Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Joint Estimation of Quantile Planes over Arbitrary Predictor Spaces

Yun Yang & Surya T. Tokdar

To cite this article: Yun Yang & Surya T. Tokdar (2016): Joint Estimation of Quantile Planes over Arbitrary Predictor Spaces, Journal of the American Statistical Association, DOI: 10.1080/01621459.2016.1192545

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1192545

Accepted author version posted online: 10 Jun 2016.
Published online: 10 Jun 2016.

Joint Estimation of Quantile Planes over Arbitrary Predictor Spaces

Yun Yang* and Surya T. Tokdar†

Florida State University and Duke University

Abstract

In spite of the recent surge of interest in quantile regression, joint estimation of linear quantile planes remains a great challenge in statistics and econometrics. We propose a novel parametrization that characterizes any collection of non-crossing quantile planes over arbitrarily shaped convex predictor domains in any dimension by means of unconstrained scalar, vector and function valued parameters. Statistical models based on this parametrization inherit a fast computation of the likelihood function, enabling penalized likelihood or Bayesian approaches to model fitting. We introduce a complete Bayesian methodology by using Gaussian process prior distributions on the function valued parameters and develop a robust and efficient Markov chain Monte Carlo parameter estimation. The resulting method is shown to offer posterior consistency under mild tail and regularity conditions. We present several illustrative examples where the new method is compared against existing approaches and is found to offer better accuracy, coverage and model fit.

Keywords: Bayesian Inference; Bayesian Nonparametric Models; Gaussian Processes; Joint Quantile Model; Linear Quantile Regression.

*Email: yyang@stat.fsu.edu
†Email: tokdar@stat.duke.edu

1. Introduction

1.1 Quantile regression

Quantile regression (QR; Koenker and Bassett, 1978) has recently gained increased recognition as a robust alternative to standard least squares regression, with applications to ecology, economics, epidemiology and climate science research, to name a few (Burgette et al., 2011; Elsner et al., 2008; Dunham et al., 2002; Abrevaya, 2001). In analyzing a scalar response Y against a vector of predictors X, Koenker and Bassett (1978) advocated replacing a linear model for the response mean with the model

$$Q_Y(\tau|X) = \beta_0(\tau) + X^T\beta(\tau), \tag{1}$$

where $\tau \in (0, 1)$ is a response proportion of interest and $Q_Y(\tau|X) = \inf\{a : P(Y \leq a|X) \geq \tau\}$ denotes the τ-th conditional quantile of Y given X. By choosing a suitable τ, researchers can detect dependency beyond the response mean or median, and, gain direct inference on how predictors affect the extremes of the response distribution. The model parameters are easily estimated by linear programming and the estimates are consistent, asymptotically Gaussian and robust against outliers. Current literature on quantile regression (QR) is both deep and diverse; see Koenker (2005) for a comprehensive overview for frequentist quantile regression. For related Bayesian approaches were developed in Yu and Moyeed (2001); Tsionas (2003); Gelfand and Kottas (2003); Kottas and Krnjajić (2009) and Thompson et al. (2010).

1.2 The issue of joint estimation

Most scientific applications of QR require inference over a dense grid of τ values, which is usually done by assimilating inference from single-τ model fits (e.g., Elsner et al., 2008). Such assimilations are often problematic because of estimated quantile curves crossing each other, a poor borrowing of information across quantiles, and, prevalence of irreconcilable differences in inference (He, 1997; Bondell et al., 2010; Reich et al., 2011; Tokdar and Kadane, 2012; Feng et al.,

2015). A congruent joint estimation of the QR intercepts and slopes for a multitude of τ values has remained a major challenge over nearly four decades.

A fundamental fact is that the linear QR specification (1), when considered simultaneously for all $\tau \in (0, 1)$, indeed gives a valid probabilistic model for the response

$$Y = \beta_0(U) + X^T\beta(U), \quad U \sim \text{Uniform}(0, 1), \tag{2}$$

parametrized by function valued intercepts and slopes $\beta_0 : (0, 1) \to \mathbb{R}$, $\beta : (0, 1) \to \mathbb{R}^p$, provided the following 'non-crossing' condition holds:

$$\beta_0(\tau_1) + x^T\beta(\tau_1) \geq \beta_0(\tau_2) + x^T\beta(\tau_2), \text{ for every pair } \tau_1 > \tau_2 \text{ and for every } x \in \mathcal{X} \tag{3}$$

where $\mathcal{X}$ is a pre-specified domain for X. Estimating the intercept and slope functions under the non-crossing constraint is challenging and the few existing methods in the literature face many shortcomings. He (1997) considers only a restricted version of the model. The use of approximate likelihood in Dunson and Taylor (2005) and Feng et al. (2015) remain debated and their methods face serious computational issues in scaling up to a dense grid of response proportions τ. Tokdar and Kadane (2012) provide a scalable solution for the univariate case, but work with a restricted single index model when X is multivariate. Bondell et al. (2010) and Reich et al. (2011) offer more comprehensive and scalable solutions by assuming the predictor domain is a hyper-rectangle in $\mathbb{R}^p$, but the current algorithmic implementation of the former struggles with extreme quantiles when sample size is small and that of the latter faces severe convergence issues when p is moderately large. More fundamentally, the assumption on $\mathcal{X}$ being a hyper-rectangle essentially encodes a seriously restricted version of the linear QR model; this is elaborated below.

1.3 The critical issue of the shape of the predictor domain

The specification of $\mathcal{X}$ is a critical model choice in QR. Without loss of generality, $\mathcal{X}$ can be chosen closed and convex because (3) holds over $\mathcal{X}$ if and only if it holds over the convex hull of $\mathcal{X}$. The convex hull of the observed predictor vectors presents the most obvious practical choice.

In spite of convexity, such an $\mathcal{X}$ may have a fairly irregular shape, causing substantial difficulty to imposing the monotonicity constraint in an analytically tractable way. To impart tractability, it might seem appealing to embed such an irregular set inside the smallest enclosing hyper-sphere or hyper-rectangle and impose non-crossing constraints on the enclosing set. But such an approach can seriously bias estimation. The original $\mathcal{X}$ may occupy only a fraction of the volume of the encompassing hyper-rectangle. Quantile planes that are required to be non-crossing over the larger hyper-rectangle will appear mostly parallel within the original $\mathcal{X}$. This is illustrated in Figure 1, where it can be seen that the methods by Bondell et al. (2010) and Reich et al. (2011), which impose non-crossing on the smallest enclosing hyper-rectangle, offers severely biased estimation when the original $\mathcal{X}$ is triangular. Unfortunately, such narrow predictor convex hulls are unavoidable whenever non-linear transforms of covariates are used or the measured covariates are naturally correlated. These are also situations where assimilation techniques exhibit dramatic crossing problems and hence a sound statistical model is most needed for joint estimation.

1.4 New contributions

In this paper we propose a novel theory that delivers the right modeling platform for joint quantile regression. Our theory covers any dimension p and any bounded convex $\mathcal{X}$ of arbitrary shape. It provides an exact reparametrization of the non-crossing linear QR model indexed by scalars, vectors and curves all but one of which are entirely constraint-free. Even the one curve with a constraint has only a mild shape restriction on it; it is required to live in the space of all CDFs on $(0, 1)$ with full support. Our reparametrization leads to an easy likelihood score calculation in the model parameters, making it ideally suited to develop practicable methods by using either penalized likelihood optimization or Bayesian techniques.

We build upon this novel theory to introduce a semiparametric Bayesian methodology for joint quantile regression over any $\mathcal{X}$ and any p, where the curve valued model parameters are assigned Gaussian process and transformed Gaussian process priors within a hierarchical setting. Asymp-

totic frequentist properties of the method are studied in Section 4 and we establish posterior consistency over a broad class of true data generating distributions with linear quantile curves. For parameter estimation, we propose a Monte Carlo technique that incorporates efficient model space discretization, adaptive Markov chain sampling (Haario et al., 1999) and reduced rank approximation (Tokdar, 2007; Banerjee et al., 2008a). We provide empirical evidence (Section 5-6) that our Gaussian process method enjoys much better estimation accuracy and coverage than the methods by Bondell et al. (2010) and Reich et al. (2011), and our estimates are comparable to regularized versions of the classical single-τ estimates. We consider the developments here make a strong case for linear quantile regression to be used as a model based inferential method rather than just an exploratory tool. A software implementation of our method is available as the R-package 'qrjoint' through CRAN: http://cran.r-project.org/web/packages/qrjoint/index.html.

2. A novel theory of joint quantile planes estimation

2.1 Characterizing non-crossing hyperplanes

We focus only on the case where the response distribution is non-atomic and admits a probability density function conditionally at every X. Then (3) is equivalent to requiring $\dot{\beta}_0(\tau) + x^T\dot{\beta}(\tau) > 0$ for all $\tau \in (0,1)$, $x \in \mathcal{X}$. Here we use notation $\dot{f}(\cdot)$ to denote the derivative of a function $f(\cdot)$. Our theory could be extended to atomic response distributions with known atoms. Assume 0 is an interior point of $\mathcal{X}$. This can be achieved without any loss of generality by a simple translation of the predictors once a suitable interior point is found within the convex hull of the observed predictors; see Appendix B.1 for more details.

The monotonicity requirement (3) is equivalent to

$$\dot{\beta}_0(\tau) > \sup_{x\in\mathcal{X}} \{ - x^T\dot{\beta}(\tau)\} = a(\dot{\beta}(\tau), \mathcal{X})\, ||\dot{\beta}(\tau)||, \quad \text{for all } \tau \in (0,1), \tag{4}$$

where $a(\cdot, \mathcal{X}) : \mathbb{R}^p \to \mathbb{R}$ is given by,

$$a(b, \mathcal{X}) = \begin{cases} \sup_{x \in \mathcal{X}} \{-x^T b\}/\|b\| & b \neq 0, \\ \text{diam}(\mathcal{X}) := \sup_{x \in \mathcal{X}} \|x\| & b = 0. \end{cases}$$

Since 0 is an interior point of $\mathcal{X}$, for any non-zero $b \in \mathbb{R}^p$ there exists an $\epsilon > 0$ such that $-\epsilon b/\|b\| \in \mathcal{X}$ and hence $a(b, \mathcal{X}) \geq \epsilon > 0$. Therefore, $a(b, \mathcal{X}) > 0$ for all $b \in \mathbb{R}^p$. Also, in (4) we have a strict inequality because $\mathcal{X}$ is closed and the supremum can always be attained at some point in $\mathcal{X}$. This characterization of $(\dot{\beta}_0(\tau), \dot{\beta}(\tau))$ leads to the following parametrization.

Theorem 1. *Let $\mathcal{X}$ be a bounded convex set in $\mathbb{R}^p$ with zero as an interior point and let $\beta_0(\tau)$ and $\beta(\tau) = (\beta_1(\tau), \cdots, \beta_p(\tau))^T$ be real, differentiable functions in $\tau \in (0, 1)$. Then $\dot{\beta}_0(\tau) + x^T \dot{\beta}(\tau) > 0$ for all $\tau \in (0, 1)$ at every $x \in \mathcal{X}$ if and only if*

$$\dot{\beta}_0(\tau) > 0, \;\; \dot{\beta}(\tau) = \dot{\beta}_0(\tau) \frac{v(\tau)}{a(v(\tau), \mathcal{X}) \sqrt{1 + \|v(\tau)\|^2}}, \;\; \tau \in (0, 1), \tag{5}$$

for some p-dimensional, vector-valued function $v(\tau) = (v_1(\tau), \cdots, v_p(\tau))^T$ in $\tau \in (0, 1)$.

Proof. If part. Suppose (5) holds. For any $\tau \in (0, 1)$ either $v(\tau) = 0$ in which case $\dot{\beta}(\tau) = 0$ and so $\dot{\beta}_0(\tau) + x^T \dot{\beta}(\tau) > 0$ at every $x \in \mathcal{X}$. Otherwise, $v(\tau) \neq 0$ and so at any $x \in \mathcal{X}$,

$$\begin{aligned} \dot{\beta}_0(\tau) + x^T \dot{\beta}(\tau) &= \dot{\beta}_0(\tau) \left\{ 1 + \frac{x^T v(\tau)}{a(v(\tau), \mathcal{X}) \sqrt{1 + \|v(\tau)\|^2}} \right\} \\ &= \dot{\beta}_0(\tau) \left\{ 1 - \sqrt{\frac{\|v(\tau)\|^2}{1 + \|v(\tau)\|^2}} \frac{\{-x^T v(\tau)\}/\|v(\tau)\|}{a(v(\tau), \mathcal{X})} \right\} \\ &\geq \dot{\beta}_0(\tau) \left\{ 1 - \sqrt{\frac{\|v(\tau)\|^2}{1 + \|v(\tau)\|^2}} \right\} > 0, \end{aligned}$$

where, the penultimate step follows since $\{-x^T b/\|b\|\}/a(b, \mathcal{X}) \leq 1$ for any $b \neq 0$ by the definition of $a(b, \mathcal{X})$.

Only if part. If $\dot{\beta}_0(\tau) + x^T \dot{\beta}(\tau) > 0$ for all $\tau \in (0, 1)$ at every $x \in \mathcal{X}$, we must have $\dot{\beta}_0(\tau) > 0$ for all $\tau \in (0, 1)$ because $\mathcal{X}$ contains 0. For any $\tau \in (0, 1)$, if $\dot{\beta}(\tau) = 0$, set $v(\tau) = 0$, which

straight away gives the desired identity $\dot{\beta}(\tau) = \dot{\beta}_0(\tau)v(\tau)/\{a(v(\tau), \mathcal{X})\sqrt{1 + ||v(\tau)||^2}\}$. Otherwise, set $v(\tau) = c(\tau)\dot{\beta}(\tau)/||\dot{\beta}(\tau)||$ where

$$c(\tau) = \left[\left\{\frac{\dot{\beta}_0(\tau)}{||\dot{\beta}(\tau)||a(\dot{\beta}(\tau), \mathcal{X})}\right\}^2 - 1\right]^{-1/2},$$

which is a well defined positive scalar since by (4), $\dot{\beta}_0(\tau) > ||\dot{\beta}(\tau)||\, a(\dot{\beta}(\tau), \mathcal{X})$. Note that $||v(\tau)|| = c(\tau)$, and, $a(v(\tau), \mathcal{X}) = a(\dot{\beta}(\tau), \mathcal{X})$. Therefore,

$$\dot{\beta}_0(\tau)\frac{v(\tau)}{a(v(\tau), \mathcal{X})\sqrt{1 + ||v(\tau)||^2}} = \frac{\dot{\beta}_0(\tau)}{||\dot{\beta}(\tau)||a(\dot{\beta}(\tau), \mathcal{X})} \times \frac{c(\tau)}{\sqrt{1 + c(\tau)^2}} \times \dot{\beta}(\tau) = \dot{\beta}(\tau),$$

by the definition of $c(\tau)$. This proves the result! □

2.2 An almost constraint-free parametrization of linear quantile regression

Theorem 1 greatly reduces the monotonicity constraint on the quantile hyperplanes to that on a single function $\beta_0(\tau)$. Construction of a single monotone function is a relatively easy task, but some care is needed in handling the range $(\beta_0(0), \beta_0(1))$, which corresponds to the support of the conditional density of Y given $X = 0$. We pursue a model for $\beta_0(\tau)$ based on a user specified or default "prior guess" $f_0(y)$ for this conditional density. In the special case where $(\beta_0(0), \beta_0(1))$ is a known finite interval, f_0 could be chosen with support equal to the same interval. In general f_0 should be chosen to have support $(-\infty, \infty)$, such as a standard normal density, or a Student-t density with a modest degrees of freedom if the response distribution is expected to have heavy tails. We focus only on this general case, although the model described below could be easily modified to f_0 supported on a fixed or variable bounded interval.

