks at minimum distance from the base model. The reference and uniform priors contract to the most complex model as they peak at maximum distance, $d = \sqrt{2}$. The uniform prior is the one that assigns less mass around $d = 0$ and it overfits according to the informal definition in [38]. A simulation study was conducted to investigate performance of the priors depicted in Figure 5. We considered a varying coefficient modelled as an AR1, focusing on two relevant cases: a first scenario (SC1) where the true VCM is close to the base model (i.e. a constant coefficient) and a flexible scenario (SC2) where the true VCM is far from the base model. See the supplementary material [17] for more details on the simulation study and full discussion of results. In summary, it was found that the three priors perform equally well in SC2, while in SC1 the PC prior outperforms the other two, especially with noisy data. The uniform achieved poorest performance, which we presume is due to the fact that this prior forces overfitting in the sense of [38]. Our findings for the AR1 case are in line with several works comparing PC priors with other prior choices for the remaining models considered in this paper, see the supplementary material for details [17].

5. Examples

In the previous section we have shown how PC priors for varying coefficient models can be derived in a unified way regardless of the model assumed for the VC. Here we illustrate their application in two spatial examples where varying coefficient models are relevant. All models are fitted within the `R-INLA` package [31] and the code is available in the supplementary material [16]. The dataset used in example 5.2 is freely available, while the data from the example in Section 5.1 cannot be published due to privacy issues, but the related `R-INLA` code is available using a simulated similar dataset.
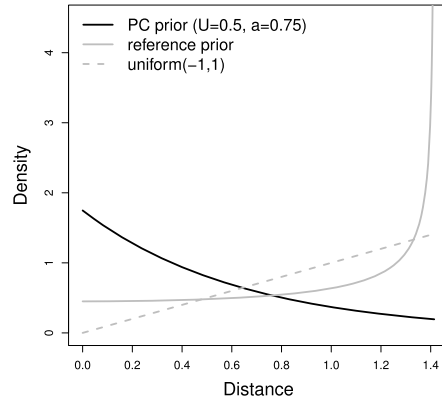
FIG 5. *The PC prior, reference and uniform prior for the lag-one correlation ρ of an AR1 model plotted in the distance scale.*

5.1. PM_{10} and hospital admissions in Torino, Italy

The goal is to estimate the effect of PM_{10} on the risk of hospitalization for respiratory causes using data on daily hospital admission 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. A reduced form of this dataset is available in the book by [8]. Daily average temperature (Kelvin degrees) and particular matter PM_{10} ($\mu g/m^3$) data are available at municipality level, the latter as estimates based on daily average PM_{10} concentration [14].

We consider the following model (all covariates are standardized), where the effect of PM_{10} is allowed to vary spatially across municipalities:

$$
\begin{aligned}
y_{i,t} &\sim \text{Poisson}(E_{i,t} \exp(\eta_{i,t})) \\
\eta_{i,t} &= \alpha_t + u_i + \gamma \text{temp}_{i,t} + \beta_0 \text{PM}_{10,i,t} + \beta_i \text{PM}_{10,i,t} \quad (15) \\
(\alpha_1, ..., \alpha_{366})^\mathsf{T} &\sim \text{cyclic RW2}(\tau_{\text{rw2}}) \quad (16) \\
(u_1, ..., u_{315})^\mathsf{T} &\sim \text{BYM}(\tau_{\text{bym}}, \gamma_{\text{bym}}) \quad (17) \\
(\beta_1, ..., \beta_n)^\mathsf{T} &\sim \text{ICAR}(\tau_{\text{icar}}) \quad (18)
\end{aligned}
$$

where $y_{i,t}$ and $E_{i,t}$ are the observed and expected number of hospitalizations in municipality $i = 1, \ldots, 315$ and day $t = 1, \ldots, 366$ respectively and $\exp(\eta_{i,t})$ is the relative risk of hospitalization in municipality i and time t. Temperature (temp) is introduced as a fixed effect, as it is well known to be a confounder for the relationship between air pollution and health. $PM_{10,i,t}$ is taken as the sum of estimated daily average concentrations in the three days before t, in region i.

