Joint quantile autoregressive modeling for univariate and spatial time-series data in a Bayesian framework

Jorge Castillo-Mateo[1], *Alan E. Gelfand*[2], *Jesus Asin*[1], *Ana C. Cebrian*[1]

[1]jorgecm@unizar.es, jasin@unizar.es, acebrian@unizar.es, Department of Statistical Methods, University of Zaragoza

[2]alan@stat.duke.edu, Department of Statistical Science, Duke University

Abstract

We propose a fully Bayesian joint quantile autoregression (QAR) modeling for time-series data. We derive a characterization of the noncrossing QAR(1) model using two monotone curves. We offer novel metrics to assess the adequacy of the QAR. Subsequently, we propose a novel spatial joint QAR for spatially referenced time-series data. We illustrate the models with an analysis of persistence in daily maximum temperature data collected in Aragón, Spain.

Keywords: Bayesian methods, daily temperature persistence, Gaussian copula process

1. Introduction

Quantile regression (QR) offers a flexible tool to capture changing explanation across quantile levels between the response and the covariates. The usual approach is the so-called *multiple* QR [1, 2], fitting a separate regression for each quantile of interest, leading to the possibility of crossing of the regression across quantiles. The approach called *joint* QR [5] avoids quantile crossing over a restricted support for the covariates. On the other hand, a seminal version of a joint quantile autoregression (QAR) model was proposed by Koenker and Xiao [4] (KX2006, hereafter). They required all the coefficients of the autoregression to be comonotonic, this is, strictly increasing functions across quantile levels $\tau \in (0, 1)$. Our contribution here is to reconsider the work by KX2006 in the context of Tokdar and Kadane [5] (TK2012, hereafter) to propose novel joint QAR modeling in a Bayesian framework with greater flexibility than KX2006. Going one step further, in the spatial setting, we introduce spatial dependence in the time-series realizations but as well, we add spatially-varying coefficients in order to obtain spatially-varying QAR's, generalizing the spatial QR model by [3] in the sense that they capture spatial dependence through a copula process but obtain a common quantile function that does not vary spatially.

2. Joint QAR model for time-series data

The support of the data. Let $\{y_t^* : t = 1, \ldots, T\}$ be the time-series data. For a noncrossing QAR(1) specification interest focuses on ensuring that the quantile curves do not cross for all values of y_{t-1}^* in a bounded interval. Although the region of interest for noncrossing must be assumed to be bounded, the variable space itself may still be unbounded. If noncrossing were desired in QAR on an unbounded domain, the result will be parallel lines, yielding the constant autoregression model. We take this interval to be $[0,1]$ and implement this by making a transformation of the data, $y_t = (y_t^* - m)/(M - m)$, where $m < \min y_t^*$ and $M > \max y_t^*$. For a convenient "automatic" strategy for selecting m and M we use basic results from the theory of order statistics where $y^*_{(1)}$ is the minimum and $y^*_{(T)}$ is the maximum of the data. We propose $m = (Ty^*_{(1)} - y^*_{(T)})/(T-1)$ and $M = (Ty^*_{(T)} - y^*_{(1)})/(T-1)$.

The model. A straightforward characterization of the required monotonicity of the QAR(1) lines is:

Theorem 1. *An autoregressive specification,* $Q_{Y_t}(\tau \mid y_{t-1}) = \theta_0(\tau) + \theta_1(\tau) y_{t-1}$ *with* $\theta_1(\tau) \in [-1,1]$ *for* $\tau \in [0,1]$, *is monotonically increasing in* τ *for* $y_{t-1} \in [0,1]$ *if and only if* $Q_{Y_t}(\tau \mid y_{t-1}) = \eta_2(\tau) + (\eta_1(\tau) - \eta_2(\tau)) y_{t-1}$ *where* $\eta_1, \eta_2 : [0,1] \to [0,1]$ *are monotonically increasing.*

