DUAL REGRESSION

RICHARD H. SPADY[†] AND SAMI STOULI[§]

Abstract. We propose dual regression as an alternative to the quantile regression process for the global estimation of conditional distribution functions under minimal assumptions. Dual regression provides all the interpretational power of the quantile regression process while avoiding the need for repairing the intersecting conditional quantile surfaces that quantile regression often produces in practice. Our approach introduces a mathematical programming characterization of conditional distribution functions which, in its simplest form, is the dual program of a simultaneous estimator for linear location-scale models. We apply our general characterization to the specification and estimation of a flexible class of conditional distribution functions, and present asymptotic theory for the corresponding empirical dual regression process.

keywords: Conditional distribution; Duality; Monotonicity; Quantile regression; Method of moments; Mathematical programming; Convex approximation.

1. Introduction

Let Y be a continuously distributed random variable and X a random vector. Then the conditional distribution function of Y given X, written $U = F_{Y|X}(Y \mid X)$, has three properties: U is standard uniform, U is independent of X, and $F_{Y|X}(y \mid x)$ is strictly increasing in y for any value x of X. We will refer to these three properties as uniformity, independence and monotonicity. For some specified mean zero and unit variance distribution function F with support the real line and inverse function F^{-1}, define $\varepsilon = F^{-1}\{F_{Y|X}(Y \mid X)\}$. Then ε satisfies independence and monotonicity, has distribution F, and is transformed to uniformity by taking $U = F(\varepsilon)$.

If we have a sample of n points $\{(x_i, y_i)\}_{i=1}^n$ drawn from the joint distribution $F_{YX}(Y, X)$, how might we estimate the n values $\varepsilon_i = F^{-1}\{F_{Y|X}(y_i \mid x_i)\}$ using only the requirement that the estimate displays independence and monotonicity, and has distribution F? We explore this question by formulating a sequence of mathematical programming problems that embodies

This version: November 13, 2017. First ArXiv version: 25 October 2012. We are indebted to Andrew Chesher for many fruitful discussions and to Roger Koenker for encouragement at a key early stage. We also thank Dennis Kristensen, Lars Nesheim, David Pacini, Jelmer Ypma, Yanos Zylberberg, the editor, the associate editor, two anonymous referees and participants to seminars and conferences for helpful comments that considerably improved the paper. Sami Stouli gratefully acknowledges the financial support of the UK Economic and Social Research Council and of the Royal Economic Society.

[†] Department of Economics, Johns Hopkins University, rspady@jhu.edu.

§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

these requirements, with each element of this sequence providing an asymptotically valid characterization of an increasingly flexible class of conditional distribution functions.

The use of dual is thus motivated by the general observation that the estimation problem for a conditional distribution function $F_{Y|X}$ indexed by a parameter θ is usually formulated in terms of a procedure that obtains θ directly and $F_{Y|X}$ as a byproduct that follows from a calculation from the representation evaluated at a specific value of θ. A classical example is the linear location shift model $F_{Y|X}(y_i \mid x_i) = F\{(y_i - \beta^{\mathrm{T}} x_i)/\sigma\}$, for which the parameter vector $\theta = (\beta, \sigma)^{\mathrm{T}}$ needs to be estimated in order to obtain the n values $\varepsilon_i = (y_i - \beta^{\mathrm{T}} x_i)/\sigma$. Here we turn that process around, obtaining ε_i first (from a mathematical programming problem) and backing out θ afterwards, if at all.

In its simplest form, dual regression augments the median regression dual programming problem (Koenker & Bassett, 1978) with global second moment orthogonality constraints, while expanding the support of parameter values from the unit interval to the real line. Adding further global orthogonality constraints gives rise to a sequence of augmented, generalized dual regression programs. Although each of these programs seeks only to find the n values $\varepsilon_i = F^{-1}(u_i)$, their first-order conditions show that the assignment of these n values corresponds to a sequence of augmented location-scale representations, the simplest element of which is a linear heteroscedastic model. Moreover, their second-order conditions are equivalent to monotonicity, so optimal dual regression solutions are free of quantile-crossing problems.