Let f_0 have support $(-\infty, \infty)$ and define its cumulative distribution function $F_0(y) = \int_{-\infty}^{y} f_0(z)dz$, quantile function $Q_0(\tau) = F_0^{-1}(\tau)$ and quantile density $q_0(\tau) = \dot{Q}_0(\tau)$. Let $\tau_0 = F_0(0)$; by the full

support assumption, $0 < \tau_0 < 1$. We pursue a model for β_0 and β as follows

$$\beta_0(\tau_0) = \gamma_0, \;\; \beta(\tau_0) = \gamma \tag{6}$$

$$\beta_0(\tau) - \beta_0(\tau_0) = \sigma \int_{\zeta(\tau_0)}^{\zeta(\tau)} q_0(u)\, du, \quad \tau \in (0,1) \tag{7}$$

$$\beta(\tau) - \beta(\tau_0) = \sigma \int_{\zeta(\tau_0)}^{\zeta(\tau)} \frac{w(u)}{a(w(u), \mathcal{X})\sqrt{1 + \|w(u)\|^2}}\, q_0(u)\, du, \quad \tau \in (0,1) \tag{8}$$

with model parameters $\gamma_0 \in \mathbb{R}$; $\gamma \in \mathbb{R}^p$; $\sigma > 0$; $w : (0,1) \to \mathbb{R}^p$, an unconstrained p-dimensional vector valued function on $(0,1)$; and $\zeta : [0,1] \to [0,1]$, a differentiable, monotonically increasing bijection, i.e., a diffeomorphism, of $[0,1]$ onto itself. We write $(\beta_0, \beta) = \mathcal{T}(\gamma_0, \gamma, \sigma, w, \zeta)$ to indicate β_0, β defined as in (6)-(8).

All model parameters, except the diffeomorphism ζ, are essentially unconstrained. The function space of ζ is simply the space of cumulative distribution functions associated with all probability densities with support $[0,1]$. Such function spaces are easy to handle for statistical model fitting; a simple approach is presented in Section 3. Note that when ζ is the identity map of $(0,1)$ onto itself and $w(\tau) \equiv 0$, we get $\beta_0 = \sigma Q_0, \beta = 0$, and the resulting joint, linear quantile regression model simplifies to a standard homogeneous, linear model: $Y_i = \gamma_0 + X_i^T \gamma + \sigma\epsilon_i$ with $\epsilon_i \sim f_0$. Various other commonly considered quantile regression structures, e.g., the linear heteroscedasticity model of He (1997), emerge as special cases by fixing one or more of the model parameters $\gamma_0, \gamma, \sigma, w, \zeta$.

Most critically, the reparametrization given in (6)-(8) indeed provides a complete representation of all β_0, β satisfying the non-crossing condition (3), subject to a matching range criterion, as detailed below[1].

Theorem 2. *Let $\beta_0 : (0,1) \to \mathbb{R}$, $\beta : (0,1) \to \mathbb{R}^p$ be differentiable with $(\beta_0(0), \beta_0(1)) = (-\infty, \infty)$, as defined in the limit. Then $\dot\beta_0(\tau) + x^T\dot\beta(\tau) > 0$ for all $\tau \in (0,1)$ at every $x \in \mathcal{X}$, if and only*

[1] As can be seen in the proof of Theorem 2, the function $v(\tau)$ of (5) is expressed as $w(\zeta(\tau))$ in the reparametrization (6)-(8). This is not a restriction of any kind because ζ is a diffeomorphism. The particular reparametrization given in (6)-(8) is convenient for both theoretical calculations and numerical implementation.

if $(\beta_0, \beta) = \mathcal{T}(\gamma_0, \gamma, \sigma, w, \zeta)$ *for some* $\gamma_0 \in \mathbb{R}$, $\gamma \in \mathbb{R}^p$, $\sigma > 0$, $w : (0,1) \to \mathbb{R}^p$, *and,* ζ, *a diffeomorphism from* $[0,1]$ *onto itself* $[0,1]$.

Proof. If $(\beta_0, \beta) = \mathcal{T}(\gamma_0, \gamma, \sigma, w, \zeta)$ then,

$$\dot\beta_0(\tau) = \sigma q_0(\zeta(\tau))\dot\zeta(\tau), \;\; \dot\beta(\tau) = \dot\beta_0(\tau)\frac{v(\tau)}{a(v(\tau), \mathcal{X})\sqrt{1 + \|v(\tau)\|^2}},$$

with $v(\tau) = w(\zeta(\tau))$. Hence, by Theorem 1, we only need establish that any real, differentiable function β_0 on $(0,1)$, with $\dot\beta_0(\tau) > 0$ for all $\tau \in (0,1)$ and $\beta_0(0) = -\infty$, $\beta_0(1) = \infty$, can be constructed as in (7) for some diffeomorphism $\zeta : [0,1] \to [0,1]$ and $\sigma > 0$. This is indeed true since one could fix $\sigma > 0$ arbitrarily and then take

$$\zeta(\tau) = F_0\left(\gamma_0 + \frac{\beta_0(\tau) - \gamma_0}{\sigma}\right), \;\; \tau \in (0,1),$$

which is differentiable and monotonically increasing in $(0,1)$ since $\dot\zeta(\tau) = \sigma^{-1} f_0(\gamma_0 + (\beta_0(\tau) - \gamma_0)/\sigma)\dot\beta_0(\tau) > 0$ for all $\tau \in (0,1)$, and, has full range, since $\zeta(0) = F_0(-\infty) = 0$ and $\zeta(1) = F_0(\infty) = 1$. □

When $\beta_0(0)$ is finite or $\beta_0(1)$ is finite (or both), we can still write β_0 as in (7), but we need either $\zeta(0) > 0$ or $\zeta(1) < 1$ (or both). While such a (β_0, β) does not strictly belong within our model space, they can be approximated arbitrarily well by model elements $\mathcal{T}(\gamma_0, \gamma, \sigma, v, \zeta)$, and, will be consistently estimated as sample size grows to infinity (Lemma 8 and Section 4).

2.3 Likelihood evaluation

A salient feature of a valid specification of $Q_Y(\tau|x)$ for all $\tau \in (0,1)$ is that it uniquely defines the conditional response density $f_Y(y|x)$ over $x \in \mathcal{X}$, given by

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

where $\tau_x(y)$ solves $Q_Y(\tau|x) = y$ in τ (Tokdar and Kadane, 2012). Consequently, one can define a valid log-likelihood score

$$\sum_i \log f_Y(y_i|x_i) = -\sum_i \log\{\dot\beta_0(\tau_{x_i}(y_i)) + x_i^T\dot\beta(\tau_{x_i}(y_i))\}. \tag{9}$$

in the model parameters based on observations (x_i, y_i), $i = 1, \ldots, n$. From (6)-(8), we could write

$$\dot{\beta}_0(\tau) + x^T\dot{\beta}(\tau) = \sigma q_0(\zeta(\tau))\dot{\zeta}(\tau)\left\{1 + \frac{x^T w(\zeta(\tau))}{a(w(\zeta(\tau)), \mathcal{X})\sqrt{1 + \|w(\zeta(\tau))\|^2}}\right\}$$

and therefore a quick evaluation of the log-likelihood score is possible once we figure out $\tau_{x_i}(y_i)$ for each $i = 1, \ldots, n$, by solving $\tau = \int_{\tau_0}^{\tau}\{\dot{\beta}_0(u) + x_i^T\dot{\beta}(u)\}du$.

With enough resources, these numbers could be found up to any desired level of accuracy through standard numerical methods for integration and root finding. But for all practical needs, model fitting and inference could be restricted to a dense grid of $\tau \in \{t_1, \ldots, t_L\}$, for which finding $\tau_{x_i}(y_i)$ requires only a simple sequential search involving trapezoidal approximations to the integral of $\dot{\beta}_0(\tau) + x^T\dot{\beta}(\tau)$. Algorithm 1 presents a pseudo-code for likelihood evaluation involving only simple matrix and vector multiplication. The code runs extremely fast when implemented in any low-level programming language with quick "for loops". In our numerical studies we used a C implementation which offered 1000 likelihood evaluations in 2 seconds on an Intel(R) Core(TM) i7-3770 machine with $n = 1000$, $p = 7$ and a grid over τ with mesh size 0.01.

A practical issue with a discrete grid of τ is that it needs to cover the image of the data range mapped into the quantile space, while ensuring the grid length L remains manageable. In our implementations we chose a data dependent grid as follows. We used equispaced grid points between $\tau = 0.01$ and $\tau = 0.99$ with an increment of 0.01. Next, on the upper tail, we augmented the grid with new grid points 0.995, 0.9975, $\cdots$ until we covered $\tau = 1 - 1/(2n)$ where n is the sample size. Same augmentation strategy with geometrically reducing increment lengths were adopted on the lower tail to reach up to $\tau = 1/(2n)$. Here, we use finer grids in the tails since quantiles $Q_Y(\tau|x)$ will tend to $-\infty$ or ∞ as $\tau \to 0$ or 1 when the support of Y is $(-\infty, \infty)$.

2.4 Specification of $\mathcal{X}$ and $a(b, \mathcal{X})$ computation

Our likelihood evaluation requires repeated calculations of the quantify $a(b, \mathcal{X})$ for many different $b \in \mathbb{R}^p$. The computational complexity of calculating $a(b, \mathcal{X})$ depends greatly on the specification of $\mathcal{X}$. For certain convex sets of regular shape, this calculation can be done analytically. For

example, if $\mathcal{X}$ is a ball centered at 0 with radius r, then $a(b, \mathcal{X}) \equiv r$. As we argued in Section 1.3, in absence of preexisting knowledge of the predictor domain, it is ideal to work with $\mathcal{X}$ = the convex hull of the observed predictor vectors $\{x_1, \ldots, x_n\}$. All numerical examples presented in this paper were worked out with this default choice of $\mathcal{X}$. For this choice, $a(b, \mathcal{X}) = \max_{1 \le i \le n}\{-x_i^T b/\|b\|\}$ because $x \mapsto -x^b/\|b\|$ is a convex function and hence its supremum must be realized at an extreme point of the convex hull. This maximization can be done efficiently, in linear time in n. More details are provided in Algorithm 1.

3. Bayesian inference with hierarchical Gaussian process priors

3.1 Prior specification

We adopt a Bayesian approach to parameter estimation with suitable prior distributions on the model parameters, including the function valued parameters ζ and $w = (w_1, \ldots, w_p)$. It is useful that w_js are completely unrestricted, allowing us to handle them with Gaussian process prior distributions. For handling ζ, we first introduce a constraint free version $w_0 : (0, 1) \to \mathbb{R}$ related to ζ through the "logistic transformation":

$$\zeta(\tau) = \frac{\int_0^\tau e^{w_0(u)} du}{\int_0^1 e^{w_0(u)} du}, \quad \tau \in (0, 1), \tag{10}$$

and use a Gaussian process prior on w_0; see Lenk (1988); Tokdar (2007) for similar uses in density estimation.

Recall that a Gaussian process $g = \{g(\tau) : \tau \in (0, 1)\}$ could be viewed as a random element of the Banach space of real valued functions on $(0, 1)$ equipped with the supremum norm. Every Gaussian process g is characterized by two functions, the mean function $m(\tau) = \mathrm{E}g(\tau)$ and the non-negative definite covariance function $c(\tau, \tau') = \mathrm{Cov}(g(\tau), g(\tau'))$, and we use the label $GP(m, c)$ to denote such a process. When $g \sim GP(m, c)$, for any finite set of points $\{\tau_1, \ldots, \tau_k\}$ the random vector $(g(\tau_1), \ldots, g(\tau_k))$ has a k-variate Gaussian distribution with mean $(m(\tau_1), \cdots, m(\tau_k))^T$ and $k \times k$ covariance matrix with elements $c(\tau_i, \tau_j)$.

Our prior specification can be expressed in the following hierarchical form:

$$w_j \sim GP(0, \kappa_j^2 c^{SE}(\cdot,\cdot|\lambda_j)), \quad j = 0, \ldots, p \tag{11}$$

$$(\kappa_j^2, \lambda_j) \sim \pi_k(\kappa_j^2)\pi_\lambda(\lambda_j), \quad j = 0, \ldots, p \tag{12}$$

$$(\gamma_0, \gamma, \sigma^2) \sim \pi(\gamma_0, \gamma, \sigma^2) \propto \frac{1}{\sigma^2}, \tag{13}$$

where $c^{SE}(\tau, \tau'|\lambda^2) = \exp(-\lambda^2(\tau - \tau')^2)$ is the so-called square exponential covariance function equipped with a rescaling parameter λ (van der Vaart and van Zanten, 2008). This particular choice of the covariance function is motivated by two facts. First, for any fixed $\lambda > 0$, the probability distribution $GP(0, c^{SE}(\cdot,\cdot|\lambda))$ assigns 100% probability to the set of all continuous functions on $(0, 1)$ and hence our prior specification does not *a-priori* rule out any valid specification of the joint linear QR model. Second, λ plays the role of a bandwidth parameter for the sample paths generated from $GP(0, c^{SE}(\cdot,\cdot|\lambda))$, with more wavy paths realized as λ gets larger. In a seminal work, van der Vaart and van Zanten (2009) show that with a suitable prior distributions specified on λ, the resulting rescaled square-exponential Gaussian process prior offers adaptively efficient estimation in nonparametric mean regression and density estimation problems by automatically adjusting λ to attain optimal smoothing.

For specifying π_λ, it is more insightful to fix a small $h > 0$ and consider the quantity $\rho_h(\lambda) = \exp(-h^2\lambda^2)$, which gives the correlation between $w_j(\tau)$ and $w_j(\tau + h)$ given $\lambda_j = \lambda$, and assign $\rho_h(\lambda)$ a $Be(a_\lambda, b_\lambda)$ prior. In our applications we use $h = 0.1$, $a_\lambda = 6$ and $b_\lambda = 4$, which assigns 95% mass to $\rho_{0.1}(\lambda) \in (0.3, 0.86)$. However, in our experience, the method shows little sensitivity to these choices. We take π_κ to be $IG(a_\kappa, b_\kappa)$, the inverse gamma pdf with shape a_κ and rate b_κ. The inverse gamma choice allows us to integrate out all κ_j parameters at the time of model fitting. In our applications, we used a 'diffuse' choice $a_\kappa = b_\kappa = 0.1$.

Our choice of π_κ and the right Haar prior on the location scale parameters $(\gamma_0, \gamma, \sigma^2)$ is partially motivated by our numerical experimentations in which we found these choices to lead to

estimates and credible intervals most similar to the Koenker-Basette estimates and confidence intervals. Other reasonable choices could be made and we discuss in Section 7 choices that offer useful shrinkage properties.

When no special information is available about the support of Y, we take f_0 to be a Student t-distribution with an unknown degrees of freedom parameter ν and assign $\nu/6$ a standard logistic prior distribution. The logistic prior is reasonably diffuse and helps the resulting method adapt well to a wide spectrum of tail behavior of the response distribution.

3.2 Model fitting via discretization and adaptive blocked Metropolis

With likelihood evaluation discretized over a grid of τ values $\{t_1, \ldots, t_L\}$ as in Algorithm 1, the curve valued parameters w_j, $1 \leq j \leq p$ are needed to be tracked only over the specified grid, reducing each curve to a parameter vector of length L. The same applies to w_0 from which $\dot{\zeta}$ and ζ could be obtained on the grid by using the trapezoidal rule of integration. While it is theoretically possible to fit the model by running a Markov chain Monte Carlo over these parameters vector and the other model parameters, such a strategy is not entirely practicable. The parameter vector derived from any w_j is conditionally an L dimensional Gaussian variable given λ_j and κ_j, and evaluating its log prior density requires factorizing or inverting a $L \times L$ covariance matrix which has an $O(L^3)$ computing complexity. Furthermore, a Markov chain sampler that operates on both these parameter vectors and the rescaling parameters λ_js run into serious mixing problems.

To overcome these difficulties, we use two sets of further discretization. First, we replace π_λ with a dense, discrete approximation covering the range $\rho_{0.1}(\lambda) \in (0.05, 0.95)$. Let π^*_λ denote the approximating probability mass function with support points $\{\lambda^*_1, \ldots, \lambda^*_G\}$. We choose the support points to be more densely packed for smaller λ values, the rational behind this and the exact manner in which the grid is chosen are discussed in Appendix B.2.

