Prior specification in one-factor mixed models applied to community ecology data

Ventrucci M[1,*], Burgazzi G[2], Cocchi D[1] and Laini A [2]

[1] *University of Bologna, Department of Statistical Sciences, massimo.ventrucci@unibo.it* [*]*Corresponding author*
[2] *Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma*

Abstract. *In community ecology studies the goal is to evaluate the effect of environmental covariates on a response variable while investigating the nature unobserved heterogeneity. We focus on one-factor mixed models in a Bayesian setting and introduce an intuitive Penalized Complexity (PC) prior to balance the variance components of the model. We start with the simple one-way anova and discuss extension to spatially structured residuals, following a Matern exponential covariance.*

Keywords. *Bayesian mixed models; Group model; Intra-class correlation; One-way anova; PC prior.*

1 Mixed models in community ecology

When modelling ecological data several authors report high levels of unexplained variation after considering the effect of environmental covariates [1]. In this cases, the linear regression framework is abandoned in favour of linear mixed models. From a statistician's point of view, accounting for lack of independence in the residuals is required to "adjust" estimates of the regression coefficients. From an ecologist's perspective, investigating the type of residual structure is important in itself to improve understanding of, or generating hypothesis on, the underlying ecological community. For instance, residuals that are correlated within some pre-specified groups/clusters of observational units can be associated to interactions between members of the community, including negative (like competition, predation and parasitism) and positive interactions (like mutualism and commensalism).

We analyze macroinvertebrate community data collected in 6 sampling campaigns carried out in three streams tributaries of the Po River (Northern Italy): Nure Stream, Parma Stream and Enza Stream. For each river a sampling area was sampled twice (in summer and winter), the spatial design including fifty random points aligned along several transects (in total, there are 38 transects). At each point, abundance of macroinvertebrates (response) and environmental covariates such as flow velocity, water depth, substrate composition and benthic organic matter were recorded. The main goals are 1) to investigate the role of the environmental covariates and 2) to assess the presence of small scale interactions within macroinvertebrate communities.

The questions above could be addressed by applying the mixed model framework. In mixed models the effect of the observed covariates and unobserved processes can be neatly separated. Assuming a

Gaussian response $\boldsymbol{Y}$ and covariates $\boldsymbol{X}$ the general formulation of a mixed model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \quad ; \quad \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_b) \quad ; \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$$

where $\boldsymbol{\beta}$ are the fixed effects and $\boldsymbol{b}$ the random effects. The common interpretation in ecology is that the $\boldsymbol{\beta}$'s account for variability explained by *observed* abiotic factors, while the $\boldsymbol{b}$'s account for variability driven by *unobserved* abiotic or biotic factors [4]. Matriz $\boldsymbol{Z}$ incorporates information about the grouping factors under consideration. In our case study, it is expected that observations tend to be similar within the same sampling campaign or the same transect, thus grouping factors to be considered in the following analysis will be *campaign* (a factors with 6 levels) and *transect* (a factor with 38 levels).

1.1 Exchangeable case: one-way anova

Assume data are grouped according to the levels of a certain *grouping factor*, with y_{ij} being the response at unit $i = 1, \dots, m_j$ within group $j = 1, \dots, n$. The simplest mixed model case is one-way anova,

$$\begin{aligned} y_{ij} &= \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + b_j + \varepsilon_{ij} \qquad i = 1, \dots, m_j \quad j = 1, \dots, n \\ b_j &\sim \mathcal{N}(0, \sigma_b^2) \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma_\varepsilon^2) \end{aligned} \tag{1}$$

where b_j's are random effects quantifying *group-specific* deviations from the intercept α and ε_{ij} are i.i.d. noise terms. It is important to note that introducing the group-specific random effects induces correlation among the residuals $(y_{ij} - (\alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta}))$. For this reason, we refer to model (1) as to the *exchangeable* case.

We note that the b_j's and ε_{ij}'s compete to capture the variance unexplained by environmental covariates. The balance between the two components is regulated by the hyper-parameters σ_b^2 and σ_ε^2. In particular, when $\sigma_b^2 = 0$ model (1) corresponds to the linear regression $y_{ij} = \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_{ij}$; the ecological conclusion would be that only environmental covariates matter and the rest is i.i.d. variation. Instead, if $\sigma_b^2 > 0$ there is a certain amount of unexplained variability in the data; the interpretation would be that covariates matters but residuals are not independent, investigating the structure in there can give useful insights on the behaviour of the ecological community.