To each element of the sequence of dual programs corresponds a convex primal problem, both nontrivial to determine and difficult to implement, the convexity of which guarantees uniqueness of optimal dual regression solutions. For a given specification of $F_{Y|X}(Y \mid X)$, the first-order conditions of the corresponding primal problem also describe necessary and sufficient conditions for independence of the associated dual solutions. Thus our dual formulation reveals a sequence of convex optimization problems, gives a feasible and direct implementation of each of them, and uniquely characterizes the family of associated globally monotone representations, which can then be used as complete estimates of a flexible class of conditional distribution functions.

2. Basics

2.1. **The dual regression problem.** We introduce the basic principles underlying our general method by first providing a new characterization of the conditional distribution function $F_{Y|X}(Y \mid X)$ associated with the linear location-scale model

$$Y = \beta_1^{\mathrm{T}} X + (\beta_2^{\mathrm{T}} X)\varepsilon, \quad \beta_2^{\mathrm{T}} X > 0, \quad \varepsilon \mid X \sim F, \tag{2.1}$$

where X is a $K \times 1$ vector of explanatory variables including an intercept, and F a mean zero and unit variance cumulative distribution function over the real line.

Suppose that we observe a sample of n identically and independently distributed realizations $\{(y_i, x_i)\}_{i=1}^n$ generated according to model (2.1). The primary population target of our analysis is

$$\varepsilon_i = \frac{y_i - \beta_1^{\mathrm{T}} x_i}{\beta_2^{\mathrm{T}} x_i} = F^{-1}\{F_{Y|X}(y_i \mid x_i)\} \quad (i = 1, \ldots, n),$$

knowledge of which is equivalent to knowledge of the n values $F_{Y|X}(y_i \mid x_i)$ up to the monotone transformation F.

Let $\lambda = (\lambda_1, \lambda_2)^{\mathrm{T}} \in \mathbb{R}^{2\times K}$ and $e_o \in \mathbb{R}^n$ satisfy the system of n equations and n inequality constraints

$$y_i = \lambda_1^{\mathrm{T}} x_i + (\lambda_2^{\mathrm{T}} x_i) e_{oi}, \quad \lambda_2^{\mathrm{T}} x_i > 0 \quad (i = 1, \ldots, n), \tag{2.2}$$

where e_o further satisfies the $2\times K$ orthogonality conditions $\sum_{i=1}^n x_i e_{oi} = 0$ and $\sum_{i=1}^n x_i(e_{oi}^2 - 1) = 0$. Since x_i includes an intercept, the sample moments of e_o and e_o^2 are 0 and 1, and e_o and e_o^2 are orthogonal to each column of the $n \times K$ matrix $(x_1, \ldots, x_n)^{\mathrm{T}}$ of explanatory variables. We propose a characterization of the sequence of vectors e_o that satisfy representation (2.2) and the associated orthogonality constraints for each n. The corresponding sequence of empirical distribution functions then provides an asymptotically valid characterization of the conditional distribution function $F_{Y|X}(Y \mid X)$ corresponding to the data-generating process (2.1). As a by-product of this approach, we simultaneously obtain a characterization of the parameter vector λ in (2.2), which then provides a consistent estimator of the population parameter β in (2.1).

