Nonparametric Distributional Regression via Quantile Regression

Cheng Peng **Stan Uryasev**
Stony Brook University

Abstract

This paper proposes a new approach to estimating the distribution of a response variable conditioned on factors. We model the conditional quantile function as a mixture (weighted sum) of basis quantile functions, with weights depending on these factors. The estimation problem is formulated as a convex optimization problem. The objective function is equivalent to the continuous ranked probability score (CRPS). This approach can be viewed as conducting quantile regressions for all confidence levels simultaneously while inherently avoiding quantile crossing. We use spline functions of factors as a primary example for the weight function. We prove an approximation property of the model. To address computational challenges, we propose a dimensionality reduction method using tensor decomposition and an alternating algorithm. Our approach offers flexibility, interpretability, tractability, and extendability. Numerical experiments demonstrate its effectiveness.

1 INTRODUCTION

This paper studies estimating the distribution of the response variable y conditioned on observing some factors (features) $\boldsymbol{x}$ given data pairs $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$. The quantile function of random variable y is defined by $Q_y(p) = \inf_q\{p \leq P(y \leq q)\}$.

Our main contributions are as follows:

- We develop a general model of conditional quantile function (Section 3). The conditional quantile function $Q_{y|\boldsymbol{x}}(p)$ is modeled by a mixture (weighted sum) of basis quantile functions. The weight of each basis quantile function is a function of the factors. We propose various possible choices of basis functions for weight and quantile.
- We formulate the parameter estimation problem as a convex optimization (Section 4). The formulation takes advantage of the linearity of the model with respect to the parameters.
- The parameter estimation problem can be viewed as constrained multiple quantile regression (Koenker, 2005) (Section 4.1). The model formulation rules out quantile crossing (Section 3.2).
- We study the approximation property of the model (Section 5).
- We propose a dimensionality reduction method using tensor decomposition(Section 6). The number of parameters grows linearly w.r.t. the dimension of factor. We propose an alternating algorithm for computation.

While the quantile functions and weight functions are arbitrary in principle, we primarily focus on quantile functions of common distributions and B-splines with nonnegative coefficients (de Boor and Daniel, 1974).

The model possesses an appealing interpretation that the factors have different impact on the body and the tail of the conditional distribution. The impact of factors on the shape of conditional quantile function can be traced analytically. Both the conditional quantile function and the quantile (hyper)surface have closed-form expressions (Section 3).

2 LITERATURE REVIEW

Mixture quantiles model (Tukey, 1962; Gilchrist, 2000) uses a combination of some basis functions functions to model the quantile function. Various basis functions have been used (Sillitto, 1969; Karvanen, 2006; Keelin, 2016).

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Mixture regression models the conditional distribution as a linear combination of basis functions (Quandt, 1972; DeSarbo and Cron, 1988; Villani et al., 2009; Yuksel et al., 2012).

Quantile regression estimates the conditional quantile of a given confidence level given some factors. Spline quantile regression has been presented in Koenker et al. (1994); He et al. (1998); Koenker (2011). Quantile crossing are addressed in Bondell et al. (2010); Chernozhukov et al. (2010).

Quantile regression process (Angrist et al., 2006; Reich et al., 2011; Lian et al., 2015; Frumento and Bottai, 2016; Yuan et al., 2017) refers to the regression coefficient as a function of the confidence level of quantile regression. Such model does not prevent quantile crossing. .

Continuous Ranked Probablity Score (CRPS, Matheson and Winkler (1976)) is a proper scoring rule that is frequently used to measure the quality of a distributional prediction and as the objective function of optimization (Hothorn et al., 2014; **?**; Gouttes et al., 2021; Berrisch and Ziel, 2021). Laio and Tamea (2006) proposes an equivalent definition that is useful to formulate convex optimization.

There is a variety of distributional regression models. To list a few, *Conditional transformation model* (Hothorn et al., 2014), *generalized additive models for location, scale and shape* (GAMLSS, Rigby and Stasinopoulos (2005)), *The linear and polynomial model* in Chernozhukov and Umantsev (2001), *Spline quantile function RNN* (Gasthaus et al., 2019).

Canonical polyadic (CP) decomposition (Hitchcock, 1927; Harshman, 1970; Carroll and Chang, 1970) for spline tensors has been applied in Computer Aided Design (CAD) to reduce model complexity (Pan et al., 2016).

Comparision Compared to existing literature, our approach has the following unique properties: *(i)* modeling the entire conditional quantile function with good approximation properties, closed-form expression and inherent prevention of quantile crossing, *(ii)* convex optimization to find the global optimum in parameter estimation, *(iii)* providing interpretability by allowing observation of how changes in factors influence the distribution's shape, and *(iv)* dimensionality reduction using tensor decomposition. A detailed literature review is in Appendix 1.

3 MODEL DESCRIPTION

This section describes the Factor Model of Mixture Quantiles and discusses noncrossing quantile and various perspectives to view the model.

- $p =$ confidence level of a quantile function
- $\boldsymbol{x} =$ vector of factors
- $\boldsymbol{a} =$ vector of parameters (unknown coefficients to be estimated)
- $Q(p) =$ quantile function
- $G(p, \boldsymbol{x}, \boldsymbol{a}) =$ model with parameters $\boldsymbol{a}$ that outputs the p-quantile of the response variable conditioned on observing factor $\boldsymbol{x}$
- $I =$ number of basis functions in the mixture
- $J =$ number of basis functions

3.1 Factor Model of Mixture Quantiles

The Factor Model of Mixture Quantiles is defined as

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = \sum_{i=0}^{I} f_i(\boldsymbol{x}, \boldsymbol{a}_i) Q_i(p) \ , \tag{1}$$

where $Q_0(p) = 1$. The basis functions $\{Q_i(p)\}_{i=0}^{I}$ are defined on the unit interval, and are linearly independent, i.e., any $Q_i(p)$ is not equal to a linear combination of other $Q_j(p)$, $j \neq i$. The weight function $f_i(\boldsymbol{x}, \boldsymbol{a}_i)$ of each basis function $Q_i(p)$ is a function of the factors $\boldsymbol{x}$.

The model formulation (1) is quite general since $\{Q_i(p)\}_{i=1}^{I}$ and $\{f_i(\boldsymbol{x}, \boldsymbol{a}_i)\}_{i=0}^{I}$ can be arbitrary functions, providing great flexibility in modeling heteroskedastic data and nonlinear relations. The weight functions determine how the factors impact the scale of basis quantile functions. It has the appealing interpretation that different factors may impact the different part of the distribution. The conditional quantile function has a closed-form expression if all basis quantile functions do. With our model, Monte Carlo simulation can be easily conducted with inverse transform sampling.

We focus on the case where $\{f_i(\boldsymbol{x}, \boldsymbol{a}_i)\}_{i=0}^{I}$ are splines with basis spline functions $\{B_{ij}(\boldsymbol{x})\}_{i,j=0}^{I,J}$. Furthermore, we use the same basis spline functions for all basis quantile functions $\{Q_i(p)\}_{i=1}^{I}$, i.e., $\forall i, j,\ B_{ij}(\boldsymbol{x}) = B_j(\boldsymbol{x})$. Then the model is defined as

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = \sum_{i=0}^{I} \sum_{j=0}^{J} a_{ij} B_j(\boldsymbol{x}) Q_i(p) \ , \tag{2}$$

where $B_0(x) = 1$, $Q_0(p) = 1$. Spline is adaptive to the data and can be optimized with the model in one shot. The linearity with respect to the coefficients not

only results in a convenient formulation of convex optimization, but also retains an interpretable factor model structure such that we can analyze the impact of each factor on the shape of the conditional quantile function.

To distinguish $\{Q_i(p)\}_{i=0}^I$ and $\{B_j(\boldsymbol{x})\}_{j=0}^J$, we hereafter refer to $\{Q_i(p)\}_{i=0}^I$ as basis quantile functions and $\{B_j(\boldsymbol{x})\}_{j=0}^J$ as basis spline functions. The meaning of the terms may be clearer if we expand the summation

$$\begin{aligned} G(p, \boldsymbol{x}, \boldsymbol{a}) =& a_{00} + \sum_{j=1}^{J} a_{0j} B_j(\boldsymbol{x}) + \sum_{i=1}^{I} a_{i0} Q_i(p) \\ &+ \sum_{i=1}^{I} \sum_{j=1}^{J} a_{ij} B_j(\boldsymbol{x}) Q_i(p) \ , \end{aligned} \tag{3}$$

where a_{00} is the constant location parameter, $\sum_{j=1}^J a_{0j} B_j(\boldsymbol{x})$ determines the conditional location, $\sum_{i=1}^I a_{i0} Q_i(p)$ is the base quantile function that does not vary with factors, $\sum_{i=1}^I \sum_{j=1}^J a_{ij} B_j(\boldsymbol{x}) Q_i(p)$ determines the conditional quantile function.

The complexity of the model is determined by the number of basis quantile functions, the number of factors and the form of weight functions. The number of parameters in the model is $(I+1) \times (J+1)$. The degrees of freedom are usually smaller due to constraints in optimization.

