Fast algorithms for semi-supervised noncrossing multiple quantile regression in high dimensions

Youngwoo Kwon[†], Ben Sherwood[‡], Aaron J. Molstad[†,1]
School of Statistics, University of Minnesota[†]
School of Business, University of Kansas[‡]

Abstract

We propose a semi-supervised method for multiple quantile regression. Traditional multiple quantile regression methods often suffer from the problem of quantile crossing, where a lower quantile estimate exceeds a higher one. We address this issue by enforcing a noncrossing constraint that preserves the monotonicity of fitted quantiles across levels. We impose the ordering constraints not only on the training covariates but also on the unlabeled test covariates. Our framework also accommodates standard penalty functions applied to the coefficient matrix. To compute our estimator, we use the Condat-Vũ primal-dual algorithm, which efficiently handles the large-scale constraints. In addition to linear quantile regression, we extend our framework to nonlinear quantile functions residing in a reproducing kernel Hilbert space. In simulation studies, we demonstrate that our method achieves improved logical consistency while maintaining comparable or better prediction accuracy relative to existing estimators. An application in chemometrics further illustrates the usefulness of our method.
Keywords. Quantile regression, non-crossing constraint, convex optimization, splitting algorithms

1 Introduction

Quantile regression, introduced by Koenker and Bassett Jr (1978), has become a fundamental tool for analyzing the relationship between covariates and the conditional distribution of a response variable. Compared with standard mean regression, quantile regression characterizes the conditional distribution more comprehensively, while remaining robust to heavy-tailed

[1]Corresponding author: amolstad@umn.edu

errors and capable of modeling heteroscedasticity. In practice, it is sometimes required to estimate several conditional quantiles simultaneously, for example in systemic risk measurement in finance (Adrian and Brunnermeier, 2016), in analyses of changing temperature distributions and extremes in climate science (Haugen et al., 2018), in modeling the effects of environmental pollutants on birth outcomes (Jin et al., 2025), and in travel time reliability analysis in transportation engineering (Ma et al., 2017).

In Section 7 we model the conditional distribution of octane rating Y given near–infrared spectral covariates X using the gasoline near–infrared spectroscopy dataset. A naïve choice is to fit each quantile level separately using linear quantile regression. The resulting conditional quantile curves cross for several gasoline samples. For example, for one sample point, the fitted 0.9 conditional quantile is smaller than the fitted 0.1 conditional quantile; see Table 1. Such a pattern cannot occur for genuine quantiles, since higher conditional quantiles must be at least as large as lower ones. This violation is the well-known problem of quantile crossing (He, 1997; Takeuchi et al., 2006; Liu and Wu, 2009; Bondell et al., 2010).

More generally, quantile crossing occurs when estimated quantile curves violate the logical monotonicity constraint required by the definition of cumulative distribution functions. Any family of conditional quantile functions must be nondecreasing in the quantile index for every covariate value, so a fitted model that produces crossing conditional quantiles cannot correspond to any valid conditional distribution. Several approaches address quantile crossing by imposing noncrossing constraints across quantile levels. He (1997) achieves noncrossing by working within a restricted location scale formulation for the conditional quantiles, which guarantees monotonicity in the quantile index by construction. Bondell

Table 1: Estimated conditional octane quantiles at conditional quantile levels $\tau = 0.1, \ldots, 0.9$ for the 5th gasoline sample under a standard unconstrained quantile linear regression fit. Note that 0.1 quantile estimate is larger than the 0.9 quantile estimate.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Fitted quantiles	88.48	88.48	88.65	88.40	88.67	88.31	88.36	87.35	87.01

et al. (2010) instead incorporates explicit linear inequality constraints into a joint estimation problem across a finite grid of quantile levels, yielding noncrossing fits over a prescribed covariate region. Although these contributions are effective in low-dimensional settings, simultaneously integrating noncrossing constraints and regularization (e.g., variable selection) in high dimensions remains an active area of research.

Some recent work has attempted to combine noncrossing constraints with sparsity-inducing penalties. Shin et al. (2024) proposed a deep neural network estimator that incorporates a group lasso penalty for variable selection while penalizing crossings via the loss function, but the noncrossing constraint is imposed through a soft penalty rather than a hard constraint, so the resulting fits do not come with a formal guarantee of monotonicity. Assuming a location–scale model, Wang et al. (2024) proposed a penalized approach that encourages similar coefficients across quantiles. This can indirectly lead to fitted quantile monotonicity, but crossings still may occur. Muggeo et al. (2013) developed a framework that combines penalized splines with constrained quantile regression. However, their method focuses on smoothing and does not incorporate structured regularization for high-dimensional coefficient estimation across many covariates or across quantiles.

Noncrossing constraints are typically enforced only on the training set. Consequently, even if the estimated curves do not cross on the training sample, they may still cross when making predictions for new observations. Ando and Li (2025) recently proposed a simplex–based parametrization of quantile regression that enforces noncrossing quantile planes over a prescribed covariate domain, improving the behavior of fitted quantiles beyond the training set. Their framework, however, focuses on low- to moderate-dimensional linear models and does not incorporate sparsity-inducing penalties or other regularization schemes.

In this paper, we propose a semi-supervised framework for high-dimensional multiple quantile regression that enforces noncrossing constraints on both training and test data. We explicitly incorporate the test design matrix into the constraint set so that the fitted quantiles remain noncrossing at both the training and test points where predictions are required. To

solve the resulting large-scale optimization problem with hard constraints and non-smooth penalties, we use the Condat-Vũ primal-dual algorithm (Condat, 2013; Vũ, 2013). This primal–dual framework naturally handles popular convex penalties in a single scheme, which allows us to implement lasso, group lasso, and nuclear-norm regularization without changing the basic algorithm.

2 Semi-supervised multiple quantile regression

2.1 Overview

We observe n pairs (y_i, x_i), where $y_i \in \mathbb{R}$ is the response and $x_i \in \mathbb{R}^{p+1}$ is the predictor vector whose first element is fixed to 1 as an intercept. Let $F(y \mid X = x) = \Pr(Y \leq y \mid X = x)$ denote the conditional cumulative distribution function of Y given $X = x$, and let $Q_\tau(x) = \inf\{y \in \mathbb{R} : F(y \mid X = x) \geq \tau\}$ be the conditional τth quantile function for $\tau \in (0, 1)$. In the linear quantile regression model $Q_\tau(x) = x^\top \beta(\tau)$, Koenker and Bassett Jr (1978) estimate the coefficient vector $\beta(\tau)$ by minimizing the empirical check loss

$$\widehat{\beta}(\tau) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta),$$

where $\rho_\tau(u) = u(\tau - 1_{u<0})$ is the check loss function.

Suppose further that we are interested in estimating quantiles at J levels $0 < \tau_1 < \cdots < \tau_J < 1$ simultaneously. Let $B = [\beta_1, \ldots, \beta_J] \in \mathbb{R}^{(p+1)\times J}$ be the matrix of coefficients, where the jth column $\beta_j := \beta(\tau_j)$ corresponds to the quantile level τ_j and the first row corresponds to the intercept term for each quantile. In this setting the conditional quantiles should preserve their natural ordering in the quantile index. For any $x \in \mathbb{R}^{p+1}$, the function $\tau \mapsto Q_\tau(x)$ is nondecreasing, so $x^\top \beta_1 \leq x^\top \beta_2 \leq \cdots \leq x^\top \beta_J$ holds for all $x \in \mathbb{R}^{p+1}$.

