Statistical Analysis of Response Distribution for Dependent Data via Joint Quantile Regression

by

Xu Chen

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Surya T. Tokdar, Advisor

_____
Amy H. Herring

_____
Mike West

_____
Alan E. Gelfand

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2021

Abstract

Statistical Analysis of Response Distribution for Dependent Data via Joint Quantile Regression

by

Xu Chen

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Surya T. Tokdar, Advisor

_____
Amy H. Herring

_____
Mike West

_____
Alan E. Gelfand

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2021

Abstract

Linear quantile regression is a powerful tool to investigate how predictors may affect a response heterogeneously across different quantile levels. Unfortunately, existing approaches find it extremely difficult to adjust for any dependency between observation units, largely because such methods are not based upon a fully generative model of the data. In this dissertation, we address this difficulty for analyzing spatial point-referenced data and hierarchical data. Several models are introduced by generalizing the joint quantile regression model of Yang and Tokdar (2017) and characterizing different dependency structures via a copula model on the underlying quantile levels of the observation units. A Bayesian semiparametric approach is introduced to perform inference of model parameters and carry out prediction. Multiple copula families are discussed for modeling response data with tail dependence and/or tail asymmetry. An effective model comparison criterion is provided for selecting between models with different combinations of sets of predictors, marginal base distributions and copula models.

Extensive simulation studies and real applications are presented to illustrate substantial gains of the proposed models in inference quality, prediction accuracy and uncertainty quantification over existing alternatives. Through case studies, we highlight that the proposed models admit great interpretability and are competent in offering insightful new discoveries of response-predictor relationship at non-central parts of the response distribution. The effectiveness of the proposed model compar-

ison criteria is verified with both empirical and theoretical evidence.

To my family

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

ALD	Bayesian quantile regression model using asymmetric Laplace error distribution by Yu and Moyeed (2001)
ALP	Spatial quantile regression model using asymmetric Laplace process by Lum and Gelfand (2012)
AR	Autoregressive
ASQR	Approximate spatial quantile regression
BQR	Bayesian linear heteroscedastic model with infinite Gaussian mixture error by Reich et al. (2010)
BS	B-splines
BSRE	Basic spatial random effects model
CDF	Cumulative distribution function
COP	Copula-based model with a three-step estimation procedure by Wang et al. (2019)
CV	Cross validation
EMP	Blocked empirical likelihood approach by Wang and Zhu (2011)
GP	Gaussian process
HS&B	High school and beyond survey
iWAIC	WAIC score corresponding to within-cluster prediction performance
JHQR	Joint hierarchical quantile regression
JQR	Joint quantile regression

JSQR	Joint spatial quantile regreesion
KB	The classical optimization based method by Koenker and Bassett (1978)
MAE	Mean absolute error
MCMC	Markov chain Monte Carlo
NNGP	Nearest neighbor Gaussian process
oWAIC	WAIC score corresponding to out-of-cluster prediction performance
pdf	Probability density function
$PM_{2.5}$	Particular matter with diameter less than 2.5 μm
QR	Quantile regression
TP	Student-t process
WAIC	Watanabe-Akaike information criteria
YT17	Yang and Tokdar (2017)

Acknowledgements

First I want to express my sincere gratitude to my advisor, Surya Tokdar, for his constant support, encouragement and guidance throughout my doctoral journey. He is always supportive of my goals and interests, and helps me identify research problems that I am excited about. He provides numerous valuable insights and inspirations to my research and I benefit a lot from observing how he approaches problems. In addition to my research, he also cares about my career development and personal well-being. I could not have asked for a better mentor.

I would also like to thank Alan Gelfand, Amy Herring and Mike West for serving on my prelim and dissertation committees. Their suggestions inspired me to think more deeply on statistical research and improved the work and presentation. I very much enjoyed our discussions.

I am also grateful to faculty members and staff in the department. Special thanks to Peter Hoff and Fan Li, for their support during my doctoral study, and to Lori Rauch and Karen Whitesell, for taking care of all administrative affairs.

I am thankful to Katherine Goodman Stern Fellowship for generously supporting my research projects. I would also like to thank National Science Foundation for supporting the research presented in this dissertation through grants DMS-1613173 and DMS-2014861.

I would love to thank all my friends for the great times we enjoyed together. Special thanks to Erika Cunningham, Yuan Gao, Sheng Jiang, Bai Li, Wenxi Liao,

Jiachang Liu, Jialiang Mao, Hanyu Song, Jiurui Tang, Lu Wang, Ye Wang, Azeem Zaman, Shuxi Zeng and Yan Zhang.

Finally, I want to express my deepest gratitude to my parents and my girlfriend Weiwei for their unconditional love and support. This work is dedicated to them.

1

Introduction and Background

1.1 Introduction

As a robust alternative to least squares regression, linear quantile regression (Koenker and Bassett, 1978) provides a powerful tool to characterize complex relations between response and predictors that may vary across different quantile levels. In quantile regression (QR, hereafter), one relates a scalar response Y to a predictor vector $X \in \mathbb{R}^p$ via the "model":

$$Q_Y(\tau \mid X) = \beta_0(\tau) + X^\top \beta(\tau), \tag{1.1.1}$$

where $Q_Y(\tau \mid x) = \inf\{a : P(Y \leq a \mid X = x) \geq \tau\}$ is the τ-th conditional response quantile given $X = x$. Given data $(X_i, Y_i)_{i=1}^n$, the regression parameters $\beta_0(\tau)$, $\beta(\tau) = (\beta_1(\tau), \ldots, \beta_p(\tau))^\top$, signifying level-specific intercept and slopes, are typically estimated by minimizing the sample averaged check loss $n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - b_0 - X_i^\top b)$ in (b_0, b), where $\rho_\tau(\varepsilon) = \varepsilon\{\tau - \mathbb{1}(\varepsilon < 0)\}$ also depends on the chosen level τ. These estimates are more robust against heteroskedasticity and outliers than least squares regression estimates. More importantly, by varying the quantile level τ, the analyst could examine predictor effects that may be present only in the tails, thus capturing

a richer variety of dependence than what is possible under ordinary mean or median regression; see Koenker (2005) for a review.

While quantile regression has gained popularity in multiple areas of scientific applications (Buchinsky, 1994; Cade et al., 1999; Elsner et al., 2008), scant methodological work exists for advancing models that are capable of accounting for additional data dependency. Nevertheless, modern scientific research is flooded with data that have various structural dependency between observations. When such dependence structures exist across observations units, ignoring noise correlation between dependent records may lead to biased estimates of regression parameters and/or overconfident uncertainty quantification. But the task of modeling, estimating and adjusting for spatial noise correlation runs into serious difficulties under the commonly held belief that the QR formulation (1.1.1) must be interpreted only locally, at a single prespecified quantile level (Koenker, 2017).

This belief does not square with common practice. Most scientific applications of QR examine whether and how the regression parameters vary with the quantile level (Cade et al., 1999; Machado and Mata, 2005; Elsner et al., 2008), typically resorting to the questionable practice of patching up estimates and p-values gathered from separate analyses (see Tokdar and Kadane, 2012, for a detailed critique). Such patch-ups are especially questionable in the presence of spatial dependence even if noise correlation was accounted for in each separate analysis. Any assumed noise correlation model induces a joint distribution for the observed response values $(Y_1, \ldots, Y_n)$. But the induced joint distributions stemming from two different choices of the quantile level are typically distinct, and hence the two sets of estimates are mutually incongruous. Put it differently, the parameter estimates at different quantile levels, though presented together, are actually obtained under very different and potentially conflicting assumptions on the same set of data.

An alternative and rapidly emerging viewpoint adopts the formulation (1.1.1)

simultaneously across all quantile levels to construct a single estimating model for the entire data set (He, 1997; Reich et al., 2011; Tokdar and Kadane, 2012; Yang and Tokdar, 2017). The regression parameters are jointly estimated as smooth functions of the quantile level under the natural *non-crossing* constraint:

$$\frac{\partial}{\partial\tau}\{\beta_0(\tau) + x^\top\beta(\tau)\} > 0 \quad \text{for all } \tau \in (0,1) \text{ and all } x \in \mathcal{X}, \tag{1.1.2}$$

improving estimation, uncertainty quantification and prediction (Yang and Tokdar, 2017, YT17, hereafter). Here $\mathcal{X} \subset \mathbb{R}^p$ is the domain of X, and is assumed to be bounded and convex. Equally important, this comprehensive view offers a novel and congruous formalism of *noise* in the context of quantile regression, taking it beyond the notion of *residual* around a single fitted line. The linear QR formulation (1.1.1) taken simultaneously for all $\tau \in (0,1)$ and restricted to the non-crossing condition (1.1.2) is equivalent to the fully generative model for the response data:

$$Y = \beta_0(U) + X^\top\beta(U), \quad U \sim \mathsf{Unif}(0,1), \tag{1.1.3}$$

where U is a random quantile level independent of X. This equivalence is a simple consequence of the so called "inverse CDF technique" for random variable generation. This new formalism enables incorporating noise correlation within QR. The equivalence between (1.1.1) and (1.1.3) remains valid even if the random quantile levels $U_1, \ldots, U_n$ associated with the observation units are mutually dependent, as long as each $U_i \sim \mathsf{Unif}(0,1)$, i.e., their joint distribution must be a *copula*.

In this dissertation, we combine various copula models with the joint QR estimation framework of Yang and Tokdar (2017) to design a novel estimation framework for dependent QR. We focus on two common types of dependency structures: spatial dependence when observations are collected at locations over a geographical region, and, hierarchical dependence when observations have a natural multi-level structure. While a combination like this is unsurprising, it breaks important new ground by making QR practicable in a broad range of applications with dependent data.

The rest of this dissertation is organized as follows. In Section 1.2 of the current chapter, we start with a brief review of the YT17 framework for independent data and introduce some notations that will be used throughout the dissertation. In Chapter 2, we focus on developing a dependent QR framework which adjusts for spatial dependence among observations. In Chapter 3, we propose three different models to model different types of within-cluster dependency for hierarchical data. In these two chapters, we thoroughly examine the proposed models from perspectives of estimation, prediction and computation. We highlight the comparisons between the proposed models and existing alternatives as well as classical linear models. In Chapter 4, we discuss the desiderata of determining the copula model and provide several copula options to account for different dependency behaviors of response. In addition, we propose information criteria for model selection according to the prediction performance of a model. We then illustrate the use of proposed models along with model selection criteria in Chapter 5 with four case studies. Chapter 6 concludes the dissertation with a summary of the contributions and future research directions.

1.2 Notations and Background

As our model is building upon the framework of Yang and Tokdar (2017), we start with reviewing this framework and introducing notations. YT17 introduce a novel parametrization which translates the non-crossing condition (1.1.2) into an almost constraint-free specification leading to efficient and embarrassingly parallel likelihood evaluation. Inference can then be proceeded using regularized optimization or Bayesian approach. In the latter category, YT17 propose a semiparametric method where function valued parameters are assigned Gaussian process priors under which posterior consistency is established.

1.2.1 Reparametrization scheme

Yang and Tokdar (2017) build a linear quantile regression model as a natural extension of standard linear model: $Y = \gamma_0 + X^\top \gamma + \sigma\varepsilon,\ \varepsilon \sim f_0$ for some probability density function (pdf) f_0 on $\mathbb{R}$ and let q_0 be the corresponding quantile density, that is, the derivative of the quantile function. Let $\tau_0 = \int_{-\infty}^{0} f_0(z)\mathrm{d}z$. One may view f_0 as a prior guess for the density, up to a scale parameter, of the level-τ_0 residual: $Y - \beta_0(\tau_0) - X^\top\beta(\tau_0)$. When the support of Y is $(-\infty, \infty)$, one may take f_0 to be the standard logistic or a standard Student-t distribution (and $\tau_0 = 0.5$); the latter being attractive for modeling heavy tailed data. These are the choices adopted in the rest of the paper. For positive valued response, one could take f_0 to be a Gamma distribution or a generalized Pareto distribution (with $\tau_0 = 0$); again the latter being appropriate for heavy tailed data.

Yang and Tokdar (2017) assume, without loss of generality, that $\mathcal{X}$ is convex and bounded and contains 0 as an interior point. The former relies on the fact that (1.1.2) holds over a set $\mathcal{X}$ if and only if it holds over the convex hall of $\mathcal{X}$. The latter can be achieved by simple translation. For any non-zero vector $b \in \mathbb{R}^p$, define the "projection radius opposite b" as $a(b, \mathcal{X}) = \sup_{x\in\mathcal{X}}\{-x^\top b\}/\|b\|$ and $a(0, \mathcal{X}) = 1$, where $\|\cdot\|$ is the Euclidean norm. Reformulate the intercept and slope functions of (1.1.3) as:

$$\beta_0(\tau_0) = \gamma_0, \quad \beta(\tau_0) = \gamma \tag{1.2.1}$$

$$\beta_0(\tau) - \beta_0(\tau_0) = \sigma \int_{\zeta(\tau_0)}^{\zeta(\tau)} q_0(u)\mathrm{d}u, \quad \tau \in (0,1) \tag{1.2.2}$$

$$\beta(\tau) - \beta(\tau_0) = \sigma \int_{\zeta(\tau_0)}^{\zeta(\tau)} \frac{\omega(u)}{a(\omega(u), \mathcal{X})\sqrt{1 + \|\omega(u)\|^2}} q_0(u)\mathrm{d}u, \quad \tau \in (0,1) \tag{1.2.3}$$

$\gamma_0 \in \mathbb{R}$; $\gamma \in \mathbb{R}^p$; $\sigma > 0$; $\omega : (0,1) \to \mathbb{R}^p$; and $\zeta : [0,1] \to [0,1]$ which is restricted to be a differentiable, monotonically increasing bijection, that is, a diffeomorphic

deformation of the unit interval. This formulation gives an exhaustive representation of all linear QR models (1.1.3) subject to the non-crossing constraint (1.1.2) and a support match between Y and f_0 as described above (Yang and Tokdar, 2017, Theorem 2).

The above reformulation embeds the non-crossing constraint mostly through the transformation (1.2.3). The function valued parameters $\omega = (\omega_1, ..., \omega_p)$ underlying the slopes are completely unconstrained. The shape restrictions on the diffeomorphism $\zeta(\cdot)$ are relieved via an additional transformation:

$$\zeta(\tau) = \frac{\int_0^\tau \exp\{\omega_0(u)\}\mathrm{d}u}{\int_0^1 \exp\{\omega_0(u)\}\mathrm{d}u}, \quad \tau \in (0,1),$$

where the function ω_0 is also completely unconstrained, except for the continuity properties needed to make the above definition valid. At this point, all function valued parameters $\omega = (\omega_1, ..., \omega_p)$ and ω_0 are unconstrained and hence well suited to be estimated with either splines using optimization-based methods or with Gaussian process priors to induce smoothness regularization.

1.2.2 Likelihood evaluation

This clever reparametrization leads to efficient likelihood evaluation. For dataset $\{(X_i, Y_i) : i = 1, \ldots, n\}$, the model specification (1.1.3) links responses and predictors through $Y_i = \beta_0(U_i) + X_i^\top \beta(U_i), \quad U_i \sim \mathsf{Unif}(0,1)$ where U_i's are unobserved quantile levels. Simple calculation gives

$$f_Y(y \mid x) = \left. \frac{1}{\frac{\partial}{\partial \tau} Q_Y(\tau \mid x)} \right|_{\tau = \tau_x(y)}$$

where $\tau_x(y)$ solves $Q_Y(\tau \mid x) = y$ in τ (Tokdar and Kadane, 2012; Yang and Tokdar, 2017). Given the configuration (1.2.1)–(1.2.3), the underlying quantile level for ith observation is essentially solved by

$$Y_i = \gamma_0 + X_i^\top \gamma + \int_{\tau_0}^{\tau} \dot{\beta}_0(u) + X_i^\top \dot{\beta}(u)\mathrm{d}u \tag{1.2.4}$$

in τ, which may be numerically approximated to arbitrary precision as follows. Fix a dense grid of points $T = \{0 = t_1 < \cdots < t_L = 1\}$ and use it to approximate $Q_i(\tau) := \beta_0(\tau) + X_i^\top \beta(\tau) = \gamma_0 + X_i^\top \gamma + \int_{\tau_0}^{\tau} \{\dot{\beta}_0(t) + X_{ij}^\top \dot{\beta}(t)\} \mathrm{d}t$ at any τ on the grid by the trapezoidal rule of integration, and, by linear interpolation for any τ in between successive grid points. Identify the index $l = l_i$ such that $Q_i(t_l) \leqslant Y_i < Q_i(t_{l+1})$ and calculate U_i by analytically inverting the linear interpolation in between these successive grid points. Note that the derivatives $\dot{\beta}_0(\tau) = \sigma q_0(\zeta(\tau))\dot{\zeta}(\tau)$ and $\dot{\beta}(\tau) = \dot{\beta}_0(\tau) h(\omega(\zeta(\tau)))$, where $h(b) = b/\{a(b, \mathcal{X}) \cdot \sqrt{1 + \|b\|^2}\}$, are required to be evaluated only for $\tau \in T$. See Yang and Tokdar (2017) for more details on the choice of the grid T and calculation of U_i's. The associated $\mathcal{O}(n)$ calculations can be carried out with any univariate root finding algorithms and are embarrassingly parallel across observation units.

After obtaining the latent quantile levels U_i's, the log-likelihood for independent observations is given by

$$\ell(\gamma_0, \gamma, \sigma, \omega, \zeta) = -\sum_{i=1}^{n} \log\{\dot{\beta}_0(U_i) + X_i^\top \dot{\beta}(U_i)\}. \tag{1.2.5}$$

2

Joint Quantile Regression for Spatial Data

2.1 Introduction

As aforementioned in Chapter 1, notwithstanding the popularity of QR methods in real applications, scant work exists in the domain of spatial data analysis. However, ignoring the spatial dependency structure in data analysis may result in serious problems in both inference and prediction. The challenge of incorporating dependency structure into the model presents a considerable difficulty in applying QR methods to spatial data analysis.

Hallin et al. (2009) address this challenge with a two-stage approach where both response and predictor data are first *spatially detrended* via smoothing and then subjected to a (locally) linear quantile regression analysis. Their approach depends fundamentally on the assumption that spatial dependency in the observed data arises because of unobserved, spatially smooth shifts added to realizations of X and Y. This is akin to assuming a spatial random intercept model, which is unfit to capture a rich class of plausible spatial dependency structures. A more comprehensive approach is offered by Lum and Gelfand (2012) who embed (1.1.1) within the widely used

asymmetric Laplace error regression model (Yu and Moyeed, 2001) and then extend it to a novel spatial process model. Despite its popularity, the asymmetric Laplace error model is known to be sensitive to heteroskedasticity and outliers, and may severely underestimate uncertainty in parameter estimation (Tokdar and Kadane, 2012). Moreover, the asymmetric Laplace process model based analysis is sensitive to prior choice and the resulting statistical performance appears sub-par at quantile levels away from the median (Sections 2.5 and 5.1).

In this chapter, we propose a novel spatially dependent QR by combining Gaussian and Student-t spatial copula processes with the joint QR estimation framework of Yang and Tokdar (2017). Our key research contributions are summarized in the following paragraphs.

Through extensive numerical experiments, the combined model is shown to greatly reduce estimation bias and improve uncertainty quantification in regression parameter estimation over existing alternatives which may or may not attempt to adjust for noise correlation (Section 2.5.1), substantiating the need and utility of adjusting for noise correlation before QR methods could be adopted to spatial data analysis. Moreover, the new model is shown to embed computationally and statistically efficient spatial smoothing which delivers model based regression quantile kriging equipped with meaningful uncertainty quantification (Sections 2.4.1, 2.5.1, 5.1) which could be a powerful tool in the analyst's toolbox to produce reliable prediction of extreme responses at new locations.

A great advantage of a model based QR analysis is that it could be customized to accommodate various scientific concerns; see Tokdar and Kadane (2012) for an application to hurricane intensity trends. A similar advantage for the proposed model is demonstrated through an important extension that targets heavy tailed response distributions and associated tail dependence across observation units, an important and challenging problem in spatial analysis. It is shown that competing flavors

of the model could be formally compared to one another by using an *information criterion* whose results are congruous with a substantially more expensive cross-validated assessment of hold-out quantile prediction accuracy (Sections 4.1, 5.1).

In two real world applications involving air quality and wildfire risk, the new model is shown to provide excellent fit to hold-out data, detect important tail effects of key predictors, and, detect the presence of heavy tails and tail dependence in noise correlation (Section 5.1). These case studies underscore the tremendous potential of QR in delivering a detailed and insightful analysis of potentially heavy tailed spatial data with heterogeneous predictor effects, over and beyond the capability of existing spatial regression models. Furthermore, our estimation method is shown to scale well to large data sets by incorporating commonly used approximation techniques from Gaussian spatial process literature, thus enabling actual deployment of the new model to serious scientific investigations (Section 2.3.3).

We end this introduction with a note that the problem of quantile regression under spatial noise correlation is quite different from the one of spatially varying quantile regression as explored by Reich et al. (2011) and Yang and He (2015). Statistical analysis in the latter problem relies on having essentially independent replications of the response measurement at identical or nearby locations and existing methods typically employ spatial smoothing of locally estimated quantile regression parameters. More fundamentally, the underlying theory posits that the structural equations representing the response-predictor relation vary across space. In contrast, our approach theorizes that a single quantile regression formulation as in (1.1.1) holds globally at all spatial locations, and, the learning of these global model parameters, while adjusting for noise correlation, is the primary goal of the analysis. However, our model does allow spatial quantile smoothing necessary for *infill* prediction at new locations where only the covariates are recorded. Although this type of smoothing has less shape flexibility than what is offered by fully spatially varying approaches,

it can still offer a superior performance under moderate spatial variation of predictor effects (Section 2.4.2).

2.2 Joint spatial quantile regression

2.2.1 Modeling spatial noise correlation

Our focus is on analyzing spatial point-referenced data where paired predictor-response observations (X_i, Y_i), $i = 1, \ldots, n$, are collected at known locations $s_i \in \mathcal{S} \subset \mathbb{R}^r$. The joint QR model postulates $Y_i = \beta_0(U_i) + X_i^\top \beta(U_i)$, where $(U_1, \ldots, U_n)$ follows a copula distribution. In the spirit of the Kolmogorov extension theorem, it is natural and useful to view the random quantile levels as $U_i = U(s_i)$, $i = 1, \ldots, n$, where $(U(s) : s \in \mathcal{S})$ is a stochastic process on $\mathcal{S}$ such that $U(s) \sim \mathsf{Unif}(0, 1)$ at every $s \in \mathcal{S}$. Such an embedding within a spatial copula process is essential to formalize prediction at new spatial locations.

Towards an interpretable and practicable statistical analysis of dependence, it is pragmatic to consider a parametric family of spatial copula processes indexed by a low dimensional parameter vector that describes both the nature and the strength of spatial correlation. For simplicity we let the correlation between any $U(s)$ and $U(s')$ depend only on the spatial distance $\|s - s'\|$. As this distance increases, the correlation must diminish to zero and near-independence must be realized within the range of the observed spatial domain. Any hope of reliable inference of the regression parameters rests on this assumption that at least at great distances within the observed spatial region, data could be taken as nearly independent realizations from the model.

We meet these modeling needs by taking $(U(s) : s \in \mathcal{S})$ to follow a Gaussian copula, induced by a stationary Gaussian spatial process. The correlation function of the Gaussian spatial process could be used to specify well structured spatial dependency models with only a small number of unknown parameters controlling the smoothness and the decay range of the correlation. The conditional copula distribu-

tions and quantile functions, key quantities needed for infill prediction and spatial interpolation, are straightforward to compute (Section 2.4.1). Equally important, the use of Gaussian process is compatible with the rich literature on spatial modeling and incorporates the popular basic spatial random effects model as a special case (Cressie, 1993; Banerjee et al., 2008, more details below).

The Gaussian copula process is specified as follows. Let $\Phi(\cdot)$ denote the cumulative distribution function (CDF) of $\mathsf{N}(0,1)$. We define

$$U(s) = \Phi(Z(s)), \quad Z(s) = W(s) + \varepsilon(s), \tag{2.2.1}$$

$$W(s) \sim \mathsf{GP}\left(0, \alpha\rho_{\mathrm{M}}(s, s'; (\nu, \phi))\right), \ \varepsilon(s_i) \overset{\text{iid}}{\sim} \mathsf{N}(0, 1-\alpha) \tag{2.2.2}$$

where

$$\rho_{\mathrm{M}}(s, s'; (\nu, \phi)) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\|s-s'\|}{\phi}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{\|s-s'\|}{\phi}\right). \tag{2.2.3}$$

is the Matérn correlation function with smoothness parameter ν and decay parameter ϕ, a widely popular choice in spatial statistics (Cressie, 1993; Stein, 2012; Banerjee et al., 2014). Here, $\Gamma(\cdot)$ is the gamma function and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind of order ν. Our approach is trivially generalized to non-Matérn correlation functions. We take ν to be of a fixed, user specified value because it is typically difficult to estimate ν from data and often a value of $\nu < 3$ is deemed sufficient for real data analysis (Banerjee et al., 2014). Thus, the spatial copula formulation is indexed by a two dimensional parameter $\theta = (\alpha, \phi)$.

In (2.2.2) the variation in the underlying quantile levels is decomposed into two parts in a similar spirit as the basic spatial random effects model (BSRE; Cressie, 1993; Banerjee et al., 2008, see Equation (5.1.1)). The process $W(s)$ captures structural spatial association while $\varepsilon(s)$ is uncorrelated pure error. The parameter $\alpha \in [0, 1]$ determines the proportion of variation that is spatially structured. When $\alpha = 1$, very little independent information is expected from nearby

observations. When $\alpha = 0$, the model reduces to the independent noise model. As a desirable byproduct, the introduction of α also improves numerical stability. Additionally, when $\beta_0(u) = \sigma\Phi^{-1}(u) + \tilde{\beta}_0$ and $\beta(u) = \tilde{\beta}$, the model becomes $Y_i = \tilde{\beta}_0 + X_i^\top \tilde{\beta} + \tilde{W}(s_i) + \tilde{\varepsilon}(s_i)$ with $\tilde{W}(s) \sim \mathsf{GP}\left(0, \alpha\sigma^2\rho(s, s'; (\nu, \phi))\right)$ and $\tilde{\varepsilon}(s_i) \overset{\text{iid}}{\sim} \mathsf{N}(0, (1-\alpha)\sigma^2)$. This special formulation coincides with BSRE.

The use of Gaussian spatial copulae leads to a fair number of computational advantages, e.g., ease of parameter estimation, analytical quantile kriging, and scalability with sample size. We shall elaborate on this presently. But Gaussian copula are limited in their ability to model extreme tail dependence, which may be present in heavy tailed response distributions. We return to this in Section 4.1 with a generalization to a t-copula for an in-depth treatment of modeling possibly heavy tailed data.

2.2.2 Prior specification for Bayesian estimation

We adopt a semiparametric Bayesian approach for making inference on model parameters (β_0, β, θ). The marginal model parameters $(\beta_0, \beta)^\top$ are first reparametrized into constraint-free parameters $(\gamma_0, \gamma, \sigma, \omega_0, \omega_1, \ldots, \omega_p)$ using the scheme described in Section 1.2. We then follow YT17 and adopt the priors below on the new model parameters:

$$\omega_j \sim \mathsf{GP}(0, \kappa_j^2 \rho_{\text{SE}}(\cdot, \cdot; \lambda_j)) \quad j = 0, \ldots, p$$

$$\kappa_j^2 \overset{\text{iid}}{\sim} \mathsf{Inv\text{-}Ga}(0.1, 0.1) \quad \lambda_j \sim \pi_\lambda(\lambda_j)$$

$$(\gamma_0, \gamma, \sigma^2) \sim \pi(\gamma_0, \gamma, \sigma^2) \propto \frac{1}{\sigma^2}$$

where $\rho_{\text{SE}}(s, s'; \lambda) = \exp\left(-\lambda^2 \|s - s'\|\right)$ is the square exponential correlation function with rescaling parameter λ. The prior specification on the function valued parameters $\eta_0, \ldots, \eta_p$ is motivated by a rich literature on Bayesian nonparametric smoothing that shows that square-exponential Gaussian process priors, equipped with a hyper-prior

on the rescaling parameter, offers optimal regularization and smoothness adaptation in a variety of function estimation problems (Tokdar and Ghosh, 2007; van der Vaart et al., 2009; Yang and Tokdar, 2015). A hyper-prior on λ is implicitly specified by choosing $r_j := \mathsf{Cor}(\eta_j(\tau), \eta_j(\tau + 0.1) \mid \lambda_j) = \exp(-0.1^2\lambda_j^2)$ to be $\mathsf{Be}(6, 4)$ distributed. The priors on $(\gamma_0, \gamma, \sigma)$ are purposely left diffuse. These parameters, corresponding to a central quantile plane (if $\tau_0 \approx 0.5$) and the overall scale of the response, are well informed by all observations in the data and require little prior regularization. In the case of dealing with positive valued response, one may choose to fix γ_0, γ at zero. See Yang and Tokdar (2017) for more details.

To address the additional copula piece of our model, we propose the following prior specification on $\theta = (\alpha, \phi)$. We choose the $\mathsf{Unif}(0, 1)$ distribution as a generic prior for α. This is not an entirely automatic "low-informative" choice. But our experimentation shows the uniform prior leads to adaptive estimation of the strength of spatial dependence (Section 2.5.2). However, if prior knowledge is available on the value or the range of the proportion of spatial correlation, a Beta or a truncated prior may be adopted.

More care is needed in choosing a prior for the correlation range parameter ϕ. It is useful to specify bounds for ϕ from the perspective of *effective range* of spatial correlation, that is, the shortest distance at which the spatial correlation falls below a small threshold (typically 0.05). As discussed earlier, any hope of reliable parameter estimation from spatially dependent data relies on the assumption of near independence between far away observation units. In any particular application, the bounds for ϕ can be determined according to one's belief of lower and upper limits of the effective range. Additionally, from a methodological perspective, the decay parameter is only weakly identified and an informative prior is needed for satisfactory and reproducible posterior inference and computation (Banerjee et al., 2008, 2014).

For a default choice, we follow the convention in the spatial modeling literature

and specify the range of ϕ so that the effective range lies between one-fourth and three-fourths of the maximal pairwise distance among all locations (Banerjee et al., 2014). The prior for ϕ is taken to be a discrete uniform distribution over a dense grid of values within the specified range. This choice of range keeps ϕ away from 0 and eliminates the identifiability issue that manifests when α approaches 0. From a computational perspective, copula density evaluation (Section 2.3.1) is expensive as it requires $\mathcal{O}(n^3)$ flops in each MCMC iteration. By using a discrete uniform prior, only finitely many correlation matrices are required to be computed and stored before initiating MCMC computation. Consequently, the computational overhead can be reduced to $\mathcal{O}(n^2)$. See Section 2.3.3 for additional reduction in computational complexity for large data.

2.3 Posterior computation

2.3.1 Likelihood evaluation

By Sklar's theorem, the joint conditional density of the response data given predictors can be partitioned into a marginal part and a copula part

$$p(Y_{1:n}, \mid X_{1:n}, s_{1:n}) = \left\{ \prod_{i=1}^{n} f_Y(y_i \mid x_i) \right\} \times c_{s_{1:n}} \left(F_Y(y_1 \mid x_1), \ldots, F_Y(y_n \mid x_n) \right)$$

where F_Y is the CDF corresponding to pdf f_Y and $c_{s_{1:n}}$ denotes the joint density of $(U(s_1), \ldots, U(s_n))$ under the spatial copula formulation on $(U(s) : s \in \mathcal{S})$. One may evaluate f_Y and F_Y via the identities

$$f_Y(y \mid x) = \left. \frac{1}{\frac{\partial}{\partial \tau} Q_Y(\tau \mid x)} \right|_{\tau = \tau_x(y)}, \quad F_Y(y \mid x) = \tau_x(y), \tag{2.3.1}$$

where $\tau_x(y)$ solves $y = Q_Y(\tau \mid x) = \beta_0(\tau) + x^\top \beta(\tau)$ in τ. Therefore, the log-likelihood score of model parameters can be expressed as

$$\ell(\gamma_0, \gamma, \sigma, \eta, \zeta, \theta) = -\sum_{i=1}^{n} \log\{\dot{\beta}_0(U_i) + X_i^\top \dot{\beta}(U_i)\} + \log c_{s_{1:n}}(U_1, \ldots, U_n \mid \theta) \tag{2.3.2}$$

where $U_i = \tau_{X_i}(Y_i)$ may be numerically approximated with the algorithm described in Section 1.2.2. Let $Z = (\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_n))^\top$ and K denote the correlation matrix with $K_{ij} = \rho_{\mathrm{M}}(s_i, s_j \mid (\nu, \phi))$. Using spectral decomposition $K = \Gamma\Lambda\Gamma^\top$, the logarithm of the Gaussian copula density is

$$\log c_{s_{1:n}}(U_1, ..., U_n \mid \theta) =$$
$$-\tfrac{1}{2}\log|\alpha\Lambda + (1-\alpha)I_n| - \tfrac{1}{2}Z^\top\Gamma((\alpha\Lambda + (1-\alpha)I_n)^{-1} - I_n)\Gamma^\top Z.$$

Note that $\alpha\Lambda + (1-\alpha)I_n$ is a diagonal matrix and hence its determinant and inverse are easy to compute. Because the prior on ϕ is finitely supported, only finitely many Λ and Γ need to be computed before running MCMC.

2.3.2 MCMC approximation

An efficient log-likelihood evaluation makes our model amenable to Metropolis type Markov chain sampling from the posterior distribution. We deal with the issue of finite representation of the function valued parameters $\omega_0, \ldots, \omega_p$, exactly as in Yang and Tokdar (2017). Briefly, a set of knots $T^* = \{\tau_1^*, \ldots, \tau_m^*\} \subset [0, 1]$ are chosen, distinct from the likelihood grid T, and with $m \ll L$. Each function ω_j is represented by its T^* based evaluations: $\omega_j^* = (\omega_j(\tau_1^*), \ldots, \omega_j(\tau_m^*))^\top \in \mathbb{R}^m$. Next, any function evaluation $\omega_j(\tau)$, needed for the likelihood evaluation, is approximated by the associated predictive process interpolation $\tilde{\omega}_j(\tau) = \mathsf{E}[\omega_j(\tau) \mid \omega_j^*]$ (Tokdar, 2007; Banerjee et al., 2008). Due to the Gaussianity of ω_j given (κ_j, λ_j) and the inverse-gamma prior on κ_j^2, the vector ω_j^* is conditionally distributed given λ_j according to a multivariate-t distribution and the conditional mean $\mathsf{E}[\omega_j(\tau) \mid \omega_j^*, \lambda_j]$ is available in closed form. The prior on λ_j is discretized over a dense finite support to write the unconditional prior on ω_j^* as a finite mixture of t distributions, and to evaluate $\tilde{\omega}_j(\tau)$ as a finite, weighted sum of the conditional means.

It may be tempting to employ a Gibbs sampler alternating between updating parameters $(\gamma_0, \gamma, \sigma, \omega, \zeta)$ of the marginal part and the copula parameters $\theta = (\alpha, \phi)$.

However, based on our experience, the scale parameter σ^2 and α can be highly correlated because both parameters affect the tightness of U_i's. Instead, we find the parametrization $(\sigma_s^2, \sigma_e^2) = (\alpha\sigma^2, (1-\alpha)\sigma^2)$ helps reduce autocorrelation between posterior samples and improves inference efficiency[1]. Notice that σ_s^2 and σ_e^2 are spatial variance and pure error variance respectively when our model degenerates to BSRE (5.1.1). With this reparametrization, one complete cycle of our Markov chain sampler could be schematically represented as follows[2]:

1. Given σ_s and ϕ, make a Gibbs update of the parameter $(\gamma_0, \gamma, \sigma_e, \eta, \zeta)$ by using the (adaptive) block Metropolis steps of Yang and Tokdar (2017).

2. Given $(\gamma_0, \gamma, \sigma_e, \eta, \zeta)$ and ϕ, update σ_s.

3. Given the current U_i's extracted from step 2 and σ_s, sample ϕ from its full conditional distribution.

4. Given other model parameters, jointly update σ_s and σ_e.

It is possible to marginalize out ϕ and run the the sampler only on other model parameters. But in our experience, such marginalization does not seem to help with mixing and instead adds to computation complexity. Also, even though σ_e and σ_s get updated in steps 1 and 3, the additional joint update in step 4 greatly improves mixing. Lastly, the latent process realization $W(s)$ is marginalized out during MCMC runs but can be recovered in post-processing for inference.

[1] Effective samples sizes are boosted by 27% and 42%, respectively for $p = 1$ and $p = 7$ in the simulation studies of Section 2.5. Mean absolute errors and coverage probabilities of regression coefficients are also improved.

[2] All function valued parameters are first updated in their finite representations and then used in likelihood evaluation as described.