With our model we are able to disentangle the mean effect of PM10 (β_0) expressing the overall change in the posterior relative risk for 1 $\mu g/m^3$ PM10 increase, from the varying cofficient β_i expressing the municipality-specific de-

viation from β_0. We impose a sum to zero constraint on the β_i's in Eq. (18) to ensure identifiability of β_0, with $\beta_0 \sim N(0, 1000)$.

The random effects (16) and (17) capture residual temporal and spatial structure, respectively. The temporal random effects are assigned a RW2 wrapped on a circle to ensure a cyclic trend over time; in practice, this is achieved by using a circulant precision matrix that constrains the first and last random effects to be the same, i.e. $\alpha_1 = \alpha_{366}$ (see [36], section 2.6.1 for details). The spatial random effect u_i is the sum of two random effects associated to municipality i, one spatially structured and one spatially unstructured, as defined by the popular BYM (Besag, York and Mollié) model [5]. We follow the BYM parametrization introduced by [35] and use the PC priors derived therein for the two hyperparameters of the BYM: a marginal precision τ_{bym}, that allows shrinkage of the risk surface to a flat field, and a mixing parameter $\gamma_{\mathrm{bym}} \in (0, 1)$, that handles the contribution from the structured and unstructured components. For ease of notation, in (17) we skip all the details and refer the reader to [35], formula (7).

Table 1 summarizes the selected U and a for all PC priors. We can use the practical rule of thumb described at the end of Section 4.2.2 to set an upper bound for the standard deviation. Weak prior knowledge suggests an upper bound for the marginal standard deviation approximately equal to $1, 3$ and 0.1 for the temporal trend (α_t), the spatial component (u_i) and the VC (β_i), respectively. For instance, the choice of $U = 0.1$ for β_i is to be interpreted as: there is roughly 95% probability that $\beta_i \in (e^{-0.1 \cdot 1.96}, e^{0.1 \cdot 1.96})$, i.e. there is little chance that the deviation in increased relative risk (associated to 1 $\mu\mathrm{g/m}^3$ increase in PM_{10}) is larger than 1.2 in a given area.

The change in the posterior relative risk for a 10 $\mu\mathrm{g/m}^3$ increase in PM_{10} is 1.002 (with 95% credible interval (0.998, 1.006)). Figure 6 (panel a) displays the posterior mean for β_i, i.e. the municipality specific deviations (in the linear predictor scale) from the mean effect of PM_{10}. Panel (b) in Figure 6 shows the posterior probability of an increased risk associated to pollution, demonstrating that changes in the varying coefficients across municipalities may only be substantial in the municipality of Turin (the *hotspot* in the south-east area). Looking at the prior vs posterior in Figure 7(a), we see that there seems to be some information in the data regarding τ_{icar} as prior and posterior are clearly separated.

From an epidemiological point of view, there seems to be two possible explanations for a spatially-varying pollution effect. First, the result might be due to 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.

Sensitivity analysis

An interesting question is how sensitive the model fit is to a change in the PC prior parameters U, a. Figure 7(b) displays posterior distributions for τ_{icar} under

three different settings (see Table 2) with increasing penalty for deviating from the base model. There does not seem to be a great effect of U on the posterior for τ_{icar} unless we impose a strong penalization for deviating from the base model (pc3). In terms of posterior relative risks, results (not reported here) 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.

TABLE 1
Summary of the PC prior parameters U and a used in model (15) for the precisions (τ) and the γ parameter.

PC prior	α_t (rw2)	u_i (BYM)	β_i (ICAR)
$\pi(\tau|U, a = 0.01)$	$U = 0.1/0.31$	$U = 3/0.31$	$U = 0.1/0.31$
$\pi(\gamma|U, a = 0.5)$	–	$U = 0.5$	–

TABLE 2
Summary of the PC prior parameters U and a for τ_{icar} used in the sensitivity analysis for Model (15).

PC prior parameters	pc1	pc2	pc3
U	$1/0.31$	$0.1/0.31$	$0.01/0.31$
a	0.01	0.01	0.01

A possible alternative could be to assume an exchangeable model for the varying coefficient. Given the large number of areas ($n = 315$) we considered it was more natural to assume the varying coefficients to be spatially structured but for similar applications with a small number of areas an exchangeable model could be used.