However, estimating each β_j separately does not guarantee this property, leading to the quantile crossing problem. Further, as discussed in the introduction, even if noncrossing is enforced on the training sample, crossings may still occur at new covariate values since noncrossing on the training covariates does not in general imply noncrossing for out-of-sample

predictions. To see how we address this problem, let $X := X_{\text{train}} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times(p+1)}$ be the design matrix of the training set with intercept for n observations, and let $X_{\text{test}} \in \mathbb{R}^{m\times(p+1)}$ be the design matrix of the test set for m observations where predictions are required. We define the augmented design matrix X_0 by stacking the training and test design matrices row-wise $X_0 = [X_{\text{train}}^\top, X_{\text{test}}^\top]^\top \in \mathbb{R}^{(n+m)\times(p+1)}$. Our strategy is to enforce noncrossing not only on the training covariate values but on all covariate values in X_0. This shrinks the feasible set relative to enforcing the constraint only on X_{train}, and optimizing over this more restrictive feasible set improves the coherence of the fitted quantile functions and estimation efficiency. Under this constraint, we pose estimation as a constrained optimization problem with a regularization term to handle high dimensionality,i.e., we estimate B by solving

$$\underset{B\in\mathbb{R}^{(p+1)\times J}}{\operatorname{argmin}} \sum_{j=1}^{J}\sum_{i=1}^{n} \rho_{\tau_j}(y_i - x_i^\top \beta_j) + r_\lambda(B) \;\text{ subject to }\; X_0\beta_j \le X_0\beta_{j+1}, \;\; j \in [J-1], \tag{1}$$

where $[J-1] := \{1, \ldots, J-1\}$. Here, as above, the inequality is understood element-wise, and $r_\lambda : \mathbb{R}^{(p+1)\times J} \to [0, \infty)$ is a convex penalty function controlled by user-specified tuning parameter λ. We do not penalize the intercept row. The choice of r_λ determines how sparsity is imposed and how information is shared across quantiles. In what follows, we discuss three popular choices for r_λ.

Lasso. The lasso penalty applies an ℓ_1 regularization to the individual entries of B, with $r_\lambda(B) = \lambda \sum_{l=2}^{p+1}\sum_{j=1}^{J} |\beta_{lj}|$. In quantile regression, the ℓ_1 penalty is well known to induce sparsity and enable variable selection under suitable sparsity conditions (Belloni and Chernozhukov, 2011). In the multi-quantile formulation, the penalty acts element-wise on the coefficient matrix, so the set of active predictors can change with the quantile level.

Group lasso. The group lasso penalty applies row-wise ℓ_2 regularization, with $r_\lambda(B) = \lambda \sum_{l=2}^{p+1} \|\beta_{l\cdot}\|_2$, where each row $\beta_{l\cdot}$ collects the coefficients of a single predictor across all J quantile levels. This penalty induces row-wise sparsity, meaning that entire rows can be set exactly to zero and variables are therefore selected or removed simultaneously across all J

quantile levels (Yuan and Lin, 2006).

Nuclear norm. The nuclear norm penalty applies regularization on the coefficient matrix with $r_\lambda(B) = \lambda\|B_{-1}\|_*$, where $\|\cdot\|_*$ denotes the sum of singular values and B_{-1} is the submatrix of B with its first row, the intercept coefficients, removed. The nuclear norm is a standard convex relaxation of matrix rank (Recht et al., 2010). In our setting, it shrinks the coefficient matrix toward a low-rank structure by encouraging sparsity in the singular values of B_{-1}. Compared with the lasso- and group-lasso–type penalties, which promote sparsity in individual entries or rows of B, the nuclear norm penalty instead retains all predictors, but constrains their effects to act through a low-dimensional set of common factors.

Convex penalties other than these could also be considered: we discuss how our algorithm can accomodate a range of penalties in Section 3.

2.2 Existing methods

We briefly review existing approaches for handling quantile crossing in multiple quantile regression. In most of the existing methods, reviewed here and in the Supplementary Material, monotonicity in the quantile index is imposed either through explicit inequality constraints over a covariate domain or through penalties that couple adjacent quantile coefficients.

Bondell et al. (2010) enforce noncrossing by requiring $x^\top\beta_1 \le \cdots \le x^\top\beta_J$ for all x in a covariate region $D \subset \mathbb{R}^{p+1}$, where D is a closed convex hull of finitely many points with the first element fixed to 1 as an intercept. They estimate the coefficients jointly across $(\tau_1, \ldots, \tau_J)$ by solving the constrained problem

$$B \in \arg\min_B \sum_{j=1}^{J} w(\tau_j) \sum_{i=1}^{n} \rho_{\tau_j}(y_i - x_i^\top \beta_j)$$
$$\text{subject to} \quad x^\top\beta_j \le x^\top\beta_{j+1} \text{ for all } x \in D,\ j \in [J-1],$$

where $w(\tau_j) > 0$ is a weight for the jth quantile. After rescaling the covariate region D to the unit cube $[0,1]^p$, they show that the global noncrossing condition can be expressed using $J-1$ linear constraints through a suitable reparameterization and solved via standard linear

programming. They also extend the linear quantile regression model to spline-based quantile functions and introduce a total variation penalty to control smoothness while maintaining the noncrossing constraint over covariates in $[0,1]^p$. A lasso penalty could be added to their objective function without much modification of their algorithm as it would remain a linear programming problem. However, this algorithm does not scale to large p and large n, limiting its applicability in high-dimensional settings.

Szendrei et al. (2024) relate noncrossing constraints to fused LASSO type regularization across the quantile index. On conditional quantiles $0 < \tau_1 < \cdots < \tau_J < 1$, they estimate B by minimizing an aggregated check loss under a noncrossing constraint over a covariate region D, similar to Bondell et al. (2010). They then introduce a scalar tuning parameter $\alpha \geq 0$ that relaxes the constraint, ranging from a weak ordering requirement at a reference covariate value when $\alpha = 0$ to a sufficient constraint that enforces noncrossing over the rescaled domain when $\alpha = 1$. For $\alpha < 1$, the method imposes a weaker requirement than enforcing noncrossing for all covariate values in D, which reduces the computational burden. They solve the resulting estimator via linear programming and select α by cross validation. The method also does not accommodate general convex high-dimensional penalties r_λ, and it only guarantees noncrossing over D when $\alpha = 1$.

Due to space constraints, we discuss other existing methods in the Supplementary Material. To summarize, none of the existing methods simultaneously enforce hard noncrossing, accommodate general convex high-dimensional penalties, and scale well in p and in the number of noncrossing constraints. Thus, our work is well motivated.

3 Computation

3.1 Overview of Condat–Vũ algorithm

We compute the estimator in (1) using the Condat–Vũ primal–dual splitting algorithm (Condat, 2013; Vũ, 2013), a first-order primal–dual splitting method for composite convex

optimization problems. It is tailored to objective functions that combine a smooth convex term, non-smooth terms, and non-smooth terms composed with linear operators. For a proper closed convex function φ and scalar $\lambda > 0$, define the proximal operator of $\lambda\varphi$ as $\mathrm{prox}_{\lambda\varphi}[w] := \underset{v}{\mathrm{argmin}}\{\lambda\varphi(v) + \|v - w\|_2^2/2\}$. To describe the algorithm, we consider the composite form

$$\min_v \; g(v) + f(v) + h(Lv), \tag{2}$$

where g and h are proper closed convex functions with tractable proximal operators, f is convex with Lipschitz gradient, and L is a linear operator. For composite optimization problems with no smooth terms, like (1), one may take $f(v) = 0$ for all v. We outline the generic version of Condat-Vũ here, then specialize to our problem in the next subsection.

A key feature of the Condat–Vũ algorithm is that it handles the linearly composed non-smooth term $h(Lv)$ by introducing a dual variable u and working with an equivalent primal–dual formulation. Let $h^\star$ denote the convex conjugate of h, defined by $h^\star(u) = \sup_z\{\langle u, z\rangle - h(z)\}$. Then by Fenchel duality, $h(Lv) = \sup_u\{\langle Lv, u\rangle - h^\star(u)\}$, which yields the saddle point representation of the solution to (2),

$$\min_v \max_u \; g(v) + f(v) + \langle Lv, u\rangle - h^\star(u).$$

Starting from this formulation, the algorithm applies a primal–dual splitting to the first order optimality conditions so that the smooth and non-smooth components are processed separately within a single iteration. The differentiable term f is handled by an explicit gradient step, and the non-smooth terms are handled by proximal updates. A Condat–Vũ iteration takes the following form. Choose step sizes $\eta > 0$ and $\sigma > 0$, a relaxation sequence (ν_t), and an initial pair $(v^{(0)}, u^{(0)})$. To obtain the t-th iterate, compute in sequence

$$\begin{aligned}
\tilde{v}^{(t+1)} &= \mathrm{prox}_{\eta g}[v^{(t)} - \eta(\nabla f(v^{(t)}) + L^\top u^{(t)})],\\
\tilde{u}^{(t+1)} &= \mathrm{prox}_{\sigma h^\star}[u^{(t)} + \sigma L(2\tilde{v}^{(t+1)} - v^{(t)})],\\
v^{(t+1)} &= \nu_t \tilde{v}^{(t+1)} + (1 - \nu_t)v^{(t)},\\
u^{(t+1)} &= \nu_t \tilde{u}^{(t+1)} + (1 - \nu_t)u^{(t)}.
\end{aligned}$$

The first order conditions characterizing a solution $(v^\star, u^\star)$ are $0 \in \partial g(v^\star) + \nabla f(v^\star) + L^\top u^\star$ and $0 \in \partial h^\star(u^\star) - Lv^\star$, where ∂g and $\partial h^\star$ denote the subdifferentials of g and $h^\star$, respectively. Each proximal step enforces one of these inclusions implicitly, while the gradient ∇f and the coupling terms L and $L^\top$ are evaluated explicitly. The extrapolated term $2\tilde{v}^{(t+1)} - v^{(t)}$ stabilizes the explicit treatment of the linear coupling, maintaining convergence.