2.3.3 Reduced rank approximation for large n

The $\mathcal{O}(n^3)$ computational complexity of factorizing a $n \times n$ matrix, and the associated $\mathcal{O}(n^2)$ cost of storage can be prohibitive for implementing the above posterior computation scheme for data with a large sample size n. Various reduced rank approximations to Gaussian processes may be used to alleviate these computation bottlenecks; prominent examples include nearest-neighbor (NN) GP (Datta et al., 2016) and predictive process (Banerjee et al., 2008) with pivoting (Foster et al., 2009). Both models only require $\mathcal{O}(n)$ computation and memory for model fitting and prediction (Section 2.4.1). We implement NNGP using the algorithm described in Finley et al. (2019). On a dataset with $16,000$ samples and 7 predictors, our algorithm under 5-NNGP with 10,000 MCMC iterations took about 55 minutes and the predicted conditional quantiles were statistically accurate.

See Appendix A.2 for a detailed comparison of computational speeds of our model with full GP and NNGP, and, an analysis of prediction accuracy of the model with NNGP.

2.4 Spatial smoothing

2.4.1 Prediction of conditional quantile

Infill prediction is one of the major objectives in spatial data analysis. Although the model specification (1.1.3) together with (2.2.1) and (2.2.2) appears to target estimating the global quantile predictor effects and spatial dependence separately, our model is able to make infill prediction that adequately accounts for spatial information, thus effectively performing spatial quantile smoothing to new locations where only the covariates have been recorded. Specifically, our model allows for a local adjustment of the quantiles of response Y^* at s^* via the conditional copula $C_{s^*|s_{1:n}}(U^* \mid U_1, \ldots, U_n, \theta)$, where $U^* = U(s^*)$. Denoting its quantile function

as $Q_{U^*}(\cdot \mid U_1, \ldots, U_n, \theta)$ and combining the marginal model, the conditional τ^*th quantile of Y^* given X^* is

$$Q_{Y^*}(\tau^* \mid X^*, s^*, U_1, \ldots, U_n) = \beta_0(\tau) + X^{*\top}\beta(\tau),\ \tau = Q_{U^*}(\tau^* \mid U_1, \ldots, U_n, \theta). \tag{2.4.1}$$

When using a Gaussian copula process, the quantile function of the conditional copula can be conveniently calculated as follows. Let K^* be a n-dimensional vector with $K_i^* = \rho_{\mathrm{M}}(s^*, s_i; (\nu, \phi))$. Based on the model specification (2.2.2), we have $Z(s^*) \mid (Z, \theta) \sim \mathsf{N}(\mu(s^*), \sigma^2(s^*))$ where $\mu(s^*) = \alpha K^{*\top}(\alpha K + (1-\alpha)I_n)^{-1}Z$ and $\sigma^2(s^*) = 1 - \alpha^2 K^{*\top}(\alpha K + (1-\alpha)I_n)^{-1}K^*$. Therefore, we obtain $Q_{U^*}(\tau^* \mid U_1, \ldots, U_n, \theta) = \Phi(\mu(s^*) + \sigma(s^*)\Phi^{-1}(\tau^*))$.

Remark 2.4.1. *The quantile smoothing detailed above assumes Y^* is generated from the model* (1.1.3) *with the underlying random level given by $U^* = U(s^*)$. In other words, Y^* is taken to be a yet unobserved unit from the same batch of response realizations as $Y_1, \ldots, Y_n$. If instead, Y^* was thought to arise from a completely fresh draw, then U^* should be treated as independent of the entire process $U(s)$, resulting in $\tau = \tau^*$ in equation* (2.4.1).

2.4.2 Spatial dependence and spatial variation

Limited spatially varying flexibility

The conditional quantile function computed above varies smoothly in s^*, resembling a spatially varying QR model. It is worthwhile to examine whether our model could be reinterpreted as $Q_Y(\tau \mid X, s) = \beta_0(s, \tau) + X^\top \beta(s, \tau)$, with conditionally independent observations $Y_1, \ldots, Y_n$, as done in Reich et al. (2011). The answer is yes, but in a very limited sense. The limitation is useful both theoretically and practically.

Let $V_i = \Phi\left(\varepsilon(s_i)/\sqrt{1-\alpha}\right)$. Given the spatial process realization $W(s)$, the data generating mechanism (1.1.3)-(2.2.2) can be equivalently expressed as

$$Y_i = \beta_0(h_{W,\alpha}(s_i, V_i)) + X_i^\top \beta(h_{W,\alpha}(s_i, V_i)),\ V_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1), 1 \leqslant i \leqslant n \tag{2.4.2}$$

where $h_{w,a}(s,t) = \Phi(w(s) + \sqrt{1-a}\Phi^{-1}(t))$ with $w : \mathcal{S} \to \mathbb{R}$, $s \in \mathcal{S}$ and $(a,t) \in (0,1)^2$. Clearly under model (2.4.2), responses $Y_1, \ldots, Y_n$ are conditionally independent given model parameters, $(W(s) : s \in \mathcal{S})$, and $X_1, \ldots, X_n$. More importantly, this model admits a quantile function in a spatially varying fashion: $\tilde{Q}_Y(\tau \mid X, s) = \beta_0\left(h_{W,\alpha}(s,\tau)\right) + X^\top \beta\left(h_{W,\alpha}(s,\tau)\right)$, because, the map $t \mapsto h_{w,a}(s,t)$ at each $s \in \mathcal{S}$ is a monotonically increasing diffeomorphism of $(0,1)$ onto itself. Notice that this formulation offers only a limited flexibility in capturing spatial variability since the location s and the quantile level τ effect the intercept and slope functions through a single, combined input value $h_{w,a}(s,\tau)$. The model is not able to fully recover the quantile functions if they get crossed with each other.

In contrast, the fully spatially varying QR model of Reich et al. (2011) offers much greater shape flexibility in spatial quantile smoothing. But it is of little use when the primary aim is to learn a global relationship between the response and the predictors, unrelated to spatial location. Additionally, a fully spatially varying QR model could be extremely difficult to estimate. Indeed, the estimation method adopted in Reich et al. (2011) is akin to carrying out a *post hoc* Bayesian smoothing of intercept and slope functions estimated locally via the Koenker-Bassett method. Such an approach could be quite useful for analyzing datasets with many repeated observations at each location, even though the two-stage estimation method is likely to offer uncertainty quantifications that are very difficult to interpret. Furthermore, in a simulation study detailed below, we observed that our model offered excellent spatial quantile smoothing, outperforming Reich et al. (2011) when the spatial variation in regression coefficients was moderate.

Computation strategy

With multiple observations available at each location, we are able to make inference of spatially varying coefficients and quantile functions. Suppose n observations are

available at each of m locations (totally $N = mn$ observations). Let Y_{ij} be the response of jth observation at location s_i with associated predictors X_{ij}. For convenience, we use s_{ij} and s_i interchangeably. Our model can be rewritten as follows.

$$Y_{ij} = \beta_0(U_{ij}) + X_{ij}^\top \beta(U_{ij}),\ U_{ij} = \Phi(Z(s_{ij})),\ Z(s_{ij}) = W(s_{ij}) + \varepsilon(s_{ij})$$

$$W(s) \sim \mathsf{GP}(0, \alpha\rho(s, s'; (\nu, \phi))),\ \varepsilon(s_{ij}) \overset{\text{iid}}{\sim} \mathsf{N}(0, 1-\alpha),\ i = 1, \ldots, m,\ j = 1, \ldots, n$$

Instead of directly applying spectral decomposition on the large $N \times N$ correlation matrix as in Section 2.3.2, we can simplify the computation using Kronecker product. Let K be a $m \times m$ matrix with $K_{kl} = \rho(s_k, s_l; (\nu, \phi))$ and 1_n be a n-dimensional vector with all entries 1. Note that $\text{vec}(Z) \sim \mathsf{N}(0, \alpha K \otimes J_n + (1-\alpha) I_N)$ where $J_n = 1_n 1_n^\top$, $\text{vec}(\cdot)$ is the vectorization function and $\otimes$ denotes Kronecker product. If the spectral decomposition of K is $\Gamma\Lambda\Gamma^\top$, then $K \otimes J_n = \tilde{\Gamma}\tilde{\Lambda}\tilde{\Gamma}^\top$ where $\tilde{\Gamma} = \Gamma \otimes 1_n/\sqrt{n}$ and $\tilde{\Lambda} = n\Lambda$. Let $R = \alpha K \otimes J_n + (1-\alpha) I_N$. Using Sherman-Morrison-Woodbury matrix inverse and Sylvester's determinant identity, we have

$$R^{-1} = \frac{1}{1-\alpha} I_N - \frac{\alpha}{(1-\alpha)^2} \tilde{\Gamma} \left(\tilde{\Lambda}^{-1} + \frac{\alpha}{1-\alpha} I_m \right)^{-1} \tilde{\Gamma}^\top$$

$$\det(R) = (1-\alpha)^N \det\left(I_m + \frac{\alpha}{1-\alpha} \tilde{\Lambda} \right).$$

By further noting that $\tilde{\Gamma}^\top \text{vec}(Z) = \text{vec}(1_n^\top Z \Gamma)/\sqrt{n}$, we substantially reduce the computation complexity for calculating copula density.

Illustration via simulated examples

To thoroughly compare JSQR against JQR and the approximate spatial quantile regression (ASQR) method proposed by Reich et al. (2011), we present two simulation studies with moderate and large spatial variations of regression coefficients respectively. In each simulation, we generated 80 replicates at each of 20 locations for each of the 100 synthetic datasets. For each replicate, we adopted the setup used

in Reich et al. (2011) to generate predictors, locations and quantiles.

$$X_i \overset{\text{iid}}{\sim} \mathsf{Unif}([0,1]^2),\ s_i \overset{\text{iid}}{\sim} \mathsf{Unif}([0,1]^2),\ U_i = \Phi(Z(s_i)),\ Z(s) \sim \mathsf{GP}(0, \rho_{\text{SE}}(s, s'; \sqrt{2}))$$

The conditional quantile functions for simulation studies were

(i). $Q_{Y_i}(\tau \mid X_i, s_i) = 2s_{i2} + (\tau+1)\Phi^{-1}(\tau) + 5s_{i1}\tau^2 X_{i2}$

(ii). $Q_{Y_i}(\tau \mid X_i, s_i) = 2s_{i2} + (\tau+1)\Phi^{-1}(\tau) + 5s_{i1}\tau^2 X_{i2} + 120(s_{i1}-0.5)(X_{i2}-0.5)$

The first example was taken from Reich et al. (2011) where the true coefficients were $\beta_0(\tau, s) = 2s_{i2} + (\tau+1)\Phi^{-1}(\tau)$, $\beta_1(\tau, s) = 0$ and $\beta_2(\tau, s) = 5s_{i1}\tau^2$. The quantile function is monotone increasing in X_{i2} with any fixed (τ, s). We added a term $120(s_{i1}-0.5)(X_{i2}-0.5)$ into the quantile function in the second example. As shown in Figure 2.1, this added term greatly increases the variation of coefficients and quantile functions across locations.

Instead of conditioning on data, ASQR adopts "approximate posterior" conditioning on the classical KB estimates and its asymptotic covariance (Koenker and Bassett, 1978). In our simulations, estimating the covariance at $\tau = 0.95$ requires at least 79 observations at each location. We also include a spatially varying version of JQR using B-splines (JQR-BS) of locations. Specifically, we adopt tensor product spline surfaces to approximate spatially varying coefficients

$$\hat{\beta}_k(\tau, (s_1, s_2)) = \sum_{q=1}^{d}\sum_{r=1}^{d} \beta_{kqr}(\tau)\psi_q(s_1)\psi_r(s_2), \quad k = 0, \ldots, p$$

where $\{\psi_r(\cdot) : r = 1, \ldots d\}$ are B-spline basis functions of order d. In our simulations, we augment the original predictors (including intercept) by including their interactions with $\{\psi_q(s_1)\psi_r(s_2) : q = 1, \ldots, d,\ r = 1, \ldots, d\}$ resulting in $(d^2+1)(p+1)$ predictors in total.

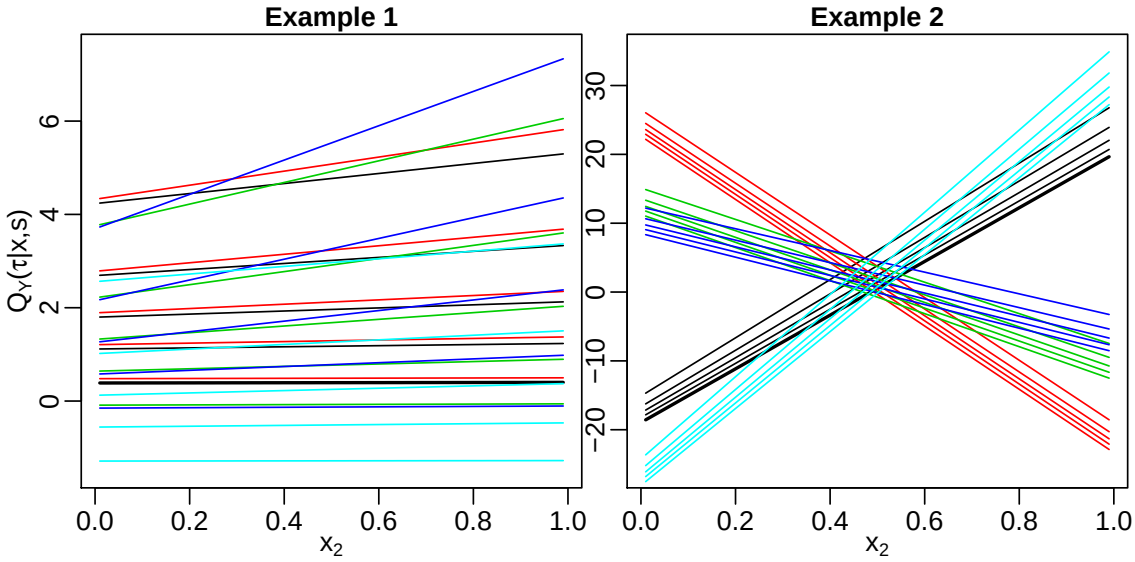

FIGURE 2.1: Quantile functions $Q_Y(\tau \mid X, s)$ against X_2 at 5 randomly simulated locations with different colors at $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

All settings for JSQR, JQR were kept same as in Section 2.5. We used B-splines with degree $d = 2$ for JQR-BS. ASQR was implemented using the code available at the author's homepage[3]. JSQR and JQR took 2.7 and 2 minutes respectively for analyzing one dataset on average while JQR-BS and ASQR took 20 and 23.2 minutes.

We compared different methods based on mean absolute errors of point estimates of regression coefficients and quantile functions. Namely, for each of 100 synthetic datasets, we calculated

$$\frac{1}{m}\sum_{i=1}^{m} |\beta_k(\tau, s_i) - \hat{\beta}_k(\tau, s_i)|, \ k = 0, \ldots, p$$

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} |Q_{Y_{ij}}(\tau \mid X_{ij}, s_i) - \hat{Q}_{Y_{ij}}(\tau \mid X_{ij}, s_i)|$$

at quantile levels $\tau \in \{0.01, 0.05, 0.1, ..., 0.9, 0.95, 0.99\}$ and then compute the mean over all datasets. All estimates are given by posterior means. The simulation results

[3] `https://www4.stat.ncsu.edu/~bjreich/code/`

are summarized in Figure 2.2. In the first example, as shown in the left panel of Figure 2.1, although the quantile functions from different locations get crossed, their slopes are similar. JSQR dominates all other competitors for estimating both coefficients and conditional quantiles. JQR provides the worst estimate for conditional quantiles while ASQR poorly estimates the coefficients. JQR-BS indeed improves upon JQR according to MAE of conditional quantiles. Nevertheless, in the second example, the performances of JSQR, JQR and JQR-BS significantly drop while the errors for ASQR are similar to those in the first example. The failure of JSQR and JQR in this example is not surprising due to the large spatial variation in regression coefficients. JQR-BS significantly reduces errors by using splines of locations with degree 2. We expect that JQR-BS has a better performance by using higher order splines. However, it also increases computational complexity.

We did not include JSQR with augmented predictors using B-splines into comparison because its estimation accuracy was even worse than JSQR itself. The poor performance may be due to drastic overparameterization. Further investigation is needed to appropriately incorporate spatially varying coefficient into JSQR model.

We note that the example 2 is used to distinguish the behaviors of our method and ASQR. In most of real data applications, we expect that the conditional distributions of the response given predictors at different locations have smaller spatial variation as in the example 1 where our method is applicable.

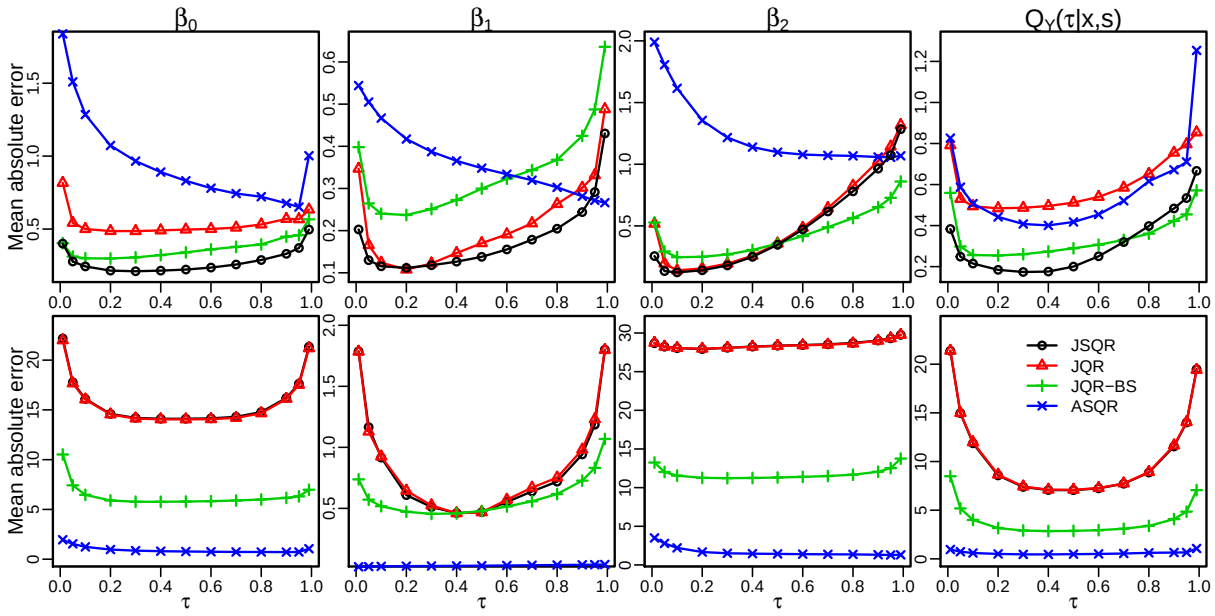

FIGURE 2.2: Mean absolute errors of regression coefficients and quantile functions at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$. The results for example 1 and 2 are in the top and bottom row respectively. JSQR: proposed joint spatial quantile regression; JQR: joint quantile regression by Yang and Tokdar (2017); JQR-BS: spatilly varying version of JQR using B-splines of locations; ASQR: approximate spatial quantile regression method by Reich et al. (2011).

2.5 Numerical experiments

We carried out several simulation studies to compare the proposed model against existing methods on inference efficiency and prediction accuracy, and to examine the adaptability of our model to different correlation strengths. In Section 2.5.1, we demonstrate that in quantile modeling for spatial data, neglecting or inappropriately incorporating spatial dependence results in suboptimal statistical performance by including competing methods that are designed for independent data or offer limited adjustment for spatial dependence. To be comprehensive on simulation design, two examples are included: one with a univariate predictor ($p = 1$) and one with multivariate predictors ($p = 7$). To investigate robustness of our model against model misspecification, two dependency patterns, one of which deviates from the Gaussian

copula process assumption, are considered for each example. In Section 2.5.2, we illustrate that our model is capable of adapting to different levels of dependence and recovering the true underlying correlation structure by assessing estimation accuracy of copula parameters and induced correlations.

2.5.1 Statistical performance assessment

Simulation setup

Spatially dependent data sets were generated on $\mathcal{S} = [0,1]^2$ by extending the simulation setting of Yang and Tokdar (2017). Spatial locations $s_1, \ldots, s_n$ were sampled uniformly from $\mathcal{S}$. A 2×2 design with 100 replications each was used with two choices of the marginal QR model and 2 choices of the spatial copula process:

M1. Simple regression with $p = 1$:

$$Q_{Y_i}(\tau \mid X_i) = 3(\tau - 0.5)\log\tfrac{1}{\tau(1-\tau)} + 4(\tau - 0.5)^2 \log\tfrac{1}{\tau(1-\tau)} X_i,\ X_i \overset{\text{iid}}{\sim} \mathsf{Unif}(-1, 1).$$

M2. Multiple regression with $p = 7$:

$$Q_{Y_i}(\tau \mid X_i) = \beta_0(\tau) + X_i^\top \beta(\tau),\ X_i \overset{\text{iid}}{\sim} \mathsf{Unif}\left(\{x \in \mathbb{R}^7 : \|x\| \leqslant 1\}\right)$$

with

$$\beta_0(0.5) = 0,\ \beta(0.5) = (0.96, -0.38, 0.05, -0.22, -0.80, -0.80, -5.97)^\top,$$

$$\dot\beta_0(\tau) = \tfrac{1}{\phi(\Phi^{-1}(\tau))},\ \dot\beta(\tau) = \tfrac{\dot\beta_0(\tau)\nu(\tau)}{\sqrt{1+\|\nu(\tau)\|^2}},\ \nu_j(\tau) = \sum_{l=1}^{3} a_{lj}\phi(\tau; \tfrac{l-1}{2}, \tfrac{1}{9}),\ 1 \leq j \leq 7,$$

$$a = \begin{pmatrix} 0 & 0 & -3 & -2 & 0 & 5 & -1 \\ -3 & 0 & 0 & 2 & 4 & 1 & 0 \\ 0 & -2 & 2 & 2 & -4 & 0 & 0 \end{pmatrix}$$

where $\phi(\cdot)$ denotes the pdf of $\mathsf{N}(0, 1)$.

C1. Asymmetric Laplace copula process (Lum and Gelfand, 2012):

$$U_i = F_{\text{AL}}\left(W(s_i); \tau\right),\ W(s_i) = \sqrt{\tfrac{2\xi_i}{\tau(1-\tau)}} Z(s_i) + \tfrac{1-2\tau}{\tau(1-\tau)}\xi_i,\ \xi_i \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$$

$$Z(s) \sim \mathsf{GP}(0, k(s, s')),\ k(s, s') = \alpha\rho_{\text{M}}(s, s'; (\nu, \phi)) + (1 - \alpha)\mathbb{1}(s = s')$$

where $F_{\mathrm{AL}}(\cdot;\tau)$ is the CDF of asymmetric Laplace distribution $\mathsf{AL}(\tau)$ with pdf $f_{\mathrm{AL}}(x;\tau) = \tau(1-\tau)\exp\{-\rho_\tau(x)\}$.

C2. Gaussian copula process:

$$U_i = \Phi(Z(s_i)),\ Z(s) \sim \mathsf{GP}(0, k(s,s'))$$
$$k(s,s') = \alpha\rho_{\mathrm{M}}(s,s';(\nu,\phi)) + (1-\alpha)\mathbb{1}(s=s')$$

The two marginal QR models are the same as in Yang and Tokdar (2017). For either copula process, we set $(\alpha,\nu,\phi) = (0.7, 2, 0.3)$. Additionally, for C1, we fixed $\tau = 0.4$. The resulting asymmetry in the true copula process was used as an opportunity to assess the robustness of symmetric Gaussian copula based JSQR under model misspecification. For M1, each synthetic data set consisted of a training set with $n = 200$ observations, and for M2, each set consisted of $n = 500$ observations. For each set, a test set containing 50 observations was used for hold-out validations.

The proposed joint spatial quantile regression model (JSQR[4]) was compared against four alternatives; KB: the classical optimization based method by Koenker and Bassett (1978), JQR: the joint quantile regression model by Yang and Tokdar (2017), ALD: Bayesian quantile regression model using asymmetric Laplace error distribution by Yu and Moyeed (2001), and, ALP: spatial quantile regression model using asymmetric Laplace process by Lum and Gelfand (2012). Among these methods, KB, JQR and ALD do not incorporate spatial dependence.

These competing methods were compared based on mean absolute errors (MAE) of point estimates of regression coefficients and corresponding coverage probabilities of 95% confidence (or credible) intervals. We also compared the MAE of predicted conditional quantiles $|Q_Y(\tau \mid X, s, U_1, \ldots, U_n) - \hat{Q}_Y(\tau \mid X, s, U_1, \ldots, U_n)|$ averaged over all observations in the test set. Here $\hat{Q}_Y(\tau \mid X, s, U_1, \ldots, U_n)$ is the predicted

[4] Without special notice, we use JSQR to denote our joint spatial quantile regression model (2.2.1) with a standard logistic density f_0 and the Gaussian copula process (2.2.2).

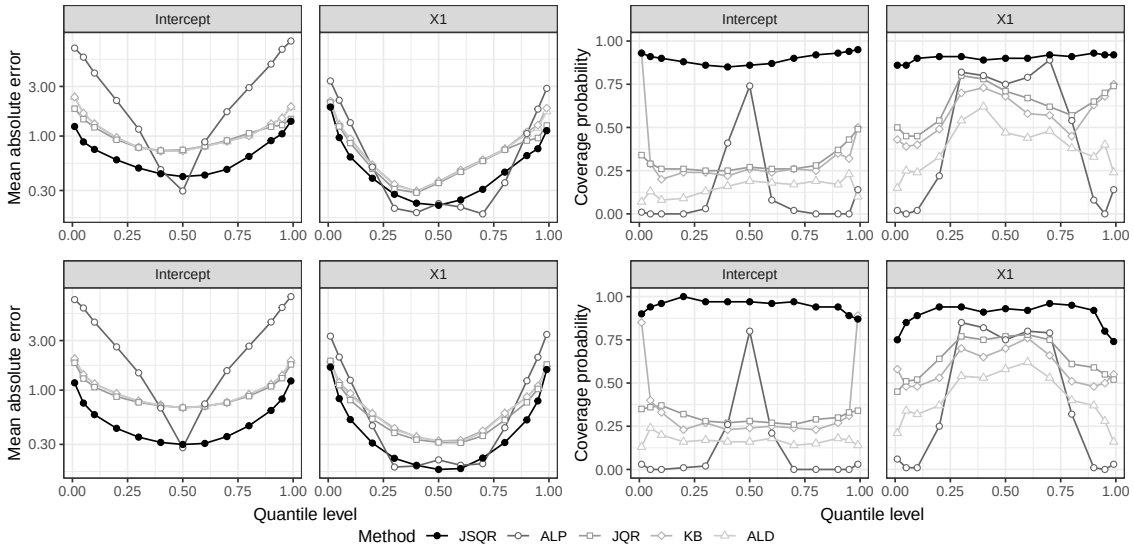

FIGURE 2.3: Inference efficiency of different methods for M1×C1 (top) and M1×C2 (bottom). In each row, the left two panels present the mean absolute errors of regression coefficients and the right two panels show the coverage probabilities of 95% confidence (or credible) intervals of regression coefficients at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

conditional quantile function given by different methods. For KB, JQR and ALD, this function is precisely $\hat{Q}_Y(\tau \mid X)$. For ALP, we adopted the spatially adjusted quantile function by conditioning on the latent process. All these comparisons were performed at quantile levels $\{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$ and averaged over all 100 data sets under each scenario. For Bayesian methods, posterior means were used as point estimates. The confidence intervals for KB were constructed by inverting a rank test proposed by Koenker (1994). Algorithm configurations are detailed in Appendix A.1. Results from M1 are presented in Figure 2.3 and those from M2×C1 are shown in 4.1. Results from M2×C2, which are very similar to M2×C1 results, are included in Appendix A.3, Figure A.3.

Key findings

In every design, JSQR provided smallest averaged errors and highest (and approximately nominal) coverage probabilities for estimating the intercept and slope functions across all quantile levels. The three non-spatial methods, KB, JQR and ALD gave distinctly worse performance, particularly on coverage probability. KB and JQR were generally similar to one another, with the latter offering better performances for certain covariates. By design, KB and ALD gave same parameter estimates, but the latter clearly had lower coverage probabilities; this issue has been raised before in Tokdar and Kadane (2012).

The estimates by ALP were significantly worse than the rest of the group for quantile levels away from 0.5, both in terms of accuracy and coverage. ALP severely distorted the estimates of the intercept function and consequently offered poor estimates of slope functions at non-central quantile levels. Since ALP assumes an asymmetric Laplace process on errors instead of on quantile levels, the model is misspecified at any given quantile level, even though data in C1 were simulated under an asymmetric Laplace copula process.

A similar picture emerged on quantile kriging accuracy at hold-out data (Figure 4.2). JSQR offered substantial improvements across all quantile levels compared to the methods that did not adjust for spatial dependence. ALP came a distant second for central quantile levels, but actually performed worse at the extremities.

Together, this empirical evidence underlines the usefulness of a proper model based spatial QR estimation for improved parameter estimation and quantile kriging. It is noteworthy that ALP, despite its limitations, improves estimation and kriging accuracy, thanks to spatial dependence adjustment, at quantile levels for which the assumed residual process is not too different from the true data generating distribution. But its limitations show up dramatically for non-central, and partic-

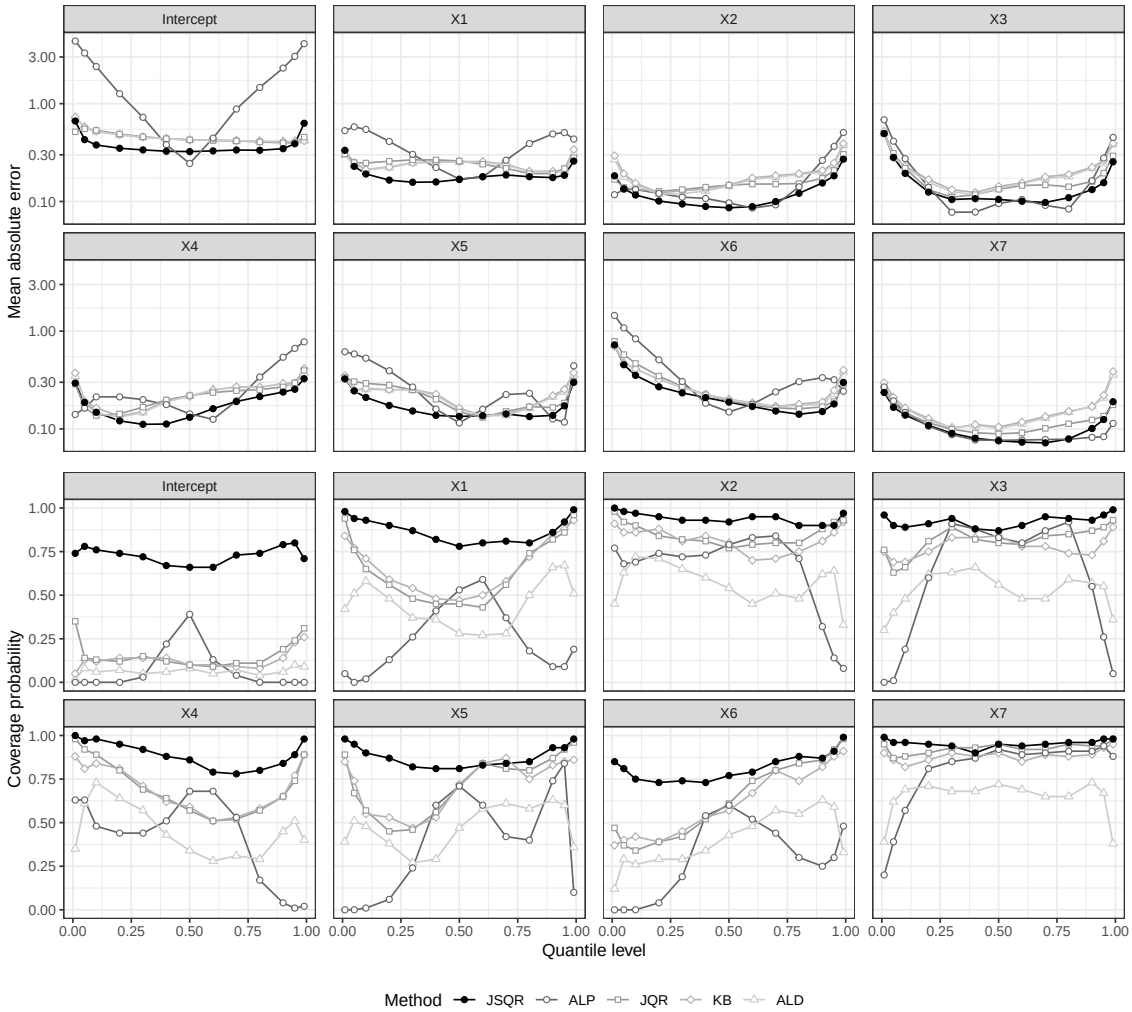

FIGURE 2.4: Inference efficiency of different methods for Example 2 with true generating process being asymmetric Laplace process. The top two rows present the mean absolute errors of regression coefficients while the bottom two rows show the coverage probabilities of 95% confidence (or credible) intervals of regression coefficients at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

ularly, extreme quantile levels. The power and flexibility of the joint QR model is evident in JSQR's superior estimation quality across all quantile levels. Moreover, JSQR appears reasonably robust against misspecification of the copula process.

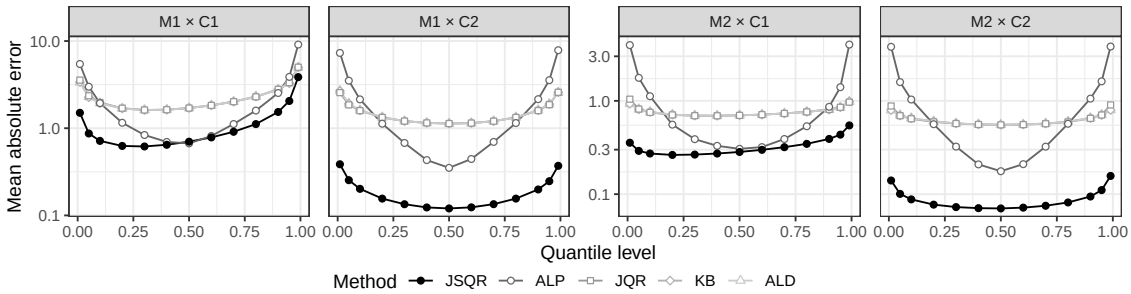

FIGURE 2.5: Mean absolute errors of conditional quantile function at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

2.5.2 Adaptation to dependence strength

In a second set of experiments, we examined the adaptability of JSQR to different strengths of spatial dependence. Specifically, 100 data sets, each with $n = 500$, were generated as in M1×C2, but for each set a distinct and random draw was made for the parameter values of (α, ϕ). The proportion parameter α was drawn from $\mathsf{Unif}(0, 1)$ and the decay parameter ϕ was randomly drawn from its prior range. The JSQR estimation method was employed exactly as in Section 2.5.1. For each dataset, we estimated α, ϕ and 10 pairwise correlations $r_{ij} = \alpha\rho(s_i, s_j; (\nu, \phi))$ between 5 randomly selected observed locations, by their respective posterior means. Mean absolute errors of estimates over 100 simulated data sets are reported in Table 2.1, which also includes the lengths and the coverage probabilities of the respective 95% credible intervals.

We note that accurately estimating induced pairwise correlations r_{ij} is potentially more important than accurately estimating the copula parameters α and ϕ. This is because, the induced correlations directly determine the dependence between observations and hence are more critical in adjusting for the spatial dependence toward a more accurate estimation of the marginal QR intercept and slope parameters. Induced correlations are also more identifiable than copula parameters because different combinations of α and ϕ may yield similar values of r_{ij}. This argument is

	Absolute error	CP of 95% CI	Length of 95% CI
α	$0.05_{0.04}$	0.96	$0.25_{0.11}$
ϕ	$0.04_{0.03}$	0.95	$0.17_{0.06}$
r_{ij}	$0.03_{0.04}$	0.954	$0.14_{0.10}$

Table 2.1: Left column: mean absolute errors of posterior means of α and ϕ and induced pairwise correlations $r_{ij} = \alpha\rho(s_i, s_j; (\nu, \phi))$. Middle and right columns: coverage probabilities and average lengths of 95% credible intervals of α, ϕ and r_{ij}. Standard deviations are shown as subscripts.

verified by the results presented in Table 2.1 which show that our model accurately estimated r_{ij}. The 95% credible interval was narrow and its coverage was close to nominal level. Both absolute errors and length of 95% credible intervals were smaller than those of α and ϕ. As the prior range for ϕ was around $(0.1, 0.4)$, both error of point estimate and uncertainty of ϕ were relatively greater than those of α and r_{ij}. This is expected as decay parameter is weakly identified. The results suggest that our model is capable of adapting to different levels of dependence and recovering the underlying correlation structure.

