Bayesian nonparametric quantile process regression and estimation of marginal quantile effects

Steven G. Xu  |  Brian J. Reich

1 INTRODUCTION

Quantile regression (QR) models a conditional quantile as a function of covariates. It allows one to analyze the statistical relationship between covariates and noncentral parts of the conditional response distribution. However, when QR is fitted separately for inference on multiple levels, the natural ordering among different quantiles cannot be ensured, and the estimated quantiles are subject to cross. Quantile crossing can be alleviated by solving a constrained optimization problem (Bondell et al., 2010; Liu and Wu, 2011) when only a grid of quantiles is modeled, but estimates based on these methods can be sensitive to the number and location of the chosen quantile grids.

Simultaneous QR (SQR) allows inference on all quantiles by specifying the full quantile process. It encourages strength borrowing across proximate quantile levels through a unified modeling approach. SQR was first proposed by He (1997) who assumes a linear heteroscedastic regression model for the response. Subsequently, linear SQR models that impose fewer restrictions on the quantile function have been developed (e.g., Reich and Smith, 2013; Yuan et al., 2017; Yang and Tokdar, 2017). These approaches enjoy great interpretability by allowing rate-of-change interpretation of quantile-dependent coefficients but cannot accommodate quantile curves with complex nonlinear trends. Furthermore, they are not suitable for high-dimensional problems since they do not implicitly account for interaction effects.

Nonlinear SQR models are a minority in the current QR literature, and existing approaches suffer from apparent shortcomings. Cannon (2018) proposed to model the...
quantile process using a feed-forward neural network (FNN). They prespecify a set of quantile levels and treat them as a monotone covariate in the model to enforce monotonicity of the quantile function. Although their approach elegantly avoids quantile crossing, additional quantiles outside the prespecified range have to be estimated via extrapolation. Das and Ghosal (2018) model the quantile process as a weighted sum of B-spline basis functions of the quantile level; the weights are further expanded by tensor products of B-spline series expansion of each covariate, and order constraints are imposed on the spline coefficients to ensure noncrossing. Although their method models the full quantile process, it does not scale well to high dimensions since the number of parameters grows exponentially with the number of covariates. Nonlinear quantile process can also be estimated by inverting (or integrating then inverting) any valid estimate of the conditional distribution function. Das and Ghosal (2018) proposed a model on the conditional cumulative distribution function (CDF) of similar form to their aforementioned quantile process model. Consequently, their CDF model also suffers from computational intractability in high dimensions. Furthermore, their model is not constrained properly to estimate a bona fide density, which often leads to poor numerical performance. Izbicki and Lee (2016) projected the conditional probability density function (PDF) onto data-dependent eigenfunctions of a kernel-based operator. Their model scales well to high dimensions and estimates a bona fide density, but the resulting quantile surfaces are often not smooth. Recently, SQR models that leverage the advancement of deep learning have also been developed (eg, Kim et al., 2021), but the primary focus of these models is on prediction rather than inference.

In this article, we propose a novel treatment to nonlinear SQR by specifying a Bayesian nonparametric model on the conditional distribution. While there exist other conditional distribution regression models (eg, Holmes et al., 2012; Li et al., 2021), we model the conditional CDF using an I-spline basis expansion; the spline coefficients are modeled as functions of covariates using neural networks with specific output activation functions that ensure the model represent a bona fide CDF. We choose to model the distribution function instead of the quantile process because the former permits analytic derivation of the likelihood function and therefore efficient Markov chain Monte Carlo (MCMC) sampling of the posterior, and for a nonparametric regression the span of potential models is the same in both cases. A spline-based model ensures the estimated conditional CDF, and therefore the estimated conditional quantile function is smooth, and the neural networks allow incorporation of complex covariate effects on the response distribution. We name this method “QR Using I-spline Neural Network (QUINN).” QUINN provides several improvements over existing nonlinear SQR models (Izbicki and Lee, 2016; Cannon, 2018; Das and Ghosal, 2018). Instead of treating the quantile levels as a monotone covariate, QUINN specifies the full CDF such that all quantiles, rather than only a subset, can be modeled without extrapolation. By imposing proper constraints on the spline coefficient functions, we ensure that the estimate is a bona fide CDF, leading to more accurate estimation of the quantile process; directly constraining the model also avoids the need of a post hoc normalization method, which often renders the final quantile estimates unsmooth. Finally, by expanding the covariate-dependent coefficient functions using FNN rather than tensor products of splines, we greatly reduce the dimensionality of the parameter space so that it only scales linearly, instead of exponentially, in the number of covariates. A relevant but distinct work is that of Smith et al. (2015) who proposed a semiparametric framework based on I-spline basis expansion for a simultaneous estimation of linear quantile planes. In contrast, QUINN models the conditional distribution nonparametrically and allows simultaneous estimation of arbitrary quantile surfaces.

A disadvantage of modeling the conditional CDF nonparametrically is that covariate effects on different quantiles are not self-explanatory. This is common for black box supervised learning models that sacrifice transparency for flexibility. To overcome this challenge, model-agnostic methods (Ribeiro et al., 2016) have been developed to extract interpretation from any supervised learning model. A recent contribution is made by Apley and Zhu (2020), who proposed accumulated local effect (ALE) plot to visualize main and second-order interaction effects of a black box supervised learning model. Their method produces reliable characterization of the covariate effects on the predicted response in a computationally efficient way. In this article, we show that ALE plots can be applied to visualize covariate effects on predicted quantiles. We also present ways to estimate feature importance of marginal quantile effects of QUINN.

