

Distributionally robust conditional quantile prediction with Wasserstein ball

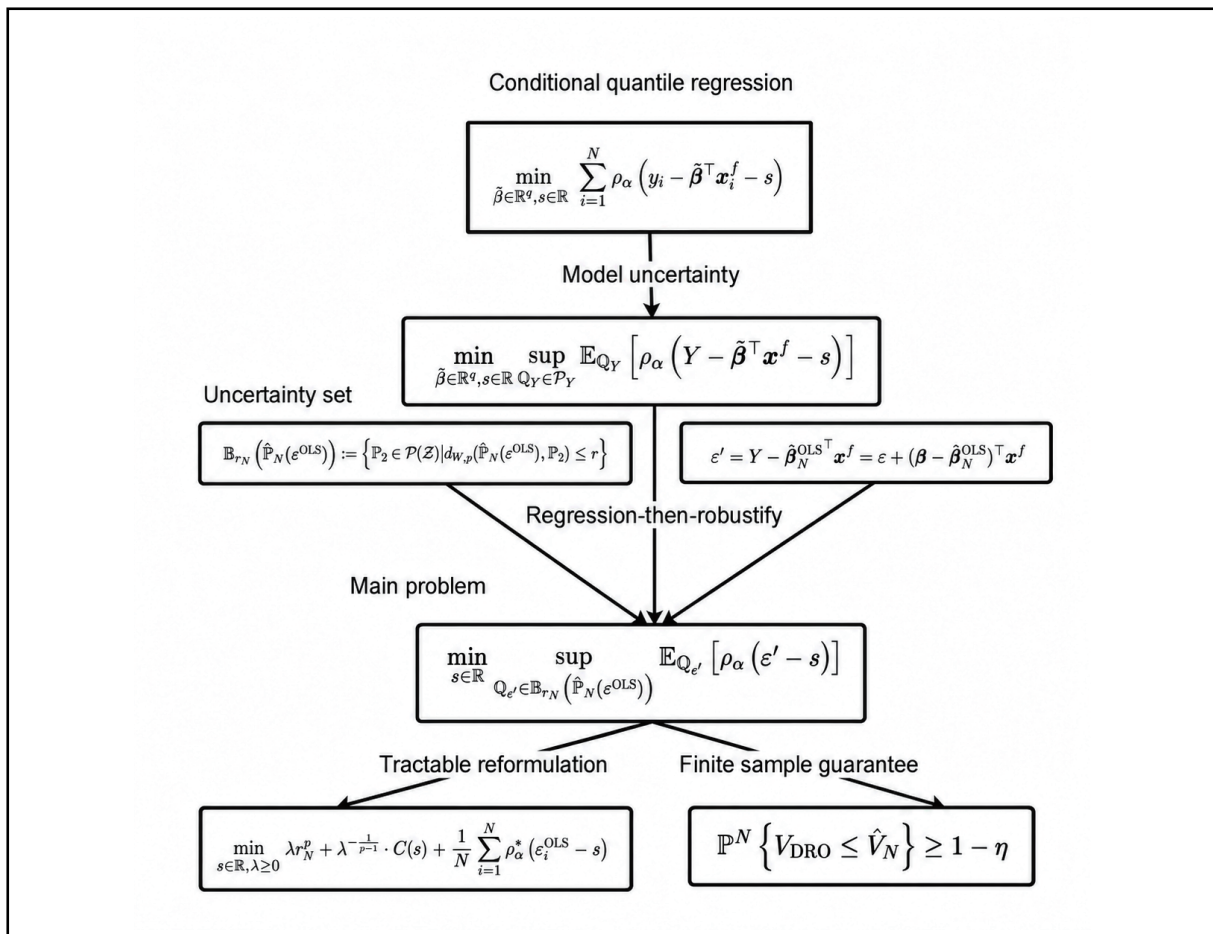
Bo Cheng, and Xinqiao Xie 

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

 Correspondence: Xinqiao Xie, E-mail: xxqyzb@mail.ustc.edu.cn

© 2026 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Graphical abstract



Research framework.

Public summary

- A new data-driven distributionally robust framework for distributionally robust conditional quantile prediction is proposed.
- The general p -order Wasserstein distance is extended to enhance the robustness of the optimization problem.
- A robust data-driven solution with finite-sample performance guarantees is provided, and numerical results are reported.

Distributionally robust conditional quantile prediction with Wasserstein ball

Bo Cheng, and Xinqiao Xie 

Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

✉ Correspondence: Xinqiao Xie, E-mail: xxqzyb@mail.ustc.edu.cn

© 2026 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Cite This: *JUSTC*, 2026, 56(X): (10pp)


Read Online

Abstract: A new data-driven distributionally robust framework was proposed for conditional quantile prediction. We extended the framework to accommodate general p -Wasserstein distances to enhance the robustness of the optimization problem and investigated quantile prediction using observed feature information. We employed a regression-then-robustify approach to obtain a robust data-driven solution to the quantile prediction problem and established finite-sample performance guarantees for this solution. We also reported the numerical performance of the proposed distributionally robust optimization (DRO) method through comparisons with other classical models.

Keywords: distributionally robust optimization; Wasserstein metric; finite-sample guarantee; conditional quantile prediction

CLC number: O221.5

Document code: A

2020 Mathematics Subject Classification: 90C17; 62G08

1 Introduction

Quantile regression has attracted considerable research attention since its introduction by Koenker and Bassett^[1]. An overview of the methodology is provided by Koenker and Hallock^[2], and a comprehensive treatment is given by Koenker^[3]. Quantile prediction has wide applications in various fields including operations research, economics and finance^[4–6]. Huang et al.^[7] provided a comprehensive review of quantile regression for different types of data and application areas.

In classical quantile regression research, the assumed distribution of the data typically follows an empirical distribution, commonly referred to as the sample average approach (SAA) method. However, in practice, determining the true distribution of the data becomes challenging when complete data are not available or only a small sample of data is obtained, making it impossible to use the empirical distribution. Consequently, this limitation can lead to poor out-of-sample performance. One approach to guarantee out-of-sample performance is to incorporate the concept of robust optimization. Motivated by this, distributionally robust optimization (DRO) framework has gained increasing scholarly attention. Scarf et al.^[8] initially introduced a min-max solution to the quantile prediction problem. This approach entails considering all possible demand values and employing the worst-case solution for decision-making. In the DRO framework, the construction of an ambiguity set of possible probability distributions is essentially important and typically based on either distributional information or historical data. The DRO quantile prediction problem has been extensively investigated under partial distribution information, including given

moment information^[9], symmetry and unimodality^[10], and assuming an ellipsoid or box discrete distribution^[11]. By incorporating historical data into the DRO framework, we consider the data-driven DRO (DD-DRO) framework. Various methods have been developed to construct ambiguity sets using historical data. For example, Wang et al.^[12] assumed that the observed data satisfied the lower empirical likelihood, whereas Bertsimas et al.^[13] employed the confidence region as an ambiguity set. Another widely employed method to construct ambiguity sets involves first determining a nominal distribution based on observed samples (e.g., empirical distribution) and subsequently incorporating all probability distributions closely related to the nominal distribution within a certain probability measure. Commonly utilized probability metrics for this purpose include the ϕ -divergence^[12,14,15], Prohorov metric^[16], and Wasserstein distance^[17,18,19,20,21]. The Wasserstein distance has gained increasing popularity because of its effectiveness in characterizing the distance between two distributions, making it the metric of choice in our study.

In many practical applications, the external features might be predetermined or designed, which means that the data generated by the response variable and the external features are not independent and identically distributed (i.i.d.). Examples include predicting goods demand via prior information such as the quantity and features of popular products or making pricing decisions according to certain strategies. For this reason, we adopt a common assumption in the prediction of an estimator known as the “fixed-design” assumption^[22] in this study. According to this assumption, sample points are considered fixed rather than random. Only the source of randomness in training is due to the response sample points. Under

the “fixed-design” assumption, we consider the DD-DRO framework and propose a distributionally robust conditional quantile prediction problem with Wasserstein ball. In this problem, we assume that there are accessible historical observations of the predetermined characteristic $\{\mathbf{x}_i'\}_{i=1}^N$ and the corresponding response variable $\{y_i\}_{i=1}^N$, which are independent but not the same distribution. Owing to the lack of i.i.d. samples, our framework is similar to that proposed by Qi et al.^[23] Based on this literature, we extend to the general p -Wasserstein distance to enhance the robustness of the optimization problem, where $p > 1$. The general p -Wasserstein distance is a more flexible metric than the 1-Wasserstein distance, allowing for a wider range of probability distributions to be compared. This flexibility can lead to improved predictions in real-world applications. Under the aforementioned assumption, we achieve an effective and practical reformulation of the quantile prediction problem. Notably, this reformulation is polynomially solvable, involving linear regression followed by the residual variables. Additionally, we provide finite-sample performance guarantees and numerical performance assessments of our proposed DRO method.