5.2. House prices in Baton Rouge, Louisiana

The dataset considered in this example is available in [1] and consists of selling prices (\$) of 70 single family homes in East Baton Rouge Parish, Louisiana, sold in June 1989. Living area (square feet) and other area (square feet) such as garden, garage, etc., are available as covariates, as well as the longitude (lon) and latitude (lat) coordinates. An extended version of this dataset is analyzed in [22]. The spatial locations of the houses sold can be seen in Figure 8, along with the border delimiting the parish of East Baton Rouge. Even though the expectation is that bigger houses with a bigger external area are more expensive than smaller ones, location plays an important role in determining the price of a house. Hence, we allow for a spatially varying effect of living area (area) and other area (Oarea) in the following model (where the covariates have been standardized):

$$\log(\text{price})_i = \alpha + \gamma_{lon}\text{long}_i + \gamma_{lat}\text{lat}_i + \beta_{a,i}\text{area} + \beta_{b,i}\text{Oarea} + \epsilon_i + e_i \quad (19)$$
$$(\beta_{a,1}, ..., \beta_{a,n})^{\mathsf{T}} \sim \mathcal{N}\left(\mathbf{0}, \tau_a^{-1}\boldsymbol{R}(\phi_a)\right) \quad (20)$$
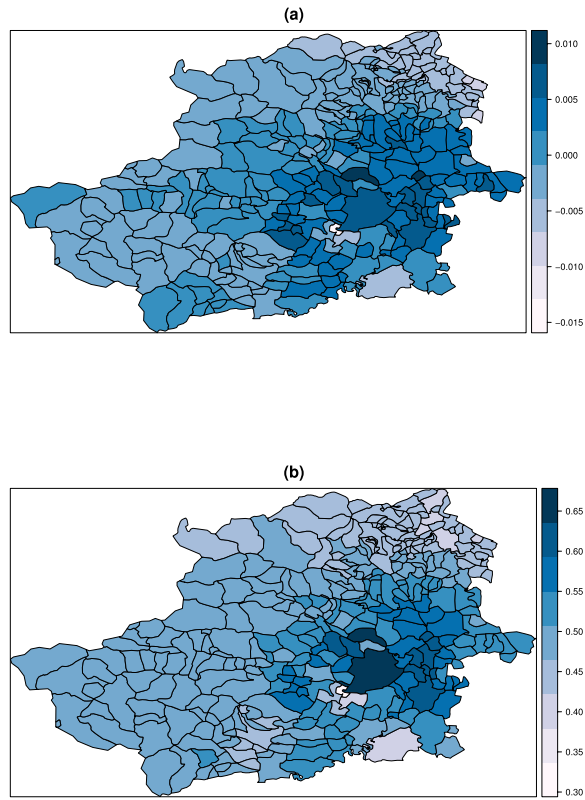
FIG 6. *Posterior mean for the varying coefficients β_i (panel a) and posterior probability $\mathbb{P}(\beta_i > 0|\boldsymbol{y})$ (panel b).*
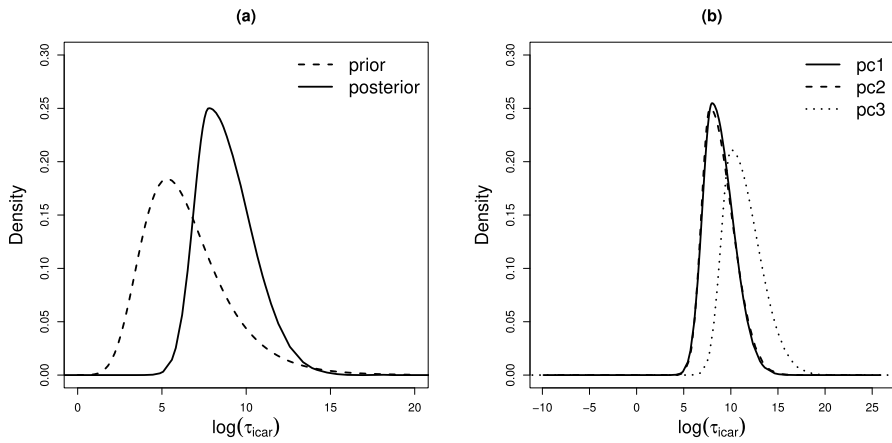