Next, we fix a set of uniformly spaced knots $\{t^*_1, \ldots, t^*_m\} \subset [0, 1]$, for some m much smaller than

L and replace each w_j curve with

$$\tilde{w}_j(\tau) := E\{w_j(\tau)|w_j(t_1^*), \ldots, w_j(t_m^*)\}, \tau \in (0,1), \quad (14)$$

which provides an interpolation approximation to w_j over $(0,1)$, passing through the points $(t_k^*, w_j(t_k^*))$, $k = 1, \ldots, m$, and determined entirely by the m-dimensional vector $W_{j*} = (w_j(t_1^*), \ldots, w_j(t_m^*))^T$, whose prior density evaluations require only $O(m^3)$ flops. Such interpolation based low rank approximations to Gaussian process priors are widely used in statistics and machine learning literature, see for example, Snelson and Ghahramani (2006); Tokdar (2007); Banerjee et al. (2008b).

Our treatment here, however, differs slightly from the above papers in that we carry out the conditional expectation in (14) after marginalizing out both λ_j and κ_j. Let $\tilde{W}_j$ denote the L-dimensional vector $(\tilde{w}_j(t_1), \ldots, \tilde{w}_j(t_L))^T$ that is needed for the likelihood evaluation. Then we can write,

$$\tilde{W}_j = \sum_{g=1}^{G} p_g(W_{j*}) A_g W_{*j}$$

where A_g denotes the $L \times m$ matrix $C_{o*}(\lambda_g)C_{**}(\lambda_g)^{-1}$ with $C_{o*}(\lambda_g) = ((c^{SE}(t_l, t_k^*|\lambda_g)))_{l,k=1}^{L,m}$ and $C_{**}(\lambda_g) = ((c^{SE}(t_l^*, {}^* t_k|\lambda_g)))_{l,k=1}^{m}$, and $p_g(W_{j*}) \propto \pi_\lambda^*(\lambda_g) p(W_{j*}|\lambda_g)$ with

$$p(W_{j*}|\lambda_g) \propto \pi_\lambda^*(\lambda_g) \left\{1 + \frac{W_{j*}^T C_{**}^{-1}(\lambda_g) W_{j*}}{2b_\kappa}\right\}^{-(a_\kappa + m/2)} \frac{\Gamma(a_\kappa + m/2) b_\kappa^{-m/2}}{\Gamma(a_\kappa)},$$

the multivariate t-density of W_{j*} given $\lambda_j = \lambda_g$. Also notice that the marginal prior density of W_{j*} is precisely $\sum_{g=1}^{G} \pi_\lambda^*(\lambda_g) p(W_{j*}|\lambda_g)$. This marginalization reduces the number of parameters and improves mixing of the Markov chain sampler without adding much to the run time of the chain because the matrices A_g and R_g, $g = 1, \ldots, G$, can be precomputed in the beginning and plugged in as needed during the actual run.

With the help of the above sets of discretization, our joint QR model is entirely determined by the $(m+1)(p+1)+2$ dimensional parameter vector $\theta = (W_{0*}^T, \ldots, W_{p*}^T, \gamma_0, \gamma^T, \sigma^2, \nu)^T$ and model fitting may be carried out by running a Markov chain sampler on θ followed by Monte Carlo approximations of posterior quantities. In our experience, an adaptive blocked Metropolis sampler has worked extremely well, offering fast mixing and reproducible results. For this sampler, we use

$p + 4$ block updates of θ per iteration of the sampler, where the first $p + 1$ blocks are given by $(W_{j*}^T, \gamma_j)^T$, $j = 0, \ldots, p$, the next two blocks are $(\gamma_0, \gamma^T)^T$ and $(\log \sigma^2, \log \nu)^T$, and the last block consists of the entire vector θ. For each block, we perform a random walk Metropolis update governed by a multivariate Gaussian proposal distribution centered at the current realization of the block and with covariance that is slowly adapted to resemble, up to a scalar multiplication, the posterior covariance matrix of the block, where the scalar multiplier is also adapted slowly to achieve a pre-specified acceptance rate. We carry out these updates according to Algorithm 4 in Andrieu and Thoms (2008).

4. Posterior consistency

Frequentist justification of Bayesian methods are often presented in the form asymptotic properties of the posterior distribution. A basic desirable property is posterior consistency: the posterior mass assigned to any fixed neighborhood of the true data generating model element should converge to 1 in probability or almost surely as sample size goes to infinity. More refined evaluations of asymptotic properties emerge through posterior convergence rate calculations, where one considers a sequence of shrinking neighborhoods and calibrates the fastest rate of shrinkage for which the posterior mass assigned to these neighborhoods still converges to 1.

We restrict only to a study of *weak* posterior consistency of the Gaussian process based QR method developed in this paper. For a formal treatment, we consider a stochastic design setting where X_is are drawn independently from a pdf f_X on $\mathcal{X}$. Since any valid specification of the quantile planes $\{Q_Y(\tau|x)\colon \tau \in (0,1), x \in \mathcal{X}\}$ uniquely corresponds to a specification of conditional response densities $\{f_Y(y|x) : y \in \mathbb{R}, x \in \mathcal{X}\}$, it also uniquely corresponds to a bivariate density function $f(x,y) = f_X(x) f_Y(y|x)$ under the stochastic design assumption. Hence our prior specification on the quantile planes induces a prior probability measure Π on the space $\mathcal{F}$ of probability density functions on $\mathcal{X} \times \mathbb{R}$. If $f^*(x,y) = f_X(x) f_Y^*(y|x)$ is the true data generating element in this space, then the posterior is said to be weakly consistent at f^* if $\Pi(U|(X_i, Y_i), i = 1, \ldots, n) \to 1$ almost surely

for every weak neighborhood U of f^* in $\mathcal{F}$.

The celebrated Schwartz Theorem (Schwartz, 1965) provides a fairly sharp sufficient condition for weak posterior consistency of Π at f^*. Let $d_{KL}(p,q) := \int p \log(p/q)$ denote the Kullback-Leibler (KL) divergence. For any $f \in \mathcal{F}$ and $\epsilon > 0$, let $K_\epsilon(f)$ denote the ϵ-KL neighborhood $\{g \in \mathcal{F} : d_{KL}(f,g) < \epsilon\}$. We say that f^* is in the KL support of Π if $\Pi(K_\epsilon(f^*)) > 0$ for all $\epsilon > 0$. Schwartz (1965) proved

Theorem 3 (Schwartz)**.** *The posterior is weakly consistent at $f^* \in \mathcal{F}$ if f^* is in the KL support of Π.*

We show that an f^* with linear conditional quantiles $Q_Y^*(\tau|x) = \beta_0^*(\tau) + x^T\beta^*(\tau)$ belongs to the KL support of Π under mild smoothness and tail conditions. Tail conditions are needed to ensure that $d_{KL}(f_Y^*(\cdot|x), f_Y(\cdot|x)) < \infty$, which holds when $f_Y^*(\cdot|x)$ has tails decaying faster than those of $f_Y(\cdot|x)$, with f generated from Π. With our choice of Π, the tails of $f_Y(\cdot|x)$ are expected to be similar to those of f_0, and hence, a minimum requirement is that the tails of $f_Y^*(\cdot|x)$ decay faster than those of f_0. We make the notion of faster tail decay more precise with the following definitions.

Definition 1. *Let f be a probability density function on $\mathbb{R}$ with quantile function Q. Take $m = Q(\tau_0)$. All statements below are interpreted with respect to a given f_0.*

1. *We say f has a type I left tail if $Q(0) > -\infty$, and, for every $\sigma > 0$,*

$$\frac{\frac{1}{\sigma} f_0(m + \frac{Q(t)-m}{\sigma})}{f(Q(t))} \to c_L(\sigma) \in (0,\infty), \text{ as } t \downarrow 0, \tag{15}$$

   with, $c_L(\sigma) \to 0$ as $\sigma \downarrow 0$.

2. *We say f has a type II left tail if for every $\sigma > 0$, $\frac{1}{\sigma} f_0(m + \frac{Q(t)-m}{\sigma}) / f(Q(t))$ diverges to ∞ as $t \downarrow 0$ and,*

$$u_L(\sigma) := \inf\left\{ t > 0 : \frac{\frac{1}{\sigma} f_0(m + \frac{Q(t)-m}{\sigma})}{f(Q(t))} \le 1 \right\} > 0, \tag{16}$$

   with, $u_L(\sigma) \to 0$ as $\sigma \downarrow 0$.

3. *Same definitions apply to the right tail, with,* $Q(1-t)$ *replacing* $Q(t)$ *in* (15), (16), *and* c_R *and* u_R *denoting the right tail counterparts of* c_L *and* u_L.

Recall that we have taken $f_0 = f_0(\cdot|\nu) = t_\nu$ with a prior on $\nu \in (0, \infty)$. Notice that an f has a type I left tail with respect to any t_ν, when $\text{supp}(f)$ is bounded from below, which is same as saying $Q(0) > -\infty$, and, $f(y)$ is bounded away from zero near $Q(0)$. If $Q(0) > -\infty$ but $f(y) \to 0$ as $y \to Q(0)$ then f has a type II left tail with respect to any t_ν. If $Q(0) = -\infty$ and $f(y)$ decays to zero as $y \to -\infty$ at a polynomial or faster rate, then, f has a type II left tail with respect to t_ν for all $\nu > 0$ sufficiently small. Our assumptions on the tails are mild—for example, when $f_0 = t_\nu$, any distribution with positive probability density function over its compact support has a type I left or right tail, and any student *t*-distribution with degrees of freedom greater than ν (the normal distribution corresponds to $\nu = \infty$) has a type II left and right tail. It is straightforward to see that $d_{KL}(f, f_0) < \infty$ whenever f has tails that are type I or type II with respect to f_0.

It turns out that a type I or II tail condition on $f_Y^*(\cdot|0)$, coupled with some regularity conditions on β_0, β are all that is needed to ensure consistency. Here is a precise statement.

Theorem 4. *Suppose* β_0^*, β^* *are differentiable on* $(0, 1)$. *Also assume* $\dot{\beta}^*/\dot{\beta}_0^*$ *can be extended to a continuous function on* $[0, 1]$, *and, there exists a* $c_0 > 0$ *such that* $\dot{\beta}_0^*(t) + x^T\dot{\beta}^*(t) \geq c_0\dot{\beta}_0^*(t)$ *for all* $x \in \mathcal{X}$ *and* $t \in (0, 1)$. *Then* f^* *belongs to the KL support of* Π *whenever* $f_Y^*(\cdot|0)$ *has type I or II tails with respect to* t_ν *for all small enough* $\nu > 0$.

A proof is given in Appendix A.2. The two regularity conditions on $(\dot{\beta}_0^*, \dot{\beta}^*)$ ensure that the conditional density functions do not exhibit pathological behaviors in the tails. Notice that the basic validity assumption $\dot{\beta}_0^*(t) + x^T\dot{\beta}^*(t) > 0$ for all $t \in (0, 1)$ automatically guarantees that $\dot{\beta}^*(t)/\dot{\beta}_0^*(t)$ is bounded for all t. To see this, notice that $\mathcal{X}$ must contain an open ball of radius $r > 0$ around origin which is an interior point. So, for any $t \in (0, 1)$ with $\dot{\beta}^*(t) \neq 0$, $u := -r\dot{\beta}^*(t)/\|\dot{\beta}^*(t)\| \in \mathcal{X}$, and hence, $0 \leq \dot{\beta}_0^*(t) + u^T\dot{\beta}^*(t) = \dot{\beta}_0^*(t) - r\|\dot{\beta}^*(t)\|$, and hence, $\|\dot{\beta}^*(t)/\dot{\beta}_0^*(t)\| \leq 1/r$. The assumption that $\dot{\beta}_0^*(t) + x^T\dot{\beta}^*(t) \geq c_0\dot{\beta}_0^*(t)$ is also mild: it is equivalent to the requirement that the quantile function $Q_Y(\tau|x)$ is monotone over an ϵ-enlargement $\mathcal{X}_\epsilon$ of the compact and convex set $\mathcal{X}$ for some $\epsilon > 0$,

where $\mathcal{X}_\epsilon = \{y : \inf_{x\in\mathcal{X}} \|y - x\| \leq \epsilon\}$. We remark that Theorem 4 applies when exact Gaussian process priors are specified for w_j, and it is still an open problem of under what conditions, such as on the number and locations of the knots, the interpolation based low rank approximation of the Gaussian process prior in Section 3.2 will also lead to posterior consistency of the method.

5. Numerical Experiments

We present two simulation studies we carried out to assess frequentist performance of the proposed method, benchmarked against the methods by Koenker and Bassett (1978), Bondell et al. (2010), and, Reich et al. (2011). The `quantreg` package in `R` was used to implement the classical method by Koenker and Bassett (1978) and confidence intervals were constructed with 200 bootstrapped samples. For Bondell et al. (2010), we used codes available from the author website[2]. We used the `BSquare` package in `R` to implement Reich et al. (2011) with 6 basis functions, with 20000 Gibbs iterations of which 500 thinned samples from the second half of the chain were used for Monte Carlo approximation of the posterior distribution. We also tried two other versions with 10 and 15 basis functions respectively. But increasing the number of basis functions resulted in a progressively poor performance. For our Gaussian process method, we used 6 equispaced knots $\tau_k^* = (k-1)/5$, $k = 1, \ldots, 6$. The quantity $a(b, \mathcal{X})$ was numerically evaluated by setting $\mathcal{X}$ as the convex hull of the observed predictors (see Algorithm 1 in the Appendix for exact details). We ran the adaptive blocked Metropolis sampler for 20000 iterations with 50% burn-in and used 500 samples from the rest for Monte Carlo. Nearly identical results were obtained with 11 equispaced knots.

Unsurprisingly, none of the joint estimation techniques came close to the algorithmic speed of the Koenker and Bassett (1978) method, but the runtimes of our method were generally quicker than the other two. For the simulation study reported in Section 5.2 with $p = 7$, $n = 1000$, and τ tracked over a dense grid with increment size 0.01, our method took on an average 7 minutes

[2] `http://www4.stat.ncsu.edu/~hdbondel/software.html`

to analyze one dataset, whereas Reich et al. (2011) took 10 minutes, and, Bondell et al. (2010) took 55 minutes. In contrast, the Koenker and Bassett (1978) method needed only 20 seconds per dataset.

5.1 Performance assessment: univariate X

In our first study we simulated 100 synthetic datasets each with $n = 1000$ observations from the model

$$X \sim \text{Uniform}(-1, 1); \quad Q_Y(\tau|X) = 3(\tau - \tfrac{1}{2}) \log \frac{1}{\tau(1-\tau)} + 4(\tau - \tfrac{1}{2})^2 \log \frac{1}{\tau(1-\tau)} X.$$

Since X was one dimensional, the shape of $\mathcal{X}$ was a not an issue. However, the nearly quadratic $\beta_1(\tau)$ function was slightly challenging to estimate. In comparing pointwise mean absolute estimation errors and coverage of Figure 2 shows comparisons of pointwise mean absolute estimation errors and the coverage of the 95% confidence/credible bands, averaged across the 100 synthetic datasets. Our method offered lowest estimation errors over the entire range of τ values, and a consistently high coverage close to the nominal target of 95% even into the tails where the methods by Bondell et al. (2010) and Reich et al. (2011) appeared to drop coverage dramatically. It is important to remember that the credible bands produced by our method and that of Reich et al. (2011) are calibrated in a Bayesian way, and so 95% credible bands are not automatically guaranteed to offer 95% coverage. But for general purpose Bayesian methods, some agreement between Bayesian and frequentist calibrations are certainly desirable.