A model for functions η_1 and η_2 induces a QAR(1) model over all valid QAR(1) specifications of $Q_{Y_t}(\tau \mid y_{t-1})$, provided the boundary conditions $Q_{Y_t}(0 \mid y_{t-1}) = 0$ and $Q_{Y_t}(1 \mid y_{t-1}) = 1$ for all $y_{t-1} \in [0,1]$ are satisfied, or equivalently, $\eta_j(0) = 0$ and $\eta_j(1) = 1$ $(j = 1,2)$. A convenient class of η's to work with are cdf's for continuous random variables with support $[0,1]$. In fact, a rich class would arise as probabilistic mixtures of such cdf's, leading to the general form $\eta(\tau) = \sum_{k=1}^{K} \lambda_k F(\tau \mid \mathbf{\Omega}_k)$, such that $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$ and $F : [0,1] \to [0,1]$ is strictly increasing for any parameters $\mathbf{\Omega}_k$. A convenient class of F's are the cdf's of the two parameter Kumaraswamy distribution. This cdf is $F(x \mid a, b) = 1 - (1 - x^a)^b$ where $x \in [0,1]$ and $a, b > 0$. The Kumaraswamy distributions are a family with behavior similar to the beta distribution but much simpler, especially in the context of simulation since the cdf can be expressed in closed form. Through simulation, we explored that $K = 1$ and $K = 2$ offer great flexibility and a higher K can lead to identification issues. We call these models QAR1K1 and QAR1K2, respectively. We conclude the model specification with the prior distribution of the parameters a's, b's, and λ's. We suggest to model the weights using the additive logistic normal transformation and the parameters of the Kumaraswamy distribution with a weak Gaussian prior in the log scale.

Likelihood evaluation and model fitting. Following the ideas of TK2012, a valid joint specification of $Q_{Y_t}(\tau \mid y_{t-1})$ for all $\tau \in (0,1)$ uniquely defines the conditional response density for $y_{t-1} \in [0,1]$,

$$f_{Y_t}(y_t \mid y_{t-1}) = \left. \frac{1}{\frac{d}{d\tau} Q_{Y_t}(\tau \mid y_{t-1})} \right|_{\tau = \tau_{y_{t-1}}(y_t)}, \tag{1}$$

where $\tau_{y_{t-1}}(y_t)$ solves $y_t = y_{t-1}\eta_1(\tau) + (1 - y_{t-1})\eta_2(\tau)$ in τ and is numerically approximated via a one-dimensional rootfinder. Consequently, given y_1, we can write a valid log-likelihood score in terms of $u_t = \tau_{y_{t-1}}(y_t)$, all of the observed data $\mathbf{y} = (y_1, \ldots, y_T)^\top$ and the model parameters $\mathbf{\Omega}$ as

$$\ell(\mathbf{\Omega} \mid \mathbf{y}) = -\sum_{t=2}^{T} \log \left\{ y_{t-1}\dot{\eta}_1(u_t) + (1 - y_{t-1})\dot{\eta}_2(u_t) \right\}. \tag{2}$$

The rootfinder used to evaluate the log-likelihood function (2) is Brent's method. We implement an adaptive block-Metropolis sampler algorithm to obtain Markov chain Monte Carlo (MCMC) samples from the posterior distribution of the parameters and the conditional quantile function.

Model adequacy and comparison. We offer two novel dimensionless metrics which assess the global adequacy and comparative performance of the conditional quantile function arising under the model. They are based on the posterior distribution of $Q_{Y_t}(\tau \mid y_{t-1})$. The first metric $\tilde{p}_v$ uses the probability that an observation is less than each conditional quantile. The second metric $\bar{R}^1$ is a generalization of $R^1(\tau)$, the analog of R^2 for the quantile loss function.

3. Joint spatial QAR model for spatio-temporal data

We focus on the analysis of spatial point-referenced time-series data where $Y_t(\mathbf{s})$ denotes the observation for time $t = 1, \dots, T$ at location $\mathbf{s} \in \mathcal{D}$, where $\mathcal{D} \subset \mathbb{R}^r$ is the study region. The joint spatial QAR model is given by

$$Y_t(\mathbf{s}) = \theta_0(U_t(\mathbf{s}); \mathbf{s}) + \theta_1(U_t(\mathbf{s}); \mathbf{s})Y_{t-1}(\mathbf{s}), \tag{3}$$

where the θ functions are quantile and spatially varying, and the vectors $(U_t(\mathbf{s}_1), \dots, U_t(\mathbf{s}_n))^\top$ are assumed to follow a spatial copula process.