The motivating example is a study analyzing the effect of pregnancy-related and demographic factors on the distribution of birth outcomes. Low birth weight (LBW) is defined as weight less than 2.5 kg. It is a leading cause of prenatal and neonatal deaths, and births of underweight infants result in long-term medical and economic costs. High birth weight (HBW) is defined as weight greater than 4 kg. It is also an emerging public health issue worldwide. Overweight infants are subject to increased risk of health problems after birth, such as obesity in early childhood. QR is a natural approach to understand the determinants of LBW and HBW by modeling the lower and upper quantiles of the birth weight distribution. Examples are the
separate QR approach by Abrevaya (2001) and SQR approach by Tokdar & Kadane (2012). These works all assume a linear regression model, which will mischaracterize effects that are nonlinear (Ngwira and Stanley, 2015). In this article, we apply QUINN to the 2019 U.S. Natality Data Set (National Center for Health Statistics, 2019) to flexibly model different quantiles of the birth weight distribution, with a primary focus on identifying the influential factors of LBW and HBW.

2 METHODS

Denote $\mathbf{X} = (X_1, \ldots, X_d) \in \mathcal{X}$ as the covariate vector and $Y$ as the scalar response. We are interested in approximating the quantile process of the response given the covariates $Q_Y(\tau | \mathbf{X} = \mathbf{x})$ for quantile level $\tau \in (0, 1)$ and all $\mathbf{x} \in \mathcal{X}$. If $Q_Y(\tau | \mathbf{x})$ is continuous and monotonically increasing in $\tau$, then for any $\mathbf{x} \in \mathcal{X}$ the conditional CDF is $F_Y(y | \mathbf{x}) = Q_Y^{-1}(\tau | \mathbf{x})$. Thus, the monotonicity constraint of $Q_Y(\tau | \mathbf{x})$ can be naturally accounted for by specifying a valid model on $F_Y(y | \mathbf{x})$ and then inverting it. Our method requires the response variable to have a lower and upper bound, which we achieve by introducing a transformed variable $Z = g(Y)$ for some monotonic function $g$ that maps to the unit interval. In this section, we will outline our method for approximating quantile process of the transformed response $Q_Z(\tau, \mathbf{x})$ whose estimate can then be back-transformed to an estimate of $Q_Y(\tau | \mathbf{x})$.

2.1 Density regression using shape-constrained splines

We propose to model the conditional density of $Z$ given $\mathbf{x}$ using shape-constrained regression splines. Specifically, we model the conditional PDF using M-splines, and the conditional CDF using I-splines. The M-spline family with degree $r$ is a set of $r + p + 1$ piecewise polynomials of degree $r$ having properties of nonnegativity and unit integral. In other words, each basis function has the properties of a PDF. Let $\mathbf{T} = \{t_1, t_2, \ldots, t_{p+2r}\}$ be an ordered sequence of knots such that $t_1 = \cdots = t_r = 0$, $t_{p+r+1} = \cdots = t_{p+2r} = 1$, and $t_{r+k} - t_{r+k-1} = 1/p$, $k \in [1, p]$. Let $\{M_{m,r}(\cdot | \mathbf{T})\}_{m=1}^{r+p+1}$ be the set of basis functions. A convex combination of M-spline basis functions, that is,

$$\sum_{m=1}^{r+p+1} \vartheta_m M_{m,r}(\cdot | \mathbf{T}) \text{ s.t. } \vartheta_m \geq 0 \forall m \text{ and } \sum_{m=1}^{r+p+1} \vartheta_m = 1,$$

is a valid model for a PDF with support on $[0,1]$. The shape of the modeled PDF can be further controlled by placing additional constraints on the coefficients $\vartheta_m$. For example, setting $\vartheta_{r+p+1} = 0$ will force the M-spline to return to 0 at unity.

I-splines are defined as the integral of M-splines

$$I_{m,r}(x|\mathbf{T}) = \int_0^x M_{m,r}(u|\mathbf{T}) du, \quad m = 1, \ldots, r + p + 1$$

and are piecewise polynomials of degree $r + 1$. Since M-splines are nonnegative and integrate to 1, I-splines are monotonically nondecreasing with range $I_{m,r}(0|\mathbf{T}) = 0$ and $I_{m,r}(1|\mathbf{T}) = 1$ for all $m$. Thus, a convex combination of I-spline basis functions, that is,

$$\sum_{m=1}^{r+p+1} \vartheta_m I_{m,r}(\cdot|\mathbf{T}) \text{ s.t. } \vartheta_m \geq 0 \forall m \text{ and } \sum_{m=1}^{r+p+1} \vartheta_m = 1, \quad (1)$$

is a valid model for a CDF with support on the unit interval.

The use of shape-constrained regression splines offers an attractive solution to density estimation problems. Many theoretical works have shown the approximation power of nonnegative splines and monotone splines. For example, Beaton (1982) shows that as the number of knots increases, the space of nonnegative splines converges to the space of nonnegative continuous functions almost as quickly as unconstrained splines. Chui et al. (1980) show an analogous result for monotonic splines on approximating continuous monotonic functions. Through numerical studies, Abrahamowicz et al. (1992) show that the asymptotic theories are not affected by the addition of simplex constraint, and that M- and I-splines yield satisfactory accuracy in density regression.

