# Exact computation of GMM estimators for instrumental variable quantile regression models[*]

Le-Yu Chen[†]

Institute of Economics, Academia Sinica

Sokbae Lee[‡]

Department of Economics, Columbia University

Centre for Microdata Methods and Practice, Institute for Fiscal Studies

September 2017

**Abstract**

We show that the generalized method of moments (GMM) estimation problem in instrumental variable quantile regression (IVQR) models can be equivalently formulated as a mixed integer quadratic programming problem. This enables exact computation of the GMM estimators for the IVQR models. We illustrate the usefulness of our algorithm via Monte Carlo experiments and an application to demand for fish.

**Keywords**: *generalized method of moments, instrumental variable, quantile regression, endogeneity, mixed integer optimization*

**JEL codes**: C21, C26, C61, C63

[*] We would like to thank the editor and three anonymous referees for helpful comments. This work was supported in part by the European Research Council (ERC-2014-CoG-646917-ROMIA), the British Academy (International Partnership and Mobility Scheme Grant, reference number PM140162), and the Ministry of Science and Technology, Taiwan (MOST105-2410-H-001-003-).

[†] E-mail: lychen@econ.sinica.edu.tw

[‡] E-mail: sl3841@columbia.edu

# 1 Introduction

The instrumental variable quantile regression (IVQR) and related models have been increasingly popular for studying the impacts of possibly endogenous covariates on the distribution of the outcome of interest. See a recent review by Chernozhukov and Hansen (2013) and references therein for the latest developments in identification, estimation, and inference as well as the list of empirical applications.

The IVQR model admits conditional moment restrictions which can be used to construct the estimating equations for the GMM estimation of the model parameters. However, the sample counterparts of the IVQR estimating equations are discontinuous in the parameters so that the resulting GMM estimation problem becomes a non-convex and computationally non-trivial optimization problem. Honoré and Hu (2004) provided a heuristic for computing the IVQR GMM estimates. Chernozhukov and Hansen (2006) developed the inverse quantile regression (QR) estimator that is not directly a GMM estimator but can be shown to be asymptotically equivalent to the IVQR GMM estimator. Xu and Burer (2017) proposed an alternative algorithm for computing the inverse QR estimator. The Markov chain Monte Carlo (MCMC) based Laplace type estimator of Chernozhukov and Hong (2003) can also be used as an approximation of the IVQR GMM estimator but it requires careful tunning in the MCMC implementation. Kaplan and Sun (2015) proposed a smoothed estimating equation approach which facilitates the GMM computation problem but requires the choice of the smoothing parameter.

In this paper, we are concerned with exact computation of the GMM estimates of the IVQR parameters. As pointed out by Andrews (1997), heuristic algorithms for computation of GMM estimates that do not guarantee to find the exact global optimum or a specific level of approximation to the global optimum may result in extremum estimators which could exhibit statistical behavior that is quite different from that established by theory. This source of computational uncertainty may impact on the empirical results. Hence, as a complement to the previous work on the IVQR computation, our paper provides a method for exact computation of the IVQR estimates within the classical GMM framework.

Our computational algorithm is based on the method of mixed integer optimization (MIO). Specifically, we show that the IVQR GMM estimation problem can be equivalently formulated as a mixed integer quadratic programming (MIQP) problem. Thanks to the developments in MIO solution algorithms and fast computing environments, this reformulation allows us to solve for the exact GMM estimates by using the modern efficient MIO solvers. Well-known numerical solvers such as CPLEX and Gurobi can be used to effectively solve large-scale MIQP problems. See Jünger, Liebling, Naddef, Nemhauser, Pulleyblank, Reinelt, Rinaldi, and Wolsey (2009), Achterberg and Wunderling (2013) and

Bertsimas, King, and Mazumder (2016, Section 2.1) for discussions on computational advances in solving the MIO problems. For classic texts on the MIO methodology and applications, see Nemhauser and Wolsey (1999) and Bertsimas and Weismantel (2005). See also Florios and Skouras (2008), Bilias, Florios, and Skouras (2013), Kitagawa and Tetenov (2015), Bertsimas, King, and Mazumder (2016) and Chen and Lee (2016) for related but distinct work on solving non-convex optimization problems in statistics and econometrics via the MIO approach.

The rest of this paper is organized as follows. In Section 2, we summarize the setup of the IVQR model and the inverse quantile regression method of Chernozhukov and Hansen (2006). In Section 3, we present the MIQP formulation of the IVQR GMM estimation problem. We conduct a simulation study of the performance of the MIQP based GMM estimates in Section 4 and illustrate the application of our computation approach in a real data exercise concerning the demand estimation in Section 5. We then conclude the paper in Section 6. Supplementary results of this paper are collated in Appendices A–C.

# 2 The instrumental variable quantile regression model

Let $Y$ be an outcome of interest. We consider the quantile regression model under endogeneity, which is characterized by the structural equation

$$Y = W'\theta(U), \tag{2.1}$$

where $U$ is an unobserved scalar random variable, $W = (D, X)$ is a vector of covariates, and $\theta(\cdot)$ is a measurable function such that the mapping $\tau \mapsto W'\theta(\tau)$ is strictly increasing in $\tau$ for almost every realization of $W$. The covariates $D$ may not be independent of $U$. We assume that there is a vector of instrumental variables, denoted as $Z$, which can be excluded from (2.1) but can influence the endogenous variables $D$ such that $\dim(Z) \geq \dim(D)$ and

$$U|X, Z \sim \text{Uniform}\,(0, 1)\,. \tag{2.2}$$

Under these assumptions, it follows that, for $\tau \in (0, 1)$,

$$P\left(Y \leq W'\theta(\tau)|X, Z\right) = P\left(U \leq \tau|X, Z\right) = \tau. \tag{2.3}$$

The model set forth so far is the well known linear IVQR model which has been studied by Chernozhukov and Hansen (2004, 2005, 2006, 2008), Chernozhukov, Hansen, and Jansson (2007, 2009), and Kaplan and Sun (2015) among many others. The value of $\theta(\tau)$ in this model captures the impact of the covariates $W$ on the outcome of an individual whose unobserved heterogeneity $U$ is fixed at $U = \tau$. In the setting with counterfactual outcomes, the quantile-specific parameter vector $\theta(\tau)$ can be causally interpreted as the structural quantile effect (Chernozhukov and Hansen, 2005). Therefore, given a random sample, $(Y_i, W_i, Z_i)_{i=1}^n$ of $n$ observations, we are interested in the estimation of $\theta(\tau)$ for some fixed values of $\tau \in (0, 1)$.

Note that, when there is no endogenous covariate, the IVQR model reduces to the linear QR model of Koenker and Bassett (1978) where $W = X$ and $W'\theta(\tau)$ is the $\tau$ quantile of the distribution of $Y$ conditional on $W$. However, in the presence of endogenous variables $D$, the $\tau$ conditional quantile of $Y$ given $W$ is not warranted to be $W'\theta(\tau)$ because of the statistical dependence between $U$ and $D$. In this case, $\theta(\tau)$ may not be consistently estimated via the conventional QR approach. Availability of instrumental variables $Z$ that satisfy (2.2) is thus useful for validating the identifying restrictions (2.3), which facilitate consistent estimation of $\theta(\tau)$. See also Chernozhukov and Hansen (2013, Section 4) for a review of alternative approaches on quantile models under endogeneity.

Chernozhukov and Hansen (2006) developed primitive conditions for the identification of $\theta(\tau)$ of the IVQR model. They also provided an inverse QR algorithm for the estimation of $\theta(\tau)$. To describe their algorithm, write $\theta = (\alpha, \beta)$ such that $W'\theta(\tau) = D'\alpha(\tau) + X'\beta(\tau)$. Let $\Psi_i = \Psi(X_i, Z_i)$ be a vector of transformations of instruments with $\dim(\Psi_i) \geq \dim(\alpha)$. Let $A$ be a given positive definite matrix. The Chernozhukov-Hansen inverse QR procedure proceeds as follows. Let

$$\widehat{\alpha}(\tau) \equiv \arg\inf_{\alpha\in\mathcal{A}} \widehat{\gamma}_\tau(\alpha)' A \widehat{\gamma}_\tau(\alpha), \tag{2.4}$$

where

$$\left(\widehat{\beta}_\tau(\alpha), \widehat{\gamma}_\tau(\alpha)\right) \equiv \arg\inf_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{G}} \frac{1}{n}\sum\nolimits_{i=1}^{n} \rho_\tau\left(Y_i - D_i'\alpha - X_i'\beta - \Psi_i'\gamma\right), \tag{2.5}$$

$\mathcal{A}$, $\mathcal{B}$ and $\mathcal{G}$ are compact parameter spaces, and the check function $\rho_\tau$ is defined by $\rho_\tau(u) = u(\tau - 1\{u < 0\})$ for $u \in \mathbb{R}$. The inverse QR estimator, denoted by $\widehat{\theta}_{IQR}(\tau)$, is then defined by

$$\widehat{\theta}_{IQR}(\tau) = \left(\widehat{\alpha}(\tau), \widehat{\beta}_\tau(\widehat{\alpha}(\tau))\right).$$

In the procedure above, the function $\Psi$ and the matrix $A$ can vary across $\tau$ and be replaced by their consistent estimates. Moreover, the QR objective function can be weighted across observations. See Chernozhukov and Hansen (2006) for further details.