3.2 Noncrossing Quantile Model

This section studies general conditions for $\{B_j(\boldsymbol{x})\}_{j=1}^J$ and $\{Q_i(p)\}_{i=1}^I$ to guarantee noncrossing quantiles. We refer to a quantile model as noncrossing if it satisfies that the conditional quantile functions of any two different confidence levels do not cross conditioned on any value of factors. That is,

$$\begin{aligned} \forall \boldsymbol{x} \in \mathcal{X}, p_1, p_2, \ 0 < p_1 < p_2 < 1, \\ G(p_1, \boldsymbol{x}, \boldsymbol{a}) \leq G(p_2, \boldsymbol{x}, \boldsymbol{a}) \ , \end{aligned} \tag{4}$$

where $\mathcal{X}$ is the set of all possible values of factors $\boldsymbol{x}$. In the literature, the condition is often relaxed to that $G(p_1, \boldsymbol{x}, \boldsymbol{a}) \leq G(p_2, \boldsymbol{x}, \boldsymbol{a})$ for a certain $\boldsymbol{x} \in \mathcal{X}$ or for a certain selected confidence levels $\{p_m\}_{m=1}^M$. With such simplification, it could still require a large number of constraints to guarantee noncrossing.

Sufficient condition Instead of relaxing the condition (4), we propose using a sufficient condition. Note that the model (2) satisfies noncrossing condition (4) if three conditions are satisfied: *(i)* $\{Q_i(p)\}_{i=1}^I$ are nondecreasing; *(ii)* $\{B_j(\boldsymbol{x})\}_{j=1}^J$ are nonnegative; *(iii)* $\{a_{ij}\}_{i \neq 0}$ are nonnegative.

Nonnegativity constraint is relatively easy to impose in optimization. Although a sufficient condition may seem too restrictive, Theorem 5.1 in Section 5 shows that the model retains good approximation ability.

Basis quantile function For $\{Q_i(p)\}_{i=1}^I$, the following two types of nondecreasing functions can be used: *(i)* quantile functions of common distributions; *(ii)* monotone basis spline functions such as I-spline (Ramsay, 1988).

Quantile functions of common distributions are prefered when the shape of the distribution is known to be close to common distributions. When the distribution is multimodal, splines offer greater flexibility. Since splines are bounded on the unit interval, it needs to be combined with common quantile functions to model distributions with long tails.

Basis weight function For $\{B_j(\boldsymbol{x})\}_{j=1}^J$, the following three types of nonnegative function can be used: *(i)* nonnegative basis spline functions such as B-spline with nonnegative coefficients (de Boor and Daniel, 1974; Papp, 2011; Papp and Alizadeh, 2014); *(ii)* arbitrary basis spline functions with constraints on coefficients that characterize nonnegative polynomials (Lukács, 1918; Papp, 2011; Papp and Alizadeh, 2014); *(iii)* nonnegative kernel functions (Marteau-Ferey et al., 2020).

While the second type is more flexible than the first one, it is more computationally expensive to optimize. Besides B-spline, other choices include Bernstein polynomial and M-spline (Curry and Schoenberg, 1988). B-spline and M-spline differ by a constant. Papp (2011), Papp and Alizadeh (2014) show that piecewise Bernstein polynomial includes B-spline as a subset. The monotone I-spline used for $\{Q_i(p)\}_{i=1}^I$ is obtained by integrating M-spline.

3.3 Examples

This subsection provides several examples of the model described in Section 3. We refer to models from other research to demonstrate the wide applicability of our approach.

Example 1. Location-scale model of normal distribution

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = a_{00} + \sqrt{2}\text{erf}^{-1}(2p-1) a_{10} \ ,$$

where $\sqrt{2}\text{erf}^{-1}(2p-1)$ is the quantile function of standard normal distribution.

Example 2. Logistic-normal mixture of quantiles without factor dependence (Karvanen, 2006; Keelin, 2016; Peng et al., 2022)

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = a_{00} + \log(\frac{p}{1-p}) a_{10} + \sqrt{2}\mathrm{erf}^{-1}(2p-1) a_{20},$$

where $\log(\frac{p}{1-p})$ is the quantile function of standard logistic distribution.

Example 3. Linear model with normal noise

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = a_{00} + x a_{01} + \sqrt{2}\mathrm{erf}^{-1}(2p-1) \ .$$

Example 4. Quantile regression process model (Reich et al., 2011; Lian et al., 2015; Frumento and Bottai, 2016; Yuan et al., 2017)

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = \sum_{k_1=1}^{K} B_{k_1}(p) x_1 a_{1k_1} + \sum_{k_2=1}^{K} B_{k_2}(p) x_2 a_{2k_2} \ ,$$

where $B_k(p)$ can be Bernstein basis polynomials or B-spline basis.

Example 5. Autoregressive conditional heteroskedasticity-1 (ARCH(1)) model with normal noise (Engle, 1982)

$$G_t(p, \boldsymbol{x}, \boldsymbol{a}) = \sqrt{2}\mathrm{erf}^{-1}(2p-1) \cdot \sqrt{a_{10} + x_{t-1}^2 a_{11}} \ ,$$

where G_t is the quantile function of x_t at time t, x_{t-1} is the response variable at at time $t-1$. At every time t, the quantile function $G_t(p, \boldsymbol{x}, \boldsymbol{a})$ depends on the previous response variable x_{t-1}.

Example 6. Dynamic quantile model (Gourieroux and Jasiak, 2008)

$$G_t(p, \boldsymbol{x}, \boldsymbol{a}) = a_{00} + x_{t-1} a_{01} + (a_{10} + x_{t-1}^2 a_{11}) \log(\frac{p}{1-p}).$$

where G_t is the quantile function of x_t at time t, x_{t-1} is the response variable at at time $t-1$. At every time t, the quantile function $G_t(p, \boldsymbol{x}, \boldsymbol{a})$ depends on past response variable x_{t-1}. The response variable at time t is generated by plugging a sample p from uniform distribution $U(0,1)$ and the past response variable x_{t-1} to $G_t(p, \boldsymbol{x}, \boldsymbol{a})$.

Example 1, 3 and 5 shows that our model incorporates some of the most common models as special cases. Maximum likelihood estimation for these models are well studied. We propose a different estimation method in Section 4. Dynamic models such as Example 5 and 6 provide useful complements to classic time series models. Although we focus on linear function in $\boldsymbol{a}$ in subsequent sections on parameter estimation, the function $f(\boldsymbol{x}, \boldsymbol{a})$ in Example 5 is nonlinear in both $\boldsymbol{x}$ and $\boldsymbol{a}$. The parameters can still be estimated by optimizing our proposed objective function, but the optimization is not convex. Gourieroux and Jasiak (2008) also uses $|x|$ and $x^+ = \max\{0, x\}$, $x^- = \max\{0, -x\}$ to replace x, so that the factors remain nonnegative.

4 PARAMETER ESTIMATION BY CONVEX OPTIMIZATION

This section includes the basics of quantile regression (Koenker and Bassett, 1978) and the convex formulation for parameter estimation problem.

4.1 Quantile Regression

Consider a model of p-quantile of response variable $y|\boldsymbol{x}$ conditioned on factor $\boldsymbol{x}$ with parameter $\boldsymbol{a}$

$$q_p(y|\boldsymbol{x}) = G(p, \boldsymbol{x}, \boldsymbol{a}) \ . \tag{5}$$

Quantile regression estimates the parameter $\boldsymbol{a}$ by minimizing the scaled Koenker-Bassett error of the residuals

$$\min_{\boldsymbol{a}} \boldsymbol{\mathcal{E}}_p(y - G(p, \boldsymbol{x}, \boldsymbol{a})) \tag{6}$$

where the error is the expected pinball loss

$$\boldsymbol{\mathcal{E}}_p(Z) = E[\rho_p(Z)] \ , \tag{7}$$

$$\rho_p(Z) = pZ\mathbb{1}_{\{Z>0\}} - (1-p)Z\mathbb{1}_{\{Z\leq 0\}} \ , \tag{8}$$

Z is a random variable, $\mathbb{1}_{\{\cdot\}}$ is the indicator function that equals 1 when the equation in the bracket is true and 0 if otherwise.

4.2 Optimization Problem Statement for Parameter Estimation

- M = number of grid points of discretized confidence level
- N = sample size
- $\boldsymbol{Y}^N = (y_1, \cdots, y_N)$ = response variables of dependent variable
- $\boldsymbol{x}_n$ = vector of factors corresponding to response variable y_n, $n = 1, \cdots, N$;
- $\boldsymbol{x}^N = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ = set of factor vectors
- $\boldsymbol{a}$ = vector of parameters
- $\mathcal{A}$ = feasible set of $\boldsymbol{a}$ determined by constraints
- $L(\boldsymbol{a}; \boldsymbol{Y}^N, \boldsymbol{x}^N, p)$ = discrete residual random variable taking with equal probabilities the following values

$$y_n - \sum_{i=0}^{I} \sum_{j=0}^{J} a_{ij} B_j(\boldsymbol{x}_n) Q_i(p) \ , \quad n = 1, \ldots, N$$

The optimization problem statement for finding optimal parameters $\boldsymbol{a}$ is formulated as follows

Problem 1.

$$\min_{\boldsymbol{a}\in\mathcal{A}} \int_0^1 w(p)\boldsymbol{\mathcal{E}}_p\left(L(\boldsymbol{a};\boldsymbol{Y}^N,\boldsymbol{x}^N,p)\right)\mathrm{d}p\,, \tag{9}$$

where $w(p)$ is a nonnegative weight function satisfying $\int_0^1 w(p)\mathrm{d}p = 1$.