The remainder of the paper is organized as follows. Section 2 introduces the definition and properties of Wasserstein distance and ordinary least squares (OLS) estimation, then presents our distributionally robust conditional quantile prediction problem with a Wasserstein ball as a minimax problem. Section 3 addresses this problem and presents a tractable reformulation. Section 4 develops the finite-sample guarantee. Section 5 presents numerical examples to validate the numerical performance of our DRO method. Finally, Section 6 concludes the paper with a brief summary.

2 Problem formulation

2.1 Wasserstein distance

We begin by defining the Wasserstein distance as introduced in Ref. [24]. The Wasserstein metric is defined as a distance function between two probability distributions on a given probability measure space.

Definition 1. (Wasserstein distance) Let $\mathcal{Z}$ denote a metric space with metric $d(\cdot, \cdot)$, $\mathcal{P}(\mathcal{Z})$ denote the set of all Borel probability measures on $\mathcal{Z}$, $\mathbb{P}_1$ and $\mathbb{P}_2$ denote two distributions supported on a set $\mathcal{P}(\mathcal{Z})$, and $\mathcal{M}(\mathbb{P}_1, \mathbb{P}_2)$ is defined as the set of all Borel probability distributions on $\mathcal{Z} \times \mathcal{Z}$ with marginal distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. For any $p \in [1, \infty)$, the p -Wasserstein distance between $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as:

$$d_{w,p}(\mathbb{P}_1, \mathbb{P}_2) := \inf_{\Pi \in \mathcal{M}(\mathbb{P}_1, \mathbb{P}_2)} \left(\int d(\xi, \xi')^p \Pi(d\xi, d\xi') \right)^{\frac{1}{p}}.$$

According to Ref. [25], one can verify that the Wasserstein distance is a metric, that is, it is nonnegative, symmetric, and subadditive, and it vanishes if and only if $\mathbb{P}_1 = \mathbb{P}_2$. Additionally, as noted in Ref. [25], $d_{w,p}(\mathbb{P}_1, \mathbb{P}_2)$ is finite whenever both $\mathbb{P}_1$ and $\mathbb{P}_2$ have finite p -order moments. In this paper, we focus on the p -Wasserstein distance as defined in Definition 1 for $p \in [1, \infty)$. Our analysis concentrates on the case in which $\mathcal{Z} = \mathbb{R}$; thus, the absolute distance $|\xi - \xi'|$ is chosen as $d(\xi, \xi')$. In addition, we introduce the definition of the dual

expression of the Wasserstein metric from Theorem 5.10 of Ref. [24], which is employed in the proof of the finite-sample guarantee.

Lemma 1. For any distributions $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathcal{Z})$, we have

$$d_{w,p}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f, g \in \mathcal{L}} \left(\int_{\mathcal{Z}} f(\xi) \mathbb{P}_1(d\xi) - \int_{\mathcal{Z}} g(\xi) \mathbb{P}_2(d\xi) \right)^{\frac{1}{p}},$$

where $\mathcal{L}$ denotes the set of all the bounded continuous functions, which satisfy $|f(\xi) - g(\xi')| \leq d(\xi, \xi')^p$ for all $\xi, \xi' \in \mathcal{Z}$, and $d(\xi, \xi')$ denotes the metric on $\mathcal{Z}$ that is used in Definition 1.

The Wasserstein distance is a natural way to compare two distributions when one is obtained from the other by perturbation. Here, we introduce the Wasserstein ball centered on a nominal distribution, which is used to characterize the distributional ambiguity in our problem.

Definition 2. For a fixed distribution $\mathbb{P}_1 \in \mathcal{P}(\mathcal{Z})$ and $r > 0$, the Wasserstein ball centered on $\mathbb{P}_1$ can be formulated as follows:

$$\mathbb{B}_r(\mathbb{P}_1) := \left\{ \mathbb{P}_2 \in \mathcal{P}(\mathcal{Z}) \mid d_{w,p}(\mathbb{P}_1, \mathbb{P}_2) \leq r \right\}.$$

2.2 Ordinary least square (OLS) estimation

Let $Y \in \mathbb{R}$ be the response variable of interest and $(X_1, \dots, X_q) \in \mathbb{R}^q$ be a vector of predictive variables. We assume that $X_1, \dots, X_q$ and Y satisfy the standard linear regression model without an interception term, that is,

$$Y = \beta_1 X_1 + \dots + \beta_q X_q + \varepsilon, \quad (1)$$

where $\beta = (\beta_1, \dots, \beta_q)^\top \in \mathbb{R}^q$ is the vector of unknown regression parameters, and $\varepsilon \in \mathbb{R}$ is a random variable of residuals. In addition, we assume that ε is independent of $(X_1, \dots, X_q)$ with zero mean and unknown variance σ^2 . The distribution of ε is not necessarily assumed to be normal.

Let $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$ be N samples having been observed from (1). Denote $(\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$ by $\mathbf{X}$ and $(y_1, \dots, y_N)^\top$ by $\mathbf{y}$. Under certain assumptions, the OLS estimator of β from the N samples, denoted by $\hat{\beta}_N^{\text{OLS}}$, is expressed as

$$\hat{\beta}_N^{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Since $\hat{\beta}_N^{\text{OLS}}$ is an essential tool in our conditional quantile regression with model uncertainty (introduced in Section 2.3), we below provide the important assumptions that ensure performance of $\hat{\beta}_N^{\text{OLS}}$ as well as some properties of the OLS estimator. Throughout this paper, we also make the same assumptions.

Assumption 1. The design matrix $\mathbf{X}$ is of full rank.

Assumption 1 ensures that the matrix $\mathbf{X}^\top \mathbf{X}$ is an invertible matrix. Subsequently, we find that the projection matrix $\mathbf{H} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is an idempotent matrix. Let $\mathbf{H} = (h_{ij})_{N \times N}$. Thus, we have $\text{trace}(\mathbf{H}) = \sum_{i=1}^N h_{ii} = q$.

Assumption 2. As sample size N goes to infinity, the $d \times d$ matrix

$$\mathbf{M} = \lim_{N \rightarrow \infty} N^{-1} \mathbf{X}^\top \mathbf{X} = \lim_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top$$

exists and is finite nonsingular.

This leads to

$$\lim_{N \rightarrow \infty} \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a} = 0$$

for any bounded vector $\mathbf{a}$; see Ref. [23] for more detailed discussions.

By applying the strong law of large numbers, we have the following lemma directly.

Lemma 2. Under Assumptions 1 and 2, the OLS estimator is strongly consistent, that is,

$$\hat{\boldsymbol{\beta}}_N^{\text{OLS}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}.$$

2.3 Conditional quantile regression with model uncertainty

In this section, we continue to consider the standard linear regression model (1). For $\alpha \in (0, 1)$, the α -quantile $Q_\alpha(Y)$ of a random variable Y is defined as

$$Q_\alpha(Y) = \inf\{y : \mathbb{P}(Y \leq y) \geq \alpha\}. \quad (2)$$

In addition to Eq. (2), $Q_\alpha(Y)$ can also be represented as the solution of a stochastic optimization problem, that is,

$$Q_\alpha(Y) = \arg \min_{v \in \mathbb{R}} \mathbb{E}[\rho_\alpha(Y - v)], \quad (3)$$

where ρ_α is the quantile loss function (also called the check function), which takes the form of

$$\rho_\alpha(u) = u(\alpha - 1_{\{u \leq 0\}}) = \begin{cases} \alpha u, & \text{if } u > 0; \\ (\alpha - 1)u, & \text{if } u \leq 0. \end{cases}$$

Recalling the “fixed-design” assumption and considering the linear regression model (1) along with the definition of the α -quantile, for $(X_1, \dots, X_q) = (x_1, \dots, x_q)^\top := \mathbf{x}^f$, the α th conditional quantile $Q_\alpha(Y|\mathbf{x}^f)$ of a random variable Y takes the following form:

$$Q_\alpha(Y|\mathbf{x}^f) = Q_\alpha(\boldsymbol{\beta}^\top \mathbf{x}^f + \varepsilon|\mathbf{x}^f) = \boldsymbol{\beta}^\top \mathbf{x}^f + Q_\alpha(\varepsilon) = \boldsymbol{\beta}^\top \mathbf{x}^f + s_\alpha, \quad (4)$$

where $s_\alpha := Q_\alpha(\varepsilon)$ is the α th quantile of ε . Let $\{(\mathbf{x}_1^f)^\top, \dots, (\mathbf{x}_N^f)^\top\}$ denotes a sequence of (row) vectors of a design matrix $\mathbf{X}$. Suppose that $\{y_1, \dots, y_N\}$ is an i.i.d. sample having been observed from Eq. (1). The quantile regression, introduced in Ref. [1], estimate a given quantile of Y conditional on $\mathbf{x}^f$. Specifically, for $\alpha \in (0, 1)$, the α th regression quantile is defined as the solution of the following problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^q, s \in \mathbb{R}} \sum_{i=1}^N \rho_\alpha(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i^f - s).$$

The aforementioned quantile regression can also be seen as follows: The conditional stochastic optimization problem from (3) applies the SAA approach.