3

Joint Hierarchical Quantile Regression

3.1 Introduction

Clustered data are widely collected and analyzed in many disciplines of scientific research (Diggle et al., 2002; Gelman and Hill, 2006). Such data typically presents a multilevel structure where individual observation units are grouped into clusters. Within a cluster, the observations share common characteristics which may not be measured and hence are potentially dependent. When analyzing data with hierarchical dependence, neglecting the dependency structure may result in suboptimal statistical inference including biased estimation, inaccurate prediction and overconfident uncertainty quantification. Moreover, the within-cluster dependency patterns may be different for different applications. For example, in each survey year of the well-known High School and Beyond study[1], the participating students from each school do not have any particular order and hence an exchangeable structure is appropriate to describe the within-cluster dependency pattern. In another example of the longitudinal study of $CD4^+$ depletion of HIV patients (Zeger and Diggle, 1994),

[1] `https://nces.ed.gov/surveys/hsb/`

the repeated measurements are recorded for each patient to monitor the cell counts over a long time course. For each patient, the records may present stronger dependence when they are collected within a shorter time interval.

Valid and efficient statistical inference relies crucially on properly adjusting for the dependency structure in data. In the context of quantile regression, several methods have been developed for cross-sectional data and longitudinal data. With the blocking strategy, Tang and Leng (2011) and Wang and Zhu (2011) treat all observations within a cluster as a whole unit and then adopt empirical likelihood to account for within-cluster dependence. These methods focus on estimating the predictor effects and cannot encode any specific dependency structure into the model. Reich et al. (2010) consider an alternative approach and propose using a Bayesian heteroskedastic linear model with an infinite mixture error term and random effects to model the dependence. However, the model capacity is limited and the inference and prediction with this model require intensive computation. A recent work by Wang et al. (2019) proposes incorporating the dependence using a copula and design a three-stage optimization based inference method. Nevertheless, since the algorithm is numerically unstable, the inference results are unreliable. The statistical performances of these methods are more comprehensively evaluated and reported in Section 3.5 and 5.2.

We propose a new joint quantile regression framework for hierarchical data by extending the model of YT17 with three prevalent dependency structures: exchangeable dependence for cross-sectional data and temporal dependencies for longitudinal data with regularly or continuously spaced timestamps. Our model characterizes population-level quantile predictor effects while adjusts for cluster-specific dependency patterns via copula. A semiparametric Bayesian inference scheme is proposed to encourage information sharing across clusters and to facilitate parameter estimation. Through extensive simulation studies (Section 3.5) and real applications

(Section 5.2), we demonstrate the superior statistical performance of the proposed model compared to existing alternatives.

We advocate adopting different copula models to adjust for potential tail dependence and/or tail asymmetry (Section 4.2). To select between model specifications, two versions of WAIC scores are introduced to compare models according to within-cluster prediction and out-of-cluster prediction accuracy. The effectiveness of the proposed model selection criterion is verified with both empirical and theoretical evidence. We highlight the model interpretation through two case studies (Section 5.2). Particularly, insightful discoveries of heterogeneous response-predictor relationships across quantile levels are provided in the High School and Beyond data and $CD4^+$ data analyses.

We conclude the introduction with a note on the connection between our model and the so-called conditional QR models (Geraci and Bottai, 2007; Kim and Yang, 2011; Geraci, 2014). This track of research adopts the viewpoint of conditional quantile regression given cluster-specific effects in the same spirit of linear mixed effects model. Nevertheless, the aforementioned approaches all postulate parametric components in either likelihood or random effects, which restricts their modeling flexibility. More importantly, as also noted in Reich et al. (2010), the inferred global regression coefficients cannot be interpreted as the population quantile predictor effects, because unlike mean effects in classical linear model, the quantile effects are not additive. Therefore, these methods become unsuitable when estimating global effects is the research objective. Compared to these conditional QR methods, however, our model intrinsically features population-level quantile predictor effects while adjusting for clustered noise dependency. Meanwhile, the estimated global predictor effects admit a cluster-varying interpretation because the estimated quantile levels of observations are clustered and may be concentrated at different regions (elaborated through case studies in Section 5.2). Furthermore, our model can be straightfor-

wardly extended for analyzing cluster-varying effects by augmenting the predictor set with their interactions with cluster indicators.

3.2 Joint hierarchical quantile regression

3.2.1 Modeling within-cluster dependence

We focus on analyzing data with a hierarchical structure where Y_{ij} and X_{ij} are the response and p-dimensional predictor of the jth observation in the ith cluster (subject), for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Denote the overall sample size as $N = \sum_{i=1}^{n} n_i$. The sample sizes n_i's across clusters can be identical or distinct corresponding to balanced or unbalanced scenarios. For such clustered data, it is natural to assume that the observations from different clusters are independent but responses within the same cluster may be dependent. Therefore, by Sklar's theorem (Sklar, 1959), the joint QR model theorizes $Y_{ij} = \beta_0(U_{ij}) + X_{ij}^\top \beta(U_{ij})$ where $(U_{i1}, \ldots, U_{in_i})$ follows a copula distribution.

Towards an interpretable and practicable statistical analysis of dependence, it is pragmatic to consider a parametric copula family indexed by a low dimensional global-local parameter pair $(\theta, \boldsymbol{\phi} = \{\phi_i\}_{i=1}^n)$. From the perspective of modeling, the global parameter θ captures the nature of within-cluster dependency shared by all clusters while the local parameters $\boldsymbol{\phi}$ account for cluster-specific characteristics of the dependency. From the perspective of estimation, this hierarchical specification guarantees the identifiability of copula parameters and facilitate the computation. Coupled with the marginal model, our joint hierarchical QR model (JHQR) is given by

$$Y_{ij} = \beta_0(U_{ij}) + X_{ij}^\top \beta(U_{ij}), \quad (U_{i1}, \ldots, U_{in_i}) \sim C_i(u_1, \ldots, u_{n_i} \mid \theta, \phi_i) \tag{3.2.1}$$

The general and versatile model formulation (3.2.1) shall be further materialized in the following subsections to adjust for different types of hierarchical dependency, by varying the ways the copula are constructed. Throughout this section, to

highlight the model structure, we narrow our focus on the Gaussian copula family mainly due to its simplicity and its close connections to classical models (detailed below). Moreover, the use of Gaussian copula also leads to a fair number of inferential and computational advantages such as the ease of parameter estimation, analytical quantile prediction, and, scalability with sample size. We shall elaborate on these presently in Section 3.3 and Section 3.4. Other choices of copula family are thoroughly investigated, to adjust for potential tail asymmetry or tail dependence, in Section 4.2.

JHQR for cross-sectional data

For modeling cross-sectional data where replications per cluster do not have a useful ordering or relational adjacencies, it is natural to assume that the observations within a cluster are exchangeable. Considering this assumption, we adopt the Gaussian copula with a permutation symmetry correlation structure. Let $\Phi(\cdot)$ denote the cumulative distribution function of $\mathsf{N}(0,1)$. We define a hierarchical copula model as

$$\text{M1:} \qquad U_{ij} = \Phi(Z_{ij}),\ Z_{ij} = W_i + \varepsilon_{ij},\ \ W_i \overset{\text{ind}}{\sim} \mathsf{N}(0,\phi_i),\ \varepsilon_{ij} \overset{\text{ind}}{\sim} \mathsf{N}(0,1-\phi_i)$$

This model specification decomposes the variation in the underlying quantiles into two parts where W_i captures the cluster-specific level while ε_{ij} is uncorrelated noise. The cluster-specific pairwise correlation between latent quantiles is governed by $\phi_i \in [0,1]$. When $\phi_i = 0$, all observations within ith cluster are independent and the JHQR degenerates to the independent model when this condition holds for all clusters. When $\phi_i = 1$, observations in the ith cluster present an extreme dependency pattern where they share the same quantile level. Moreover, when $\beta_0(u) = \sigma\Phi^{-1}(u) + \tilde{\beta}_0$, $\beta(u) = \tilde{\beta}$ and $\phi_i = \phi$ for all $i \in \{1,\ldots,n\}$, M1 becomes $Y_{ij} = \tilde{\beta}_0 + X_{ij}^\top\tilde{\beta} + \tilde{W}_i + \tilde{\varepsilon}_{ij}$ with $\tilde{W}_i \overset{\text{iid}}{\sim} \mathsf{N}(0,\phi\sigma^2)$ and $\tilde{\varepsilon}_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0,(1-\phi)\sigma^2)$. This special model formulation coincides with the classical random intercepts model.

JHQR for longitudinal data

When the paired predictor-response observation is collected at a known timestamp t_{ij}, we advocate incorporating the timestamp information into the JHQR through the marginal model and/or the copula model, depending on the nature of the association between the response and time. When the response exhibits a time trend, timestamp and/or its nonlinear transformations may be included in the marginal model as predictors (see Section 5.2.1 for a real example). If the timestamps are recorded only at a few time points, one may consider including them as dummy predictors. For modeling autocorrelation among observations over time, it is natural and useful to view the random quantile levels as realizations of a temporal process. Specifically, let $U_{ij} = U(t_{ij})$ where $(U(t_{i\cdot}) : t_{i\cdot} \in \mathcal{T} \subset \mathbb{R})$ is a stationary stochastic process on $\mathcal{T}$ for every $i \in \{1, \ldots, n\}$ such that $U(t_{i\cdot}) \sim \mathsf{Unif}(0,1)$ at every $t_{i\cdot} \in \mathcal{T}$. Such an embedding within a temporal copula process is essential to formalize forecasting at future time points.

If timestamps are on a regular time grid, we propose modeling the temporal dependence using a discrete time Markov copula process. Due to the Markovian property, the joint copula can be factored into the product of a series of bivariate copula. Particularly, we consider the first order autoregressive structure

$$\text{M2:} \qquad U(t_{ij}) = \Phi(Z(t_{ij})),\ Z(t_{ij}) = \phi_i Z(t_{i,j-1}) + \sqrt{1-\phi_i^2}\varepsilon(t_{ij}),$$

$$Z(t_{i1}) \overset{\text{iid}}{\sim} \mathsf{N}(0,1),\ \varepsilon(t_{ij}) \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$$

where $\phi_i \in [0,1]$ describes the cluster-specific, positive lag-1 autocorrelation between quantile levels.[2] When $\beta_0(u) = \sigma^2\Phi^{-1}(u) + \mu$ and $\beta(u) \equiv 0$, our model fully recovers the blocked version of the stationary Gaussian AR(1) model: $Y_{ij} = \phi_i Y_{i,j-1} + (1-\phi_i)\mu + \sigma^2\sqrt{1-\phi_i^2}\varepsilon_{ij}$ where $Y_{i1} \sim \mathsf{N}(\mu, \sigma^2)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$.

[2] If it is reasonable to assume negative autocorrelations, one can extend the support of ϕ_i to be $[-1,1]$.

Higher order AR processes are not pursued here due to the sophisticated constraint on AR coefficients imposed by the stationarity condition. Coupled with marginal quantile regression model, however, the seemingly simplified first order Markovian dependency is capable of inducing time series with different kinds of extreme and dependency patterns; see Section 4.2.3 for more details.

Remark 3.2.1. *The idea of quantile autoregression has been explored by Koenker and Xiao (2006) and Chen et al. (2009), among others. The former models the temporal dependency by including lagged responses as predictors in the linear quantile regression equation while the latter employs parametric marginal and bivariate copula models to incorporate more flexible dependencies.* JHQR *can be considered as a combination of the two because our model adopts a linear QR margin of the former with bivariate copula of the latter. Therefore, when responses at preceding time points are included in the marginal model,* JHQR *is able to model both lagged dependency in response and persistency in noise, which echoes the core characteristic of the autoregressive moving average (ARMA) model. Particularly, M2 includes Koenker and Xiao (2006) as a special case when data only contains a single cluster with* $\phi_1 = 0$.

When timestamps are unequally spaced, we adopt a continuous time stochastic process to capture the temporal dependence. The stationarity condition requires that the correlation between any $U(t_{i\cdot})$ and $U(t'_{i\cdot})$ depend only on their temporal distance $|t_{i\cdot} - t'_{i\cdot}|$. As this distance increases, the correlation must diminish to zero. Given these desiderata, we define

M3:
$$U(t_{ij}) = \Phi(Z(t_{ij})),\ Z(t_{ij}) = W(t_{ij}) + \varepsilon(t_{ij}),$$
$$W(t_{i\cdot}) \sim \mathsf{GP}\left(0, \phi_i \rho_{\mathrm{M}}(t_{i\cdot}, t'_{i\cdot}; (\nu, \ell))\right),\ \varepsilon(t_{ij}) \overset{\text{ind}}{\sim} \mathsf{N}(0, 1 - \phi_i)$$

where

$$\rho_{\mathrm{M}}(t_{i\cdot}, t'_{i\cdot}; (\nu, \ell)) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|t_{i\cdot} - t'_{i\cdot}|}{\ell} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{|t_{i\cdot} - t'_{i\cdot}|}{\ell} \right)$$

is the Matérn correlation function. Here $\Gamma(\cdot)$ is the Gamma function and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind of order ν. Our model can be trivially extended to other choices of correlation functions. The global parameter pair (ν, ℓ) controls the smoothness and decay rate of the process shared across clusters. Notably, when $\nu = 1/2$, the resulting Ornstein-Uhlenbeck process (Uhlenbeck and Ornstein, 1930) is a Markov process. As ν is typically weakly identifiable from data with a value less than 3 considered sufficient for real data analysis (Banerjee et al., 2014), we hence fix ν in the model with a user-specified value. The local parameter ϕ_i determines the cluster-varying proportion of variation in quantiles that is temporally structured.

3.2.2 Prior specification for Bayesian estimation

Prior for marginal parameters

A semiparametric Bayesian approach is adopted for making inference on parameters in the proposed JHQR model, which comprises a marginal component and a copula component. The model parameters from the marginal component are first reparametrized using the strategy in Section 1.2. We then follow YT17 and adopt the priors below on the new model parameters:

$$\omega_j \sim \mathsf{GP}(0, \kappa_j^2 \rho_{\mathrm{SE}}(\cdot, \cdot; \lambda_j)) \quad j = 0, \ldots, p$$

$$\kappa_j^2 \overset{\text{iid}}{\sim} \mathsf{Inv\text{-}Ga}(0.1, 0.1) \quad \lambda_j \sim \pi_{\lambda}(\lambda_j)$$

$$(\gamma_0, \gamma, \sigma^2) \sim \pi(\gamma_0, \gamma, \sigma^2) \propto \frac{1}{\sigma^2}$$

where $\rho_{\mathrm{SE}}(s, s'; \lambda) = \exp\left(-\lambda^2 \|s - s'\|\right)$ is the square exponential correlation function with rescaling parameter λ. A hyper-prior on λ is implicitly specified by choosing

$r_j := \mathsf{Cor}(\omega_j(\tau), \omega_j(\tau + 0.1) \mid \lambda_j) = \exp(-0.1^2\lambda_j^2)$ to be $\mathsf{Be}(6, 4)$ distributed. See Yang and Tokdar (2017) for more details.

Prior for copula parameters

All three variations of the JHQR model contain cluster-specific parameters $\{\phi_i\}_{i=1}^n$. To allow for information borrowing across clusters, we assume that $\{\phi_i\}_{i=1}^n$ are independent and identically distributed according to a shrinkage prior $\mathsf{Beta}(\mu, \psi)$. The prior is parameterized in terms of the mean $\mu = \alpha/(\alpha + \beta)$ and "sample size" $\psi = \alpha + \beta$, where α and β are shape parameters in the classical parametrization of Beta distribution. This meaningful parametrization facilitates the hyperprior specification. When no prior knowledge is available on the average or variation of $\{\phi_i\}_{i=1}^n$, we recommend adopting generic, uninformative priors to allow for data-driven inference. Particularly, we specify $\mu \sim \mathsf{Un}(0, 1)$ and $\psi \sim \mathsf{Ex}(1)$. As their natural extensions, Beta and Gamma distributions may be justified as well if some clues of (μ, ψ) are accessible.

To specify a prior for the decay parameter ℓ in the temporal copula process, we take the perspective of *effective range* from the spatial smoothing literature (Gelfand et al., 2003; Banerjee et al., 2014) described in Section 2.2.2. The same reasoning also applies to temporal modeling, with one key difference in prediction. In time-series analysis, forecasting at future time points is often of interest. Such prediction, in contrast to the in-fill kriging in spatial modeling, can be deemed as *extrapolation*. It is hence reasonable to extend the support of the prior to allow for adequate temporal correlation when carrying out forecasting. We specify a default uniform prior with a support of which the lower and upper limits are set such that the effective range is between 1/8 and 7/8 of the maximal within-cluster time span. We further discretize the prior over a dense grid of values within the specified range. Our choice of prior has several additional advantages. From a modeling perspective, it keeps ℓ away from

0 and avoids the identifiability issue that exhibits when all ϕ_i's approach 0. From a computational perspective, copula density evaluation (see Section 3.3 for details) requires inverting within-cluster correlation matrices in each MCMC iterations. Our discrete uniform prior greatly reduces the computational burden by computing and storing finitely many correlation matrices before initiating MCMC computation.

3.3 Posterior Computation

3.3.1 Likelihood evaluation

For hierarchical data, the likelihood can be factored by clusters. Within each cluster, the joint conditional density of the response given predictors can be partitioned into a marginal part and a copula part, due to Sklar's theorem

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{x}_i)$$

$$p(\boldsymbol{y}_i \mid \boldsymbol{x}_i) = \prod_{j=1}^{n_i} f_Y(y_{ij} \mid x_{ij}) \times c_i \left(F_Y(y_{i1} \mid x_{i1}), \ldots, F_Y(y_{i,n_i} \mid x_{i,n_i})\right)$$

where $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^n$, $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i=1}^n$, $\boldsymbol{x}_i = \{x_{ij}\}_{j=1}^{n_i}$ and $\boldsymbol{y}_i = \{y_{ij}\}_{j=1}^{n_i}$. The CDF F_Y corresponds to pdf f_Y and c_i denotes the joint density of $(U_{i1}, \ldots, U_{in_i})$. One may evaluate f_Y and F_Y via the identities

$$f_Y(y \mid x) = \frac{1}{\frac{\partial}{\partial \tau} Q_Y(\tau \mid x)}\bigg|_{\tau = \tau_x(y)}, \quad F_Y(y \mid x) = \tau_x(y), \tag{3.3.1}$$

where $\tau_x(y)$ solves $y = Q_Y(\tau \mid x) = \beta_0(\tau) + x^\top \beta(\tau)$ in τ. Therefore, the log-likelihood score of model parameters can be expressed as

$$\ell(\gamma_0, \gamma, \sigma, \omega, \zeta, \theta, \boldsymbol{\phi}) = -\sum_{i=1}^{n} \left\{ \sum_{j=1}^{n_i} \log\{\dot{\beta}_0(U_{ij}) + X_{ij}^\top \dot{\beta}(U_{ij})\} + \log c_i(U_{i1}, \ldots, U_{in_i} \mid \theta, \phi_i) \right\} \tag{3.3.2}$$

where $U_{ij} = \tau_{X_{ij}}(Y_{ij})$.

The logarithm of the Gaussian copula density for cluster i given a correlation matrix R_i is

$$\log c_i(U_{i1}, \ldots, U_{in_i} \mid R_i) = -\frac{1}{2}\left(\log|R_i| + Z_i^\top (R_i^{-1} - I_{n_i}) Z_i\right) \tag{3.3.3}$$

where $Z_i = (\Phi^{-1}(U_{i1}), \ldots, \Phi^{-1}(U_{in_i}))^\top$. The computations associated with matrix inversion and determinant evaluation can be simplified by employing the correlation structures induced by our model formulations. For M1 with exchangeable correlation structure, (3.3.3) is

$$-\frac{1}{2}\left((n_i - 1)\log(1-\phi_i) + \log(1 + (n_i-1)\phi_i) + \frac{\phi_i}{1-\phi_i}\left\{Z_i^\top Z_i - \frac{(Z_i^\top 1_{n_i})^2}{1 + (n_i - 1)\phi_i}\right\}\right).$$

For M2 with the first order autoregressive correlation structure,

$$(3.3.3) = -\frac{1}{2}\left((n_i - 1)\log(1-\phi_i^2) + \frac{\sum_{j=2}^{n_i}\left((Z_{ij} - \phi_i Z_{i,j-1})^2 - (1-\phi_i^2)Z_{ij}^2\right)}{1-\phi_i^2}\right).$$

For M1 and M2, the simplifications reduce the computational complexity of copula density from $\mathcal{O}(n_i^3)$ to $\mathcal{O}(n_i)$, resulting in a $\mathcal{O}(N)$ overall complexity.

For M3 with the continuous temporal copula process, let K_i denote the correlation matrix with $K_{ijk} = \rho_{\mathrm{M}}(t_{ij}, t_{ik} \mid (\nu, \ell))$. Using spectral decomposition $K_i = \Gamma_i \Lambda_i \Gamma_i^\top$, we have

$$(3.3.3) = -\frac{1}{2}\left(\log|\phi_i\Lambda + (1-\phi_i)I_{n_i}| + Z_i^\top \Gamma_i((\phi_i\Lambda_i + (1-\phi_i)I_{n_i})^{-1} - I_{n_i})\Gamma_i^\top Z_i\right).$$

Note that $\phi_i\Lambda_i + (1-\phi_i)I_{n_i}$ is a diagonal matrix and hence its determinant and inverse are easy to compute. Because the prior on ℓ is finitely supported, only finitely many Λ_i and Γ_i need to be computed before running MCMC. The complexity reduces to $\mathcal{O}(n_i^2)$ for ith cluster and the overall complexity is between $\mathcal{O}(N)$ and $\mathcal{O}(N^2)$

3.3.2 MCMC approximation

An efficient log-likelihood evaluation makes our model amenable to Metropolis type Markov chain sampling from the posterior distribution. We augment the adaptive

blocked Metropolis sampling scheme (Haario et al., 2001; Andrieu and Thoms, 2008) of Yang and Tokdar (2017) by alternating between updating parameters $(\gamma_0, \gamma, \sigma, \omega, \zeta)$ of the marginal part, the copula parameters $(\theta, \{\phi_i\}_{i=1}^n)$ and the hyperparameters (μ, ψ). The cluster-specific parameters $\{\phi_i\}_{i=1}^n$ can be either jointly or individually updated. We compare the two schemes across all different simulation scenarios in Section 3.5 and find that the joint updating scheme achieves higher values of the posterior and offers increased effective posterior sample sizes of global parameters. Taking scenario S1×C1 as an example, due to the joint update, the effective posterior sample sizes increase 11% and 307% respectively for global parameters from the marginal part and the copula part. Although the joint updating scheme introduces higher autocorrelations between samples of cluster-specific parameters $\{\phi_i\}_{i=1}^n$, leading to a 43% drop in average effective size, it improves the overall model fitting and the mixing of Markov chains–the posterior samples rapidly arrive at regions with higher posterior values (Figure 3.1). Therefore, we adopt the joint updating scheme for $\{\phi_i\}_{i=1}^n$ throughout the remaining sections. With this joint updating scheme, a generic MCMC algorithm for posterior sampling of the JHQR model is provided in Algorithm 1.

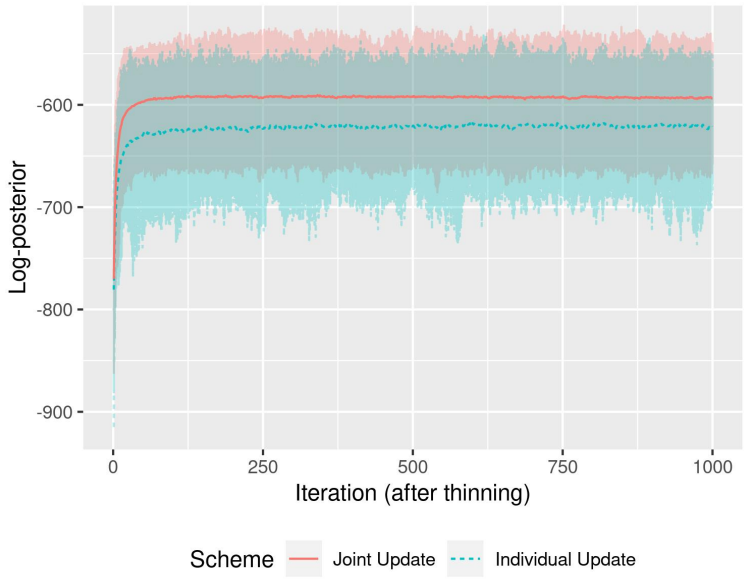

FIGURE 3.1: Log posterior value trajectory against MCMC iteration. The shadow areas represent the trace plots of MCMC runs over 100 datasets in scenario S1×C1 (Section 3.5), while the solid curves represent the corresponding averages.

3.4 Bayesian Predictive Inference

Bayesian prediction of a future observation is performed via the posterior predictive distribution, that is, the conditional distribution of a future observation given observed data. Suppose predicting a new response y^* given predictors x^* in cluster g is of interest. The posterior predictive distribution under the JHQR model is

$$p(y^* \mid x^*, \boldsymbol{x}, \boldsymbol{y}) = \int \underbrace{p(y^* \mid x^*, \boldsymbol{x}, \boldsymbol{y}, \beta_0, \beta, \theta, \boldsymbol{\phi})}_{\text{(I)}} \underbrace{p(\beta_0, \beta, \theta, \boldsymbol{\phi} \mid \boldsymbol{x}, \boldsymbol{y})}_{\text{(II)}} \mathrm{d}\beta_0 \mathrm{d}\beta \mathrm{d}\theta \mathrm{d}\boldsymbol{\phi} \tag{3.4.1}$$

Since the second part of the integrand is the posterior, the overall integration is numerically performed with the posterior samples of model parameters obtained

[3] For example, the degree of freedom of the t copula (Section 4.2) is allowed to take value in $(0.5, +\infty)$ while the set of possible values of the decay parameter in our temporal copula process is finite.

Algorithm 1: Adaptive blocked Metropolis posterior sampler for JHQR

Input: Initial values of parameters: $(\gamma_0^{(0)}, \gamma^{(0)}, \sigma^{(0)}, \omega^{(0)}, \zeta^{(0)})$, $(\theta^{(0)}, \{\phi_i^{(0)}\}_{i=1}^n)$, $(\mu^{(0)}, \psi^{(0)})$; Total number of iterations: T.

Output: T posterior samples of model parameters, induced latent quantile levels.

for $t \leftarrow 1$ **to** T **do**

1. Given $(\theta^{(t-1)}, \{\phi_i^{(t-1)}\}_{i=1}^n)$, draw a new sample $(\gamma_0^{(t)}, \gamma^{(t)}, \sigma^{(t)}, \omega^{(t)}, \zeta^{(t)})$ using the Gibbs-type blocked sampling scheme of Yang and Tokdar (2017);
2. Given the current latent quantile levels $U_{ij}^{(t)}$ extracted from step 1 and $(\theta^{(t-1)}, \mu^{(t-1)}, \psi^{(t-1)})$, jointly sample $\{\phi_i^{(t)}\}_{i=1}^n$;
3. Given $U_{ij}^{(t)}$'s and $\{\phi_i^{(t)}\}_{i=1}^n$, sample $\theta^{(t)}$ using either the adaptive Metropolis sampler if the posterior of θ takes an interval as support, or, from a categorical distribution if the support is finite[3];
4. Given $\{\phi_i^{(t)}\}_{i=1}^n$, jointly sample $(\mu^{(t)}, \psi^{(t)})$;
5. Update proposal distributions of model parameters.

end

return $\{(\gamma_0^{(t)}, \gamma^{(t)}, \sigma^{(t)}, \omega^{(t)}, \zeta^{(t)}), (\theta^{(t)}, \{\phi_i^{(t)}\}_{i=1}^n), (\mu^{(t)}, \psi^{(t)})\}_{t=1}^T$ and $\{U_{ij}^{(t)}, i = 1, \ldots, n; j = 1, \ldots, n_i\}_{t=1}^T$

using the MCMC algorithm described in Section 3.3.2. The evaluation of the conditional density in the first part can be further particularized depending on the specific prediction target. Two important predictive tasks, within-cluster prediction and out-of-cluster prediction, are routinely performed in scientific investigations in many different areas (Moretti, 2004; Sweeting and Thompson, 2012; Kirkbride et al., 2013). To illuminate the difference between two tasks, suppose we have math scores of students from 10 schools. The former type of prediction addresses the question: what is the score for a new student in one of these ten schools, while the latter concerns with a different, but equally important question: what is the score for a (new) student from a new school that is not one of those ten schools.

The two prediction tasks lead to different prediction procedures. The within-cluster prediction posits that g is an existing cluster. Due to the assumption of

conditional independence across clusters, part (I) can be rewritten as

$$p(y^* \mid x^*, \boldsymbol{x}_g, \boldsymbol{y}_g, \beta_0, \beta, \theta, \phi_g) = \frac{p(y^*, \boldsymbol{y}_g \mid x^*, \boldsymbol{x}_g, \beta_0, \beta, \theta, \phi_g)}{p(\boldsymbol{y}_g \mid x^*, \boldsymbol{x}_g, \beta_0, \beta, \theta, \phi_g)}$$

$$= f_Y(y^* \mid x^*) \times c\left(u^* \mid u_{g1}, \ldots, u_{gn_g}, \theta, \phi_g\right) \qquad (3.4.2)$$

where $u^* = F_Y(y^* \mid x^*)$ and $u_{gj} = F_Y(y_{gj} \mid x_{gj})$. It is clear from (3.4.2) that the within-cluster dependency is adjusted through the conditional copula. The quantile of this conditional distribution is derived in Lemma 3.4.1.

Lemma 3.4.1. *Under the JHQR model, the conditional τ^* quantile of Y^* in an existing cluster g, given x^*, model parameters and observed data, is*

$$Q_{Y^*}(\tau^* \mid x^*, \boldsymbol{x}, \boldsymbol{y}, \beta_0, \beta, \theta, \boldsymbol{\phi}) = \beta_0(\tau) + x^{*\top}\beta(\tau), \quad \tau = Q_{U^*}(\tau^* \mid u_{g1}, \ldots, u_{gn_g}, \theta, \phi_g)$$

where $Q_{U^*}(\cdot \mid u_{g1}, \ldots, u_{gn_g}, \theta, \phi_g)$ is the quantile function of the conditional copula with density $c\left(\cdot \mid u_{g1}, \ldots, u_{gn_g}, \theta, \phi_g\right)$.

Proof. One can think of equation (3.4.2) from the perspective of change of variable where $u^* = F_Y(y^* \mid x^*)$ and hence $f_Y(y^* \mid x^*)$ is the Jacobian term. Note that the transformation is strictly monotone increasing and invertible. Hence if τ is the τ^*th quantile of the conditional copula, then the inverse map of τ, that is, $Q_Y(\tau \mid x^*)$, is the τ^*th quantile of the corresponding conditional distribution. □

Remark 3.4.1. *The calculation of τ can be analytically carried out using conditional Gaussian distribution. Denote R_g^* be the n_g-dimensional vector where each element is the correlation between $Z^* := \Phi^{-1}(U^*)$ and Z_g. Then the conditional distribution $Z^* \mid Z_g, \theta, \phi_g \sim \mathsf{N}(\mu^*, (\sigma^*)^2)$ where $\mu^* = R_g^{*\top} R_g^{-1} Z_g$ and $(\sigma^*)^2 = 1 - R_g^{*\top} R_g^{-1} R_g^*$. For the exchangeable correlation structure, these expressions can be condensed into*

$$\mu^* = \frac{\phi_g 1_{n_g}^\top Z_g}{1 + (n_g - 1)\phi_g}, \ (\sigma^*)^2 = 1 - \frac{n_g \phi_g^2}{1 + (n_g - 1)\phi_g}.$$

Similarly, for the autoregressive scenario, the general formulation of the conditional quantile can be simplified by utilizing the Markov property. For a future time point, only the quantile level of the latest observation is conditioned on. Specifically, for the k-step forecasting the correlation between Z^* and Z_{gn_g} is ϕ_g^k and hence $\mu^* = \phi_g^k Z_{gn_g}$, $(\sigma^*)^2 = 1 - \phi_g^{2k}$. After obtaining $(\mu^*, (\sigma^*)^2)$, the conditional quantile level τ can be calculated as $\Phi\left(\mu^* + \sigma^* \Phi^{-1}(\tau^*)\right)$.

In contrast to within-cluster prediction, quantile level adjustment is not required for out-of-cluster prediction. Part (I) in equation (3.4.1) hence reduces to $p(y^* \mid x^*, \beta_0, \beta)$ and the corresponding conditional τ^*th quantile of Y^* given x^*, model parameters and observed data is essentially $\beta_0(\tau^*) + x^{*\top}\beta(\tau^*)$. As the copula parameters do not explicitly exhibit in the expression, one may wonder how the copula model contributes to the out-of-cluster prediction. We note that the inclusion of the copula model greatly improves the inference of the parameters (β_0, β) of the marginal part, as we shall demonstrate in Section 3.5.

3.5 Numerical Experiments

3.5.1 Simulation setup

We present three sets of simulation studies to assess the estimation quality and prediction accuracy of the proposed JHQR model, compared against BQR: Bayesian linear heteroscedastic model with infinite Gaussian mixture error by Reich et al. (2010)[4]; COP: copula-based model with a three-step estimation procedure by Wang et al. (2019); EMP: blocked empirical likelihood approach by Wang and Zhu (2011)[5], and, KB: the classical optimization-based method by Koenker and Bassett (1978). KB is not able to adjust for within-cluster dependencies and serves as a benchmark.

[4] The code is available at `https://www4.stat.ncsu.edu/~bjreich/software`.

[5] The code is available at `https://blogs.gwu.edu/judywang/software`.

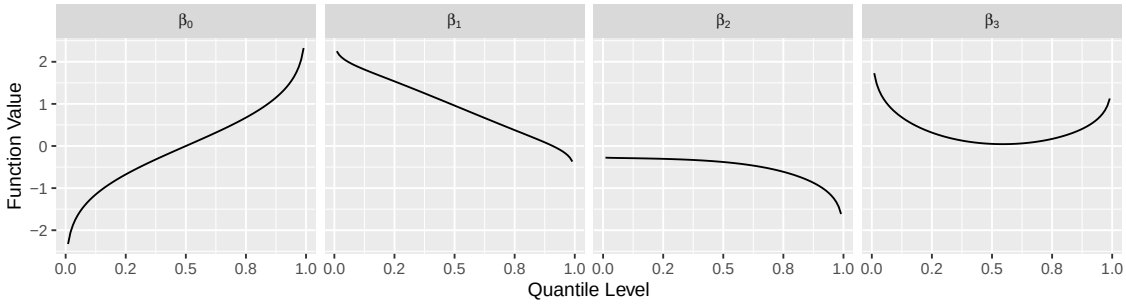

FIGURE 3.2: Visualization of heterogeneous predictor effects of the data generating model (3.5.1) in simulation studies.

The three simulation scenarios were configured with various copula models corresponding to the three dependency patterns introduced in Section 3.2.1, but with the same marginal model adapted from Yang and Tokdar (2017):

$$Q_{Y_{ij}}(\tau \mid X_{ij}) = \beta_0(\tau) + X_{ij}^\top \beta(\tau),\ X_{ij} \overset{\text{iid}}{\sim} \mathsf{Unif}\left(\{x \in \mathbb{R}^3 : \|x\| \leqslant 1\}\right) \tag{3.5.1}$$

with

$$\beta_0(0.5) = 0,\ \beta(0.5) = (0.96, -0.38, 0.05)^\top,$$

$$\dot{\beta}_0(\tau) = \frac{1}{\phi(\Phi^{-1}(\tau))},\ \dot{\beta}(\tau) = \frac{\dot{\beta}_0(\tau)\nu(\tau)}{\sqrt{1 + \|\nu(\tau)\|^2}},$$

$$\nu_k(\tau) = \sum_{l=1}^{3} a_{lk}\phi(\tau; (l-1)/2, 1/9),\ 1 \leqslant k \leqslant 3,$$

$$a = \begin{pmatrix} 0 & 0 & -3 \\ -3 & 0 & 0 \\ 0 & -2 & 2 \end{pmatrix}$$

where $\phi(\cdot)$ denotes the pdf of $\mathsf{N}(0,1)$ inducing a standard normal base distribution of Y_{ij} at $X_{ij} = 0$. It may be difficult to perceive the heterogeneous predictor effects from the mathematical specification and hence we visualize (β_0, β) in Figure 3.2.