The algorithm is guaranteed to converge to a primal–dual solution under standard step size conditions determined by the Lipschitz constant of ∇f and the operator norm $\|L\|_{\mathrm{op}}$, together with a relaxation parameter. From a practical standpoint, each Condat–Vũ iteration consists of evaluations of ∇f, applications of L and $L^\top$, and proximal updates for g and $h^\star$. Since the algorithm does not require solving linear systems, the per-iteration cost is typically dominated by the matrix multiplications induced by L, with the remaining steps often reducing to closed-form shrinkage or projection operations.

In our setting, the framework applies directly because the objective can be written as a sum of convex terms composed with the linear maps $B \mapsto XB$ and $B \mapsto X_0 B$ together with the penalty r_λ. We make this formulation explicit in the next subsection.

3.2 Application to semi-supervised multiple quantile regression

To see how the algorithmic framework of Condat-Vũ applies to (1), we specialize the generic template (2) to our setting, taking the primal variable to be the coefficient matrix B and partitioning the dual variable as $u = (u_0^\top, u_1^\top)^\top$. First note that we can express (1) equivalently in the unconstrained form

$$\underset{B \in \mathbb{R}^{(p+1)\times J}}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{i=1}^{n} \rho_{\tau_j}(y_i - x_i^\top \beta_j) + \sum_{j=1}^{J-1} \mathcal{I}_{\mathbb{R}_+^{n+m}}(X_0\beta_{j+1} - X_0\beta_j) + r_\lambda(B), \tag{3}$$

where, taking $\infty \times 0 = 0$ by convention, $\mathcal{I}_{\mathbb{R}_+^{n+m}}(\nu) = \infty \times \mathbf{1}(\nu \notin \mathbb{R}_+^{n+m})$ is the inclusion indicator of the nonnegative orthant $\mathbb{R}_+^{n+m}$, which is a closed convex set. Because of this indicator function, if any quantiles cross, the objective function value in (3) will be infinite. Consequently, a solution to (3) must satisfy the noncrossing constraint.

To express (3) in the form (2), take $f(B) = 0$, $g(B) = r_\lambda(B)$ and $h(LB) = h_0(X_0 B) +$

$h_1(XB)$ where $h_1(XB) = \sum_{j=1}^{J}\sum_{i=1}^{n}\rho_{\tau_j}(y_i - x_i^\top\beta_j)$ and $h_0(X_0B) = \sum_{j=1}^{J-1}\mathcal{I}_{\mathbb{R}_+^{n+m}}(X_0\beta_{j+1} - X_0\beta_j)$. Here, $L = (X_0^\top, X^\top)^\top \in \mathbb{R}^{(2n+m)\times(p+1)}$ and the dual variable $u = (u_0^\top, u_1^\top)^\top \in \mathbb{R}^{(2n+m)\times J}$. This representation is useful because the proximal operators of g, h_0, and h_1 can be computed efficiently. For g and h_1, the proximal operators have a closed form, whereas for h_0, the proximal operator requires computing the Euclidean projection onto the monotone cone $= \{w \in \mathbb{R}^J \,:\, w_j \le w_{j+1} \text{ for } j \in [J-1]\}$, denoted `Iso` (for isotonic regression). Many efficient algorithms exist for computing this projection, perhaps the best known being the pool adjacent violators algorithm (PAVA). Then, once the proximal operator of h is obtained, we can use the Moreau decomposition to compute the proximal operator of $h^\star$ according to the identity $\operatorname{prox}_{\sigma h^\star}[v] = v - \sigma\operatorname{prox}_{h/\sigma}[v/\sigma]$.

The Condat-Vũ algorithm applied to (3) thus has iterates

$$
\begin{aligned}
\tilde{B}^{(t+1)} &= \operatorname{prox}_{\eta r_\lambda}[B^{(t)} - \eta L^\top u^{(t)}], && (4)\\
\tilde{u}_0^{(t+1)} &= \operatorname{prox}_{\sigma h_0^\star}[u_0^{(t)} + \sigma X_0(2\tilde{B}^{(t+1)} - B^{(t)})]\\
&= u_0^{(t)} + \sigma X_0(2\tilde{B}^{(t+1)} - B^{(t)}) - \sigma\,\texttt{Iso}[\sigma^{-1}u_0^{(t)} + X_0(2\tilde{B}^{(t+1)} - B^{(t)})], && (5)\\
\tilde{u}_1^{(t+1)} &= \operatorname{prox}_{\sigma h_1^\star}[u_1^{(t)} + \sigma X(2\tilde{B}^{(t+1)} - B^{(t)})]\\
&= u_1^{(t)} + \sigma X(2\tilde{B}^{(t+1)} - B^{(t)}) - \sigma\operatorname{prox}_{h_1/\sigma}[\sigma^{-1}u_1^{(t)} + X(2\tilde{B}^{(t+1)} - B^{(t)})], && (6)\\
B^{(t+1)} &= \nu_t\tilde{B}^{(t+1)} + (1-\nu_t)B^{(t)},\\
u^{(t+1)} &= \nu_t\tilde{u}^{(t+1)} + (1-\nu_t)u^{(t)}.
\end{aligned}
$$

First, the update (4) is typically available in closed form for many common penalties r_λ, including the lasso, group lasso, and nuclear norm. For the lasso, this is element-wise soft-thresholding $[|Z_{jk}| - \eta\lambda]_+\operatorname{sign}(Z_{jk})$ applied to each non-intercept entry of Z. For the group lasso, it is the block soft-threshold $[1 - \eta\lambda/\|Z_{j\cdot}\|_2]_+ Z_{j\cdot}$ applied row-wise to the non-intercept rows. For the nuclear norm, it is singular-value soft-thresholding of Z_{-1}, requiring one SVD of a $p \times J$ matrix per iteration. Second, (5) requires only evaluating `Iso`: $\mathbb{R}^{(n+m)\times J} \to \mathbb{R}^{(n+m)\times J}$. Notably, monotonicity need only be applied row-wise, so the projection is computed by

applying PAVA to each row of the matrix $V_0 = \sigma^{-1}u_0^{(t)} + X_0(2\tilde{B}^{(t+1)} - B^{(t)})$ in parallel. Finally, the dual update for $\tilde{u}_1$ requires computing $\text{prox}_{h_1/\sigma}$, the proximal operator of the scaled check loss. Because h_1 is separable across entries of XB, for $V = \sigma^{-1}u_1^{(t)} + X(2\tilde{B}^{(t+1)} - B^{(t)})$, the entire update reduces to the element-wise clipping $[\sigma(V - \text{prox}_{h_1/\sigma}[V])]_{ij} = \min\{1 - \tau_j, \max\{-\tau_j, \sigma(V_{ij} - y_i)\}\}$. The complexity of clipping is $O(nJ)$, which is dominated by the matrix–matrix multiplication XB, $O(npJ)$. To summarize, the per-iteration cost is dominated by the matrix–matrix multiplications X_0B and XB, which scale as $O\{(n+m)pJ\}$ and $O(npJ)$, respectively. Therefore, the overall per-iteration complexity is $O\{(n+m)pJ\}$, which is linear in the number of augmented covariate points $n+m$, the number of predictors p, and the number of quantile levels J.

We now state conditions under which the iterates converge. As noted above, the optimization problem in (1) fits the framework of Condat (2013). Specializing their result to our notation yields the following convergence theorem.