For implementation, Chernozhukov and Hansen (2006) proposed to solve the outer optimization problem (2.4) by the grid search method. The inner optimization problem (2.5) is the standard quantile regression problem, which can be solved very efficiently. Thus, when $\dim(\alpha) = 1$, the inverse QR method is computationally appealing because its implementation amounts to solving convex optimization sub-problems within a low-dimensional global search procedure. However, this computational merit diminishes rapidly with the increase of the number of endogenous variables. Instead of performing grid search, Xu and Burer (2017) proposed an alternative method to compute the inverse QR estimator. Their approach is based on exact minimization of the quadratic norm as in (2.4) subject to the optimality conditions for the linear programming formulation of the QR problem of (2.5). Xu and Burer (2017) showed that the resulting computational problem reduces to a quadratic programming problem subject to complementarity constraints for which they developed a branch-and-bound algorithm to compute the exact solution.

# 3 Exact computation of the GMM based IVQR estimator via the mixed integer optimization approach

The conditional moment restriction (2.3) can be used to form estimating equations for the GMM estimation of $\theta\left(\tau\right)$. That is,

$$E\left[\left(1\left\{Y \leq W'\theta\left(\tau\right)\right\} - \tau\right)L\right] = 0, \tag{3.1}$$

where $L$ is a vector of instruments consisting of functions of $X$ and $Z$. As noted by Chernozhukov and Hansen (2006), the inverse QR estimator, which is not directly a GMM estimator, can be shown to be asymptotically equivalent to the GMM estimator with the instruments $L_{\mathrm{CH}} \equiv [X', \Psi\left(X,Z\right)']'$.

In this paper, we provide an algorithm for directly computing the GMM based IVQR estimator using the orthogonality conditions (3.1). Let $s_\tau(t)$ denote the vector $(s_{\tau,i}(t))_{i=1}^{n}$, where $s_{\tau,i}(t) \equiv 1\left\{Y_i \leq W_i't\right\} - \tau$ for $i \in \{1,...,n\}$. Let $G$ be the $n$-by-dim$(L)$ matrix whose $i$th row vector is $L_i'$. Let $\widehat{Q}$ be a given positive definite matrix of dimension dim$(L)$. The GMM based IVQR estimator of $\theta\left(\tau\right)$, denoted by $\widehat{\theta}_{GMM}\left(\tau\right)$, is given by

$$\widehat{\theta}_{GMM}\left(\tau\right) = \arg\inf\nolimits_{\theta\in\Theta} s_\tau\left(\theta\right)' G\widehat{Q}G's_\tau\left(\theta\right), \tag{3.2}$$

where $\Theta$ is the compact parameter space of $\theta$.[1]

We now present our computational algorithm, which is based on the method of mixed integer optimization. Let $c_\tau = (\tau, ..., \tau)$ denote the $n$ dimensional vector of constants all taking the same value specified by the quantile index $\tau$. We note that the optimization problem (3.2) can be equivalently formulated as the following constrained optimization

[1] Compactness of the parameter space is a standard assumption for deriving consistency of the GMM estimator (see e.g., Newey and McFadden, 1994, Theorem 2.6). This assumption is difficult to relax when the GMM objective function is non-convex. In practice, knowledge of the parameter space also helps to tighten the global optimization problem in the GMM estimation.

problem:

$$\inf_{e=(e_1,...,e_n),\theta\in\Theta} (e-c_\tau)' G\widehat{Q}G' (e-c_\tau) \tag{3.3}$$

subject to

$$e_i(-M_i-\epsilon) < Y_i - W_i'\theta \leq (1-e_i) M_i,\ i\in\{1,...,n\}, \tag{3.4}$$

$$e_i \in \{0,1\},\ i\in\{1,...,n\}, \tag{3.5}$$

where $\epsilon$ is a given small and positive real scalar (e.g. $\epsilon = 10^{-6}$ as in our simulation study), and

$$M_i \equiv \max_{\theta\in\Theta} |Y_i - W_i'\theta|,\ i\in\{1,...,n\}. \tag{3.6}$$

Since (3.3) is a quadratic form and matrix $G\widehat{Q}G'$ is positive semi-definite, the objective function is therefore convex in the control variables $e$. Given that these variables take binary values, the formulation (3.3) results in a mixed integer quadratic programming (MIQP) problem.

We now explain the equivalence between (3.2) and (3.3). Note that, for a given value of $\theta\in\Theta$, the sign constraints (3.4) and the dichotomization constraints (3.5) enforce that $e_i = 1\{Y_i \leq W_i'\theta\}$ for $i\in\{1,...n\}$. Therefore, solving the constrained MIQP problem (3.3) is equivalent to solving the GMM estimation problem (3.2). This equivalence enables us to employ the modern MIQP solvers to exactly compute the GMM estimator $\widehat{\theta}_{GMM}(\tau)$. For the implementation, note that the values $(M_i)_{i=1}^n$ in the inequality constraints (3.4) can be computed by formulating the maximization problem in (3.6) as linear programming problems, which can be efficiently solved by modern optimization solvers. Hence these values can be easily computed and stored as the inputs to the MIQP formulation (3.3).

We can perform inference on $\theta(\tau)$ using the GMM estimator $\widehat{\theta}_{GMM}(\tau)$. As noted by Chernozhukov, Hansen, and Jansson (2009), we can take

$$\widehat{Q} = \left[\tau(1-\tau)n^{-1}\sum_{i=1}^n L_i L_i'\right]^{-1} \tag{3.7}$$

as a convenient and natural choice of the GMM weight matrix. By (2.3), this weight matrix equals the inverse of the variance of $n^{-1/2}\sum_{i=1}^n s_{\tau,i}(\theta(\tau))L_i$ conditional on $(L_i)_{i=1}^n$. Let $\varepsilon_\tau \equiv Y - W'\theta(\tau)$. In the GMM estimation (3.2) with $\widehat{Q}$ given by (3.7), it is straightforward to establish via empirical process theory (see e.g., Pakes and Pollard, 1989) that

$$\sqrt{n}(\widehat{\theta}_{GMM}(\tau) - \theta(\tau)) \xrightarrow{d} N(0,\Omega), \tag{3.8}$$

where the asymptotic variance matrix $\Omega$ is given by

$$\Omega = \tau(1-\tau)\left[\Sigma_{WL}\Sigma_{LL}^{-1}\Sigma_{WL}'\right]^{-1}, \Sigma_{WL} = E\left[f_{\varepsilon_\tau}(0|W,Z)\,WL'\right], \Sigma_{LL} = E\left[LL'\right]. \tag{3.9}$$

We can estimate $\Sigma_{LL}$ by the sample analog $\widehat{\Sigma}_{LL} \equiv n^{-1}\sum_{i=1}^{n} L_i L_i'$. Let $\widehat{\varepsilon}_{\tau,i} \equiv Y_i - W_i'\widehat{\theta}_{GMM}(\tau)$. Following Powell (1986), $\Sigma_{WL}$ can be consistently estimated by

$$\widehat{\Sigma}_{WL} \equiv n^{-1}\sum_{i=1}^{n} \left[K\left(\widehat{\varepsilon}_{\tau,i}/h_n\right)/h_n\right] W_i L_i', \tag{3.10}$$

where $K(\cdot)$ is a kernel function and $h_n$ is a bandwidth sequence satisfying that $h_n \longrightarrow 0$ and $\sqrt{n}h_n \longrightarrow \infty$. Specific rule-of-thumb choices of $h_n$ can be based on Koenker (1994). See also Chernozhukov and Hansen (2006, Section 3.4) and Chernozhukov and Hansen (2008, Section 4.4) for the estimation of the IVQR variance components. Based on these results, it is therefore straightforward to construct the confidence interval estimates of $\theta(\tau)$ within the GMM framework.