2.2 QR using I-splines and neural network (QUINN)

Let $F_Z(z | \mathbf{x}, \mathcal{W})$ denote the conditional CDF of the transformed response variable $Z$ given $\mathbf{x}$. Following (1), a flexible model for $F_Z(z | \mathbf{x})$ can be expressed as

$$F_Z(z | \mathbf{x}, \mathcal{W}) = \sum_{m=1}^{r+p+1} \vartheta_m(\mathbf{x}, \mathcal{W}) I_{m,r}(z | \mathbf{T}) \text{ s.t. } \vartheta_m(\mathbf{x}, \mathcal{W}) \geq 0 \forall m \text{ and } \sum_{m=1}^{r+p+1} \vartheta_m(\mathbf{x}, \mathcal{W}) = 1,$$

where the covariates affect the conditional CDF through the spline coefficient functions $\vartheta_m(\mathbf{x}, \mathcal{W})$ parameterized by $\mathcal{W}$. The coefficient functions govern the covariate effect
on the conditional CDF and therefore should be flexible
eough to capture complex nonlinear trends and allow
for high-order interaction effects. They also need to be
properly constrained so that \( F_\beta(z|x, \mathcal{W}) \) has the
properties of a valid CDF. To satisfy these two requirements,
we model \( \theta_m(x, \mathcal{W}) \) using an FNN with softmax output
activation,

\[
\theta_m(x, \mathcal{W}) = \frac{\exp(u_m(x, \mathcal{W}))}{\sum_{i=1}^{r+p+1} \exp(u_i(x, \mathcal{W}))}
\]

\[
u_m(x, \mathcal{W}) = W_{2m0} + \sum_{l=1}^{V} W_{2ml} \phi \left( W_{1l0} + \sum_{j=1}^{d} W_{1lj} x_j \right)
\]

where \( \mathcal{W} = \{W_{uvw}\} \) are the unknown weights and \( \phi \) is
the known activation function. Throughout this article,
\( \phi \) is taken to be the hyperbolic tangent function. Profi-
ing from its universal approximation theorem (Hornik et al., 1989), FNN allows the unconstrained coefficient
functions \( u_m(x, \mathcal{W}) \) to describe arbitrary complex covari-
ate effects. While the softmax activation naturally projects
\( u_m(x, \mathcal{W}) \) to the unit simplex and overcomes the challenge
of parameter estimation under monotonicity constraints.
For simplicity, we describe the FNN with a single hidden
layer with \( V \) neurons, but extensions to deeper networks
are straightforward.

The proposed model can approximate any continuous
conditional CDF. Following the results of Chui et al.
(1980) and Abrahamowicz et al. (1992), with a large
enough \( p \), we can assume for any \( x \) there exists a set of
nonnegative coefficients \( \{\alpha_1, \alpha_2, \ldots, \alpha_{r+p+1}\} \) satisfying the
constraint \( \sum_{m=1}^{r+p+1} \alpha_m = 1 \) such that \( \sum_{m=1}^{r+p+1} \alpha_m I_{m,r}(z|T) \)
approximates the conditional CDF \( F(z|x) \) arbitrarily well.
In QUINN, the mapping \( \psi : x \rightarrow \{\alpha_1, \alpha_2, \ldots, \alpha_{r+p+1}\} \) is
modeled by a single-hidden-layer FNN with softmax out-
put which is a non-constant, bounded, and continuous
function. Then by the universal approximation theorem
(Hornik et al., 1989), there exist weights \( \mathcal{W} \) such that
the single-hidden-layer FNN \( \theta_m(x, \mathcal{W}) \) approximates the
mapping \( \psi \) arbitrarily well for all \( x \), provided that the
number of hidden neurons \( V \) is large enough. Thus, by
leveraging the approximation power of I-splines and FNN,
the model \( \sum_{m=1}^{r+p+1} \theta_m(x, \mathcal{W}) I_{m,r}(z|T) \) can approximate any
conditional CDF \( F(z|x) \).

We adopt a Bayesian framework to estimate the weights
\( \mathcal{W} \) by assigning them prior distributions. Compared to
its frequentist counterpart, Bayesian neural network mod-
celing captures uncertainty in both the fitted model and
weight parameters, and avoids overfitting when the sam-
ple size is small. Zero-mean Gaussian distributions are
the most commonly used prior on weights and have been
explored in many classic works (MacKay, 1992; Neal, 1993).

Their popularity arise from their “weight-decay” regu-
larization effect, which prevents individual nodes from
having extreme value. For QUINN, we set \( W_{1vw} \sim
\mathcal{N}(0, \sigma^2_w), W_{2vw} \sim \mathcal{N}(0, \gamma^2) \) so that weights in input-
hidden layer have feature-wise variances, and weights in
hidden-output layer share a common variance. The scale
hyperparameters \( \sigma_w \) and \( \gamma \) are also treated as unknown
and assigned hyperpriors, so that their values can be opti-
imized by the data. Gelman (2006) recommend half-\( t \) fam-
ilies with a small degrees of freedom. These distributions
allow the variance to be arbitrarily close to 0, which reg-
ularizes the complexity of the model. In practice however,
the heavy tailedness of half-\( t \) distributions make them too
broad and often cause difficulty in convergence. Therefore,
we set \( \sigma_w, \gamma \sim \mathcal{N}^+(0, a^2) \) to follow the half-Gaussian dis-
tribution for simpler posterior geometry. The variance of
half-Gaussian prior is set to be \( a^2 = 900 \) so it is still rela-
tively noninformative. Experiments show that our model
is not sensitive to moderately large values of \( a \). The like-
lihood function of QUINN has a closed-form expression,
therefore MCMC algorithms can be used to explore the
posterior. However, traditional methods such as random-
walk Metropolis and Gibbs sampler do not scale well to
high-dimensional posterior with complex geometry. In this
article, we use No-U-Turn sampler (NUTS) (Hoffman and
Gelman, 2014) that uses gradient information to sam-
ple efficiently from high-dimensional posterior. Webb
Appendix A describes the MCMC algorithm used to
approximate the posterior.