Modeling spatial dependence. *Spatially varying quantiles.* For the spatially-varying coefficients, we consider one cdf for each $\eta(\tau; \mathbf{s})$. In fact, at location $\mathbf{s}$, let assume $\eta_j(\tau; \mathbf{s}) = 1 - (1 - \tau^{a_j(\mathbf{s})})^{b_j(\mathbf{s})}$ with parameters $a_j(\mathbf{s})$ and $b_j(\mathbf{s})$ $(j = 1, 2)$. We introduce four independent GP's for the a's and b's on the log scale. In particular, we model $\log a_j(\mathbf{s}) \sim GP(a_j, \sigma^2_{a_j}\rho(\mathbf{s}, \mathbf{s}'; \phi_{a_j}))$ and $\log b_j(\mathbf{s}) \sim GP(b_j, \sigma^2_{b_j}\rho(\mathbf{s}, \mathbf{s}'; \phi_{b_j}))$ where $\rho(\mathbf{s}, \mathbf{s}'; \phi)$is an exponential correlation functions with decay ϕ.

The spatial copula process. With regard to the copula model for (3), we take the processes $U_t(\mathbf{s})$'s to follow a Gaussian copula for each t, induced by a stationary spatial GP. In the spirit of [3], we define

$$U_t(\mathbf{s}) = \Phi(Z_t(\mathbf{s})), \quad Z_t(\mathbf{s}) = W_t(\mathbf{s}) + \epsilon_t(\mathbf{s}), \quad W_t(\mathbf{s}) \sim GP(0, \gamma\rho(\mathbf{s}, \mathbf{s}'; \phi)), \quad \epsilon_t(\mathbf{s}) \sim \text{IID } N(0, 1 - \gamma). \tag{4}$$

The process $W_t(\mathbf{s})$ captures spatial dependence while $\epsilon_t(\mathbf{s})$ is independent pure error. The parameter $\gamma \in [0, 1]$ determines the proportion of spatial and independent variation. With this approach, the Gaussian copula density has correlation matrix $R \equiv \gamma R(\phi) + (1 - \gamma)\mathbf{I}_n$ where $R(\phi)$ is the $n \times n$ correlation matrix induced by $\rho(\mathbf{s}, \mathbf{s}'; \phi)$.

Likelihood evaluation and spatial interpolation. We are interested in the likelihood under model (3) and (4). It is convenient to first obtain the joint distribution for all data, $\mathbf{y}$. By Sklar's theorem, the joint conditional density of $\mathbf{y}$ can be partitioned into a marginal part and a copula part. Subsequently, we find the expression of the log-likelihood function for the spatial QAR, and after giving weakly informative priors, inference proceeds in a similar way as in the univariate case. With the proposed model we can interpolate conditional quantiles to any desired location in the study region given any proposed or reference value for the previous day's temperature at that location.

4. Application to Temperature Data

The analyses consider daily maximum temperature (°C) data at $n = 18$ sites around the Comunidad Autónoma de Aragón provided by the Agencia Estatal de Meteorología (AEMET) in northeastern Spain. We use data at a daily scale in 2015, but we focus the analyses on the warm months from May 1 to September 30.

Illustratively, Figure 1 shows the posterior mean of the functions θ_0 and θ_1 in Zaragoza for the

