

Bayesian structural equation modeling for factors influencing residential radon levels

Ippoliti Luigi, Fontanella Lara, Valentini Pasquale, Sarra Annalina, Palermi Sergio

Abstract Motivated by the problem of modelling high-dimensional multivariate referenced data, arising in many areas of research, this article proposes a Generalised Latent Spatial Quantile Regression (GLSQR) model as a reliable solution in studying the effects of some covariates across the quantiles of the response distribution. In addition, the discussed model warrants consideration when the matrix of the explanatory variables is defined through a set of spatial common latent factors. The latent factors and quantile regression components are estimated through a hierarchical Bayesian procedure and MCMC algorithms are used to provide full probabilistic inference. We illustrate the use of our GLSQR model with application to radon data, as motivating example coming from the environmental protection research.

Abstract *In questo articolo si introduce un modello (GLSQR) che combina la regressione spaziale quantilica con un modello ad equazioni strutturali generalizzato. Il modello GLSQR è adatto per studiare gli effetti di alcune covariate sui quantili condizionati della distribuzione della variabile risposta. La modellizzazione in esame consente, inoltre, di poter specificare la matrice delle variabili esplicative attraverso fattori latenti spaziali comuni. I fattori latenti e la componente di regressione quantilica del modello sono stimati con una procedura bayesiana gerarchica che prevede il ricorso ad algoritmi di tipo MCMC. La metodologia proposta è illustrata attraverso un'applicazione a dati relativi alle misure di concentrazioni di radon indoor.*

Key words: Bayesian inference, factor analysis, quantile regression, indoor radon

Ippoliti L., Fontanella L., Valentini P., Sarra A.

Department of Economics, University "G.d'Annunzio", V.le Pindaro 42, 65127 Pescara,
e-mail: ippoliti@unich.it, e-mail: lfontan@unich.it, e-mail: pvalent@unich.it, e-mail:
asarra@dmqte.unich.it

Palermi S.

Agency of Environmental Protection of Abruzzo (ARTA), V.le Marconi, 178; 65127 Pescara
(Italy), e-mail: s.palermi@artaabruzzo.it

1 Introduction

It is common in many areas of research to have high-dimensional multivariate spatially referenced data. Hereby, in spatial settings, there is the need for flexible explanatory stochastic data models in order to improve estimation precision and provide simpler description of the complex relationships among variables. A crucial problem consists in describing and analysing the influence of some covariates on a real valued response variable. In most practical cases, the conventional regression analysis, based on conditional means, may ignore some essential features of the dependence of the outcome on a set of predictors. For instance, we could be interested in investigating the changing effect of the covariates across quantiles. In this respect, a quantile regression approach is worthwhile since it allows the covariates to affect the entire conditional distribution, rather than just the mean. Albeit there is a vast literature discussing quantile regression approach [see for example, Section 2 of Lum and Gelfand, 2012, for a recent review], few researches [see Reich et al., 2012, Lum and Gelfand, 2012] have to date defined a conditional quantile regression which incorporates spatial dependence. A semiparametric modelling approach where a type of spatial autoregressive model is used to enrich the structure of the quantile regression function, has been formulated by Borgoni [2011]. Following the lines of Lum and Gelfand [2012], in this paper we build a fully model-based approach to examine quantile effects together with spatial dependence. Our contribution is the introduction of a Generalised Latent Spatial Quantile Regression (GLSQR) model which integrates the spatial dependence through the definition of the spatial asymmetric Laplace process. This model differs from commonly used quantile regression models in several aspects. Both from a computational and interpretative perspective there can be advantages to working with the specification of the GLSQR model. Rather than focusing on a single variable, our approach allows the data to have a multivariate and multidimensional structure. A remarkable flexibility of our model specification regards the possibility to specify the required design matrix through a set of spatial latent factors. Furthermore, the GLSQR model formulation confers the distinct advantage to take into account the spatial nature of the variables and facilitate the identification of spatial clusters in the latent factors. In such a way, the GLSQR model provides an effective way to address the curse of dimensionality, usually present in large spatial data. In addition, of particular interest in our model specification, is the possibility for the latent factors to be not restricted to normally distributed variables, since more types of observed data from the exponential family can be handled. The GLSQR model is implemented in a full bayesian framework in which the latent variables and the spatial quantile regression components are treated jointly, overcoming the issue of underestimating the uncertainty, arising when a two-step estimation strategy is adopted. Bayesian computation for this model is performed by using MCMC methods. The rest of paper is organized as follows. Section 2 is devoted to illustrate the general approach providing a background on spatial quantile regression and latent variables analysis. In Section 3 we describe inference for the parameters in a fully bayesian context. Finally, Section 4 gives an example using indoor radon data.