$w(p)$ can be chosen to focus on the distribution tail or body. Inherent from the pinball loss, the estimator is robust to outliers. While we consider the case where $f_i(\boldsymbol{x},\boldsymbol{a})$ are spline functions, the estimation method works for the general case (1).

Next, we discretize the problem (9) by using a grid on p. The resultant optimization problem is still a convex programming problem.

Problem 2. Discrete variant:

$$\min_{\boldsymbol{a}\in\mathcal{A}} \sum_{m=1}^{M} w_m\boldsymbol{\mathcal{E}}_{p_m}\left(L(\boldsymbol{a};\boldsymbol{Y}^N,\boldsymbol{x}^N,p_m)\right)\,, \tag{10}$$

where w_m are nonnegative weights satisfying $\sum_{m=1}^{M} w_m = 1$.

Similar to quantile regression, the problem can be reduced to linear programming. Note that although we select only a finite number of confidence levels in the discrete variant, the model still produces the entire conditional quantile function and ensures noncrossing quantiles. The estimation method also allows one to focus on the tail of the distribution by assigning higher weights on errors with tail confidence levels.

4.3 Equivalence to Constrained Joint Quantile Regression

Define M discrete random variables $\widehat{L}_m(\boldsymbol{a};\boldsymbol{Y}^N,\boldsymbol{x}^N)$, each taking with equal probabilities the values $y_n - \sum_{j=0}^{J}\lambda_{jm}B_j(\boldsymbol{x}_n)$, $n = 1,\ldots,N$, where λ is defined below. We can write (10) equivalently as

Problem 3.

$$\min_{\boldsymbol{a}\in\mathcal{A},\,\lambda} \sum_{m=1}^{M} w_m\boldsymbol{\mathcal{E}}_{p_m}\left(\widehat{L}_m(\boldsymbol{a};\boldsymbol{Y}^N,\boldsymbol{x}^N)\right) \tag{11}$$

$$\text{subject to } \sum_{i=0}^{I} Q_i(p_m)a_{ij} = \lambda_{jm}\,, \tag{12}$$

$$m = 1,\ldots,M,\ j = 0,\cdots,J\,.$$

This problem formulation makes it clear that the estimation in Problem 2 is equivalent to conducting several quantile regressions in one shot with constraints on the parameters. Since $\sum_{j=0}^{J}\lambda_{jm}B_j(\boldsymbol{x}_n)$ is a spline of scalar factor x_j, the objective function in (11) is the sum of objective functions of spline quantile regressions with confidence levels $\{p_m\}_{m=1}^{M}$. $\{\lambda_{jm}\}_{j=0,m=1}^{J,M}$ are associated by systems of equations (12), where the coefficients $\boldsymbol{a}$ are constrained by feasible set $\mathcal{A}$.

As J systems of equations, (12) can be written in matrix format $\boldsymbol{QA} = \boldsymbol{\Lambda}$, where $\boldsymbol{Q} = [Q_i(p_m)]'_{M\times(I+1)}$, $\boldsymbol{A} = [a_{ij}]_{(I+1)\times(J+1)}$, $\boldsymbol{\Lambda} = [\lambda_{jm}]'_{M\times(J+1)}$.

If the solution to the unconstrained problem (11) is in the feasible set defined by $\mathcal{A}$ and (12), we can solve Problem 3 with the following two steps.

1. Solve M spline quantile regressions separately

$$\min_{\boldsymbol{\lambda}} \boldsymbol{\mathcal{E}}_{p_m}\left(\widehat{L}_m(\boldsymbol{a};\boldsymbol{Y}^N,\boldsymbol{x}^N)\right),\ m = 1,\cdots,M. \tag{13}$$

2. Solve the systems of equations with constraints

$$\boldsymbol{a}\in\mathcal{A},\ \ \boldsymbol{QA} = \boldsymbol{\Lambda}\,. \tag{14}$$

In both steps, the solution can be nonunique. The uniqueness of solution to quantile regression is discussed in Koenker (2005). Portnoy (1991) shows that the number of distinct solutions to quantile regression when the confidence level p varies in $(0,1)$ is $\mathcal{O}(N\log N)$ in probability, which grows slower than the upper bound $\binom{N}{M\times(J+1)}$. The uniqueness of solution to the systems of equations depends on $\boldsymbol{Q}$ when $\mathcal{A} = \mathrm{R}^{(I+1)\times(J+1)}$. For proper choice of independent basis quantile functions $\{Q_i(p)\}_{i=1}^{I}$, $\boldsymbol{Q}$ has full column rank when $M \geq I+1$. To check feasibility when $\mathcal{A} = \mathrm{R}_{\geq 0}^{(I+1)\times(J+1)}$, we can solve the linear programming problem $\min_{\boldsymbol{a}\geq 0}||\boldsymbol{QA}-\boldsymbol{\Lambda}||_1$.

4.4 Equivalence to CRPS Minimization

Continuous Ranking Probability Score (CRPS) is frequently used to evaluate the quality of a distributional forecast. We show that Problem 1 is equivalent to CRPS minimization.

For a CDF F and an observation y_i of the response variable y, CRPS is defined by

$$CRPS(F,y_i) = \int_{\mathcal{R}}\left(F(x) - 1_{\{y_i\leq x\}}\right)^2\mathrm{d}x\,. \tag{15}$$

This is the squared distance between F and the CDF of a single observation y_i of the response variable y. For

the corresponding quantile function Q, i.e., the generalized inverse function of F, and the response variable y_i, CRPS has the following equivalent definition (Laio and Tamea, 2006; Fakoor et al., 2021)

$$CRPS(Q, y_i) = 2 \int_0^1 \rho_p \left(y_i - Q(p)\right) \mathrm{d}p \,. \tag{16}$$

Consider estimating the parameters by minimizing the sum of CRPS of the response variables

$$\min_{\boldsymbol{a}} \sum_{n=1}^{N} \int_0^1 \rho_p \left(y_n - G(p, \boldsymbol{x}_n, \boldsymbol{a})\right) \mathrm{d}p \,. \tag{17}$$

We see that (17) is equal to the objective function in Problem 1 with uniform weight $w(p)$ by exchanging the integral and sum.

Consider the special case where the basis quantile functions $\{Q_i(p)\}_{i=0}^{I}$ is $Q_0(p) = 1$. The model $G(p, \boldsymbol{x}, \boldsymbol{a})$ is not dependent on p and reduces to point estimation given $\boldsymbol{x}$. Since F becomes a step function, CRPS (15) reduces to $CRPS(F, y_i) = |y - y_i|$, where y is the discontinuous point of F. Thus CRPS minimization reduces to the least absolute deviation regression.

5 APPROXIMATION THEOREM

Our model can approximate any bounded conditional quantile model to arbitrary precision as the number of knots tends to infinity. Since the considered model has nonnegative parameters, classic approximation theorem of polynomials cannot be directly applied. The proof is in Appendix 2.

We adopt the following notations. $\mathcal{X}$ = hyperrectangular domain of factors. $T = (t_{ij})_{i=1,\cdots,I+1, j=1,\cdots,J}$ = matrix representation of subdivision of $[0,1] \times \mathcal{X}$. $||T|| = \max_{i,j} |t_{i+1,j} - t_{i,j}|$ mesh size of subdivision Z. cone(U) = cone of nonnegative linear combination of functions in a set of basis functions U. int = interior of a set. $\mathcal{P}(U, T)$ = set of piecewise functions where each piece (in the scaled representation) defined on a subdivision defined by Z is in cone(T). $G_{[0,1]\times\mathcal{X}}$ = cone of all continuous functions defined on $[0,1] \times \mathcal{X}$ that is nondecreasing in the first variable.

Theorem 5.1. *Consider I-spline basis $\{Q_i(p)\}_{i=0}^{\mathrm{I}}$ defined on $[0,1]$ and B-spline basis $\{B_j(\boldsymbol{x})\}_{j=0}^{\mathrm{J}}$ defined on $\mathcal{X}$. Furthermore, let T_i be an asymptotically nested sequence of subdivisions with mesh sizes approaching zero. Then the set $\cup_i \mathcal{P}(Q_i(p) \otimes B_j(\boldsymbol{x}), T_i)$ is a dense subcone of $G_{[0,1]\times\mathcal{X}}$.*

6 DIMENSIONALITY REDUCTION BY REDUCED RANK TENSOR

This section introduces the basics of CP decomposition of tensors and an alternating algorithm for dimensionality reduction. Multivariate B-spline is obtained by tensor product of univariate B-splines. In high dimensions, the tensor product B-spline has a gigantic number of basis functions. We propose an dimensionality reduction method by CP decomposition. The number of parameters of the reduced model grows linearly with respect to the dimension. It can be viewed as creating new basis functions by optimal linear combination of original ones. Our proposed alternating algorithm efficiently solves a smaller convex optimization problem in each iteration.