Theorem 1. *(Condat, 2013) Let $\eta > 0$ and $\sigma > 0$ be the primal and dual step sizes. Suppose $\sum_{t=0}^{\infty} \nu_t(2 - \nu_t) = +\infty$ and $\eta\sigma(\|X\|_{\text{op}}^2 + \|X_0\|_{\text{op}}^2) < 1$, where $\|\cdot\|_{\text{op}}$ denotes the spectral norm. Suppose that r_λ is a proper closed convex function. Then the sequence of primal–dual iterates $(B^{(t)}, u_0^{(t)}, u_1^{(t)})$ converges to a saddle point of the associated Lagrangian. Consequently, the primal sequence $B^{(t)}$ converges to a global minimizer of* (1).

For practical implementation, the regularization parameter λ is selected via five-fold cross-validation, and we choose the value that minimizes the validation check loss. To satisfy the condition in Theorem 1, we set $\eta = \sigma = 0.9(\|X\|_{\text{op}}^2 + \|X_0\|_{\text{op}}^2)^{-1/2}$ and $\nu_t = 1.5$ for all t.

4 Statistical perspective

We now consider the statistical properties of the constrained estimator defined in (1). The noncrossing constraint defines the closed convex set $\Theta = \{B \in \mathbb{R}^{(p+1)\times J} : [X_0\beta_j]_i \leq [X_0\beta_{j+1}]_i \text{ for all } i \in [n+m] \text{ and } j \in [J-1]\}$. Under the linear quantile regression model,

the true coefficient matrix B^* satisfies $X_0\beta_j^* \leq X_0\beta_{j+1}^*$ element-wise, since the conditional quantile function is nondecreasing in the quantile index. In particular, $B^* \in \Theta$, so the constrained estimator optimizes over a smaller feasible set that still contains the target. This observation has a useful consequence: any oracle inequality established for the unconstrained penalized estimator carries over to the constrained estimator without modification.

Remark 1. *Let $\widehat{B}$ denote the constrained estimator defined in* (1)*, and let $\widehat{B}_{\text{unc}}$ denote the unconstrained penalized estimator obtained by removing the noncrossing constraint. Since $B^* \in \Theta$, the constrained estimator satisfies the basic inequality*

$$\sum_{j=1}^{J}\sum_{i=1}^{n}\rho_{\tau_j}(y_i - x_i^\top \widehat{\beta}_j) + r_\lambda(\widehat{B}) \leq \sum_{j=1}^{J}\sum_{i=1}^{n}\rho_{\tau_j}(y_i - x_i^\top \beta_j^*) + r_\lambda(B^*). \tag{7}$$

This is the same basic inequality satisfied by $\widehat{B}_{\text{unc}}$. Consequently, any rate of convergence or oracle inequality for the unconstrained estimator that is derived from (7) *together with conditions on the design matrix X and the error distribution holds for $\widehat{B}$ with the same bound.*

Because $\widehat{B}$ minimizes the penalized check loss over Θ and $B^* \in \Theta$, inequality (7) holds. The unconstrained estimator $\widehat{B}_{\text{unc}}$ satisfies the same inequality because $B^* \in \mathbb{R}^{(p+1)\times J}$. The proofs of oracle inequalities for penalized quantile regression (e.g., see Belloni and Chernozhukov (2011) or Kato (2011)) proceed from this basic inequality using only properties of the design matrix and the error distribution. Since these arguments do not depend on the feasible set, the same bounds apply to $\widehat{B}$.

Remark 1 implies, for instance, that under the restricted eigenvalue and other regularity conditions in Belloni and Chernozhukov (2011), the lasso-penalized version of $\widehat{B}$ satisfies the same error bounds as the unconstrained lasso estimator. Since $\widehat{B} \in \Theta$ and $B^* \in \Theta$, while it is possible that $\widehat{B}_{\text{unc}} \notin \Theta$, it is natural to wonder if $\widehat{B}$ is a more efficient estimator. Asymptotically this is not the case. Theorem 1 from Bondell et al. (2010) demonstrates for fixed p and without a penalty that the unconstrained and constrained cases have the same limiting distribution. Their proof relies on the fact that the probability $\widehat{B}_{\text{unc}} \notin \Theta$ will go to zero as $n \to \infty$ and thus the unconstrained and constrained estimators share the

same asymptotic properties. This same logic would hold for the high-dimensional estimators suggesting that the rates of convergence will be the same for the constrained and unconstrained estimators. However, for finite sample size it is possible that a difference would be observed and we do see some evidence of that in the simulations in Section 6.

5 Extension to nonlinear multiple quantile regression

5.1 Overview

The linear model $Q_{\tau_j}(x) = x^\top \beta_j$ may be misspecified when the conditional quantile function depends on the covariates in a nonlinear way. We extend our semi-supervised framework to this setting using reproducing kernel Hilbert spaces (RKHS). Let $z_i \in \mathbb{R}^p$ denote the non-intercept portion of x_i for $i \in [n]$, and let $z_i^{\text{test}} \in \mathbb{R}^p$ denote the corresponding portion of each test covariate. Define $Z = (z_1, \ldots, z_n)^\top \in \mathbb{R}^{n \times p}$, $Z_{\text{test}} = (z_1^{\text{test}}, \ldots, z_m^{\text{test}})^\top \in \mathbb{R}^{m \times p}$ and $Z_0 = (Z^\top, Z_{\text{test}}^\top)^\top \in \mathbb{R}^{(n+m) \times p}$. Let $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be a symmetric positive definite kernel, and let $\mathcal{H}_\kappa$ denote the associated RKHS with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$ and norm $\| \cdot \|_{\mathcal{H}_\kappa}^2 = \langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$. For each quantile level τ_j, we model $Q_{\tau_j}(x) = \mu_j + f_j(z)$, where $\mu_j \in \mathbb{R}$ is an intercept and $f_j \in \mathcal{H}_\kappa$. We control the complexity of f_j by penalizing $\|f_j\|_{\mathcal{H}_\kappa}^2$.

Define the $n \times n$ Gram matrix K with entries $K_{ij} = \kappa(z_i, z_j)$, and the $(n+m) \times n$ cross-kernel matrix K_0 with entries $[K_0]_{ij} = \kappa([Z_0]_{i\cdot}, z_j)$. By the representer theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001), any minimizer of the regularized objective admits the finite expansion $f_j(\cdot) = \sum_{i=1}^n \alpha_{ji} \kappa(\cdot, z_i)$ for some coefficient vector $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jn})^\top \in \mathbb{R}^n$. Under this representation, $\|f_j\|_{\mathcal{H}_\kappa}^2 = \alpha_j^\top K \alpha_j$. Letting $\alpha = [\alpha_1, \ldots, \alpha_J] \in \mathbb{R}^{n \times J}$ and $\mu = (\mu_1, \ldots, \mu_J)^\top \in \mathbb{R}^J$, we estimate $(\widehat{\alpha}, \widehat{\mu})$ using

$$(\widehat{\alpha}, \widehat{\mu}) \in \underset{\substack{\alpha \in \mathbb{R}^{n \times J} \\ \mu \in \mathbb{R}^J}}{\operatorname{argmin}} \sum_{j=1}^J \sum_{i=1}^n \rho_{\tau_j}(y_i - [K\alpha_j]_i - \mu_j) + \frac{\lambda}{2} \sum_{j=1}^J \alpha_j^\top K \alpha_j$$

$$\text{subject to} \quad K_0 \alpha_j + 1_{n+m} \mu_j \;\le\; K_0 \alpha_{j+1} + 1_{n+m} \mu_{j+1}, \quad j \in [J-1], \tag{8}$$

where inequalities are element-wise. The noncrossing constraints enforce monotonicity each row of Z_0, i.e., at all training and test covariate values where predictions are required.