To comprehensively examine the statistical performance of JHQR when the model is correctly specified and is misspecified, both Gaussian copula and Laplace

copula were adopted to generate synthetic datasets under each scenario. Specifically, the 3×2 design of copula model of the latent quantile levels were

S1. Exchangeable structure:

$$Z_i \sim \mathsf{N}(0, R_i),\ (R_i)_{jj'} = \phi_i + (1-\phi_i)\mathbb{1}(j = j')$$

S2. Autoregressive structure:

$$Z_i \sim \mathsf{N}(0, R_i),\ (R_i)_{jj'} = \phi_i^{|j-j'|}$$

S3. Temporal stochastic process:

$$Z(t_{i\cdot}) \sim \mathsf{GP}\left(0, \phi_i \rho_{\mathrm{M}}(t_{i\cdot}, t'_{i\cdot}; (\nu, \ell)) + (1-\phi_i)\mathbb{1}(t_{i\cdot} = t'_{i\cdot})\right),\ t_{ij} \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1)$$

C1. Gaussian copula: $U_{ij} = \Phi(Z_{ij})$.

C2. Laplace copula: $U_{ij} = F_{\mathrm{L}}(2\sqrt{2\xi_{ij}} Z_{ij}),\ \xi_{ij} \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$ where $F_{\mathrm{L}}(\cdot)$ is the CDF of Laplace distribution with pdf $f_{\mathrm{L}}(x) = \frac{1}{4}\exp\left(-\frac{|x|}{2}\right)$.

To encourage moderate within-cluster dependency, we set $\phi_i \overset{\text{iid}}{\sim} \mathsf{Beta}(2,2)$ in all designs and fixed $(\nu, \ell) = (2, 0.3)$ for S3. As the implementation of COP only works under the balanced clustering scheme, the numbers of observations were set to be equal across all clusters.[6] For each design, 100 replications of datasets were generated. Each replicated dataset consisted of $n = 50$ clusters with 11 observations in each cluster, among which 10 were in the training set and the remaining one was in the test set. This testing sample was random selected for S1 and was chosen with the largest t_{ij} for S2 and S3 to reflect forecasting at future time points.

The competing methods were compared based on mean absolute errors (MAE) of point estimates of regression coefficients and corresponding coverage probabilities of 95% confidence (or credible) intervals. By assessing the estimation accuracy

[6] JHQR works under both balanced and unbalanced designs, see Section 4.2 and 5.2 for the use of JHQR in unbalanced scenario.

of regression coefficients, we also indirectly evaluated the out-of-cluster prediction performance of models, because the observations from difference clusters were assumed to be conditionally independent given regression coefficients. To examine the within-cluster prediction quality, we also compared the MAE of predicted conditional quantiles $|Q_Y(\tau \mid X, \boldsymbol{x}_g, \boldsymbol{y}_g) - \hat{Q}_Y(\tau \mid X, \boldsymbol{x}_g, \boldsymbol{y}_g)|$ averaged over all observations in the test set. Here $\hat{Q}_Y(\tau \mid X, \boldsymbol{x}_g, \boldsymbol{y}_g)$ is the predicted conditional quantile function for cluster g given by different methods. For clarity, model parameters are omitted from the conditioning set in this expression as different methods contain different parameters. All these comparisons were performed at quantile levels $\{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$ and averaged over all 100 data sets under each scenario. For Bayesian methods, posterior means were used as point estimates. The simulation results on estimation quality for S1×C1 and S1×C2 are presented in Figure 3.3. The results of prediction accuracy for S1 and S2 are summarized in Figure 3.4. Other results are similar and are included in Appendix B.1.

3.5.2 A summary of simulation results

In terms of estimation quality, JHQR offered smallest averaged errors and highest (and approximately nominal) coverage probabilities for estimating the intercept and slopes across all quantile levels in all simulation designs, except for a slightly smaller coverage of the credible interval for intercept compared to COP and EMP. The superiority of JHQR was more pronounced at extreme quantile levels, at which the performances of other methods substantially dropped.

COP provided satisfactory performances in S1 but the estimation procedure was numerically unstable. In each design of S1 and S2, the procedure failed to offer confidence intervals of slopes for 5 datasets (out of 100). Such instability was severe in S3 where the method was not able to provide either point estimates or interval estimates for 89 and 95 datasets respectively for C1 and C2. Therefore, this

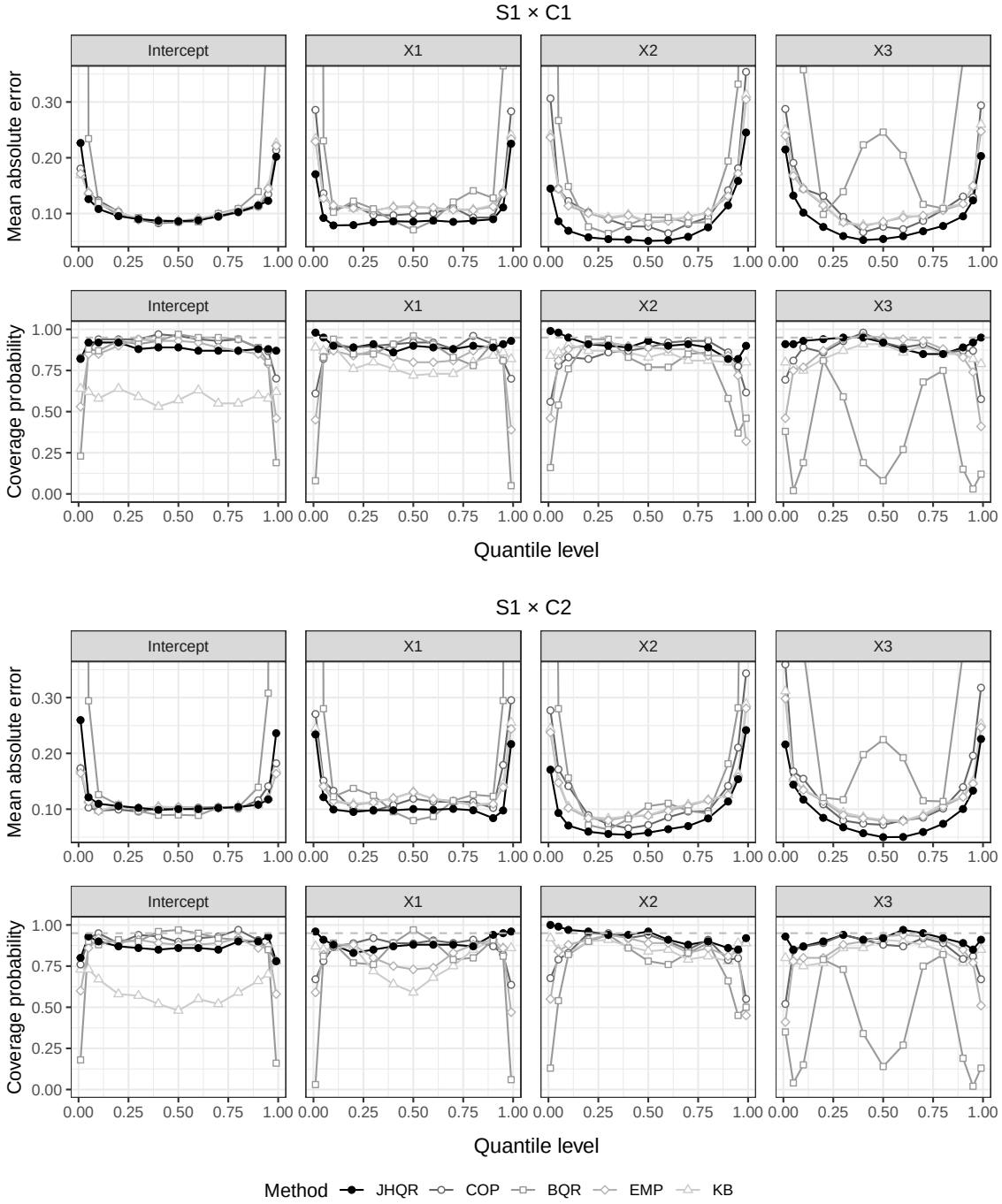

FIGURE 3.3: Estimation quality of different methods for S1×C1 (top panel) and S1×C2 (bottom panel). In each panel, mean absolute errors of point estimates of regression coefficients are presented in the top row, and, coverage probabilities of 95% confidence (credible) intervals of regression coefficients are presented in the bottom row. The evaluations are performed over quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

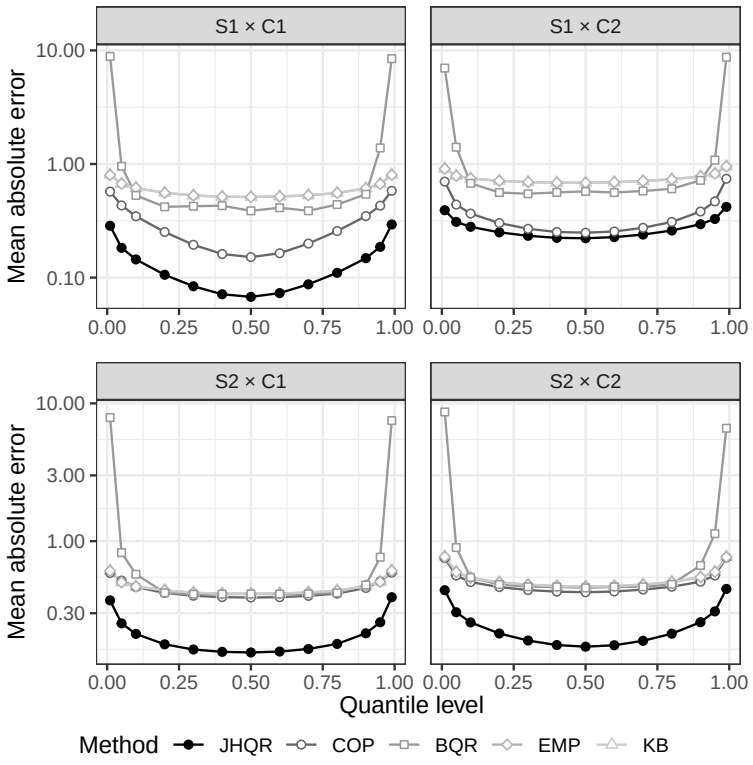

FIGURE 3.4: Within-cluster prediction accuracy of different methods for S1 (top) and S2 (bottom). Mean absolute errors of predicted conditional quantiles are presented over quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

method was not included for comparison in S3 (Appendix B.1, Figure B.1, B.3). EMP provided similar average errors as KB but the former gave better confidence intervals. Both point estimate and interval estimate given by BQR were worse than other methods at non-central quantile levels. The average errors given by the method at extreme quantile levels were not presented due to their large scales. Particularly, the method found it challenging to estimate non-monotonic predictor effect of X_3, as shown in the last panel of Figure 3.2. Such poor performance manifests that the linear heteroscedasticity assumption is restrictive in capturing heterogeneous predictor effects.

The within-cluster prediction accuracies of models were consistent with their estimation performances. JHQR substantially improved over other methods across

all quantile levels in all designs. Excluding JHQR model, COP outperformed the other three methods in S1 but the performances of COP, EMP and KB were indistinguishable in S2. As EMP is not able to model the dependency structure, the method cannot take the dependence into account when performing within-cluster prediction. It was unsurprising that BQR provided the worst prediction at extreme quantile levels. However, the method offered smaller errors compared to EMP and KB at quantile levels from 0.1 to 0.9 in S1, demonstrating that the method was able to adjust for within-cluster dependence, though in a limited sense. Such advantage did not extend to S2 or S3 because BQR assumed an exchangeable within-cluster correlation structure, aligning with the setup in S1.

In summary, the results of extensive numerical experiments evidently suggest the importance of a model based QR approach with proper within-cluster dependence adjustment for predictor effect estimation and conditional quantile prediction. The superior performance of JHQR in all designs, across all quantile levels highlights its competency and flexibility in adjusting for heterogeneous within-cluster dependencies with different structures. Furthermore, JHQR presents adequate robustness against copula model misspecification.

4

Copula Selection and Model Comparison

4.1 Modeling spatial dependence with heavy tailed response data

4.1.1 Desiderata for copula determination

To appropriately incorporate spatial dependence, a copula family must satisfy the following principles. Since the observation units from different locations do not have any particular orders, the dependency induced by the copula should also not be affected by the sampling order of observations. In addition, the dependency pattern should be invariant regardless of whether a particular sample is included in the analysis or not. These two principles share a similar spirit with the Kolmogorov extension theorem and guarantee the compatibility of spatially infill prediction.

We mathematically formalize these principles into the following conditions.

(i) (Permutation invariance) For any two permutations π_1 and π_2 of $\{1, \ldots, n\}$, denote the copulas constructed according to a same procedure with data $\{(Y_{\pi_1(i)}, X_{\pi_1(i)}, s_{\pi_1(i)})\}_{i=1}^n$ and $\{(Y_{\pi_2(i)}, X_{\pi_2(i)}, s_{\pi_2(i)})\}_{i=1}^n$ as $C_{\pi_1}(\cdot, \ldots, \cdot \mid \theta_{\pi_1})$ and

$C_{\pi_2}(\cdot, \ldots, \cdot \mid \theta_{\pi_2})$ respectively. We have

$$C_{\pi_1}(u_{\pi_1(1)}, \ldots, u_{\pi_1(n)} \mid \theta_{\pi_1}) = C_{\pi_2}(u_{\pi_2(1)}, \ldots, u_{\pi_2(n)} \mid \theta_{\pi_2})$$
$$\text{for any } (u_1, \ldots, u_n) \in (0,1)^n$$

(ii) (Closure under marginalization) If $(U_1, \ldots, U_n) \sim C(u_1, \ldots, u_n \mid \theta)$, then $(U_{t_1}, \ldots, U_{t_k}) \sim C(u_{t_1}, \ldots, u_{t_k} \mid \theta)$ for any subset $\{t_1, \ldots, t_k\} \subseteq \{1, \ldots, n\}$.

(iii) (Stationarity) $C(U(s_{t_1}), \ldots, U(s_{t_k}) \mid \theta) = C(U(s_{t_1} + h), \ldots, U(s_{t_k} + h) \mid \theta)$ for any subset $\{t_1, \ldots, t_k\} \subseteq \{1, \ldots, n\}$ and any $h \in \mathbb{R}^r$ such that $\{s_{t_i} + h\}_{i=1}^k \subset \mathcal{S}$.

(iv) Any two observations should be less dependent when their distance becomes larger. Particularly, $C(U_i, U_j \mid \theta) \to U_i U_j$ as $\|s_i - s_j\| \to +\infty$ for any i and j. In the limit, two observations are highly dependent, such that the copula attains the Fréchet-Hoeffding upper bound, when their distance is tiny. Conversely, the observations are essentially independent when they are distant.

In the first condition, θ_π may be invariant when copula parameter θ governs the overall dependency pattern between observations. Examples include the Archimedean copula family and the Gaussian copula process with smoothness and decay parameters. In other cases, θ_π may depend on the specific permutation of observations, such as Gaussian copula with a "correlation matrix" as parameter.

This desiderata rules out the use of many prevalent copula families. For example, multivariate Archimedean copulas have exchangeable dependence structures, that is, all pairwise dependencies are exactly identical. The pairwise correlation induced by the Farlie-Gumbel-Morgenstern copula has a limited range and decreases as sample size increases (Mari and Kotz, 2001). The dependency between observations produced by the structured factor copula model (Krupskii and Joe, 2015) depends on the order of samples. As the computation complexity of the multivariate χ^2 copula (Bárdossy, 2006) density is $O(2^n)$, it has limited practical appeal in our context.

While Vine copulas are flexible in modeling complex dependence structures, a careful design of the $n(n-1)/2$ bivariate margins would be required to satisfy our desiderata, and the resulting $O(n^2)$ parameters would be extremely challenging to estimate from limited data.

Although it seems challenging to construct a copula to satisfy our desiderata, a natural and attractive option which complies with all these conditions is elliptical copula process. In addition to the Gaussian copula process described in 2.2, t copula process is another key member of the family for modeling tail dependence.

4.1.2 Tail dependence and t copula

One of the biggest appeals of QR is that it enables analyzing predictor effects at extreme quantile levels, which could be particularly relevant for analyzing heavy tailed response data. In order to model independent data with heavy tailed response, it is straightforward to adapt JQR by adopting a heavy tailed base distribution f_0. However, the same adaptation alone for JSQR is inadequate for modeling heavy tailed, spatially dependent response because Gaussian copula is *tail independent* (Sibuya, 1960) and hence would fail to account for possible dependence between extreme outcomes at nearby locations. This deficiency, which could lead to considerably biased prediction of extreme quantiles, may be rectified by employing a heavy tailed base distribution in conjunction with a spatial copula process that admits tail dependence. We achieve this by using the so called t-copula process (Embrechts et al., 2001), a tail dependence encoding generalization of the Gaussian spatial copula.

Following Rasmussen and Williams (2006), we say f is a t process on $\mathcal{X}$ with parameter $\psi > 0$, mean function $m(\cdot) : \mathcal{X} \to \mathbb{R}$, covariance function $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if any finite collection of function values $(f(x_1), ..., f(x_n))^\top \sim t_n(\psi, m, K)$ where $K_{ij} = k(x_i, x_j)$ and $m_i = m(x_i)$. We denote $f \sim \mathsf{TP}(\psi, m, k)$. Here $t_d(\psi, \mu, \Sigma)$ denote a multivariate t distribution on $\mathbb{R}^d$ with location μ, scale matrix Σ and degree

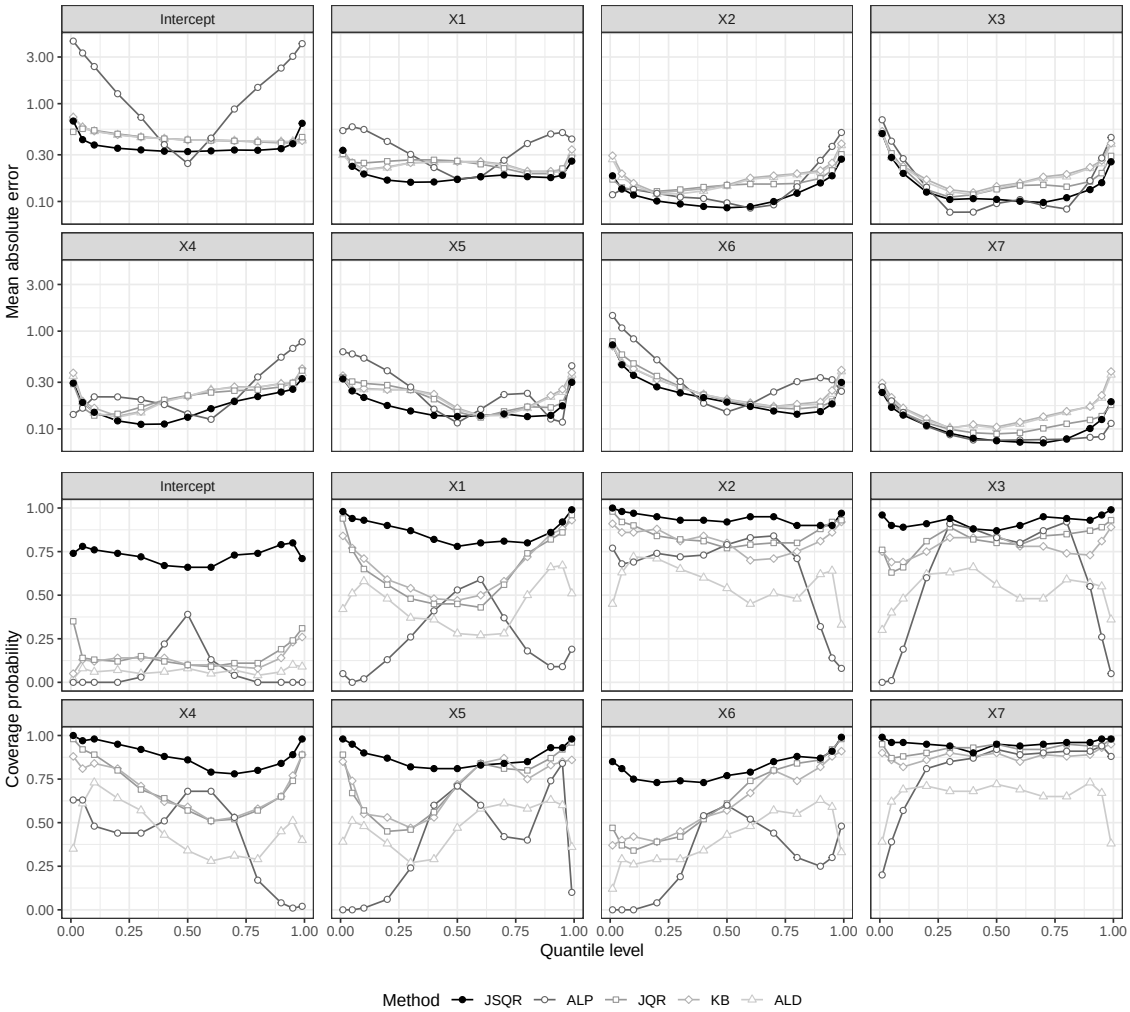

FIGURE 4.1: Inference efficiency of different methods for Example 2 with true generating process being asymmetric Laplace process. The top two rows present the mean absolute errors of regression coefficients while the bottom two rows show the coverage probabilities of 95% confidence (or credible) intervals of regression coefficients at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

of freedom ψ with pdf

$$g(z) = \frac{\Gamma((\psi+d)/2)}{\Gamma(\psi/2)(\psi\pi)^{d/2}|\Sigma|^{1/2}} \left\{1 + \frac{(z-\mu)^\top \Sigma^{-1}(z-\mu)}{\psi}\right\}^{-(\psi+d)/2}.$$

Let T_ψ denote the CDF of univariate, standard t distribution with degrees of freedom ψ. We take the copula process to be $U(s) = T_\psi(Z(s))$, $s \in \mathcal{S}$, where

$$Z(s) \sim \mathsf{TP}(\psi, 0, k(s,s')),\ k(s,s') = \alpha\rho(s,s';(\nu,\phi)) + (1-\alpha)\mathbb{1}(s=s'). \qquad (4.1.1)$$

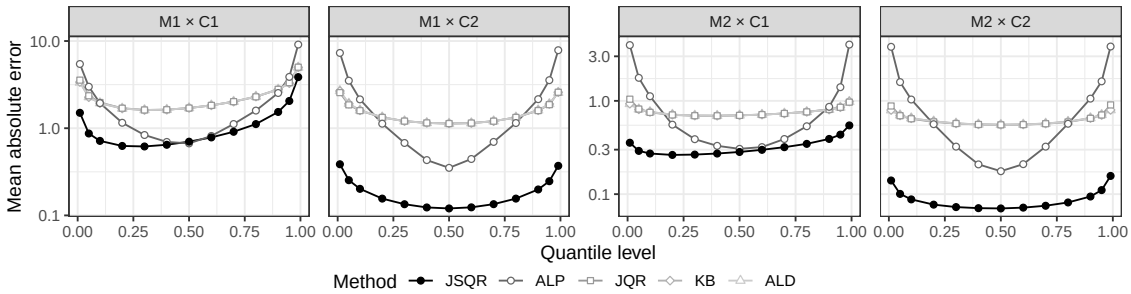

FIGURE 4.2: Mean absolute errors of conditional quantile function at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

An additive decomposition similar to (2.2.2) can be retrieved via an equivalent Gaussian process mixture representation

$$Z(s_i) = W(s_i) + \varepsilon(s_i),\ W(s) \sim \mathsf{GP}(0, \alpha\rho(s, s'; (\nu, \phi))/\varphi)$$

$$\varepsilon(s_i) \overset{\text{iid}}{\sim} \mathsf{N}(0, (1-\alpha)/\varphi),\ \varphi \sim \mathsf{Gamma}(\psi/2, \psi/2).$$

The amount of tail dependence is controlled by ψ with smaller values inducing greater dependence. The t spatial process approaches a Gaussian spatial process in the limit as $\psi \to \infty$, with the asymptotic equivalence kicking in for moderately large values of ψ. We restrict ψ to be larger than 0.5 to avoid numerical instability and adopt $\mathsf{Exp}(0.1)$ as prior for $(\psi - 0.5)$. Other priors (e.g, uniform) are also examined and the results show little sensitivity to the prior choice. Prediction of conditional quantile can be performed similarly to that of the Gaussian copula process and is detailed in the Appendix C.2.1.

Although the t copula admits tail dependence, neither it nor the Gaussian copula can accommodate different lower and upper tail behaviors. To this end, skew-elliptical copulae, based on skew normal distributions (Azzalini and Valle, 1996) and skew t distributions (Azzalini and Capitanio, 2003) may be considered, especially for applications where tail asymmetry is a primary concern, e.g., in finance applications (McNeil et al., 2005). However, our results in Section 2.5.1 indicate that our model with the Gaussian copula is robust against mild asymmetries in true copula process.

Interested readers are referred to Demarta and McNeil (2005) and Smith et al. (2012) on the construction of skew-copulae.

4.1.3 Model assessment and comparison

Because a JSQR analysis could be carried out with various combinations of the marginal and dependence models, a natural question arises as to how to compare such competing models and select one with better predictive quality. A desirable model comparison tool here should be capable of comparing models, with different sets of predictors, base distributions and copula processes, using a single, unified criteria. Apropos of this consideration, we advocate using the Watanabe-Akaike information criteria (Watanabe, 2010, WAIC) for model comparison and selection. WAIC is particularly attractive in complex Bayesian modeling like ours for at least two reasons. From a theoretical perspective, WAIC has the property of averaging over posterior distributions and is asymptotically equivalent to a Bayesian leave-one-out cross validation. From a practical view point, WAIC is computationally efficient as it can be simply calculated by reusing posterior samples (Gelman et al., 2014).

There is an important technical hurdle to applying WAIC directly in our context. The construction of WAIC relies on independence among observations so that the log-likelihood score can be decomposed as the sum of n individual contributions. This is not the case for JSQR because of the copula piece which captures dependence between observation. We circumvent this problem by invoking conditional independence between observation units given the smooth spatial process W, which may be treated as an additional model parameter for WAIC calculation.

According to equation (2.3.1), the log-likelihood for ith observation with $W(s)$ as an additional parameter is

$$-\log\left\{\dot{\beta}_0(\Phi(Z(s_i))) + X_i^\top \dot{\beta}(\Phi(Z(s_i)))\right\} - \log \dot{h}_{W,\alpha}(s_i, V_i) \tag{4.1.2}$$

where $\dot{h}_{W,\alpha}(s_i, V_i) = \frac{\sqrt{1-\alpha}\phi(Z(s_i))}{\phi(\Phi^{-1}(V_i))}$. The process realization can be recovered according to its posterior

$$W(s) \mid Z(s), \theta \sim \mathsf{N}\left(\frac{1}{1-\alpha}\left(\frac{1}{\alpha}K^{-1} + \frac{1}{1-\alpha}I_n\right)^{-1} Z(s), \left(\frac{1}{\alpha}K^{-1} + \frac{1}{1-\alpha}I_n\right)^{-1}\right)$$

and then $V_i = \Phi\left(\left(Z(s_i) - W(s_i)\right)/\sqrt{1-\alpha}\right)$. The derivation for t copula process is similar and is detailed in Appendix C.2.2.

Following Watanabe (2010), we establish an approximate mathematical equivalence between the conditional WAIC score proposed here and the conditional Bayesian leave-one-out cross validated log probability density score; details are provided in Appendix C. Using the proposed conditional WAIC, the results of model comparison for both simulated (Section 4.1.4) and real data (Section 5.1) demonstrate that the criteria is effective in picking up the true model and generally agrees with the actual prediction performance. Since our work focuses on modeling spatial extremes and spatial dependence, we highlight model comparisons between joint quantile regression models with different base distributions (Gaussian or t) and dependency structures (independence, Gaussian or t copula processes). By contrasting these three dependency structures, we are able to determine whether incorporating spatial dependence and whether using the t copula process result in model improvement. As selecting predictors is not our focus here, we refer interested readers to Cunningham et al. (2020) for the use of WAIC in that regard.

Remark 4.1.1. *The idea of conditional WAIC has been investigated in Li et al. (2016) and Millar (2018) in some widely used Bayesian hierarchical models (e.g., finite mixture model, random effects model) where they argue that the conditional WAIC could be unstable and an alternative "integrated WAIC" (by further marginalizing out the latent variables associated with hold-out observations) is suggested. However, we do not observe any instability of the conditional WAIC under the proposed model and there is little difference between values of the conditional WAIC and*

"integrated WAIC". Therefore, we suggest using the conditional WAIC in comparing JSQR models due to its simplicity.

Remark 4.1.2. *Several methods exist for selecting a copula model (Chen and Fan, 2006; Genest et al., 2006; Huard et al., 2006; Dissmann et al., 2013; Smith et al., 2010). However these techniques are unsuitable for our model with multivariate elliptical copula and latent quantile levels. Additionally, a unified criteria is required for determining various components of the JSQR model, for which a single numerical model scoring method, as given by the WAIC, is practically appealing.*

4.1.4 Illustration: model comparison using WAIC

Does the use of a t copula process improve predictive accuracy in analyzing heavy tailed response data? Among various flavors of the JSQR model, does the one with best prediction accuracy get picked when the model with the lowest WAIC score is selected? We compared JQR, JSQR with the Gaussian copula process (JSQR-GP) and JSQR with the t copula process (JSQR-TP). All three models adopted a t base distribution with an unknown degree of freedom to be learned from data. We simulated synthetic data from the linear QR formulation with an univariate predictor $X \sim \mathsf{Unif}(-1, 1)$, where

$$\beta_0(0.5) = 0,\ \beta(0.5) = 0,\ \dot{\beta}_0(\tau) = \frac{3}{t_3(T_3^{-1}(\tau),0,1)},\ \dot{\beta}(\tau) = \frac{\dot{\beta}_0(\tau)\nu(\tau)}{\sqrt{1+\|\nu(\tau)\|^2}},\ \nu(\tau) = 3\,(\tau - 0.5)\,.$$

Three dependency structures were considered: (i) independence, (ii) the Gaussian copula process with $(\alpha, \nu, \phi) = (0.7, 2, 0.3)$, and, (iii) the t copula process with $(\alpha, \nu, \phi, \psi) = (0.7, 2, 0.3, 3)$. Notice that each generating model matched exactly one estimation model under comparison; the rest were either inadequate or redundant. For each scenario, we simulated 100 training sets of sample size 500 and 100 test sets of sample size 50.

Truth	JQR		JSQR-GP		JSQR-TP	
	ΔWAIC	Selection	ΔWAIC	Selection	ΔWAIC	Selection
Independent noise	-	78%	+1.8	8%	+2.1	14%
Gaussian copula	+442.9	0	-	56	+0.5	44
t copula	+402.7	0	+1.8	20	-	80

Table 4.1: Average WAIC scores (relative to the matching model) and model selection rates.

In each scenario, we calculated the differences in WAIC of other two models and that of the matching model. According to Table 4.1, JSQR-GP and JSQR-TP generally had comparable WAICs because they provided similar model fits if the data only contains a few extreme observations. Due to the penalization of more effective model parameters, the WAICs provided by two JSQR models were just slightly larger than JQR when observations were independent. The mean absolute errors of predicted condition quantiles were computed over observations in the test set and presented in Figure 4.3, the last panel of which highlighted the differences between JSQR-GP and JSQR-TP in scenario (iii). It is clear that t copula process improves over Gaussian copula process at extreme quantile levels by 10% to 20% when data is generated from a heavy tailed marginal distribution with a tail dependent copula.

The WAIC and MAE of predicted conditional quantile were consistent as their differences between any two models were small or large simultaneously. Also, the base models provided the smallest WAICs for most data sets in all three scenarios according to Table 4.1. These evidence suggest that WAIC accurately reflects the predictive performance of a model and is suitable for comparing joint QR models with different dependency structures.

4.2 Copula Selection for JHQR

The choice of copula model is crucial to adequately adjusting for different dependency structures in hierarchical data. From a conceptual viewpoint, we first present several

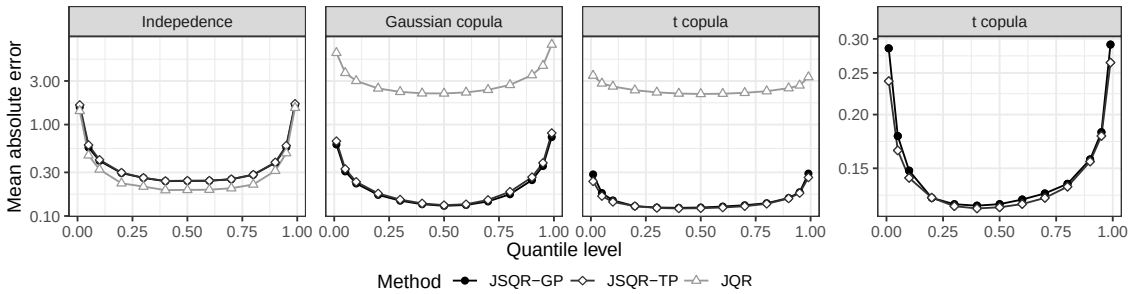

FIGURE 4.3: First three panels: mean absolute errors of predicted conditional quantiles at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$. The fourth panel: A zoomed version of the third panel highlighting the comparison between JSQR-GP and JSQR-TP for scenario (iii).

requirements that the copula family must satisfy to properly model the within-cluster dependency in Section 4.2.1. Complying with these basic requirements, we further discuss the options of copula from the perspective of tail dependence and reflection symmetry, which are important in real data applications. Particularly, we introduce t copula in Section 4.2.2, as a natural yet important generalization of Gaussian copula, to account for tail dependence for all three dependency structures introduced in Section 3.2.1. In addition, we propose using Gumbel copula and Clayton copula in Section 4.2.3 to adjust for reflection asymmetry and tail asymmetry in the scenario of Markov dependency structure for observations on a regular time grid. The discussions above depend fundamentally on the assumption that prior knowledge is available on the dependency pattern or one has a particular type of dependency to detect. If such information is inaccessible, an automatic model selection procedure with WAIC (Watanabe, 2010; Gelman et al., 2014) is provided in Section 4.2.4.

4.2.1 Desiderata for choosing copula

Copula for exchangeable dependence

To model hierarchical data where observations in each cluster do not have any orderings, we naturally assume an exchangeable within-cluster dependency structure of

replicated measurements. Furthermore, the dependency pattern should be invariant regardless of whether a particular replication is included into the analysis or not. In light of these assumptions, the copula distribution must be permutation symmetric and closed under marginalization. Mathematically, this desiderata can be formalized as follows.

Definition 1 (Permutation symmetry, Nelsen (2007))**.** *Let* $(U_1, \ldots, U_d)$ *follows a copula* C*. Then* C *is called permutation symmetric if for any permutation* π *of* $\{1, \ldots, d\}$*, we have* $(U_{\pi(1)}, \ldots, U_{\pi(d)}) \sim C$*.*

Definition 2 (Closure under marginalization, Joe (1997))**.** *Let* $(U_1, \ldots, U_d)$ *follows a* d*-dimensional copula* C_θ*, where* C_θ *belongs to a parametric copula family* $\mathscr{C} = \{C_\theta : \theta \in \Omega\}$ *with parameter* θ*. Then* C_θ *is called to be closed under marginalization if for any subset* $1 \leqslant i_1 < \cdots < i_k \leqslant d$*, we have* $(U_{i_1}, \ldots, U_{i_k})$ *follows a* k*-dimensional copula* C_θ *within the same copula family* $\mathscr{C}$ *and the same parameter* θ*.*

Commonly used copula families that enjoy both properties include elliptical copula and multivariate Archimedean copula (Joe, 2014). The former option is elaborated in detail in Section 4.2.2. For the latter copula family, the copula density becomes increasingly complicated and computationally unstable as dimensionality grows (Hofert et al., 2012). Therefore, we do not pursue in this direction further and refer interested readers to Hofert et al. (2012) and Hofert et al. (2013) for detailed discussions on likelihood-based inference for multivariate Archimedean copula.

Copula for temporal dependence

For modeling the first order Markovian dependency, any bivariate copula can be employed provided that it can induce a flexible range of "positive dependency". We formalize this idea by requiring that the Kendall's τ (Kendall, 1938) induced by the copula can be any value between 0 and 1. As a more general measure of *concordance*

between two random variables than commonly used Pearson's correlation, Kendall's τ does not require finite second moments and is invariant to monotone transformations. This requirement rules out some widely-used copula such as Ali–Mikhail–Haq copula and Farlie–Gumbel–Morgenstern copula.

To appropriately incorporate temporal dependence for observations at unequally spaced timestamps, we adopt the stationarity assumption that the dependence between any $U(t_{i\cdot})$ and $U(t'_{i\cdot})$ is a decreasing function of their temporal distance $|t_{i\cdot} - t'_{i\cdot}|$. In addition, the closure under marginalization property should also hold in this scenario because, for example, whether the physical exam result of a patient in any particular month is missing should not affect the dependency pattern of the results in remaining months. These conditions promote the use of elliptical copula processes.