The SAA approach has commendable performance under the “random-design” assumption because the empirical distribution converges to the real underlying data generating distribution as the sample size increases. However, it becomes inefficient under the “fixed-design” assumption, as this convergence is no longer guaranteed. This motivates us to find a more effective approach to solve the quantile regression problem. Additionally, in practice the distribution obeyed by the data cannot be determined if the complete data are not avail-

able or if a small sample of data is obtained and an empirical distribution function cannot be used. Motivated by these findings, we consider including distributional ambiguity in the quantile regression problem. See Qi et al. [23] for a more detailed discussion of the motivation. As a result, we follow the conditional quantile regression with model uncertainty, proposed by Qi et al. [23], to incorporate the ambiguity of distribution and utilize historical data. To incorporate the ambiguity of distribution and utilize historical data, it is formulated as follows:

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^q, s \in \mathbb{R}} \sup_{Q_Y \in \mathcal{P}_Y} \mathbb{E}_{Q_Y} [\rho_\alpha(Y - \tilde{\boldsymbol{\beta}}^\top \mathbf{x}^f - s)], \quad (5)$$

where Q_Y is a distribution of Y conditional on $(X_1, \dots, X_q) = (x_1, \dots, x_q)^\top := \mathbf{x}^f$ and $\mathcal{P}_Y$ is an ambiguity set constructed by observed data points $\{(\mathbf{x}_i^f, y_i)\}_{i=1}^N$.

2.4 Uncertainty set construction

The uncertainty set in our work is constructed using the p -Wasserstein distance and the OLS estimator. We follow the regress-then-robustify method introduced by Qi et al. [23] to simplify (5). Let $\tilde{\varepsilon} = Y - \tilde{\boldsymbol{\beta}}^\top \mathbf{x}^f = \varepsilon + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^\top \mathbf{x}^f$, conditional quantile regression with model uncertainty problem (5) can be rewritten as

$$\min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^q, s \in \mathbb{R}} \sup_{Q_{\tilde{\varepsilon}} \in \mathcal{P}_{\tilde{\varepsilon}}} \mathbb{E}_{Q_{\tilde{\varepsilon}}} [\rho_\alpha(\tilde{\varepsilon} - s)], \quad (6)$$

where $Q_{\tilde{\varepsilon}}$ and $\mathcal{P}_{\tilde{\varepsilon}}$ denote the candidate distribution and ambiguity set of $\tilde{\varepsilon}$, respectively. Note that $\mathcal{P}_{\tilde{\varepsilon}}$ also depends on $\tilde{\boldsymbol{\beta}}$. We consider the empirical distribution as the nominal distribution to estimate the true probability distribution. Denote $\tilde{\varepsilon}_i = y_i - \tilde{\boldsymbol{\beta}}^\top \mathbf{x}_i^f$, then, the uncertainty set is constructed as a Wasserstein ball centered at the empirical distribution (i.e., $\mathcal{P}_{\tilde{\varepsilon}} := \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\tilde{\varepsilon}))$ with $\hat{\mathbb{P}}_N(\tilde{\varepsilon}) = \frac{1}{N} \sum_{i=1}^N \delta_{\tilde{\varepsilon}_i}$, where the definition of the Wasserstein ball is provided in Section 2.1 and r_N is determined by Theorem 3.4 of Ref. [18]). Subsequently, we provide a good estimator of $\boldsymbol{\beta}$ to replace $\tilde{\boldsymbol{\beta}}$ and have an approximated distributionally robust problem of (6). Since the OLS estimator $\hat{\boldsymbol{\beta}}_N^{\text{OLS}}$ of $\boldsymbol{\beta}$ has a good closed-form and essential asymptotic property under certain assumptions, we consider replacing $\tilde{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_N^{\text{OLS}}$. More specifically, defining $\varepsilon' = \varepsilon + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_N^{\text{OLS}})^\top \mathbf{x}^f$, we obtain the corresponding samples $\{\varepsilon_i^{\text{OLS}}\}_{i=1}^N$ with $\varepsilon_i^{\text{OLS}} = y_i - (\hat{\boldsymbol{\beta}}_N^{\text{OLS}})^\top \mathbf{x}_i^f$ and the corresponding empirical distribution $\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}) = \frac{1}{N} \sum_{i=1}^N \delta_{\varepsilon_i^{\text{OLS}}}$. The empirical distribution leads to a Wasserstein ball $\mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}))$. It is important to note that the ambiguity set does not center on the empirical distribution of ε' but rather on the empirical distribution constructed by OLS residuals, which is different from former researches. With this clarification, we now formulate our distributionally robust conditional quantile prediction problem under the Wasserstein ball:

① The superscript “ f ” means that we adopt a “fixed-design” setting in our model.

② In the “fixed-design” assumption, although $\{(\mathbf{x}_1^f)^\top, \dots, (\mathbf{x}_N^f)^\top\}$ cannot be seen as i.i.d. samples, we still assume that the design matrix $\mathbf{X}^f$ satisfies Assumption 2 and then has the asymptotic property.

$$\min_{s \in \mathbb{R}} \sup_{Q_{\varepsilon'} \in \mathcal{B}_{r_N}(\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}))} \mathbb{E}_{Q_{\varepsilon'}} [\rho_{\alpha}(\varepsilon' - s)], \quad (7)$$

where $Q_{\varepsilon'}$ and $\mathcal{B}_{r_N}(\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}))$ denote the candidate distribution and ambiguity set of ε' , respectively. In the next section, we derive a tractable reformulation of problem (7).

3 Tractable reformulation

In this section, our goal is to solve the distributionally robust conditional quantile prediction problem (7). As the notation in Section 2.1, we consider the metric space $\mathcal{Z}$ to be $\mathbb{R}$. We prove that it can be rewritten as a convex program with decoupled components and can thus be solved by OLS regression. Theorem 1 provides a tractable reformulation.

Theorem 1. For any $r_N \geq 0$, $p > 1$, Problem (7) is equivalent to the following problem:

$$\min_{s \in \mathbb{R}, \lambda \geq 0} \lambda r_N^p + \lambda^{-\frac{1}{p-1}} \cdot C(s) + \frac{1}{N} \sum_{i=1}^N \rho_{\alpha}^*(\varepsilon_i^{\text{OLS}} - s), \quad (8)$$

where

$$C(s) = (p-1) \left(\frac{\alpha}{p} \right)^{\frac{p}{p-1}} I_1 + (p-1) \left(\frac{1-\alpha}{p} \right)^{\frac{p}{p-1}} I_2,$$

$$I_1 = \frac{1}{N} \sum_{i=1}^N 1_{\{\varepsilon_i^{\text{OLS}} \geq s+a\}}, \quad I_2 = 1 - I_1,$$

$$a = (\lambda p)^{-\frac{1}{p-1}} \left(1 - \frac{1}{p} \right) \left[(1-\alpha)^{\frac{p}{p-1}} - \alpha^{\frac{p}{p-1}} \right],$$

and

$$\rho_{\alpha}^*(u) = u(\alpha - 1_{\{u \leq a\}}) = \begin{cases} \alpha u & \text{if } u > a; \\ (\alpha - 1)u & \text{if } u \leq a. \end{cases}$$

Proof. We first focus on tackling the inner problem of (7):

$$\sup_{Q_{\varepsilon'} \in \mathcal{B}_{r_N}(\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}))} \mathbb{E}_{Q_{\varepsilon'}} [\rho_{\alpha}(\varepsilon' - s)]. \quad (9)$$