5.2 Performance assessment: multivariate X

For assessing performance in the multivariate case, we ran another simulation study with synthetic data generated from the model:

$$X \sim \text{Uniform}(\{x \in \mathbb{R}^7 : \|x\| \leq 1\}); \quad Q_Y(\tau|X) = \beta_0(\tau) + X^T\beta(\tau)$$

with β_0 and β specified by the equations

$$\beta_0(0.5) = 0, \quad \beta(0.5) = \begin{pmatrix} 0.96 & -0.38 & 0.05 & -0.22 & -0.80 & -0.80 & -5.97 \end{pmatrix}^T,$$

$$\dot{\beta}_0(\tau) = \frac{1}{\tau(1-\tau)}; \quad \dot{\beta}(\tau) = \frac{\dot{\beta}_0(\tau) v(\tau)}{\sqrt{1 + \|v(\tau)\|^2}}, \quad \tau \in (0,1),$$

where $v_j(\tau) = \sum_{l=0}^{2} a_{lj} \cdot \phi(\tau; l/2, 1/(3^2))$, $1 \leq j \leq 7$, with $\phi(\cdot|\mu, \sigma^2)$ denoting the $N(\mu, \sigma^2)$ density function and

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

These specifications define a valid model by Theorem 1 because $a(b, \mathcal{X}) = 1$ for any non-zero b when $\mathcal{X}$ is the unit ball centered at zero. Also note that with these specifications, $Q_Y(\tau|0)$ is precisely the quantile function of the standard logistic distribution. For simulating an (X, Y) from the model we set $X = U_1 Z/\|Z\|$ and $Y = Q_Y(U_2|X)$ where $U_1 \sim \text{Beta}(8,1)$, $U_2 \sim \text{Uniform}(0,1)$ and $Z \sim N_7(0, I_7)$, drawn independently of each other. We evaluated each instance of $Q_Y(U_2|X)$ to a precision of 10^{-16} by numerically integrating $\dot{\beta}_0$ and $\dot{\beta}$ between 0.5 and U_2 with the integrate() function in R.

Figure 3 compares the estimation error and coverage of the three methods averaged across 100 datasets of size $n = 200$ generated from the above simulation model. Figure 4 presents the same but from 100 datasets of size $n = 1000$ each. Again, our method offered nearly the lowest estimation errors for either sample size across all τ values. The coverage of its 95% credible bands was a little below the nominal value for $n = 200$ but improved considerably for $n = 1000$. As in the univariate case, our method offered robust estimation and coverage into the tails. The method by Reich et al. (2011) did not perform very well. We noticed a drop in the coverage of its 95% credible bands when n was increased from 200 to 1000. This could have been due to a poorer mixing of the Gibbs

sampler for the larger data size. The method by Bondell et al. (2010) performed sub-optimally compared to our method or the classical Koenker and Bassett (1978) approach. It particularly suffered at the tails: failing to provide estimates at $\tau = 0.01$ or $\tau = 0.99$ for $n = 200$, and, generally dropping coverage at τ close to 0 or 1, even for the larger sample size of $n = 1000$.

6. Case studies

6.1 Plasma concentration of beta-carotene

Nierenberg et al. (1989) present a study of the association of beta-carotene plasma concentrations with dietary intakes and drugs use for nonmelanoma skin cancer patients. The Statlib database (`http://lib.stat.cmu.edu/datasets/Plasma_Retinol`) hosts a subset of the data from 315 patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. This dataset has been analyzed in the literature (Kai et al., 2011) to assess how personal characteristics, smoking and dietary habits as well as dietary intake of beta-carotene affects concentration levels of beta-carotene in the plasma.

We analyzed the same data with our joint QR model with plasma beta-carotene concentration (ng/ml) as the response and 11 covariates consisting of age (years), sex (1=Male, 2=Female), smoking status (1=Never, 2=Former, 3=Current Smoker), Quetelet index or BMI (weight/(height^2)), vitamin use[3] (1=No, 2=Yes, not often, 3=Yes, fairly often), and daily consumption of calories, fat (g), fiber (g), alcohol (number of drinks), cholesterol (mg) and dietary beta-carotene (mcg). These covariates gave rise to 13 predictors when the categorical variables (sex, smoking status and vitamin use) were coded with dummy indicators. Estimated intercept and slope curves, with 95% credible bands are shown in Figure 5. We used $k = 6$ knots as in Section 5. We ran two parallel Markov chain samplers, each for 100,000 iterations, to ascertain convergence. Estimates were obtained by merging together 1000 thinned draws each from the latter halves of the two chains. That

[3] relabeled for better clarity as '3 – the original label'

longer runs were needed for good mixing than in the simulation setting of Section 5.2 was perhaps due to a stronger correlation between the predictors in the present case.

The estimated intercept curve strongly suggests a longer right tail for the response distribution. The slope curve estimates indicate that being female, use of vitamin and consumption of fiber have reasonably strong positive effect on plasma concentration of beta-carotene, whereas, smoking and BMI have reasonably strong negative effect. Calories, fat, alcohol or cholesterol consumption appears to have little effect. Dietary intake of beta-carotene appears to have a positive effect, but the inference is not conclusive. The slope estimates in Figure 5 suggest more dramatic effects of some predictors on the upper quantiles, but the credible bands paint a more modest picture. However, credible bands for $\beta_j(0.9) - \beta_j(0.1)$ and $\beta_j(0.9) - \beta_j(0.5)$, constructed directly from the posterior draws, indeed suggest more enhanced positive effect of heavy vitamin use at the upper quantiles (Table 1).

We also performed a sixteen fold validation study to assess how well our joint model captured the intricacies of the beta-carotene data. In each fold of the study, we randomly partitioned the 315 observations into training and test sets at roughly 2:1 ratio. We fitted our joint model on the training data and obtained estimates $\hat{\beta}_j$ of β_j, $j = 0, \ldots, p$ in the form of posterior means. These estimates were then used to evaluate the training and test data "check" loss at every $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$ by averaging $\rho_\tau(Y_i - \hat{\beta}_0(\tau) - X_i^T\hat{\beta}(\tau))$ over, respectively, all training and all test set observations (X_i, Y_i), where $\rho_\tau(r) = r\{\tau - I(r < 0)\}$. The same was done with Koenker-Bassette, Bondell et al. (2010), Reich et al. (2011) and standard least squares estimates for linear regression. The relative accuracy of a method at any τ was calculated as the reciprocal of its check loss at that τ relative to the least square method. Our method offered the best test data accuracy across all τ values and maintained its advantage over least squares at the upper quantiles where the other three quantile regression methods appeared to suffer a sharp loss of efficiency (Figure 5).

6.2 Survival analysis under right censoring

Joint estimation of quantile regression parameters could be particularly beneficial for survival analysis with censored response. A greater borrowing of information may help cover the information gaps left by censoring. A crossing-free estimation of the quantile functions means that the estimated survival curves are proper and interpretable. Also, a joint estimation offers an automatic way to quantify estimation uncertainty of the entire survival curves by simple inversions of estimated quantile functions. The probabilistic modeling framework of our joint quantile regression approach makes it particularly straightforward to handle right-censoring. The log-likelihood score calculation (9) now changes to

$$\sum_i [(1-c_i)\log f_Y(y_i|x_i) + c_i \log\{1 - F_Y(y_i|x_i)\}]$$
$$= \sum_i \left[c_i \log\{1-\tau_{x_i}(y_i)\} - (1-c_i)\log\{\dot\beta_0(\tau_{x_i}(y_i)) + x_i^T \dot\beta(\tau_{x_i}(y_i))\}\right], \qquad (17)$$

where c_i is the censoring status (1= right censored, 0 = observed). With this single change, the same prior specification and Markov chain Monte Carlo parameter estimation as detailed in Section 3 remain applicable.

We illustrate these points with a reanalysis of the University of Massachusetts Aids Research Unit IMPACT Study data (UIS, Hosmer and Lemeshow, 1998, Table 1.3) in which we estimated the conditional quantiles of the logarithm of the time to return to drug use (Y) as linear functions of current treatment assignment (TREAT, 1 = Long course, 0 = Short course), number of prior drug treatments (NDT), recent intravenous drug use (IV3, 1 = Yes, 0 = No), Beck depression score (BECK), a compliance factor measuring length of stay in the treatment relative to the course length (FRAC), race of the subject (RACE, 1 = Non-white, 0 = White), age (AGE) and treatment site (SITE). For model fitting, we used the 575 complete observations available in the uis data set of the R package quantreg. Return times were right censored for 111 of these subjects.

Figures 6-7 show parameter and survival curves (for 9 randomly chosen subjects) estimation

with our joint quantile regression approach and also with the censored quantile regression approach described in Koenker (2008). The latter was implemented by using the crq function in R-package quantreg which uses a technique by Portnoy (2003). For our method, we fixed the base probability density f_0 to be $N(0,1)$ instead of a t_ν, since the tails of the distribution of log return time are expected to be fast decaying. We used $k = 6$ knots for the Gaussian process implementation and derived estimates from a Markov chain sampler run of 20000 iterations.

The two sets of parameter estimates are comparable, except in the upper tails. The Portnoy method fails to produce an estimate beyond $\tau = 0.88$, and confidence intervals get extremely wide for τ close to this limit. In contrast, the credible bands from joint estimation are much more stable across the entire range of τ. Estimated survival curves are remarkably similar, though for the Portnoy method, the issue of quantile crossing manifests in the form of estimated survival curves that are not strictly decreasing (e.g., subject # 313 in Figure 7).

7. Discussion

We have introduced a complete and practicable theoretical framework for simultaneous estimation of linear quantile planes in any dimension and over arbitrarily shaped convex predictor domains. Although we have pursued here a specific estimation procedure, our modeling platform is extremely broad and parameter estimation could be done in a variety of other manners. For example, one could choose to use spline based estimation of the basic functions $w_0, \ldots, w_p$ via penalized likelihood maximization or Bayesian averaging. Also, a variety of specifications could be used on the diffeomorphism parameter ζ, e.g., one could model ζ as a mixture of beta cumulative distribution functions, or try estimating ζ directly by adding isotonic regression type constraints.

A number of interesting features could be added to the Bayesian parameter estimation method we have pursued here. An important consideration is shrinkage for large p. For moderately large p, any standard shrinkage prior could be used on γ, and the resulting posterior could be explored by the same Markov chain sampler as in Section 3.2 as long as the prior density on γ is available

in an explicit form up to a normalizing constant. Shrinkage could also be applied on the curve valued parameters w_j, $j = 1, \ldots, p$, by choosing appropriate prior distributions on $(\kappa_1^2, \ldots, \kappa_p^2)$. An attractive choice is to replace the single gamma prior distribution we used in Section 3 with a spike-slab type mixture of gamma distributions, e.g., $\kappa_j^{-2} \sim 0.5Ga(a_\kappa, b_\kappa) + 0.5Ga(a_\kappa, b_\kappa/100)$. Such a specification still allows integrating out κ_j in (14) and hence could be explored by the same Markov chain sampler as before.

The primary computational bottleneck of our method is that the likelihood evaluation involves a search over the grid of τ values for each observation. While our current implementation easily scales to thousands of observations, scaling it to even larger datasets will require further computing innovations. Fortunately, the likelihood evaluation is embarrassingly parallel in the observations and involves very simple arithmetic operations, and thus, it should be possible to obtain many fold speed ups by the use of graphics processing units; such an implementation is currently underway.

Acknowledgments

This research was supported by grants ES017436 and R01ES020619 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). We thank the Associate Editor and the referees for their suggestions to improving the article.

A. Technical details

This section presents a proof of Theorem 4, starting with a few fundamental results that allow comparing two probability density functions given information on their corresponding quantile density functions. We adopt the following notation in the remainder of this section: by a 'probability function quartet' we mean a four-tuple (Q, q, F, f) of real valued functions where $Q : (0, 1) \to \mathbb{R}$ is a non-atomic quantile function that admits a strictly positive derivative $q = \dot{Q}$, $F = Q^{-1}$ is the associated cumulative distribution function with probability density function $f = \dot{F}$. Recall the identities $f(y) = 1/q(F(y))$ and $q(t) = 1/f(Q(t))$.

A.1 Auxiliary results

Lemma 5. *Let* (Q_1, q_1, F_1, f_1) *and* (Q_2, q_2, F_2, f_2) *be two probability function quartets and take* $m_j = Q_j(\tau_0)$, $j = 1, 2$. *If there exist* $0 < c_1 \leq 1 \leq c_2 < \infty$ *such that* $c_1 q_2(t) \leq q_1(t) \leq c_2 q_2(t)$, *for all* $t \in (0, 1)$, *then,*

$$f_2(y) = f_1(y + \Delta_1(y)) \cdot \Delta_2(y), \quad \textit{for all } y \in \text{supp}(f_2),$$

for two real valued functions Δ_1, Δ_2 *satisfying* $|\Delta_1(y)| \leq \max(1 - c_1, c_2 - 1)|y - m_2| + |m_1 - m_2|$, *and,* $\Delta_2(y) \in [c_1, c_2]$.

Proof. By the assumption on q_1, q_2, for every $y \in \text{supp}(f_2)$,

$$c_1 y + b_1 \leq Q_1 F_2(y) \leq c_2 y + b_2,$$

where $b_1 = m_1 - c_1 m_2$, $b_2 = m_1 - c_2 m_2$. Then, $\Delta_1(y) := Q_1 F_2(y) - y$ satisfies, $||\Delta_1(y)|| \leq \max(1 - c_1, c_2 - 1)|y - m_2| + |m_1 - m_2|$. Since, $f_i(y) = 1/q_i(F_i(y))$, $i = 1, 2$, we have, for any $y \in \text{supp}(f_2)$, $f_1(Q_1(F_2(y)))q_1(F_2(y)) = 1$, and hence,

$$f_2(y) = \frac{f_1(Q_1(F_2(y)))q_1(F_2(y))}{q_2(F_2(y))} = f_1(y + \Delta_1(y)) \cdot \frac{q_1(F_2(y))}{q_2(F_2(y))},$$

which proves the result. □

Lemma 6. *Let* f^* *satisfy the conditions of Theorem 4. Given any* $\delta, \sigma, c_1, c_2 > 0$, *there exists an* $\epsilon > 0$ *such that ,*

$$\sup_{x \in \mathcal{X}} \int_{[Q_Y^*(\epsilon|x), Q_Y^*(1-\epsilon|x)]} f_Y^*(y|x) \left| \log\{ f_0(y/\sigma + \Delta(y))/\sigma \} \right| dy < \delta$$

for every $\Delta : \mathbb{R} \to \mathbb{R}$ *satisfying* $|\Delta(y)| < c_1|y| + c_2$ *for all* $y \in \mathbb{R}$.

Proof. By the assumption on f^*, $q_Y^*(t|x)/q_Y^*(t|0) = 1 + x^T \dot{\beta}^*(t)/\dot{\beta}_0^*(t)$ is bounded away from zero and infinity. Hence, by Lemma 5, there are constants $a, b > 0$ such that for every $x \in \mathcal{X}$, $Q_Y^*(F_Y^*(y|0)|x) = y + \Delta_{1,x}(y)$, with $|\Delta_{1,x}(y)| \leq a|y| + b$ for all $y \in A_0 := \text{supp}(f_Y^*(\cdot|0))$. Fix any

$x \in \mathcal{X}$. By the change of variable $z = Q_Y^*(F_Y^*(y|x)|0)$,

$$\begin{aligned}\int f_Y^*(y|x) \left|\log f_0\left(\frac{y}{\sigma} + \Delta(y)\right)\right| dy \\ &= \int_{A_0} f_Y^*(z|0) \left|\log f_0\left(\frac{Q_Y^*(F_Y^*(z|0)|x)}{\sigma} + \Delta(Q_Y^*((F_Y^*(z|0)|x))\right)\right| dz \\ &= \int_{A_0} f_Y^*(z|0) \left|\log f_0(z/\sigma + \Delta_2(z))\right| dz,\end{aligned}$$

with $|\Delta_2(z)| \leq |\Delta_1(z)|/\sigma + |\Delta(z + \Delta_1(z))| \leq a_1|z| + b_1$ where $a_1, b_1 > 0$ depend only on δ, σ, c_1 and c_2. The tail assumption on $f_Y^*(\cdot|0)$ implies that the last integral is finite, proving the result! □

Lemma 7. *Fix $\gamma_0 \in \mathbb{R}$, $\gamma \in \mathbb{R}^p$, $\sigma > 0$, $w : (0,1) \to \mathbb{R}^p$, and two differentiable, monotonically increasing functions $\zeta, \zeta^\dagger : [0,1] \to [0,1]$, with $[\zeta(0), \zeta(1)] \subset [\zeta^\dagger(0), \zeta^\dagger(1)]$. Let $(\beta_0, \beta) = \mathcal{T}(\gamma_0, \gamma, \sigma, w, \zeta)$, $(\beta_0^\dagger, \beta^\dagger) = \mathcal{T}(\gamma_0^\dagger, \gamma^\dagger, \sigma, w, \zeta^\dagger)$, where,*

$$\gamma_0^\dagger = \gamma_0 + \sigma \int_{\zeta(\tau_0)}^{\zeta^\dagger(\tau_0)} q_0(u)du, \quad \gamma^\dagger = \gamma + \sigma \int_{\zeta(\tau_0)}^{\zeta^\dagger(\tau_0)} q_0(u)h(u)du,$$

with $h(\tau) := w(\tau)/\{a(w(\tau), \mathcal{X})\sqrt{1 + ||w(\tau)||^2}\}$, $\tau \in (0,1)$. Fix any $x \in \mathcal{X}$ and consider the probability function quartets $(Q_Y(\cdot|x), q_Y(\cdot|x), F_Y(\cdot|x), f_Y(\cdot|x))$, $(Q_Y^\dagger(\cdot|x), q_Y^\dagger(\cdot|x), F_Y^\dagger(\cdot|x), f_Y^\dagger(\cdot|x))$ where $Q_Y(\tau|x) = \beta_0(t) + x^T\beta(\tau)$, $Q_Y^\dagger(\tau|x) = \beta_0^\dagger(\tau) + x^T\beta^\dagger(\tau)$. Then, $f_Y(y|x)/f_Y^\dagger(y|x) = \dot{\zeta}^\dagger(F_Y^\dagger(y|x))/\dot{\zeta}(F_Y(y|x))$ for all $y \in \text{supp}(f_Y(\cdot|x))$.