4.2.2 Modeling tail dependence with t copula

Tail dependence is especially important when analyzing heavy-tailed response data. Failing to account for tail dependence may result in substantially biased prediction of extreme response quantiles when cluster-specific dependency is high and/or observations are collected at neighboring time points. Unfortunately, despite its flexibility and simplicity, Gaussian copula is tail independent, inducing seemingly independent extreme events; see the top row of Figure 4.4 for an illustration. To adequately adjust for extreme outcome dependence, we propose using a t base marginal distribution jointly with a t copula. As a tail dependent extension of Gaussian copula, t copula admit symmetric tail dependence at both lower and upper quantiles. Specifically, let $t_d(\eta, \mu, \Sigma)$ denote a d-dimensional t distribution with location parameter μ, scale matrix Σ and degree of freedom η with pdf

$$f(z) = \frac{\Gamma\left((\eta + d)/2\right)}{\Gamma(\eta/2)(\eta\pi)^{d/2}|\Sigma|^{1/2}} \left\{1 + \frac{(z-\mu)^\top \Sigma^{-1}(z-\mu)}{\eta}\right\}^{-(\eta+d)/2}.$$

Let T_η denote the CDF of univariate, standard t distribution with degrees of freedom η. For cluster i, a general formulation of the t copula model for all dependency structures can be written as

$$U_{ij} = T_\eta(Z_{ij}),\ (Z_{i1}, \ldots, Z_{in_i})^\top \sim t_{n_i}(\eta, 0, R_i) \tag{4.2.1}$$

where R_i is the induced correlation matrix which has the same structure as that in the Gaussian copula model introduced in Section 3.2.1. Therefore, the t copula model inherits all model parameters from the Gaussian copula model with an additional parameter η controlling tail dependence. The t copula is approximately equivalent to Gaussian copula when η is moderately large ($\sim$50) and eventually degenerates to Gaussian copula in the limit as $\eta \to \infty$. To avoid numerical instability, we restrict η to be larger than 0.5 and assign $\mathsf{Exp}(0.1)$ as its prior. Similar simplifications as described in Section 3.3.1 can be adopted for copula density evaluation.

The within-cluster prediction can be carried out using the conditional t copula. For cluster g, define $Z_g = \left(T_\eta^{-1}(U_{g1}), \ldots, T_\eta^{-1}(U_{gn_g})\right)^\top$. Let R_g^* be the n_g-dimensional vector of induced correlations between $Z^* = T_\eta^{-1}(U^*)$ and Z_g. Based on model specification (4.2.1), we have

$$Z^* \mid Z_g, \eta \sim t_1\left(\eta + n_g, \mu^*, \frac{\eta + d}{\eta + n_g}(\sigma^*)^2\right)$$

where $d = Z_g^\top R_g^{-1} Z_g$, $\mu^* = R_g^{*\top} R_g^{-1} Z_g$ and $(\sigma^*)^2 = 1 - R_g^{*\top} R_g^{-1} R_g^{*\top}$. Following the notations in Lemma 3.4.1, the quantile function of the conditional copula is

$$Q_{U^*}(\tau^* \mid u_{g1}, \ldots, u_{gn_g}, \theta, \phi_g) = T_\eta\left(\mu^* + \sqrt{\frac{\eta + d}{\eta + n_g}}\sigma^* T_{\eta+n_g}^{-1}(\tau^*)\right).$$

4.2.3 Modeling reflection asymmetry with Gumbel and Clayton copula

Although t copula is capable of modeling tail dependence, both Gaussian and t copula admit symmetric lower and upper quantile behaviors and hence are inadequate

to adjust for the so-called "reflection asymmetry"[1], which is prevalent in financial applications (Ang and Chen, 2002; Hong et al., 2007; Rodriguez, 2007). We propose modeling such reflection asymmetry with Gumbel and Clayton copula because together they cover a broad range of different dependency patterns. From the perspective of flexibility, both copula are able to accommodate for positive dependency with various strengths where Kendall's τ varies from 0 to 1. In terms of tail property, Gumbel and Clayton respectively admit upper and lower tail dependencies (Table 4.2). Computationally, their densities are analytically available and the resulting conditional quantiles are easy to obtain (details below).

We focus on the Markovian dependency structure where the joint copula can be decomposed into a series of bivariate copula. Using the Markovian property, the copula model is

$$U(t_{ij}) \sim C(\cdot \mid U(t_{i,j-1}), \phi_i),\ j = 2, \ldots, n_i \quad U(t_{i1}) \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1),\ i = 1, \ldots n$$

where $C(\cdot \mid U(t_{i,j-1}), \phi_i)$ is the conditional copula of $U(t_{ij})$ given the lag-1 quantile level and cluster-specific copula parameter ϕ_i. Again, to encourage information sharing and shrinkage across clusters, we impose a hierarchical prior on $\{\phi_i\}_{i=1}^n$. To facilitate and unify the prior specifications for different copula models, we propose specifying a prior directly on Kendall's τ instead of on copula parameters because their meanings are not always clear and their ranges are different (Table 4.2). Following the same recipe described in Section 3.2.2, we adopt a $\mathsf{Beta}(\mu, \psi)$ prior with hyperpriors on (μ, ψ).

[1] A bivariate copula C is called reflection symmetric if $C(u,v) = C(1-u, 1-v)$ for any $u, v \in [0,1]$.

Copula	Tail dependence	Reflection symmetry	Kendall's τ	Range of ϕ
Gaussian	$\lambda_L = \lambda_U = 0$	Yes	$\frac{2}{\pi} \arcsin \phi$	$[0, 1]$
t	$\lambda_L = \lambda_U = 2T_{\eta+1}\left(-\sqrt{\frac{(\eta+1)(1-\phi)}{(1+\phi)}}\right)$	Yes	$\frac{2}{\pi} \arcsin \phi$	$[0, 1]$
Gumbel	$\lambda_L = 0$, $\lambda_U = 2 - 2^{1/\phi}$	No	$1 - 1/\phi$	$[1, +\infty)$
Clayton	$\lambda_L = 2^{-1/\phi}$, $\lambda_U = 0$	No	$1 - 2/(2 + \phi)$	$(0, +\infty)$

Table 4.2: Summary of key properties of four different copula models adopted in the Markovian dependency scenario.

The procedure of performing within-cluster prediction is greatly determined by the scheme of sampling from a copula. Unlike Gaussian or t copula, where the quantile levels can be obtained by simple transformations of samples drawn from multivariate Gaussian or t distributions, quantile function of conditional copula is the key for generating data under the Gumbel and Clayton copula models with Markovian structure. For a bivariate copula $C(u, v)$, the conditional copula of u given v is $C(u \mid v) = \partial C(u, v)/\partial v$. By using the "inverse-CDF" method, the quantile levels can be serially generated. Specifically, the data generating process for any cluster under the JHQR model can be summarized in Algorithm 2. In this algorithm, the

Algorithm 2: Data generating process under JHQR with Markovian dependency structure.

Input: Sample size n; Predictor values: $\{x_i\}_{i=1}^n$; Regression coefficients: (β_0, β); Initial quantile level: $u_1 \sim \mathsf{Unif}(0, 1)$.

Output: n observations of the response.

Calculate $y_1 = \beta_0(u_1) + x_1^\top \beta(u_1)$

for $i \leftarrow 2$ **to** n **do**

1. Generate $w_i \sim \mathsf{Unif}(0, 1)$;
2. Obtain $u_i = C^{-1}(w_i \mid u_{i-1})$;
3. Generate $y_i = \beta_0(u_i) + x_i^\top \beta(u_i)$.

end

return $\{y_i\}_{i=1}^n$

conditional quantile function C^{-1} for Clayton copula is analytically available. For Gumbel copula, however, the quantile level u_i must be calculated by numerically solving $w_i = C(u_i \mid u_{i-1})$. Note that $C(\cdot \mid u_{i-1})$ is a monotone increasing function with domain $[0, 1]$. The root finding can be efficiently carried out using, for example, bisection method. The associated computations of copula densities and conditional copula are detailed in Appendix B.2.

To demonstrate the data generating process and to showcase different dependency patterns induced by different copula models, we present a numerical example using the setting in Section 4.2.4. As our focus is on the differences between copula, we implemented Algorithm 2 with sample size $n = 500$ and $x_i = 0$ for all i. We simulated the same set of $w_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0, 1)$ for all four copula models with 0.75 being the same value of Kendall's τ[2] such that the generated response values are directly comparable. We visualize the simulated quantile levels and responses in Figure 4.4.

Except for the Clayton copula, which presents a quite different upper quantile behavior, it is difficult to distinguish between the other three copula from the lag-1 quantile plots (the left column in Figure 4.4) though some differences exist. Their key characteristics are clearly pronounced in the tail behaviors of the generated response values. The lag-1 response plots in the middle column are consistent with the tail dependence coefficients in Table 4.2. For example, the Gumbel copula admits upper tail dependence but is tail independent in the lower tail. Consequently, there are several unstructured extreme responses in the lower left corner but not in the upper right corner. The actual time series presented in the right column further highlights their differences. In the green regions, response values under difference copula are all relatively larger than the average. The response values simulated with t and Gumbel present more dependence but some extremes are generated with Gaussian as well as

[2] This corresponds to setting the correlation ϕ to be 0.92 for Gaussian copula, to give a sense of the dependence strength.

Clayton and appear to be much larger than neighboring responses. Such phenomenon is expected because both Gaussian and Clayton copula are upper tail independent which suggests that extreme events with high response occur independently. The patterns exhibited in the purple regions can be explained by a similar argument. In summary, the great variety of dependency patterns presented in this numerical example manifests the great modeling flexibility of JHQR when adopting different copula. Particularly, the JHQR model with Markovian dependency structure is able to generate stationary time series with persistent bursts, which resembles the key feature of the generalized autoregressive conditional heteroskedasticity (GARCH) model.

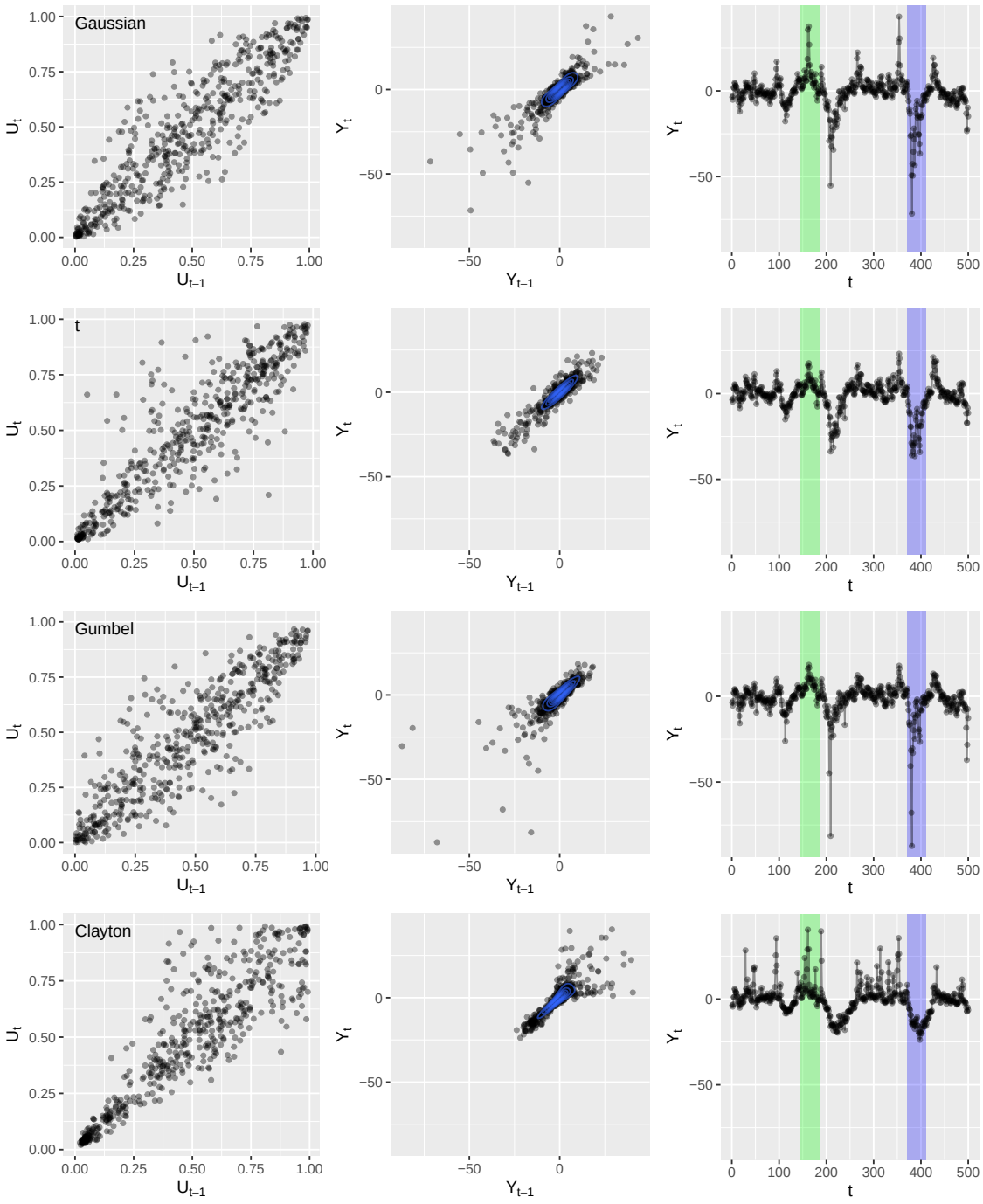

FIGURE 4.4: Illustration of (a) data generating process under JHQR with copula admitting first order Markovian property, and, (b) different dependency patterns of quantile levels (hidden) and actual responses (observed) induced by different bivariate copula.

4.2.4 Joint model selection with WAIC

Although we have demonstrated that the proposed JHQR model is capable of incorporating different dependency structures by adopting different copula, it is typically difficult to determine the dependency pattern in advance before model fitting in real applications. It is hence practically useful to develop an automatic procedure for selecting an "optimal" model for the data, from a set of candidate models. The optimality can be defined in many possible ways and here we take the perspective of Bayesian predictive inference described in Section 3.4, where we focus on the prediction performance of model on future data while taking the uncertainty of model parameters into account.

The in-sample approximation of the out-of-sample prediction is typically performed using cross validation (CV). Despite its popularity, CV requires model refittings and hence is computationally expensive. As a cheaper approximation, Watanabe-Akaike information criteria (Watanabe, 2010) is asymptotically equivalent to Bayesian leave-one-out CV and only demands fitting model once on the whole dataset. Therefore the computation of WAIC relies solely on the single model fitting which is exactly the procedure needed for performing inference on model parameters. It is particularly advantageous to adopt WAIC in our context as the posterior samples can be recycled to calculate the marginal likelihood using Monte Carlo integration.

Following the discussion in Section 3.4, we propose two variants of the standard WAIC, oWAIC and iWAIC, respectively aiming for out-of-cluster and within-cluster predictions. Careful design of the leave-one-out scheme is crucial to properly accommodate for hierarchical dependency structure and to accurately reflect the prediction target.

Remark 4.2.1. *Both Millar (2018) and Merkle et al. (2019) use the terminologies "marginal WAIC" and "conditional WAIC" to distinguish between whether the*

cluster-specific parameters are conditioned on in WAIC calculation. As we shall elaborate presently, such procedure-based terms are ambiguous in our context because the computation of out-of-cluster WAIC under JHQR *also requires conditioning on cluster-specific parameters* $\{\phi_i\}_{i=1}^n$*. Therefore, we instead adopt objective-based terms to directly reflect the prediction tasks that the WAIC targets on.*

Out-of-cluster prediction

For out-of-cluster prediction, future data is assumed to come from a new cluster that is not present in the training data. The corresponding in-sample analogy is "leave-one-cluster-out" (Millar, 2018; Merkle et al., 2019) where all observation units in a cluster are jointly considered as a single sample. Recall the notations in Section 3.3.1 and denote $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ as the whole dataset. Let $\mathcal{D}_{-i}$ be the partial data without observations in the ith cluster. Therefore, the targeting Bayesian leave-one-cluster-out predictive density is $p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \mathcal{D}_{-i})$. Conditioning on both global and local model parameters, n samples are independent and the log-likelihood score for the ith observation is essentially

$$-\sum_{j=1}^{n_i} \log\{\dot{\beta}_0(U_{ij}) + X_{ij}^\top \dot{\beta}(U_{ij})\} - \log c_i(U_{i1}, \ldots, U_{in_i} \mid \theta, \phi_i). \tag{4.2.2}$$

By summing (4.2.2) over all samples with a correction term (Gelman et al., 2014), the oWAIC can be calculated with almost no extra computation because the log-likelihood scores are tracked in MCMC sampling. Note that the computation of oWAIC is not affected by the within-cluster dependency and can be carried out for all dependency scenarios with any copula model.

Within-cluster prediction

Compared to oWAIC, the computation procedures of iWAIC are distinct for different dependency structures and copula models. For clustered data without orderings, the

future data from an existing cluster is assumed to be exchangeable with training observations in this cluster. The associated "leave-one-unit-out" predictive density is $p(y_{ij} \mid x_{ij}, \mathcal{D}_{-ij})$, $i = 1, \ldots, n$; $j = 1, \ldots, n_i$. For M1 with Gaussian copula, conditioning on the cluster-specific level W_i in addition to model parameters, the observations are independent and the model can be equivalently written as

$$Y_{ij} = \beta_0(h_{W_i,\phi_i}(V_{ij})) + X_{ij}^\top \beta(h_{W_i,\phi_i}(V_{ij})),\ V_{ij} \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1) \tag{4.2.3}$$

where $h_{w,\phi}(v) = \Phi(w + \sqrt{1-\phi}\Phi^{-1}(v))$ with $(w, \phi, v) \in \mathbb{R} \times (0,1)^2$. The corresponding conditional log-likelihood is

$$-\log\left\{\dot{\beta}_0(h_{W_i,\phi_i}(V_{ij})) + X_{ij}^\top \dot{\beta}(h_{W_i,\phi_i}(V_{ij}))\right\} - \log \dot{h}_{W_i,\phi_i}(V_{ij}) \tag{4.2.4}$$

where $\dot{h}_{W_i,\phi_i}(V_{ij}) = \frac{\sqrt{1-\phi_i}\phi(Z_{ij})}{\phi(\Phi^{-1}(V_{ij}))}$. The latent level W_i is marginalized out in our sampling Algorithm 1 but can be recovered with

$$W_i \mid Z_i, \phi_i \sim \mathsf{N}\left(\frac{\phi_i 1_{n_i}^\top Z_i}{1 + (n_i - 1)\phi_i}, \frac{\phi_i(1-\phi_i)}{1 + (n_i - 1)\phi_i}\right)$$

and then $V_{ij} = \Phi\left((Z_{ij} - W_i)/\sqrt{1-\phi_i}\right)$.

The computation of iWAIC for t copula model is similar by utilizing the Gaussian mixture representation of model (4.2.1)

$$\begin{aligned} U_{ij} &= T_\eta(Z_{ij}),\ Z_{ij} = W_i + \varepsilon_{ij} \\ W_i &\sim \mathsf{N}\left(0, \phi_i/\varphi_i\right),\ \varepsilon_{ij} \overset{\text{ind}}{\sim} \mathsf{N}\left(0, (1-\phi_i)/\varphi_i\right) \\ \varphi_i &\overset{\text{iid}}{\sim} \mathsf{Ga}(\eta/2, \eta/2). \end{aligned}$$

The observations are conditionally independent given cluster-specific parameters (W_i, φ_i, ϕ_i) and global parameters (β_0, β, η). The computation of the log-likelihood is detailed in the Appendix C.3.

When observations in each cluster have temporal orderings, the prediction task typically concerns with forecasting at future time points. To resemble the task using in-sample approximation, the general leave-one-out strategy is not appropriate

because it is unreasonable to predict previous events given future data. Instead, the leave-one-unit-out scheme must be adjusted such that the latest observation is held out. Therefore, the target predictive density is $p(y_{in_i} \mid x_{in_i}, \mathcal{D}_{-in_i})$ for $i = 1, \ldots, n_i$. For the continuous temporal copula process model (M3), observations are conditionally independent given model parameters and realizations of the latent process $W(t_{ij})$. As the calculations of the conditional log-likelihood for both Gaussian and t copula processes are similar to those of Chen and Tokdar (2019) for spatial copula process, we omit detailed derivations here. The key difference is that the computation here is only carried out for the latest observation in each cluster.

Remark 4.2.2. *Unfortunately, for* JHQR *with discrete Markovian dependence (M2), the observations are still dependent even when conditioning on cluster-specific parameters. Therefore, the proposed iWAIC is not applicable for this scenario.*

Remark 4.2.3. *Although this "leave-final-unit-out" scheme is conceptually legitimate, it may provide a poor approximation with large bias when data only contains a limited number of clusters. Under this circumstance and the discrete Markovian scenario, we recommend adopting the leave-future-out CV with Pareto smoothed importance sampling (Bürkner et al., 2020) to evaluate the within-cluster prediction quality of the model. Further investigations in this direction are out of the scope of the present work and the interested readers may consult the reference for more details.*

Effectiveness of the proposed WAIC scores

To verify the proposed oWAIC and iWAIC indeed provide truthful approximations of the corresponding Bayesian in-sample predictive losses, we establish a mathematical connection between the two quantities. Particularly, following Li et al. (2016) and Watanabe (2010), we show that the Taylor's expansions of the two quantities agree

in the leading 3 terms, for all three dependency scenarios and all copula models. The proof is detailed in the Appendix C.4.

To thoroughly examine the empirical performance of the proposed WAIC scores for model selection, we performed a set of simulation studies with the Markovian dependency structure because 4 different copula models could be adopted under this scenario. The synthetic data was simulated with the same marginal QR model under 4 different copula models. Specifically, the marginal model was

$$\beta_0(0.5) = 0,\ \beta(0.5) = 0,\ \dot{\beta}_0(\tau) = \frac{5}{t_2(T_2^{-1}(\tau))},$$

$$\dot{\beta}(\tau) = \frac{\dot{\beta}_0(\tau)\nu(\tau)}{\sqrt{1 + \|\nu(\tau)\|^2}},\ \nu(\tau) = 5(\tau - 0.5).$$

The setups of the Gaussian and t copula models were kept same as those in Section 3.5.1 with the additional degree of freedom parameter of the t copula specified as 2. For Gumbel and Clayton copula, we first generated the cluster-specific Kendall's τ i.i.d. from $\mathsf{Beta}(2, 2)$ and then transformed back to corresponding $\{\phi_i\}_{i=1}^n$. We simulated 100 datasets with each copula model. Each dataset contained 50 clusters and the cluster sizes i.i.d. followed a Poisson distribution with 20 as the expected value.

We fitted JHQR with the t base distribution and 4 different copula models under each copula setting and compare the associated oWAIC scores. A fruitful set of conclusions can be drawn from the simulation results summarized in Figure 4.5. First, the matching models are selected most frequently in all simulation settings. Although oWAIC is targeting for the prediction performance of the model, it is desirable for the criteria to be capable of picking the model that is correctly specified. Second, the scales of the differences between oWAIC scores are consistent with the copula properties. From the perspective of tail dependence, t copula is the most flexible model (among the 4 copula we considered) as it admits both lower and upper tail

dependencies. Therefore, it gets selected the second often in all three settings where t copula is misspecified. Particularly, as a generalization of the Gaussian copula, t copula provides a similar selection percentage and oWAIC scores as Gaussian copula, when the latter is the true model. As displayed in the lower left corner panel of Figure 4.4, the upper quantile behavior of the Clayton copula is quite different from other copula models and hence it is seldom selected when misspecified.

Remark 4.2.4. *As unified model selection criteria, the proposed WAIC scores are also capable of comparing models with different sets of predictors and/or different base distributions, though the focus here is copula selection using oWAIC. With two real applications in Section 5.2, we highlight the use of iWAIC for selecting predictors and copula models, and, the consistency between iWAIC and actual within-cluster prediction performance. We refer interested readers to Cunningham et al. (2020) and Chen and Tokdar (2019) for complementing the present discussions of WAIC. Specifically, the former provides a comprehensive case study in environmental science where WAIC is used to select predictors and determine their formats (e.g., interactions, transformations) in the joint QR model. The latter emphasizes the advantage of t copula over Gaussian copula in tail quantile estimation and connects such advantage with the difference on WAIC scores, in the context of spatial data analysis.*

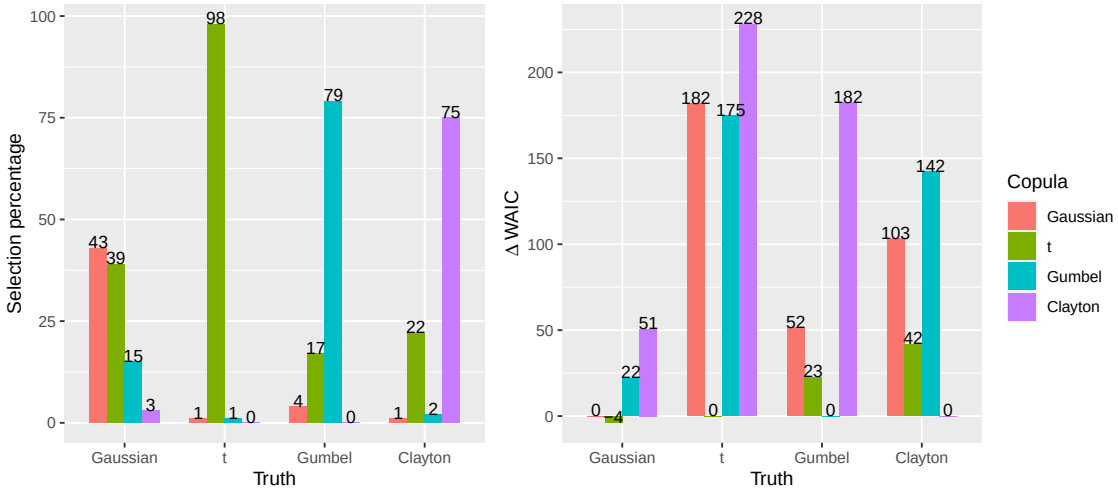

FIGURE 4.5: Model selection percentages and corresponding relative WAIC scores for different simulation scenarios.

5

Case Studies

5.1 Spatial data analysis with JSQR

Two case studies are presented to demonstrate the potential of JSQR models in real applications. In both examples, we showcase the superior prediction performance of the proposed models compared to existing others, and, the effectiveness of WAIC in selecting base distributions and copula processes. In the analysis of $PM_{2.5}$ concentration data (Section 5.1.1) with 339 observations, we highlight the insightful heterogeneous predictor effects offered by JSQR in contrast to the classical BSRE model. In the wildfire risk analysis (Section 5.1.2), we feature the t-copula process in capturing spatial tail dependence with a substantially larger dataset with 1855 observations.

We clarify that this section serves as a validation of the proposed model in real data analysis. Both WAICs and prediction performances are evaluated using multi-fold validation studies to reduce randomness. However, readers who intend to use the JSQR model in their applications should fit the model on the whole dataset (without split) regardless of the purposes of model selection, predictor effect inference

or quantile prediction.

5.1.1 Analysis of $PM_{2.5}$ concentration data

Exposure to fine particulate matter (diameter less than 2.5 μm) is positively associated with the risk for lung cancer and cardiovascular disease morbidity and mortality (Dockery et al., 1993; Pope and Dockery, 2006; Brook et al., 2010). Various regression models, e.g., generalized additive models (Yanosky et al., 2008b), Cox hazard models (Pope et al., 1995; Jerrett et al., 2005) and conditional autoregressive models (Paciorek, 2013), have been proposed to predict $PM_{2.5}$ concentration. These methods focus only on the response mean prediction. But, arguably, predicting high response quantiles is of a greater concern here. High $PM_{2.5}$ concentration is more harmful as acute exposure could trigger cardiovascular morbidity (Bell et al., 2005). Additionally, the relationship between $PM_{2.5}$ and many important predictors (detailed below) appears more complex than a homoskedastic linear dependence (Figure 5.1), rendering classical mean regression analyses rather inadequate. Quantile regression has been recently employed to analyze $PM_{2.5}$ concentration (Barmpadimos et al., 2012; Porter et al., 2015). However, analyses based on KB estimates clearly overlook the spatial dependence of data from different monitoring stations as presented in Figure 5.1(a).

We reanalyzed data[1] on $PM_{2.5}$ reported in Paciorek (2013) using the proposed JSQR models. The dataset contained averaged daily $PM_{2.5}$ concentrations (shown in Figure 5.1) from 2001 to 2002 monitored at 339 stations in the northeastern United States. In Yanosky et al. (2008a) and Paciorek et al. (2009), the following 5 covariates have been shown to have significant associations with $PM_{2.5}$ or improve predictions and hence were included into our analysis: (i) logarithm of population density at county level (LCYPOP), (ii) logarithm of distance to A1 class roads (primary roads,

[1] The dataset is available at `https://www.stat.berkeley.edu/users/paciorek/code/ejs/`.

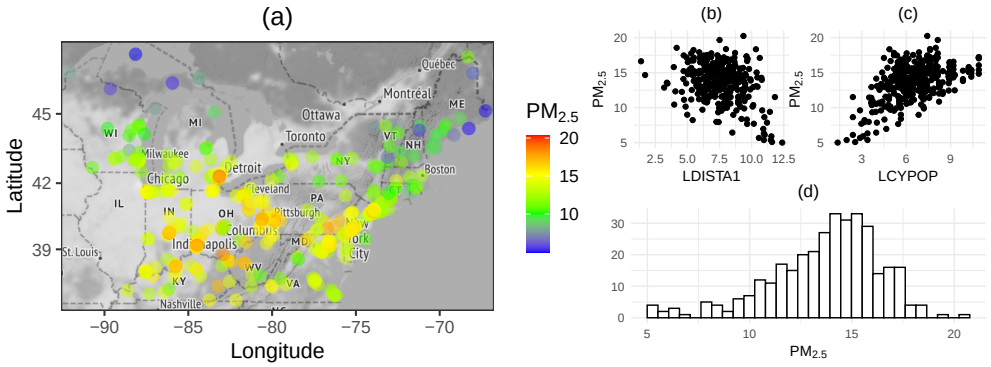

FIGURE 5.1: Averaged daily $PM_{2.5}$ concentration ($\mu g/m^3$) during 2001-2 across monitoring stations in northeastern United States (a) and against two potential relevant predictors (b)-(c); response data does not appear heavy tailed (d).

typically interstates) (LDISTA1), (iii) proportion of urban land use within 1 kilometer of the station location (URB), (iv) logarithm of $PM_{2.5}$ emissions within a 10 kilometer buffer (L25E10), and, (v) elevation of the station (ELEV). We refer interested readers to Yanosky et al. (2008a) for a detailed description of the dataset.

Assessment of infill quantile prediction

To examine the existences of heavy-tailedness and tail dependence, we adopted three JSQR models with different combinations of base distributions and copula processes: JSQR-GP1 with a logistic base and Gaussian copula, JSQR-GP2 with a t base and Gaussian copula, and, JSQR-TP with a t base and a t copula. The methods in Section 2.5.1 were also applied for comparison. A 10-fold validation study was conducted to assess how well different spatial and non-spatial quantile regression methods performed in predicting response quantiles at hold-out locations. In each fold, the data was randomly partitioned into a training set and a test set with 270 and 69 observations respectively. All prior specifications and MCMC settings were kept same as those in Section 2.5. The accuracy of quantile prediction at a level τ was evaluated by the check loss $\rho_\tau(Y_i - \hat{Q}_{Y_i}(\tau \mid X_i, s_i, U_1, \ldots, U_n))$ averaged over all observations in test sets.

Among the four joint QR models, JSQR-GP1 had the smallest average WAIC (808), closely followed by JSQR-GP2 and JSQR-TP (both at 810) and trailed by JQR (1124). This relative ordering suggests the existence of spatial noise correlation among observation units, but the distribution of $PM_{2.5}$ does not appear to be heavy tailed (consistent with Figure 5.1(d)) or exhibit tail dependence. The order of WAICs were coherent with actual prediction performances (not reported), JSQR-GP1 offered slightly smaller average check losses than other JSQR models and provided a clear improvement over JQR. Therefore, we adopted JSQR-GP1 as a representative of JSQR models for all subsequent analyses.

Prediction accuracy of QR models, measured by averaged check loss relative to BSRE, are summarized in Figure 5.3. For QR models, JSQR-GP1 consistently offered the smallest averaged loss across all quantile levels on test sets. The losses given by JQR, KB and ALD were similar and typically the largest except for extreme quantile levels. ALP had smaller losses than the three non-spatial methods from $\tau = 0.2$ to 0.8 but provided larger losses out of this range. Overall, these results are consistent with our findings in Section 2.5.1 and underline the necessity of adjusting for spatial dependence toward a more accurate QR analysis of $PM_{2.5}$ concentration. We note that BSRE and JSQR offered comparable prediction accuracy for this data set. This is perhaps due to the fact that our data consisted of daily measurements averaged across two years, thus boosting Gaussianity of the joint distribution of predictors and response. However, JSQR did offer smaller prediction error at very high quantile levels ($\geqslant 0.9$). Furthermore, JSQR offered a more nuanced view of the relationship between certain predictors and $PM_{2.5}$, especially at high quantiles, which we turn to next.

Heterogeneous predictor effect on response quantiles

To understand how predictors may impact $PM_{2.5}$ concentration heterogeneously across different quantile levels, we refitted the entire dataset with JSQR-GP1 and compared parameter estimates against those from the basic spatial random effects model which only allows making inference on mean predictor effects. To recall, the basic spatial random effects model (Cressie, 1993) is

$$Y_i = \beta_0 + X_i^\top \beta + W(s_i) + \varepsilon(s_i), \tag{5.1.1}$$

where $W(s) \sim \mathsf{GP}\left(0, \sigma_s^2 \rho(s, s'; \theta)\right)$ can be considered as spatial random effects and $\varepsilon(s) \sim \mathsf{N}(0, \sigma_e^2)$ is independent pure error. To contrast with JSQR, notice that the conditional response quantiles under BSRE are given by $Q_{Y_i}(\tau \mid X_i) = \beta_0 + \sigma_e \Phi^{-1}(\tau) + X_i^\top \beta$.

Both BSRE and JSQR detected strong spatial noise correlation. The proportion of spatial variation $\alpha = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2}$ was estimated to be 81% by BSRE and 86% by JSQR. The corresponding estimates of the decay parameter ϕ were 0.43 and 0.46. BSRE and JSQR provided similar median and interval estimates for the intercept function across all quantiles (Figure 5.2). The two models exhibited agreement to some extent on the effects of population density (LCYPOP) and elevation (ELEV). JSQR estimated the effects of these predictors to be relatively stable across all quantile levels. The signs of the regression coefficients matched expectation and were in line with existing literature (Yanosky et al., 2008b). Given other predictors in the model, the distance between monitoring station and A1 class roads (LDISTA1) was not found to have a significant effect by either model.

A difference between JSQR and BSRE emerged in the estimates of regression coefficients of the percentage of urban land use (URB) and $PM_{2.5}$ emission within 10 kilometers (L25E10). BSRE estimated the effects of both these predictors to be significant and positive. JSQR estimated URB to have a severe impact at low and

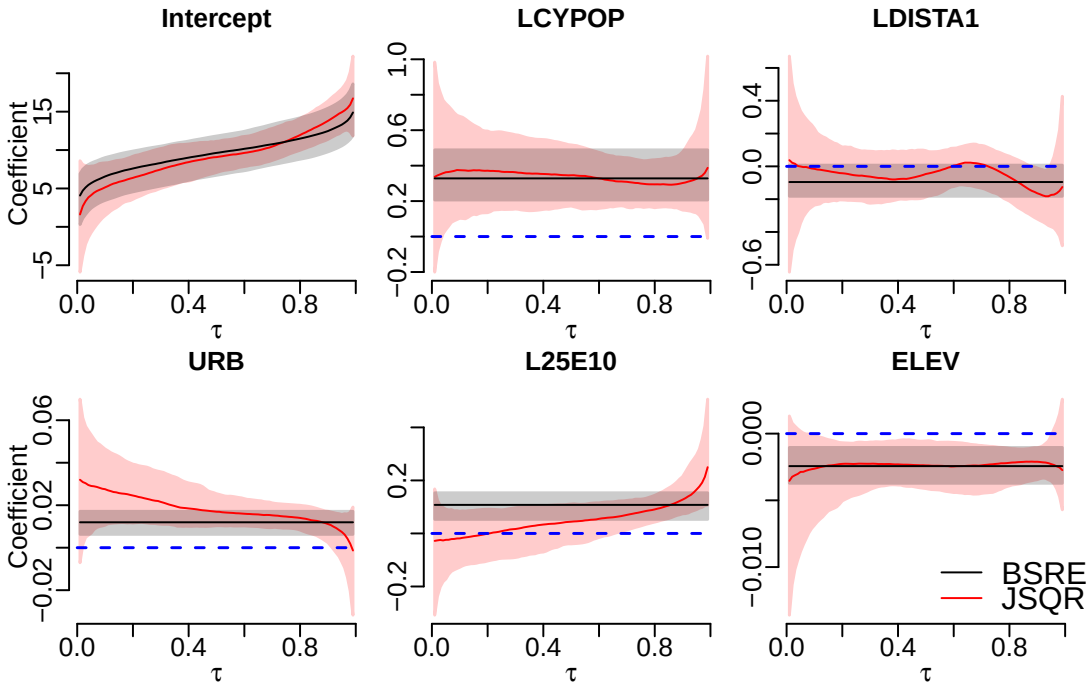

FIGURE 5.2: $PM_{2.5}$ regression coefficient estimates (and 95% bands) by BSRE and JSQR.

middle quantile levels but little effect on the high quantiles. It also estimated L25E10 to have no significant effect at low and middle quantile levels, but a fairly pronounced positive effect at high and extremely high levels. These differential effects were significant; the difference between the slopes at $\tau = 0.9$ and $\tau = 0.1$ was estimated to be -0.02 (95% credible interval $= [-0.04, 0.00]$) for URB and 0.15 ([0.03.0.29]) for L25E10.