2 Prior specification in one-factor mixed models

Because the estimates of the variance components σ_b^2, σ_ε^2 drive most of the ecological interpretations on the behaviour of underlying communities, the choice of priors for the hyper-parameters σ_b^2, σ_ε^2 is an important aspect of model specification. [3] address this issue in a general class of one-factor Bayesian mixed models: their proposal is to tackle the choice of priors for the variance components jointly, by specifying a prior on the intraclass correlation (ICC) parameter, $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_\varepsilon^2)$. This prior is derived under the Penalized Complexity (PC) prior framework [2]. By definition, a PC prior is an exponential distribution with rate parameter λ defined on a *distance* scale, d. Such distance d quantifies the increased complexity of the model under consideration w.r.t. to its *base model*, in our case the base model being the linear regression $y_{ij} = \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_{ij}$. Thus λ is a scaling parameter controlling the degree of shrinkage to the base model and needs to be specified by the expert user/ecologist. Once the PC prior has been scaled according to a given λ, the prior on the original parameter, ρ, can be computed by the change of

variable rule. For a detailed discussion of the principles underpinning the construction of PC priors and their properties see [2].

In [3] group models assuming different correlation structures for the within group residuals are presented and the prior for the associated correlation parameter (e.g., ρ in the exchangeable case) is always derived under the same principles. There are several practical advantages for the user/ecologist. First, the PC prior ensures proper shrinkage to the base model, thus avoiding overfitting. Second, in the exchangeable case, the scaling parameter λ can be elicited upon a prior statement on the ICC, i.e. the proportion of total variance explained by the grouping factor; for instance, one may compute λ such that $\mathbb{P}(\rho < 0.5) = 0.5$. Third, the prior for ρ is actually defined on an underlying distance scale, which is common to all group models (e.g. exchangeable residuals within transects, serially correlated residuals within transect, they both are extension of the same base model). Thus, the intuitive choice of λ based on eliciting the ICC, is one that can be applied in general for any group model.

2.1 Spatially correlated case

In the present work we extend to the case of spatially correlated residuals, according to a Matern exponential covariance. The *spatially correlated* group model is

$$\begin{aligned} y_{ij} &= \alpha + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \theta_{ij} \qquad i = 1,\dots,m_j \quad j = 1,\dots,n \\ (\theta_{ij},\dots,\theta_{mj})^{\mathsf{T}} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{R}_j(\phi)) \end{aligned}$$

where the correlation matrix depends on a range parameter $\phi > 0$,

$$\boldsymbol{R}_j(\phi) = \begin{bmatrix} 1 & \exp(-u_{1,2}/\phi) & \cdots & \cdots & \exp(-u_{1,m}/\phi) \\ \exp(-u_{2,1}/\phi) & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & \exp(-u_{m-1,m}/\phi) \\ \exp(-u_{m,1}/\phi) & \cdots & \cdots & \exp(-u_{m,m-1}/\phi) & 1 \end{bmatrix}. \tag{2}$$

Notation $u_{i,h}$ in matrix (2) indicates the euclidean distance between spatial units i and h. We note that the base model is achieved at $\phi = 0$, in which case we are back to i.i.d. residual case. The PC prior for ϕ can be derived numerically.

3 Concluding remarks

We emphasize that a very intuitive aspect of the proposed PC prior on ϕ is that the scaling parameter λ can be chosen according to a prior statement on the ICC, like in the exchangeable case. We believe this intuitive way to define λ provides an easy-to-elicit prior. The user is then able to balance variance components in an intuitive manner, even in complex models where the variance parameters are difficult to interpret.

The poster presentation will focus in particular n the benefits of using PC priors for residual correlation parameters in a model comparison setting. We will discuss comparison of two different one-factor

mixed models, having different residual structures: *exchangeable* residuals within campaign versus *spatially correlated* residuals within campaign. This comparison would provide insights into the strength of spatial correlation in the residuals, as a preliminary answer to the main questions under study in our motivating example, the one about the presence of possible interactions between members of the ecological community.

Acknowledgments. Daniela Cocchi and Massimo Ventrucci are supported by the PRIN 2015 grant project n. 20154X8K23 (EPHASTAT); Gemma Burgazzi is supported by the PRIN 2015 grant project n. 201572HW8F (NOACQUA). Both projects are funded by the Italian Ministry of Education and University.

References

[1] Lamouroux N., Dolédec S and Gayraud S (2004). Biological traits of stream macroinvertebrate communities: effects of microhabitat, reach, and basin filters. *Journal of the North American Benthological Society* **23(3)**: 449–466.

[2] Simpson D, Rue H, Riebler A, Martins T.G. and Sorbye S.H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science* **32**: 1–28.

[3] Ventrucci M, Burgazzi G, Cocchi D and Laini A (2019). PC priors for residual correlation parameters in one-factor mixed models. *arXiv:1902.08828*.

[4] Warton et al., (2015) So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution* **30(12)**: 766-779.