Proof. Let $\tau_1 = \zeta(\tau_0)$ and, define,

$$q_0(\tau|x) = q_0(\tau)\{1 + x^T h(\tau)\}, \quad \tau \in (0,1), x \in \mathcal{X}.$$

Then $q_0(\tau|x)$ is strictly positive, and, hence,

$$Q_0(\tau|x) := \int_{\tau_1}^{\tau} q_0(u|x)du, \quad \tau \in (0,1), x \in \mathcal{X},$$

defines valid quantile planes on $\mathcal{X}$. Denote the associated conditional distribution and density functions by $F_0(\cdot|x)$ and $f_0(\cdot|x)$. By definition of (β_0, β), $q_Y(\tau|x) = \dot{\beta}_0(\tau) + x^T\dot{\beta}(\tau) =$

$\sigma q_0(\zeta(\tau))\left\{1 + x^T h(\zeta(\tau))\right\}\dot{\zeta}(\tau) = \sigma q_0(\zeta(\tau)|x)\dot{\zeta}(\tau)$, and, hence,

$$Q_Y(\tau|x) = \gamma_0 + x^T\gamma + \int_{\tau_0}^{\tau} q_Y(u|x)du = \gamma_0 + x^T\gamma + \sigma Q_0(\zeta(\tau)|x). \quad (18)$$

Similarly, $q_Y^\dagger(\tau|x) = \sigma q_0(\zeta^\dagger(\tau)|x)\dot{\zeta}^\dagger(\tau)$, and, hence,

$$Q_Y^\dagger(\tau|x) = \gamma_0^\dagger + x^T\gamma^\dagger + \int_{\tau_0}^{\tau} q_Y^\dagger(u|x)du = \gamma_0 + x^T\gamma + \sigma Q_0(\zeta^\dagger(\tau)|x). \quad (19)$$

by the definitions of $\gamma_0^\dagger$, $\gamma^\dagger$. Inverting (18), we get, for every $x \in \mathcal{X}$,

$$\zeta(\tau) = F_0\left(\frac{Q_Y(\tau|x) - \gamma_0 - x^T\gamma}{\sigma}\middle| x\right), \tau \in (0,1).$$

Therefore, if $y \in (Q_Y(0|x), Q_Y(1|x))$, then,

$$f_Y(y|x) = \frac{1}{q_Y(F_Y(y|x)|x)} = \frac{1}{\sigma q_0(\zeta(F_Y(y|x))|x)\dot{\zeta}(F_Y(y|x))} = \frac{f_0(\frac{y-\gamma_0-x^T\gamma}{\sigma}|x)}{\sigma\dot{\zeta}(F_Y(y|x))}.$$

Similarly, $f_Y^\dagger(y|x) = f_0(\frac{y-\gamma_0-x^T\gamma}{\sigma}|x)/\{\sigma\dot{\zeta}^\dagger(F_Y^\dagger(y|x))\}$, proving the result! □

A.2 Approximating f^* within assumed model space

Let Π_ν denote the conditional prior distribution on f under Π given ν. Theorem 4 is proved in two stages. Let $\nu_0 > 0$ such that the tails of $f_Y^*(\cdot|0)$ are of type I or II with respect to $f_0(\cdot|\nu)$ for every $0 < \nu \leq \nu_0$. First we show that for any such ν and any given $\delta > 0$, there exists an $f^\dagger \in K_\delta(f^*)$ within our model space with nicely behaved underlying w_j curves. Next we show $\Pi_\nu(f : \|\log(f^\dagger/f)\|_\infty < \delta) > 0$ which leads to the claim of Theorem 4. The following lemma gives a precise statement of the first step.

Lemma 8. *Let f^* satisfy the conditions of Theorem 4. For any small $\delta > 0$ and $0 < \nu \leq \nu_0$, there exists an $f^\dagger \in K_\delta(f^*)$ associated with $(\beta_0^\dagger, \beta^\dagger) = \mathcal{T}(\gamma_0^\dagger, \gamma^\dagger, \sigma^\dagger, w^\dagger, \zeta^\dagger)$ where $w^\dagger : [0,1] \to \mathbb{R}^p$ is bounded continuous and $\zeta^\dagger : [0,1] \to [0,1]$ is a diffeomorphism with $\dot{\zeta}^\dagger(t) \in [e^{-B}, e^B]$ for all $t \in [0,1]$ for some finite $B > 0$.*

Proof. Fix a $\nu \in (0, \nu_0)$ and a $\delta \in (0, \tau_0)$. All calculations below are carried out for this particular value of ν and we suppress ν from the notation $f_0(\cdot|\nu)$.

Let $\gamma_0^* = \beta_0^*(\tau_0)$. Fix a $\sigma_L > 0$ such that $c_L(\sigma_L) \leq 1/2$ if $f_Y^*(\cdot|0)$ has a type I left tail with respect to f_0, or, $u_L(\sigma_L)\log\{1/u_L(\sigma_L)\} \leq \delta/2$ if the left tail is of type II. Similarly fix σ_R and take $\sigma^* = \min\{\sigma_L, \sigma_R\}$. Define $\zeta^* : [0,1] \to [0,1]$ as

$$\zeta^*(t) = F_0\left(\gamma_0^* + \frac{\beta_0^*(t) - \gamma_0^*}{\sigma^*}\right), t \in [0,1],$$

which is differentiable and monotonically increasing, and, whose derivative can be written as,

$$\dot{\zeta}^*(t) = \frac{1}{\sigma^*} f_0\left(\gamma_0^* + \frac{\beta_0^*(t) - \gamma_0^*}{\sigma^*}\right)\dot{\beta}_0^*(t) = \frac{(1/\sigma^*) f_0\left(\gamma_0^* + \frac{\beta_0^*(t)-\gamma_0^*}{\sigma^*}\right)}{f_Y^*(\beta_0^*(t)|0)},\ t \in (0,1),$$

since $\beta_0^*(t) = Q_Y^*(t|0)$.

Because ζ^* has a continuously differentiable inverse on $[\zeta^*(0), \zeta^*(1)]$, the relation $h^*(\zeta^*(u)) = \dot{\beta}^*(u)/\dot{\beta}_0^*(u)$ defines a map $h^* : [\zeta^*(0), \zeta^*(1)] \to \mathbb{R}^p$ that is bounded and continuous by the assumption of Theorem 4, and hence, can be extended to a bounded continuous function $h^* : [0,1] \to \mathbb{R}^p$. Define $w^* : [0,1] \to \mathbb{R}^p$ as follows, essentially repeating the construction in the "Only if part" of the proof of Theorem 1. If $h^*(t) = 0$ then set $w^*(t) = 0$. Otherwise, take $c(t) = [[1/\{\|h^*(t)\|a(h^*(\tau), \mathcal{X})\}]^2 - 1]^{-1/2}$ and set $w^*(t) = c(t)h^*(t)/\|h^*(t)\|$. By the assumption on $\dot{\beta}^*/\dot{\beta}_0^*$, w^* is a bounded continuous function on $[0,1]$.

By construction $(\beta_0^*, \beta^*) = \mathcal{T}(\gamma_0^*, \gamma^*, \sigma^*, w^*, \zeta^*)$. However this parameter vector may not be in our model space since we may have either $[\zeta^*(0), \zeta^*(1)] \neq [0,1]$ or $\|\dot{\zeta}^*\|_\infty = \infty$. We correct this by introducing a proper diffeomorphism $\zeta^\dagger$ on $[0,1]$ with $\dot{\zeta}^\dagger$ bounded away from 0 and infinity, such that $\zeta^\dagger(t) = \zeta^*(t)$ for $t \in [\delta_L, 1 - \delta_R]$ for suitably chosen small numbers $\delta_L, \delta_R > 0$. This is the crux of the approximation argument.

If the left tail is type I, then $\zeta^*(0) > 0$ and $\lim_{t\downarrow 0}\dot{\zeta}^*(t) = c_L(\sigma^*) \in (0, 1/2]$. So one can fix $\delta_L > 0$ small enough such that $\zeta^*(\delta_L) > \delta_L$, $\dot{\zeta}^*(t) \in (c_L(\sigma^*)/2, 1]$ for all $t \in (0, \delta_L]$, and, $\delta_L \log[4/\{c_L(\sigma^*)\delta_L\}] < \delta/2$. Otherwise, the left tail is type II, and in that case choose $\delta_L = u_L(\sigma^*)$, which automatically ensures $\dot{\zeta}^*(t) \geq 1$ for all $t \in (0,1)$ with $\dot{\zeta}^*(\delta_L) = 1$, and, $\delta_L \log(1/\delta_L) \leq \delta/2$. Since $\zeta^*(0) \geq 0$, we also must have $\zeta^*(\delta_L) \geq \delta_L$. Fix δ_R by repeating the same steps with the right

tail. Define $\zeta^\dagger : [0,1] \to \mathbb{R}$ as,

$$\zeta^\dagger(t) = \begin{cases} \zeta^*(t), & t \in [\delta_L, 1-\delta_R], \\ a_L t^2 + b_L t, & t \in [0, \delta_L) \\ 1 - a_R(1-t)^2 - b_R(1-t), & t \in (1-\delta_R, 1], \end{cases}$$

where,

$$a_L = \frac{\delta_L \dot{\zeta}^*(\delta_L) - \zeta^*(\delta_L)}{\delta_L^2}, \;\; b_L = \frac{2\zeta^*(\delta_L) - \delta_L \dot{\zeta}^*(\delta_L)}{\delta_L},$$
$$a_R = \frac{\delta_R \dot{\zeta}^*(1-\delta_R) - \{1 - \zeta^*(1-\delta_R)\}}{\delta_R^2}, \;\; b_R = \frac{2\{1 - \zeta^*(1-\delta_R)\} - \delta_R \dot{\zeta}^*(1-\delta_R)}{\delta_R}.$$

By choice of δ_L and δ_R, $a_L < 0$, $a_R < 0$ and $b_L \in [\dot{\zeta}^*(\delta_L), 2/\delta_L]$, $b_R \in [\dot{\zeta}^*(1-\delta_R), 2/\delta_R]$. It is straightforward to verify that $\zeta^\dagger$ defines a diffeomorphism from $[0,1]$ onto $[0,1]$, with $\dot{\zeta}^\dagger(t) \in [\dot{\zeta}^*(\delta_L), b_L]$ for all $t \in [0, \delta_L]$ and $\dot{\zeta}^\dagger(t) \in [\dot{\zeta}^*(1-\delta_R), b_R]$ for all $t \in [1-\delta_R, 1]$. Therefore there exists a $B > 0$ such that $\dot{\zeta}^\dagger(t) \in [e^{-B}, e^B]$ for all $t \in [0,1]$.

Take $(\beta_0^\dagger, \beta^\dagger) = \mathcal{T}(\gamma_0^*, \gamma^*, \sigma^*, w^*, \zeta^\dagger)$ with valid conditional quantile planes $Q_Y^\dagger(\cdot|x)$ and associated cumulative distribution and probability density functions given by $F_Y^\dagger(\cdot|x)$ and $f_Y^\dagger(\cdot|x)$. By construction of $\zeta^\dagger$, $Q_Y^\dagger(\tau|x) = Q_Y^*(\tau|x)$ for all $\tau \in [\delta_L, 1-\delta_R]$ and hence $F_Y^\dagger(y|x) = F_Y^*(y|x)$ for all $y \in [Q_Y^*(\delta_L|x), Q_Y^*(1-\delta_R|x)]$. Hence, by Lemma 7,

$$d_{KL}(f_Y^*(\cdot|x), f_Y^\dagger(\cdot|x)) = \int_{y \in Q_Y^*(\delta_L|x), Q_Y^*(1-\delta_R|x)]^c} f_Y^*(y|x) \log \frac{\dot{\zeta}^\dagger(F_Y^\dagger(y|x))}{\dot{\zeta}^*(F_Y^*(y|x))} dy.$$

Split the integral above into two integrals, one over $y < Q_Y^*(\delta_L|x)$ and the other over $y > Q_Y^*(1-\delta_R|x)$. When $y < Q_Y^*(\delta_L|x)]$, both $F_Y^\dagger(y|x) < \delta_L$, and, $F_Y^*(y|x) < \delta_L$. Clearly, $\dot{\zeta}^\dagger(F_Y^\dagger(y|x)) \le b_L \le 2/\delta_L$. If left tail is type I then, $\dot{\zeta}^*(F_Y^*(y|x)) \ge c_L(\sigma^*)/2$ and hence,

$$\int_{Q_Y^*(0|x)}^{Q_Y^*(\delta_L|x)} f_Y^*(y|x) \log \frac{\dot{\zeta}^\dagger(F_Y^\dagger(y|x))}{\dot{\zeta}^*(F_Y^*(y|x))} dy \le \left[\log \frac{4}{\delta_L c_L(\sigma^*)}\right] \int_{Q_Y^*(0|x)}^{Q_Y^*(\delta_L|x)} f_Y^*(y|x) dy$$
$$= \delta_L \log \frac{4}{\delta_L c_L(\sigma^*)} \le \delta/2$$

by the choice of δ_L for the type I left tail. On the other hand, if the left tail is type II, then $\dot{\zeta}^*(F_Y^*(y|x)) \geq 1$ and hence

$$\int_{Q_Y^*(0|x)}^{Q_Y^*(\delta_L|x)} f_Y^*(y|x) \log \frac{\dot{\zeta}^\dagger(F_Y^\dagger(y|x))}{\dot{\zeta}^*(F_Y^*(y|x))} dy \leq \delta_L \log \frac{2}{\delta_L} \leq \delta/2,$$

again by the choice of δ_L for this case. Same arguments apply to the integral over $y \in [Q_Y^*(1-\delta_R|x), Q_Y^*(1|x)]$, and hence, for every $x \in \mathcal{X}$, $d_{KL}(f_Y^*(\cdot|x), f_Y^\dagger(\cdot|x)) \leq \delta$. Therefore $d_{KL}(f^*, f^\dagger) = \int f_X(x) d_{KL}(f_Y^*(\cdot|x), f_Y^\dagger(\cdot|x)) dx \leq \delta$. □

A.3 Proof of Theorem 4

Since the prior on ν has full support, it suffices to show that given any $\delta > 0$ and $\nu < \nu_0$, the conditional prior $\Pi_\nu = \Pi(\cdot|\nu)$ assigns positive mass to the event $\{f : d_{KL}(f^*, f) < 3\delta\}$. Fix any $\delta > 0$, and $\nu < \nu_0$. By Lemma 8, there is a $(\beta_0^\dagger, \beta^\dagger) = \mathcal{T}(\gamma_0^\dagger, \gamma^\dagger, \sigma^\dagger, w^\dagger, \zeta^\dagger)$ with the associated probability density function $f^\dagger$ satisfying $d_{KL}(f^*, f^\dagger) < \delta$, where, $w^\dagger : [0,1] \to \mathbb{R}^p$ is bounded continuous, and, $\zeta^\dagger : [0,1] \to [0,1]$ is a diffeomorphism with $\|\log \dot{\zeta}^\dagger\|_\infty < \infty$.