The JSQR estimates could be interpreted as saying that while higher urban land use results in an increased baseline $PM_{2.5}$ concentration, it is the amount of local $PM_{2.5}$ emission that largely determines extreme levels of $PM_{2.5}$ concentration. At the 95th percentile level, the regression coefficient of L25E10 was estimated to be 0.17. Accordingly, a rise of $PM_{2.5}$ emission by 36 times within a 10km radius (one standard deviation increase of L25E10) would result in an increase of 0.6 $\mu g/m^3$

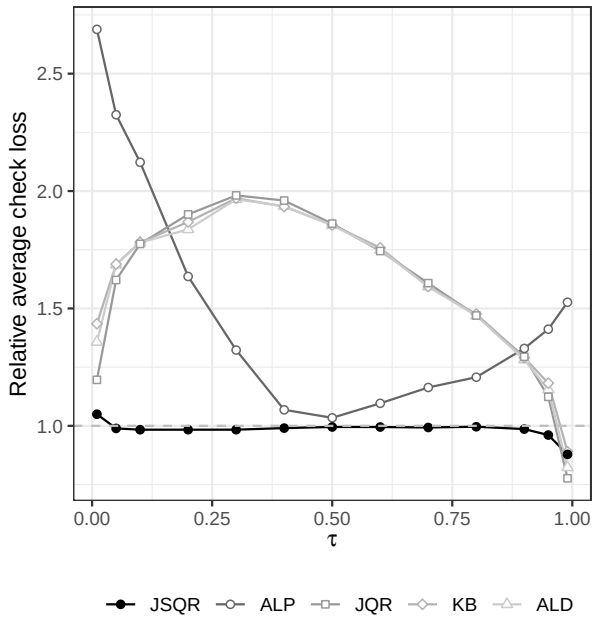

FIGURE 5.3: Average check loss given by QR methods relative to BSRE at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

of the 95th percentile of $PM_{2.5}$ concentration which corresponds to a 7% (or 4%) increase for states such as Maine, New Hampshire and Vermont with low (or, Ohio, Pennsylvania and DC with high) $PM_{2.5}$ concentrations.

5.1.2 Wildfire risk analysis

Wildfires cause significant, lasting damage to ecology (Moritz et al., 2014), economy (Butry et al., 2001) and human health (Reid et al., 2016). The burning index (BI) is a metric commonly used by National Fire Danger Rating System to monitor and forecast risk due to wildfires. The calculation of BI is sophisticated and requires measuring fuel characteristics, particle properties and moisture (Cohen and Deeming, 1985). Since collecting such measurements requires field-specific knowledge and equipment and is impossible to conduct for every location, model-based approaches are essential for analyzing and predicting BI with data that can be conveniently obtained. Kreuzer et al. (2017) propose a simultaneous autoregressive model with

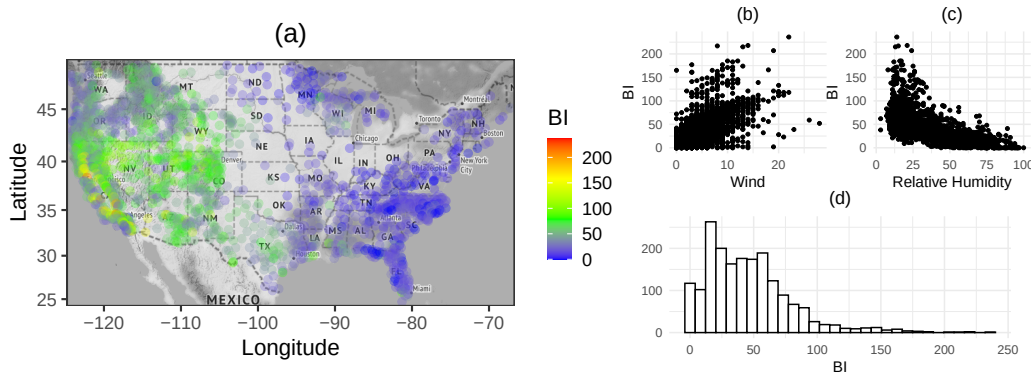

FIGURE 5.4: Burning Index measured on August 19, 2020 across contiguous US (a), and plotted against wind speed (b) and relative humidity (c). Response data appears heavy tailed (d).

t-distributed, spatially dependent random errors to analyze BI. While such a model is able to adjust for the heavy tails of the mean residuals, it fails to provide a detailed reconstructions of the extreme right tail quantiles of BI which are of direct interest because of the associated heightened risk.

We analyzed daily wildfire risk data collected by Wildland Fire Assessment System (`https://www.wfas.net/`) on Auguest 19, 2020. The data contained fire weather and fire danger observations at 1855 stations across United States (excluding Alaska and Hawaii). As expected, high values of BI were concentrated in California and the data suggested the potential existence of spatial dependence (Figure 5.4(a)). We adopted elevation and weather information as the only predictors because they are easy to collect and hence are particularly useful for surrogate BI calculation. Specifically, we included (i) elevation (ELEV), (ii) temperature (TEMP), (iii) relative humidity (RH), (iv) wind speed (WIND) and (v) precipitation (PPT) into our analysis. The focus on non-central quantiles of BI, the existence of spatial dependence, and, the heteroskedasticity of BI (Figure 5.4(b, c)) simultaneously motivated the use of JSQR models.

Similar to Section 5.1.1, we performed a 10-fold validation study with the same

set of methods. In each fold, the data was randomly partitioned into a training set and a test set with 1000 and 855 observations respectively. Among the four joint QR models, JSQR-TP had the smallest averaged WAIC score (7957), followed by JSQR-GP2 (7970), JSQR-GP1 (7976) and JQR (8482), lending support to incorporating a heavy-tailed marginal distribution and spatial tail dependence into the analysis. JSQR-TP was chosen as a representative of JSQR models for subsequent analyses.

The regression coefficient estimates provided by JSQR-TP and BSRE were substantially distinct (Figure 5.5). The signs of estimated regression coefficients given by the two models generally agreed. Elevation, relative humidity and precipitation had non-positive effects at all quantile levels, whereas temperature and wind speed had non-negative effects. JSQR-TP detected clear heterogeneous predictor effects on TEMP, RH and WIND, estimating much stronger effects on the right tail despite increased uncertainty. These variables are known to be associated with the two major components determining BI, namely the energy release which increases with higher temperature and lower humidity, and, spread which is enhanced by high winds (Cohen and Deeming, 1985). Our QR analysis suggests that these three predictors have a particularly strong impact on the severity of extreme wildfires.

Figure 5.6 compares the prediction accuracy of JSQR against other competitors (all figures shown relative to BSRE). JSQR-TP provided the smallest averaged loss consistently at all quantile levels. A key distinction from the $PM_{2.5}$ study is that the QR models provided much smaller losses than BSRE, especially for non-central quantile levels. Particularly, JSQR-TP offered 20 to 40% improvement over BSRE at both lower and upper quantile levels. The Gaussian noise assumption underlying BSRE creates a handicap in modeling skewed, heavy tailed daily BI measurement. But, even if this shortcoming were rectified with heavy tailed error distribution model, the heterogeneous predictor effect estimates offered by JSQR could not be accomplished with such extensions.

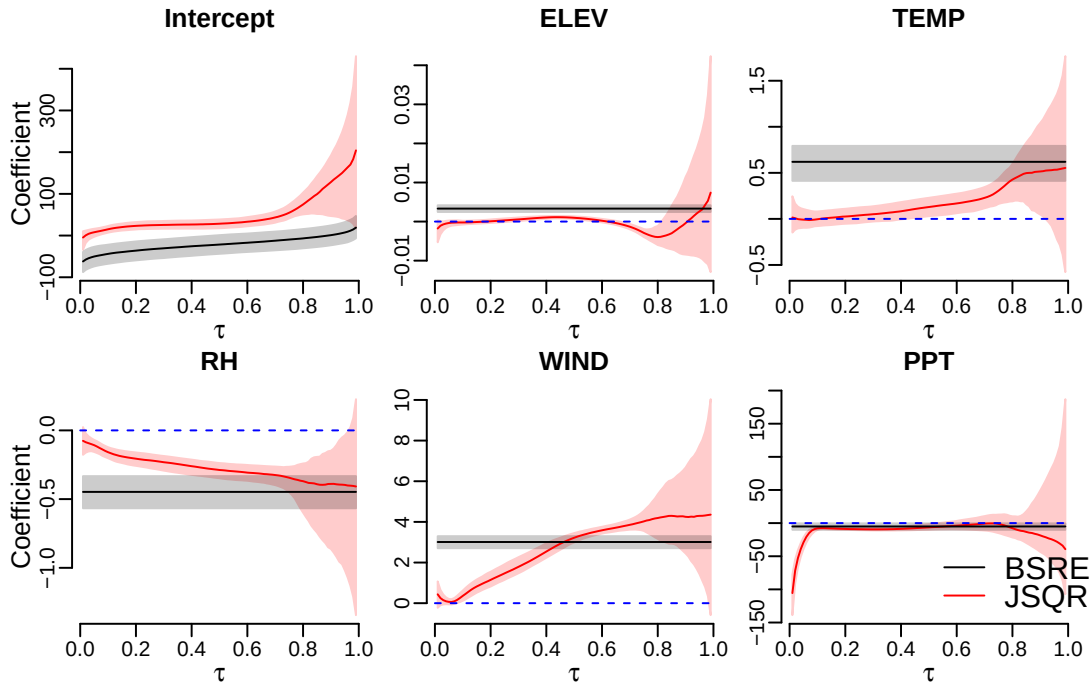

FIGURE 5.5: Wildfire risk regression coefficient estimates (and 95% bands) by BSRE and JSQR.

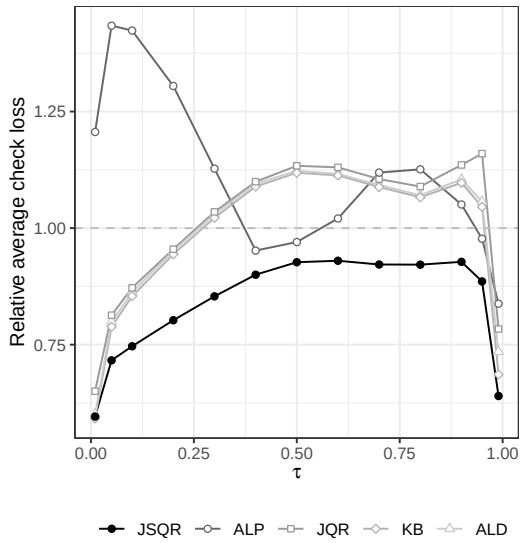

FIGURE 5.6: Average check loss given by QR methods relative to BSRE at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

5.2 Hierarchical data analysis with JHQR

5.2.1 Monitoring HIV progression with CD4$^+$ data analysis

Human immunodeficiency virus (HIV) attacks the immune system by reducing the number of vital cells of a human body, such as CD4$^+$ T cells. As the number of CD4$^+$ cells decreases with time from infection, the cell count is typically used to monitor the progression of the disease. Extensive statistical research has been conducted to make inference of the longitudinal time trajectory of CD4$^+$ cell depletion, and, to understand the associations between cell count and individual lifestyle risk and mental health factors (Zeger and Diggle, 1994; Fan and Zhang, 2000; Lin and Ying, 2001, to name a few). Although the nonparametric components in their model formulations are flexible to characterize the complex relationships between the response and predictors, the restrictive focus on conditional mean leads to inadequacy for studying varying predictor effects at tail areas of the response distribution. To overcome such inadequacy, quantile regression has been employed on CD4$^+$ datasets to analyze the non-central response-predictor associations (Lipsitz et al., 1997; Wang et al., 2009; Yuan and Yin, 2010). However, the assembled level-specific analyses lacks the strength of information sharing across quantiles.

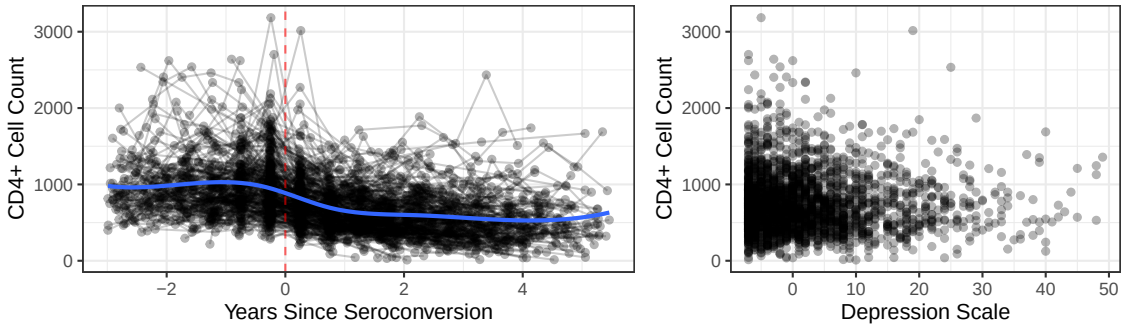

FIGURE 5.7: Data visualization of the CD4$^+$ dataset. In the left panel, the red dashed vertical line at 0 marks the time of seroconversion (time when HIV is detectable) while the blue trajectory is the fitted curve provided by linear regression with B-spline function of time with degree of freedom 6.

We reanalyzed the classical $CD4^+$ dataset (Zeger and Diggle, 1994) which contains 2376 records of $CD4^+$ cell count for 369 infected males enrolled in the Multicenter AIDS Cohort Study. We visualize the $CD4^+$ count against years since seroconversion (time when HIV is detectable) in the spaghetti plot (left panel) in Figure 5.7. It is evident that the average cell counts are approximately constant[2] before seroconversion and decrease rapidly thereafter. The dataset also contains several explanatory variables: recreational drug use (yes/no); smoking status (packs per day); number of sexual partners; depressive symptoms measured by Center for Epidemiological Studies-Depression (CESD), with high scores indicating greater depressive symptoms. As shown in the right panel of Figure 5.7, the cell counts present a decreasing trend and heteroskedasticity with respect to depression scale.

Since time might present different relationships with cell count before and after seroconversion, we augmented the time variable with two new variables respectively encoding the negative and positive parts, that is, $t^- = \max(-t, 0)$ and $t^+ = \max(t, 0)$. As the dataset contains several extreme cell counts (more than 3000, as shown in Figure 5.7), we fitted JHQR with both Gaussian and t copula. To comprehensively compare the JHQR models with different combinations of formats of the time variable, base distributions and copula models, we conducted a 10-fold validation study. In each fold, the dataset was first randomly split into 300 training subjects ($\sim 80\%$) and 68 test subjects. For training subjects, we held out the latest records of each subject. Therefore, the test dataset contained both out-of-cluster and within-cluster observations. We included either the original form or the augmented form of time in addition to other predictors. The fitted models with average WAIC scores are summarized in Table 5.1.

Benchmarked by the WAIC scores (oWAIC: 23386, iWAIC: 23367) provided by JQR (Yang and Tokdar, 2017), it is clear that accounting for the hierarchical depen-

[2] An unaffected person has around 1100 $CD4^+$ cells in blood per millilitre (Diggle et al., 2002).

dency structure greatly improves model fitting and prediction (see also Figure 5.8). In addition, we observe that the models with t base distribution-t copula pair offer smaller WAIC scores, which may be due to the existence of extreme values and tail dependence. Complying with the exploratory data analysis, augmenting time predictors is indeed crucial as it allows for different predictor effects of time before and after seroconversion. The difference between the first column and the last column of Table 5.1 suggests that the continuous temporal process provides a more accurate adjustment of the dependence by utilizing temporal distances between observations, compared to the discrete temporal process which only exploits the relative orderings of observations.

We selected M3 with t base and t copula as the representative from the JHQR category and compared with other methods in Section 3.5. We fitted all other models using the augmented time predictors with the same 10-fold datasets. The average check losses of the out-of-cluster and within-cluster predicted conditional quantiles are summarized in Figure 5.8. JHQR dominates all other methods on the within-cluster prediction performance. BQR provides consistently higher losses for both within-cluster and out-of-cluster predictions across most of quantile levels. The other three methods behave similarly. Overall, as the test data for the within-cluster prediction comes from the clusters that exist in the training data, the corresponding prediction losses are smaller than out-of-cluster prediction losses. Although KB and JQR are not able to adjust for within-cluster dependence, they offer similar out-of-cluster prediction performances as JHQR. This is because the dataset contains a large number of clusters that are conditionally independent given marginal model parameters (β_0, β) and hence the methods provide similar estimates of the regression coefficients.

Model	M2	M3	M3	M3	M3
Base-copula	t-t	log-gauss	log-gauss	t-t	t-t
Augmented time	✓	✗	✓	✗	✓
Average oWAIC	22823	22833	22838	22777	**22775**
Average iWAIC	–	22400	22396	22270	**22263**

Table 5.1: Average WAIC scores (smaller score indicates better fit) over the 10-fold validation study. t-t and log-gauss respectively denote the t base-t copula pair and logistic base-Gaussian copula pair.

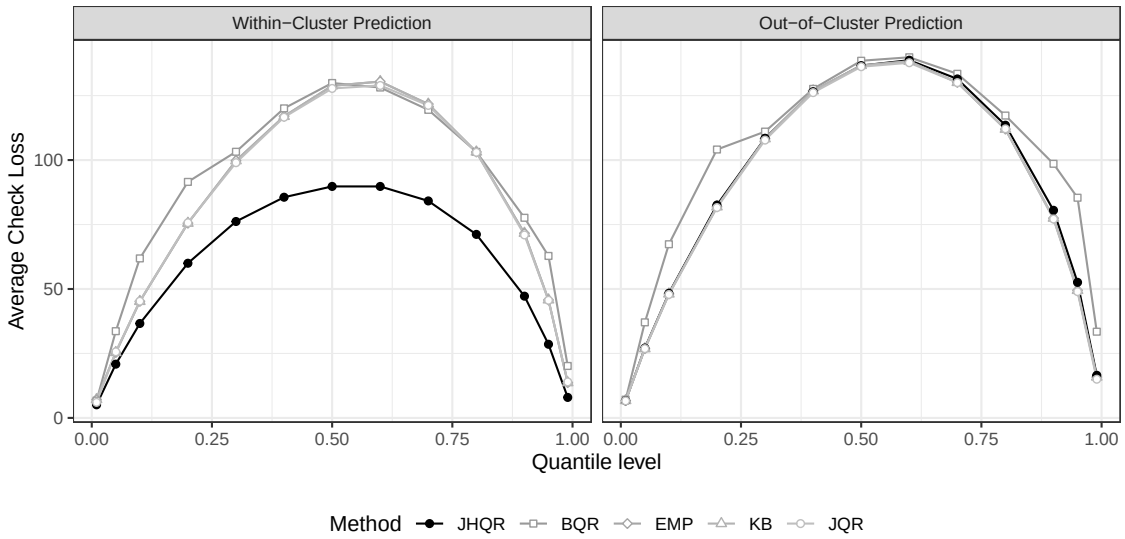

FIGURE 5.8: Average check losses of predictions provided by different methods at quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$ for the CD4$^+$ dataset.

The point and interval estimates offered by JHQR are summarized in Figure 5.9, based on the whole dataset. The monotone trend of time effect[3] suggests that the depletion of CD4$^+$ cells is more rapid initially when CD4$^+$ cell counts are large (-115 counts/year at 0.95 quantile level) and then becomes slower (-59 counts/year at 0.05 quantile level). Similarly, the depression scale also presents a greater impact at high quantile levels.

Remark 5.2.1. *The positive, increasing effect of smoking status may be seemingly*

[3] Recall $t^- = \max(-t, 0)$ and hence the positive coefficient here indicates the CD4$^+$ counts decrease with respect to time.

counterintuitive. However, inconsistent results on the association between smoking status and HIV progression exist in immunology literature (Crothers et al., 2005; Galai et al., 1997; Royce and Winkelstein Jr, 1990). For the CD4$^+$ dataset we analyzed, the sample correlation between CD4$^+$ count and smoking status is 0.25 and such positive association is also reflected on the predictor effect, which agrees with the result reported in Tang and Leng (2011). We refer interested readers to Kabali et al. (2011) and the references therein for related biological mechanisms and existing results based on other datasets.

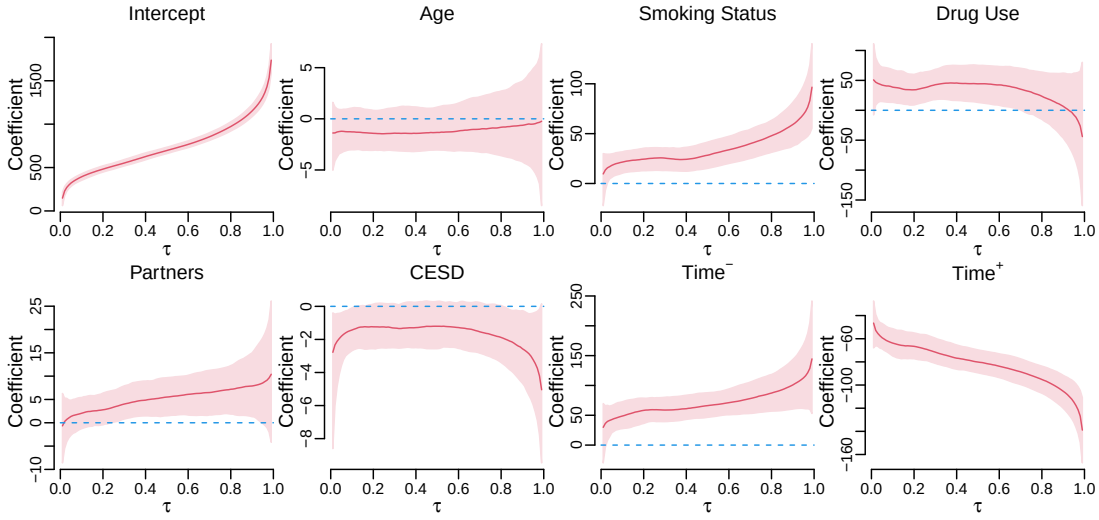

FIGURE 5.9: Point (solid red curve) and 95% credible interval (red bands) estimates of regression coefficients provided by JHQR for the CD4$^+$ dataset.

5.2.2 Improving math achievement with HS&B data analysis

The importance of math education and achievement is widely recognized since early math performance of a student is a key factor of future academic attainment (Ritchie and Bates, 2013), major selection (Wang, 2013) and career development (Mau, 2003). Therefore, identifying what and how student, family and school factors affect math achievement is crucial for improving education equity and efficiency. A vast amount

of earlier literature focuses on investigating the associations and causal relations between various characteristics of a student and corresponding academic achievement (Willms, 1985; Lee and Bryk, 1989; Ehrenberg and Brewer, 1994), using classical linear mixed effects model. Notwithstanding the popularity of linear models in developmental science, the narrow focus on mean effects overlooks the possible heterogeneous predictor effects on the outcome, depending on the quantile. However, such heterogeneous effects are important for improving personalized learning experiences for students with low or high academic achievements. In light of the limitations of classical linear models, development science has recently embraced quantile regression in data analysis, such as Petscher and Logan (2014); Simzar et al. (2015); Susperreguy et al. (2018), using the KB method. Nevertheless, the potential hierarchical dependency structure is not properly adjusted for in these analyses.

We revisited the High School and Beyond (HS&B) data analysis reported in Petscher and Logan (2014). The dataset consists of 7185 student records from 160 schools and is a subset of the large survey conducted by National Center for Education Statistics. In addition to math achievement score (MathAch), the data also contains student-level binary characteristics including Minority and Gender respectively indicating whether a student is identified as minority (yes=1) and whether a student is female (yes=1); family-level factor SES: socioeconomic status which is a composite score of parent education, occupation and income; and school-level factor Size: number of students, Sector: a binary covariate with catholic school encoded as 1 and public school encoded as 0, PrAcad: percentage of students on the academic track, DisClim: disciplinary climate score of the school, higher value indicates better climate, HighMinority: a binary covariate with 1 if minority enrollment exceeds 40%, 0 otherwise, MeanSES: average SES of students in the school[4].

We first removed 94 records with extremely low (below -2.5) or high (above 1.5)

[4] See Lee and Bryk (1989) for more details on these variables.

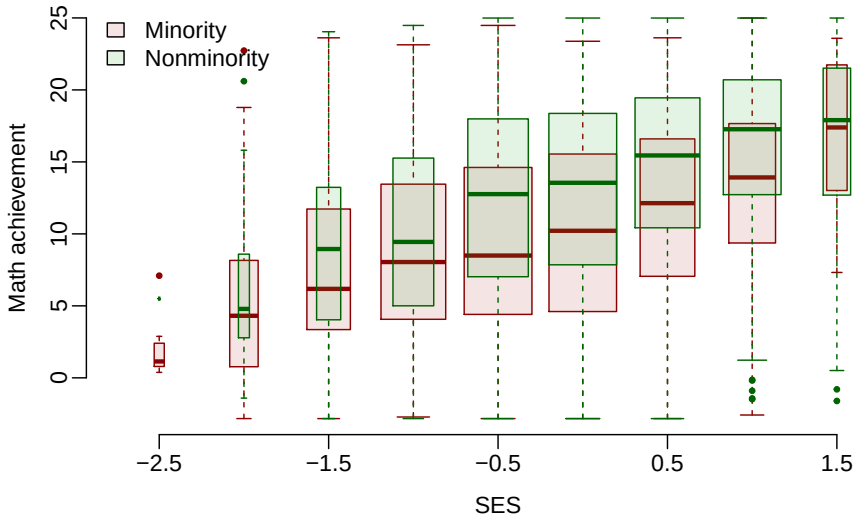

FIGURE 5.10: Boxplot of math achievement with respect to SES by ethnicity.

SES and trimmed dataset by only preserving the schools with moderate percentage (between 5% and 95%) of minority students. The remaining 4636 observations from 106 schools are visualized in Figure 5.10. Overall, MathAch presents clear marginal associations with both Minority and SES. However, the students from families with high SES (the rightmost boxes) had similar math achievements regardless of ethnicity. Also, some students from the nonminority group had low MathAch scores.

In addition to SES and Minority, whose associations with MathAch are investigated in Petscher and Logan (2014), we further included Gender and interactions between Minority and DisClim, SES, Sector into our analysis. We fitted the models in Section 3.5 and JQR with this set of predictors. We purposely excluded school-level predictors for the JHQR model to emphasize its competence on adjusting for within-school dependency and to highlight the comparison with JQR (detailed below). We also fitted JQR with an augmented predictor set by furthering including the main effects of school-level covariates: DisClim, HighMinority, MeanSES, PrAcad, Sector, Size. The corresponding model fit is denoted as JQR-S. For JHQR models, we considered both Gaussian copula and t copula. As expected, we found that they offered

similar WAIC scores and prediction performances because the distribution of MathAch (Figure 5.10) is light-tailed. Therefore, we adopted the logistic base-Gaussian copula pair for JHQR models and compared it with other methods.

Similar to $CD4^+$ data analysis, we performed a 10-fold validation study to compare different models. In each fold the training set consisted of 85 randomly selected schools and the observations in the remaining 21 schools were in the out-of-school test set. For each school in the training set, we randomly split the students into a training set and a within-school test set with 80% and 20% samples.

The prediction results are summarized in Figure 5.12. Overall, different methods perform similarly on this dataset and the difference between average losses for within-school prediction and out-of-school prediction is minor. These results suggest that the within-school dependency is not considerable, which is verified by the histogram of estimated school-specific correlation parameters in Figure 5.11. Almost all correlations between quantile levels within clusters are below 0.5. However, in terms of within-school prediction performance, we still observe that JHQR offers smallest losses at most of quantile levels. JQR-S provides slightly worse results than JHQR, but both outperform other methods.

For joint QR models, the order of their within-school prediction performances agrees with the order of their average iWAIC scores: JHQR: 18990, JQR-S: 19099 and JQR: 19190. For out-of-school prediction, JQR-S provides lower losses than other methods, but with 6 more school-level covariates in the model. The comparison between the joint QR models suggests that JHQR is capable of accounting for the (unknown) shared information within a cluster via the dependence between quantile levels, which is properly modeled using copula. JQR ignores the dependency structure and hence is incompetent to incorporate the shared information. Such deficiency may result in undesirable consequences from the perspectives of both inference and prediction. In this HS&B example, JQR provides worse within-cluster prediction

performance than JHQR. To match the prediction performance of JHQR, JQR-S needs to include the 6 school-level predictors as main effects into the model, which together serve as a useful surrogate of the shared information within schools. However, such covariates may be unavailable in other applications because it is difficult, if not impossible, to know what covariates have predictive power to the response until a statistical analysis is carried out. Exhaustive search of such covariates may lead to overfitting at best, and, suspicious predictor effects and hence misleading interpretations and conclusions at worst.

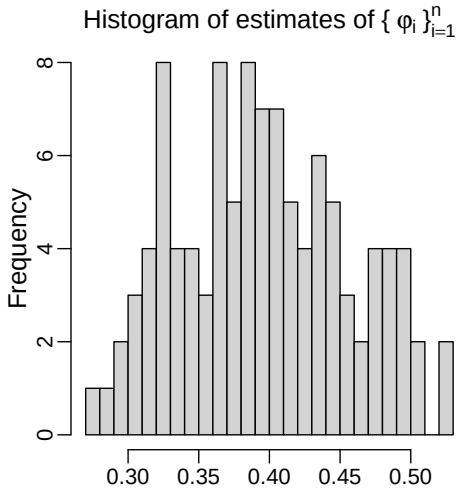

FIGURE 5.11: Histogram of posterior means of $\{\phi\}_{i=1}^{n}$ provided by JHQR, based on the whole dataset.

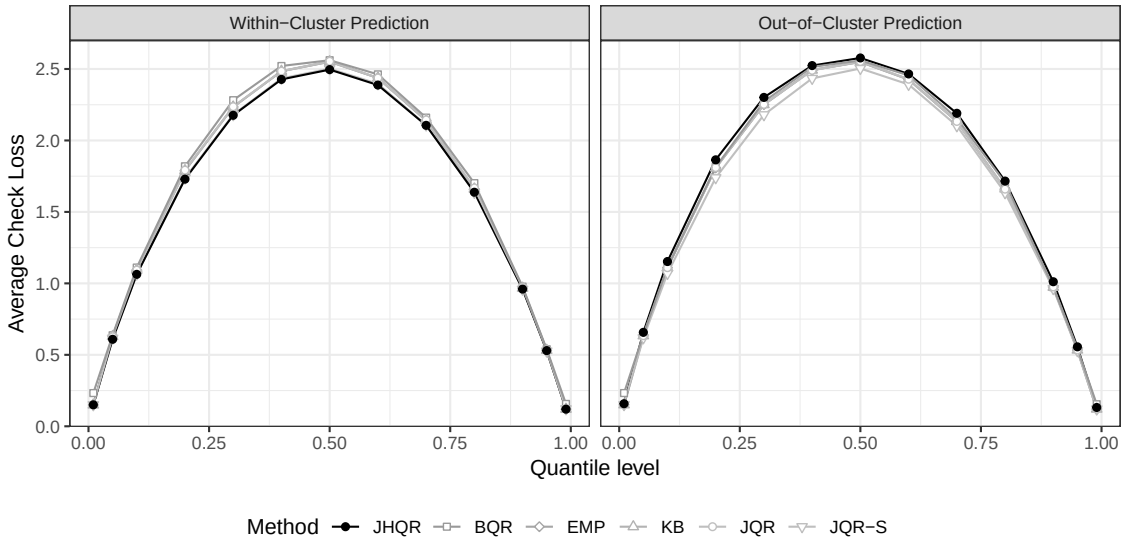

FIGURE 5.12: Average check losses of predictions provided by different methods at quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$ for the HS&B dataset.

The raw heterogeneous predictor effects given by JHQR are summarized in Figure 5.13. The effects of Minority and SES comply with the results of Petscher and Logan (2014). Particularly, the near-zero effects of Minority at high quantile levels suggest that Minority is not an effective predictor for math achievement of students with top performances. A similar interpretation applies to Gender whose effect closes to 0 at both lower and upper quantile levels. We further visualize the derived quantile effects separately for students from minority and nonminority groups in Figure 5.14. The effect of SES is consistently positive across quantile levels for nonminority students. However, we observe that SES has lesser impact for minority students at lower tails (0-0.2) of math achievement relative to nonminority students. Since our model does not include the main effects of DisClim and Sector, the corresponding derived effects are 0 for nonminority students.

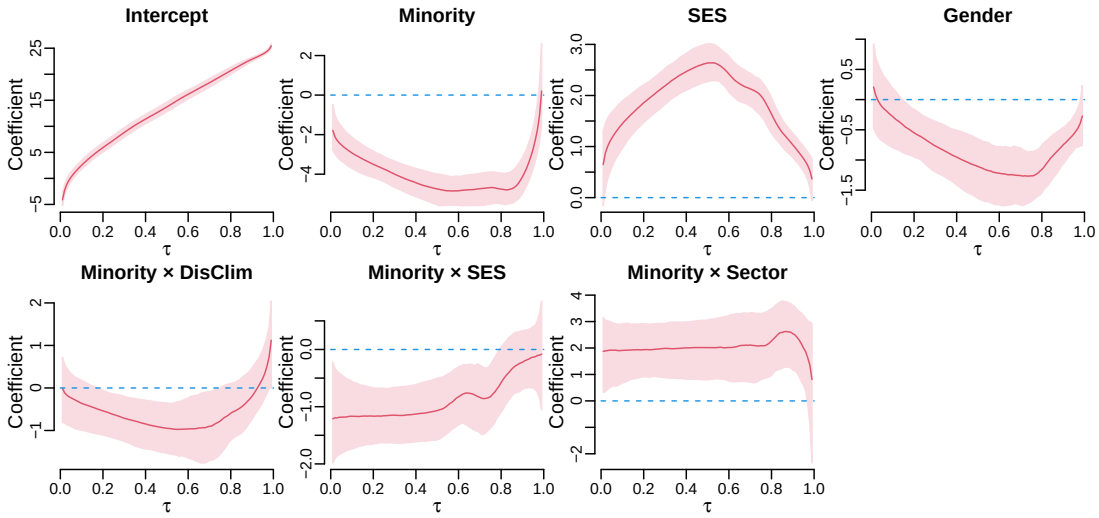

FIGURE 5.13: Point (solid curve) and 95% credible interval (bands) estimates of regression coefficients provided by JHQR for the HS&B dataset.

We can also gain useful insights from the estimated quantile levels of students (Figure 5.15) from multiple resolutions. From a coarse level of resolution, students from different schools have different dependency levels–some schools (e.g., 8193, 8367) have more concentrated quantile levels while others are more diffused (e.g., 3533, 7734). Since the bulks of the quantile levels from different schools are distributed at distinct intervals between 0 and 1, JHQR can also be viewed as a random slopes model with structurally modeled predictor effects obeying the monotonicity constraint. From a fine scale of resolution, the estimated quantile level $\hat{u}$ of a student has an intuitive and useful interpretation as follows: the math achievement of this student is approximately at $\hat{u}$ percentile among the group of students with similar characteristics that are considered in the model.

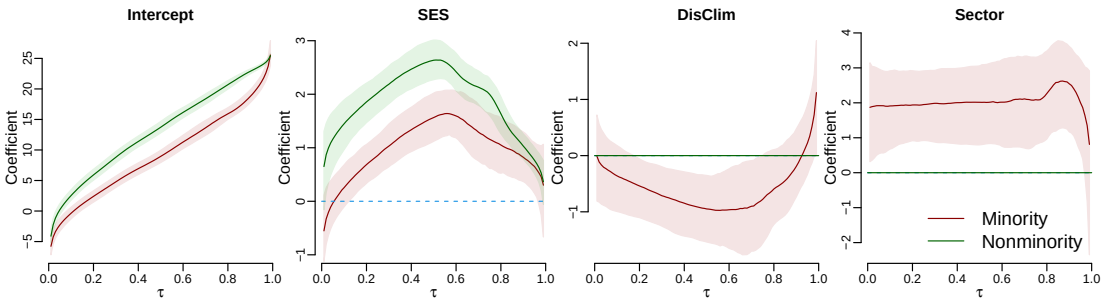

FIGURE 5.14: Derived point (solid curve) and 95% credible interval (bands) estimates of regression coefficients for students from minority and nonminority groups provided by JHQR for the HS&B dataset.