For each x_i, with the scale function $\lambda_2^{\mathrm{T}} x_i > 0$, y_i is an increasing function of e_{oi}, and to representation (2.2) corresponds a convex function

$$C(x_i, e_{oi}, \lambda) = \int_0^{e_{oi}} \{\lambda_1^{\mathrm{T}} x_i + (\lambda_2^{\mathrm{T}} x_i)s\} ds = (\lambda_1^{\mathrm{T}} x_i) e_{oi} + \frac{1}{2}(\lambda_2^{\mathrm{T}} x_i) e_{oi}^2 \quad (e_{oi} \in \mathbb{R}),$$

and whose quadratic form corresponds to a location-scale representation for $F_{Y|X}(Y \mid X)$. Letting y be the $n \times 1$ vector of dependent variable values, and assuming knowledge of λ and e_o, we consider assigning a value e_i to each observation in the sample by maximizing the correlation between y and $e = (e_1, \ldots, e_n)^{\mathrm{T}}$ subject to a constraint that embodies the properties of e_o:

$$\max_{e \in \mathbb{R}^n} \left\{ y^{\mathrm{T}} e : \sum_{i=1}^n C(x_i, e_i, \lambda) = \sum_{i=1}^n C(x_i, e_{oi}, \lambda) \right\}. \tag{2.3}$$

Problem (2.3) describes the assignment of e values to y values in a sample generated according to a location-scale model, and it admits $e = e_o$ as its only solution. Since e_o and λ are unknown, the assignment problem (2.3) is infeasible: we thus introduce the equivalent, feasible formulation

$$\text{(D)} \quad \max_{e \in \mathbb{R}^n} \left\{ y^{\mathrm{T}} e : \sum_{i=1}^n x_i e_i = 0,\ \frac{1}{2}\sum_{i=1}^n x_i(e_i^2 - 1) = 0 \right\},$$

the dual regression program.

2.2. **Solving the dual program.** The solution to (D) is easily found from the Lagrangian

$$\mathscr{L} = \sum_{i=1}^{n} y_i e_i - \lambda_1 \sum_{i=1}^{n} x_i e_i - \frac{1}{2}\lambda_2 \sum_{i=1}^{n} x_i (e_i^2 - 1).$$

Differentiating with respect to e_i, we obtain n first-order conditions:

$$\frac{\partial \mathscr{L}}{\partial e_i} = y_i - \lambda_1^{\mathrm{T}} x_i - (\lambda_2^{\mathrm{T}} x_i) e_i = 0 \quad (i = 1, \ldots, n).$$

Upon rearranging we obtain the closed-form solution

$$e_i = \frac{y_i - \lambda_1^{\mathrm{T}} x_i}{\lambda_2^{\mathrm{T}} x_i} \quad (i = 1, \ldots, n),$$

which is of the location-scale form $e_i = \{y_i - \mu(x_i)\}/\sigma(x_i)$, with $\mu(x_i)$ and $\sigma(x_i)$ linear in x_i.

Another view is obtained by writing the first-order conditions as

$$y_i = \lambda_1^{\mathrm{T}} x_i + (\lambda_2^{\mathrm{T}} x_i) e_i \quad (i = 1, \ldots, n), \tag{2.4}$$

a linear location-scale representation, with corresponding quantile regression representation

$$y_i = (\lambda_1 + \lambda_2 e_i)^{\mathrm{T}} x_i = \{\lambda_1 + \lambda_2 F_n^{-1}(u_i)\}^{\mathrm{T}} x_i \equiv \beta(u_i)^{\mathrm{T}} x_i \quad (i = 1, \ldots, n). \tag{2.5}$$

Program (D) thus provides a complete characterization of linear representations of the form (2.4) and (2.5), as they arise from its first-order conditions. Moreover, the parameters of these representations are the Lagrange multipliers λ_1 and λ_2 of an optimization problem with solution $e = e_o$.