Our ultimate goal is to estimate the quantile process of
the original response variable \( Q_Z(\tau|x) \). Let \( \hat{F}_Z(z|x) \) denote
the conditional CDF estimator and \( D_Z = \{z_1, z_2, \ldots, z_N\} \)
denote a dense grid on the unit interval. Nonparametric
estimate of the quantile process \( Q_Z(\tau|x) \) can be easily
obtained by first evaluating \( \hat{F}_Z(z|x) \) on \( D_Z \) and then per-
forming linear interpolation on a dense percentile grid by
treating \( \{\hat{F}_Z(z|x)\}_{i=1}^{N} \) as the input values and \( D_Z \) as the
functional output values. Because of the one-to-one corre-
spondence between quantile function and CDF, the result-
ing quantile process estimator will also inherit the approxi-
mator property of the proposed CDF estimator. Finally,
the estimated quantile process of the original response is given by
\( \hat{Q}_Y(\tau|x) = g^{-1}[\hat{Q}_Z(\tau|x)] \).

As discussed in Section 1, the proposed model has
several advantages over existing nonlinear SQR models
(Izbicki and Lee, 2016; Cannon, 2018; Das and Ghosal,
2018). The combination of I-splines and FNN leads to a
valid probability model that spans a wide class of condi-
tional distribution functions. Also, as described below and
shown later by simulation studies, this combination leads
to efficient computation and fully-Bayesian inference on
quantile effects.
3 | SUMMARIZING COVARIATE EFFECTS

QUINN includes a flexible FNN model for covariate effects across quantile levels. FNN is a “black box” supervised learning model that excel in flexibility but lack transparency. Unlike linear QR models, which enable rate-of-change interpretation of the 𝜋-dependent coefficients, the proposed FNN-based model does not characterize the covariate effects on the predicted quantile in a self-explanatory way. This is inconvenient, since QR models are often used for data exploratory purposes.

Fortunately, research on model agnostic methods has allowed post hoc analysis of main effects and second-order interaction effects (we omit consideration of higher order effects as they cannot be visualized or interpreted meaningfully) of the covariates on the predictions made by “black box” models. The most popular model agnostic method is partial dependence plot (PD plot), which visualizes the average marginal effect a (pair of) covariate(s) have on the predictions. The PD plot is straightforward to implement and intuitive to interpret, but is expensive to compute. Recently, Apley and Zhu (2020) proposed the accumulative local effects (ALEs) plot that provides the same level of interpretation in a more computationally efficient way. In this section, we provide a brief review of the definitions of ALE plot and explain how it can be applied to QR models. We will later demonstrate using a multivariate simulation study how ALE plot can be utilized to extract interpretation from QUINN.

The sensitivity of 𝑄_𝑻(𝜏|𝐱) to covariate 𝑗 is naturally quantified by the derivative 𝑞_𝑻(𝜏,𝐱) = 𝜕𝑄_𝑻(𝜏|𝐱)/𝜕𝐱_𝑗. In a linear QR, the derivative is the scalar effect of covariate 𝑗 on quantile level 𝜏, but for a nonlinear regression function the derivative depends on 𝐱. The ALE begins by averaging 𝑞_𝑻(𝜏,𝐱) over 𝐱 conditioned on 𝑋_𝑗 = 𝒙_𝒋, giving {̄𝑞_𝑻(𝜏,𝐱) = 𝐸_𝑋[𝑞_𝑻(𝜏,𝐱)|𝑋_𝑗 = 𝒙_𝒋]. The uncentered ALE main effect function of 𝑋_𝑗 is then defined as

\[ \tilde{𝑄}_𝑗^\gamma(𝜏,𝑥_𝑗) = \int_{𝑥_{min,𝑗}}^{𝑥_𝑗} \tilde{q}_𝑗(𝜏,𝑢_𝑗)d𝑢_𝑗, \]

The function \( \tilde{Q}_j^\gamma(𝜏,𝑥_j) \) can be interpreted as the ALE of 𝑋_𝑗 in the sense that it is an accumulation of local effects \( \tilde{q}_j(𝜏,𝑢_j) \) averaged over the distribution of 𝐱. The uncentered ALE effect does not have a straightforward interpretation because the derivative is invariant to scalar addition, which leads to the definition of the (centered) ALE main effect function \( \hat{Q}_j^\gamma(𝜏,𝑥_j) \) that is the same as \( \tilde{Q}_j^\gamma(𝜏,𝑥_j) \) except centered to have mean 0 with respect to the marginal distribution of 𝑋_𝑗.

Analogous formulas define the second-order ALE for 𝑋_𝑗 and 𝑋_𝑙. Consider the second-order partial derivative \( q_{jl}(𝜏,𝐱) = 𝜕^2𝑄_𝑗^\gamma(𝜏|𝐱)/𝜕𝐱_𝑗 𝜕𝐱_𝑙 \). The local effect at 𝑋_𝑗 = 𝑥_𝑗 and 𝑋_𝑙 = 𝑥_𝑙, averaging over the other covariates, is \( \hat{q}_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) = \mathbb{E}_𝐗[𝐪_{jl}(𝜏,𝐗)|𝐗_𝑗 = 𝒙_𝒋, 𝊶态 = 𝒙_𝒍] \). The uncentered second-order ALE is then

\[ \hat{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) = \int_{𝑥_{min,𝑗}}^{𝑥_𝑗} \int_{𝑥_{min,𝑙}}^{𝑥_𝑙} \hat{q}_{jl}(𝜏,𝑢_𝑗,𝑢_𝑙)d𝑢_𝑗d𝑢_𝑙, \]

and the second-order ALE function \( \tilde{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) \) is mean-centered with respect to the marginal distribution of \( (𝐗_𝑗,𝐗_𝑙) \). The second-order ALE \( \tilde{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) \) describes the joint effects of the two covariates, which consist of both their main effects and interaction effect. In cases where assessment of only the interaction effect is of interest, main effects of 𝑋_𝑗 and 𝑋_𝑙 can be further subtracted from \( \tilde{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) \) to obtain the pure interaction effect \( \hat{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) \).