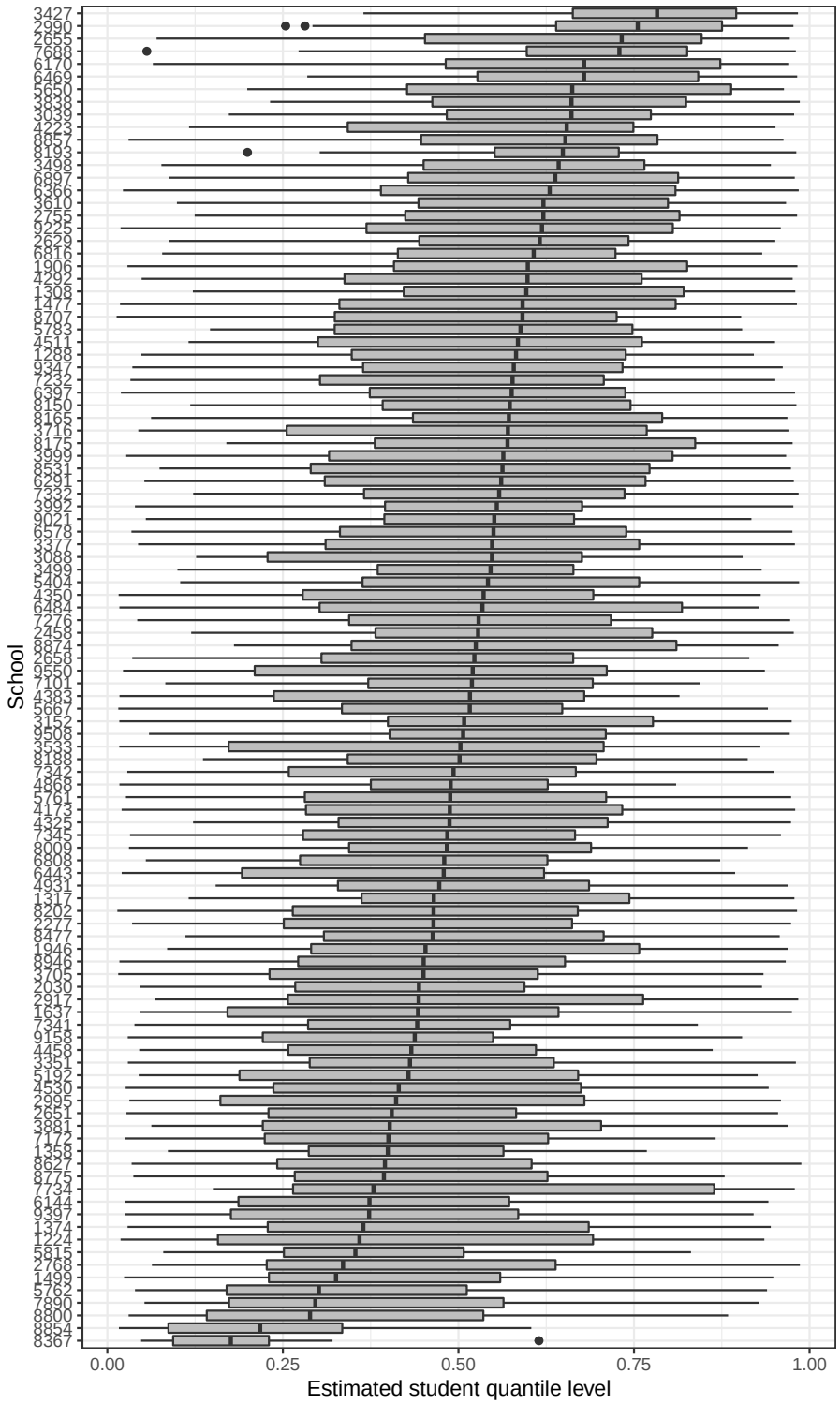

FIGURE 5.15: Boxplot of estimated quantile levels (posterior mean) of students by school. The schools are ordered according to the median of estimated quantile levels.

6

Concluding Remarks

We have introduced here a novel generalization of the joint quantile regression model of Yang and Tokdar (2017) to adjust for various dependency structures for analyzing point-referenced spatial data, cross-sectional data and longitudinal data. We have provided ample evidence that such adjustment improves parameter estimation, uncertainty quantification and prediction. As extensions of some classical models such as BSRE and random intercepts model, the proposed JSQR model offers a comprehensive and practicable solution to analyzing complex response-predictor relationships at non-central parts of the response distribution. Through several case studies we have shown that both JSQR and JHQR produce interpretable estimates of potentially heterogeneous predictor effects, offers excellent fit to hold-out data and can successfully adapt to heavy tailed response, tail dependence and tail asymmetry. An accompanying R package is under development; but a ready-to-deploy version of JSQR maybe accessed at `https://github.com/xuchenstat/JSQR`.

Our development crucially depends on accepting the linear QR model assumption (1.1.1) simultaneously for all quantile levels $\tau \in (0, 1)$. This strong assumption, a "leap of faith" according to Koenker (2017), should be given its due diligence before

adopting any joint quantile regression analysis of data. It is conceivable that the global linear QR model is adequate only when a sufficiently rich set of nonlinear transformations of the original covariates are included as predictors. In particular, any serious application with JQR, JSQR or JHQR should include diagnostics of model fit, which could be performed by assessing uniformity of the residual random levels $V_1, \ldots, V_n$ or V_{ij}'s introduced in Section 4.1.3 and 4.2.4. Additionally, iterative and interactive model improvement could be sought via addition of nonlinear transformations of original covariates as new or alternative predictors, and a formal model selection may be performed by using WAIC. See Cunningham et al. (2020) for recent development along this line.

Joint quantile regression models are a class of conditional density estimation models and are naturally related to the so called "density regression" methods which attempt to directly estimate the conditional response density given predictors (De Iorio et al., 2004; Dunson et al., 2007; Tokdar et al., 2010). Clearly, one could obtain estimates of conditional quantiles under a density regression approach, just as one could easily obtain estimates of conditional densities under a joint QR model. Which approach is selected for a particular data analysis task depends primarily on the goals of the analysis. A major appeal of the joint QR approach is interpretability. The estimates of the slope functions give a clear picture of potentially heterogeneous predictor effects, which may be vital in a careful scientific analysis of predictor-response relationship (Cade and Noon, 2003; Abrevaya and Dahl, 2008; Wasko and Sharma, 2014). Density regression, on the other hand, offers more flexible quantile shape adjustment which could be more useful for applications focused on prediction. However, even for pure prediction purposes, joint QR, via nonlinear transformations of covariates, may be competitive against density regression methods which often struggle to account for sharp changes in the spread and skewness of the conditional densities, a hallmark of extremely heterogeneous predictor effects; see Tokdar (2013). Other than

density regression, joint quantile regression is also a member of the distributional regression family, which posits distributional statistics of the response as functions of predictors (Foresi and Peracchi, 1995; Firpo et al., 2009; Klein et al., 2015, to name a few). The choice of statistics substantially determines flexibility and interpretation of the model. As our focus is on quantile regression, direct comparison with other distributional regression models is out of the scope of this work; see Koenker et al. (2013) for an interesting discussion.

Our results here are of empirical nature. A rigorous mathematical treatment of asymptotic properties of JSQR and JHQR remains the subject of a future work. The assumption of dependence between observations makes it quite challenging to study posterior consistency properties of the proposed Bayesian method. However, the conditional independence implied by the formulation (2.4.2) and (4.2.3) offers a path to addressing this challenge under the asymptotic settings where the latent process W (or the cluster-specific level W_i's for JHQR) is an additional parameter to be estimated along with $\beta_0(\cdot)$ and $\beta(\cdot)$.

Finally, we note that extensions similar to JSQR and JHQR could be sought to address a variety of other dependency structures within a joint quantile regression model, with applications to time series data or data where the response is vector valued. Such developments crucially depend on the choice of the copula families that adequately address relevant aspects of statistical estimation and computing.

Appendix A

Supplemental material for Chapter 2

A.1 Algorithm configurations

The algorithm configurations are as follows. The base density function f_0 was specified to be the pdf of standard logistic distribution for both JQR and JSQR. We followed the prior specifications for JSQR in Section 2.2.2 with smoothness parameter $\nu = 2$ and employed 10 discrete values for decay parameter ϕ. ALD was implemented with default non-informative priors (Benoit et al., 2017). We implemented ALP according to the model fitting procedure described in Section 5 and Appendix A in Lum and Gelfand (2012). We found that their method is very sensitive to the prior choice. In all numerical experiments, a $\mathsf{Be}(1, 6)$ prior was assigned for the proportion of spatial variation α instead of uniform prior stated in their paper to achieve better estimation accuracy. The underlying Gaussian process configuration was kept same as JSQR.

KB, JQR and ALD were implemented using R packages quantreg, qrjoint and bayesQR respectively. JSQR and ALP were implemented in C with R wrapper. We ran all Bayesian methods for 20,000 MCMC iterations with first 10,000 samples as

burn-in. We thinned the remaining samples and preserved 500 samples for posterior inference. Except for ALP, we tracked the estimates over a grid of τ from 0 to 1 with increment 0.01. We use the simulation M2 to illustrate the computing time for each algorithm to analyze one dataset. On a desktop with CPU Intel(R) Core(TM) i7-4790 3.60GHz, JSQR, JQR and ALD took 6.3, 2.6 and 9.2 minutes respectively. KB took only 0.07 seconds. ALP took 36 minutes to perform an analysis at one quantile level and hence took 468 minutes in total for the quantile levels used for summary.

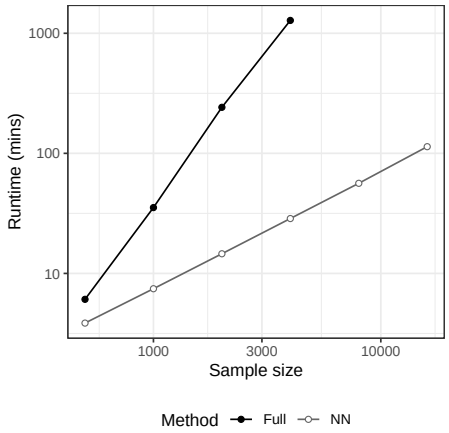

FIGURE A.1: Running time comparison between JSQR with full Gaussian copula process and with 5-NN Gaussian copula process. Each point represents the average over 10 independent runs.

A.2 JSQR with NN Gaussian copula process

A simulation study is presented to showcase the computational gain and prediction accuracy given by our model with NN Gaussian copula process. NNGP is implemented by using the Pseudocode 2 and 3 detailed in Finley et al. (2019). Both inference and prediction have $\mathcal{O}(nm^3)$ computational complexity, where m is the number of neighbors adopted in NNGP. The simulation setup is same as M2 in Section 2.5.1 with varying sample sizes from 500 to 16,000. Using a single CPU, we run our model with full GP and 5-NNGP with 20,000 MCMC iterations. According to average running times presented in Figure A.1, a substantial computational gain is obtained by adopting NN Gaussian copula process. The results are consistent with our statement of computational complexities of two algorithms in Section 2.3.3.

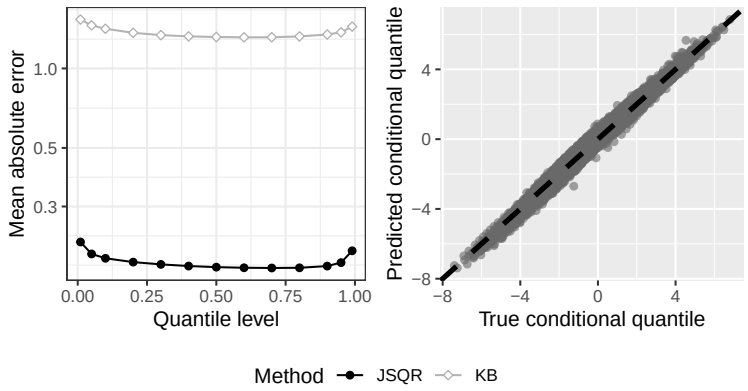

FIGURE A.2: Prediction performance of JSQR with 5-NN Gaussian copula process. Left: mean absolute errors of predicted conditional quantiles; right: true conditional quantiles against predicted conditional quantiles.

The prediction accuracy of JSQR with NN Gaussian copula process is examined under the same setting with $16,000$ and 50 samples in the training set and test set respectively. With 10 replications of datasets, we calculate the mean absolute errors of predicted conditional quantiles across a grid of quantile levels. Here we only compare our model with 5-NN Gaussian copula process against KB (left panel in Figure A.2) because KB, ALD and JQR have similar prediction performances and ALP requires expensive computations. Combining the comparison between the predicted conditional quantiles and true conditional quantiles (right panel in Figure A.2), it is clear that our model with NN Gaussian copula process is able to provide accurate spatial prediction and outperforms other methods.

A.3 Simulation results of Example 2

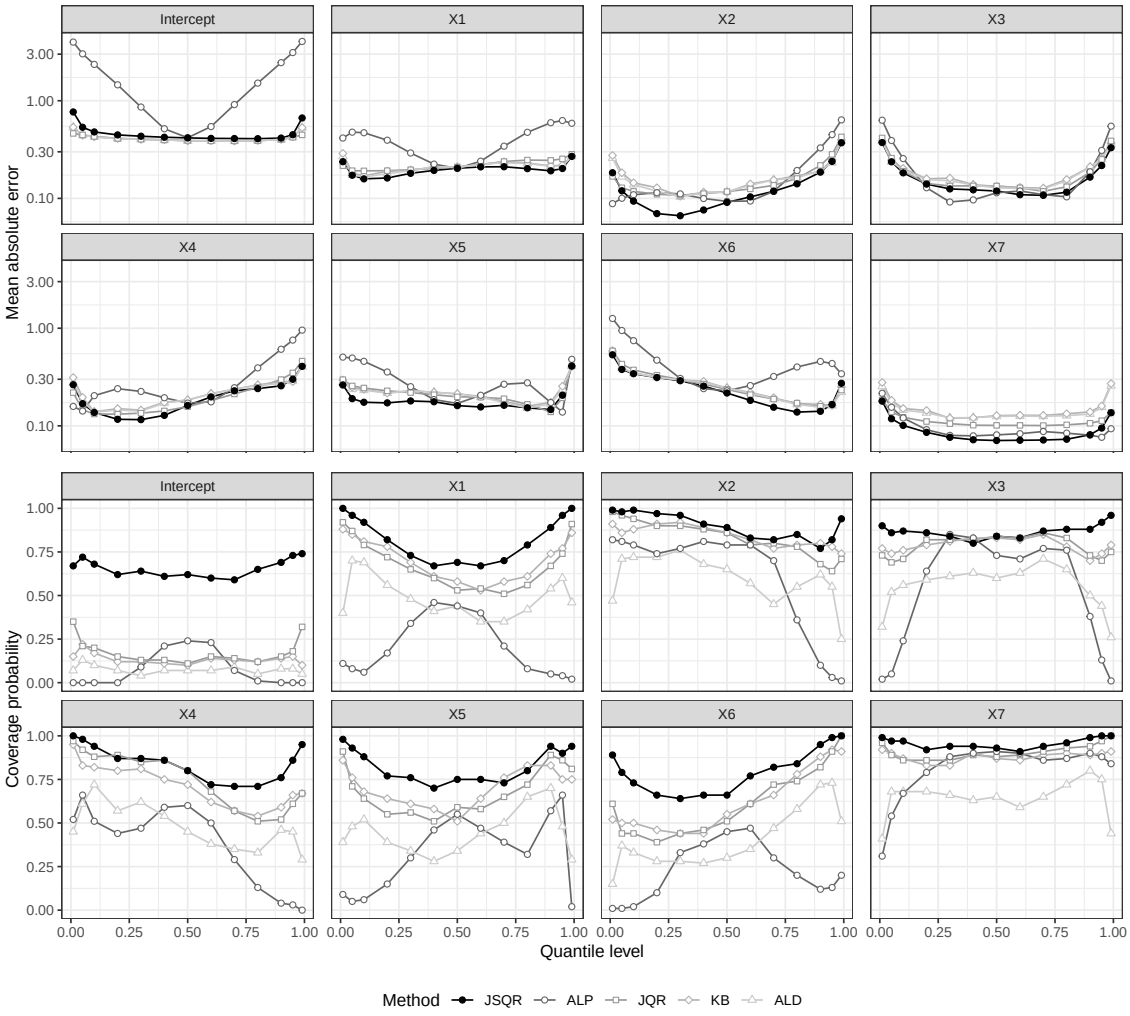

FIGURE A.3: Inference efficiency of different methods for Example 2 with true generating process being Gaussian process. The top two rows present the mean absolute errors of regression coefficients while the bottom two rows show the coverage probabilities of 95% confidence (or credible) intervals of regression coefficients at $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

Appendix B

Supplemental material for Chapter 3

B.1 Simulation results for S2 and S3

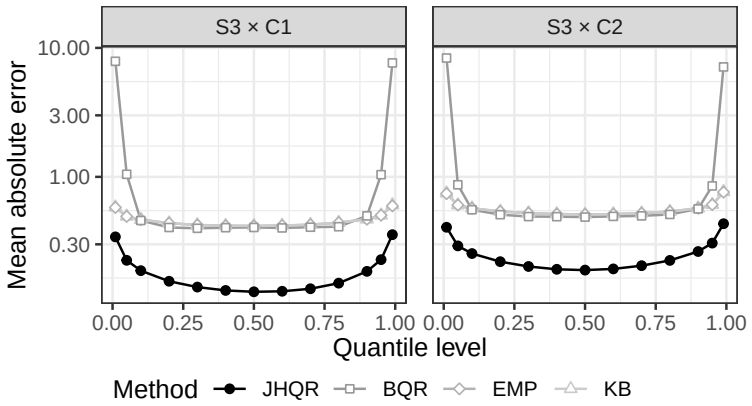

FIGURE B.1: Within-cluster prediction accuracy of different methods for S3×C1 (left panel) and S3×C2 (right panel). Mean absolute errors of predicted conditional quantiles are presented over quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

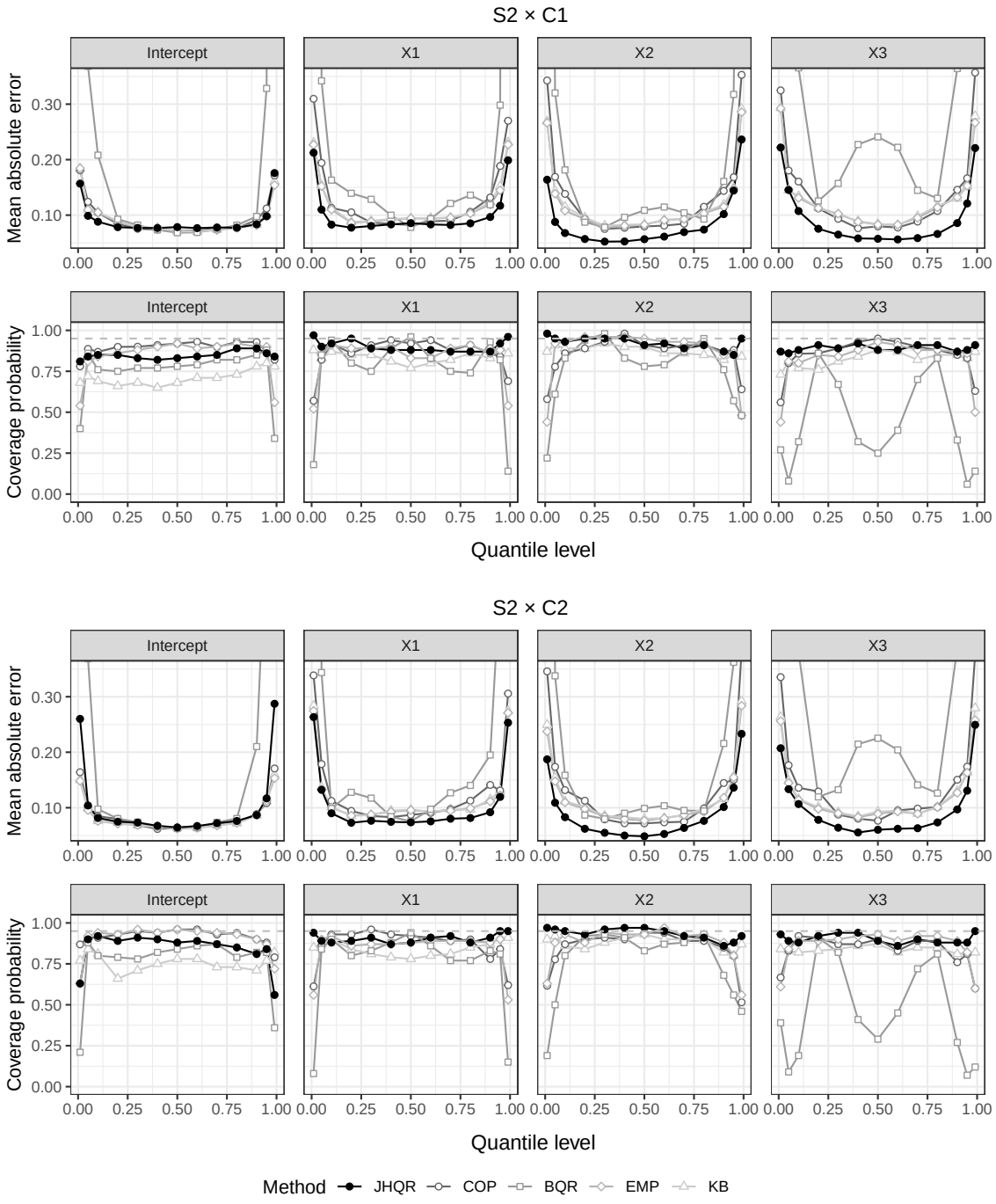

FIGURE B.2: Estimation quality of different methods for S2×C1 (top panel) and S2×C2 (bottom panel). In each panel, mean absolute errors of point estimates of regression coefficients are presented in the top row, and, coverage probabilities of 95% confidence (credible) intervals of regression coefficients are presented in the bottom row. The evaluations are performed over quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

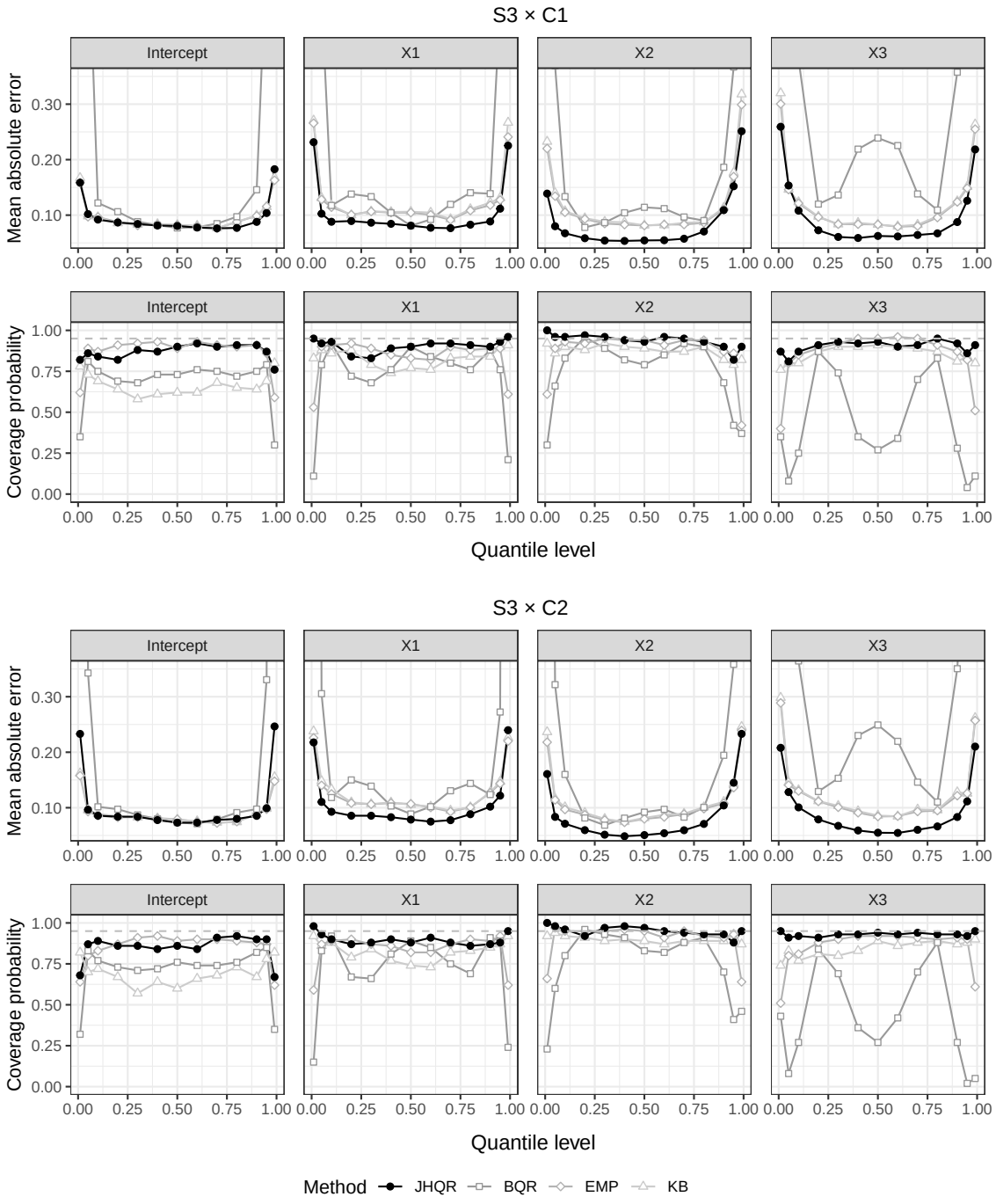

FIGURE B.3: Estimation quality of different methods for S3×C1 (top panel) and S3×C2 (bottom panel). In each panel, mean absolute errors of point estimates of regression coefficients are presented in the top row, and, coverage probabilities of 95% confidence (credible) intervals of regression coefficients are presented in the bottom row. The evaluations are performed over quantile levels $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

B.2 Density and quantile functions of Gumbel and Clayton copula

The bivariate Clayton and Gumbel copula functions $C(u, v \mid \phi)$ are respectively given by

$$\text{Clayton:} \quad \left(u^{-\phi} + v^{-\phi} - 1\right)^{-1/\phi}$$

$$\text{Gumbel:} \quad \exp\left\{-\left[(-\log u)^{\phi} + (-\log v)^{\phi}\right]^{1/\phi}\right\}.$$

The associated conditional copula $C(u \mid v, \phi) = \partial C(u, v \mid \phi)/\partial v$ are

$$\text{Clayton:} \quad v^{-\phi-1}\left(u^{-\phi} + v^{-\phi} - 1\right)^{-1/\phi-1}$$

$$\text{Gumbel:} \quad v^{-1}(-\log v)^{\phi-1}\left[(-\log u)^{\phi} + (-\log v)^{\phi}\right]^{1/\phi-1} \times \exp\left\{-\left[(-\log u)^{\phi} + (-\log v)^{\phi}\right]^{1/\phi}\right\}.$$

Therefore, a random variable pair (u, v) can be drawn from a Clayton copula by inverting the conditional CDF as follows. First draw $v, w \overset{\text{iid}}{\sim} \mathsf{Unif}(0, 1)$ and then solve $w = C(u \mid v, \phi)$ for u with

$$u = C^{-1}(w \mid v, \phi) = \left(w^{-\frac{\phi}{\phi+1}} v^{-\phi} - v^{-\phi} + 1\right)^{-1/\phi}.$$

Since $C^{-1}(w \mid v, \phi)$ is not analytically available for Gumbel copula, we need to numerically solve $w = C(u \mid v, \phi)$ for u.

The density functions $c(u, v \mid \phi) = \partial^2 C(u, v \mid \phi)/\partial u \partial v$ used for log-likelihood calculation are

$$\text{Clayton:} \quad (1+\phi)u^{-1-\phi}v^{-1-\phi}\left(-1 + u^{-\phi} + v^{-\phi}\right)^{-2-1/\phi}$$

$$\text{Gumbel:} \quad u^{-1}v^{-1}\left(-\log u\right)^{\phi-1}\left(-\log v\right)^{\phi-1}\exp\left\{-\left((-\log u)^{\phi} + (-\log v)^{\phi}\right)^{1/\phi}\right\} \times \left(\left((-\log u)^{\phi} + (-\log v)^{\phi}\right)^{1/\phi} + \phi - 1\right)\left((-\log u)^{\phi} + (-\log v)^{\phi}\right)^{1/\phi-2}.$$

Appendix C

Supplemental material for Chapter 4

C.1 Connection between conditional WAIC and Bayesian leave-one-out CV

We establish the connection between the conditional WAIC and Bayesian leave-one-out cross validation by showing that the Taylor's expansions of the two quantities agree in the leading 3 terms. The proof consists of two parts: (i) rewriting the Bayesian leave-one-out cross validation using importance weighting, and, (ii) calculating and comparing the leading terms of the Taylor's expansions of the two quantities. Part (i) and (ii) adapt respectively from the derivations in Section 4.1 and 4.2 of Li et al. (2016) and Watanabe (2010).

Part (i): Let $\mathcal{D} = \{(X_i, Y_i, s_i)\}_{i=1}^n$ denote the dataset. The Bayesian leave-one-out cross validation of the log predictive density (Geisser and Eddy, 1979; Gneiting and Raftery, 2007) is given by

$$\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid x_i, s_i, \mathcal{D}_{-i}) \tag{C.1.1}$$

where $\mathcal{D}_{-i}$ denotes the set of all observations except (X_i, Y_i, s_i). Our model contains

marginal parameters $\xi = (\beta_0, \beta)$ and copula parameters $\theta = (\alpha, \phi)$. In the following derivations, spatial locations s_i are omitted for brevity. According to our model structure, we have,

$$p(y_i \mid x_i, \mathcal{D}_{-i}) = \int p(y_i \mid x_i, \theta, \xi, W_i, \mathcal{D}_{-i}) p(W_i, \theta, \xi \mid \mathcal{D}_{-i}) \mathrm{d}W_i \mathrm{d}\theta \mathrm{d}\xi \tag{C.1.2}$$

Using the idea of importance sampling (Gelfand et al., 1992; Gelfand, 1996), we have

$$(\text{C.1.2}) = \frac{\int p(y_i \mid x_i, \theta, \xi, W_i) \frac{p(W_i, \theta, \xi \mid \mathcal{D}_{-i})}{p(W_i, \theta, \xi \mid \mathcal{D})} p(W_i, \theta, \xi \mid \mathcal{D}) \mathrm{d}W_i \mathrm{d}\theta \mathrm{d}\xi}{\int \frac{p(W_i, \theta, \xi \mid \mathcal{D}_{-i})}{p(W_i, \theta, \xi \mid \mathcal{D})} p(W_i, \theta, \xi \mid \mathcal{D}) \mathrm{d}W_i \mathrm{d}\theta \mathrm{d}\xi}$$

Note that

$$p(W_i, \theta, \xi \mid \mathcal{D}_{-i}) = \int \frac{1}{C_1} \prod_{j \neq i}^{n} p(y_j \mid x_j, \theta, \xi, W_j) p(W_{1:n} \mid \theta) p(\theta, \xi) \mathrm{d}W_{-i}$$

where C_1 is a normalizing constant only involving $\mathcal{D}_{-i}$. Similarly,

$$p(W_i, \theta, \xi \mid \mathcal{D}) = \int \frac{1}{C_2} \prod_{j=1}^{n} p(y_j \mid x_j, \theta, \xi, W_j) p(W_{1:n} \mid \theta) p(\theta, \xi) \mathrm{d}W_{-i}$$

where C_2 is a normalizing constant only involving $\mathcal{D}$. Their ratio is

$$\frac{p(W_i, \theta, \xi \mid \mathcal{D}_{-i})}{p(W_i, \theta, \xi \mid \mathcal{D})} = \frac{C_2}{C_1} \frac{1}{p(y_i \mid x_i, \theta, \xi, W_i)}$$

Therefore,

$$\begin{aligned}
(\text{C.1.2}) &= \frac{\int p(y_i \mid x_i, \theta, \xi, W_i) \frac{C_2}{C_1} \frac{1}{p(y_i \mid x_i, \theta, \xi, W_i)} p(W_i, \theta, \xi \mid \mathcal{D}) \mathrm{d}W_i \mathrm{d}\theta \mathrm{d}\xi}{\int \frac{C_2}{C_1} \frac{1}{p(y_i \mid x_i, \theta, \xi, W_i)} p(W_i, \theta, \xi \mid \mathcal{D}) \mathrm{d}W_i \mathrm{d}\theta \mathrm{d}\xi} \\
&= \frac{1}{\int \frac{1}{p(y_i \mid x_i, \theta, \xi, W_i)} p(W_i, \theta, \xi \mid \mathcal{D}) \mathrm{d}W_i \mathrm{d}\theta \mathrm{d}\xi} \\
\log(\text{C.1.2}) &= -\log \mathsf{E}_{W_i, \theta, \xi \mid \mathcal{D}}[p(y_i \mid x_i, \theta, \xi, W_i)^{-1}]
\end{aligned}$$

Note that log of $p(y_i \mid x_i, \theta, \xi, W_i)$ is precisely given by equation (6.2).

Part (ii): The conditional WAIC for the ith observation (up to a constant) is given by

$$\text{WAIC}_i = \log \mathsf{E}_{W_i,\theta,\xi|\mathcal{D}}[p(y_i \mid x_i, \theta, \xi, W_i)] - \mathsf{Var}_{W_i,\theta,\xi|\mathcal{D}}[\log p(y_i \mid x_i, \theta, \xi, W_i)]$$

Let

$$F(k) = -\log \mathsf{E}_{W_i,\theta,\xi|\mathcal{D}}[p(y_i \mid x_i, \theta, \xi, W_i)^k]$$

Then $F(-1) = \log p(y_i \mid x_i, \mathcal{D}_{-i})$. Using the Taylor's expansion of $F(k)$ at $k = 0$, we have

$$F(-1) = F(0) - F'(0) + \frac{F''(0)}{2} - \frac{F'''(0)}{6} + \sum_{i=4}^{\infty} \frac{(-1)^i F^{(i)}(0)}{i!} \tag{C.1.3}$$

$$F(1) = F(0) + F'(0) + \frac{F''(0)}{2} + \frac{F'''(0)}{6} + \sum_{i=4}^{\infty} \frac{F^{(i)}(0)}{i!}$$

Clearly, $F(0) = 0, F(1) = -\log \mathsf{E}_{W_i,\theta,\xi|\mathcal{D}}[p(y_i \mid x_i, \theta, \xi, W_i)]$. Also, it is not difficult to verify that

$$F'(0) = -\mathsf{E}_{W_i,\theta,\xi|\mathcal{D}}[\log p(y_i \mid x_i, \theta, \xi, W_i)]$$
$$F''(0) = -\mathsf{Var}_{W_i,\theta,\xi|\mathcal{D}}[\log p(y_i \mid x_i, \theta, \xi, W_i)]$$

Therefore,

$$\text{WAIC}_i = -F(1) + F''(0) = -F'(0) + \frac{1}{2}F''(0) - \frac{F'''(0)}{6} - \sum_{i=4}^{\infty} \frac{F^{(i)}(0)}{i!} \tag{C.1.4}$$

Comparing equations (C.1.3) and (C.1.4), we prove that the Taylor's expansions of the Bayesian leave-one-out cross validation and conditional WAIC agree in the leading 3 terms.