FIG 7. *Prior* vs *posterior comparison for the precision parameter τ_{icar} as specified in Table 1 (panel a) and posterior for τ_{icar} for each setting in Table 2 (panel b).*

$$(\beta_{b,1}, ..., \beta_{b,n})^\mathsf{T} \sim \mathcal{N}\left(\mathbf{0}, \tau_b^{-1}\boldsymbol{R}(\phi_b)\right) \tag{21}$$

$$\epsilon_i \sim \mathcal{N}(0, \tau_\epsilon^{-1}\boldsymbol{R}(\phi_\epsilon)) \tag{22}$$

$$e_i \sim \mathcal{N}(0, \tau_e^{-1}) \tag{23}$$

with $\boldsymbol{R}(\phi)$ as in Eq. (8). PC priors for the parameters of the *Matérn* covariance functions ϕ_a, τ_a, ϕ_b, τ_b and $\phi_\epsilon, \tau_\epsilon$ were scaled as follows. The maximum distance between observed locations is 5.12, so we set $U_\phi = 2$ and $a_\phi = 0.5$ so that $\mathbb{P}(\phi < 2) = 0.5$ for all ϕ_a, ϕ_b and ϕ_ϵ. Regarding the marginal standard deviation, prior knowledge on the scale of the response and of the covariates can be used to select U_τ and a_τ in a reasonable way; we set $U_\tau = 0.1/0.31$ and $a_\tau = 0.01$ for τ_a and τ_b (i.e. $\mathbb{P}(1/\sqrt{\tau} > 0.1/0.31) = 0.01$) and $U_\tau = 0.4/0.31$ and $a_\tau = 0.01$ for τ_ϵ (i.e. $\mathbb{P}(1/\sqrt{\tau} > 0.4/0.31) = 0.01$).

The posterior varying coefficient estimates for living area and other area are shown in Figure 8. The effect of living area on log selling price (panel a) changes depending on location and is greater than that of other area (panel b); in particular, there are two hot-spots where the effect appears to be greatest. The one on the left roughly corresponds to the area where Baton Rouge, capital of the state of Louisiana, is located. The bottom right corner corresponds to a district where household income is greater than that of the region as a whole.

The effect of other area on log selling price also varies spatially as it can be seen in Figure 8(b). In particular, the red spot on the left hand side is roughly located on downtown Baton Rouge, the historic area of the city. On the other hand, it seems plausible that for houses located on the outskirts of the main cities in the region, the variable other area does not have such a strong impact on house price.

A small sensitivity analysis (see Table 3), was carried out in order to assess the impact of varying U and a. The results (not shown here) seldom vary unless a PC prior for τ with nearly all the mass concentrated on the base model (`pc.b`) is used (as already observed in example 5.1). In practice, it is not possible to disentangle the effect of the range and marginal variance of a GRF. This results in sometimes different posterior means and distributions for the parameters under the remaining prior specifications in Table 3 but with essentially no differences in the fitted surfaces with respect to those shown in Figure 8. Given this difficulty in separating the effect of parameters ϕ and τ we opted to use an informative prior for the marginal variance, where U and a can be set in a more intuitive way, and a less informative prior for the range parameter.

TABLE 3

Summary of the PC prior parameters U and a used in model (19) for the precisions (τ) and the ϕ parameters in the sensitivity analysis.

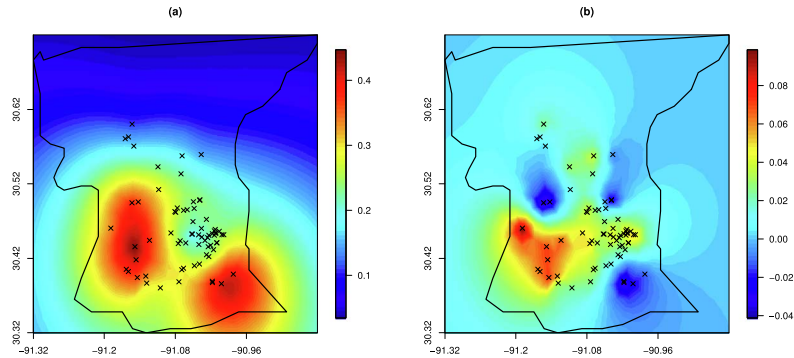|											scenario			
		a_i			pc.a		pc.b			pc.c			pc.d	
	a_ϕ	a_τ	U_ϕ	U_τ	U_ϕ	U_τ	U_ϕ	U_τ	U_ϕ	U_τ				
$\beta_{a,i}$	0.5	0.01	2	1/0.31	2	0.01/0.31	0.5	0.1/0.31	5	0.1/0.31				
$\beta_{b,i}$	0.5	0.01	2	1/0.31	2	0.01/0.31	0.5	0.1/0.31	5	0.1/0.31				
ϵ_i	0.5	0.01	2	4/0.31	2	0.04/0.31	0.5	0.4/0.31	5	0.4/0.31				