For a fixed s , Problem (9) is equivalent to

$$\sup_{Q_{\varepsilon'} \in \mathcal{P}(\mathbb{R})} \mathbb{E}_{Q_{\varepsilon'}} [\rho_{\alpha}(\varepsilon' - s)] \quad \text{s.t.} \quad d_{W,p}(Q_{\varepsilon'}, \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) \leq r_N. \quad (10)$$

By rewriting the objective as an integral with respect to the joint distribution Π and using the definition of the p -Wasserstein distance, we obtain the following equivalent formulation:

$$\begin{aligned} & \sup_{\Pi \in \mathcal{P}(\mathbb{R} \times \mathbb{R})} \int_{\mathbb{R}} \rho_{\alpha}(\varepsilon' - s) Q_{\varepsilon'}(d\varepsilon') \\ & \text{s.t.} \quad \int_{\mathbb{R} \times \mathbb{R}} |\varepsilon' - \xi|^p \Pi(d\varepsilon', d\xi) \leq r_N^p, \\ & \text{where } \varepsilon' \sim Q_{\varepsilon'}, \xi \sim \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}), \\ & \text{and } \Pi(\cdot \times \mathbb{R}) = Q_{\varepsilon'}, \Pi(\mathbb{R} \times \cdot) = \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}). \end{aligned} \quad (11)$$

For $i = 1, \dots, N$, given that $\xi = \varepsilon_i^{\text{OLS}}$, denote the distribution of ε' conditional on $\xi = \varepsilon_i^{\text{OLS}}$ by Q^i . One can verify that for any (ε', ξ) 's joint distribution Π , there exists a series of distributions $\{Q^i\}_{i=1}^N$ such that $\Pi = \frac{1}{N} \sum_{i=1}^N \delta_{\varepsilon_i^{\text{OLS}}} \otimes Q^i$. As a result, (10) is equivalent to

$$\begin{aligned} & \sup_{Q^i \in \mathcal{P}(\mathbb{R}), i=1, \dots, N} \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} \rho_{\alpha}(\varepsilon' - s) Q^i(d\varepsilon') \\ & \text{s.t.} \quad \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} |\varepsilon' - \varepsilon_i^{\text{OLS}}|^p Q^i(d\varepsilon') \leq r_N^p. \end{aligned} \quad (12)$$

By applying Lagrangian multiplier method, Problem (12) can be rewritten as

$$\begin{aligned} & \sup_{Q^i \in \mathcal{P}(\mathbb{R}), i=1, \dots, N} \inf_{\lambda \geq 0} \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} \rho_{\alpha}(\varepsilon' - s) Q^i(d\varepsilon') + \\ & \lambda \left(r_N^p - \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} |\varepsilon' - \varepsilon_i^{\text{OLS}}|^p Q^i(d\varepsilon') \right). \end{aligned} \quad (13)$$

Since strong duality holds for Problem (13), Problem (13) is equivalent to

$$\begin{aligned} & \inf_{\lambda \geq 0} \sup_{Q^i \in \mathcal{P}(\mathbb{R}), i=1, \dots, N} \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} \rho_{\alpha}(\varepsilon' - s) Q^i(d\varepsilon') + \\ & \lambda \left(r_N^p - \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} |\varepsilon' - \varepsilon_i^{\text{OLS}}|^p Q^i(d\varepsilon') \right). \end{aligned} \quad (14)$$

Then we turn to solve the inner problem of (14). One can verify that the optimal $\{Q^i\}_{i=1}^N$, that solves

$$\sup_{Q^i \in \mathcal{P}(\mathbb{R}), i=1, \dots, N} \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} (\rho_{\alpha}(\varepsilon' - s) - \lambda |\varepsilon' - \varepsilon_i^{\text{OLS}}|^p) Q^i(d\varepsilon'), \quad (15)$$

is a series of the Dirac distributions $\{\delta_{\varepsilon_i^*}\}_{i=1}^N$, where $\varepsilon_i^* = 1, \dots, N$, solves

$$\sup_{\varepsilon' \in \mathbb{R}} \rho_{\alpha}(\varepsilon' - s) - \lambda |\varepsilon' - \varepsilon_i^{\text{OLS}}|^p, \quad \text{for } i = 1, \dots, N. \quad (16)$$

If $\varepsilon_i^{\text{OLS}} \geq s+a$, we can solve (16) with $\varepsilon_i^* = \varepsilon_{i1} := \varepsilon_i^{\text{OLS}} + \left(\frac{\alpha}{\lambda p} \right)^{\frac{1}{p-1}}$. If $\varepsilon_i^{\text{OLS}} < s+a$, the solution of (16) is $\varepsilon_i^* = \varepsilon_{i2} := \varepsilon_i^{\text{OLS}} - \left(\frac{1-\alpha}{\lambda p} \right)^{\frac{1}{p-1}}$. As a result, Problem (14) can be reduced to

$$\begin{aligned} & \inf_{\lambda \geq 0} \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} \rho_{\alpha}(\varepsilon' - s) Q^i(d\varepsilon') + \\ & \lambda \left(r_N^p - \frac{1}{N} \sum_{i=1}^N \int_{\mathbb{R}} |\varepsilon' - \varepsilon_i^{\text{OLS}}|^p Q^i(d\varepsilon') \right) = \\ & \inf_{\lambda \geq 0} \lambda r_N^p + \frac{1}{N} \sum_{i=1}^N \left[\rho_{\alpha}(\varepsilon_{i1} - s) - \lambda \left(\frac{\alpha}{\lambda p} \right)^{\frac{p}{p-1}} I_{\{\varepsilon_i^{\text{OLS}} \geq s+a\}} + \right. \\ & \left. \frac{1}{N} \sum_{i=1}^N \left[\rho_{\alpha}(\varepsilon_{i2} - s) - \lambda \left(\frac{1-\alpha}{\lambda p} \right)^{\frac{p}{p-1}} I_{\{\varepsilon_i^{\text{OLS}} < s+a\}} \right] = \right. \\ & \left. \inf_{\lambda \geq 0} \lambda r_N^p + \lambda^{-\frac{1}{p-1}} \left[(p-1) \left(\frac{\alpha}{p} \right)^{\frac{p}{p-1}} I_1 + (p-1) \left(\frac{1-\alpha}{p} \right)^{\frac{p}{p-1}} I_2 \right] + \right. \\ & \left. \frac{1}{N} \sum_{i=1}^N \rho_{\alpha}^*(\varepsilon_i^{\text{OLS}} - s) \right]. \end{aligned} \quad (17)$$

The desired result follows.

Theorem 1 provides a tractable reformulation of Problem (7), which can be solved numerically. When $\alpha = 0.5$, Theorem 1 can further lead to a general closed-form solution where

$$\hat{\lambda}_N = \left(\frac{C(s)}{r_N(p-1)} \right)^{\frac{p-1}{p}},$$

$$\hat{s}_N = \arg \min_s r_N^{\frac{1}{p}} C(s)^{\frac{p-1}{p}} p(p-1)^{\frac{1-p}{p}} + \frac{1}{N} \sum_{i=1}^N \rho_\alpha(\varepsilon_i^{\text{OLS}} - s). \quad (18)$$

Remark 1. When we consider general p' -quantile regression, its expression is as follows:

$$\rho_\alpha(u) = \begin{cases} \alpha u^{p'}, & \text{if } u > 0; \\ (1-\alpha)(-u)^{p'}, & \text{if } u \leq 0. \end{cases}$$

Problem (16) has no feasible solution when $p < p'$; when $p = p'$, A similar result can be obtained by referring to the above proof. If $\varepsilon_i^{\text{OLS}} \geq s$, we can solve Problem (16) with $\varepsilon_i^* = \varepsilon_{i1} := \frac{\varepsilon_i^{\text{OLS}} - a_1 s}{1 - a_1}$, where $a_1 = \left(\frac{\alpha}{\lambda}\right)^{\frac{1}{p-1}}$; if $\varepsilon_i^{\text{OLS}} < s$, the solution of problem (16) is $\varepsilon_i^* = \varepsilon_{i2} := \frac{\varepsilon_i^{\text{OLS}} - a_2 s}{1 - a_2}$, where $a_2 = \left(\frac{1-\alpha}{\lambda}\right)^{\frac{1}{p-1}}$. Then, It can be easily shown that Problem (14) can be reduced to

$$\min_{s \in \mathbb{R}, \lambda \geq \max\{\alpha, 1-\alpha\}} \lambda r_N^p - \lambda |\varepsilon_i^{\text{OLS}} - s|^p \left[\left(\frac{a_1}{1-a_1} \right)^p I_1 + \left(\frac{a_2}{1-a_2} \right)^p I_2 \right] + \frac{1}{N} \sum_{i=1}^N \rho_\alpha^{**}(\varepsilon_i^{\text{OLS}} - s), \quad (19)$$

where $I_1 = \frac{1}{N} \sum_{i=1}^N I_{\{\varepsilon_i^{\text{OLS}} \geq s\}}$, $I_2 = 1 - I_1$, and the function ρ_α^{**} is

$$\rho_\alpha^{**}(u) = \begin{cases} \alpha \left(\frac{u}{1-a_1} \right)^p & \text{if } u > 0; \\ (1-\alpha) \left(\frac{-u}{1-a_2} \right)^p & \text{if } u \leq 0. \end{cases}$$

Our paper has provided a tractable solution for the distributionally robust conditional quantile optimization problem. Compared with Theorem 1 of Qi et al.^[23], Problem (7), which is constructed under the general p -Wasserstein metric, is more robust, because its solution is related to the Wasserstein ball radius r_N , and thus related to the constructed ambiguity set.