5.2 Application of Condat–Vũ algorithm to nonlinear setting

To express (8) in the composite form (2), define the augmented kernel matrices $\widetilde{K} = [1_n,\, K] \in \mathbb{R}^{n\times(n+1)}$, $\widetilde{K}_0 = [1_{n+m},\, K_0] \in \mathbb{R}^{(n+m)\times(n+1)}$, and let the primal variable be $\theta = (\mu, \alpha^\top)^\top \in \mathbb{R}^{(n+1)\times J}$, so that $\widetilde{K}\theta = K\alpha + 1_n\mu^\top$ and $\widetilde{K}_0\theta = K_0\alpha + 1_{n+m}\mu^\top$. Taking $g \equiv 0$ and defining $f(\theta) = \frac{\lambda}{2}\sum_{j=1}^J \alpha_j^\top K\alpha_j$, $h_0(\widetilde{K}_0\theta) = \sum_{j=1}^J \mathcal{I}_{\mathbb{R}_+^{n+m}}(\widetilde{K}_0\theta_{j+1} - \widetilde{K}_0\theta_j)$, and $h_1(Z) = \sum_{j=1}^J\sum_{i=1}^n \rho_{\tau_j}(y_i - Z_{ij})$, the objective in (8) takes the form $f(\theta)+h_0(\widetilde{K}_0\theta)+h_1(\widetilde{K}\theta)$ with linear operator $L = (\widetilde{K}_0^\top, \widetilde{K}^\top)^\top$. The gradient of f is $\nabla f(\theta) = (\mathbf{0}_{1\times J}^\top, \lambda(K\alpha)^\top)^\top$, which is Lipschitz with constant $\lambda\|K\|_{\mathrm{op}}$. Letting the dual variable be $u = (u_0^\top, u_1^\top)^\top$ with $u_0 \in \mathbb{R}^{(n+m)\times J}$ and $u_1 \in \mathbb{R}^{n\times J}$, and using $\mathrm{prox}_{\eta g} = \mathrm{Id}$ since $g \equiv 0$, the Condat–Vũ iterates become

$$
\begin{aligned}
\widetilde{\theta}^{(t+1)} &= \theta^{(t)} - \eta(\nabla f(\theta^{(t)}) + \widetilde{K}_0^\top u_0^{(t)} + \widetilde{K}^\top u_1^{(t)}),\\
\tilde{u}_0^{(t+1)} &= u_0^{(t)} + \sigma\widetilde{K}_0(2\widetilde{\theta}^{(t+1)} - \theta^{(t)}) - \sigma\,\mathtt{Iso}[\sigma^{-1}u_0^{(t)} + \widetilde{K}_0(2\widetilde{\theta}^{(t+1)} - \theta^{(t)})],\\
\tilde{u}_1^{(t+1)} &= u_1^{(t)} + \sigma\widetilde{K}(2\widetilde{\theta}^{(t+1)} - \theta^{(t)}) - \sigma\,\mathrm{prox}_{h_1/\sigma}[\sigma^{-1}u_1^{(t)} + \widetilde{K}(2\widetilde{\theta}^{(t+1)} - \theta^{(t)})],\\
\theta^{(t+1)} &= \nu_t\widetilde{\theta}^{(t+1)} + (1-\nu_t)\theta^{(t)},\\
u^{(t+1)} &= \nu_t\tilde{u}^{(t+1)} + (1-\nu_t)u^{(t)}.
\end{aligned}
$$

It follows from Condat (2013) that the iterates converge to a solution of (8) provided that $\eta, \sigma > 0$ and (ν_t) satisfy $\eta^{-1} - \sigma(\|\widetilde{K}_0\|_{\mathrm{op}}^2 + \|\widetilde{K}\|_{\mathrm{op}}^2) \geq \frac{\lambda}{2}\|K\|_{\mathrm{op}}$, $\nu_t \in (0,\delta)$ for all t, where $\delta = 2 - (\lambda/2)\|K\|_{\mathrm{op}}(\eta^{-1} - \sigma(\|\widetilde{K}_0\|_{\mathrm{op}}^2 + \|\widetilde{K}\|_{\mathrm{op}}^2))^{-1} \in [1,2)$, and $\sum_{t=0}^\infty \nu_t(\delta - \nu_t) = \infty$. The per-iteration cost is $O((n+m)nJ)$, dominated by multiplication with the $(n+m)\times n$ cross-kernel matrix K_0, since the Gram matrix K has the same column dimension.

6 Numerical studies

6.1 Non-crossing with augmented data points

We conduct simulation studies to evaluate the proposed noncrossing estimators by examining how augmenting the covariate points used to impose the noncrossing constraint affects prediction accuracy under different penalty choices. In each replication, we set $p = 50$,

training sample size $n = 100$, and test sample size $m = 100$. We generate latent Gaussian covariates with independent rows $Z_i \sim \mathrm{N}_p(0, \Sigma)$, where $\Sigma_{ll'} = 0.3^{|l-l'|}$ for $l, l' \in \{1, \ldots, p\}$. Let Φ denote the standard normal CDF, and define $w_i = \Phi(Z_i)$ componentwise. We fix the coefficient vector β_* to be sparse with the first five entries equal to 2 and the remaining entries equal to 0. We generate responses according to $Y_i = 1 + w_i^\top \beta_* + \epsilon_i$ with $\epsilon_i \sim \mathrm{N}(0, 1)$. We set $J = 9$ with $\tau = (0.1, \ldots, 0.9)$, and repeat the experiment for 50 replications.

In this setting, we fit two sparse penalized noncrossing estimators, based on the lasso and group lasso penalties, which are both designed for sparse coefficient settings. For each estimator, we vary the set of covariate points at which the noncrossing constraint is imposed. In the train-only setting, the constraint is enforced using only the training covariates. In the train-test setting, the constraint is enforced using both the training and test covariates. In the augmented settings, we additionally include either 200 or 1000 extra covariate points from the same covariate distribution for enforcing the constraint.

Because the error distribution is known, the true quantile coefficient matrix $B^* \in \mathbb{R}^{(p+1)\times J}$ can be constructed for the quantile levels under consideration. Let $\widehat{B}$ denote the corresponding estimated coefficient matrix. We measure coefficient estimation accuracy by $\mathrm{Err}_B = ((p+1)J)^{-1/2}\|B^* - \widehat{B}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm. We also compare the true and estimated conditional quantiles at the test covariates. Let $Q^* = X_{\mathrm{test}} B^*$ and $\widehat{Q} = X_{\mathrm{test}}\widehat{B}$. We measure quantile prediction accuracy by $\mathrm{Err}_Q = (mJ)^{-1/2}\|Q^* - \widehat{Q}\|_F$.

Figure 1 indicates that enlarging the set of covariate points used to impose the noncrossing constraint improves both coefficient estimation and quantile prediction accuracy. Relative to enforcing noncrossing only on the training covariates, enforcing the constraint on both the training and test covariates reduces the mean coefficient error by 7.18% for the lasso and 5.99% for the group lasso. The corresponding mean quantile prediction error decreases by 8.94% and 7.47%, respectively. This suggests that enforcing monotonicity at the prediction locations yields an immediate benefit. Adding further augmented covariate points provides additional improvements for both sparse penalties. With 1000 augmented points, the coefficient error

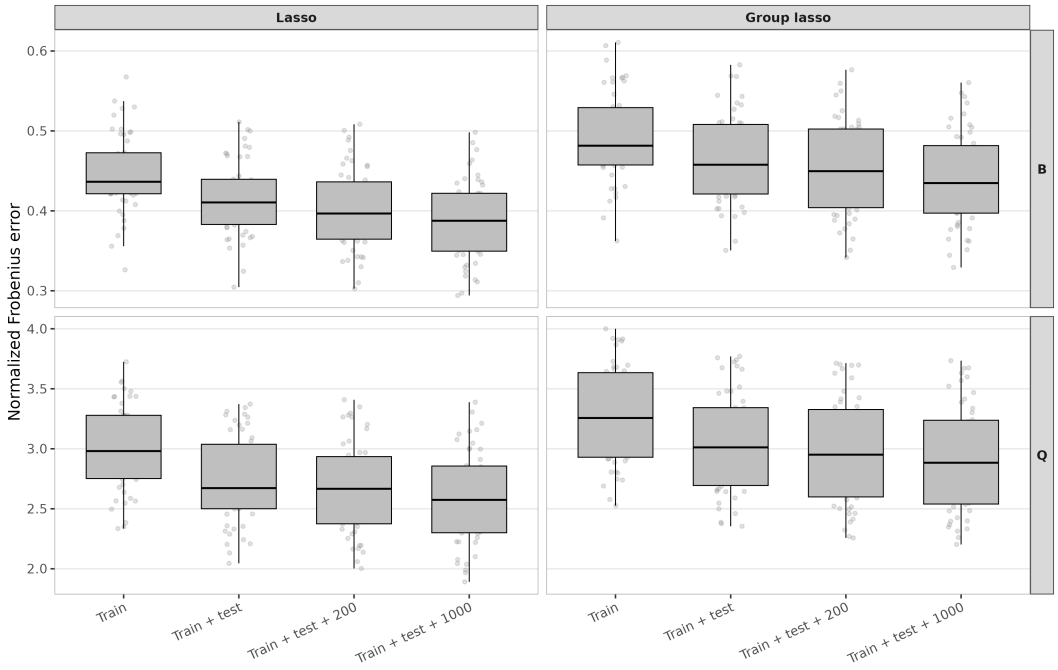

Figure 1: Simulation 1 with $n = 100$, $p = 50$, and $m = 100$. The top and bottom rows report the normalized coefficient error $\text{Err}_B = \|B^* - \widehat{B}\|_F / \sqrt{(p+1)J}$ and the normalized quantile prediction error $\text{Err}_Q = \|Q^* - \widehat{Q}\|_F / \sqrt{mJ}$, respectively. The noncrossing constraint is imposed using the training covariates, the training and test covariates, or the training and test covariates together with additional unlabeled covariate points.