FIG 8. *Posterior mean for the varying coefficient of area $\beta_{a,i}$ (panel a) and other area $\beta_{b,i}$ (panel b). Observed locations are marked with a cross.*

6. Discussion

Most of the vast literature available on varying coefficient models considers case-specific prior distributions depending on the type of effect modifier and pays little attention to the risk of overfitting entailed by the increased flexibility that these models offer. In this paper we present varying coefficient models as a single class of models and use a unified approach for setting priors, regardless of the model assumed on the coefficients. The definition of the varying coefficient model as a flexible extension of a simpler model *calls for* eliciting priors that allow the simpler base model to arise. PC priors guarantee this; since the mode is at the base model, overfitting, a common aspect in complex hierarchical models, is avoided by construction. PC priors follow specific principles that remain unchanged no matter the model choice for the varying coefficient $\pi(\boldsymbol{\beta}|\xi)$. This means we can address prior specification for any varying coefficient model in the same way using well defined principles.

We have illustrated the use of PC priors for varying coefficients in two different applications. Whether the covariate is standardized or not obviously makes an impact on the scale of the varying coefficient, thus the user should be careful in defining the value U for the precision parameter τ and change it accordingly if the scale of the covariate is transformed. In our experience the choice of U does not impact much the posterior for $\boldsymbol{\beta}$, unless almost all the probability mass is assigned deliberately to the base model, i.e. unless an unreasonable prior is used, meaning a prior that is against our prior knowledge on the behaviour of the VC. Building a prior on the distance from a base model allows the level of informativeness of the prior to be set according to the actual amount of prior information. In the VCM case, for instance, the PC prior can be set as a weakly informative prior for the precision as we usually have a reasonable guess on the scale of the varying coefficient (depending on the link function of the model, the scale of the data and of the covariate).

With the aim of covering the most popular varying coefficient models we assumed $\boldsymbol{\beta}$ as a Gaussian process with marginal precision τ and focused on sev-

eral structures for the covariance matrix, reflecting different behaviours for the varying coefficient. The class of models presented here does not include the scale mixture of normals, $\beta_t|\tau_t \sim \mathcal{N}(0, \tau_t^{-1}), t = 1, \ldots, n$, where the precision parameter varies over the range of the effect modifier, leading to a more complex and computationally involved VCM. These models are useful in specific situations like sparse regression [10] and adaptive smoothing [37]. In a VCM setting, this kind of models could be useful when the varying coefficient is thought to be a smooth function with non constant degree of smoothness.

To conclude, choice of the prior $\pi(\xi)$ is difficult in practice, because there is typically no prior information on the hyperparameters in hierarchical models. Moreover, the empirical information available to estimate the posterior for ξ is less compared to that available for the parameters in the linear predictor. This means that the prior for ξ is deemed to have a large impact on the model, especially in cases where data are poorly informative. In our opinion, this represents a further good reason for using PC priors in varying coefficient models, as we can be more confident that no overfitting takes place when there is not enough information in the data. Even though we do not know much at prior about suitable values for ξ, we often know exactly what a hyperparameter *does* in terms of shrinkage to a simpler model.

Appendix A: Derivation of the PC prior

A.1. The unstructured case

The varying coefficient model in the exchangeable case is

$$\eta_t = \alpha + \beta_t x_t \qquad t = 1, \ldots, n,$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{R}(\rho)),$$

with

$$\boldsymbol{R}(\rho) = \begin{bmatrix} 1 & \rho & \ldots & & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ . & & . & & . \\ . & & & . & . \\ \rho & \rho & \ldots & \rho & 1 \end{bmatrix}$$

and base model $\rho = 1$ (i.e. $\beta_t = \beta \ \forall t$). To evaluate the distance from the base model we need to use a limiting argument and consider a fixed value of $\rho = \rho_0$ close to 1 under the base model. For zero-mean multivariate normal densities, the KLD simplifies to:

$$\mathrm{KLD}(f_1(\rho)||f_0) = \frac{1}{2}\left(tr(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) - n - \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right)$$

where $\boldsymbol{\Sigma}_0 = \boldsymbol{R}(\rho_0)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{R}(\rho)$, $\rho < \rho_0$, are the variance-covariance matrices under the base and flexible model respectively. In this case, the KLD:

$$\text{KLD}(f_1(\rho)||f_0)$$
$$= \frac{1}{2}\left(\frac{n(1+(n-2)\rho_0 - (n-1)\rho\rho_0)}{(1-\rho_0)((n-1)\rho_0+1)} - n - \log\frac{(1+(n-1)\rho)(1-\rho)^{n-1}}{(1+(n-1)\rho_0)(1-\rho_0)^{n-1}}\right)$$

Considering the limiting value as $\rho_0 \to 1$, the distance

$$d(\rho) = \lim_{\rho_0 \to 1} \sqrt{2\text{KLD}(f_1(\rho)||f_0)} = \lim_{\rho_0 \to 1}\sqrt{\frac{(n-1)(1-\rho)}{1-\rho_0}} = c\sqrt{1-\rho}, \quad 0 \le \rho < 1$$

for a constant c that does not depend on ρ. Since $0 \le d(\rho) \le c$, assigning a truncated exponential with rate λ on $d(\rho)$ we have

$$\pi(d(\rho)) = \frac{\lambda \exp(-\lambda c\sqrt{1-\rho})}{1 - \exp(-\lambda c)}, \qquad 0 \le d(\rho) \le c, \quad \lambda > 0.$$

Reparametrizing $\theta = \lambda c$ leads to the PC prior for ρ:

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{2\sqrt{1-\rho}(1 - \exp(-\theta))}, \quad 0 \le \rho < 1, \quad \theta > 0.$$

A.2. The autoregressive model of first order

The varying coefficient model in the AR1 case is

$$\eta_t = \alpha + \beta_t x_t \qquad t = 1,...,n,$$
$$\boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{R}(\rho)),$$

with $\boldsymbol{R}(\rho)_{ij} = (\rho^{|i-j|})$ and base model $\rho = 1$. Using a limiting argument similar to that of Appendix A.1, the distance to the base model is

$$d(\rho) = c\sqrt{1-\rho}, \qquad |\rho| < 1 \tag{24}$$

where c is a constant. Note that (24) is upper bounded, $0 \le d(\rho) \le c\sqrt{2}$, so that the PC prior for $d(\rho)$ is

$$\pi(d(\rho)) = \frac{\lambda \exp(-\lambda c\sqrt{1-\rho})}{1 - \exp(-\lambda c\sqrt{2})}, \qquad 0 \le d(\rho) \le c\sqrt{2}, \quad \lambda > 0.$$

Reparametrizing $\lambda = \theta/c$ and using the change of variable formula it follows that the PC prior on the ρ scale is [40]

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{2\sqrt{1-\rho}(1 - \exp(-\sqrt{2}\theta))}, \qquad |\rho| < 1, \quad \theta > 0.$$

A.3. Random walk model of order one and two

The varying coefficient has a joint distribution given by

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^{-1} \boldsymbol{K}^{-1})$$

with $\boldsymbol{K}$ symmetric semi-positive definite matrix. Let $f_0 = \pi(\boldsymbol{\beta}|\tau_0 = \infty)$ and $f_1 = \pi(\boldsymbol{\beta}|\tau)$ denote the base and flexible models, with precisions τ_0 and τ, respectively. [38] show that $\text{KLD}(f_1||f_0)$ goes to $\frac{\tau_0 n}{2\tau}$, for τ much lower than τ_0 and $\tau_0 \to \infty$, so that $d(\tau) = \sqrt{2\text{KLD}(f_1||f_0)} = \sqrt{\tau_0 n/\tau}$ and $d(\tau) \sim \exp(\lambda)$, $\lambda > 0$.

By a change of variable and setting the rate $\lambda = \theta/\sqrt{n\tau_0}$, [38] derive the PC prior for τ as

$$\pi(\tau) = \frac{\theta}{2} \tau^{-3/2} \exp\left(-\theta/\sqrt{\tau}\right), \quad \tau > 0, \theta > 0, \tag{25}$$

which is a $\text{Gumbel}(1/2, \theta)$ type 2 distribution.