6.1 Tensor and CP Decomposition

We adopt the following notations. $\boldsymbol{A}$ = parameter tensor. K = number of scalar factors. R = rank of parameter tensor. $\otimes$ = tensor product of two vectors. $\bigotimes_{k=1}^{K}$ = tensor product of K vectors. $< \cdot, \cdot >_F$= Frobenius inner product of two tensors.

A tensor is a multidimensional array. CP decomposition for a rank-R tensor $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = \sum_{r=1}^{R} \boldsymbol{u}_r^{(0)} \otimes \cdots \otimes \boldsymbol{u}_r^{(K)} \,, \tag{18}$$

where $\{\boldsymbol{u}_r^k\}_{r=1,k=0}^{R,K}$ are vectors.

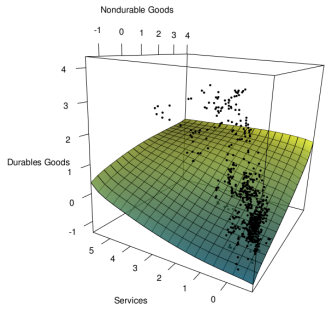
The following equation is useful for subsequent interpretation of reduced rank method

$$\bigotimes_{k=0}^{K} \boldsymbol{u}_r^{(k)} \cdot \bigotimes_{k=0}^{K} \boldsymbol{v}_r^{(k)} = \prod_{k=0}^{K} \boldsymbol{u}_r^{(k)} \cdot \boldsymbol{v}_r^{(k)} \,. \tag{19}$$

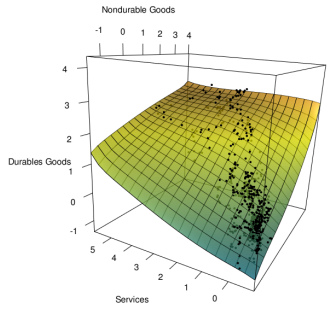
6.2 Reduced Rank Method for Parameter Tensor

We adopt the following notations. K = number of scalar factors. L = number of basis spline functions for each scalar factor. $J = K \times L$. $\boldsymbol{A}$ = parameter tensor with the same elements as parameter vector $\boldsymbol{a}$. $\boldsymbol{B}_k(x) = (B_{k1}(x), \cdots, B_{kL}(x))$ = vector of univariate spline basis of scalar factor x_k. $\bigotimes_{k=1}^{K} \boldsymbol{B}_k(x_k)$ = tensor product of K vectors of univariate spline basis of scalar factors. $\boldsymbol{Q}(p) = (Q_0(p), \cdots, Q_I(p))$ = vector of basis quantile functions. $\geq$ represents elementwise inequality when used on vector, matrix and tensor.

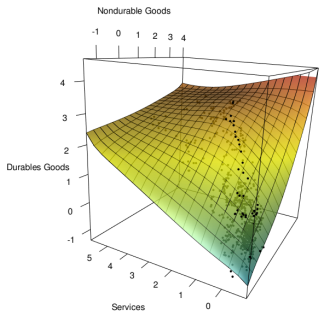
For B-spline, L = degree of univariate polynomial+number of knots+1. Multivariate B-spline basis is obtained by tensor product of univariate B-spline basis.



Quantile surface, confidence level $p = 0.05$



Quantile surface, confidence level $p = 0.5$



Quantile surface, confidence level $p = 0.95$

Figure 1: Quantile surfaces of confidence levels $p = 0.05, 0.5, 0.95$ in Experiment 2. The dots are the data points. The quantile surface plots the conditional p-quantile of the response variable as a function of the factors.

Table 1: Average coverage rates of the prediction intervals in 10-fold cross-validation

Model	0.98	0.9	0.7	0.5	0.3	0.1	Ave. Diff.
Factor model of mixture quantiles	0.980	0.904	0.706	0.484	0.279	0.087	**0.007**
Gaussian mixture regression	0.986	0.907	0.690	0.479	0.277	0.091	0.013
Generalized additive model	0.984	0.920	0.711	0.492	0.276	0.090	0.013
NGBoost	0.970	0.894	0.696	0.490	0.288	0.094	0.008

Result of Experiment 1. The coverage rate is calculated by the percentage of the response variable that falls within the predicted conditional interval of two specified quantiles. The average differences between the realized coverage rate and the target coverage rate are calculated. The target coverage rates are obtained by confidence intervals $(0.01, 0.99)$, $(0.05, 0.95)$, $(0.15, 0.85)$,$(0.25, 0.75)$,$(0.35, 0.65)$,$(0.45, 0.55)$.

With the above notations, we can write the model (2) in tensor format

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = \left\langle \boldsymbol{A},\ \boldsymbol{Q}(p) \otimes \bigotimes_{k=1}^{K} \boldsymbol{B}_k(x_k) \right\rangle_F . \tag{20}$$

Note that in our model formulation, $\boldsymbol{A} \geq 0$. Consider a nonnegative version of CP decomposition (18) on the parameter tensor $\boldsymbol{A}$ where $\boldsymbol{u}_r^{(k)} \geq 0$. With (18)(19), we have

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = \sum_{r=1}^{R} \left(\boldsymbol{u}_r^{(0)} \cdot \boldsymbol{Q} \right) \prod_{k=1}^{K} \left(\boldsymbol{u}_r^{(k)} \cdot \widehat{\boldsymbol{B}}_K \right).$$

Thus reduced rank method has a straightforward interpretation. The model is reduced to a linear combination of R functions from $(I+1)(J+1)$. In the extreme case where $R = 1$, the model reduced to a heteroscedastic model where the conditional quantile function is $\boldsymbol{u}_r^{(0)} \cdot \boldsymbol{Q}$, whose scale is controlled by a scalar function $\prod_{k=1}^{K} \boldsymbol{u}_r^{(k)} \cdot \boldsymbol{B}_k$. It can be regarded as obtaining R new basis quantile functions $\{\boldsymbol{u}_r^{(0)} \cdot \boldsymbol{Q}\}_{r=1}^{R}$, likewise for basis spline functions. The new basis, having a smaller number, can still have undesirable smoothness condition. We expect better performance when it is used along with penalties in P-spline. The reduced basis functions are a linear combination of all original basis functions. This is in contrast to sparse optimization which leads to a subset of basis functions by forcing zeros among the parameters. We find that the solution to Problem 2 is often sparse with many small nonzero values. Thus the low rank decomposition is expected to produce good approximation, although the decomposition (18) is not always valid for any $\boldsymbol{A}$ for a low rank R.

6.3 Alternating Algorithm

We propose an alternating algorithm to find $\boldsymbol{A}$. Let $\boldsymbol{U}_k = (\boldsymbol{u}_1^{(k)}, \cdots, \boldsymbol{u}_R^{(k)})$, $k = 1, \cdots, K$. Define $\widetilde{L}\left(\{\boldsymbol{U}_k\}_{k=0}^{K}; \boldsymbol{Y}^N, \boldsymbol{x}^N, p\right)$ = discrete residual random variable taking with equal probabilities the following

values for $n = 1, \ldots, N$

$$y_n - \sum_{r=1}^{R} \left(u_r^{(0)} \cdot Q(p) \right) \prod_{k=1}^{K} \left(u_r^{(k)} \cdot B_k(x_n) \right) .$$

Each step in the algorithm is a convex optimization problem. The objective function is nonincreasing in each step. The algorithm will find a local minimum at the end. Different initialization is needed for finding a smaller local minimum.

Algorithm 1 Alternating Algorithm for Estimating Reduced Rank Parameter Tensor

Input: Data $\{(y_i, x_i)\}_{i=1}^N$; confidence levels $\{p_m\}_{m=1}^M$; model $G(p, \theta, a)$; error function $\mathcal{E}_p$; threshold ϵ

1: Initialize $\{U_k\}_{k=0}^K$
2: **repeat**
3: **for** $k = 0$ to K **do**
4: Update $U_k^{(s)}$ to $U_k^{(s+1)}$ with fixed $\{U_{k_t}^{(s)}\}_{k_t \neq k}$

$$U_k^{(s+1)} = \underset{U_k \geq 0}{\operatorname{argmin}} \sum_{m=1}^{M} w_m \mathcal{E}_{p_m} \left(\widetilde{L}(\{U_k^{(s)}\}_{k=0}^K; Y^N, x^N, p_m) \right)$$

5: **end for**
6: $s += 1$
7: **until** $||A^{(s+1)} - A^{(s)}||_F \leq \epsilon$

Output: Parameter tensor A

7 NUMERICAL EXPERIMENTS

Experiment 1 This experiment compares the performance of our model[1] with three benchmarks. The dataset is Combined Cycle Power Plant from UCI Repository (4 features, 1 target and 9568 data points). We fit our model and three benchmarks in 10-fold cross-validation and compare CRPS and coverage rate. The coverage rate is calculated by the percentage of the response variable that falls within the predicted conditional interval of two specified quantiles. The detailed setup of all experiments is in Appendix 4.

The results are summarized in Table 1 and Table 2. Our approach demonstrates an improved average out-of-sample coverage rate. The average out-of-sample CRPS for our method is lower than that of GMR and GAMLSS but higher than NGBoost. Our approach has the shortest fitting time.