decreases by 13.15% for the lasso and 10.78% for the group lasso, while the quantile prediction error decreases by 13.52% and 11.21%, respectively. Overall, these results support augmenting the constraint set as a practical way to improve recovery of the conditional quantile functions, particularly for the lasso and group lasso estimators.

6.2 Non-crossing in simulation

We conduct additional simulation studies to evaluate the empirical performance of the proposed noncrossing estimator under a range of data-generating mechanisms. We benchmark eight estimators, summarized in Table 2. These estimators differ in the choice of regularization and in how the noncrossing requirement is enforced. The regularization choices include ridge, lasso and group lasso penalties. We include a ridge-penalized estimator as a baseline method

Table 2: Method labels for Simulation 2.

Code	Penalty	Noncrossing constraint	Tuning parameter
A	Ridge	Unconstrained	Shared
B	Lasso	Unconstrained	Shared
C	Lasso	Unconstrained	By quantile
D	Lasso	Training covariates	Shared
E	Lasso	Training and test covariates	Shared
F	Group lasso	Unconstrained	Shared
G	Group lasso	Training covariates	Shared
H	Group lasso	Training and test covariates	Shared

because it remains well posed even when $n < p$. For the unconstrained lasso estimator, Method B uses a tuning parameter shared across all quantile levels, whereas Method C uses separate tuning parameters by quantile level. For the lasso and group lasso estimators, we compare unconstrained versions with noncrossing versions that impose the constraint either on the training covariates only or on the training and test covariates. Since this simulation focuses on sparse linear settings below, we do not include nuclear-norm regularization here. This method is instead considered in the real-data analysis in Section 7.

We consider a simulation setup that follows from the previous simulation in Section 6.1 for the noncrossing constraint for the augmented data. We generate latent Gaussian covariates Z_i and transformed covariates w_i as in the previous simulation. For estimation, we use the augmented vector $x_i = (1, w_i^\top)^\top \in \mathbb{R}^{p+1}$, and define the design matrix $X = (x_1, \ldots, x_n)^\top$. We fix the coefficient vector β_* to have its first five entries equal to 2 and the remaining entries equal to 0. We consider four data generating settings.

- Setting 1: location shift model with Gaussian errors, $Y_i = 1 + w_i^\top \beta_* + \epsilon_i$ and $\epsilon_i \sim N(0, 1)$.
- Setting 2: location shift model with heavy tailed errors, $Y_i = 1 + w_i^\top \beta_* + \epsilon_i$ and $\epsilon_i \sim t(2)$.
- Setting 3: asymmetric location shift model, $Y_i = 1 + w_i^\top \beta_* + \epsilon_i$ and $\epsilon_i \sim \chi^2(3) - 3$.
- Setting 4: location-scale shift model, $Y_i = 1 + w_i^\top \beta_* + (1 + w_i^\top \xi)\epsilon_i$, with $\epsilon_i \sim N(0, 1)$, $\xi = (2, 1, 0, \ldots, 0)^\top$.

Table 3: Summary of quantile crossing on the test data points in Simulation 2. Perc. of crossing is the percentage of test quantile entries that exhibit crossing. Q1, median, Q3, and Max summarize positive crossing magnitudes across test points, quantile levels, and replications. Rows aggregate methods by constraint setting.

Setting	Constraint setting	Perc. of crossing	Q1	Median	Q3	Max
$(n, p) = (100, 50)$	Unconstrained	22.20%	8.280e-02	2.006e-01	4.070e-01	4.887e+00
	Training	13.00%	5.040e-02	1.267e-01	2.742e-01	3.053e+00
	Training + test	0.00%	0	0	0	0
$(n, p) = (200, 50)$	Unconstrained	11.80%	5.670e-02	1.401e-01	3.011e-01	3.072e+00
	Training	7.45%	3.570e-02	8.870e-02	1.829e-01	3.164e+00
	Training + test	0.00%	0	0	0	0
$(n, p) = (50, 100)$	Unconstrained	30.44%	1.319e-01	3.072e-01	6.016e-01	5.560e+00
	Training	19.80%	8.920e-02	2.493e-01	5.594e-01	5.021e+00
	Training + test	0.00%	0	0	0	0

We fix the quantile grid at $J = 9$ with $\tau \in \{0.1, \ldots, 0.9\}$ and set the correlation parameter to $\phi = 0.3$. For each replication, we generate an independent test set of size $m = 100$. We consider three configurations of (n, p) given by $(100, 50)$, $(200, 50)$, and $(50, 100)$. Each setting is replicated 50 times. We quantify crossing by the entry-wise crossing $\delta_{ij} = \max\{m_{ij} - R_{ij},\, R_{ij} - M_{ij},\, 0\}$, where $R = X_{\text{test}}B$, R_{ij} is the (i, j) entry, $m_{ij} = \max_{\ell \le j} R_{i\ell}$, and $M_{ij} = \min_{\ell \ge j} R_{i\ell}$. Thus, for fixed i, $\delta_{ij} = 0$ for all j if and only if the fitted quantiles are nondecreasing in j for the ith row of X_{test}.

Table 3 reports the percentage of fitted quantile entries on the test data points that exhibit quantile crossing and summarizes the distribution of positive δ_{ij} across test points, quantile levels, and replications. The results show that enforcing noncrossing only on the training data points reduces crossings on the test data points, but does not eliminate them. In contrast, enforcing noncrossing on both the training and test covariate points eliminates crossings on the test data points up to numerical tolerance. Estimators that do not enforce noncrossing continue to exhibit crossings with non-negligible frequency and magnitude.

Figure 2 illustrates the distribution of test normalized Frobenius errors of eight different estimators from four data-generating mechanisms. Results show that imposing the noncrossing constraint yields modest but systematic gains over its unconstrained counterpart. The median