# 4 Simulation study

In this section, we study the performance of the GMM estimator $\widehat{\theta}_{GMM}(\tau)$ in finite-sample simulations. We used the MATLAB implementation of the Gurobi Optimizer to solve the MIO problems for all numerical results of this paper.[2] All computations were done on a desktop PC (Windows 7) equipped with 32 GB RAM and a CPU processor (Intel i7-5930K) of 3.5 GHz.

We generated $n = 100$ observations from the following simple location scale model:

$$\begin{aligned} Y &= 1 + D_1 + D_2 + D_3 + (0.5 + D_1 + 0.25D_2 + 0.15D_3)\varepsilon, \\ D_1 &= \Phi(Z_1 + v_1), D_2 = 2\Phi(Z_2 + v_2), D_3 = 1.5\Phi(Z_3 + v_3), \end{aligned} \tag{4.1}$$

where $\Phi$ denotes the cumulative distribution function of the standard normal random variable, $Z_1$, $Z_2$ and $Z_3$ are independent standard normal random variables, and $(\varepsilon, v_1, v_2, v_3)$ is generated independently of $(Z_1, Z_2, Z_3)$ from multivariate normal distribution with mean zero and variance $0.25V$ where

$$V = \begin{bmatrix} 1 & 0.4 & 0.6 & -0.2 \\ 0.4 & 1 & 0 & 0 \\ 0.6 & 0 & 1 & 0 \\ -0.2 & 0 & 0 & 1 \end{bmatrix}.$$

[2] The MATLAB codes for computing $\widehat{\theta}_{GMM}(\tau)$ are available from the authors via the website `https://github.com/LeyuChen/IVQR-GMM-computation-codes`. This implementation requires the Gurobi Optimizer, which is freely available for academic purposes.

By Skorohod representation, we can rewrite the model (4.1) as

$$Y = \theta_0(U) + \theta_1(U)D_1 + \theta_2(U)D_2 + \theta_3(U)D_3,$$

where $U = F_\varepsilon(\varepsilon)$ with $F_\varepsilon$ being the cdf of the unobservable $\varepsilon$, and

$$\theta_0(\tau) = 1+0.5F_\varepsilon^{-1}(\tau), \theta_1(\tau) = 1+F_\varepsilon^{-1}(\tau), \theta_2(\tau) = 1+0.25F_\varepsilon^{-1}(\tau), \theta_3(\tau) = 1+0.15F_\varepsilon^{-1}(\tau).$$

We used 500 simulation repetitions for the simulation experiments. In the GMM estimation, we took $W = (1, D_1, D_2, D_3)$ and $L = (1, Z_1, Z_2, Z_3)$. The GMM weight matrix $\widehat{Q}$ was constructed based on (3.7). We set the parameter space $\Theta$ in the MIQP problem (3.3) to be the product of the intervals $[\widehat{\theta}_{j,2SLS} - 10\widehat{\sigma}_{j,2SLS}, \widehat{\theta}_{j,2SLS} + 10\widehat{\sigma}_{j,2SLS}]$, where for $j \in \{0, 1, 2, 3\}$, $\widehat{\theta}_{j,2SLS}$ and $\widehat{\sigma}_{j,2SLS}$, respectively denote the parameter estimate and its estimated heteroskedasticity-robust standard error from the two-stage least square regression of $Y$ on the covariates $W$ using $L$ as the instruments. The value of $\epsilon$ in (3.4) was set to be $10^{-6}$.

We now present the simulation results. First, we report the computational performance of our MIQP algorithm for computing the IVQR GMM estimator. Table 1 gives the summary statistics of the MIQP computation time in CPU seconds across simulation repetitions. From this table, we can see that the MIQP problems (3.3) were solved very efficiently in these simulations which incorporated three endogenous covariates. For the two cases with $\tau \in \{0.25, 0.75\}$, the computation time was comparable. Both cases could be easily solved with the mean and median computation time not exceeding 100 seconds and the maximum time below 200 seconds. The case of $\tau = 0.5$ appeared to be the most computationally demanding but its maximum time remained capped within 17 minutes.

Table 1: MIQP computation time (CPU seconds)

| $\tau$ | mean | min | median | max |
|---|---|---|---|---|
| 0.25 | 94 | 37 | 92 | 197 |
| 0.5 | 348 | 104 | 333 | 989 |
| 0.75 | 86 | 33 | 84 | 186 |

We note that the computation time for the case of $\tau = 0.5$ exceeded that for the other two cases uniformly over the four types of descriptive statistics in Table 1. This can be intuitively explained as follows. For the MIQP problem (3.3), the binary control variables $e_i$ at optimum, denoted as $\widehat{e}_i$, should satisfy that $\widehat{e}_i = 1\{Y_i \leq W_i'\widehat{\theta}_{GMM}(\tau)\}$ for $i \in \{1, ...n\}$. In Appendix A, we show that, for $i \in \{1, ...n\}$, $\widehat{e}_i$ converges in probability to $e_i^* \equiv 1\{Y_i \leq W_i'\theta(\tau)\}$. By (2.3), the indicator $e_i^*$ is a *Bernoulli*$(\tau)$ random variable whose variance is $\tau(1-\tau)$. This variance as a function of $\tau$ is inverted-U shaped with the peak occurring at $\tau = 0.5$, thus suggesting that, for the case of $\tau = 0.5$, the MIQP solver would have more difficult time in predicting and adjusting accordingly its search direction for the optimizers $\widehat{e}_i$.

Following this argument, we further present in Table 2, for a broader range of $\tau$ values, the ratio of a summary statistic of the time for computing $\widehat{\theta}_{GMM}(\tau)$ over that for computing $\widehat{\theta}_{GMM}(0.5)$.[3] It is evident that the results of Table 2 cohere well with our conjecture that, for solving the problem (3.3), the MIQP computation time as a function of $\tau$ is also inverted-U shaped with the most computationally difficult case occurring at $\tau = 0.5$.

Table 2: Computational performance comparison across quantiles relative to median

| $\tau$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| mean ratio | 0.015 | 0.103 | 0.274 | 0.480 | 1 | 0.553 | 0.311 | 0.125 | 0.019 |
| min ratio | 0.031 | 0.172 | 0.497 | 0.652 | 1 | 0.691 | 0.327 | 0.177 | 0.024 |
| median ratio | 0.016 | 0.107 | 0.288 | 0.509 | 1 | 0.547 | 0.317 | 0.130 | 0.020 |
| max ratio | 0.013 | 0.103 | 0.230 | 0.427 | 1 | 0.548 | 0.301 | 0.151 | 0.018 |

We now study the statistical performance of the IVQR GMM estimator. In Table 3, we report the mean and median biases, root mean squared error (RMSE) and median absolute error (MAE) of the GMM estimators $\widehat{\theta}_{GMM}(\tau)$ for $\tau \in \{0.25, 0.5, 0.75\}$. From these results, we find that the GMM estimators performed quite well in terms of estimation bias. Across the three quantile cases, the estimators for $\theta_1(\tau)$ appeared to have larger dispersion in terms of both RMSE and MAE.

It is also interesting to assess how well the finite-sample behavior of the IVQR GMM estimator can be approximated by asymptotic theory. For this purpose, our exact GMM estimator can be used to eliminate the unquantified uncertainty on the solution inaccuracy that might emerge in a heuristic optimization procedure. In Table 4, we estimated the asymptotic standard error based on the formula (3.9) evaluated at true parameter values of the simulation design. This quantity was then compared to standard deviation of $\widehat{\theta}_{GMM}(\tau)$ in simulations. The results of Table 4 indicate that the finite-sample standard error of the GMM estimator in this simulation setup, though being slightly larger, can be well approximated by the asymptotic standard error.

In practice, for carrying out inference, the asymptotic variance of the GMM estimator has to be estimated. We used the Gaussian kernel in the estimation of $\Sigma_{WL}$. The band-

[3] To save computational time, we reduced the number of simulation repetitions to 100 for computing the results of Table 2. Results of all the other tables of Section 4 remained to be based on 500 simulation repetitions.