[1]The R implementation is available at `https://github.com/ch-pg/distributional_regression`.

Table 2: CRPS and time cost

Model	FMMQ	GMR	GAMLSS	NGBoost
CRPS	0.080	0.088	0.081	**0.075**
Std CRPS	**0.002**	0.003	0.003	**0.002**
Time(s)	**11.1**	109.0	29.0	12.0

FMMQ = Factor model of mixture quantiles, GMR = Gaussian mixture regression, GAMLSS = generalized additive models for location, scale and shape.std CRPS = standard deviation of CRPS in 10-fold cross-validation. Time(s) = average time cost in seconds.

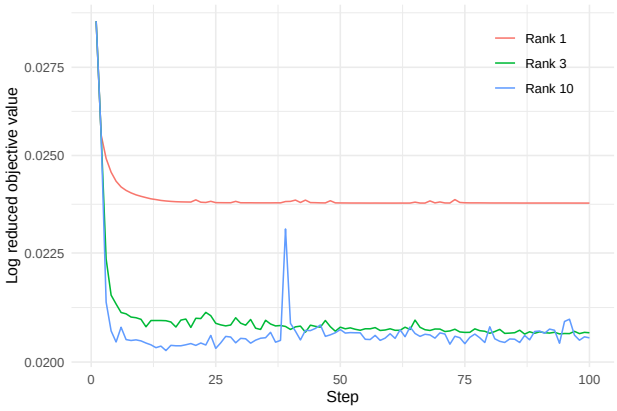


Figure 2: Experiment 3: Semilog plot of training loss using Algorithm 1 on models with rank 1, 3 and 10.

Experiment 2 Figure 1 visualizes the quantile surfaces of confidence levels $q = 0.05, 0.5, 0.95$. The dataset is monthly CPI of three categories (McCracken and Ng, 2016; Zimmermann, Zimmermann).

Experiment 3 The experiment tests the convergence of the alternating algorithm and loss of fit of the dimensionality reduction method for models with rank 1, 3, and 10. We calculate the log of reduction in the objective function during training, i.e., initial training loss minus training loss at current step.

The log reduced objective value at each step is displayed in Figure 2. We observe that in all cases, the value converges in less than 20 iterations, demonstrating the effectiveness of the proposed algorithm. The optimal objective values of models with rank 3 and 10 are very close, while the model with rank 1 has a larger objective value as expected. The nonmonotonicity of the curves is caused by numerical errors.

8 CONCLUSION

The proposed nonparametric distributional regression has a strong approximation property, can be ef-

ficiently estimated by convex optimization, and allows for dimensionality reduction by tensor decomposition. We analysis the model specifications and equivalent formulations, and propose an alternating algorithm for computation.

References

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006). Quantile regression under misspecification, with an application to the u.s. wage structure. *Econometrica 74*(2), 539–563.

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

Benaglia, T., D. Chauveau, D. R. Hunter, and D. S. Young (2009). mixtools: An r package for analyzing mixture models. *Journal of Statistical Software 32*(6), 1–29.

Berrisch, J. and F. Ziel (2021). Crps learning. *Journal of Econometrics*.

Bondell, H. D., B. J. Reich, and H. Wang (2010). Noncrossing quantile regression curve estimation. *Biometrika 97*(4), 825–838.

Bremnes, J. B. (2020). Ensemble postprocessing using quantile function regression based on neural networks and bernstein polynomials. *Monthly Weather Review 148*(1), 403 – 414.

Carroll, J. D. and J. J. Chang (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika 35*, 283–319.

Chen, L., J. J. Dolado, and J. Gonzalo (2021). Quantile factor models. *Econometrica 89*(2), 875–910.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Chernozhukov, V. and L. Umantsev (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics 26*(1), 271–292.

Curry, H. B. and I. J. Schoenberg (1988). *On Pólya Frequency Functions IV: The Fundamental Spline Functions and their Limits*, pp. 347–383. Boston, MA: Birkhäuser Boston.

de Boor, C. and J. W. Daniel (1974). Splines with nonnegative *b*-spline coefficients. *Mathematics of Computation 28*(126), 565–568.

DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification 5*(2), 249–282.

Eilers, P. H. and B. D. Marx (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems 66*(2), 159–174.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89 – 121.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica 50*(4), 987–1007.

Fakoor, R., T. Kim, J. Mueller, A. J. Smola, and R. J. Tibshirani (2021). Flexible model aggregation for quantile regression.

Frumento, P. and M. Bottai (2016). Parametric modeling of quantile regression coefficient functions. *Biometrics 72*(1), 74–84.

Gasthaus, J., K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski (2019, 16–18 Apr). Probabilistic forecasting with spline quantile function rnns. In K. Chaudhuri and M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Volume 89 of *Proceedings of Machine Learning Research*, pp. 1901–1910. PMLR.

Ghahramani, Z., G. E. Hinton, et al. (1996). The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Gilchrist, W. (2000). Statistical modelling with quantile functions.

Gourieroux, C. and J. Jasiak (2008). Dynamic quantile models. *Journal of Econometrics 147*(1), 198–205. Econometric modelling in finance and risk management: An overview.

Gouttes, A., K. Rasul, M. Koren, J. Stephan, and T. Naghibi (2021). Probabilistic time series forecasting with implicit quantile networks.

Harshman, R. A. (1970). Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-model factor analysis.

He, X., P. Ng, and S. Portnoy (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(3), 537–550.

Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics 6*(1-4), 164–189.

Hothorn, T., T. Kneib, and P. Bühlmann (2014). Conditional transformation models. *Journal of the*

Royal Statistical Society. Series B (Statistical Methodology) 76(1), 3–27.

Karvanen, J. (2006). Estimation of quantile mixtures via l-moments and trimmed l-moments. *Computational Statistics & Data Analysis 51*(2), 947–959.

Keelin, T. W. (2016). The metalog distributions. *Decision Analysis 13*(4), 243–277.

Koenker, R. (1984). A note on l-estimates for linear models. *Statistics & Probability Letters 2*(6), 323–325.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics 25*(3), 239 – 262.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika 81*(4), 673–680.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*(3), 455–500.

Laio, F. and S. Tamea (2006). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences 11*, 1267–1277.

Lian, H., J. Meng, and Z. Fan (2015). Simultaneous estimation of linear conditional quantiles with penalized splines. *Journal of Multivariate Analysis 141*, 1–21.

Lian, H., W. Zhao, and Y. Ma (2019). Multiple quantile modeling via reduced-rank regression. *Statistica Sinica 29*(3), 1439–1464.

Lukács, F. (1918, September). Verschärfung des ersten Mittelwertsatzes der Integralrechnung für rationale Polynome. *Mathematische Zeitschrift 2*(3), 295–305.

Marteau-Ferey, U., F. Bach, and A. Rudi (2020). Nonparametric models for non-negative functions. *Advances in neural information processing systems 33*, 12816–12826.

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*(10), 1087–1096.

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier.

Pan, M., W. Tong, and F. Chen (2016). Compact implicit surface reconstruction via low-rank tensor approximation. *Computer-Aided Design 78*, 158–167. SPM 2016.

Papp, D. (2011). Optimization models for shape-constrained function estimation problems involving nonnegative polynomials and their restrictions.

Papp, D. and F. Alizadeh (2014). Shape-constrained estimation using nonnegative splines. *Journal of Computational and Graphical Statistics 23*(1), 211–231.

Peng, C., Y. Li, and S. Uryasev (2022). Mixture quantiles calibrated with constrained linear regression.

Portnoy, S. (1991). Asymptotic behavior of the number of regression quantile breakpoints. *SIAM Journal on Scientific and Statistical Computing 12*(4), 867–883.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association 67*(338), 306–310.

Rabanser, S., O. Shchur, and S. Günnemann (2017). Introduction to tensor decompositions and their applications in machine learning.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science 3*(4), 425–441.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological bulletin 86*(3), 446.

Reich, B. J., M. Fuentes, and D. B. Dunson (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association 106*(493), 6–20. PMID: 23459794.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics 54*, 507–554.

Rockafellar, R. T. and S. Uryasev (2000). Optimization of Conditional Value-at-Risk. *Journal of Risk 2*, 21–41.

Rockafellar, R. T. and S. Uryasev (2002). Conditional Value-at-Risk for general loss distributions. *Journal of Banking and Finance*, 1443–1471.

Sillitto, G. P. (1969). Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika 56*(3), 641–650.

Sottile, G. and P. Frumento (2022). Robust estimation and regression with parametric quantile functions. *Computational Statistics & Data Analysis 171*, 107471.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics 33*(1), 1–67.

Villani, M., R. Kohn, and P. Giordani (2009). Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics 153*(2), 155–173.

Vincent, S. B. (1912). *The Functions of the Vibrissae in the Behavior of the White Rat...*, Volume 1. University of Chicago.

Yuan, Y., N. Chen, and S. Zhou (2017). Modeling regression quantile process using monotone b-splines. *Technometrics 59*(3), 338–350.