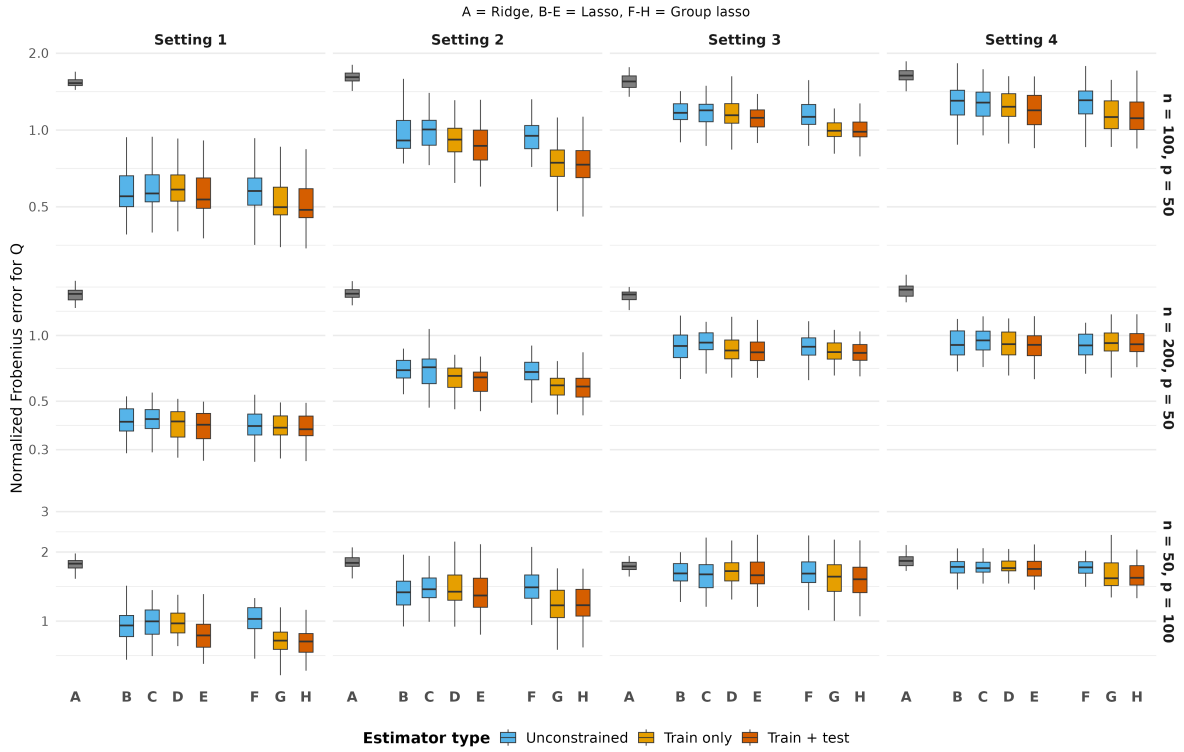

Figure 2: Normalized Frobenius error $\mathrm{Err}_Q = \|Q^* - \widehat{Q}\|_F / \sqrt{mJ}$ for the eight estimators across 50 replications in Simulation 2. The columns Setting 1–Setting 4 correspond to Gaussian location shift, heavy-tailed location shift, asymmetric location shift, and location-scale data-generating mechanisms, respectively. The rows correspond to the three sample-size and dimension settings. The method labels A–H are defined in Table 2.

test errors are lower and interquartile ranges are shorter with fewer extreme points. Typically the advantage is most visible in the group lasso estimators. Notably, the proposed noncrossing estimators maintain competitive predictive performance for most of the results, suggesting that the hard constraints do not over-restrict the model space but rather act as an effective regularizer.

6.3 Non-crossing in nonlinear multiple quantile regression

We next examine the nonlinear RKHS extension from Section 5. We consider three nonlinear location-shift models with covariate dimension $p = 3$. For each observation, we generate the covariate vector $x_i = (x_{i1}, \ldots, x_{ip})^\top$ with components independently and uniformly distributed

on $[-1, 1]$. The response is generated from $y_i = f(p^{-1/2} \sum_{j=1}^{p} x_{ij}) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.4$, and f is defined as either (Quadratic) $f(z) = 2z^2$, (Sine) $f(z) = 1.5 \sin(\pi z)$, or (Localized Gaussian) $f(z) = 2 \exp(-2z^2)$.

We fix the quantile grid at $J = 9$ with $\tau \in \{0.1, \ldots, 0.9\}$. For each replication, we generate independent training and test sets of sizes $n = 100$ and $m = 100$, respectively. Each setting is replicated 50 times.

We compare a linear baseline, a cubic polynomial estimator, and the RKHS estimator from Section 5, each fitted either without the noncrossing constraint or with the constraint imposed on both the training and test covariates. The cubic polynomial estimator uses all polynomial terms up to degree three. The RKHS estimator uses the Gaussian radial basis kernel $\kappa(z, z') = \exp\{-\gamma \|z - z'\|^2\}$, following the general framework of kernel quantile regression (Takeuchi et al., 2006). The kernel bandwidth γ and regularization parameter λ are selected by cross-validation, with the selected bandwidth shared by the unconstrained and noncrossing kernel estimators. We evaluate prediction by the normalized Frobenius error $\mathrm{Err}_Q = \|Q^* - \widehat{Q}\|_F / \sqrt{mJ}$, where Q^* and $\widehat{Q}$ are the true and estimated test quantile matrices.

Figure 3 reports the normalized Frobenius error for estimating the true conditional quantile matrix. The linear estimators have substantially larger errors across all three nonlinear settings, reflecting misspecification of the linear conditional quantile model. The full cubic estimators improve substantially over the linear estimators, especially in the quadratic setting. However, they show larger variability in the sine and localized Gaussian settings, which reflects the difficulty of estimating a comparatively rich polynomial basis when the nonlinear signal is not well represented by a cubic polynomial model. Across all three settings, the kernel estimators achieve the smallest errors. Comparing the two kernel estimators, imposing the noncrossing constraint reduces the mean error by 12.6%, 8.8%, and 4.7% in the quadratic, sine, and localized Gaussian settings, respectively. Overall, these results show that the RKHS extension captures nonlinear conditional quantile structure more effectively than the linear and polynomial alternatives, while the noncrossing constraint

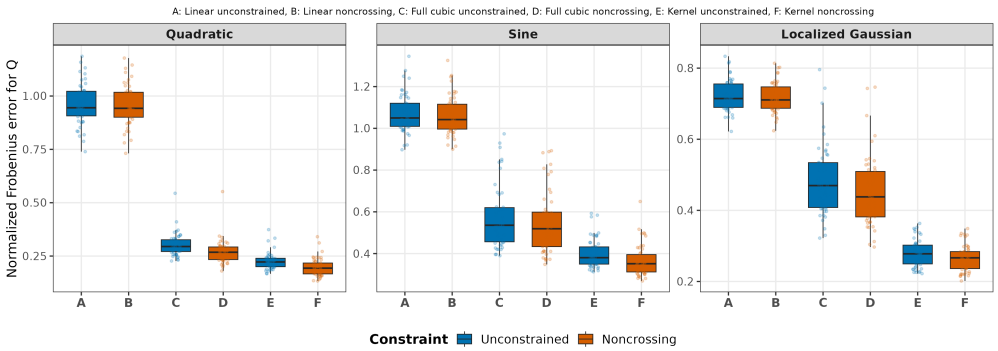

Figure 3: Normalized Frobenius error $\mathrm{Err}_Q = \|Q^* - \widehat{Q}\|_F/\sqrt{mJ}$ for the six estimators across 50 replications in Simulation 3. The columns correspond to the quadratic, sine, and localized Gaussian data-generating mechanisms. The method labels A–F denote linear unconstrained, linear noncrossing, cubic unconstrained, cubic noncrossing, kernel unconstrained, and kernel noncrossing, respectively.

further improves or maintains estimation accuracy. Though not shown here, unconstrained estimators tended to have greater than 25% of test quantile entries that exhibited crossing.

7 Octane ratings data analysis

We illustrate our method on the gasoline near–infrared (NIR) spectroscopy dataset from the `pls` R package (Mevik and Wehrens, 2007). The gasoline NIR calibration task is a standard example of spectroscopic prediction problems in chemometrics, where the goal is to predict octane rating from near-infrared spectra. Octane rating affects engine performance and must satisfy regulatory standards, so NIR-based prediction models have been studied as faster alternatives to laboratory assays (Kelly et al., 1989; Bohács et al., 1998). Quantile regression is attractive in this setting because it describes the full conditional distribution of octane given the spectrum, rather than only a single mean prediction, and it can accommodate heteroskedasticity and outliers. Noncrossing conditional quantile curves yield interpretable prediction bands at different risk levels, which crossing curves cannot.