The functions \( \hat{Q}_j^\gamma(𝜏,𝑥_𝑗), \tilde{Q}_j^\gamma(𝜏,𝑥_𝑗,𝑥_𝑙), \) and \( \hat{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) \) can be plotted to understand each main and interaction effect. When plotted against 𝑥_𝑗, the main ALE \( \hat{Q}_j^\gamma(𝜏,𝑥_𝑗) \) quantifies the difference between average prediction conditioned on \( 𝑋_𝑗 = 𝒙_𝒋 \) and the average prediction over \( 𝐸_𝐗[\hat{Q}_j^\gamma(𝜏,𝐗)|𝐗_𝑗 = 𝒙_𝒋] \). When plotted against \( 𝑥_𝑗 \) and \( 𝑥_𝑙 \), the second-order ALE \( \hat{Q}^\gamma_{jl}(𝜏,𝑥_𝑗,𝑥_𝑙) \) quantifies the difference between average prediction conditioned on \( (𝐗_𝑗,𝐗_𝑙) = (𝑥_𝑗,𝑥_𝑙) \) and the average prediction over \( 𝐸_𝐗[\hat{Q}^\gamma_{jl}(𝜏,𝐗,𝐗)] \). For categorical covariates, the standard deviation is replaced by one fourth of the range. These VI scores (and the intermediate functions \( \hat{Q}_j^\gamma, \tilde{Q}_j^\gamma, \) and \( \hat{Q}^\gamma_{jl} \)) can be approximated using the partitioning schemes of Apley and Zhu (2020) as described in Web Appendix B.

Although for notational simplicity we have omitted the dependence of the quantile function on the parameters \( 𝑊 \), in practice the posterior uncertainty in \( 𝑊 \) leads to posterior uncertainty in the sensitivity metrics such as \( \hat{Q}_j^\gamma(𝜏) \) and \( \tilde{V}_j^\gamma(𝜏) \). We account for this uncertainty by computing the sensitivity measures for many MCMC samples from the posterior distribution of \( 𝑊 \), giving a Monte Carlo approximation of the posterior distribution of the sensitivity measures.
4 | SIMULATION

We investigate the numerical performance of our model in four scenarios. The details of each simulation design are provided below.

**Design 1.** The covariate and response are generated as $X \sim \text{Uniform}(0, 5)$ and

$$Y = X + \sin(2X) + 3\varepsilon; \varepsilon \sim \text{Skew-Normal}(0, 1, 4).$$

The quantile curves are parallel, and the data exhibit strong right-skewness.

**Design 2.** The covariate and response are generated as $X \sim \text{Uniform}(0, 1)$ and

$$Y = 3X + [0.5 + 2X + \sin(3\pi X + 1)]\varepsilon; \varepsilon \sim \text{Normal}(0, 1).$$

The data exhibit strong heteroscedasticity. The quantile curves are linear at the median but have strong curvature at the extremes.

**Design 3.** The covariates $X_j, j = 1, 2$ are generated from Uniform([0, 1] $\times$ [0, 1]). The response variable $Y$ is given by

$$Y = \sin(2\pi X_1) + \cos(2\pi X_2) + \sqrt{2(X_1^2 + X_2^2)}\varepsilon; \varepsilon \sim \text{Student's } t(3).$$

The data exhibit both heteroscedasticity and heavy-tailedness.

**Design 4.** The covariates $X_j, j = 1, 2, .., d$ are generated from Uniform([0, 1]$^d$). The quantile function $Q_Y(\tau | \mathbf{X})$ is given by

$$Q_Y(\tau | \mathbf{X}) = 3(\tau - 0.5) \left( X_1 + \frac{3}{5} \right)^3$$

$$+ 15 \left[ X_2 + 4 \left( X_2 - \frac{1}{2} \right)^2 \right]$$

$$\times \exp \left( -X_2^2 \right)$$

$$+ 12 \exp \left[ \left( X_3 + \frac{1}{2} \right)^2 \left( X_4 - \frac{1}{2} \right)^2 \right]$$

$$+ 5(\tau - 1) \left( X_5 + \frac{2}{5} \right)^2 \left( X_6 + \frac{1}{2} \right)^2$$

$$+ 0.25\Phi^{-1}(\tau),$$

where $\Phi^{-1}(\cdot)$ is the standard normal quantile function and $d \in \{10, 20, 40\}$. The response variable is generated by sampling $U \sim \text{Uniform}(0, 1)$ and setting $Y = Q(U | \mathbf{X})$. The quantile process has a complex structure with strong interaction effects. The model is sparse as only the first six covariates affect the quantile function.

For Designs 1–3, we generate samples of sizes $n \in \{50, 100, 200\}$ and for Design 4 we use $n = 200$. The proposed model is compared to four nonlinear SQR methods: the monotone composite QR neural network (MCQRNN) of Cannon (2018), the nonparametric simultaneous QR (NPSQR) of Das and Ghosal (2018), the nonparametric distribution function simultaneous QR (NPDFSQR) also of Das and Ghosal (2018), and the spectral series conditional density estimator (seriesCDE) of Izbicki and Lee (2016). MCQRNN is implemented in the qrnn package in R; codes for NPSQR and NPDFSQR are available from the second author’s webpage; and codes for seriesCDE are available from the supplemental material of their online paper. Implementation details including model selection for the competing methods are given in Web Appendix C. For QUINN, we first map the response variable to the unit interval using min–max normalization. The covariates are not required to be normalized. However, it is a common practice to normalize the inputs to an FNN when optimizing its parameters using a gradient-based approach (Bishop, 1995). In this article, we always map the covariate vector to the unit interval, even if it is one-dimensional. Posterior distribution of QUINN is approximated by 1900 MCMC samples that are obtained by running NUTS for 20,000 iterations, discarding the first 1000 iterations as burn-in and saving every 10th draw from the remaining iterations. Convergence of MCMC is monitored by trace plots of log-likelihood from multiple independent chains as shown in Web Figure 1, and popular diagnostic statistics as described in Web Appendix A.6. The performance of QUINN depends on the number of spline knots $p$ and hidden neurons $V$, so we use a grid search approach and select the best combination of $p, V \in \{5, 8, 10\}$ based on Watanabe-Akaike Information Criterion (WAIC). (Watanabe, 2013). We choose WAIC over other information criteria (eg, Akaike information criterion [AIC] and Deviance information criterion [DIC]) because it is fully Bayesian, uses the entire posterior distribution, and is asymptotically equal to Bayesian leave-one-out cross-validation (Vehtari et al., 2017). Web Figure 2 plots the distribution of out-of-sample RMISE against ranking of WAIC. The result shows that model chosen by WAIC is in favor of a higher out-of-sample prediction accuracy. We also observe that the
FIGURE 1  RMISE ($\tau$) for the simulation studies at quantile levels $\tau \in \{0.05, 0.1, \ldots, 0.95\}$. The training sample sizes are $n = 100$ for Designs 1–3 and $n = 200$ for Design 4