Yuksel, S. E., J. N. Wilson, and P. D. Gader (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems 23*(8), 1177–1193.

Zabarankin, M., S. Uryasev, et al. (2016). *Statistical decision problems*. Springer.

Zhang, Q., A. Makur, and K. Azizzadenesheli (2022). Functional linear regression of cdfs.

Zimmermann, C. Astonishingly different inflation across goods categories. Accessed: 2023-03-14.

Zou, H. and M. Yuan (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics 36*(3), 1108 – 1126.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

1 Detailed Literature Review

Mixture quantiles model uses a combination of some basis functions functions to model the quantile function. The idea is studied in Tukey (1962); Gilchrist (2000). Various basis functions have been used, such as orthogonal polynomials (Sillitto, 1969), a mixture of normal and Cauchy distributions with linear and quadratic terms (Karvanen, 2006), and a modified logistic distribution (Keelin, 2016). Vincent (1912); Ratcliff (1979) have a different motivation to obtain population quantile function by averaging individual quantile functions. However, these methods lack factor dependence and some do not guarantee that the resulting quantile function is nondecreasing. Our model uses mixture quantiles as the conditional quantile function (Section 3). Note that mixture quantiles and mixture densities are two different families of distributions.

Mixture regression shares a similar idea to our approach, where the conditional distribution is modeled as a linear combination of basis functions. The conditional density function of the widely applied Gaussian mixture regression is modeled by a mixture (weighed sum) of Gaussian densities, where the weight, mean and variance of each component can be modeled as functions of factors. Gaussian mixture regression has been studied under different names, such as switching regression model, mixture of experts (Quandt, 1972; DeSarbo and Cron, 1988; Villani et al., 2009; Yuksel et al., 2012). Similarly, in our model formulation, the weights of quantile functions are functions of the factors (Section 3). Mixture regression is generally computationally expensive to estimate, while our parameter estimation can be solved by convex optimization. The weight function in mixture regression must sum up to one, which is typically achieved by the softmax transformation, while our model only requires nonnegative weight functions.

Quantile regression estimates the conditional quantile of a given confidence level given some factors. The application of spline functions to conduct nonparametric quantile regression has been presented in Koenker et al. (1994); He et al. (1998); Koenker (2011). However, conducting multiple quantile regressions separately can result in quantile crossing, particularly in nonlinear models, which impedes the interpretation of the results. To mitigate this problem, various methods have been proposed, such as imposing extra constraints (Bondell et al., 2010) or rearrangement (Chernozhukov et al., 2010). The constraints are often imposed at observed data points or at some confidence levels. We present a novel approach to quantile regression that guarantees noncrossing quantiles without any post-processing. Our method estimates the entire conditional distribution of the response variable, not just the quantiles at several confidence levels. The proper choice of basis functions ensures the validity of the noncrossing property at all confidence levels (Section 3). The objective function in our approach is a linear combination of pinball losses at different confidence levels (Section 4), which has been adopted in various parameter estimation methods (Koenker, 1984; Zou and Yuan, 2008; Sottile and Frumento, 2022).

Quantile regression process refers to the regression coefficient as a function of the confidence level of quantile regression. Angrist et al. (2006) shows that the rescaled quantile regression process converges to a zero-mean Gaussian process. A number of independent but closely related studies have explored the use of various basis functions for modeling the quantile regression process, such as Bernstein basis polynomials (Reich et al., 2011), P-spline basis (Lian et al., 2015), common parameterized functions (Frumento and Bottai, 2016), and monotone B-spline (Yuan et al., 2017). By rearranging the model formulation in Section 3 and viewing each polynomial term as a factor, it can be shown that our model coincides with the quantile regression process models.

There are several notable differences between our model and existing models on quantile regression process. First, the parameter estimation in Frumento and Bottai (2016) and Reich et al. (2011) involves numerical integration and Bayesian method, respectively, while our approach uses convex optimization, which is more efficient. Second, it it can be challenging to ensure noncrossing quantiles when directly modeling the linear quantile regression process. Reich et al. (2011) introduces prior latent unconstrained variables; Frumento and Bottai (2016) checks the nonnegativity of derivative; Yuan et al. (2017) uses linear constraints when the feasible set is a

bounded convex polytope; Lian et al. (2015) does not guarantee noncrossing quantiles. In contrast, our model guarantees noncrossing conditional quantiles at any two distinct confidence levels and any given factor value. Third, using a spline to model quantile regression process results in bounded conditional quantile function, which may lead to severe underestimation of tail risk. Thus it is important to supplement splines with quantile functions of common distributions. Furthermore, all aforementioned quantile regression process models study linear models and can be regarded as special cases of our approach. To ensure that a linear quantile model does not have quantile crossing given any factor values, the coefficients of each factor must be equal for all confidence levels. This condition leads to a homoscedastic model.

Continuous Ranked Probablity Score (CRPS, Matheson and Winkler (1976)) is a proper scoring rule that is frequently used to measure the quality of a distributional prediction when the response variable is a scalar. Laio and Tamea (2006) proposes an equivalent definition that is useful to formulate convex optimization. Hothorn et al. (2014); Gouttes et al. (2021); Berrisch and Ziel (2021) use CRPS as the objective function of optimization in learning tasks. Zhang et al. (2022) proposes a related model where the CDF is a linear combination of basis CDFs. Although not mentioned in the paper, their estimation procedure minimizes CRPS.

Quantile model aggregation has a similar model formulation, but aims to aggregate multiple conditional quantile models to improve overall performance. The aggregation can be performed across different factor values and different quantile levels. Berrisch and Ziel (2021) uses B-splines as weight function across confidence levels, while Fakoor et al. (2021) uses neural networks as weight function across different factor values and different confidence levels. Viewing each model as a basis function, our approach can be immediately adapted for quantile model aggregation across different factor values. On the other hand, Fakoor et al. (2021) considers a regression model called the deep quantile regression by using constant functions as individual models. The conditional quantile function is a neural network, which requires extra efforts to ensure monotonicity.

Conditional transformation model (Hothorn et al., 2014) finds the optimal conditional spline transformation of the conditional distribution, such that the quantile function of the transformed conditional distribution can be modeled by a predetermined quantile function. In contrast, our approach can be viewed as modeling the inverse of the conditional transformation. Our model is also related to the *generalized additive models for location, scale and shape* (GAMLSS, Rigby and Stasinopoulos (2005)). When there is only one basis quantile function, e.g., quantile function of normal distribution, the mean and variance are modeled as splines of the factor, which is equivalent to generalized additive models for location and scale. *The linear and polynomial model* in Chernozhukov and Umantsev (2001) can be regarded as a special case of our model. The extension of our approach to neural networks is related to *Spline quantile function RNN* (Gasthaus et al., 2019), which models the parameter of piecewise linear splines with neural networks and estimate the parameters by CRPS minimization. Similarly, the model ensemble method in Bremnes (2020) uses Bernstein polynomials.

Multivariate B-spline is obtained by tensor product of univariate B-splines. The CP decomposition for spline tensors has been applied in Computer Aided Design (CAD) to reduce model complexity (Pan et al., 2016). For a comprehensive review of tensor decomposition, the readers are referred to Kolda and Bader (2009); Rabanser et al. (2017). This study applies the technique to statistical estimation. In high dimensions, the tensor product B-spline has a gigantic number of basis functions. The dimensionality reduction method by CP decomposition can be viewed as creating new basis functions by optimal linear combination of original ones. Our approach is different from existing work on dimensionality reduction for quantile regression. Lian et al. (2019) proposes reduced matrix rank regression for homoscedastic multiple quantile modeling. Chen et al. (2021) proposes a different approach to dimensionality reduction for quantile regression determining different principle components for different confidence levels. Our approach also results in different factors for different confidence levels (Section 6).

Distributional estimation has been extensively studied. *Conformal prediction* (**?**) post-processes a point forecast to obtain prediction interval with guarantees on coverage. *Conformalized quantile regression* (**?**) conducts quantile regression, then conformalizes the quantile estimate. Our method, when focused on a specific confidence level, can be conformalized using **?**.

Tree-based methods and neural networks are very effective on tabular data, see for example **?????**. However, they often do not provide a closed-form expression for the conditional quantile function. *NGBoost* (**?**) proposes a modular boosting approach consisting of base learner, parametric distribution, and score function. Our model formulation is equivalent to NGBoost model formulation when the base learner is the basis weight functions

$\{a_j B_j(\boldsymbol{x_n})\}_{j=1}^J$ with unknown parameter a_j; the parametric distribution has quantile function $\sum_{i=1}^I a_i Q_i(p)$; the scoring rule is CRPS. Similarly, the distribution parameters are obtained by an additive combination of base learner outputs, that is, $a_i = \sum_{j=1}^J a_{ij} B(\boldsymbol{x_i})$.