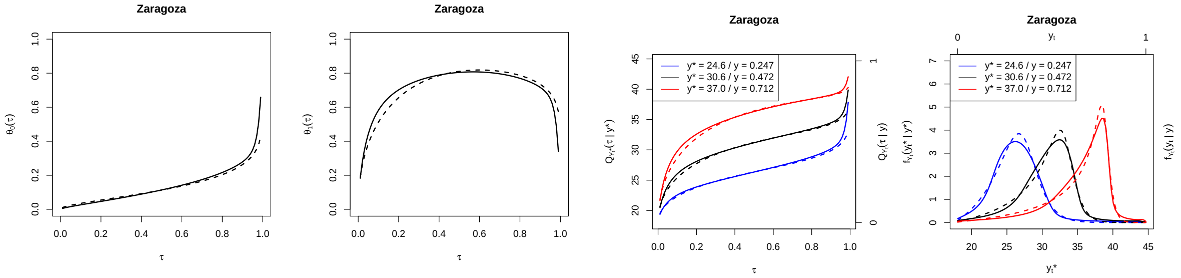

Figure 1: Posterior mean of (1st) $\theta_0(\tau)$, (2nd) $\theta_1(\tau)$, (3rd) quantile function $Q_{Y_t}(\tau \mid y)$ vs. τ; and (4th) density function $f_{Y_t}(x \mid y)$. y is the empirical τ-marginal quantile, $\tau = 0.1$ (blue), 0.5 (black), 0.9 (red).

models QAR1K1 (dashed) and QAR1K2 (solid). The intercepts on the original scale can be recovered as $\theta_0^*(\tau) = m(1 - \eta_1(\tau)) + M\eta_2(\tau)$. Mainly, note that θ_1 is nonmonotonic with smaller values in the extremes, this means that the previous day's temperature is less influential for extreme quantiles. This characteristic in the persistence of temperature was observed in [1, 2]. It cannot be reproduced by KX2006. Although higher K offers more flexibility, both curves offer similar results. Additionally, Figure 1 shows the posterior mean of the conditional quantile functions $Q_{Y_t}(\tau \mid y)$ for three situations where y is the empirical τ-marginal quantile for $\tau = 0.1, 0.5, 0.9$. The figure also shows the posterior mean of the conditional density function in (1) under the same conditions. The shape of the distribution changes according to the value on which we condition.

The spatial QAR model is fitted to the $n = 18$ series jointly. The posterior mean of γ, the proportion of spatial dependence in (4), is 0.95 with a 95% credible interval of $(0.93, 0.97)$. This result indicates very strong spatial dependence in the quantile levels. Results about spatial GP's for the parameters of the Kumaraswamy cdf (not shown) suggest that the GP of $a_2(\mathbf{s})$ might be not necessary but the spatial variability of $a_1(\mathbf{s})$ is higher and it could be related to distance to coast. We notice that $b_1(\mathbf{s})$ and $b_2(\mathbf{s})$ show approximately negative spatial correlation against each other because $b_1(\mathbf{s})$ has the highest values where $b_2(\mathbf{s})$ has the lowest.

5. Extensions and Future Work

The complete work also includes an approach for the QAR(p) case and a novel multivariate QAR for multivariate time-series data using a copula process. A future direction will consider a proper implementation of covariates in the joint QAR setting. Another interesting direction is to build a bivariate spatial QAR model for daily maximum and minimum temperature.

6. Acknowledgments

This work was partially supported by the Ministerio de Ciencia e Innovación under Grant PID2020-116873GB-I00 and TED2021-130702B-I00; Gobierno de Aragón under Research Group E46_20R: Modelos Estocásticos; and J. C.-M. was supported by Gobierno de Aragón under Doctoral Scholarship ORDEN CUS/581/2020. This work was done while J. C.-M. was a Visiting Scholar at Duke University.

7. Bibliography

[1] Castillo-Mateo J., Asín J., Cebrián A. C., Gelfand A. E. and Abaurrea J. (*in press*). Spatial quantile autoregression for season within year daily maximum temperature data. *Annals of Applied Statistics*. `https://doi.org/10.1214/22-AOAS1719`

[2] Castillo-Mateo J., Gelfand A. E., Asín J. and Cebrián A. C. (2022). Spatio-temporal quantile autoregression for detecting changes in daily temperature in northeastern Spain. In S. Cabras, I. Cascos, M. E. Castellanos, M. Durbán (Eds.), *Book of Abstracts XVIII Congreso de Biometría CEB-MADRID* (pp. 122–126). Universidad Carlos III de Madrid.

[3] Chen X. and Tokdar S. T. (2021). Joint quantile regression for spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4), 826–852.

[4] Koenker R. and Xiao Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, 101(475), 980–990.

[5] Tokdar S. T. and Kadane J. B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis*, 7(1), 51–72.