Table 3: Finite-sample performance of the GMM estimator

| | mean bias | RMSE | median bias | MAE |
|---|---|---|---|---|
| $\theta_0\,(0.25)$ | 0.0109 | 0.2436 | 0.0012 | 0.1643 |
| $\theta_1\,(0.25)$ | -0.0327 | 0.3554 | -0.0048 | 0.2309 |
| $\theta_2\,(0.25)$ | 0.0003 | 0.1642 | 0.0031 | 0.1008 |
| $\theta_3\,(0.25)$ | 0.0064 | 0.2232 | -0.0068 | 0.1522 |
| $\theta_0\,(0.5)$ | 0.0161 | 0.2498 | -0.0037 | 0.1724 |
| $\theta_1\,(0.5)$ | -0.0412 | 0.3241 | -0.0316 | 0.2315 |
| $\theta_2\,(0.5)$ | -0.0012 | 0.1561 | 0.0066 | 0.1038 |
| $\theta_3\,(0.5)$ | 0.0012 | 0.2047 | 0.0031 | 0.1396 |
| $\theta_0\,(0.75)$ | 0.0187 | 0.3046 | 0.0055 | 0.1849 |
| $\theta_1\,(0.75)$ | -0.0358 | 0.3425 | -0.0264 | 0.2235 |
| $\theta_2\,(0.75)$ | -0.0022 | 0.1820 | 0.0062 | 0.1181 |
| $\theta_3\,(0.75)$ | 0.0035 | 0.2393 | -0.0016 | 0.1538 |

Table 4: Comparison with asymptotic approximation

| | standard deviation in simulations | asymptotic standard error |
|---|---|---|
| $\theta_0\,(0.25)$ | 0.2434 | 0.2297 |
| $\theta_1\,(0.25)$ | 0.3539 | 0.3256 |
| $\theta_2\,(0.25)$ | 0.1642 | 0.1572 |
| $\theta_3\,(0.25)$ | 0.2231 | 0.2059 |
| $\theta_0\,(0.5)$ | 0.2493 | 0.2296 |
| $\theta_1\,(0.5)$ | 0.3215 | 0.3049 |
| $\theta_2\,(0.5)$ | 0.1561 | 0.1474 |
| $\theta_3\,(0.5)$ | 0.2047 | 0.1994 |
| $\theta_0\,(0.75)$ | 0.3040 | 0.2744 |
| $\theta_1\,(0.75)$ | 0.3406 | 0.3400 |
| $\theta_2\,(0.75)$ | 0.1820 | 0.1664 |
| $\theta_3\,(0.75)$ | 0.2393 | 0.2283 |

width sequence $h_n$ in (3.10) was based on the Hall-Sheather bandwidth choice, which was suggested by Koenker (1994) and also used by Chernozhukov, Hansen, and Jansson (2009). We also checked the sensitivity of the inference results with respect to this bandwidth choice. Specifically, we reported in Table 5 the finite-sample coverage probabilities of the 95% confidence interval (CI) estimates for $\theta\,(\tau)$, which were constructed based on the normal approximation theory described in Section 3 with three different bandwidth choices: $h_n \in \{0.8h_{n,HS}, h_{n,HS}, 1.2h_{n,HS}\}$, where $h_{n,HS}$ denotes the Hall-Sheather bandwidth sequence. From Table 5, we find that the coverage probabilities results were not very sensitive across bandwidth values although the CI estimates were slightly undersized. We also notice that the CI estimates based on taking $h_n = h_{n,HS}$ or $h_n = 1.2h_{n,HS}$ performed quite well in terms of overall performance.

Table 5: Coverage probabilities (95% CI)

| | $0.8h_{n,HS}$ | $h_{n,HS}$ | $1.2h_{n,HS}$ |
|---|---|---|---|
| $\theta_0\,(0.25)$ | 0.930 | 0.940 | 0.952 |
| $\theta_1\,(0.25)$ | 0.906 | 0.914 | 0.918 |
| $\theta_2\,(0.25)$ | 0.912 | 0.924 | 0.934 |
| $\theta_3\,(0.25)$ | 0.916 | 0.926 | 0.938 |
| $\theta_0\,(0.5)$ | 0.958 | 0.962 | 0.970 |
| $\theta_1\,(0.5)$ | 0.936 | 0.944 | 0.950 |
| $\theta_2\,(0.5)$ | 0.938 | 0.944 | 0.952 |
| $\theta_3\,(0.5)$ | 0.950 | 0.958 | 0.966 |
| $\theta_0\,(0.75)$ | 0.896 | 0.916 | 0.928 |
| $\theta_1\,(0.75)$ | 0.922 | 0.938 | 0.944 |
| $\theta_2\,(0.75)$ | 0.892 | 0.896 | 0.908 |
| $\theta_3\,(0.75)$ | 0.918 | 0.928 | 0.942 |

# 5 An illustrative empirical example: estimating the demand for fish

We illustrate usefulness of our method for exact computation of the IVQR GMM estimator in an empirical study of the demand for fish. We used the dataset constructed by Graddy (1995) on the transactions of whiting in the Fulton fish market in New York. The data were also previously studied in Chernozhukov and Hansen (2008) and Chernozhukov, Hansen, and Jansson (2009) to illustrate the econometric methods developed for quantile regression models with endogeneity. In what follows, we mainly focused on analyzing the results estimated by the MIQP approach and comparing them to the inverse QR estimation results.

The data consist of 111 observations on the price and quantity of whiting transactions aggregated by day. The outcome variable $Y$ is the logarithm of total amount of whitings sold on each day and the endogenous explanatory variable $D$ is the logarithm of the average daily price. The exogenous explanatory variables include the indicators ($Monday$, $Tuesday$, $Wednesday$ and $Thursday$) for days of the week. The instrumental variables are indicators ($Stormy$ and $Mixed$) for weather conditions at sea. These instruments capture the wave height and wind speed, which should affect the supplied quantity of fish and hence the price in the market but are unlikely to influence the demand for fish. See Graddy (1995, 2006) for further details on the operation of the Fulton fish market, and the data and variables used for this study.

Following Chernozhukov, Hansen, and Jansson (2009), we considered the simple demand equation

$$Y = \theta_0 (U) + \theta_1 (U) D \tag{5.1}$$

for the estimation of $\theta_1$, the price elasticity of the demand, which may vary across the demand level $U$. We also augmented the specification (5.1) by incorporating the day effect variables as additional controls, and then performed the estimation. Table 6 presents the estimation results for $\theta_1 (\tau)$ under these two different specifications. For GMM estimation results, we took $L = (1, \mathit{Stormy}, \mathit{Mixed})$ as instruments and configurated the MIQP setting in the same fashion as in Section 4. We used the Gaussian kernel and the Hall-Sheather bandwidth choice for estimating the standard deviation of the GMM estimator and constructing the 95% CI for $\theta_1 (\tau)$. We also performed some sensitivity check and found that the results were not very sensitive to the bandwidth choice. Moreover, we also extracted the inverse QR and the corresponding 95% asymptotic CI estimation results provided by Chernozhukov, Hansen, and Jansson (2009, Table 1) on the same estimating model specifications and listed them in Table 6 for comparison.

We now summarize the results in Table 6. First, we find that, for both model specifications, the point estimates of the demand elasticity were all negative but the magnitudes varied across quantile indices. Moreover, the CI results based on both the GMM and inverse QR methods indicate that the negativity of $\theta_1 (\tau)$ was significant for $\tau \in \{0.25, 0.75\}$ but we could not reject the case of $\theta_1 (\tau)$ being zero at $\tau = 0.5$. We note that, for the same quantile index $\tau$, the GMM and inverse QR estimates of $\theta_1(\tau)$ were somewhat different with the exception that both estimates nearly coincided for the case of $\tau = 0.5$ under the basic specification (5.1). When the day effect variables were included as additional controls, the values of $\widehat{\theta}_1(\tau)$ across these two estimation methods differed to a larger extent in the case of $\tau = 0.25$. The differences between the GMM and inverse QR approaches in the estimation results could be due to the issue that the sample size is not large enough for this empirical application. On the other hand, in the over-identification context where the instruments outnumber the endogenous variables, the asymptotic variance of the inverse QR estimator is generally different from the variance form given by (3.9) (see Chernozhukov and Hansen, 2008, Proposition 2). This fact also helps to explain the differences between the two estimation approaches in the results of Table 6.