4 Finite-sample guarantee

In the distributionally robust conditional quantile optimization problem (7), we construct the ambiguity set by the general p -Wasserstein metric and the empirical distribution. This ambiguity set offers an attractive performance guarantee under the common light tail assumption of generating a distribution over unknown data. In this section, we present three lemmas and use them to derive a concentration inequality for the p -Wasserstein distance. The inequality provide the finite sample guarantee based on p -Wasserstein ball. To clarify our analysis, we first introduce the following notations:

V^* : the optimal expected cost that we target, that is,

$$V^* = \min_{\beta \in \mathbb{R}^d, s \in \mathbb{R}} \mathbb{E}_{Y|X^f} [\rho_\alpha(Y - \tilde{\beta}^\top X^f - s)] = \min_{\beta \in \mathbb{R}^d, s \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_\varepsilon} [\rho_\alpha(\tilde{\varepsilon} - s)],$$

where $\tilde{\varepsilon} = Y - \tilde{\beta}^\top X^f$ and $\mathbb{P}_\varepsilon$ is the true probability distribution of $\tilde{\varepsilon}$.

$\hat{V}_N$: the in-sample cost achieved by the distributionally robust optimization problem (7), that is,

$$\hat{V}_N = \min_{s \in \mathbb{R}} \sup_{Q_{\varepsilon'} \in \mathcal{P}(\varepsilon')} \mathbb{E}_{Q_{\varepsilon'}} [\rho_\alpha(\varepsilon' - s)],$$

where $\tilde{\varepsilon} = \varepsilon + (\beta - \hat{\beta}_N^{\text{OLS}})^\top X^f$.

V_{DRO} : the expected out-of-sample cost achieved by the solution of our DRO method, $(\hat{s}_N, \hat{\beta}_N^{\text{OLS}})$, that is,

$$V_{\text{DRO}} = \mathbb{E}_{\mathbb{P}_{\varepsilon'}} [\rho_\alpha(\varepsilon' - \hat{s}_N)],$$

where $\mathbb{P}_{\varepsilon'}$ is the true probability distribution of ε' .

Assumption 3. (Light-tailed assumption) There exists an exponent $a > p$ such that

$$A := \mathbb{E}_{\mathbb{P}_\varepsilon} [\exp(\|\varepsilon\|^a)] < \infty.$$

We also assume that $A > 1$.

Assumption 3 essentially requires that the tails of the probability distribution decay at an exponential rate. According to Lemma 4 of Qi et al.^[23], we can assume that there exists a constant κ such that $\sigma^2 \leq \kappa$, where σ^2 denotes a variance of ε . We first give some lemmas before proving the finite-sample guarantee of our results. For the observed data points $\{(\mathbf{x}_i^f, y_i)\}_{i=1}^N$, denoting the true probability distribution and the empirical distribution of ε by $\mathbb{P}_\varepsilon$ and $\hat{\mathbb{P}}_N(\varepsilon)$, respectively, we have the following lemma from Ref. [26, Theorem 2].

Lemma 3. If Assumption 3 holds, then for any $\eta_1 \in (0, 1)$, $N \geq 1$, there exists r_{N1} such that:

$$\mathbb{P}_\varepsilon^N \left\{ d_{W,p}(\mathbb{P}_\varepsilon, \hat{\mathbb{P}}_N(\varepsilon)) \geq r_{N1} \right\} \leq \eta_1,$$

where $\hat{\mathbb{P}}_N(\varepsilon) = \frac{1}{N} \sum_{i=1}^N \delta_{\varepsilon_i}$ with $\varepsilon_i = y_i - \beta^\top \mathbf{x}_i^f$,

$$r_{N1}(\eta_1) := \begin{cases} \left(\frac{\log(c_1 \eta_1^{-1})}{c_2 N} \right)^{\frac{1}{m_1}} & \text{if } N \geq \frac{\log(c_1 \eta_1^{-1})}{c_2}; \\ \left(\frac{\log(c_1 \eta_1^{-1})}{c_2 N} \right)^{\frac{p}{a}} & \text{if } N < \frac{\log(c_1 \eta_1^{-1})}{c_2}. \end{cases} \quad (20)$$

Here, c_1 and c_2 are positive constants depending only on a , A , and m ; $m_1 = \max\{2, m/p\}$, and m is the dimension of the set $\mathcal{P}(\mathcal{Z})$.

For the observed data points $\{(\mathbf{x}_i^f, y_i)\}_{i=1}^N$, denoting the OLS estimated empirical distribution by $\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})$, we have the following result of the Wasserstein distance between $\hat{\mathbb{P}}_N(\varepsilon)$ and $\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})$.

Lemma 4. If Assumptions 1, 2, and 3 hold, then for any $\eta_2 \in (0, 1)$, and $N \geq 1$, we have

$$r_{N2} = \begin{cases} \frac{N^{\frac{1}{2} - \frac{1}{p}} \sqrt{\kappa q}}{\eta_2} & \text{if } 1 < p < 2; \\ \frac{N^{-\frac{1}{p}} \sqrt{\kappa q}}{\eta_2} & \text{if } p \geq 2; \end{cases}$$

such that

$$\mathbb{P}_\varepsilon^N \{d_{w,p}(\hat{\mathbb{P}}_N(\varepsilon), \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) \geq r_{N2}\} \leq \eta_2,$$

where $\hat{\mathbb{P}}_N(\varepsilon) = \frac{1}{N} \sum_{i=1}^N \delta_{\varepsilon_i}$ with $\varepsilon_i = y_i - \beta^\top \mathbf{x}_i^f$ and $\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}) = \frac{1}{N} \sum_{i=1}^N \delta_{\varepsilon_i^{\text{OLS}}}$ with $\varepsilon_i^{\text{OLS}} = y_i - (\hat{\beta}_N^{\text{OLS}})^\top \mathbf{x}_i^f$.

Proof. Applying the dual representation of the Wasserstein distance from Lemma 1, we have

$$\begin{aligned} (d_{w,p}(\hat{\mathbb{P}}_N(\varepsilon), \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})))^p &= \\ \sup_{f, g \in \mathcal{L}} \left\{ \frac{1}{N} \sum_{i=1}^N f(\varepsilon_i) - \frac{1}{N} \sum_{i=1}^N g(\varepsilon_i^{\text{OLS}}) \right\} &\leq \\ \frac{1}{N} \sum_{i=1}^N |\varepsilon_i - \varepsilon_i^{\text{OLS}}|^p &= \frac{1}{N} \sum_{i=1}^N |(\hat{\beta}_N^{\text{OLS}} - \beta)^\top \mathbf{x}_i^f|^p = \\ \frac{1}{N} \sum_{i=1}^N |(\mathbf{x}_i^f)^\top (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}|^p, \end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^\top$ denotes the real residuals. For simplicity, we define the random variables as

$$Z_i = (\mathbf{x}_i^f)^\top (X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}, \quad i = 1, \dots, N.$$

One can verify that $\mathbb{E}[Z_i] = 0$ and

$$\mathbb{E}[Z_i^2] = \text{Var}(Z_i) = \sigma^2 (\mathbf{x}_i^f)^\top (X^\top X)^{-1} \mathbf{x}_i^f \leq \kappa h_{ii},$$

where κ is the bound of the variance of ε , and h_{ii} is the diagonal element of the projection matrix $\mathbf{H} := X(X^\top X)^{-1} X^\top$, $i = 1, \dots, N$, and $\sum_{i=1}^N h_{ii} = q$. Thus, we have

$$\mathbb{P}_\varepsilon^N \{d_{w,p}(\hat{\mathbb{P}}_N(\varepsilon), \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) \geq r_{N2}\} \leq \mathbb{P} \left\{ \left(\frac{1}{N} \sum_{i=1}^N |Z_i|^p \right)^{\frac{1}{p}} \geq r_{N2} \right\}.$$

Using Markov's inequality, we have

$$\mathbb{P} \left\{ \left(\frac{1}{N} \sum_{i=1}^N |Z_i|^p \right)^{\frac{1}{p}} \geq r_{N2} \right\} \leq \frac{\mathbb{E} \left[\left(\frac{1}{N} \sum_{i=1}^N |Z_i|^p \right)^{\frac{1}{p}} \right]}{r_{N2}}.$$

Subsequently, we consider the upper bound of $\mathbb{P} \left(\left(\frac{1}{N} \sum_{i=1}^N |Z_i|^p \right)^{\frac{1}{p}} \geq r_{N2} \right)$ in the following two cases.