2 The GLSQR model

In this section we describe the GLSQR model. Our approach considers a conditional quantile regression which integrates spatial dependence through the definition of the spatial asymmetric Laplace process (ALP). This strategy is well discussed in Lum and Gelfand [2012]. Our model-based approach for spatial quantile regression has the additional key feature to deem some of the relevant explanatory variables (predictors in the quantile regression) as indicators of spatial latent factors. The GLSQR model performs estimation of quantile regression and latent spatial factors components hierarchically. Let $Y(\mathbf{s}_i)$ be a spatial process representing some response of interest and $\mathbf{X}(\mathbf{s}_i)$ a $(k \times 1)$ related vector of explanatory random variables. We assume that the variables are measured at n sites, $\mathbf{s}_i \in S, i = 1, \dots, n$, with S a spatial domain in two dimensional Euclidean space $\mathcal{R}^2$. For a generic site $\mathbf{s}$, let $q_p[\mathbf{X}(\mathbf{s})]$ denote the p th ($0 < p < 1$) quantile regression function of $Y(\mathbf{s})$ given $\mathbf{X}(\mathbf{s})$. A general form for a spatial quantile regression is given by $Y(\mathbf{s}) = q_p[\mathbf{X}(\mathbf{s})] + \varepsilon_p(\mathbf{s})$ where $\varepsilon_p(\mathbf{s})$ must satisfy the constraint that $Pr(\varepsilon_p(\mathbf{s}) \leq 0) = p$. The distribution of $\varepsilon_p(\mathbf{s})$ is thus restricted to have the p th quantile equal to zero.

Apart from including spatially referenced covariates in $q_p[\mathbf{X}(\mathbf{s})]$, spatial dependence can be incorporated in the $\varepsilon_p(\mathbf{s})$ which, as shown by Lum and Gelfand [2012], is assumed to be an asymmetric Laplace process, i.e. $\varepsilon_p(\mathbf{s}) \sim AL(p, 0, \tau)$. By using its mixture representation [Kuzobowski and Podgorski, 2000], we can thus write

$$\varepsilon_p(\mathbf{s}) = \sqrt{\frac{2\xi(\mathbf{s})}{\tau p(1-p)}} Z(\mathbf{s}) + \frac{1-2p}{p(1-p)} \xi(\mathbf{s}) \quad (1)$$

where $Z(\mathbf{s})$ is a zero mean Gaussian process with valid correlation function, $\rho_Z(\mathbf{s}, \mathbf{s}' | \theta)$, and $\xi(\mathbf{s})$ a marginally exponentially distributed process with rate τ [Lum and Gelfand, 2012]. Different modelling options are possible for $\xi(\mathbf{s})$. However, as discussed in Lum and Gelfand [2012], the assumption that the $\xi(\mathbf{s})$'s are independent and identically distributed exponential variables has been found adequate in practice and offers computational simplicity and good behavior in the fitting procedure. In this case, they show that

$$Cor[\varepsilon_p(\mathbf{s}), \varepsilon_p(\mathbf{s}')] = \rho_Z(\mathbf{s}, \mathbf{s}') \frac{\frac{\pi}{2} p(1-p)}{1-2p+2p^2}$$

and that, therefore, the covariance of the quantile process model is a scaled version of that of the Gaussian process.

The first level of the model is completed by assuming that at each site $\mathbf{s}$, the relationship between $q_p[\mathbf{X}(\mathbf{s})]$ and $\mathbf{X}(\mathbf{s})$ can be modelled as $q_p[\mathbf{X}(\mathbf{s})] = \mathbf{X}(\mathbf{s})' \boldsymbol{\beta}_p$, where $\boldsymbol{\beta}_p$ is a vector of unknown parameters of interest. Since conditionally on $\xi(\mathbf{s})$, the specification of $\varepsilon_p(\mathbf{s})$ in equation (1) gives a normal linear regression model, it is not difficult to derive the full conditional density of $\boldsymbol{\beta}_p$. At the second level of hierarchy

we consider a generalized spatial structural equation analysis. Following the discussion in Wang and Wall [2003], we assume there are two hidden common factors underlying the observed variables. Let $\mathbf{X}_g(\mathbf{s}) = (X_{g_1}(\mathbf{s}), \dots, X_{g_n}(\mathbf{s}))'$ be a multivariate spatial process observed at a site $\mathbf{s}$, referring to a set of ordinal variables. We express the generalized common spatial factor model as a linear model for a continuously distributed unobserved latent variable $X_{g_j}^*(\mathbf{s})$, as