The performance of QUINN is generally robust to different values of the two parameters except for some particularly bad combinations.

To compare the different approaches, 100 data sets are simulated. For each sample size, the performance of each method is measured by the root mean integrated square error (RMISE) between the actual and estimated (posterior mean) quantile processes. We first divide the domain of each dimension of $\mathbf{X}$ by $g$ equidistant gridpoints, giving $G = gd$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_G$ that span the range of $\mathbf{X}$. For Designs 1 and 2, we set $g = G = 101$; for Design 3, we set $g = 21$ and thus $G = 21^2$. The RMISE is then approximated as

$$\text{RMISE}(\tau_k) = \sqrt{\frac{1}{G} \sum_{i=1}^{G} \left( Q_Y(\tau_k | \mathbf{x}_i) - \hat{Q}(\tau_k, \mathbf{x}_i) \right)^2}$$

for quantile level $\tau_k \in \{0.05, 0.10, \ldots, 0.95\}$ and

$$\text{RMISE}_{QP} = \sqrt{\frac{1}{19} \sum_{k=1}^{19} \text{RMISE}(\tau_k)^2}$$

for the entire quantile process.
The average RMISE_{QP} over 100 simulated data sets along with their standard errors are shown in Web Table 1 in Web Appendix D. The results show that QUINN yields significantly smaller average RMISE_{QP} in all settings when compared to NPSQR, NPDFSQR, and seriesCDE; and smaller or similar average RMISE_{QP} in all but one setting when compared to MCQRNN. In particular, QUINN is robust to data sparsity as seen from its small variance. We also plot the average RMISE(\tau) for cases when n = 100 in Figures 1a–c. The results show that QUINN gives the best estimation of intermediate quantiles in all cases, whereas MCQRNN gives better estimation of extreme quantiles when the data exhibit significant heavy-tailedness.

For Design 4, we compare QUINN with MCQRNN and seriesCDE only. We omit NPSQR and NPDFSQR because expanding each covariate of a d-dimensional \( \mathbf{X} \) using quadratic B-spline basis functions results in a parameter space of dimension \( p_{DG}(p_{DG} + 2)^d \). Even with \( d \) as few as 10, fitting NPSQR and NPDFSQR becomes computationally infeasible. We generate sample of size \( n = 400 \) and split into 200 training and 200 testing data points. Each model is first fit to the training data, then conditional quantile predictions at \( \tau \in \{0.05, 0.1, \ldots, 0.95\} \) are calculated for each given \( \mathbf{x} \) of the testing data. Posterior distribution of QUINN is approximated by 2400 samples that are obtained by running NUTS for 50,000 iterations, discarding the first 2000 iterations as burn-in and saving every 20th draw from the remaining iterations. We choose the best model configuration of \( p \in \{5, 8, 10\} \) and \( V \in \{8, 10, 15\} \) using WAIC. We generate 100 replicates and compare different approaches based on RMISE(\tau) and RMISE_{QP} between actual and predicted quantiles.
conditioned on the testing data points. The average RMISE\(_{QP}\) for \(d = 10\) are 2.47 for QUINN, 3.21 for MCQRNN, and 3.37 for seriesCDE and the average RMISE(\(\tau\)) are plotted in Figure 1d. The result shows that QUINN gives substantially better estimation of the quantile process than MCQRNN and seriesCDE. To further investigate the performance of QUINN in high-dimension setting, we repeat Design 4 with \(X\) of dimensions \(d = 20, 40\) with the additional covariates being independent of the response. The average RMISE\(_{QP}\) are 2.73 and 3.09, respectively. Therefore, QUINN shows promising performance when the quantile process is high-dimensional, has complex interaction effects, and has a sparse structure.

