Prior distributions for latent Gaussian models

May 20, 2021

1 Introduction

Latent Gaussian models (Rue et al., 2009) constitutes a general class of Bayesian hierarchical regression models widely used in several applied fields. They are defined assuming an additive combination of effects for the linear predictor, and a Gaussian prior is specified as prior distribution for each parameter.

Using suitable design matrices and structured covariance matrices for the Gaussian priors, it is possible to let the model account for different relationships between the response and the covariates. For instance, flexible non-linear relationships (Lang and Brezger, 2004), spatially correlated errors (Besag et al., 1991) and temporal effects can be included as effects in latent Gaussian models. It is common to specify the covariance matrices of the Gaussian priors in order to define a Gaussian Markov Random Field (GMRF), characterized by a sparse precision matrix that leads to relevant computational advantages (Rue and Held, 2005). Moreover, Intrinsic-GMRF priors, in which the covariance matrix is not full-rank, are often chosen.

The structured covariance matrix characterizing such priors for the model parameters is multiplied by a scale factor, a scalar hyper-parameter that requires a further prior distribution, if a fully Bayesian framework is chosen. The issue of specifying the prior distribution for scale parameters has an important tradition in the statistical literature and is crucial when structured matrices are specified since they can impact the amount of prior variance assigned to each model component. Among the others see Sørbye and Rue (2014), Klein et al. (2016), Simpson et al. (2017) and Fuglstad et al. (2020).

2 Background and statement of the problem

A latent Gaussian model is based on the assumption that the response variable $\mathbf{y} \in \mathbb{R}^n$ follows an exponential family distribution with the i-th $(i = 1, ..., n)$ conditional expectation μ_i that is related to the linear predictor η_i through a link function $g(\mu_i) = \eta_i$. The most general structure of the predictor is:

$$\eta_i = \beta_0 + \sum_{l=1}^{p} \beta_l c_{li} + \sum_{j=1}^{q} f_j(\mathbf{z}_{ji}),$$

where β_0 is the intercept, β_j are the fixed effect coefficients related to the covariate values $\mathbf{c}_l$, $f_j(\cdot)$ are smooth flexible functions that allow to model the dependence between the response and a set of covariates. Exploiting the mixed model representation (Klein et al., 2016), it is possible to write the flexible functions as $\mathbf{z}_{ji}^T\boldsymbol{\gamma}_j$, where $\boldsymbol{\gamma}_j$ is a vector of coefficients for which the following Gaussian prior is assumed:

$$\boldsymbol{\gamma}_j \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\gamma_j}\mathbf{K}^-_{\gamma_j}\right).$$

The main focus of the work is the elicitation of the scale parameter's prior $\pi\left(\sigma^2_{\gamma_j}\right)$ introducing available information information on the phenomenon and taking into account the fixed structure of the precision matrix $\mathbf{K}^-_{\gamma_j}$. In past works, Sørbye and Rue (2014) proposed to scale the covariance matrix using a generalized variance, then the prior on $\sigma^2_{\gamma_j}$ is specified treating the random effect as an unstructured one. It represents an interesting operational solution, but the developed framework is still conditioned with respect to the scale parameter. A solution based on the marginalization of the prior distribution of the random effects prior with respect to $\sigma^2_{\gamma_j}$ was presented in Klein et al. (2016).

3 Research question or hypothesis, aim, objectives and deliveries

The research work relies on the idea of specifying the prior distribution on the random vector $\boldsymbol{\nu}_j = \mathbf{Z}_j\boldsymbol{\gamma}_j$ considering its sample variance defined as follows:

$$V_{\boldsymbol{\nu}_j} = \frac{\boldsymbol{\nu}_j^T\mathbf{M}\boldsymbol{\nu}_j}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\nu_{ij} - \bar{\nu}_j\right)^2,$$

where $\mathbf{M}$ is the centering matrix. Conditionally on the scale parameter, $V_{\boldsymbol{\nu}_j}$ is a random variable distributed as a quadratic form in Gaussian variable:

$$V_{\boldsymbol{\nu}_j}|\sigma^2_{\gamma_j} \sim \frac{\sigma^2_{\gamma_j}}{n-1}\sum_{k=1}^{n}\lambda_{kj}X_k; \quad X_k \sim \chi^2_1,\ k = 1,\dots,n,$$

where $\{\lambda_{1j}, \lambda_{2j}, \dots, \lambda_{nj}\}$ are the eigenvalues of $\mathbf{K}^-_{\nu_j}$. Eventually, the elicitation of the prior on $V_{\boldsymbol{\nu}_j}$ is carried out on its marginal distribution, after that an integral equation is solved.

Starting from this general idea, the following developments are expected.

- Theoretical progresses: the available results need to be formalized. The final goal is proposing a self-contained framework in which the user must specify the prior variability on the linear predictor only, then an equal subdivision among the different components is automatically performed. In particular, the impact that the developed prior elicitation framework has on the fixed covariates and the effects of the presence of improper priors (i.e. IGMRF) must be further investigated.

- Software: MCMC algorithms are required in order to retrieve samples from the posterior distribution of the model parameters. In particular, Metropolis-Hastings algorithms must be implemented to estimate latent Gaussian models under the proposed prior for scales parameters. To encourage the use of the provided prior elicitation tool, a R package containing the functions required for the model estimation should be developed.

- Applications: two different applied problems are expected to be tackled under the proposed methodology.

  - Analysis of INVALSI data: INVALSI data provide rich information concerning school performances over the last years, and they can be extremely helpful in evaluating the effect of COVID-19 pandemic on student performances. In this context, multilevel models have been widely used for estimating the impact of individual and group-level socio-economic variables. Latent Gaussian models can be a useful tool to this aim because of their ability to capture complex non-linear relationship and to model interactions between variables.
  - Disease mapping: the BYM model (Besag et al., 1991) is a popular tool used in environmental epidemiology for which the prior elicitation problem has a long tradition (see, e.g, Bernardinelli et al. (1995) and Wakefield (2006)). The developed proposal is expected to be adapted to this applied framework.

The research activity will follow the plan reported below:

- Survey and synthesis of the scientific literature (month 1);

- Methodological developments concerning the proposed prior specification setting: priors on fixed effects (months 2-6);

- Software implementation, development of an R package (months 4-12);

- Application of the developed methodologies: the BYM model in disease mapping and INVALSI data (months 6-12).

The fellow is expected to deliver at least two working papers and to disseminate the advancements of his research in two international conferences.

4 Participants in the study and the role they play

The research will involve some members of the Department of Statistical Sciences:

- Fedele Greco will be involved as a specialist in the field of prior elicitation in Bayesian random effect models.

- Daniela Cocchi will be involved as expert of Bayesian inference.

- Mariagiulia Matteucci will be involved as expert in the analysis of education data in the application of the developed methodologies to the INVALSI data.

References

L. Bernardinelli, D. Clayton, and C. Montomoli. Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, 14(21-22):2411–2431, 1995.

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

G.-A. Fuglstad, I. G. Hem, A. Knight, H. Rue, A. Riebler, et al. Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15(4):1109–1137, 2020.

N. Klein, T. Kneib, et al. Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, 11(4):1071–1106, 2016.

S. Lang and A. Brezger. Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212, 2004.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications.* CRC press, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 71(2):319–392, 2009.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pages 1–28, 2017.

S. H. Sørbye and H. Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2): 158–183, 2006.