$$X_{g_j}^*(\mathbf{s}) = v_j + \psi_j f_g(\mathbf{s}) + e_{g_j}(\mathbf{s}), \quad j = 1, \dots, n \quad (2)$$

where v_j is the intercept, ψ_j is the factor loading, $f_g(\mathbf{s})$ is the common spatial factor process underlying the n observed variables and $e_{g_j}(\mathbf{s})$ is a zero mean Gaussian error with variance σ_e^2 . Additionally, suppose to have another multivariate spatial process observed at a site $\mathbf{u}$, denoted as $\mathbf{X}_r(\mathbf{u}) = (X_{r_1}(\mathbf{u}), X_{r_2}(\mathbf{u}), \dots, X_{r_m}(\mathbf{u}))'$. It is assumed that, conditionally on their mean, $\mu_{x_{r_j}}(\mathbf{u})$, and their variance, $\sigma_{x_{r_j}}^2$, the variables $X_{r_j}(\mathbf{u})$, $j = 1, 2, \dots, m$, are independent in the one-parameter natural exponential family with distribution F and a link function, $g_{x_r}(\cdot)$. The generalized common spatial factor model for this data is thus defined through the linear predictor

$$g_{x_r}(\mu_{x_{r_j}}(\mathbf{u})) = m_j + \lambda_j f_r(\mathbf{u}), \quad j = 1, 2, \dots, m \quad (3)$$

where m_j is the intercept, λ_j is the factor loading and $f_r(\mathbf{u})$ is the common spatial factor process underlying the observed variables. For these variables we use a Gamma distribution and a logarithmic link instead of the canonical link function. As displayed in the Section 4, we relate the two common factors through a linear regression function.

3 Bayesian inference

Full probabilistic inference for the model parameters is carried out by eliciting the following priors distributions. In the quantile model stage, reasonable assumptions for the priors are: $\xi(\mathbf{s}) \stackrel{iid}{\sim} \text{Exp}(\tau)$, $\boldsymbol{\beta}_p \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ with $\boldsymbol{\mu}_\beta = \mathbf{0}$, $\boldsymbol{\Sigma}_\beta = \sigma_\beta^2 \mathbf{I}$ and σ_β^2 large. The prior choice for τ depends upon the necessity to ensure a constant process variance across quantiles. Recall that $\text{Var}[\varepsilon_p(\mathbf{s})] = 1/\kappa^2$, with $\kappa = \tau(\frac{2}{p(1-p)} + a_p)^{-0.5}$. In this respect, the specification of a Gamma prior for κ allows to keep $\text{Var}[\varepsilon_p(\mathbf{s})]$ the same at all quantiles. Further details can be found in Lum and Gelfand [2012]. As for $Z(\mathbf{s})$ an exponential correlation function with decay parameter θ is adopted. To ensure that θ takes values only on a plausible interval determined by the scale of data locations, we use a discrete uniform prior for this

parameter as well as for the other spatial processes $f_r(\mathbf{u})$ and $f_g(\mathbf{s})$. In the latent stage of the model, we typically set $\boldsymbol{\mu}_{f_g} = \mathbf{0}$ and $\boldsymbol{\mu}_{f_r} = \delta_1 \mathbf{f}_g$ with $\delta_0 = 0$. For all other parameters, m_j , λ_j , v_j , ψ_j and δ_1 , we assume a Normal prior distribution with mean equal to zero and variance chosen to be large to make the priors relatively non-informative. A uniform prior is used for the thresholds in $\boldsymbol{\gamma}_j$, whereas to all unknown error precision model parameters are assigned a Gamma prior. Posterior inference for the proposed model is facilitated by adapting standard MCMC algorithms to our model specification. The full conditional distributions for the spatial quantile regression model parameters are easy to obtain and are given in Lum and Gelfand [2012]. Also the conditional distributions involved in the latent stage of our model specification are not difficult to derive. In particular, for the sampling of the spatial common factors, we refer to Wang and Wall [2003] and Liu et al. [2005]. Instead, the Metropolis-Hastings algorithm is employed when the full conditional distributions are not available in closed form. Finally, it is worth noting that in the latent stage of the GLSQR model specification we are obliged to solve some identifiability issues linked both to scale uncertainty of the latent factors ($f_r(\mathbf{u})$ and $f_g(\mathbf{s})$) and the Probit formulation of equation (2) which does not allow a distinction between the error precision parameter $e_{g_j}(\mathbf{s})$ and threshold parameters. In the latter respect, we follow a standard solution of assuming that $e_{g_j}(\mathbf{s}) \sim N(0, 1)$, as suggested in Koop [2003].