We now demonstrate how ALEs plot can be used with QUINN to visualize main and second-order interaction effects on its predicted quantiles. Design 4 is constructed such that \(\bar{Q}_1(\tau, x_1), \ldots, \bar{Q}_6(\tau, x_6), \bar{Q}_{3,4}(\tau, x_3, x_4),\) and \(\bar{Q}_{5,6}(\tau, x_5, x_6)\) are nonzero functions of \(\tau\). In addition, \(\bar{Q}_j(\tau, x_j), j \in \{1, 5, 6\},\) and \(\bar{Q}_{5,6}(\tau, x_5, x_6)\) are nonconstant functions of \(\tau\). To evaluate the sensitivity of QUINN in identifying these marginal quantile effects, we generate 100 replicates of sample size 5000 from Design 4. For each replicate, we estimate the ALE main effect for each covariate and interaction effect for each pair of covariates at quantile levels \(\tau \in \{0.05, 0.10, \ldots, 0.95\}\) based on the fitted QUINN. The estimated ALE main effects at these quantile levels along with their ground truths are shown in Figure 2, where the thin black lines represent individual estimates based on the 100 simulated data sets, and the thick gray line represents the true ALE effect calculated from the generating model. The estimated ALEs show that QUINN successfully captures the main effects of each covariate. For ALE interaction effects, since it is impossible to visualize all estimated surfaces in one plot, we instead show estimates from eight randomly selected replicates. For each replicate, estimated \(\bar{Q}_{jl}(\tau, x_j, x_l)\) and \(\bar{Q}_{jl}(\tau, x_j, x_l)\) at quantile levels \(\tau \in \{0.05, 0.5, 0.95\}\) are visualized using contour plots (see Web Appendix D) and compared with that of the ground truth. The results show that QUINN also successfully recovers the complex interaction effects of the generating model.

To investigate whether QUINN is capable of recovering the relative importance of the marginal effects, we calculate VI scores for each ALE main and interaction effect. Because only eight marginal effects contribute to the conditional quantile function in Design 4, we demonstrate the sensitivity of QUINN by showing the rank plot of the top eight estimated marginal effects with the highest VI in Figure 3. The results show that at quantile levels \(\tau \in \{0.05, 0.5, 0.95\}\), the estimated VIs and their ranking of the top eight marginal effects resemble the ground truth. The sensitivity analysis shows that QUINN is able to identify the relative order of the important covariate effects.

The simulation results presented in this article are only selected ones that demonstrate the predictive accuracy and sensitivity of QUINN. Additional results, including plots of estimated quantile curves and conditional densities, as well as comparison between estimated and actual ALE second-order effects, are available in Web Appendix D of the online Supporting Information.
5 APPLICATION TO BIRTH WEIGHT DATA

To illustrate the practical effectiveness of QUINN, we study the effect of pregnancy-related factors on infant birth weight (weight, in grams) quantiles. Our data (Xu, 2021) consist of 10,000 randomly chosen entries from the 2019 U.S. Natality Data Set (National Center for Health Statistics, 2019) on singleton live births to mothers recorded as Black or White, in the age group 18–45, with height between 59 and 73 inches, and smoke no more than 20 cigarettes daily during pregnancy. The list of covariates contains demographic characteristics, maternal behavior and health characteristics, as well as infant health characteristics. For demographic characteristics, we include indicator of age above 40 years old (fatherAge) for the father; and age (motherAge, in years), indicators of Black (Black), education attainment up to high school graduate (highSchool) and at least college graduate (collegeGraduate), and parity greater than 1 (Parity) for the mother. For maternal behavior and health characteristics, we include body mass index (BMI), height (Height, in inches), weight gain (wtGain, in pounds), indicator of smoking before pregnancy (Smoker), average daily number of cigarettes during pregnancy (Cigarettes), indicators of not receiving prenatal care (noPrenatal), preexisting diabetes (preDiab) and hypertension (preHype), gestational diabetes (gestDiab) and hypertension (gestHype), no infections present and/or treated during pregnancy (noInfec), and infertile treatment (infTreat).

For infant health characteristics, we include gestational age (week) and indicator of boy (Boy).

The response variable and all continuous covariates are mapped to the unit interval using min–max normalization. We fit QUINN with \( V \in \{10, 20, 30\} \) hidden neurons and \( p \in \{10, 15, 20\} \) spline knots. We approximate its posterior distribution using 2400 samples obtained by running NUTS for 50,000 iterations, discarding the first 2000 iterations as burn-in, and selecting every 20th draw from the remaining iterations. The best model configuration is chosen based on WAIC.

To determine which covariates have the most significant impact on the birth weight quantiles, we calculate the ALE-induced VI score for each covariate across different quantile levels. In particular, we are interested in identifying the covariates that most impact LBW (represented by the 0.05 quantile), typical birth weight (TBW, represented by the 0.5 quantile), and HBW (represented by the 0.95 quantile). Figure 4 shows the ranking of ALE main effects at \( \tau \in \{0.05,0.50,0.95\} \). The main effects of week, height, BMI, wtGain, Cigarette, Black, preDiab, Boy, and Smoker have the highest VI measure at all three quantiles and therefore are most influential on the birth weight distribution. In particular, week has a dominant effect on all three quantiles, Cigarette is most influential on LBW, and Height and BMI are more influential on TBW and HBW.

To understand the functional relationship between the top covariates and the predicted birth weight quantiles of QUINN, we plot their estimated ALE main effects \( \hat{Q}_i(\tau) \) at \( \tau \in \{0.05,0.50,0.95\} \) in Figures 5a–i. The results show that higher values of week, height, BMI, wtGain, preDiab, and Boy are associated with higher predicted birth weight, whereas higher values of Cigarette, Black, and Smoker are associated with lower predicted birth weight. Furthermore, the effects of week, height, BMI, wtGain, Cigarette, and preDiab on birth weight are significantly nonconstant across quantiles, and the effects of week, BMI, wtGain, Cigarette are highly nonlinear. For example, as Cigarette increases, LBW and TBW display a consistent downward trend, whereas HBW plateaus when Cigarette is greater than 13; as wtGain increases, HBW and TBW display a consistent upward trend, whereas LBW plateaus when wtGain is greater than 65.

The results in Figures 5a–i also provide numerical quantification of the main effects on the predicted birth weight quantiles. For example, Figure 5b shows that compared to mothers who do not smoke during pregnancy, mothers who smoke as many as 20 cigarettes daily are associated with a more than 250-g decrease in predicted LBW, on average. Figure 5g shows that compared to mothers who do not have preexisting diabetes, mothers who have preexisting diabetes are associated with a 254-g increase in predicted HBW, on average.