Table 6: IVQR estimation of demand elasticity

| | $\tau = 0.25$ | $\tau = 0.5$ | $\tau = 0.75$ |
|---|---|---|---|
| *Specification (5.1)* | | | |
| Estimation method: GMM via the MIQP implementation | | | |
| $\widehat{\theta}_1(\tau)$ | -1.0880 | -0.8876 | -0.9755 |
| std. dev. | 0.4773 | 0.5056 | 0.3027 |
| 95% CI | $(-2.0234, -0.1525)$ | $(-1.8787, 0.1034)$ | $(-1.5689, -0.3822)$ |
| Estimation method: Inverse QR | | | |
| $\widehat{\theta}_1(\tau)$ | -1.3680 | -0.8860 | -1.2685 |
| std. dev. | 0.5704 | 0.4673 | 0.3911 |
| 95% CI | $(-2.486, -0.250)$ | $(-1.802, 0.030)$ | $(-2.035, -0.502)$ |
| *Specification (5.1) augmented with day fixed effects* | | | |
| Estimation method: GMM via the MIQP implementation | | | |
| $\widehat{\theta}_1(\tau)$ | -0.6915 | -0.7152 | -1.0904 |
| std. dev. | 0.3253 | 0.4828 | 0.2465 |
| 95% CI | $(-1.3290, -0.0540)$ | $(-1.6616, 0.2312)$ | $(-1.5735, -0.6074)$ |
| Estimation method: Inverse QR | | | |
| $\widehat{\theta}_1(\tau)$ | -1.3635 | -0.5950 | -1.1790 |
| std. dev. | 0.5304 | 0.4398 | 0.3653 |
| 95% CI | $(-2.403, -0.324)$ | $(-1.457, 0.267)$ | $(-1.895, -0.463)$ |

# 6 Conclusions

In this paper, we have proposed a mixed integer quadratic programming approach for estimating the IVQR model within the GMM framework. Our computational approach can be used to find the exact global solution in the IVQR GMM estimation problem. Modern mixed integer optimization solvers employ branch-and-bound type algorithms which maintain along the solution process both the feasible solutions and lower bounds on the optimal objective function value. For computationally demanding applications, this feature enables us to solve for an approximate IVQR GMM estimator with a guaranteed approximation error bound, thus facilitating the design of an early stopping rule as described in Chen and Lee (2016, Section 4.3). Development of such a theoretically justified early stopping rule for the IVQR estimation problem is therefore a useful further research topic.

One possible application of our approach is panel data quantile regression for group-level treatments (Chetverikov, Larsen, and Palmer, 2016). To deal with group-level unobservables, the estimation procedure in Chetverikov, Larsen, and Palmer (2016) consists of group-by-group quantile regression followed by two-stage least squares. They mention (in their footnote 10) that the latter step could be replaced by an IV median regression, if one is willing to replace the usual assumption that the group-level errors are uncorrelated with instruments with median uncorrelation (Komarova, Severini, and Tamer, 2012). This alternative step can be computed using our computation algorithm. It is an interesting topic for future research to fully develop this alternative to IV quantile regression for group-level treatments.

Our approach is limited to GMM estimators for parametric IVQR models. One may consider semiparametric models with endogeneity. For example, Chen, Linton, and Van Keilegom (2003) considered partially linear median regression with endogenous regressors as one of their examples. Their proposed estimator consists of a two-step procedure: in the first step, nonparametric median regression is carried out given the parameter of interest and in the second step, GMM estimation is implemented with the first step estimates as inputs. Our proposed algorithm is not directly applicable because of the first nonparametric step. It is another interesting topic for future research to develop an algorithm to compute this kind of two-step semiparametric quantile IV estimators.

# A Asymptotic analysis of the binary-valued solutions to the MIQP problem (3.3)

Let $\widehat{e} = (\widehat{e}_i)_{i=1}^n$ denote the solution of the binary control variables $e = (e_i)_{i=1}^n$ to the MIQP problem (3.3). We present in the following theorem the asymptotic properties of these binary optimizers.

**Theorem 1.** *Let $e_i^* \equiv 1\{Y_i \leq W_i'\theta(\tau)\}$ for $i \in \{1,...,n\}$. Suppose the support of the distribution of $W$ is bounded. Then, given (3.8), it holds that $\widehat{e}_i \xrightarrow{p} e_i^*$ for $i \in \{1,...,n\}$.*

*Proof.* Let $\|\cdot\|$ denote the Euclidean norm. Let $\varepsilon_n$ be a sequence which tends to zero at the rate slower than $n^{-1/2}$. By (3.8), we have that

$$P\left(\left\|\widehat{\theta}_{GMM}(\tau) - \theta(\tau)\right\| > \varepsilon_n\right) \longrightarrow 0. \tag{A.1}$$

By (A.1) and the boundedness condition on the distribution of the covariates, it follows that

$$P\left(\max_{i\in\{1,...,n\}}\left|W'\left(\widehat{\theta}_{GMM}(\tau) - \theta(\tau)\right)\right| > \varepsilon_n\right) \longrightarrow 0. \tag{A.2}$$

Note that

$$
\begin{aligned}
& P\left(W_i'\theta(\tau) < Y_i \leq W_i'\widehat{\theta}_{GMM}(\tau)\right) \\
\leq \ & P\left(W_i'\theta(\tau) < Y_i \leq W_i'\widehat{\theta}_{GMM}(\tau), \max_{i\in\{1,...,n\}}\left|W'\left(\widehat{\theta}_{GMM}(\tau)-\theta(\tau)\right)\right| \leq \varepsilon_n\right) \\
& + P\left(\max_{i\in\{1,...,n\}}\left|W'\left(\widehat{\theta}_{GMM}(\tau)-\theta(\tau)\right)\right| > \varepsilon_n\right) \\
\leq \ & P\left(W_i'\theta(\tau) < Y_i \leq W_i'\theta(\tau)+\varepsilon_n\right) + P\left(\max_{i\in\{1,...,n\}}\left|W'\left(\widehat{\theta}_{GMM}(\tau)-\theta(\tau)\right)\right| > \varepsilon_n\right).
\end{aligned}
$$

By (A.2) and the continuity property of probability, we therefore have that $P(W_i'\theta(\tau) < Y_i \leq W_i'\widehat{\theta}_{GMM}(\tau)) \longrightarrow 0$ for $i \in \{1,...,n\}$. Following similar arguments, it is straightforward to see that $P\left(W_i'\widehat{\theta}_{GMM}(\tau) < Y_iW_i' \leq W_i'\theta(\tau)\right)$ also tends to zero for $i \in \{1,...,n\}$. By the analysis of the MIQP formulation (3.3) given in Section 3, we can deduce that $\widehat{e}_i = 1\{Y_i \leq W_i'\widehat{\theta}_{GMM}(\tau)\}$ for $i \in \{1,...n\}$. Using these results, we thus have that

$$
P(\widehat{e}_i \neq e_i^*) = P\left(W_i'\theta(\tau) < Y_i \leq W_i'\widehat{\theta}_{GMM}(\tau)\right) + P\left(W_i'\widehat{\theta}_{GMM}(\tau) < Y_iW_i' \leq W_i'\theta(\tau)\right) \longrightarrow 0.
$$

Hence, the statement $\widehat{e}_i \xrightarrow{p} e_i^*$ holds for $i \in \{1,...,n\}$. □

# B Additional simulation results

## B.1 Finite sample performance comparison of the MIQP based estimation approach to the inverse QR approach

In Appendix B.1, we present simulation results on the comparison of the finite sample performance of our MIQP approach to that of the inverse QR approach. The simulations were based on the data generating design (4.1) as described in Section 4. In the GMM estimation, we took $W = (1, D_1, D_2, D_3)$ and $L = (1, Z_1, Z_2, Z_3)$. For the inverse QR approach, we set the values $\Psi_i$ in (2.5) to be $(Z_{1i}, Z_{2i}, Z_{3i})$ such that both our MIQP based GMM estimator and the inverse QR estimator are asymptotically equivalent in distribution (Chernozhukov and Hansen, 2006, Theorem 3), thus facilitating the performance comparison for these two approaches. For implementation, both our MIQP and the inverse QR approaches require a specification of the parameter space. In this simulation study, we considered two different parameter spaces: $\widehat{\Theta}(10)$ and $\widehat{\Theta}(20)$. Here, $\widehat{\Theta}(t)$ denotes the product of the intervals $[\widehat{\theta}_{j,2SLS} - t\widehat{\sigma}_{j,2SLS}, \widehat{\theta}_{j,2SLS} + t\widehat{\sigma}_{j,2SLS}]$, where for $j \in \{0, 1, 2, 3\}$, $\widehat{\theta}_{j,2SLS}$ and $\widehat{\sigma}_{j,2SLS}$, respectively denote the parameter estimate and its estimated heteroskedasticity-robust standard error from the two-stage least square regression of $Y$ on the covariates $W$ using $L$ as the instruments.

We solved the inverse QR outer optimization problem (2.4) by searching over an evenly spaced grid in a three-dimensional cube induced by the constructed parameter space. The step size along each dimension in the grid search was set to be 0.05. To reduce the computational cost, we set the sample size to be 100 and used 100 simulation repetitions for computing the finite sample performance results.