2 Proof of Theorem 5.1

Proof. We want to prove that for any $G(p, \boldsymbol{x}, \boldsymbol{\theta}) \in G_{[0,1]}$ and any $\epsilon > 0$, there exists $F(p, \boldsymbol{x}) \in \cup_i \mathcal{P}(Q_i(p) \otimes B_j(\boldsymbol{x}), T_i)$ such that

$$|F(p, \boldsymbol{x}) - G(p, \boldsymbol{x}, \boldsymbol{\theta})| < \epsilon$$

for $\boldsymbol{x} \in \mathcal{X}$. Note that $\forall G(p, \boldsymbol{x}, \boldsymbol{\theta}) \in G_{[0,1]\times\mathcal{X}}$, we have $\frac{\partial}{\partial p} G(p, \boldsymbol{x}, \boldsymbol{\theta}) \geq 0$, since quantile function is nondecreasing in p. By Theorem 4.1 of Papp (2011), for any $\epsilon > 0$, there exists a tensor-product B-spline $f(p, \boldsymbol{x})$ with nonnegative coefficients such that

$$|\frac{\partial}{\partial p} G(p, \boldsymbol{x}, \boldsymbol{\theta}) - f(p, \boldsymbol{x})| < \epsilon.$$

Integrating $f(p, \boldsymbol{x})$ with respect to p, we obtain

$$\int f(p, \boldsymbol{x}) \mathrm{d}p = \widetilde{F}(p, \boldsymbol{x}) + \widetilde{t}(\boldsymbol{x}),$$

where $t(\boldsymbol{x})$ is a differentiable function. Since I-spline is defined by the integral of B-spline, we have $\widetilde{F}(p, \boldsymbol{x}) \in \cup_i \mathcal{P}(Q_i(p) \otimes B_j(\boldsymbol{x}), T_i)$. For any $\frac{1}{2}\epsilon > 0$, there exists a B-spline $t(\boldsymbol{x})$ such that

$$|\widetilde{F}(0.5, \boldsymbol{x}) + t(\boldsymbol{x}) - G(0.5, \boldsymbol{x}, \boldsymbol{\theta})| < \frac{1}{2}\epsilon.$$

Let $\widetilde{t}(\boldsymbol{x}) = t(\boldsymbol{x})$. We have $\widetilde{F}(p, \boldsymbol{x}) + \widetilde{t}(\boldsymbol{x}) \in \cup_i \mathcal{P}(Q_i(p) \otimes B_j(\boldsymbol{x}), T_i)$. We also have

$$\begin{aligned}
&|\widetilde{F}(p, \boldsymbol{x}) + \widetilde{t}(\boldsymbol{x}) - G(p, \boldsymbol{x}, \boldsymbol{\theta}) - (\widetilde{F}(0.5, \boldsymbol{x}) + \widetilde{t}(\boldsymbol{x}) - G(0.5, \boldsymbol{x}, \boldsymbol{\theta}))| \\
<& \max\left\{ \int_0^{\frac{1}{2}} |\frac{\partial}{\partial p} G(p, \boldsymbol{x}, \boldsymbol{\theta}) - f(p, \boldsymbol{x})| \mathrm{d}p, \int_{\frac{1}{2}}^1 |\frac{\partial}{\partial p} G(p, \boldsymbol{x}, \boldsymbol{\theta}) - f(p, \boldsymbol{x})| \mathrm{d}p \right\} \\
<& \frac{1}{2}\epsilon
\end{aligned}$$

Hence for any $\boldsymbol{x} \in \mathcal{X}$, we have

$$|\widetilde{F}(p, \boldsymbol{x}) + \widetilde{t}(\boldsymbol{x}) - G(p, \boldsymbol{x}, \boldsymbol{\theta})| < \frac{1}{2}\epsilon + \frac{1}{2}\epsilon = \epsilon.$$

□

3 Asymptotic Property

This section contains the asymptotic properties of the estimator. We adopt the following notations. $\widehat{\mathcal{E}}_N(\boldsymbol{a}) = \sum_{m=1}^M w_m \boldsymbol{\mathcal{E}}_{p_m}\left(L(\boldsymbol{a}; \boldsymbol{Y}^N, \boldsymbol{x}^N, p_m)\right)$. $\widehat{\boldsymbol{a}}_N = \mathrm{argmin}_{\boldsymbol{a}\in\mathcal{A}} \widehat{\mathcal{E}}_N(\boldsymbol{a})$. $\boldsymbol{a}_0 =$ true parameter.

Theorem 3.1, 3.2 in the following from Frumento and Bottai (2016) are applications of Newey and McFadden (1994).

Theorem 3.1. *Assume that $\mathcal{A}$ is a compact set. If there is a function $\widehat{\mathcal{E}}_0(\boldsymbol{a})$ such that (i) $\widehat{\mathcal{E}}_N(\boldsymbol{a})$ converges uniformly in probability to $\widehat{\mathcal{E}}_0(\boldsymbol{a})$; (ii) $\widehat{\mathcal{E}}_0(\boldsymbol{a})$ is uniquely minimized by $\widehat{\boldsymbol{a}}$; (iii) $\widehat{\mathcal{E}}_0(\boldsymbol{a})$ is continuous at $\widehat{\boldsymbol{a}}$. Then $\widehat{\boldsymbol{a}}_\mathrm{N} \xrightarrow{d} \boldsymbol{a}_0$.*

Theorem 3.2. *Suppose that the conditions Theorem 3.1 are satisfied, and (i) $\boldsymbol{a}_0$ is an interior point of $\mathcal{A}$; (ii) $\widehat{\mathcal{E}}_\mathrm{N}(\boldsymbol{a})$ is twice continuously differentiable in a neighborhood of $\boldsymbol{a}_0$; (iii) $\sqrt{n}\nabla_{\boldsymbol{a}}\widehat{\mathcal{E}}_\mathrm{N}(\boldsymbol{a}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Omega)$; (iv) there is $H(\boldsymbol{a})$ that is continuous at $\boldsymbol{a}_0$ and $\sup_{\boldsymbol{\theta}_0\in\mathcal{A}} ||\nabla_{\boldsymbol{aa}}\widehat{\mathcal{E}}_\mathrm{N}(\boldsymbol{a}_0) - H(\boldsymbol{a})|| \xrightarrow{p} \boldsymbol{0}$; (v) $H = H(\boldsymbol{a}_0)$ is nonsingular. Then*

$$\sqrt{N}(\widehat{\boldsymbol{a}}_\mathrm{N} - \boldsymbol{a}_0) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, H^{-1}\Omega H^{-1}) \,. \tag{1}$$

4 Details of Numerical Experiments

Experiments were conducted on a 13-inch MacBook Pro (2020) equipped with a 2 GHz Quad-Core Intel Core i5 processor, 16 GB of 3733 MHz LPDDR4X memory, and integrated Intel Iris Plus Graphics with 1536 MB of VRAM, running macOS Sequoia 15.1.1.

4.1 Experiment 1

Data The dataset is Combined Cycle Power Plant from UCI Repository. The data are standardized to have zero median and unit interquartile range. The data is randomly shuffled before experiments.

Model Based on preliminary examination of the dataset, we choose the following model for efficient computation.

- $Q_{\mathtt{N}}(p) = \sqrt{2}\mathrm{erf}^{-1}(2p-1) =$ quantile function of standard normal distribution

Consider the Two-Factor Model of Mixture Quantiles

$$G(p, \boldsymbol{x}, \boldsymbol{a}) = f_0(\boldsymbol{x}) + Q_{\mathtt{N}}(p) f_1(\boldsymbol{x}) \,, \tag{2}$$

For the B-splines f_0, f_1, we use basis of degree 2 with 3 equidistant knots. Hyperparameter of the model includes the number of basis quantile functions, the degree of basis weight functions, the position of knots, etc. Hyperparameter tuning is not pursued in this study.

Parameter Estimation The confidence levels used in the optimization are $0.01, 0.33, 0.66, 0.99$. The weights for the confidence levels are $0.5, 0.2, 0.2, 0.5$ and normalized such that they sum up to 1. We add a penalty term on the squared difference between adjacent coefficients (Eilers and Marx, 1996, 2003) for the spline function f_1 to reduce overfit. The penalty coefficients are fixed at 0.1, which is not optimized by cross-validation. The optimization for parameter estimation is conducted with R package Portfolio Safeguard (Zabarankin et al., 2016).

Benchmarks We choose the Gaussian mixture regression, the generalized additive model (GAM) and NGBoost as the benchmarks. For the Gaussian mixture regression, we use 3 mixture components. The weights of the components are constants in the model. For the generalized additive model, the mean, standard deviation of the normal distribution are functions of the factors. The additive formula of the functions are sum of univariate cubic splines of each factor. Adding additional interaction terms often leads to convergence issue. The univariate cubic splines are penalized on the second derivative of the functions. For the three modules in NGBoost, we use trees with depth 3 as the base learner, normal distribution as the parametric function, and CRPS as the score function. The benchmark models are implemented by R packages Mixtools (Benaglia et al., 2009), GAMLSS (Rigby and Stasinopoulos, 2005) and Python library NGBoost, respectively.

Performance Measure We conduct 10-fold cross-validation and use the out-of-sample CRPS and coverage rate as performance measures. The CRPS is calculated by sum of CRPS of distributional prediction on response variable. The coverage rate is calculated by the percentage of the response variable that falls within the predicted conditional interval of two specified quantiles. A small CRPS and a coverage rate close to the target indicate a high-quality distributional prediction.

Results Results are presented in the main part of the paper.