(i) For $1 < p < 2$, applying Karamata's inequality, Jensen's inequality, and Hölder's inequality, respectively, we have

$$\begin{aligned} \frac{\mathbb{E} \left[\left(\frac{1}{N} \sum_{i=1}^N |Z_i|^p \right)^{\frac{1}{p}} \right]}{r_{N2}} &\leq \frac{N^{-\frac{1}{p}} \sum_{i=1}^N \mathbb{E}[|Z_i|]}{r_{N2}} \leq \\ \frac{N^{-\frac{1}{p}} \sum_{i=1}^N \sqrt{\text{Var}(Z_i)}}{r_{N2}} &\leq \frac{N^{-\frac{1}{p}} \sum_{i=1}^N \sqrt{\kappa h_{ii}}}{r_{N2}} \leq \\ \frac{N^{\frac{1}{2}-\frac{1}{p}} \sqrt{\kappa \sum_{i=1}^N h_{ii}}}{r_{N2}} &= \frac{N^{\frac{1}{2}-\frac{1}{p}} \sqrt{\kappa q}}{r_{N2}} = \eta_2. \end{aligned}$$

(ii) For $p \geq 2$, applying Karamata's inequality and Jensen's inequality, we have

$$\begin{aligned} \frac{\mathbb{E} \left[\left(\frac{1}{N} \sum_{i=1}^N |Z_i|^p \right)^{\frac{1}{p}} \right]}{r_{N2}} &\leq \frac{N^{-\frac{1}{p}} \mathbb{E} \left[\left(\sum_{i=1}^N |Z_i|^2 \right)^{\frac{1}{2}} \right]}{r_{N2}} \leq \\ \frac{N^{-\frac{1}{p}} \sqrt{\mathbb{E} \left[\sum_{i=1}^N |Z_i|^2 \right]}}{r_{N2}} &= \frac{N^{-\frac{1}{p}} \sqrt{\sum_{i=1}^N \mathbb{E}[Z_i^2]}}{r_{N2}} \leq \\ \frac{N^{-\frac{1}{p}} \sqrt{\sum_{i=1}^N \kappa h_{ii}}}{r_{N2}} &= \frac{N^{-\frac{1}{p}} \sqrt{\kappa q}}{r_{N2}} = \eta_2. \end{aligned}$$

Combining the two cases, the desired result follows.

We have the following result for the Wasserstein distance between the true probability distributions of ε and $\tilde{\varepsilon}$.

Lemma 5. If Assumption 3 holds, then for any $\eta_3 \in (0, 1)$, $N \geq 1$, and

$$r_{N3} = \sqrt{\frac{\kappa}{\eta_3} ((\mathbf{x}^f)^\top (X^\top X)^{-1} \mathbf{x}^f)},$$

we have

$$\mathbb{P}_\varepsilon^N \{d_{w,p}(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) \geq r_{N3}\} \leq \eta_3.$$

Proof. Recall that $\tilde{\varepsilon} = \varepsilon + (\beta - \hat{\beta}_N^{\text{OLS}})^\top \mathbf{x}^f$. Thus, we find that $d_{w,p}(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) = |(\beta - \hat{\beta}_N^{\text{OLS}})^\top \mathbf{x}^f|$. Applying Chebyshev's inequality, we have

$$\begin{aligned} \mathbb{P}_\varepsilon^N \{d_{w,p}(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) \geq r_{N3}\} &= \mathbb{P}_\varepsilon^N \{ |(\beta - \hat{\beta}_N^{\text{OLS}})^\top \mathbf{x}^f| \geq r_{N3} \} \leq \\ \frac{\text{Var}(|(\beta - \hat{\beta}_N^{\text{OLS}})^\top \mathbf{x}^f|)}{(r_{N3})^2} &\leq \\ \frac{\text{Var}((\beta - \hat{\beta}_N^{\text{OLS}})^\top \mathbf{x}^f)}{(r_{N3})^2} &= \\ \frac{\text{Var}((X^\top X)^{-1} X^\top \boldsymbol{\varepsilon})^\top \mathbf{x}^f}{(r_{N3})^2} &= \\ \frac{\sigma^2 (\mathbf{x}^f)^\top (X^\top X)^{-1} \mathbf{x}^f}{(r_{N3})^2} &\leq \\ \frac{\kappa}{(r_{N3})^2} ((\mathbf{x}^f)^\top (X^\top X)^{-1} \mathbf{x}^f) &= \eta_3. \end{aligned}$$

The desired result follows.

The aforementioned lemmas give rise to the following finite-sample guarantee.

Proposition 1. (Finite-sample guarantee) Suppose that Assumptions 1, 2, and 3 hold, then for any $\eta \in (0, 1)$, there exists a radius r_N such that $r_N \rightarrow 0$ as $N \rightarrow \infty$ and $\hat{V}_N$ upper bounds V_{DRO} with high probability, that is,

$$\mathbb{P}^N \{V_{\text{DRO}} \leq \hat{V}_N\} \geq 1 - \eta. \quad (21)$$

Proof. For any $\eta \in (0, 1)$ and $N \geq 1$, let $\eta_i = \eta/3$, r_{Ni} be as defined in Lemmas 3, 4, and 5, $i = 1, 2, 3$, and $r_N = r_{N1} + r_{N2} + r_{N3}$. We observe that

$$d_{w,p}(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) \leq d_{w,p}(\mathbb{P}_{\varepsilon'}, \mathbb{P}_{\varepsilon}) + d_{w,p}(\mathbb{P}_{\varepsilon}, \hat{\mathbb{P}}_N(\varepsilon)) + d_{w,p}(\hat{\mathbb{P}}_N(\varepsilon), \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})).$$

By Lemmas 3, 4, and 5, we have

$$\begin{aligned} & \mathbb{P}^N \{d_{w,p}(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) \leq r_N\} \geq \\ & 1 - \mathbb{P}_{\varepsilon} \{d_{w,p}(\mathbb{P}_{\varepsilon'}, \mathbb{P}_{\varepsilon}) > r_{N3}\} - \\ & \mathbb{P}_{\varepsilon}^N \{d_{w,p}(\mathbb{P}_{\varepsilon}, \hat{\mathbb{P}}_N(\varepsilon)) > r_{N1}\} - \\ & - \mathbb{P}^N \{d_{w,p}(\hat{\mathbb{P}}_N(\varepsilon), \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) > r_{N2}\} \geq \\ & 1 - \eta. \end{aligned}$$

One can verify that $\mathbb{P}_{\varepsilon'} \in \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}))$ implies that $V_{\text{DRO}} \leq \hat{V}_N$. Then we have

$$\begin{aligned} \mathbb{P}^N \{V_{\text{DRO}} \leq \hat{V}_N\} & \geq \mathbb{P}^N \{\mathbb{P}_{\varepsilon'} \in \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}}))\} = \\ & \mathbb{P}^N \{d_{w,p}(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\varepsilon^{\text{OLS}})) \leq r_N\} \geq 1 - \eta. \end{aligned}$$

Therefore, the finite-sample guarantee that (21) holds, this completes the proof.

5 Numerical experiments

In this section, we conduct numerical experiments using a dataset simulated from real-world data to validate our theoretical findings. Our objective is to predict the total demand for bicycles at a certain time of day via external factors such as the wind speed and temperature. Accurately predicting the quantiles of the stochastic demand process will enhance bicycle inventory management. Section 5.1 describes the data generation process, while Section 5.2 presents a comprehensive performance comparison between our DRO method based on the p -Wasserstein metric and the standard SAA method. We compare our DRO method, which is based on the p -Wasserstein metric, with the methods in Ref. [23].