The quantile regression representation of the first-order conditions of (D) sheds additional light on the monotonicity property of dual regression solutions, when there are no repeated X values. For $u, u' \in (0,1)$, $u' > u$, the no-crossing property of conditional quantiles requires

$$\beta(u')^{\mathrm{T}} x_i - \beta(u)^{\mathrm{T}} x_i > 0, \quad (i = 1, \ldots, n),$$

which is satisfied if $\lambda_2^{\mathrm{T}} x_i$ is strictly positive for each i, and coincides with the n second-order conditions of program (D):

$$\frac{\partial^2 \mathscr{L}}{\partial e_i \partial e_i} = -\lambda_2^{\mathrm{T}} x_i < 0, \quad (i = 1, \ldots, n).$$

Therefore, an optimal e solution that violates the monotonicity property is ruled out by the requirement that for an observation with X value x_i, the ordering of the Y values $\beta(u')^{\mathrm{T}} x_i$ and $\beta(u)^{\mathrm{T}} x_i$ must correspond to the ordering of the u values. Hence the correlation criterion of system (D) suffices to impose monotonicity, with optimality of a solution then being equivalent to monotonicity at the n sample points. Dual regression is thus able to incorporate this property in the estimation procedure, which facilitates extrapolation beyond the empirical support of X, and yields significant finite-sample improvements in the estimation of conditional quantile functions as illustrated by our simulations in §5.2.

2.3. **Formal duality.** By Lagrangian duality arguments (Boyd & Vandenberghe, 2004, Chapter 5), the objective function of the dual of problem (D) is

$$Q_n(\lambda) = \sup_{e \in \mathbb{R}^n} y^{\mathrm{T}} e - \sum_{i=1}^{n} \left\{ C\left(x_i, e_i, \lambda\right) - C\left(x_i, e_{oi}, \lambda\right) \right\},$$

defined for all $\lambda \in \Lambda_0$, where $\Lambda_0 = \Lambda_1 \times \Lambda_2$, with $\Lambda_1 = \mathbb{R}^K$ and $\Lambda_2 = \{\lambda_2 \in \mathbb{R}^K : \inf_{i \leq n} \lambda_2^{\mathrm{T}} x_i > 0\}$. Under the conditions of Theorem 1 below, $Q_n(\lambda)$ has a closed-form expression, is strictly convex over Λ_0, and minimizing $Q_n(\lambda)$ over Λ_0 is equivalent to solving (D). Given a vector $\omega \in \mathbb{R}^n$, we let $\mathrm{diag}(\omega_i)$ denote the $n \times n$ diagonal matrix with diagonal elements $\omega_1, \ldots, \omega_n$.

Condition 1. *The random variable Y is continuously distributed conditional on X, with conditional density $f_{Y|X}(y \mid X)$ bounded away from 0.*

Condition 2. *For a specified vector $\omega \in \mathbb{R}^n$, the matrix $diag(\omega_i)$ is nonsingular and the matrix $\sum_{i=1}^n \omega_i^{-1} x_i x_i^T = M_n$ is finite, positive definite, and has rank K.*

Condition 3. *There exists $(\lambda, e_o) \in \Lambda_0 \times \mathbb{R}^n$ such that $y_i = \lambda_1^T x_i + (\lambda_2^T x_i) e_{oi}$ with $\inf_{i \leq n} \lambda_2^T x_i \geq \tau$ for some constant $\tau > 0$, and $\sum_{i=1}^n x_i e_{oi} = 0$ and $\sum_{i=1}^n x_i (e_{oi}^2 - 1) = 0$.*

Theorem 1 summarizes our finite-sample analysis of dual regression. The proofs of all formal results in the paper are given in the Supplementary Material.

Theorem 1. *If Conditions 1–3 hold with $\omega = (\lambda_2^T x_1, \ldots, \lambda_2^T x_n)$, for all $\lambda_2 \in \Lambda_2$, then problem (2.3) admits the equivalent feasible formulation (D), with solution and multipliers e^* and λ^*, respectively. Moreover, for program (D) the following holds:*

(i) Primal problem: the dual of (D) is

$$(P) \quad \min_{\lambda \in \Lambda_0} \sum_{i=1}^{n} \frac{1}{2} \left\{ \left(\frac{y_i - \lambda_1^T x_i}{\lambda_2^T x_i} \right)^2 + 1 \right\} \left(\lambda_2^T x_i \right),$$

the primal dual regression problem, with solution λ_n.