For any $\lambda > 0$, let A_λ denote the set of $(\gamma_0, \gamma, \sigma, w, \zeta)$ such that $|\gamma_0 - \gamma_0^\dagger| < \lambda$, $\|\gamma - \gamma^\dagger\| < \lambda\sigma^\dagger/\text{diam}(\mathcal{X})$, $|\sigma/\sigma^\dagger - 1| < \lambda$, $w : [0,1] \to \mathbb{R}^p$ is continuous and $\sup_t \|w(t) - w^\dagger(t)\| < \lambda$, $\zeta : [0,1] \to [0,1]$ is a diffeomorphism and $\|\log \dot{\zeta} - \log \dot{\zeta}^\dagger\|_\infty < \lambda$. By construction and because of the full support properties of Gaussian processes (Tokdar and Ghosh, 2007), the conditional prior Π_ν assigns a positive mass to the set of f associated with $(\beta_0, \beta) = \mathcal{T}(\gamma_0, \gamma, \sigma, w, \zeta)$, $(\gamma_0, \gamma, \sigma, w, \zeta) \in A_\lambda$, for every $\lambda > 0$. So, it suffices to show that $\lambda > 0$ could be chosen small enough such that any f associated with a $(\gamma_0, \gamma, \sigma, w, \zeta) \in A_\lambda$ satisfies $\int f_Y^*(y|x) \log\{f_Y^\dagger(y|x)/f_Y(y|x)\} dy \leq 2\delta$ for all $x \in \mathcal{X}$.

Let $b = \|\log \dot{\zeta}^\dagger\|_\infty + 1$. Since $w^\dagger$ is bounded on $[0,1]$, there exists a $B > 0$ such that $h^\dagger(t) := w^\dagger(t)/\{a(w^\dagger(t), \mathcal{X})\sqrt{1 + \|w^\dagger(t)\|^2}\}$ satisfies

$$1 + x^T h^\dagger(t) \in \left[e^{-B}, e^B\right], \text{ for all } t \in [0,1].$$

Clearly, there exists a $\lambda_0 \in (0, 1/2)$ such that $\|\log \dot{\zeta} - \log \dot{\zeta}^\dagger\|_\infty < \lambda_0$ implies $\|\log \dot{\zeta}\|_\infty < 2b$, and, $\sup_t \|w(t) - w^\dagger(t)\| < \lambda_0$ implies $1 + x^T w(t)/\{a(w(t), \mathcal{X})\sqrt{1 + \|w(t)\|^2}\} \in [e^{-2B}, e^{2B}]$ for all $t \in [0,1]$.

Take $c_1 = 1 + e^{2B}$, $c_2 = c_1\{|\gamma_0^\dagger| + \|\gamma^\dagger\| + 1)\} + 1/2$, and, $\tilde{c}_1 = (1 + 2c_1)/\sigma^\dagger$, $\tilde{c}_2 = 2(c_2 + |\gamma_0^\dagger| + 1)/\sigma^\dagger$. By Lemma 6 there exists an $0 < \epsilon < \delta/\max\{12b, 6(B+1)\}$ such that,

$$\sup_{x\in\mathcal{X}} \int_{[Q_Y^*(\epsilon|x), Q_Y^*(1-\epsilon|x)]} f_Y^*(y|x)|\log f_0(y/\sigma^\dagger + \Delta(y))/\sigma^\dagger|dy < \delta/3.$$

for every $\Delta : \mathbb{R} \to \mathbb{R}$ satisfying $|\Delta(y)| < \tilde{c}_1|y| + \tilde{c}_2$ for all $y \in \mathbb{R}$

Take any $(\gamma_0, \gamma, \sigma, w, \zeta) \in A_{\lambda_0}$ and let $(\beta_0, \beta) = \mathcal{T}(\gamma_0, \gamma, \sigma, w, \zeta)$, $(\beta_0^{e\dagger}, \beta^{e\dagger}) = \mathcal{T}(\gamma_0^{e\dagger}, \gamma^{e\dagger}, \sigma^\dagger, w^\dagger, e)$, and, $(\beta_0^e, \beta^e) = \mathcal{T}(\gamma_0^e, \gamma^e, \sigma, w, e)$, where e denotes the identity function on $[0,1]$ onto itself and

$$\gamma_0^{e\dagger} = \gamma_0^\dagger + \sigma^\dagger \int_{\zeta^\dagger(\tau_0)}^{\tau_0} q_0(u)du, \quad \gamma^{e\dagger} = \gamma^\dagger + \sigma^\dagger \int_{\zeta^\dagger(\tau_0)}^{\tau_0} q_0(u)h^\dagger(u)du,$$
$$\gamma_0^e = \gamma_0 + \sigma \int_{\zeta(\tau_0)}^{\tau_0} q_0(u)du, \quad \gamma^e = \gamma + \sigma \int_{\zeta(\tau_0)}^{\tau_0} q_0(u)h(u)du.$$

with $h(t) = w(t)/\{a(w(t), \mathcal{X})\sqrt{1 + \|w(t)\|^2}\}$, $t \in [0,1]$. The definitions of $\gamma_0, \gamma_0^e, \gamma_0^{e\dagger}, \gamma^{e\dagger}$ match the requirements of Lemma 7. Let $(Q_Y(\cdot|x), q_Y(\cdot|X), F_Y(\cdot|x), f_Y(\cdot|x))$ denote the probability function quartet of (β_0, β), and the same symbols with appropriate superscripts denote the same quantities associated with the other three pairs (β_0^e, β^e), $(\beta_0^\dagger, \beta^\dagger)$ and $(\beta_0^{e\dagger}, \beta^{e\dagger})$.

Consider the following factorization in log-scale

$$\log \frac{f_Y^\dagger(y|x)}{f_Y(y|x)} = \log \frac{f_Y^\dagger(y|x)}{f_Y^{e\dagger}(y|x)} + \log \frac{f_Y^e(y|x)}{f_Y(y|x)} + \log \frac{f_Y^{e\dagger}(y|x)}{f_Y^e(y|x)}.$$

By Lemma 7, $|\log\{f_Y^\dagger(y|x)/f_Y^{e\dagger}(y|x)\}| = |\log \dot{\zeta}^\dagger(F_Y^\dagger(y|x))| \le 2b$, and, $|\log\{f_Y^e(y|x)\ /f_Y(y|x)\}| = |\log \dot{\zeta}(F_Y(y|x))| \le 2b$. Since $1 + x^T h(t) \in [e^{-2B}, e^{2B}]$ for all $t \in [0,1]$, we have, for any $x \in \mathcal{X}$, $q_Y^e(t|x) = q_Y^e(t|0) \cdot [e^{-2B}, e^{2B}]$ for all $t \in (0,1)$, and hence, by Lemma 5, $f_Y^e(y|x) = f_Y^e(y + \Delta_{1,x}(y)|0)\Delta_{2,x}(y)$, with $|\Delta_{1,x}(y)| \le c_1|y| + c_2$ for all $y \in \mathbb{R}$, and, $\|\log \Delta_{2,x}\|_\infty \le 2B$. But, $f_Y^e(y|0) = f_0((y - \gamma_0)/\sigma)/\sigma$ and so, $f_Y^e(y|x) = f_0(y/\sigma^\dagger + \tilde{\Delta}_{1,x}(y))\tilde{\Delta}_{2,x}(y)/\sigma^\dagger$ with $|\tilde{\Delta}_{1,x}(y)| \le \tilde{c}_1|y| + \tilde{c}_2$, for all $y \in \mathbb{R}$, and, $\|\log \tilde{\Delta}_{2,x}\|_\infty \le B + 1$. The same calculations work for $f_Y^{e\dagger}$ because $(\gamma_0^\dagger, \gamma^\dagger, \sigma^\dagger, w^\dagger, \zeta^\dagger) \in$

A_{λ_0}. Therefore,

$$\int_{[Q_Y^*(\epsilon|x),Q_Y^*(1-\epsilon|x)]^c} f_Y^*(y|x)\log\frac{f_Y^\dagger(y|x)}{f_Y(y|x)}dy < \delta,$$

for every $x \in \mathcal{X}$.

The map $(x,y) \mapsto \log f_Y^\dagger(y|x)$ is equicontinuous on $\{(x,y) : x \in \mathcal{X}, y \in [Q_Y^*(\epsilon|x), Q_Y^*(1-\epsilon|x)]\}$, and hence there exist a $\kappa > 0$ such that $\log|f_Y^\dagger(y+z|x)/f_Y^\dagger(y|x)| < \delta/2$ for all $x \in \mathcal{X}$, $y \in [Q_Y^*(\epsilon|x), Q_Y^*(1-\epsilon|x)]$, $|z| < \kappa$. Fix a small $0 < \eta < \kappa/2$ such that

$$\max(e^\eta - 1, 1 - e^{-\eta}) \cdot \sup_{x\in\mathcal{X}} \max\{|Q_Y^*(\epsilon|x) - \gamma_0^\dagger - x^T\gamma^\dagger|, |Q_Y^*(1-\epsilon|x) - \gamma_0^\dagger - x^T\gamma^\dagger|\} < \frac{\kappa}{2}.$$

By the equicontinuity of the maps $s \mapsto \log q_0(e^s)$ and $s \mapsto h^\dagger(e^s)$ on the interval $[\log\epsilon, \log(1-\epsilon)]$, and the continuity of the transformation $v \mapsto v/\{a(v,\mathcal{X})\sqrt{1+\|v\|^2}\}$, one can fix $0 < \lambda < \min(\lambda_0, \kappa/4)$ such that for any $(\gamma_0, \gamma, \sigma, w, \zeta) \in A_\lambda$,

$$\frac{q_Y(t|x)}{q_Y^\dagger(t|x)} = \frac{\sigma}{\sigma^\dagger} \times \frac{q_0(\zeta(t))}{q_0(\zeta^\dagger(t))} \times \frac{1 + x^T h(\zeta(t))}{1 + x^T h^\dagger(\zeta^\dagger(t))} \times \frac{\dot\zeta(t)}{\dot\zeta^\dagger(t)} \in [e^{-\eta}, e^\eta],$$

for every $t \in [\epsilon, 1-\epsilon]$ and $x \in \mathcal{X}$. Consequently, by Lemma 5 for every $x \in \mathcal{X}$ and $y \in [Q_Y^*(\epsilon|x), Q_Y^*(1-\epsilon|x)]$, $|\log\{f_Y(y|x)/f_Y^\dagger(y|x)\}| < \delta$. This proves the result.

B. Computational details

B.1 Centering the predictors

A preprocessing step of our method is to center the observed predictors $\{x_1, \ldots, x_n\}$ around an interior point of their convex hull (Figure 8). While the sample mean vector automatically gives an interior point, it may lie too close to the hull boundary and lead to poorer model fit. A better strategy is to use the mean of the extreme points of the data cloud, but finding the extreme points becomes computationally intensive for $p > 2$. Instead, we employ a fast algorithm that recursively identifies $p+1$ points $x_1^*, \ldots, x_{p+1}^*$, from the data cloud that are close to the boundary and far away from each other.

input : Model parameters
scalars: γ_0, σ,
vectors: $\gamma = (\gamma_1, \cdots, \gamma_p)^T$,
functions: $w = (w_1, \cdots, w_p) : (0,1) \to \mathbb{R}^p$, diffeomorphism ζ on $[0,1]$
output: Log-likelihood score

// Basic quantities
for $l = 1 : m$ // Could be parallelized in l
  set $v_l = \omega(\zeta(t_l))$ and $\dot{b}_{0,l} = \sigma q_0(\zeta(t_l))\dot{\zeta}(t_l)$;
  for $i = 1 : n$ calculate $a_i = x_i^T v_l$ calculate $a_\chi = \max_{1 \le i \le n}\{-a_i\}/\sqrt{||v_l||}$;
  for $i = 1 : n$ set $\tilde{a}_{il} = a_i/\{a_\chi \cdot \sqrt{1 + ||v_l||^2}\}$
endfor
// Calculatelog likelihood score by sequencing through obs
set $\ell\ell = 0$; // initialize the log likelihood
for $i = 1 : n$ // Could be parallelized in i
  calculate $Q_0 = \gamma_0 + \gamma^T x_i$;
  if $Y_i > Q_0$ **then**
    set $Q_U = Q_0$ and $l = k$;
    while $Y_i > Q_U$ **do**
      set $Q_L = Q_U$ and $l = l + 1$;
      if $l \le m$ **then** calculate $Q_U = Q_L + (\delta/2) \cdot \{\dot{b}_{0,l-1}(1 + \tilde{a}_{i,l-1}) + \dot{b}_{0,l}(1 + \tilde{a}_{i,l})\}$ **else** set $Q_U = \infty$
    end
  else
    set $Q_L = Q_0$ and $l = k + 1$;
    while $Y_i \le Q_L$ **do**
      set $Q_U = Q_L$ and $l = l - 1$;
      if $l \ge 2$ **then** calculate $Q_L = Q_U - (\delta/2) \cdot \{\dot{b}_{0,l-1}(1 + \tilde{a}_{i,l-1}) + \dot{b}_{0,l}(1 + \tilde{a}_{i,l})\}$ **else** set $Q_L = -\infty$
    end
  end
  if $Q_L = -\infty$ **or** $Q_U = \infty$ **then**
    set $\ell\ell = -\infty$
  else
    calculate $\alpha = (Y_i - Q_L)/(Q_U - Q_L)$;
    set $\ell\ell = \ell\ell - \log\{(1 - \alpha)\dot{b}_{0,l-1}(1 + \tilde{a}_{i,l-1}) + \alpha\dot{b}_{0,l}(1 + \tilde{a}_{i,l})\}$
  end
endfor
return $\ell\ell$

Algorithm 1: Log-likelihood evaluation

Consider a Gaussian process f on $\mathbb{R}^p$ with covariance function $C(x, x') = \exp\{-\|\Delta^{-1}(x-x')\|^2\}$, where Δ is the $p \times p$ diagonal matrix with j-th element equaling the observed range of the j-th predictor, $j = 1, \ldots, p$. Take $x_1^* = x_1$ and recursively select x_j^* as the $x \in \{x_1, \ldots, x_n\}$ with maximum $\mathrm{Var}(f(x)|x_1^*, \ldots, x_{j-1}^*)$, $j = 2, \ldots, p+1$. This recursive selection can be carried out extremely fast, with computational complexity of the order $(p+1)n \log n$ flops, by carrying out a rank-$(p+1)$ incomplete, pivoted Cholesky decomposition of the $n \times n$ non-negative definite matrix $K = ((C(x_i, x_j)))$, for example, by using the inchol function of the R package kernlab. Such implementations depend on the order in which the x_is are stored. To encourage selection close from the boundary, we prearrange the x_is in decreasing order of their Mahalanobis distance $\|S^{-1}(x_i - \bar{x})\|$ from mean $\bar{x}$, where S denotes the sample covariance.

B.2 Choosing λ grid points

In choosing the grid points λ_g, $g = 1, \ldots, G$, for λ_j, it is important to ensure that the conditional prior distributions $N(0, \kappa_j^2 C_{**}(\lambda_g))$ remain sufficiently overlapped for neighboring λ_g values, since otherwise, the grid based discretization of the prior on λ may lead to poor mixing of the Markov chain sampler. If overlap is measured by the Kullback-Leibler divergence $d(\lambda, \lambda') := d_{KL}(N(0, \kappa_j^2 C_{**}(\lambda)), N(0, \kappa_j^2 C_{**}(\lambda')))$, which does not depend on κ_j, it is easy to see that one must use a non-uniform grid of λ values since for a given $\Delta > 0$, $d(\lambda, \lambda + \Delta)$ is much larger for a small λ than a large one. To choose this non-uniform grid, we set λ_1 to be the smallest value in the predetermined range, one that gives $\rho_{0.1}(\lambda_1) = 0.99$, and then increment λ recursively so that $d(\lambda_{g-1}, \lambda_g) = 1$, $g = 2, 3, \ldots$, until the whole range is covered.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics 26*, 247–257.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive mcmc. *Statistical Computing 18*, 343–373.

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

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

Bondell, H. D., B. J. Reich, and H. Wang (2010). Non-crossing quantile regression curve estimation. *Biometrika 97*, 825–838.