5.1 Data selection

The data used in the simulation experiment are a real dataset obtained from the UCI Machine Learning Repository curated by Dua and Graff^[27]. We utilized the Seoul bicycle sharing demand dataset, resulting in a dataset containing 8760 instances. This dataset includes the number of public bicycles rented per hour in the Seoul bicycle management system, along with corresponding weather data and holiday information. To address multicollinearity issues, some features were removed, resulting in a design matrix $\mathbf{X}$ with 7 features. The dataset encompasses weather information such as temperature, humidity, wind speed, visibility, dew point temperature, solar radiation, and rainfall.

Since the observations of the features correspond to periods of time, it is not appropriate to treat the observations as i.i.d. data. Therefore, we adopt the “fixed-design” assumption. First, a linear regression model was applied to the pre-processed dataset mentioned above. Next, we discarded the target column of the dataset, retaining only the feature columns. A new dataset was subsequently constructed by simulating the reconstructed target column. Specifically, in the simulation, we treated the estimated coefficients as potential true linear coefficients (β). The demand observations were then generated by adding i.i.d. simulated noise

$$y_i = \beta^T x_i^f + \varepsilon_i. \quad (22)$$

Given that our data are time-series related, the observed values are divided into a training set and a test set based on time periods. For any specific size of N , the training set comprises these N observations, while the subsequent 500 time periods' observations are utilized as the test set. The data processing method follows a similar approach to that of Ref. [23].

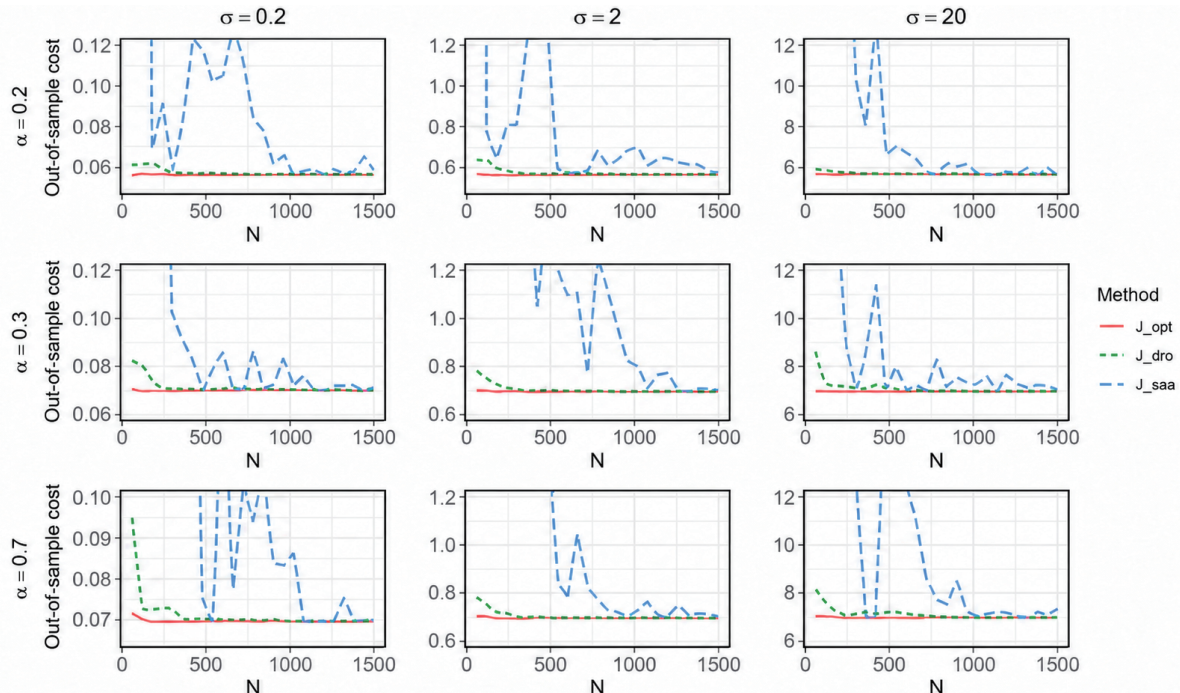


Fig. 1. Out-of-sample costs versus training sample size under Gaussian noise.

Table 1. Out-of-sample costs and relative cost differences under Gaussian noise.

σ	N	J_{opt}			J_{dro}			J_{ssa}			J_{droqi}								
		$\alpha = 0.2$	0.3	0.7	$\alpha = 0.2$	0.3	0.7	$\alpha = 0.2$	0.3	0.7	$\alpha = 0.2$	0.3	0.7						
0.2	60	0.056	0.070	0.072	0.061	0.082	0.095	2.416	(-97.5 %)	0.177	(-53.6%)	0.134	(-29.4 %)	0.092	(-34.0 %)	0.094	(-13.1 %)	0.112	(-15.6 %)
	300	0.056	0.070	0.069	0.057	0.071	0.073	0.059	(-3.7 %)	0.102	(-30.9 %)	0.389	(-81.3 %)	0.068	(-16.3 %)	0.078	(-9.6 %)	0.081	(-10.1 %)
	600	0.056	0.070	0.070	0.057	0.071	0.070	0.105	(-46.0 %)	0.086	(-17.5 %)	0.135	(-47.9 %)	0.062	(-9.2 %)	0.075	(-6.3 %)	0.074	(-5.3 %)
	900	0.056	0.070	0.070	0.056	0.070	0.070	0.062	(-8.9 %)	0.072	(-2.6 %)	0.084	(-16.3 %)	0.058	(-3.2 %)	0.072	(-2.9 %)	0.071	(-0.9 %)
	1200	0.056	0.070	0.070	0.056	0.070	0.070	0.056	(-0.0 %)	0.070	(-0.1 %)	0.070	(0.0 %)	0.057	(-1.4 %)	0.071	(-2.0 %)	0.071	(-1.2 %)
2.0	1500	0.056	0.070	0.070	0.056	0.070	0.070	0.059	(-3.9 %)	0.071	(-1.8 %)	0.070	(-0.4 %)	0.057	(-0.7 %)	0.070	(-0.4 %)	0.071	(-1.3 %)
	60	0.565	0.696	0.699	0.632	0.779	0.778	30.471	(-97.9 %)	8.907	(-91.3 %)	10.099	(-92.3 %)	0.790	(-20.0 %)	1.016	(-23.4 %)	1.090	(-28.6 %)
	300	0.560	0.695	0.691	0.570	0.703	0.696	0.808	(-29.5 %)	1.489	(-52.8 %)	4.515	(-84.6 %)	0.698	(-18.3 %)	0.791	(-11.1 %)	0.726	(-4.1 %)
	600	0.561	0.691	0.691	0.573	0.697	0.693	0.560	(2.2 %)	1.098	(-36.5 %)	0.774	(-10.5 %)	0.668	(-14.2 %)	0.731	(-4.6 %)	0.704	(-1.5 %)
	900	0.560	0.692	0.691	0.564	0.694	0.692	0.635	(-11.1 %)	0.958	(-27.5%)	0.718	(-3.5 %)	0.597	(-5.4 %)	0.702	(-1.1 %)	0.696	(-0.5 %)
20.0	1200	0.560	0.691	0.691	0.561	0.693	0.692	0.641	(-12.5 %)	0.772	(-10.2 %)	0.699	(-1.0 %)	0.575	(-2.5 %)	0.699	(-0.9 %)	0.695	(-0.4 %)
	1500	0.560	0.691	0.690	0.561	0.693	0.692	0.574	(-2.3%)	0.699	(-0.8 %)	0.700	(-1.2 %)	0.563	(-0.5 %)	0.699	(-0.9 %)	0.695	(-0.5 %)
	60	5.626	6.917	6.998	5.833	8.590	8.119	97.176	(-94.0 %)	18.464	(-53.5 %)	31.833	(-74.5%)	6.084	(-4.1 %)	9.564	(-10.2 %)	9.373	(-13.4 %)
	300	5.621	6.917	6.934	5.659	7.075	7.105	10.269	(-44.9 %)	7.024	(0.7%)	12.938	(-45.1 %)	6.129	(-7.7 %)	7.611	(-7.0 %)	7.865	(-9.7 %)
	600	5.613	6.924	6.944	5.642	6.981	7.098	6.697	(-15.8 %)	6.978	(0.0 %)	12.547	(-43.4 %)	5.732	(-1.6 %)	7.426	(-6.0 %)	7.249	(-2.1 %)
20.0	900	5.616	6.930	6.935	5.643	6.957	7.953	6.009	(-6.1 %)	7.525	(-7.6 %)	8.474	(-17.9 %)	5.646	(-0.1 %)	7.095	(-2.0 %)	7.056	(-1.5 %)
	1200	5.612	6.922	6.938	5.626	6.936	6.957	5.777	(-2.6%)	7.619	(-9.0 %)	7.154	(-2.8 %)	5.636	(-0.2 %)	6.999	(-0.9 %)	6.970	(-0.2 %)
	1500	5.612	6.917	6.934	5.623	6.926	6.946	5.617	(-0.1 %)	7.014	(-1.3 %)	7.320	(-5.1 %)	5.634	(-0.2 %)	6.988	(-0.9 %)	6.966	(-0.3 %)

5.2 Comparison of out-of-sample cost

In this section, we first evaluate the performance of our 2-Wasserstein DRO method and compare it to the SAA method and the optimal solution, which can be evaluated due to the known distribution of noise ε and β . Therefore, we can compare the out-of-sample costs of these methods.