In addition to covariate effects on specific quantiles, QUINN also allows direct characterization of covariate effects on the whole conditional distribution thanks to its density regression nature. To illustrate this property of QUINN, Figure 5j plots the predicted birth weight density for week \( \in \{33, 34, \ldots, 42\} \). The result shows that gestational age has a prominent effect on the location of the predicted density. The shifting of the density is most significant when gestational age increases from 33 to 37 and gradually plateaus when the pregnancy term further increases. For each predicted density, we also highlight the region of LBW (weight <2500) and HBW (weight >4000). The result indicates that preterm (week <37) and postterm (week >40) are determinant factors of LBW and HBW, respectively. Figure 5k plots the predicted birth weight CDF for different levels of preDiab. Compared to mothers who do not have preexisting diabetes, mothers with preexisting diabetes are associated with a 10% increase in probability of giving birth to an overweight infant.
We further analyze the second-order interaction effects between the top covariates that have significant main effects. For each combination, we estimate its ALE joint effect $\bar{Q}_{j,l}(\tau)$ as well as its pure interaction effect $\bar{Q}_{j,l}^\prime(\tau)$ at $\tau \in \{0.05, 0.5, 0.95\}$. The significance of each interactions is determined by the estimated VI measure $\hat{V}I_{j,l}(\tau)$. Among all interaction considered, the interaction between gestational age and average daily number of cigarettes during pregnancy ($\text{Week} \times \text{Cigarette}$) yields the highest VI measure on all three quantiles. To understand the functional relationship between $\text{Week} \times \text{Cigarette}$ and the birth weight quantiles, Figure 6 plots the contour of the joint effects of $\text{Week}$ and $\text{Cigarette}$, as quantified by $\bar{Q}_{j,l}(\tau)$. The result indicates that higher gestational age is associated with higher birth weight regardless of maternal smoking habit, but the effect is clearly amplified for nonsmokers.
Figure 5 Marginal main effect estimates of the birth weight data. (a)–(i) Posterior mean ALE main effects at quantile levels \( \tau \in \{0.05, 0.50, 0.95\} \) for the top nine important covariates. For continuous covariates, black dashed line represents the value 0. (j)–(i) Conditional distribution estimates by gestational age \((\text{Week})\) and prepregnancy diabetes indicator \((\text{preDiab})\), respectively, with all other covariates fixed at their median (continuous covariates) or mode (binary covariates).

In addition, heavier maternal smoking is associated with lower birth weight only for births that occur after the 37th week of pregnancy.

6 Conclusion

In this article, we propose a novel nonlinear SQR model that leverages the flexibility of spline and neural network. We adopt a Bayesian framework by assigning prior distributions to the weight parameters and utilize the state-of-art NUTS to sample efficiently from the high-dimensional posterior. Compared to existing works, our method models the full quantile process, does not involve constrained optimization, and scales to high-dimensional setting. We also show that our model can yield meaningful interpretation via ALE plots and VI scores. Simulation studies show that our model better recover high-dimensional quantile process with complex structure and is robust to data and model sparsity. Sensitivity analysis shows that our
The proposed method was used to analyze the relationship between birth weight and pregnancy-related factors of U.S. newborns and specifically to identify influential effects on LBW and HBW. Our results showed that LBW is primarily associated with prematurity and heavy maternal smoking; whereas HBW is primarily associated with high maternal body mass index, maternal height, maternal weight gain, and having preexisting diabetes. Future extension can focus on accommodating spatial and/or temporal correlation between observations and variable/model selection using sparsity-inducing priors.

DATA AVAILABILITY STATEMENT
The data that support the findings of this article are openly available in Open Science Framework at https://doi.org/10.17605/OSF.IO/3GFHE. These data were derived from the following resources available in the public domain: 2019 Natality, https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/DVS/natality/.

OPEN RESEARCH BADGES
This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://doi.org/10.17605/OSF.IO/UC9QE.

REFERENCES
Abrahamowicz, M., Clampl, A. and Ramsay, J. O. (1992) Nonparametric density estimation for censored survival data: regression-spline approach. Canadian Journal of Statistics, 20, 171–185.
Abrevaya, J. (2001) The effects of demographics and maternal behavior on the distribution of birth outcomes. Empirical Economics, 26, 247–257.
Apley, D. W. and Zhu, J. (2020) Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B, Statistical Methodology, 82, 1059–1086.
Beatson, R. (1982) Restricted range approximation by splines and variational inequalities. SIAM Journal on Numerical Analysis, 19, 372–380.
Bishop, C. M. (1995) Neural Networks for Pattern Recognition. Oxford, UK: Oxford University Press.
Bondell, H. D., Reich, B. J. and Wang, H. (2010) Noncrossing quantile regression curve estimation. Biometrika, 97, 825–838.
Cannon, A. J. (2018) Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stochastic Environmental Research and Risk Assessment, 32, 3207–3225.
Chui, C., Smith, P. and Ward, J. (1980) Degree of $L_p$ Approximation by Monotone Splines. SIAM Journal on Mathematical Analysis, 11, 436–447.
Das, P. and Ghosal, S. (2018) Bayesian non-parametric simultaneous quantile regression for complete and grid data. Computational Statistics and Data Analysis, 127, 172–186.
Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, 515–534.
Greenwell, B. M., Boehmke, B. C. and McCarthy, A. J. (2018) A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755.
He, X. (1997) Quantile curves without crossing. American Statistician, 51, 186–192.
Hoffman, M. D. and Gelman, A. (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research, 15, 1593–1623.
Xu, S. G. and Reich, B. J. (2023) Bayesian nonparametric quantile process regression and estimation of marginal quantile effects. *Biometrics*, 79, 151–164. https://doi.org/10.1111/biom.13576