Burgette, L., J. Reiter, and M. Miranda (2011). Exploratory quantile regression with many covariates: an application to adverse birth outcomes. *Epidemiology 22*, 859–66.

Dunham, J., B. Cade, and J. Terrell (2002). Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. *Transactions of the American Fisheries Society 131*, 86–98.

Dunson, D. B. and J. A. Taylor (2005). Approximate Bayesian inference for quantiles. *J. Nonparametr. Stat. 17*, 385–400.

Elsner, J. B., J. P. Kossin, and T. H. Jagger (2008). The increasing intensity of the strongest tropical cyclones. *Nature 455*, 92–95.

Feng, Y., Y. Chen, and X. He (2015). Bayesian quantile regression with approximate likelihood. *Bernoulli 21*(2), 832–850.

Gelfand, A. E. and A. Kottas (2003). Bayesian semiparametric regression for median residual life. *Scandinavian Journal of Statistics 30*, 651–665.

Haario, H., E. Saksman, and J. Tamminen (1999). Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistic 14*, 375–395.

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

Hosmer, D. W. and S. Lemeshow (1998). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York, NY: John Wiley and Sons Inc.

Kai, B., R. Li, and H. Zou (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of statistics 39*(1), 305.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press, Cambridge.

Koenker, R. (2008). Censored quantile regression redux. *Journal of Statistical Software 27*, 1–25.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Kottas, A. and M. Krnjajić (2009). Bayesian nonparametric modeling in quantile regression. *Scandinavian Journal of Statistics 36*, 297–319.

Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of American Statistical Association 83*, 509–516.

Nierenberg, D. W., T. A. Stukel, J. A. Baron, B. J. Dain, and E. R. Greenberg (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology 130*(3), 511–521.

Portnoy, S. (2003). Censored quantile regression. *Journal of American Statistical Association 98*, 1001–1012.

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

Schwartz, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 4*, 10–26.

Snelson, E. and Z. Ghahramani (2006). Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems*, Volume 18, Cambrisge, Massachussetts. The MIT Press.

Thompson, P., Y. Cai, R. Moyeed, D. Reeve, and J. Stander (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics and Data Analysis 54*, 1138–1150.

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

Tokdar, S. T. and J. K. Ghosh (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference 137*, 34–42.

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

Tsionas, G. (2003). Bayesian quantile inference. *Journal of Statistical Computation and Simulation 73*, 659–674.

van der Vaart, A. W. and J. H. van Zanten (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Annals of Statistics 36*, 1435–1463.

van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annal of Statistics 37*, 2655–2675.

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

j	Predictor	$\beta_j(0.1)$	$\beta_j(0.5)$	$\beta_j(0.9)$	$\beta_j(0.9)-\beta_j(0.1)$	$\beta_j(0.9)-\beta_j(0.5)$
1	Age	0.46 (−0.17,1.02)	1.11 (0.56,1.57)	2.07 (0.18,4.44)	1.6 (−0.33,4.07)	0.97 (−0.81,3.34)
2	Sex2	23.61 (−0.66,52.76)	20.19 (−0.96,52.1)	28.5 (−28.5,87.6)	4.6 (−50.75,57.83)	6.5 (−42.93,57.27)
3	SmokStat2	1.09 (−17.6,16.41)	−6.46 (−25.57,7.76)	−12.35 (−68.97,38.82)	−13.43 (−69.5,38)	−5.94 (−55.69,43.85)
4	SmokStat3	−11.73 (−40.03,9.97)	−10.62 (−36.61,11.28)	−32.03 (−109.98,22.54)	−19.11 (−94.84,31.17)	−20.9 (−90,26.57)
5	Quetelet	−2.14 (−3.55,−1.11)	−2.61 (−3.72,−1.75)	−4.16 (−7.44,−1.34)	−1.93 (−5.61,0.91)	−1.51 (−4.79,1.02)
6	VitUse1	20.01 (2.56,36.48)	17.66 (1.37,32.01)	33.93 (−7.76,99.93)	14.37 (−26.87,81.3)	16.65 (−22.72,81.83)
7	VitUse2	−0.12 (−18.87,21.84)	18.97 (−3.58,37.57)	112.52 (24.58,205.53)	113.05 (19.86,209.36)	93.53 (12.62,185.14)
8	Calories	0 (−0.04,0.03)	0 (−0.03,0.03)	0 (−0.04,0.04)	0 (−0.02,0.02)	0 (−0.02,0.02)
9	Fat	−0.24 (−0.81,0.46)	−0.23 (−0.82,0.49)	−0.26 (−0.92,0.55)	−0.02 (−0.37,0.36)	−0.02 (−0.37,0.32)
10	Fiber	3.03 (1.08,5.24)	3.1 (1.44,5.06)	2.85 (−0.21,6.29)	−0.16 (−3.31,3.19)	−0.22 (−3.27,2.85)
11	Alcohol	0.01 (−0.94,0.9)	0 (−0.82,0.88)	0.07 (−1.07,1.51)	0.04 (−0.74,1.28)	0.05 (−0.62,1.11)
12	Cholesterol	−0.05 (−0.11,0.02)	−0.04 (−0.12,0.04)	−0.04 (−0.19,0.1)	0 (−0.12,0.13)	0 (−0.12,0.12)
13	BetaDiet	0 (−0.01,0.01)	0.01 (0,0.02)	0.01 (0,0.03)	0.01 (0,0.04)	0.01 (−0.01,0.03)

Table 1: Estimates of slopes and slope differentials at $\tau = 0.1, 0.5, 0.9$. Estimates are posterior medians and the intervals below the estimates are 95% posterior credible intervals.

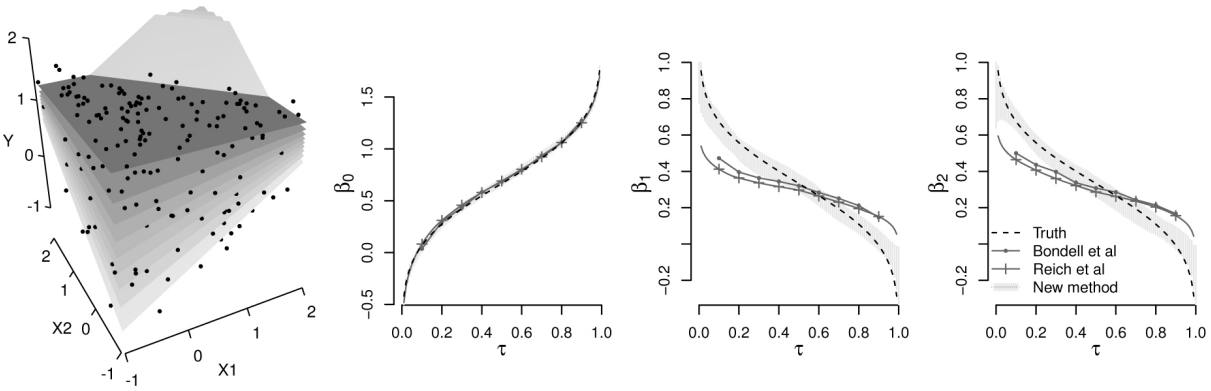

Figure 1: A toy example where $\mathcal{X}$ is the triangle in $\mathbb{R}^2$ with vertices $(-1,1)$, $(-1,2)$ and $(2,-1)$ and $Q_Y(\tau|X) = \frac{1-(X_1+X_2)}{3} Q_N(\tau|0,1) + \frac{2+X_1+X_2}{3} Q_N(\tau|1,0.2^2)$ where $Q_N(\tau|\mu,\sigma^2)$ denotes the τ-th quantile of the $N(\mu,\sigma^2)$ distribution. These quantile hyperplanes (left panel) are correctly ordered within $\mathcal{X}$ but cross each other inside the smallest embedding rectangle $[-1,2]\times[-1,2]$. This negatively impacts slope estimation by the Bondell et al. (2010) and Reich et al. (2011) methods, which cannot adapt to the triangular shape of the predictor space. In contrast, the new method proposed here works on the convex hull of the observed predictors and can retrieve the true parameter curves with a much higher accuracy. The estimates shown here are based off a single synthetic data set of sample size 1000, where X was sampled uniformly from $\mathcal{X}$.

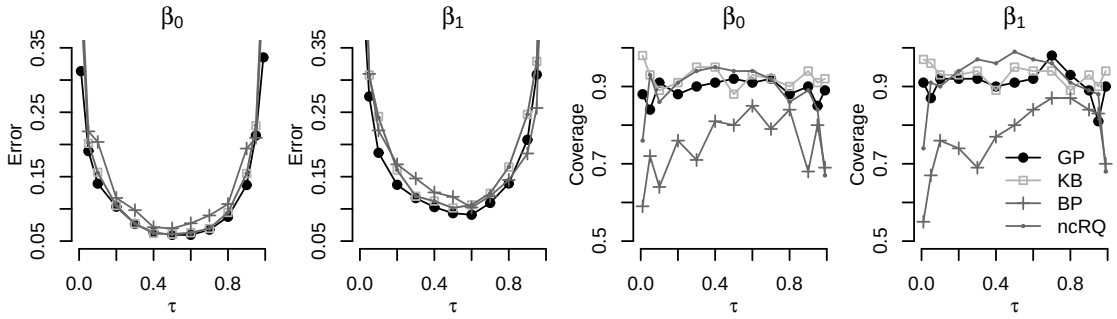

Figure 2: Assessing performance with univariate X. Left two panels show mean absolute estimation errors and right two panels show coverage by 95% confidence or credible bands, for $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$. KB: classical Koenker-Basette estimates; ncRQ: non-crossing regression quantile estimates from Bondell et al. (2010); BP: Bayesian estimates from Bernstein polynomial model of Reich et al. (2011); GP: estimates from the proposed Gaussian process method.

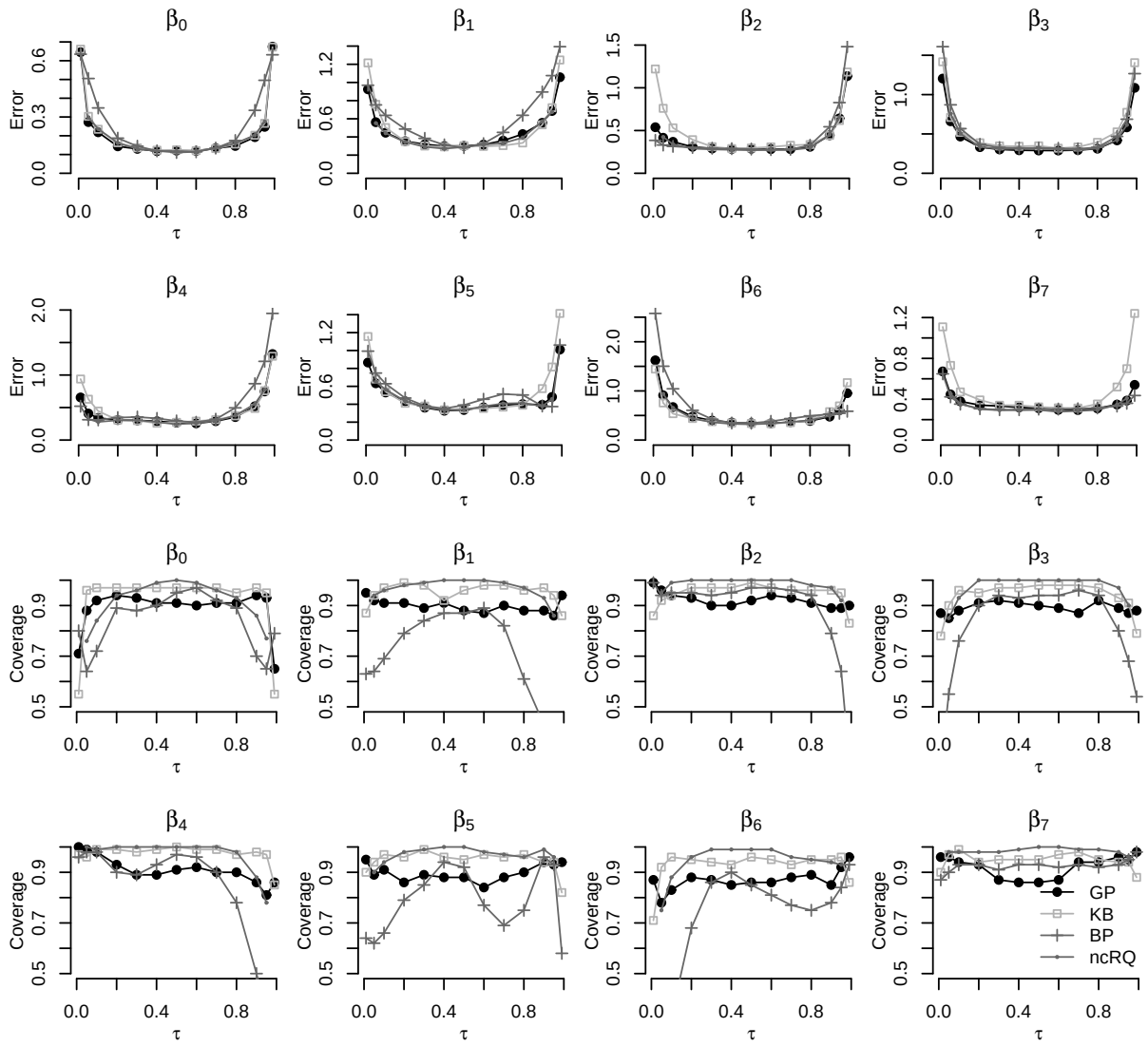

Figure 3: Assessing performance with a 7 dimensional X ($n = 200$). Top two rows show estimation errors and bottom two rows show coverage by 95% confidence or credible bands for $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$. The Bondell et al. (2010) method (ncRQ) could not produce estimates at $\tau = 0.01$ or $\tau = 0.99$.

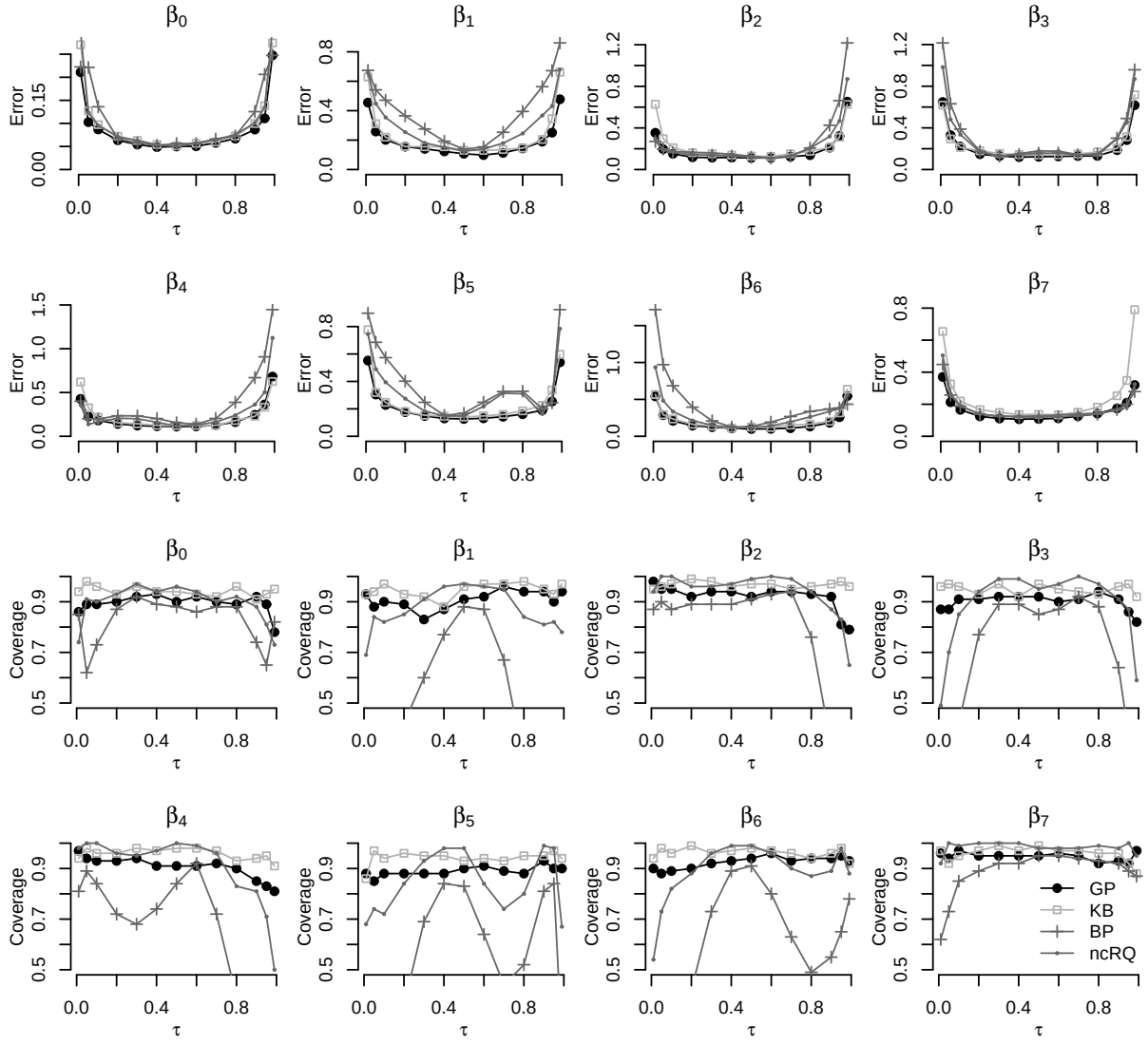

Figure 4: Assessing performance with a 7 dimensional X ($n = 1000$). Top two rows show estimation errors and bottom two rows show coverage by 95% confidence or credible bands for $\tau \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$.