Table 7 gives the statistical performance results for the inverse QR and the MIQP based GMM estimation approaches for the cases of $\tau \in \{0.25, 0.5, 0.75\}$. From Table 7, we find that both the inverse QR and the MIQP based GMM estimators on the whole performed comparably well. We now compare in Table 8 the computational performance of these two estimation approaches. The results of Table 8 indicate that the computation time for the inverse QR approach was of similar magnitude across the quantile indices. Specifically, the case of $\tau = 0.5$ under this approach appeared to be least computationally intensive. This finding is not surprising because the number of quantile regression subproblems (2.5) in the inverse QR procedure amounts to that of the grid points used for solving (2.4), and solving the standard quantile regression for the upper or lower quantile case is generally more computationally costly than that for the median case. By contrast, the results of Table 8 concerning the MIQP performance echo with those of Table 2 and reveal that solving the MIQP based IVQR GMM estimation problem for the upper and lower quantile cases is far more computationally simpler than that for the median case. We also find from Table 8 that, when we doubled the radius of the parameter space, the increase in computation time under our MIQP approach appeared to be much milder than that under the inverse QR approach. For the latter, this sharp increase in computation time reflected the fact that the number of grid points used in the inverse QR procedure increased exponentially in the radius of the parameter space.

## B.2 Impacts of the sample size on the performance of the MIQP based IVQR GMM estimator

In Appendix B.2, we conduct a simple simulation study to assess impacts of the sample size on the performance of our MIQP based IVQR GMM estimation approach. Specifically, we doubled the sample size used in the simulations described in Section 4 and then investigated how the MIQP computation time would scale up in this setting. For simplicity, we used 100 simulation repetitions. Moreover, we enforced a computation time limit of one hour above which we would terminate the MIQP solver and then take the estimate $\widehat{\theta}_{GMM}(\tau)$ to be the best feasible solution discovered by the solver upon termination. Except for these modifications, the simulation and computing configurations in this simulation study were the same as those used in Section 4.

We now present in Table 9 the summary statistics of the MIQP computation time in CPU seconds and the percentage of the simulated datasets for which the MIQP solver

Table 7: Finite sample performance results for the MIQP and inverse QR approaches

| | Inverse QR | | | | MIQP | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | | median | | mean | | median | |
| | bias | RMSE | bias | MAE | bias | RMSE | bias | MAE |
| Small parameter space ($\widehat{\Theta}\,(10)$) | | | | | | | | |
| $\theta_0\,(0.25)$ | -0.0169 | 0.2167 | -0.0204 | 0.1575 | -0.0131 | 0.2327 | -0.0279 | 0.1682 |
| $\theta_1\,(0.25)$ | 0.0196 | 0.3114 | 0.0715 | 0.1782 | 0.0122 | 0.3146 | 0.0285 | 0.1777 |
| $\theta_2\,(0.25)$ | 0.0275 | 0.1419 | 0.0479 | 0.0945 | 0.0240 | 0.1406 | 0.0272 | 0.0780 |
| $\theta_3\,(0.25)$ | -0.0063 | 0.2148 | 0.0095 | 0.1257 | -0.0071 | 0.2171 | 0.0066 | 0.1588 |
| $\theta_0\,(0.5)$ | 0.0147 | 0.2143 | -0.0017 | 0.1670 | 0.0114 | 0.2180 | 0.0044 | 0.1445 |
| $\theta_1\,(0.5)$ | -0.0241 | 0.2889 | 0.0061 | 0.2102 | -0.0124 | 0.2882 | 0.0495 | 0.2059 |
| $\theta_2\,(0.5)$ | 0.0167 | 0.1406 | 0.0304 | 0.0854 | 0.0152 | 0.1385 | 0.0287 | 0.0832 |
| $\theta_3\,(0.5)$ | -0.0304 | 0.1852 | -0.0414 | 0.1189 | -0.0266 | 0.1806 | -0.0313 | 0.1334 |
| $\theta_0\,(0.75)$ | -0.0046 | 0.2779 | -0.0234 | 0.1934 | -0.0066 | 0.2737 | 0.0067 | 0.1844 |
| $\theta_1\,(0.75)$ | -0.0109 | 0.3455 | 0.0152 | 0.2045 | -0.0167 | 0.3327 | -0.0290 | 0.2317 |
| $\theta_2\,(0.75)$ | 0.0149 | 0.1562 | 0.0251 | 0.1077 | 0.0136 | 0.1571 | 0.0182 | 0.1079 |
| $\theta_3\,(0.75)$ | -0.0120 | 0.2332 | -0.0104 | 0.1674 | -0.0082 | 0.2281 | -0.0013 | 0.1602 |
| Large parameter space ($\widehat{\Theta}\,(20)$) | | | | | | | | |
| $\theta_0\,(0.25)$ | -0.0180 | 0.2173 | -0.0457 | 0.1512 | -0.0075 | 0.2418 | -0.0056 | 0.1705 |
| $\theta_1\,(0.25)$ | 0.0182 | 0.3140 | 0.0483 | 0.1791 | 0.0038 | 0.3156 | 0.0000 | 0.1738 |
| $\theta_2\,(0.25)$ | 0.0270 | 0.1434 | 0.0350 | 0.0903 | 0.0223 | 0.1422 | 0.0167 | 0.0865 |
| $\theta_3\,(0.25)$ | -0.0026 | 0.2150 | -0.0026 | 0.1313 | -0.0103 | 0.2258 | -0.0044 | 0.1705 |
| $\theta_0\,(0.5)$ | 0.0207 | 0.2189 | -0.0053 | 0.1625 | 0.0110 | 0.2196 | 0.0116 | 0.1523 |
| $\theta_1\,(0.5)$ | -0.0325 | 0.2864 | 0.0109 | 0.2215 | -0.0042 | 0.2935 | 0.0379 | 0.2093 |
| $\theta_2\,(0.5)$ | 0.0167 | 0.1411 | 0.0326 | 0.0961 | 0.0172 | 0.1374 | 0.0168 | 0.0896 |
| $\theta_3\,(0.5)$ | -0.0316 | 0.1852 | -0.0394 | 0.1259 | -0.0265 | 0.1801 | -0.0233 | 0.1193 |
| $\theta_0\,(0.75)$ | -0.0035 | 0.2734 | -0.0294 | 0.1891 | 0.0015 | 0.2713 | 0.0058 | 0.1982 |
| $\theta_1\,(0.75)$ | -0.0128 | 0.3525 | 0.0105 | 0.2215 | -0.0166 | 0.3410 | 0.0290 | 0.2305 |
| $\theta_2\,(0.75)$ | 0.0154 | 0.1545 | 0.0283 | 0.1039 | 0.0138 | 0.1561 | 0.0338 | 0.0940 |
| $\theta_3\,(0.75)$ | -0.0112 | 0.2299 | -0.0057 | 0.1643 | -0.0156 | 0.2326 | -0.0279 | 0.1710 |

Table 8: Computation time (CPU seconds) for the MIQP and inverse QR approaches

| | Inverse QR | | | | MIQP | | | |
|---|---|---|---|---|---|---|---|---|
| $\tau$ | mean | min | median | max | mean | min | median | max |
| Small parameter space ($\widehat{\Theta}\,(10)$) | | | | | | | | |
| 0.25 | 392 | 146 | 363 | 959 | 99 | 39 | 100 | 197 |
| 0.5 | 331 | 119 | 304 | 806 | 374 | 107 | 350 | 777 |
| 0.75 | 386 | 151 | 359 | 958 | 91 | 38 | 89 | 186 |
| Large parameter space ($\widehat{\Theta}\,(20)$) | | | | | | | | |
| 0.25 | 3074 | 1174 | 2823 | 7441 | 132 | 64 | 129 | 201 |
| 0.5 | 2584 | 952 | 2401 | 6254 | 497 | 219 | 485 | 779 |
| 0.75 | 3101 | 1204 | 2848 | 7645 | 156 | 64 | 153 | 317 |

converged and $\widehat{\theta}_{GMM}\,(\tau)$ was exactly computed within the specified time limit. From Table 9, we find that the solver converged to a global solution for all simulations with $\tau \in \{0.25, 0.75\}$. However, the case of $\tau = 0.5$ remained the most computationally demanding and the solver for computing this case could not converge within the one-hour time limit in 36% of the simulation repetitions. Comparing the computational performance results in Tables 1 and 9, we also notice that, when we doubled the sample size, the time for computing the MIQP problem (3.3) could scale up more than ten times. We note that the computational scalability issue revealed in these results is not unexpected because the MIQP problem is known to be in the class of *NP* (Non-deterministic polynomial time) complete problems (see e.g., Del Pia, Dey, and Molinaro, 2017). Namely, the computational complexity for solving the MIQP problem may increase rapidly in the problem size. However, in view of the advance of modern computing technology, it is expected that in near future, the MIQP approach will become more tractable and computationally attractive for solving the exact IVQR GMM computation problems with moderately large sample size. Moreover, the MIQP solution method is based on branch-and-bound type algorithms, which themselves can be well adapted to the parallel computing framework thereby enhancing the computational performance for solving large scale problem instances. See e.g., Gendron and Crainic (1994) for a classic survey on parallel branch-and-bound algorithms.