The dataset consists of $n = 60$ gasoline samples with near–infrared spectra. For each sample we record $p = 401$ intensity values at wavelengths from 900 to 1700 nm in steps

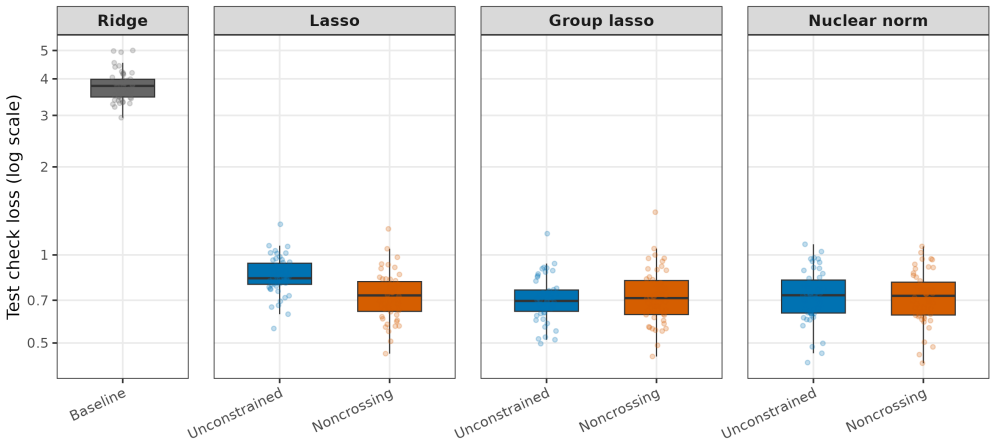

Figure 4: Test errors of seven estimators on the gasoline near-infrared spectroscopy data. Boxplots display the distribution of test check loss for the lasso, group-lasso, and nuclear-norm penalties, each fitted with either crossing allowed or a noncrossing constraint enforced, together with the ridge baseline.

of 2 nm, and we collect these predictors in the matrix `NIR` of size 60×401. The response variable `octane` gives the octane rating for each sample. In our analysis we select the 50 spectral channels with the largest sample variance, standardize these predictors, and add an intercept column. We then take `octane` as the response, draw a random 70%/30% train–test split, fit $J = 9$ quantile regression models at levels $\tau_j = j/(j+1)$ for $j \in [J]$, and evaluate performance on the test set.

We compare seven estimators on the gasoline data: Ridge quantile regression as a baseline, three crossing estimators with lasso, group lasso, and nuclear norm penalties, and their three noncrossing counterparts. Table 4 summarizes the magnitude of quantile crossings. Models

Table 4: Summary of quantile crossing on the test data points for the gasoline near-infrared spectroscopy data. Perc. of crossing reports the percentage of fitted test quantile entries that exhibit crossing, and Q1, median, Q3, and Max summarize positive crossing magnitudes. Rows aggregate estimators by constraint setting.

Model	Perc. of Crossing	Q1	Median	Q3	Max
Baseline (Ridge)	0.9%	3.072e-02	7.133e-02	9.474e-02	1.422e-01
Unconstrained	17.6%	8.220e-03	2.760e-02	6.395e-02	5.786e-01
Noncrossing	0%	0	0	0	0

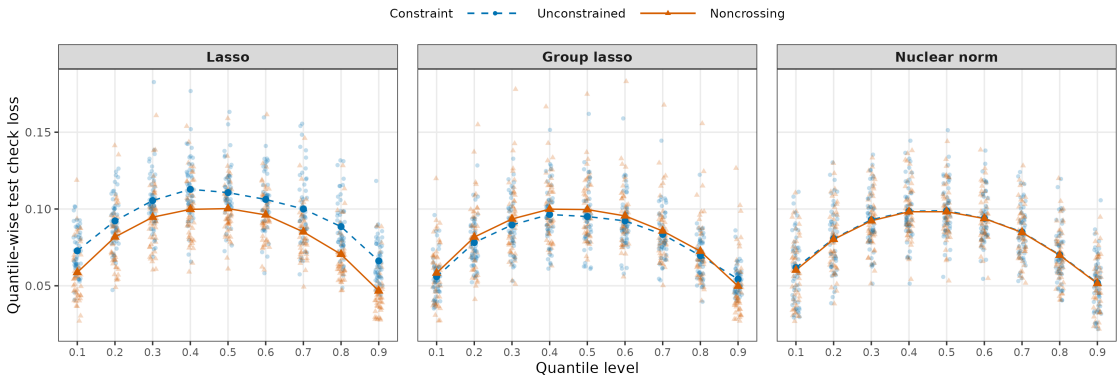

Figure 5: Test check loss by quantile level on the gasoline near-infrared spectroscopy data. For each quantile level $\tau \in \{0.1, \ldots, 0.9\}$, lines show the average loss for the lasso, group lasso, and nuclear norm penalties, comparing models with and without the noncrossing constraint.

that permit crossing retain nonzero violations, whereas imposing the noncrossing constraint drives these values exactly to zero, effectively eliminating crossings in this dataset. This reduction in crossing violations is important for interpreting the spectral data as the fitted conditional octane ratings should be nondecreasing consistently as the quantile level increases.

Figure 4 reports the test error for each penalty. Enforcing noncrossing leads to clear improvements for lasso penalty, with lower medians and shorter upper tails. For the group lasso and nuclear-norm penalty, the noncrossing and unconstrained fits show very similar test error distributions. Overall, the effect of the noncrossing constraint on prediction accuracy depends on the penalty, while the crossing violations are eliminated in all noncrossing fits.

Similarly, Figure 5 shows the test check loss at each quantile level. For the lasso penalty, the noncrossing curve lies below the unconstrained curve for most $\tau \in [0.1, 0.9]$, and the individual points are more tightly clustered. For the group-lasso penalty, the unconstrained curve is slightly below the noncrossing curve across quantile levels, except at $\tau = 0.9$, where the noncrossing curve is lower. For the nuclear-norm penalty, the two curves are nearly indistinguishable.

References

Tobias Adrian and Markus K. Brunnermeier. CoVaR. *American Economic Review*, 106(7): 1705–1741, 2016.

Tomohiro Ando and Ker-Chau Li. Simplex quantile regression without crossing. *The Annals of Statistics*, 53(1):144–169, 2025.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Gy. Bohács, Z. Ovádi, and A. Salgó. Prediction of gasoline properties with near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 6(1):341–348, 1998.

Howard D. Bondell, Brian J. Reich, and Huixia Wang. Noncrossing Quantile regression curve estimation. *Biometrika*, 97(4):825–838, 2010.

Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

Matz A. Haugen, Michael L. Stein, Elisabeth J. Moyer, and Ryan L. Sriver. Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression. *Journal of Climate*, 31(20):8573–8588, 2018.

Xuming He. Quantile Curves without Crossing. *The American Statistician*, 51(2):186–192, 1997.

Yisen Jin, Aaron J Molstad, Ander Wilson, and Joseph Antonelli. Smooth and shape-constrained quantile distributed lag models. *Biometrics*, 81(3):ujaf101, 2025.

Kengo Kato. Group lasso for high dimensional sparse quantile regression models. *arXiv preprint arXiv:1103.1458*, 2011.

Jeffrey J. Kelly, Clyde H. Barlow, Thomas M. Jinguji, and James B. Callis. Prediction of gasoline octane numbers from near-infrared spectral features in the range 660–1215 nm. *Analytical Chemistry*, 61(4):313–320, 1989.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

Yufeng Liu and Yichao Wu. Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and its Interface*, 2(3):299–310, 2009.

Zhenliang Ma, Sicong Zhu, Haris N. Koutsopoulos, and Luis Ferreira. Quantile regression analysis of transit travel time reliability with automatic vehicle location and farecard data. *Transportation Research Record: Journal of the Transportation Research Board*, 2652(1): 19–29, 2017.

Bjørn-Helge Mevik and Ron Wehrens. The `pls` package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2):1–23, 2007.

Vito M. R. Muggeo, Mariangela Sciandra, Agostino Tomasello, and Sebastiano Calvo. Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and Ecological Statistics*, 20(4):519–531, Dec 2013. ISSN 1573-3009.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *Computational Learning Theory*, pages 416–426. Springer, 2001.

Jungmin Shin, Seunghyun Gwak, Seung Jun Shin, and Sungwan Bang. Simultaneous estimation and variable selection for a non-crossing multiple quantile regression using deep neural networks. *Statistics and Computing*, 34(3):102, Mar 2024. ISSN 1573-1375.

Tibor Szendrei, Arnab Bhattacharjee, and Mark E. Schaffer. Fused LASSO as non-crossing quantile regression. IZA Discussion Paper 17149, IZA Institute of Labor Economics, July 2024.

Ichiro Takeuchi, Quoc V Le, Timothy D Sears, and Alexander J Smola. Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7:1231–1264, 2006.

Bằng C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

Mingqiu Wang, Xiaoning Kang, Jiajuan Liang, Kun Wang, and Yuanshan Wu. Heteroscedasticity identification and variable selection via multiple quantile regression. *Journal of Statistical Computation and Simulation*, 94(2):297–314, 2024.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67, 2006.