4.2 Experiment 2

The experiment focuses on the Factor Model of Mixture Quantiles (Section 3) and its parameter estimation by convex optimization (Section 4). The model is used to estimate the conditional distribution of Durable Goods as a function of Nondurable Goods and Services.

Data The response variable is the percentage change from previous year of the Consumer Price Index for All Urban Consumers: Durables in US City Average, while the factors are the percentage change from previous year of the Consumer Price Index for All Urban Consumers: Nondurables in US City Average and the percentage change from previous year of the Consumer Price Index for All Urban Consumers: Services in US City Average. For convenience, we denote them by Durable Goods, Nondurable Goods and Services in subsequent sections. The monthly data, covering the period from January 1957 to January 2023, is described in McCracken and Ng (2016) and downloaded from Zimmermann (Zimmermann). The data are standardized to have zero median and unit interquartile range. The data is randomly shuffled before experiments.

Model

- $Q_{\mathtt{N}}(p) = \sqrt{2}\mathrm{erf}^{-1}(2p-1)$ = quantile function of standard normal distribution
- $Q_{\mathtt{E}_1}(p) = -\ln(4-4p)$ = quantile function of right-side transformed exponential distribution that is nonzero in $(0.75, 1)$
- $Q_{\mathtt{E}_1}(p) = \ln(4p)$ = quantile function of left-side transformed exponential distribution that is nonzero in $(0, 0.25)$

Consider the Two-Factor Model of Mixture Quantiles

$$\begin{aligned} G(p, \boldsymbol{x}, \boldsymbol{a}) =& f_0(x_1, x_2) + Q_{\mathtt{N}}(p) f_1(x_1, x_2) \\ &+ Q_{\mathtt{E}_1}(p) f_2(x_1, x_2) + Q_{\mathtt{E}_2}(p) f_3(x_1, x_2) \,, \end{aligned} \tag{3}$$

For each bivariate B-spline $\{f_i\}_{i=0,1,2,3}$, we use basis of degree three with 10 equidistant knots. The interpretation is that $f_1(x_1, x_2)$ determines the body of the distribution, while $f_2(x_1, x_2)$ and $f_3(x_1, x_2)$ determine the left and right tails, respectively.

Parameter Estimation The confidence levels used in the optimization are $0.01, 0.05, 0.15, 0.25, \cdots, 0.95, 0.99$. The weights for the confidence levels are $20, 10, 1, , : \cdots, 1, 10, 20$ and normalized such that they sum up to 1. We add a penalty term on the squared difference between adjacent coefficients (Eilers and Marx, 1996, 2003) for each spline function $\{f_i\}_{i=0,1,2,3}$ to reduce overfit. The penalty coefficients are fixed at 0.01 and not optimized. The optimization for parameter estimation is conducted with R package Portfolio Safeguard (Zabarankin et al., 2016).

Benchmarks We choose the Gaussian mixture regression and the generalized additive model (GAM) as the benchmarks. For the Gaussian mixture regression, we use BIC as the model selection criteria, which is one of the built-in methods in the package Mixtools. The weights of the components are constants in the model. For the generalized additive model, the mean, standard deviation, shape and skewness of the Skew t type 2 distribution (Azzalini and Capitanio, 2003) are functions of the factors. The additive formula of the functions are sum of univariate cubic splines of each factor with an additional linear interaction term. The univariate cubic splines are penalized on the second derivative of the functions. The benchmark models are implemented by R packages Mixtools (Benaglia et al., 2009) and GAMLSS (Rigby and Stasinopoulos, 2005), respectively.

Performance Measure We conduct 10-fold cross-validation and use the out-of-sample CRPS and coverage rate as performance measures. The CRPS is calculated by sum of CRPS of distributional prediction on response variable. The coverage rate is calculated by the percentage of the response variable that falls within the predicted conditional interval of two specified quantiles. A small CRPS and a coverage rate close to the target indicate a high-quality distributional prediction.

Results Results are presented in Table 1 and Table 2. In general, our approach has comparable performance with the benchmarks. Our approach exhibits improved average out-of-sample coverage rate. The average out-of-sample CRPS of our approach is lower than the Gaussian mixture regression but higher than the generalized additive model. However, the generalized additive model often yield poor extrapolation results when the test data lies outside the domain of training data, resulting in nonexistent second moments and therefore nonexistent CRPS values.

Table 1: Average coverage rates of the prediction intervals in 10-fold cross-validation for three models

Model	0.98	0.9	0.7	0.5	0.3	0.1	Ave. Diff.
Factor model of mixture quantiles	0.977	0.895	0.741	0.508	0.297	0.089	**0.005**
Gaussian mixture regression	0.969	0.907	0.789	0.603	0.372	0.113	0.046
Generalized additive model	0.975	0.903	0.728	0.527	0.315	0.112	0.013

Result of Experiment 1. The coverage rate is calculated by the percentage of the response variable that falls within the predicted conditional interval of two specified quantiles. The average differences between the realized coverage rate and the target coverage rate are calculated. The target coverage rates are obtained by confidence intervals $(0.01, 0.99)$, $(0.05, 0.95)$, $(0.15, 0.85)$,$(0.25, 0.75)$,$(0.35, 0.65)$,$(0.45, 0.55)$.

Table 2: Average out-of-sample Continuous Ranked Probability Score (CRPS) for three models: Factor model of mixtures, Gaussian mixture regression, and generalized additive model, obtained by 10-fold cross-validation procedure. The mean and standard deviation of the CRPS values are reported for each model. The asterisk $*$ denotes instances in which a model produces predictions of nonexistent second moment and therefore nonexistent CRPS values for some data points. The CRPS is calculated only for predictions that have finite CRPS values.

Model	1	2	3	4	5	6
Factor model of mixture quantiles	0.199	0.241	0.236	0.220	0.204	0.219
Gaussian mixture regression	0.209	0.262	0.247	0.238	0.215	0.241
Generalized additive model	0.186	0.213	0.210	0.204	0.193*	0.208*

Model	7	8	9	10	Mean	Std
Factor model of mixture quantiles	0.197	0.255	0.193	0.186	0.215	0.023
Gaussian mixture regression	0.205	0.268	0.214	0.188	0.229	0.027
Generalized additive model	0.187*	0.226	0.176	0.253	0.206	0.023

We can obtain the p-quantile surface of the response variable by varying the values of (x_1, x_2) with a fixed p. We use the model fit in the 10th fold of the cross-validation as an example. Figure 1 shows three selected quantile surfaces for visualization, demonstrating the nonlinear relations incorporated in the model. The surfaces do not to cross each other.

4.3 Experiment 3

Data We randomly select two thirds of the data in Experiment 1.

Model We use I-spline basis of degree three with five knots and exponential distribution as basis quantile functions. for splines $\boldsymbol{B}_k(x_k)$ in weight function, we still use B-spline of degree three with six knots.

Parameter estimation The initial parameter for three models are the same. No penalty is applied to the smoothness of splines in this experiment, since the aim is to compare the convergence speed and loss of precision using varying ranks. The nonnegativity constraint is applied in each step of the algorithm. However, there should be no sign constraint on the coefficients $\boldsymbol{a}_0$ of the spline $f_0(\boldsymbol{x}, \boldsymbol{a}_0)$ that determines the conditional location, as shown in (3). To address this, we add a positive constant to all observations before applying the alternating algorithm. This ensures that the model $G(p, \boldsymbol{x}, \boldsymbol{a})$ is positive, and $\boldsymbol{a}_0$ satisfy the nonnegativity constraint. We stop the iteration at step 100. Other stopping rules can be used as well, such as the number of iterations, the difference between consecutively updated objective values.

Note that the low-rank representation alone does not guarantee improvements in out-of-sample test, since the smoothness condition is unconstrained and may lead to overfit. The combination of the low-rank representation with P-spline is not pursued in this experiment. Due to this reason, we report the out-of-sample error during training only in the appendix for reader's reference. The validation set is the unselected one third of the data from Experiment 1.

We observe in Figure 1 that the out-of-sample errors decrease fairly fast in the first few iterations. Then, models with rank 1 keeps decreasing slowly and stably. The other two models shows wibbling but overall increasing out-of-sample error.

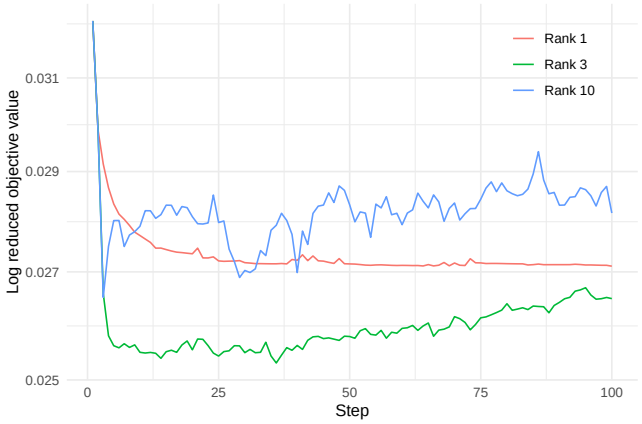


Figure 1: Experiment 2: Semilog plot of validation loss using alternating algorithm on models with rank 1, 3 and 10.