C.2 Spatial smoothing and WAIC calculation with t copula process

C.2.1 Conditional quantile function under the t copula process

The conditional t copula can be evaluated using conditional t distribution. Let s^* be a new location. Define $Z = \left(T_\psi^{-1}(U_1), \ldots, T_\psi^{-1}(U_n)\right)^\top$. Let K be a $n \times n$ matrix with $K_{ij} = \rho(s_i, s_j; (\nu, \phi))$ and K^* be a n-dimensional vector with $K_i^* = \rho(s^*, s_i; (\nu, \phi))$. Based on model specification (4.1.1), we have

$$Z(s^*) \mid Z, \theta \sim t_1\left(\psi + n, \mu(s^*), \frac{\psi + d}{\psi + n}\sigma^2(s^*)\right)$$

where

$$\begin{aligned} d &= Z^\top(\alpha K + (1-\alpha)I_n)^{-1}Z \\ \mu(s^*) &= \alpha K^{*\top}(\alpha K + (1-\alpha)I_n)^{-1}Z \\ \sigma^2(s^*) &= 1 - \alpha^2 K^{*\top}(\alpha K + (1-\alpha)I_n)^{-1}K^* \end{aligned}$$

Therefore, the conditional τ^*th quantile of response Y^* given predictors X^* at location s^* is

$$\begin{aligned} Q_{Y^*}(\tau^* \mid X^*, s^*, U_1, \ldots, U_n) &= \beta_0(\tau) + X^{*\top}\beta(\tau) \\ \tau &= T_\psi\left(\mu(s^*) + \sqrt{\frac{\psi + d}{\psi + n}}\sigma(s^*)T_{\psi+n}^{-1}(\tau^*)\right) \end{aligned}$$

C.2.2 Log-likelihood score with $W(s)$ and φ as additional model parameters for t copula process

The JSQR model with t copula process is given by

$$\begin{aligned} Y_i &= \beta_0(U_i) + X_i^\top\beta(U_i) \\ U_i &= T_\psi(Z(s_i)),\ Z(s_i) = W(s_i) + \varepsilon(s_i) \\ W(s) &\sim \mathsf{GP}(0, \alpha\rho(s, s'; (\nu, \phi))/\varphi),\ \varepsilon(s_i) \overset{\text{iid}}{\sim} \mathsf{N}(0, (1-\alpha)/\varphi) \\ \varphi &\sim \mathsf{Gamma}(\psi/2, \psi/2) \end{aligned}$$

Let $V_i = \Phi\left(\sqrt{\varphi/(1-\alpha)}\varepsilon(s_i)\right)$ and write the model conditioning on $W(s)$ and φ

$$Y_i = \beta_0\left(h_{(W,\alpha,\varphi,\psi)}(s_i,V_i)\right) + X_i^\top \beta\left(h_{(W,\alpha,\varphi,\psi)}(s_i,V_i)\right),\ V_i \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1), 1 \leqslant i \leqslant n$$

where $h_{(w,a,\varphi,\psi)}(s,t) = T_\psi\left(w(s) + \sqrt{(1-\alpha)/\varphi}\Phi^{-1}(t)\right)$. Therefore, according to (3.3.1), the log-likelihood for ith observation with $W(s)$ and φ as additional parameters is

$$-\log\left\{\dot\beta_0(U_i) + X_i^\top \dot\beta(U_i)\right\} - \log \dot h_{(W,\alpha,\varphi,\psi)}(s_i,V_i) \tag{C.2.1}$$

where

$$\dot h_{(W,\alpha,\varphi,\psi)}(s_i,V_i) = \frac{\sqrt{(1-\alpha)/\varphi}\, t_\psi(Z(s_i))}{\phi(\Phi^{-1}(V_i))}.$$

The latent parameter φ and the process realization $W(s)$ can be recovered according to their respective posteriors

$$\varphi \mid Z(s),\theta,\psi \sim \mathsf{Gamma}\left(\frac{\psi+n}{2}, \frac{\psi + Z(s)^\top(\alpha K + (1-\alpha)I_n)^{-1}Z(s)}{2}\right)$$

$$W(s) \mid Z(s),\theta,\varphi,\psi \sim \mathsf{N}\left(\frac{1}{1-\alpha}\left(\frac{1}{\alpha}K^{-1} + \frac{1}{1-\alpha}I_n\right)^{-1} Z(s), \frac{1}{\varphi}\left(\frac{1}{\alpha}K^{-1} + \frac{1}{1-\alpha}I_n\right)^{-1}\right)$$

where K is a $n \times n$ matrix with $K_{ij} = \rho(s_i,s_j;(\nu,\phi))$. We then have

$$V_i = \Phi(\sqrt{\varphi/(1-\alpha)}(Z(s_i) - W(s_i))).$$

C.3 Log-likelihood computation associated with t copula

Using the Gaussian mixture representation, the JHQR model with t-copula for cross-sectional data can be written as

$$Y_{ij} = \beta_0(h_{W_i,\phi_i,\varphi_i,\eta}(V_{ij})) + X_{ij}^\top \beta(h_{W_i,\phi_i,\varphi_i,\eta}(V_{ij})),\ V_{ij} \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1)$$

where $h_{w,\phi,\varphi,\eta}(v) = T_\eta(w + \sqrt{(1-\phi)/\varphi}\Phi^{-1}(v))$ with $(w,\phi,\varphi,\eta,v) \in \mathbb{R} \times (0,1) \times \mathbb{R}^+ \times \mathbb{R}^+ \times (0,1)$. The corresponding conditional log-likelihood is

$$-\log\left\{\dot{\beta}_0(h_{W_i,\phi_i,\varphi_i,\eta}(V_{ij})) + X_{ij}^\top \dot{\beta}(h_{W_i,\phi_i,\varphi_i,\eta}(V_{ij}))\right\} - \log \dot{h}_{W_i,\phi_i,\varphi_i,\eta}(V_{ij}) \qquad \text{(C.3.1)}$$

where

$$\dot{h}_{W_i,\phi_i,\varphi_i,\eta}(V_{ij}) = \frac{\sqrt{(1-\phi_i)/\varphi_i}\, t_\eta(Z_{ij})}{\phi(\Phi^{-1}(V_{ij}))}.$$

The latent level W_i and the latent parameter φ_i can be recovered according to their respective full conditional posteriors

$$\varphi_i \mid Z_i, \phi_i, \eta \sim \mathsf{Gamma}\left(\frac{\eta + n_i}{2}, \frac{\eta + \frac{1}{1-\phi_i}\left(Z_i^\top Z_i - \frac{\phi_i (Z_i^\top 1_{n_i})^2}{1+(n_i-1)\phi_i}\right)}{2}\right)$$

$$W_i \mid Z_i, \phi_i, \varphi_i, \eta \sim \mathsf{N}\left(\frac{\phi_i 1_{n_i}^\top Z_i}{1+(n_i-1)\phi_i}, \frac{\phi_i(1-\phi_i)}{(1+(n_i-1)\phi_i)\varphi_i}\right)$$

and then $V_{ij} = \Phi\left(\sqrt{\varphi_i/(1-\phi_i)}(Z_{ij} - W_i)\right)$.

C.4 Approximating Bayesian cross validation losses for hierarchical data with variants of WAIC

C.4.1 Leave-one-cluster-out cross validation and oWAIC

An approximate mathematical equivalence between Bayesian leave-one-cluster-out cross validation loss and oWAIC score under the JHQR model is established by showing the Taylor's expansions of the two quantities agree in the leading 3 terms. In the first part of the proof, we rewrite the Bayesian leave-one-cluster-out cross validation loss using importance weighting, and then in the second part we calculate and compare the leading terms of the Taylor's expansions of the two quantities. Two parts adapt respectively from the derivations in Section 4.1 and 4.2 of Li et al. (2016) and Watanabe (2010).

Part (i): Recall the whole dataset is $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$. The Bayesian leave-one-cluster-out cross validation loss of the log predictive density (Geisser and Eddy, 1979; Gneiting and Raftery, 2007) is

$$\frac{1}{n}\sum_{i=1}^{n} \log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \mathcal{D}_{-i}) \tag{C.4.1}$$

where $\mathcal{D}_{-i}$ is the partial dataset without observations in the ith cluster. Our model contains parameters $\eta_m = (\beta_0, \beta)$ of the marginal part and copula parameters $\eta_c = (\theta, \boldsymbol{\phi})$. Since observations are conditionally independent across clusters given model parameters, we have

$$\begin{aligned} p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \mathcal{D}_{-i}) &= \int p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \eta_c, \mathcal{D}_{-i}) p(\eta_m, \eta_c \mid \mathcal{D}_{-i}) \mathrm{d}\eta_m \mathrm{d}\eta_c \\ &= \int p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i) p(\eta_m, \theta, \phi_i \mid \mathcal{D}_{-i}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i \end{aligned} \tag{C.4.2}$$

Applying the idea of importance sampling (Gelfand et al., 1992; Gelfand, 1996), we

have

$$(\text{C.4.2}) = \frac{\int p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i) \frac{p(\eta_m, \theta, \phi_i \mid \mathcal{D}_{-i})}{p(\eta_m, \theta, \phi_i \mid \mathcal{D})} p(\eta_m, \theta, \phi_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i}{\int \frac{p(\eta_m, \theta, \phi_i \mid \mathcal{D}_{-i})}{p(\eta_m, \theta, \phi_i \mid \mathcal{D})} p(\eta_m, \theta, \phi_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i}$$

Note that

$$p(\eta_m, \theta, \phi_i \mid \mathcal{D}_{-i}) = \int \frac{1}{C_1} \prod_{j \neq i}^{n} p(\boldsymbol{y}_j \mid \boldsymbol{x}_j, \eta_m, \theta, \phi_j) p(\eta_m) p(\theta) p(\boldsymbol{\phi} \mid \xi) p(\xi) \mathrm{d}\phi_{-i} \mathrm{d}\xi$$

where ξ contains all hyperparameters of the hyperprior for $\boldsymbol{\phi}$ and C_1 is a normalizing constant only involving $\mathcal{D}_{-i}$. Similarly,

$$p(\eta_m, \theta, \phi_i \mid \mathcal{D}) = \int \frac{1}{C_2} \prod_{j=1}^{n} p(\boldsymbol{y}_j \mid \boldsymbol{x}_j, \eta_m, \theta, \phi_j) p(\eta_m) p(\theta) p(\boldsymbol{\phi} \mid \xi) p(\xi) \mathrm{d}\phi_{-i} \mathrm{d}\xi$$

where C_2 is a normalizing constant only involving $\mathcal{D}$. Their ratio is

$$\frac{p(\eta_m, \theta, \phi_i \mid \mathcal{D}_{-i})}{p(\eta_m, \theta, \phi_i \mid \mathcal{D})} = \frac{C_2}{C_1} \frac{1}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)}$$

Therefore,

$$\begin{aligned}
(\text{C.4.2}) &= \frac{\int p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i) \frac{C_2}{C_1} \frac{1}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)} p(\eta_m, \theta, \phi_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i}{\int \frac{C_2}{C_1} \frac{1}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)} p(\eta_m, \theta, \phi_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i} \\
&= \frac{1}{\int \frac{1}{p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)} p(\eta_m, \theta, \phi_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i} \\
\log (\text{C.4.2}) &= -\log \mathsf{E}_{\eta_m, \theta, \phi_i \mid \mathcal{D}} [p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)^{-1}]
\end{aligned}$$

Note that log of $p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)$ is precisely given by Equation (4.2.2).

Part (ii): The oWAIC for the ith cluster (up to a constant) is essentially

$$\text{oWAIC}_i = \log \mathsf{E}_{\eta_m, \theta, \phi_i \mid \mathcal{D}} [p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)] - \mathsf{Var}_{\eta_m, \theta, \phi_i \mid \mathcal{D}} [\log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)]$$

Let

$$F(k) = -\log \mathsf{E}_{\eta_m, \theta, \phi_i \mid \mathcal{D}} [p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)^k]$$

Then $F(-1) = \log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \mathcal{D}_{-i})$. Taking the Taylor's expansion of $F(k)$ at $k = 0$, we have

$$F(-1) = F(0) - F'(0) + \frac{F''(0)}{2} - \frac{F'''(0)}{6} + \sum_{i=4}^{\infty} \frac{(-1)^i F^{(i)}(0)}{i!} \tag{C.4.3}$$

$$F(1) = F(0) + F'(0) + \frac{F''(0)}{2} + \frac{F'''(0)}{6} + \sum_{i=4}^{\infty} \frac{F^{(i)}(0)}{i!}$$

With straightforward calculations, we can verify that

$$F'(0) = -\mathsf{E}_{\eta_m,\theta,\phi_i|\mathcal{D}}[\log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)]$$

$$F''(0) = -\mathsf{Var}_{\eta_m,\theta,\phi_i|\mathcal{D}}[\log p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)]$$

Since $F(0) = 0$ and $F(1) = -\log \mathsf{E}_{\eta_m,\theta,\phi_i|\mathcal{D}}[p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \eta_m, \theta, \phi_i)]$, we have

$$\text{oWAIC}_i = -F(1) + F''(0) = -F'(0) + \frac{1}{2}F''(0) - \frac{F'''(0)}{6} - \sum_{i=4}^{\infty} \frac{F^{(i)}(0)}{i!} \tag{C.4.4}$$

Comparing equations (C.4.3) and (C.4.4), we prove that the Taylor's expansions of the Bayesian leave-one-cluster-out cross validation loss and oWAIC score agree in the leading 3 terms. Particularly, we note that the proof does not rely on assumptions of dependency patterns or copula models and hence this approximate equivalence holds broadly for hierarchical data under the JHQR model with any copula models.

C.4.2 Leave-one-unit-out cross validation and iWAIC

Similarly, we show the Taylor's expansions of the Bayesian leave-one-unit-out cross validation loss and iWAIC agree in the leading 3 terms, for exchangeable dependency scenario (M1) and continuous temporal copula process scenario (M3). As the proof is similar to that of oWAIC, we only present essential steps.

With exchangeable dependency for cross-sectional data, the Bayesian leave-one-unit-out cross validation loss of the log predictive density is given by

$$\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log p(y_{ij} \mid x_{ij}, \mathcal{D}_{-ij}) \tag{C.4.5}$$

where $\mathcal{D}_{-ij}$ denotes the set of all observations except (x_{ij}, y_{ij}). The observations in the same cluster are conditionally independent given the latent quantile level W_i in addition to model parameters $\eta_m = (\beta_0, \beta)$ of the marginal part and copula parameters $\eta_c = (\theta, \boldsymbol{\phi})$. We have

$$\begin{aligned} p(y_{ij} \mid x_{ij}, \mathcal{D}_{-ij}) &= \int p(y_{ij} \mid x_{ij}, \eta_m, \eta_c, W_i, \mathcal{D}_{-ij}) p(\eta_m, \eta_c, W_i \mid \mathcal{D}_{-ij}) \mathrm{d}\eta_m \mathrm{d}\eta_c \mathrm{d}W_i \\ &= \int p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i) p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}_{-ij}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i \mathrm{d}W_i \\ &= \frac{\int p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i) \frac{p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}_{-ij})}{p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D})} p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i \mathrm{d}W_i}{\int \frac{p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}_{-ij})}{p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D})} p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i \mathrm{d}W_i} \end{aligned} \tag{C.4.6}$$

Note that

$$p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}_{-ij}) = \int \frac{1}{C_1} \frac{\prod_{k=1}^{n} \prod_{l=1}^{n_k} p(y_{kl} \mid x_{kl}, \eta_m, \theta, \phi_k, W_k)}{p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)} p(\eta_m) p(\theta) \times$$

$$\prod_{k=1}^{n} p(W_k \mid \phi_k) p(\boldsymbol{\phi} \mid \xi) p(\xi) \mathrm{d}\phi_{-i} \mathrm{d}\xi \mathrm{d}W_{-i}$$

where ξ contains all hyperparameters of the hyperprior for $\boldsymbol{\phi}$ and C_1 is a normalizing constant only involving $\mathcal{D}_{-ij}$. Similarly,

$$p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}) = \int \frac{1}{C_1} \prod_{k=1}^{n} \prod_{l=1}^{n_k} p(y_{kl} \mid x_{kl}, \eta_m, \theta, \phi_k, W_k) p(\eta_m) p(\theta) \times$$

$$\prod_{k=1}^{n} p(W_k \mid \phi_k) p(\boldsymbol{\phi} \mid \xi) p(\xi) \mathrm{d}\phi_{-i} \mathrm{d}\xi \mathrm{d}W_{-i}$$

where C_2 is a normalizing constant only involving $\mathcal{D}$. Their ratio is

$$\frac{p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}_{-ij})}{p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D})} = \frac{C_2}{C_1} \frac{1}{p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)}$$

Therefore,

$$(\text{C.4.6}) = \frac{1}{\int \frac{1}{p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)} p(\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}) \mathrm{d}\eta_m \mathrm{d}\theta \mathrm{d}\phi_i \mathrm{d}W_i}$$

$$\log (\text{C.4.6}) = -\log \mathsf{E}_{\eta_m, \theta, \phi_i, W_i \mid \mathcal{D}} [p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)^{-1}]$$

Note that log of $p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)$ is precisely given by Equation (4.2.4) and (C.3.1) for Gaussian and t copula respectively.

The iWAIC score for the jth observation in cluster i (up to a constant) is given by

$$\begin{aligned}\text{iWAIC}_{ij} = &\log \mathsf{E}_{\eta_m,\theta,\phi_i,W_i|\mathcal{D}}[p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)] - \\ &\mathsf{Var}_{\eta_m,\theta,\phi_i,W_i|\mathcal{D}}[\log p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)]\end{aligned}$$

Let

$$F(k) = -\log \mathsf{E}_{\eta_m,\theta,\phi_i,W_i|\mathcal{D}}[p(y_{ij} \mid x_{ij}, \eta_m, \theta, \phi_i, W_i)^k]$$

With the Taylor's expansion of $F(k)$ at $k = 0$, we have

$$\log p(y_{ij} \mid x_{ij}, \mathcal{D}_{-ij}) = F(-1) = F(0) - F'(0) + \frac{F''(0)}{2} - \frac{F'''(0)}{6} + \sum_{i=4}^{\infty} \frac{(-1)^i F^{(i)}(0)}{i!} \tag{C.4.7}$$

$$\text{iWAIC}_{ij} = -F(1) + F''(0) = -F'(0) + \frac{1}{2}F''(0) - \frac{F'''(0)}{6} - \sum_{i=4}^{\infty} \frac{F^{(i)}(0)}{i!} \tag{C.4.8}$$

Comparing equations (C.4.7) and (C.4.8), we prove that the Taylor's expansions of the Bayesian leave-one-unit-out cross validation loss and iWAIC score under JHQR model with exchangeable dependency structure (M1) agree in the leading 3 terms.

The proof is basically same for continuous temporal copula process with longitudinal data at irregular time grid, with the only difference being the target predictive loss. In this scenario, the Bayesian leave-final-unit-out cross validation loss of the log predictive density is

$$\frac{1}{n}\sum_{i=1}^{n} \log p(y_{in_i} \mid x_{in_i}, \mathcal{D}_{-in_i}).$$

Bibliography

Abrevaya, J. and Dahl, C. M. (2008). "The effects of birth inputs on birthweight: evidence from quantile estimation on panel data." *Journal of Business & Economic Statistics*, 26(4): 379–397.

Andrieu, C. and Thoms, J. (2008). "A tutorial on adaptive MCMC." *Statistics and computing*, 18(4): 343–373.

Ang, A. and Chen, J. (2002). "Asymmetric correlations of equity portfolios." *Journal of financial Economics*, 63(3): 443–494.

Azzalini, A. and Capitanio, A. (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 367–389.

Azzalini, A. and Valle, A. D. (1996). "The multivariate skew-normal distribution." *Biometrika*, 83(4): 715–726.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). "Gaussian predictive process models for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4): 825–848.

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." *Water Resources Research*, 42(11).

Barmpadimos, I., Keller, J., Oderbolz, D., Hueglin, C., and Prévôt, A. (2012). "One decade of parallel fine (PM 2.5) and coarse (PM 10–PM 2.5) particulate matter measurements in Europe: trends and variability." *Atmospheric Chemistry and Physics*, 12(7): 3189–3203.

Bell, M. L., Dominici, F., and Samet, J. M. (2005). "A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study." *Epidemiology (Cambridge, Mass.)*, 16(4): 436.

Benoit, D. F., Van den Poel, D., et al. (2017). "bayesQR: A Bayesian approach to quantile regression." *Journal of Statistical Software*, 76(7): 1–32.

Brook, R. D., Rajagopalan, S., Pope III, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., et al. (2010). "Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association." *Circulation*, 121(21): 2331–2378.

Buchinsky, M. (1994). "Changes in the US Wage Structure 1963-1987: Application of Quantile Regression." *Econometrica: Journal of the Econometric Society*, 405–458.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). "Approximate leave-future-out cross-validation for Bayesian time series models." *Journal of Statistical Computation and Simulation*, 90(14): 2499–2523.

Butry, D. T., Mercer, E., Prestemon, J. P., Pye, J. M., and Holmes, T. P. (2001). "What is the price of catastrophic wildfire?" *Journal of Forestry*, 99(11): 9–17.

Cade, B. S. and Noon, B. R. (2003). "A gentle introduction to quantile regression for ecologists." *Frontiers in Ecology and the Environment*, 1(8): 412–420.

Cade, B. S., Terrell, J. W., and Schroeder, R. L. (1999). "Estimating effects of limiting factors with regression quantiles." *Ecology*, 80(1): 311–323.

Chen, X. and Fan, Y. (2006). "Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification." *Journal of econometrics*, 135(1-2): 125–154.

Chen, X., Koenker, R., and Xiao, Z. (2009). "Copula-based nonlinear quantile autoregression." *The Econometrics Journal*, 12: S50–S67.

Chen, X. and Tokdar, S. T. (2019). "Joint Quantile Regression for Spatial Data." *arXiv preprint arXiv:1910.13119*.

Cohen, J. D. and Deeming, J. E. (1985). *The national fire-danger rating system: basic equations*, volume 82. US Department of Agriculture, Forest Service, Pacific Southwest Forest and Range Experiment Station.

Cressie, N. A. (1993). *Statistics for spatial data, 2nd*. New York: Wiley.

Crothers, K., Griffith, T. A., McGinnis, K. A., Rodriguez-Barradas, M. C., Leaf, D. A., Weissman, S., Gibert, C. L., Butt, A. A., and Justice, A. C. (2005). "The impact of cigarette smoking on mortality, quality of life, and comorbid illness among HIV-positive veterans." *Journal of general internal medicine*, 20(12): 1142–1145.

Cunningham, E., Tokdar, S. T., and Clark, J. S. (2020). "Chapter 2 - A vignette on model-based quantile regression: analysing excess zero response." In Fan, Y., Nott, D., Smith, M. S., and Dortet-Bernadet, J.-L. (eds.), *Flexible Bayesian Regression Modelling*, 27 – 64. Academic Press.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111(514): 800–812.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99(465): 205–215.

Demarta, S. and McNeil, A. J. (2005). "The t copula and related copulas." *International statistical review*, 73(1): 111–129.

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Heagerty, P. J., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford University Press.

Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). "Selecting and estimating regular vine copulae and application to financial returns." *Computational Statistics & Data Analysis*, 59: 52–69.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G., and Speizer, F. E. (1993). "An association between air pollution and mortality in six US cities." *New England journal of medicine*, 329(24): 1753–1759.

Dunson, D. B., Pillai, N., and Park, J.-H. (2007). "Bayesian density regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 163–183.

Ehrenberg, R. G. and Brewer, D. J. (1994). "Do school and teacher characteristics matter? Evidence from high school and beyond." *Economics of education review*, 13(1): 1–17.

Elsner, J. B., Kossin, J. P., and Jagger, T. H. (2008). "The increasing intensity of the strongest tropical cyclones." *Nature*, 455(7209): 92.

Embrechts, P., Lindskog, F., and McNeil, A. (2001). "Modelling dependence with copulas." *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Fan, J. and Zhang, J.-T. (2000). "Two-step estimation of functional linear models with applications to longitudinal data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2): 303–322.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). "Efficient algorithms for Bayesian nearest neighbor Gaussian processes." *Journal of Computational and Graphical Statistics*, 28(2): 401–414.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). "Unconditional quantile regressions." *Econometrica*, 77(3): 953–973.

Foresi, S. and Peracchi, F. (1995). "The conditional distribution of excess returns: An empirical analysis." *Journal of the American Statistical Association*, 90(430): 451–466.

Foster, L., Waagen, A., Aijaz, N., Hurley, M., Luis, A., Rinsky, J., Satyavolu, C., Way, M. J., Gazis, P., and Srivastava, A. (2009). "Stable and efficient gaussian process calculations." *Journal of Machine Learning Research*, 10(Apr): 857–882.

Galai, N., Park, L. P., Wesch, J., Visscher, B., Riddler, S., and Margolick, J. B. (1997). "Effect of smoking on the clinical progression of HIV-1 infection." *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 14(5): 451–458.

Geisser, S. and Eddy, W. F. (1979). "A predictive approach to model selection." *Journal of the American Statistical Association*, 74(365): 153–160.

Gelfand, A. E. (1996). "Model determination using sampling-based methods." *Markov chain Monte Carlo in practice*, 145–161.

Gelfand, A. E., Dey, D. K., Chang, H., Ior, A., et al. (1992). "Model determination using predictive distributions with implementation via sampling-based-methods (with Discussion." In *In Bayesian Statistics 4*. Citeseer.

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

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6): 997–1016.

Genest, C., Quessy, J.-F., and Rémillard, B. (2006). "Goodness-of-fit procedures for copula models based on the probability integral transformation." *Scandinavian Journal of Statistics*, 33(2): 337–366.

Geraci, M. (2014). "Linear quantile mixed models: the lqmm package for Laplace quantile regression." *Journal of Statistical Software*, 57(1): 1–29.

Geraci, M. and Bottai, M. (2007). "Quantile regression for longitudinal data using the asymmetric Laplace distribution." *Biostatistics*, 8(1): 140–154.

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American statistical Association*, 102(477): 359–378.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). "An adaptive Metropolis algorithm." *Bernoulli*, 7(2): 223–242.

Hallin, M., Lu, Z., Yu, K., et al. (2009). "Local linear spatial quantile regression." *Bernoulli*, 15(3): 659–686.

He, X. (1997). "Quantile curves without crossing." *The American Statistician*, 51(2): 186–192.

Hofert, M., Mächler, M., and Mcneil, A. J. (2012). "Likelihood inference for Archimedean copulas in high dimensions under known margins." *Journal of Multivariate Analysis*, 110: 133–150.

Hofert, M., Mächler, M., and McNeil, A. J. (2013). "Archimedean copulas in high dimensions: Estimators and numerical challenges motivated by financial applications." *Journal de la Société Française de Statistique*, 154(1): 25–63.

Hong, Y., Tu, J., and Zhou, G. (2007). "Asymmetries in stock returns: Statistical tests and economic evaluation." *The Review of Financial Studies*, 20(5): 1547–1581.

Huard, D., Évin, G., and Favre, A.-C. (2006). "Bayesian copula selection." *Computational Statistics & Data Analysis*, 51(2): 809–822.

Jerrett, M., Burnett, R. T., Ma, R., Pope III, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E., et al. (2005). "Spatial analysis of air pollution and mortality in Los Angeles." *Epidemiology*, 727–736.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.

— (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

Kabali, C., Cheng, D. M., Brooks, D. R., Bridden, C., Horsburgh Jr, C. R., and Samet, J. H. (2011). "Recent cigarette smoking and HIV disease progression: no evidence of an association." *AIDS care*, 23(8): 947–956.

Kendall, M. G. (1938). "A new measure of rank correlation." *Biometrika*, 30(1/2): 81–93.

Kim, M.-O. and Yang, Y. (2011). "Semiparametric approach to a random effects quantile regression model." *Journal of the American Statistical Association*, 106(496): 1405–1417.

Kirkbride, J. B., Jackson, D., Perez, J., Fowler, D., Winton, F., Coid, J. W., Murray, R. M., and Jones, P. B. (2013). "A population-level prediction tool for the incidence of first-episode psychosis: translational epidemiology based on cross-sectional data." *BMJ open*, 3(2): e001998.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015). "Bayesian structured additive distributional regression for multivariate responses." *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 569–591.

Koenker, R. (1994). "Confidence intervals for regression quantiles." In *Asymptotic statistics*, 349–359. Springer.

— (2005). *Quantile Regression (Econometric Society Monographs; no. 38)*. Cambridge University Press.

— (2017). "Quantile regression: 40 years on." *Annual Review of Economics*, 9: 155–176.

Koenker, R. and Bassett, G. (1978). "Regression quantiles." *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R., Leorato, S., and Peracchi, F. (2013). "Distributional vs. quantile regression."

Koenker, R. and Xiao, Z. (2006). "Quantile autoregression." *Journal of the American statistical association*, 101(475): 980–990.

Kreuzer, A., Erhardt, T., Nagler, T., and Czado, C. (2017). "Heavy tailed spatial autocorrelation models." *arXiv preprint arXiv:1707.03165*.

Krupskii, P. and Joe, H. (2015). "Structured factor copula models: Theory, inference and computation." *Journal of Multivariate Analysis*, 138: 53–73.

Lee, V. E. and Bryk, A. S. (1989). "A multilevel model of the social distribution of high school achievement." *Sociology of education*, 172–192.

Li, L., Qiu, S., Zhang, B., and Feng, C. X. (2016). "Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC." *Statistics and Computing*, 26(4): 881–897.

Lin, D. and Ying, Z. (2001). "Semiparametric and nonparametric regression analysis of longitudinal data." *Journal of the American Statistical Association*, 96(453): 103–126.

Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., and Zhao, L. P. (1997). "Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4): 463–476.

Lum, K. and Gelfand, A. E. (2012). "Spatial quantile multiple regression using the asymmetric Laplace process." *Bayesian Analysis*, 7(2): 235–258.

Machado, J. A. and Mata, J. (2005). "Counterfactual decomposition of changes in wage distributions using quantile regression." *Journal of applied Econometrics*, 20(4): 445–465.

Mari, D. D. and Kotz, S. (2001). *Correlation and dependence*. World Scientific.

Mau, W.-C. (2003). "Factors that influence persistence in science and engineering career aspirations." *The Career Development Quarterly*, 51(3): 234–243.

McNeil, A. J., Frey, R., Embrechts, P., et al. (2005). *Quantitative risk management: Concepts, techniques and tools*, volume 3. Princeton university press Princeton.

Merkle, E. C., Furr, D., and Rabe-Hesketh, S. (2019). "Bayesian comparison of latent variable models: Conditional versus marginal likelihoods." *psychometrika*, 84(3): 802–829.

Millar, R. B. (2018). "Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation." *Statistics and Computing*, 28(2): 375–385.

Moretti, E. (2004). "Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data." *Journal of econometrics*, 121(1-2): 175–212.

Moritz, M. A., Batllori, E., Bradstock, R. A., Gill, A. M., Handmer, J., Hessburg, P. F., Leonard, J., McCaffrey, S., Odion, D. C., Schoennagel, T., et al. (2014). "Learning to coexist with wildfire." *Nature*, 515(7525): 58–66.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Paciorek, C. J. (2013). "Spatial models for point and areal data using Markov random fields on a fine grid." *Electronic Journal of Statistics*, 7: 946–972.

Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F., Suh, H. H., et al. (2009). "Practical large-scale spatio-temporal modeling of particulate matter concentrations." *The Annals of Applied Statistics*, 3(1): 370–397.

Petscher, Y. and Logan, J. A. (2014). "Quantile regression in the study of developmental sciences." *Child development*, 85(3): 861–881.

Pope, C. A. and Dockery, D. W. (2006). "Health effects of fine particulate air pollution: lines that connect." *Journal of the air & waste management association*, 56(6): 709–742.

Pope, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., Heath, C. W., et al. (1995). "Particulate air pollution as a predictor of mortality in a prospective study of US adults." *American journal of respiratory and critical care medicine*, 151(3): 669–674.

Porter, W. C., Heald, C. L., Cooley, D., and Russell, B. (2015). "Investigating the observed sensitivities of air-quality extremes to meteorological drivers via quantile regression." *Atmospheric Chemistry and Physics*, 15(18): 10349–10366.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian process for machine learning*. MIT press.

Reich, B. J., Bondell, H. D., and Wang, H. J. (2010). "Flexible Bayesian quantile regression for independent and clustered data." *Biostatistics*, 11(2): 337–352.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). "Bayesian spatial quantile regression." *Journal of the American Statistical Association*, 106(493): 6–20.

Reid, C. E., Brauer, M., Johnston, F. H., Jerrett, M., Balmes, J. R., and Elliott, C. T. (2016). "Critical review of health impacts of wildfire smoke exposure." *Environmental health perspectives*, 124(9): 1334–1343.

Ritchie, S. J. and Bates, T. C. (2013). "Enduring links from childhood mathematics and reading achievement to adult socioeconomic status." *Psychological science*, 24(7): 1301–1308.

Rodriguez, J. C. (2007). "Measuring financial contagion: A copula approach." *Journal of empirical finance*, 14(3): 401–423.

Royce, R. A. and Winkelstein Jr, W. (1990). "HIV infection, cigarette smoking and CD4+ T-lymphocyte counts: preliminary results from the San Francisco Men's Health Study." *AIDS (London, England)*, 4(4): 327–333.

Sibuya, M. (1960). "Bivariate extreme statistics, I." *Annals of the Institute of Statistical Mathematics*, 11(2): 195–210.

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., and Conley, A. M. (2015). "Raising the stakes: How students' motivation for mathematics associates with high-and low-stakes test achievement." *Learning and individual differences*, 39: 49–63.

Sklar, A. (1959). "Fonstions de Répartition à n dimensions et leurs Marges." *Publications de l'Institut de Statistique de l'Université de Paris*, 8: 229–231.

Smith, M., Min, A., Almeida, C., and Czado, C. (2010). "Modeling longitudinal data using a pair-copula decomposition of serial dependence." *Journal of the American Statistical Association*, 105(492): 1467–1479.

Smith, M. S., Gan, Q., and Kohn, R. J. (2012). "Modelling dependence using skew t copulas: Bayesian inference and applications." *Journal of Applied Econometrics*, 27(3): 500–522.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Susperreguy, M. I., Davis-Kean, P. E., Duckworth, K., and Chen, M. (2018). "Self-concept predicts academic achievement across levels of the achievement distribution: Domain specificity for math and reading." *Child development*, 89(6): 2196–2214.

Sweeting, M. and Thompson, S. (2012). "Making predictions from complex longitudinal data, with application to planning monitoring intervals in a national screening programme." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2): 569–586.

Tang, C. Y. and Leng, C. (2011). "Empirical likelihood and quantile regression in longitudinal data analysis." *Biometrika*, 98(4): 1001–1006.

Tokdar, S. T. (2007). "Towards a faster implementation of density estimation with logistic Gaussian process priors." *Journal of Computational and Graphical Statistics*, 16(3): 633–655.

— (2013). "Contributed Discussion on Article by Muller and Mitra [Bayesian Nonparametric Inference–Why and How]." *Bayesian Analysis*, 323—-356.

Tokdar, S. T. and Ghosh, J. K. (2007). "Posterior consistency of logistic Gaussian process priors in density estimation." *Journal of statistical planning and inference*, 137(1): 34–42.

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

Tokdar, S. T., Zhu, Y. M., Ghosh, J. K., et al. (2010). "Bayesian density regression with logistic Gaussian process and subspace projection." *Bayesian analysis*, 5(2): 319–344.

Uhlenbeck, G. E. and Ornstein, L. S. (1930). "On the theory of the Brownian motion." *Physical review*, 36(5): 823.

van der Vaart, A. W., van Zanten, J. H., et al. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth." *The Annals of Statistics*, 37(5B): 2655–2675.

Wang, H. J., Feng, X., and Dong, C. (2019). "Copula-based quantile regression for longitudinal data." *Statistica Sinica*, 29(1): 245–264.

Wang, H. J. and Zhu, Z. (2011). "Empirical likelihood for quantile regression models with longitudinal data." *Journal of statistical planning and inference*, 141(4): 1603–1615.

Wang, H. J., Zhu, Z., and Zhou, J. (2009). "Quantile regression in partially linear varying coefficient models." *The Annals of Statistics*, 3841–3866.

Wang, X. (2013). "Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support." *American Educational Research Journal*, 50(5): 1081–1121.

Wasko, C. and Sharma, A. (2014). "Quantile regression for investigating scaling of extreme precipitation with temperature." *Water Resources Research*, 50(4): 3608–3614.

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, 11(Dec): 3571–3594.

Willms, J. D. (1985). "Catholic-school effects on academic achievement: New evidence from the high school and beyond follow-up study." *Sociology of education*, 98–114.

Yang, Y. and He, X. (2015). "Quantile regression for spatially correlated data: An empirical likelihood approach." *Statistica Sinica*, 261–274.

Yang, Y. and Tokdar, S. T. (2015). "Minimax-optimal nonparametric regression in high dimensions." *The Annals of Statistics*, 43(2): 652–674.

— (2017). "Joint estimation of quantile planes over arbitrary predictor spaces." *Journal of the American Statistical Association*, 112(519): 1107–1120.

Yanosky, J. D., Paciorek, C. J., Schwartz, J., Laden, F., Puett, R., and Suh, H. H. (2008a). "Spatio-temporal modeling of chronic PM10 exposure for the Nurses' Health Study." *Atmospheric Environment*, 42(18): 4047–4062.

Yanosky, J. D., Paciorek, C. J., and Suh, H. H. (2008b). "Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the Northeastern and Midwestern United States."

Yu, K. and Moyeed, R. A. (2001). "Bayesian quantile regression." *Statistics & Probability Letters*, 54(4): 437–447.

Yuan, Y. and Yin, G. (2010). "Bayesian quantile regression for longitudinal studies with nonignorable missing data." *Biometrics*, 66(1): 105–114.

Zeger, S. L. and Diggle, P. J. (1994). "Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters." *Biometrics*, 689–699.