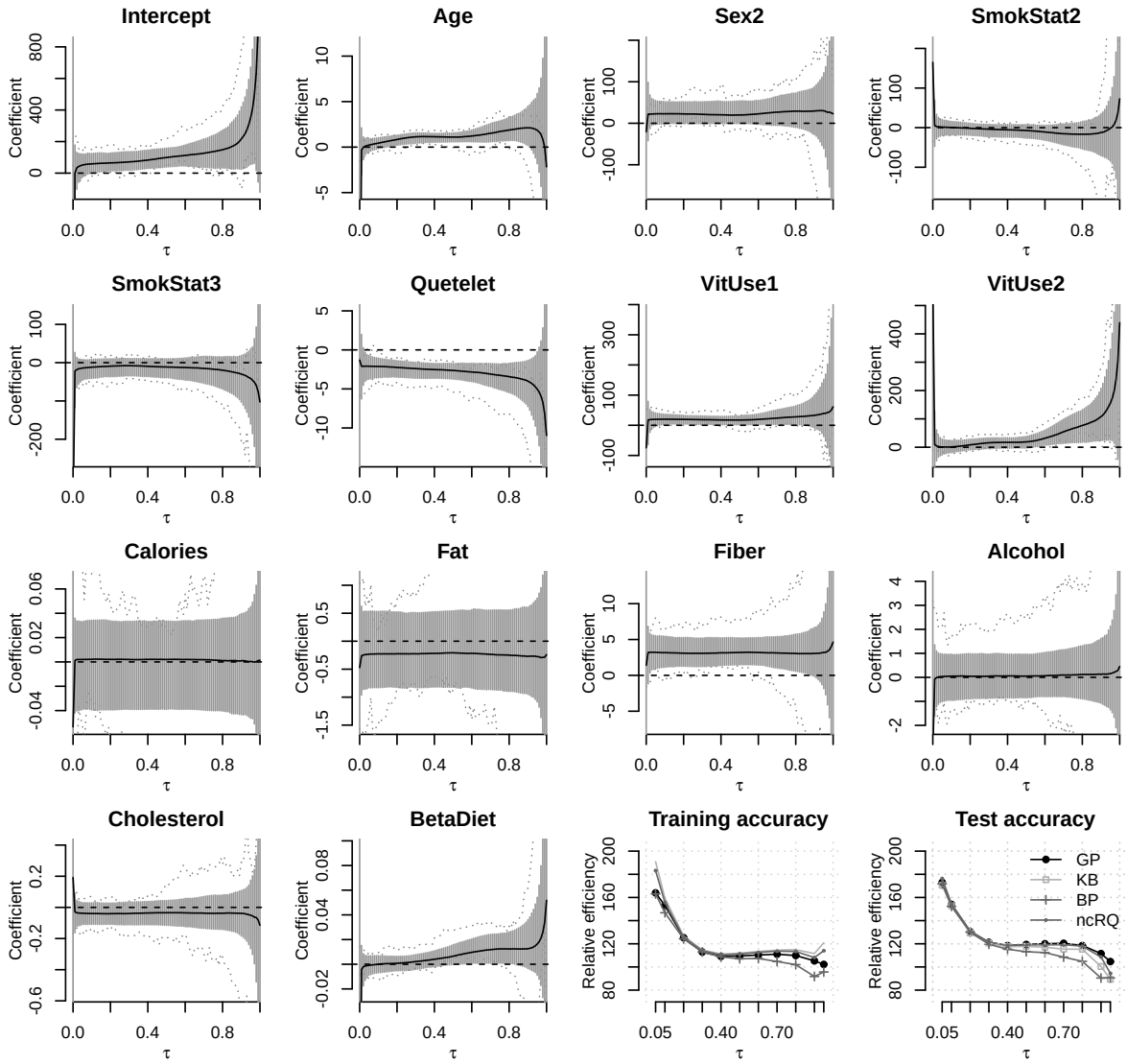

Figure 5: Parameter estimation and fit assessment for plasma data analysis. First 14 plots show posterior means and pointwise 95% credible bands for intercept and slope curves (gray), overlaid with the 95% bootstrap confidence bands obtained from single-τ Koenker-Bassett fits (dotted black lines). The last two plots show cross-validated assessment of quantile plane estimation.

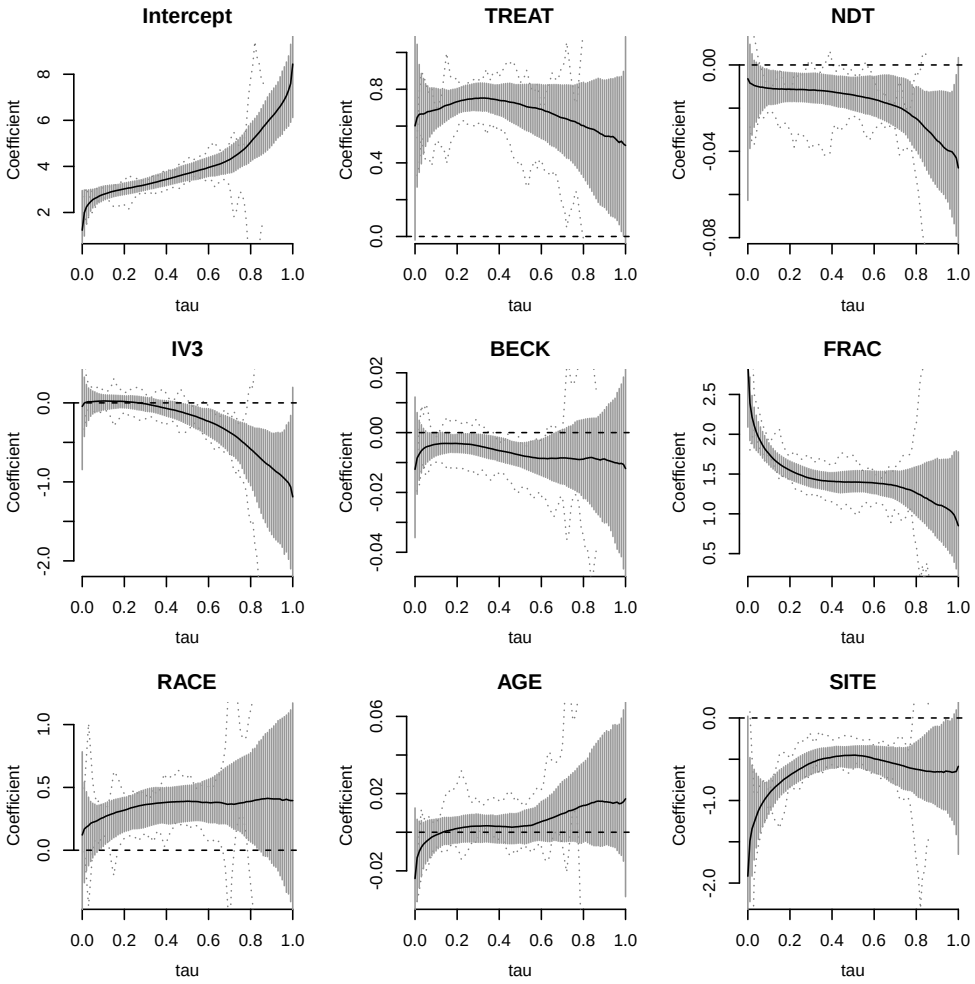

Figure 6: Parameter estimation for UIS data analysis. Solid black lines and gray bands give posterior means and pointwise 95% credible bands for intercept and slope curves, overlaid with the 95% bootstrap confidence bands (dotted lines) obtained from the Portnoy approach.

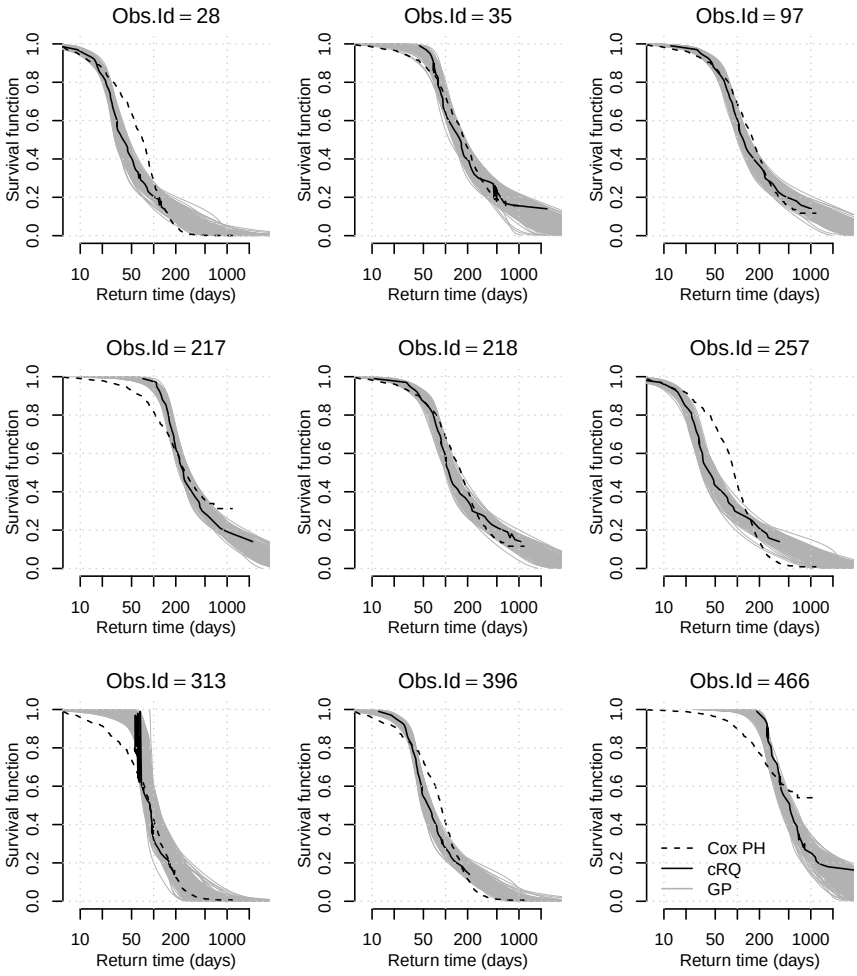

Figure 7: Estimated survival curves for 9 random sampled subjects in the UIS study. In each subject's plot, the gray lines are the posterior draws of the survival curves, the solid black line is the estimate from the Portnoy approach, and the dashed line is the estimate under the Cox proportional hazard model.

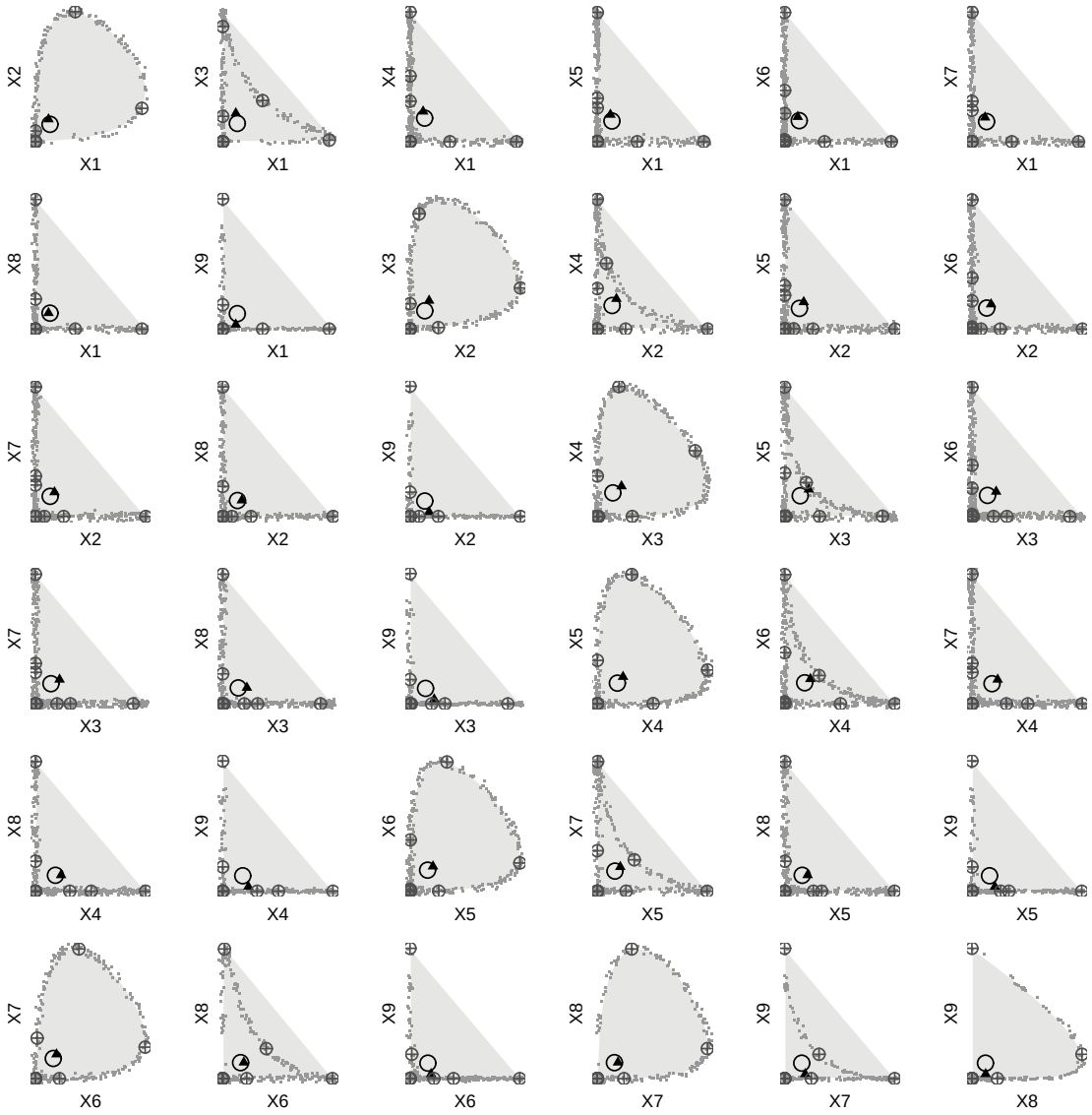

Figure 8: A toy demonstration of finding an interior point of the convex hull of observed predictors. The observed predictor vectors are 9 dimensional B-spline transforms of 500 uniform random drawn from $[0, 1]$, shown as dark gray dots (with some jitter to improve visibility) in the pairwise plots. The filled black triangle denotes the projection of the sample mean, and the large open circle denotes the projection of the interior point found by our preprocessing method. The crosshairs are the selected 10 points.