For each experiment, given the training set size, the response variable value in the training set is generated via (22) along with the training feature vector. Both the 2-Wasserstein DRO method and the SAA method use a training set. Subsequently, i.i.d. noise is simulated 50 times for each of the 500 feature vectors in the test set, resulting in a total of 25000 data samples. The worst-case out-of-sample performances of the optimal solution and solutions of the SAA method and the 2-Wasserstein DRO method are evaluated using the average cost on the test set, which are denoted by J_{opt} , J_{saa} , and J_{dro} , respectively.

Fig. 1 illustrates that the out-of-sample performance of the different approaches in different configurations of the quantile level α and the residual standard deviation σ . From Fig. 1, we observe that our 2-Wasserstein DRO method outperforms the SAA method, particularly with a small sample size. Additionally, we find that the SAA method exhibits unstable performance under the “fixed-design” setting. As the sample size increases, both the DRO and SAA methods converge to the optimal solution and the rate of convergence of the DRO method is much higher than that of the SAA method. However, owing to the nonrandom generation of data points in the training and test sets, the performance guarantee no longer holds for the SAA method; for further details, see Ref. [28]. Therefore, the 2-Wasserstein DRO method performs better and is more robust than the SAA method. Table 1 summarizes the quantified out-of-sample performance. For comparison, we incorporate the performance of the DRO method proposed by Ref. [23], which is denoted by $J_{\text{dro}_{\text{qi}}}$. In Table 1, the percentage ratio described in parentheses corresponds to the improvement of our DRO method with the p -Wasserstein distance relative to the SAA method and the DRO method, calculated as $(J_{\text{dro}} - J_{\text{saa}})/J_{\text{saa}}$ and $(J_{\text{dro}} - J_{\text{dro}_{\text{qi}}})/J_{\text{dro}_{\text{qi}}}$. These values indicate that compared with the SAA method and the DRO method proposed by Ref. [23], our DRO strategy with 2-Wasserstein distance achieves better quantile prediction.

6 Conclusions

In this paper, we propose a new DRO framework by extending the classical Wasserstein measure to conditional quantile prediction and constructing the problem under the general p -Wasserstein metric. The feasibility and reconfigurability of the conditional quantile prediction DRO problem under the quantile loss function are studied. By using OLS estimators of linear coefficients, we develop a regression and then robust framework, and derive new concentrated inequalities. We bind the expected out-of-sample cost to the probability of any expectation. In addition, we show that the proposed distributionally robust framework implements a simple and practical reformulation that can be solved in polynomial time. Finally, a numerical experiment was conducted to evaluate the per-

formance of the proposed DRO method, along with comparisons against the SAA method and the DRO method developed by Ref. [23] under the 1-Wasserstein metric. The results demonstrate that our DRO method achieves superior performance over the other two methods.

This work provides an alternative to relying solely on the empirical residual distribution for conditional quantile prediction. By combining a residual-based Wasserstein ambiguity set with finite-sample guarantees and a tractable reformulation, the proposed framework offers a practical way to address distributional uncertainty when only limited data are available. Future research could extend the regression-then-robustify approach to nonlinear regression models and investigate adaptive selection of the Wasserstein order and radius. An important consideration in these extensions is how to account for regression estimation error and data-dependent calibration while preserving computational tractability and valid finite-sample guarantees.

Acknowledgements

We are grateful to Prof. Tiantian Mao for valuable discussions concerning model design and theoretical proofs.

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Bo Cheng received the M.S. degree in Statistics from University of Science and Technology of China. He received his B.S. degree in Economics from Anhui University. His research mainly focuses on financial probability and risk management.

Xinqiao Xie received the Ph.D. degree in Statistics from University of Science and Technology of China (USTC). She received her B.S. degree in Mathematics and Applied Mathematics from USTC in 2020. Her research mainly focuses on risk management, investment strategy, and financial probability.

References

- [1] Koenker R, Bassett Jr G. Regression quantiles. *Econometrica*, **1978**, 46 (1): 33–50.
- [2] Koenker R, Hallock K F. Quantile regression. *Journal of Economic Perspectives*, **2001**, 15 (4): 143–156.
- [3] Koenker R. Quantile Regression. Cambridge, UK: Cambridge University Press, **2005**.
- [4] Abrevaya J. The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, **2001**, 26: 247–257.
- [5] Dunham J B, Cade B S, Terrell J W. Influences of spatial and temporal variation on fish-habitat relationships defined by regression quantiles. *Transactions of the American Fisheries Society*, **2002**, 131 (1): 86–98.
- [6] Burgette L F, Reiter J P, Miranda M L. Exploratory quantile regression with many covariates: an application to adverse birth outcomes. *Epidemiology*, **2011**, 22 (6): 859–866.
- [7] Huang Q, Zhang H, Chen J, et al. Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, **2017**, 8 (3): 1–6.
- [8] Scarf H E, Arrow K J, Karlin S. A min-max solution of an inventory

- problem. Santa Monica, CA, USA: RAND Corporation, 1957.
- [9] Zhu Z, Zhang J, Ye Y. Newsvendor optimization with limited distribution information. *Optimization Methods and Software*, **2013**, 28 (3): 640–667.
 - [10] Perakis G, Roels G. Regret in the newsvendor model with partial information. *Operations Research*, **2008**, 56 (1): 188–203.
 - [11] Qiu R, Shang J, Huang X. Robust inventory decision under distribution uncertainty: A CVaR-based optimization approach. *International Journal of Production Economics*, **2014**, 153: 13–23.
 - [12] Wang Z, Glynn P W, Ye Y. Likelihood robust optimization for data-driven problems. *Computational Management Science*, **2016**, 13: 241–261.
 - [13] Bertsimas D, Gupta V, Kallus N. Robust sample average approximation. *Mathematical Programming*, **2018**, 171: 217–282.
 - [14] Bayraksan G, Love D K. Data-driven stochastic programming using phi-divergences. In: *The Operations Research Revolution*. Catonsville, MD, USA: INFORMS, **2015**: 1–19.
 - [15] Jiang R, Guan Y. Data-driven chance constrained stochastic program. *Mathematical Programming*, **2016**, 158 (1): 291–327.
 - [16] Erdoğan E, Iyengar G. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, **2006**, 107: 37–61.
 - [17] Gao R, Kleywegt A. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, **2023**, 48 (2): 603–655.
 - [18] Mohajerin Esfahani P, Kuhn D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, **2018**, 171 (1): 115–166.
 - [19] Zhao C, Guan Y. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, **2018**, 46 (2): 262–267.
 - [20] Blanchet J, Murthy K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, **2019**, 44 (2): 565–600.
 - [21] Blanchet J, Kang Y, Murthy K. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, **2019**, 56 (3): 830–857.
 - [22] Rosset S, Tibshirani R J. From Fixed-X to Random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, **2020**, 115 (529): 138–151.
 - [23] Qi M, Cao Y, Shen Z. Distributionally robust conditional quantile prediction with fixed design. *Management Science*, **2022**, 68 (3): 1639–1658.
 - [24] Villani C. *Optimal Transport: Old and New*. Berlin, Heidelberg: Springer, **2009**.
 - [25] Villani C. *Topics in Optimal Transportation*. Providence, RI, USA: American Mathematical Society, **2021**.
 - [26] Fournier N, Guillin A. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, **2015**, 162 (3): 707–738.
 - [27] Dua D, Graff C. UCI machine learning repository. **2017**. <http://archive.ics.uci.edu/ml>. Accessed April 1, 2024.
 - [28] Ban G Y, Rudin C. The big data newsvendor: Practical insights from machine learning. *Operations Research*, **2019**, 67 (1): 90–108.