4 Application to indoor radon data

The model discussed in Section 2 is applied to understanding the impact of specific variables on indoor radon concentrations. Owing to the well-established link with malignant diseases, indoor radon concentration has been the subject of numerous studies. Our illustrative analysis explicitly recognizes the multifactorial dependence of residential radon concentration and examines the relationships existing between indoor radon concentrations and some covariates linked to buildings, geological and radiometric features. The response variable of interest are the annual average radon concentrations in 710 buildings, mainly homes, mostly at ground level, collected by the regional Agency of Environmental Protection (ARTA) of Abruzzo, in the L'Aquila district. Data used in this study also include the physical characteristics of each building, the natural radioactivity content of the topsoil (equivalent uranium, equivalent thorium and potassium), the underlying geological characteristic (permeability, porosity, fracturation, karst) of the territory, some lithological information of L'Aquila district, and the altitude, integrated in the analysis as a crude proxy of climate-related variables, which have an influence on the emanation and diffusion radon process. Further details about this dataset can be found in Pasculli et al. [2014]. In our model specification altitude, building characteristics and lithotypes are directly included in the model, whereas radiometric and geological data are viewed as indicators of two interrelated latent factor. In the second stage of GLSQR specification, we consider a generalised spatial structural equation analysis, which

assumes that there are two latent common factors underlying the geological and radiometric variables (say f_g and f_r , respectively). Besides, since the distribution of radiometric factors depends on the geological characteristics of the area, we relate the two spatial common factor through a linear regression function. It is assumed that $\mathbf{f}_r | \mathbf{f}_g \sim N(\mathbf{1}\delta_0 + \delta_1 \mathbf{f}_g, \boldsymbol{\Sigma}_{f_r}(\boldsymbol{\omega}))$, where $\mathbf{1}$ is a vector of ones and $\boldsymbol{\Sigma}_{f_r}(\boldsymbol{\omega})$ a parameterized spatial covariance matrix that captures the spatial structure of the radiometric spatial common factor score. Notice that radiometric and geological data come from two independent sources. As a result, the linear regression model can be only estimated once the spatial factor scores $f_g(s)$ are interpolated on the sites u of the radiometric network. To obtain posterior summaries for the parameters the MCMC algorithm was run for 100000 iterations. Posterior inference was based on the last 50000 draws using every 5th member of the chain to avoid autocorrelation within the sampled values. Results of the analysis indicate a clear effect of the building characteristics on the variation of indoor radon concentration with an impact across quantiles of floor, building typology and type of soil connection. The regression coefficients associated to the estimated common spatial factor $\mathbf{f}_r$ (which synthesizes the radiometric variable) highlight a positive relationship with radon and suggest that the increments of the level of the factor also produce significant increments of the upper quantiles of the conditional indoor radon distribution. For these data the direct effect of the geological factor on indoor radon measures seems to be weak and constant across the quantiles while its indirect effect, estimated through the regression with the radiometric factor, is found significant and positive.

References

- R. Borgoni. A Quantile Regression Approach to Evaluate Factors Influencing Residential Indoor Radon Concentration. *Environ Model Assess*, 16:239–250, 2011.
- G. Koop. *Bayesian Econometrics*. John Wiley & Sons., Chichester, 2003.
- T.J. Kuzobowski and K. Podgorski. A multivariate and asymmetric generalization of laplace distribution. *Computation Stat*, 15(4):531–540, 2000.
- X. Liu, M.M. Wall, and J.S. Hodges. Generalized spatial structural equation modeling. *Epidemiology*, 6:539–557, 2005.
- K. Lum and E. Gelfand. Spatial Quantile Multiple Regression Using the Asymmetric Laplace Process. *Bayesian Anal*, 7(2):235–258, 2012.
- A. Pasculli, S. Palermi, A. Sarra, T. Piacentini, and E. Miccadei. A modelling methodology for the analysis of radon potential based on environmental geology and geographically weighted regression. *Environ Modell Softw*, 54:165–181, 2014.
- B.J. Reich, M. Fuentes, and D.B. Dunson. Bayesian Spatial Quantile Regression. *J Am Stat Assoc*, 106(493):6–20, 2012.
- F. Wang and M. M. Wall. Generalized common spatial factor model. *Biostatistics*, 4:569–582, 2003.