In Table 10, we report the statistical performance of the MIQP based IVQR GMM estimator in this simulation setting. Comparing this table to Table 3 of Section 4, we find that the precision of the estimate $\widehat{\theta}_{GMM}\,(\tau)$ generally increased with the sample size. It is worth noting that the statistical performance of the MIQP based estimation approach remained quite good for the case of $\tau = 0.5$ even though only 64% of the MIQP problems (3.3) in the simulation for this estimation case could be exactly solved within the specified

Table 9: MIQP computational performance statistics (sample size $n = 200$)

| $\tau$ | mean | min | median | max | solver convergence |
|---|---|---|---|---|---|
| 0.25 | 1009 | 490 | 920 | 2739 | 100% |
| 0.5 | 3007 | 1654 | 3120 | 3600 | 64% |
| 0.75 | 1243 | 506 | 1185 | 2452 | 100% |

time limit. This result echoes with the findings in the MIO literature that the MIO solver could often locate the exact solution quickly yet it can take a longer time for the solver to verify that the solution status is indeed exact (see e.g., Florios and Skouras, 2008). In Table 11, we further compare the estimator bias and dispersion results for the cases where the MIQP solver could not converge within the one-hour time limit and those where the solver found the global solution within this time limit. These results also suggest that, even in the presence of severe computation constraints, we may still get satisfactorily approximate IVQR GMM estimates via the MIQP approach.

Table 10: Statistical performance of the GMM estimator (sample size $n = 200$)

| | mean | | median | |
|---|---|---|---|---|
| | bias | RMSE | bias | MAE |
| $\theta_0\,(0.25)$ | -0.0160 | 0.1637 | -0.0059 | 0.1177 |
| $\theta_1\,(0.25)$ | 0.0186 | 0.2765 | 0.0279 | 0.2127 |
| $\theta_2\,(0.25)$ | -0.0001 | 0.1140 | -0.0014 | 0.0702 |
| $\theta_3\,(0.25)$ | 0.0175 | 0.1576 | 0.0133 | 0.0895 |
| $\theta_0\,(0.5)$ | -0.0047 | 0.1693 | -0.0130 | 0.1272 |
| $\theta_1\,(0.5)$ | -0.0012 | 0.2325 | 0.0094 | 0.1466 |
| $\theta_2\,(0.5)$ | 0.0024 | 0.0961 | 0.0165 | 0.0652 |
| $\theta_3\,(0.5)$ | 0.0132 | 0.1529 | 0.0265 | 0.1079 |
| $\theta_0\,(0.75)$ | 0.0185 | 0.1891 | 0.0134 | 0.0995 |
| $\theta_1\,(0.75)$ | -0.0194 | 0.2425 | -0.0137 | 0.1572 |
| $\theta_2\,(0.75)$ | 0.0024 | 0.1204 | 0.0104 | 0.0682 |
| $\theta_3\,(0.75)$ | -0.0127 | 0.1594 | 0.0077 | 0.0999 |

# C Alternative MIO based formulations of the IVQR GMM estimation problem

In Appendix C, we provide two alternative MIO based formulations of the IVQR GMM estimation problem. These alternative formulations complement the main MIQP formulation (3.3) of this paper.

Table 11: Performance comparison for the convergence and early termination cases

| | Solver did not converge within one hour | | | | Solver converged within one hour | | | |
|---|---|---|---|---|---|---|---|---|
| | mean | | median | | mean | | median | |
| | bias | RMSE | bias | MAE | bias | RMSE | bias | MAE |
| $\theta_0\,(0.5)$ | 0.0250 | 0.1668 | 0.0317 | 0.1067 | -0.0213 | 0.1706 | -0.0312 | 0.1412 |
| $\theta_1\,(0.5)$ | -0.0352 | 0.2672 | -0.0173 | 0.2031 | 0.0180 | 0.2105 | 0.0350 | 0.1328 |
| $\theta_2\,(0.5)$ | -0.0035 | 0.0925 | 0.0076 | 0.0672 | 0.0057 | 0.0982 | 0.0239 | 0.0647 |
| $\theta_3\,(0.5)$ | 0.0020 | 0.1620 | 0.0185 | 0.0965 | 0.0195 | 0.1475 | 0.0412 | 0.1082 |

## C.1 MIQP formulation based on special ordered set constraints

Suppose that the distribution of $Y$ conditional on $W$ is absolutely continuous with respect to Lebesgue measure. Under this assumption, we can construct another MIQP based formulation of the problem (3.2), which does not exploit the quantities $(M_i)_{i=1}^n$ of (3.6) and can hence be robust to numerical computation issues that may arise when $(M_i)_{i=1}^n$ take extremely large values.

Recall that $c_\tau = (\tau, ..., \tau)$ denotes the $n$ dimensional vector of the same constant values specified by the quantile index $\tau$. If the outcome $Y$ is continuously distributed conditional on $W$, we note that the optimization problem (3.2) can be equivalently formulated as the following MIQP problem:

$$\inf_{e=(e_i)_{i=1}^n, r=(r_i)_{i=1}^n, s=(s_i)_{i=1}^n, \theta\in\Theta} (e-c_\tau)' G\widehat{Q}G' (e-c_\tau) \tag{C.1}$$

$$\text{s.t. (3.5) and}$$

$$r_i - s_i = Y_i - W_i'\theta,\ i \in \{1, ..., n\}, \tag{C.2}$$

$$(r_i, e_i) : \text{SOS-1},\ i \in \{1, ..., n\}, \tag{C.3}$$

$$(s_i, 1-e_i) : \text{SOS-1},\ i \in \{1, ..., n\}, \tag{C.4}$$

$$r_i + s_i > 0,\ i \in \{1, ..., n\}, \tag{C.5}$$

$$r_i \geq 0,\ s_i \geq 0,\ i \in \{1, ..., n\}, \tag{C.6}$$

where, for a pair of variables $(x, y)$, the statement $(x, y)$ : SOS-1 is a shorthand statement for the condition that these two variables should be subject to the constraint of Type 1 special ordered set (SOS-1), which means at most one of them can take non-zero value.

We now explain the equivalence between the problems (3.2) and (C.1). By (C.5) and (C.6), the values $r_i$ and $s_i$ are non-negative but cannot be both zero. Thus, by (3.5) and the SOS-1 constraints (C.3) and (C.4), it follows that

$$\begin{aligned} e_i &= 1 \iff r_i = 0 \text{ and } s_i > 0, && \text{(C.7)} \\ e_i &= 0 \iff r_i > 0 \text{ and } s_i = 0. && \text{(C.8)} \end{aligned}$$

By continuity of the distribution of $Y$ given $W$, for any given value of $\theta \in \Theta$, the difference $Y - W'\theta$ is almost surely non-zero. Using this continuity assumption, and (C.2), (C.7) and (C.8), we can deduce that the equations $e_i = 1\{Y_i \leq W_i'\theta\}$ for $i \in \{1, ...n\}$ also hold almost surely. Therefore, the MIQP problem (C.1) is almost surely equivalent to the GMM estimation problem (3.2).

We note that mixed integer linear and quadratic optimization problems subject to SOS-1 constraints can be exactly solved via the branch-and-bound type algorithms. Hence, modern MIO solvers can also be used to compute the exact solution to the MIQP problem (C.1). For implementation, we replace (C.5) by $r_i + s_i \geq \eta$, where $\eta$ is a small positive number (e.g. $\eta = 10^{-5}$ as in our numerical study) that is larger than the solver's constraint feasibility tolerance level. This ensures that strict positivity of $r_i + s_i$ is well enacted in numerical computation.