A.4. Continuous spatial variation

The spatially varying coefficient can be seen as a realization of a Gaussian random field (GRF)

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \tau^{-1} \boldsymbol{R}(\phi)\right)$$

with *Matérn* correlation function as in (14). PC priors for the range and marginal variance parameters of a GRF with *Matérn* covariance function have been derived by [20]. Here we only summarize the main results on the computation of the PC prior, while for further details the reader is referred to [20]. Deriving PC priors for these parameters is more complex that in the previous situations considered in this paper due to the infinite-dimensional nature of GRFs. Following [20] and setting $d = 2$, parameters ϕ and τ are conveniently reparametrized as:

$$\kappa = \frac{\sqrt{8\nu}}{\phi} \qquad \psi = \sqrt{\tau^{-1}} \phi^\nu \sqrt{\frac{\Gamma(\nu + 1)4\pi}{\Gamma(\nu)}}$$

Since the parameter ψ depends on κ, the joint PC prior is built as $\pi(\psi, \kappa) = \pi(\kappa)\pi(\psi|\kappa)$, which can then be transformed into a joint PC prior for (ϕ, τ). In this case, the base model corresponds to $\phi = \infty$ (or equivalently, $\kappa = 0$), i.e. the spatial correlation is so strong that we have a constant field and $\tau = \infty$ ($\psi = 0$), i.e. no marginal variance. The PC prior $\pi(\psi|\kappa)$ is built based on the observations available at n locations, while the PC prior $\pi(\kappa)$ is based on the infinite-dimensional GRF to avoid a model-dependent prior; see [20] for details.

The PC prior for κ:

$$\pi(\kappa) = \lambda_1 \exp\left(-\lambda_1 \kappa\right), \quad \kappa > 0, \tag{26}$$

and $\lambda_1 > 0$. The user can set U_1 and a_1 such that $\mathbb{P}(\phi < U_1) = a_1$, so that $\lambda_1 = - \left(\frac{U_1}{\sqrt{8\nu}} \right) \log(a_1)$.

The PC prior for $\psi | \kappa$ follows an exponential distribution:

$$\pi(\psi | \kappa) = \lambda_2 \exp(-\lambda_2 \psi), \quad \psi > 0 \tag{27}$$

where, as before, $\lambda_2 > 0$ can be selected based on the user-selected values U_2 and a_2 such that $\mathbb{P}(1/\sqrt{\tau} > U_2 | \kappa) = a_2$, which leads to $\lambda_2(\kappa) = -\kappa^{-\nu} \sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu+1)4\pi}} \frac{\log(a_2)}{U_2}$.

The joint PC prior $\pi(\kappa, \psi) = \pi(\kappa)\pi(\psi | \kappa)$, and by a change of variable (setting $\lambda_\phi = \sqrt{8\nu}\lambda_1$ and $\lambda_\tau = \kappa^\nu \sqrt{\frac{\Gamma(\nu+1)4\pi}{\Gamma(\nu)}}\lambda_2$) it follows that the PC prior for τ, ϕ:

$$\pi(\tau, \phi) = \pi(\phi)\pi(\tau | \phi) = \lambda_\phi \phi^{-2} \exp\left(-\lambda_\phi \phi^{-1}\right) \frac{\lambda_\tau}{2} \tau^{-3/2} \exp\left(-\frac{\lambda_\tau}{\sqrt{\tau}}\right),$$
$$\tau > 0, \phi > 0 \tag{28}$$

where, once the user fixes U_ϕ, a_ϕ, U_τ, a_τ such that $\mathbb{P}(\phi < U_\phi) = a_\phi$, $\mathbb{P}(1/\sqrt{\tau} > U_\tau) = a_\tau$ the parameters λ_ϕ, λ_τ are calculated as

$$\lambda_\phi = - \log(a_\phi)U_\phi, \qquad \lambda_\tau = - \frac{\log(a_\tau)}{U_\tau}.$$

Acknowledgements

Supplementary Material

Supplementary material A for the paper: "A unified view on Bayesian varying coefficient models"
(doi: 10.1214/19-EJS1653SUPPA; .zip).