## C.2 MILP formulation based on linearization of the IVQR GMM objective function

We now present another MIO based formulation of the problem (3.2). This formulation is based on linearizing the MIQP objective function of (3.3) and thereby results in a mixed integer linear programming (MILP) problem.[4]

Let $\widehat{\Gamma} \equiv G\widehat{Q}G'$. Note that, for $e = (e_i)_{i=1}^n \in \{0,1\}^n$,

$$\begin{aligned} & (e - c_\tau)' \, \widehat{\Gamma} \, (e - c_\tau) \\ = \; & \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \widehat{\Gamma}_{ij} e_i e_j - 2e'\widehat{\Gamma} c_\tau + c_\tau' \widehat{\Gamma} c_\tau \\ = \; & e' \left( diag(\widehat{\Gamma}) - 2\widehat{\Gamma} c_\tau \right) + 2 \sum\nolimits_{i=1}^{n-1} \sum\nolimits_{j=i+1}^{n} \widehat{\Gamma}_{ij} e_i e_j + c_\tau' \widehat{\Gamma} c_\tau, \qquad \text{(C.9)} \end{aligned}$$

where $diag(\widehat{\Gamma})$ denotes the diagonal vector of $\widehat{\Gamma}$. By (C.9), we can thus rewrite the MIQP

[4] We thank one referee for pointing out this MILP formulation of the IVQR GMM estimation problem.

problem (3.3) as the following MILP problem:

$$\inf_{e=(e_i)_{i=1}^n,(x_{ij})_{1\le i<j\le n},\theta\in\Theta} e'\left(diag(\widehat{\Gamma})-2\widehat{\Gamma}c_\tau\right)+2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\widehat{\Gamma}_{ij}x_{ij}+c_\tau'\widehat{\Gamma}c_\tau \quad \text{(C.10)}$$

s.t. (3.4), (3.5) and

$$x_{ij}\le e_i,\ 1\le i<j\le n, \quad \text{(C.11)}$$

$$x_{ij}\le e_j,\ 1\le i<j\le n, \quad \text{(C.12)}$$

$$e_i+e_j-x_{ij}\le 1,\ 1\le i<j\le n, \quad \text{(C.13)}$$

$$x_{ij}\in\{0,1\},\ 1\le i<j\le n. \quad \text{(C.14)}$$

It is straightforward to see that $x_{ij}=e_ie_j$ under the dichotomization constraints (3.5) and (C.14), and inequality constraints (C.11), (C.12) and (C.13). Hence, it follows that the MILP problem (C.10) is equivalent to the MIQP problem (3.3) and therefore also equivalent to the IVQR GMM estimation problem (3.2).

## C.3 Computational performance comparison of the MIO formulations (3.3), (C.1) and (C.10)

The MIQP formulation (C.1) does not require computation of the bounding quantities $(M_i)_{i=1}^n$, albeit at the cost of incurring $2n$ additional control variables in the optimization problem. While optimization over a linear objective function can be computationally simpler, the MILP formulation (C.10) consists of $2n+3n(n-1)/2$ inequality constraints and $n+n(n-1)/2$ binary controls whereas the MIQP formulation (3.3) has $2n$ inequality constraints and only $n$ binary controls. Therefore, compared to the formulation (3.3), the computational performance of the alternative formulations (C.1) and (C.10) can be more compromised as the sample size gets large.

We assess the computational performance of the MIO formulations (3.3), (C.1) and (C.10) in a small scale simulation study. We use the same simulation data generating design and computing configuration as in Section 4. To save computational time, we used 100 simulation repetitions for which each simulated dataset contained $n=40$ observations. From the simulation results, we note that the optimal objective function values in all the three MIO formulations are identical across all the simulation repetitions. Moreover, these minimized objective function values are identical to the GMM objective function values of (3.2) evaluated at $\widehat{\theta}_{GMM}(\tau)$ computed via these three MIO formulations. These numerical results consist with the mathematical equivalence between the formulations (3.2), (C.1) and (C.10). However, as reported in Table 12, the MIQP formulation (3.3) substantially outperformed the other two MIO formulations in terms of CPU seconds required for computing the estimates $\widehat{\theta}_{GMM}(\tau)$.

Table 12: Computation time (CPU seconds)

| | MIQP (3.3) | | | MIQP (C.1) | | | MILP (C.10) | | |
|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | 0.25 | 0.5 | 0.75 | 0.25 | 0.5 | 0.75 | 0.25 | 0.5 | 0.75 |
| mean | 1.08 | 2.75 | 0.95 | 29.22 | 110.87 | 27.05 | 38.12 | 183.73 | 57.56 |
| min | 0.37 | 0.99 | 0.39 | 4.90 | 8.47 | 4.01 | 23.87 | 122.94 | 20.68 |
| median | 1.08 | 2.59 | 0.83 | 28.29 | 105.19 | 26.88 | 37.19 | 183.87 | 57.46 |
| max | 2.31 | 5.44 | 1.91 | 51.84 | 209.21 | 93.00 | 66.91 | 254.44 | 98.75 |

# References

Achterberg, T., and R. Wunderling (2013): "Mixed integer programming: Analyzing 12 years of progress," in *Facets of combinatorial optimization*, pp. 449–481. Springer.

Andrews, D. W. (1997): "A stopping rule for the computation of generalized method of moments estimators," *Econometrica*, pp. 913–931.

Bertsimas, D., A. King, and R. Mazumder (2016): "Best subset selection via a modern optimization lens," *Annals of Statistics*, 44(2), 813–852.

Bertsimas, D., and R. Weismantel (2005): *Optimization over integers*, vol. 13. Dynamic Ideas Belmont.

Bilias, Y., K. Florios, and S. Skouras (2013): "Exact Computation of Censored Least Absolute Deviations Estimators," Available at SSRN: http://ssrn.com/abstract=2372588.

Chen, L.-Y., and S. Lee (2016): "Best Subset Binary Prediction," arXiv:1610.02738.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.

Chernozhukov, V., and C. Hansen (2004): "The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis," *Review of Economics and statistics*, 86(3), 735–751.

——— (2005): "An IV model of quantile treatment effects," *Econometrica*, 73(1), 245–261.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132(2), 491–525.

——— (2008): "Instrumental variable quantile regression: A robust inference approach," *Journal of Econometrics*, 142(1), 379–398.

——— (2013): "Quantile Models with Endogeneity," *Annual Review of Economics*, 5(1), 57–81.

Chernozhukov, V., C. Hansen, and M. Jansson (2007): "Inference approaches for instrumental variable quantile regression," *Economics Letters*, 95(2), 272–277.

——— (2009): "Finite sample inference for quantile regression models," *Journal of Econometrics*, 152(2), 93–103.

Chernozhukov, V., and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115(2), 293–346.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade," *Econometrica*, 84(2), 809–833.

Del Pia, A., S. S. Dey, and M. Molinaro (2017): "Mixed-integer quadratic programming is in NP," *Mathematical Programming*, 162(1-2), 225–240.

Florios, K., and S. Skouras (2008): "Exact computation of max weighted score estimators," *Journal of Econometrics*, 146(1), 86–91.

Gendron, B., and T. G. Crainic (1994): "Parallel branch-and-branch algorithms: Survey and synthesis," *Operations research*, 42(6), 1042–1066.

Graddy, K. (1995): "Testing for imperfect competition at the Fulton fish market," *The RAND Journal of Economics*, pp. 75–92.

——— (2006): "Markets: the Fulton fish market," *The Journal of Economic Perspectives*, 20(2), 207–220.

Honoré, B. E., and L. Hu (2004): "On the performance of some robust instrumental variables estimators," *Journal of Business and Economic Statistics*, 22(1), 30–39.

Jünger, M., T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey (2009): *50 years of integer programming 1958-2008: From the early years to the state-of-the-art*. Springer Science & Business Media.

Kaplan, D. M., and Y. Sun (2015): "Smoothed estimating equations for instrumental variables quantile regression," *Econometric Theory*, pp. 1–53.

Kitagawa, T., and A. Tetenov (2015): "Who should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," Cemmap Working Paper CWP10/15.

Koenker, R. (1994): "Confidence intervals for regression quantiles," in *Asymptotic statistics*, pp. 349–359. Springer.

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

Komarova, T., T. Severini, and E. Tamer (2012): "Quantile uncorrelation and instrumental regressions," *Journal of Econometric Methods*, 1(1), 2–14.

Nemhauser, G. L., and L. A. Wolsey (1999): *Integer and combinatorial optimization*. Wiley-Interscience.

Newey, W. K., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, pp. 1027–1057.

Powell, J. L. (1986): "Censored regression quantiles," *Journal of econometrics*, 32(1), 143–155.

Xu, G., and S. Burer (2017): "A branch-and-bound algorithm for instrumental variable quantile regression," *Mathematical Programming Computation*, pp. 1–27.