Supplementary material B for the paper: "A unified view on Bayesian varying coefficient models"
(doi: 10.1214/19-EJS1653SUPPB; .pdf).

References

[1] 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. MR3362184

[2] Barndorff-Nielsen, O., and Schou, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3:408–419. MR0343510

[3] Berger, J. O., and Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory*, 10:461–482. MR1309107

[4] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B*, 36(2):192–225. MR0373208

[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. MR1105822

[6] Biller, C. and Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, 1(3):195–211.

[7] Bitto, A. and Frühwirth-Schnatter, S. (2018). Achieving Shrinkage in a Time-Varying Parameter Model Framework. arXiv:1611.01310. MR3944764

[8] Blangiardo, M. and Cameletti, M. (2017). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley. MR3364017

[9] Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30:93–11. MR1963895

[10] Carvalho C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480. MR2650751

[11] Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14:715–745. MR2087971

[12] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27:1491–1518. MR1742497

[13] Ferguson, C., Bowman, A., Scott, E., and Carvalho, L. (2007). Model comparison for a complex ecological system. *Journal of the Royal Statistical Society Series A*, 170(3):691–711. MR2380593

[14] Finazzi, F., Scott, M., and Fasso, A. (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society Series C*, 62(2):287–308. MR3045878

[15] Finley, A. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2:143–154.

[16] Franco-Villoria, M., Ventrucci, M., and Rue, H. (2019). Supplement A to "A unified view on Bayesian varying coefficient models". DOI: 10.1214/08-AOS99GSUPPA.

[17] Franco-Villoria, M., Ventrucci, M., and Rue, H. (2019). Supplement B to "A unified view on Bayesian varying coefficient models". DOI: 10.1214/08-AOS99GSUPPB.

[18] Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models.

Journal of Econometrics, 154(1):85–100. MR2558953

[19] Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In *J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.)*, pages 165–200. Bayesian Statistics 9, Oxford. MR3204006

[20] Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*. MR3941267

[21] Gamerman, D., Moreira, A. R., and Rue, H. (2003). Space-varying regression models: specifications and simulation. *Computational Statistics & Data Analysis*, 42(3):513–533. MR2005406

[22] Gelfand, A., Kim, J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396. MR1995715

[23] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 3:515–534. MR2221284

[24] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55(4):757–796. MR1229881

[25] Hoover, D., Rice, J., and Wu, C. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822. MR1666699

[26] Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502. MR0254999

[27] Kowal, D. R., Matteson, D. S., and Ruppert, D. (2018). Dynamic Shrinkage Processes. arXiv:1707.00763. MR3997101

[28] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86. MR0039968

[29] Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

[30] Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics*, 35(4):691–700. MR2468870

[31] Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67(0):68–83. MR3079584

[32] Marx, B. (2010). P-spline varying coefficient models for complex data. In *T. Kneib and G. Tutz (Eds.)*. Statistical Modelling and Regression Structures, Physica-Verlag HD. MR2664626

[33] Mu, J., Wang, G., and Wang, L. (2018). Estimation and inference in spatially varying coefficient models. *Environmetrics*, 29. MR3749556

[34] Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135:370–384. MR0375592

[35] Riebler, A., Sørbye, 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. MR3541089

[36] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC. MR2130347

[37] Scheipl, F. and Kneib, T. (2009). Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior. *Computational Statistics & Data Analysis*, 53:3533–3552. MR2751944

[38] Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28. MR3634300

[39] Sørbye, S. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51. MR3326820

[40] Sørbye, S. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38:923–935. MR3714116

[41] Staubach, C., Schmid, V., Knorr-Held, L., and Ziller, M. (2002). A Bayesian model for spatial wildlife disease prevalence data. *Preventive Veterinary Medicine*, 56:75–87.

[42] Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York. MR1697409

[43] Tian, L., Zucker, D., and Wei, L. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469):172–183. MR2156827

[44] Waller, L., Zhu, L., Gotway, C., Gorman, D., and Gruenewald, P. (2007). Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21:573–588. MR2380676

[45] Warnes, J. and Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74(3):640–642. MR0909370

[46] Yue, Y. R., Simpson, D., Lindgren, F. and Rue, H. (2017). Bayesian adaptive smoothing splines using stochastic differential equations. *Bayesian Analysis*, 2:397–424. MR3217001

[47] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261. MR2054303