(ii) First-order conditions: program (D) admits the method-of-moments representation

$$\sum_{i=1}^{n} x_i \left(\frac{y_i - \lambda_1^T x_i}{\lambda_2^T x_i} \right) = 0, \quad \frac{1}{2} \sum_{i=1}^{n} x_i \left\{ \left(\frac{y_i - \lambda_1^T x_i}{\lambda_2^T x_i} \right)^2 - 1 \right\} = 0, \tag{2.6}$$

the first-order conditions of (P).

(iii) With probability 1: (a) uniqueness: the pair (λ_n, e^*) is the unique optimal solution to (P) and (D), and $\lambda_n = \lambda^*$; (b) strong duality: the value of (D) equals the value of (P).

Theorem 1 establishes formal duality of our initial assignment problem under first and second moment orthogonality constraints and the global M-estimation problem (P). Convexity of (P) guarantees that to a unique assignment of e values corresponds a unique linear representation of the form (2.2). Uniqueness further implies that if e_o satisfies independence, then

the orthogonality conditions in (2.6) are both necessary and sufficient for the dual solution e^* to satisfy independence.

The primal problem (P) is a locally heteroscedastic generalization of a simultaneous location-scale estimator proposed by Huber (1981) and further analyzed in Owen (2001). The linear heteroscedastic model of equation (2.4) has been previously encountered in the quantile regression literature: see Koenker & Zhao (1994) and He (1997). The former consider the efficient estimation of (2.4) via L-estimation while the latter develops a restricted quantile regression method that prevents quantile crossing. Compared to these quantile-based methods, dual regression trades local estimation and the convenient linear programming formulation of quantile regression for simultaneous estimation of location and scale parameters.

2.4. **Connection with the dual formulation of quantile regression.** The dual problem of the linear $0 \cdot 5$ quantile regression of Y on X is (cf., Koenker, 2005, p. 87, equation 3.12):

$$\max_u \left\{ y^{\mathrm{T}} u : \sum_{i=1}^{n} x_i \left(u_i - \frac{1}{2} \right) = 0, \quad u \in [0,1]^n \right\}. \tag{2.7}$$

The solution to problem (2.7) produces values of u that are largely 0 and 1, with K sample points being assigned u values that are neither 0 nor 1. The points that are assigned 1 fall above the median quantile regression; the points receiving 0's fall below; and the remaining points fall on the median quantile regression plane. One direction of extension of (2.7) is to replace the 1/2 with values α that fall between 0 and 1 to obtain the α quantile regression.

Another extension is to augment problem (2.7) by adding K more constraints:

$$\max_u \left\{ y^{\mathrm{T}} u : \sum_{i=1}^{n} x_i \left(u_i - \frac{1}{2} \right) = 0, \ \sum_{i=1}^{n} x_i \left(u_i^2 - \frac{1}{3} \right) = 0, \ u \in [0,1]^n \right\}. \tag{2.8}$$

It is apparent that the solution to (2.7) does not satisfy (2.8): the variance of u around 0 in the solution to (2.7) is approximately 1/2, not 1/3. To satisfy program (2.8), the u's have to be moved off 0 and 1. Since x_i contains an intercept, the sample moments of u and u^2 will be 1/2 and 1/3; u and u^2 will be orthogonal to the columns of the matrix $(x_1, \ldots, x_n)^{\mathrm{T}}$, relations that are necessary but not sufficient for uniformity and independence.

Both systems (2.7) and (2.8) demand monotonicity by maximally correlating y and u. A violation of monotonicity requires there to be two observations that share the same X values but have different y values, with the lower of the two y values having the weakly higher value of u. However, a solution characterized by such a violation could be improved upon by exchanging the u assignments. Thus violation of monotonicity in program (2.7) arises because the set of admissible exchanges in u assignments is overly restricted: (2.7) is dual to a linear program well-known to have solutions at which K observations are interpolated when K parameters are being estimated, i.e., the hyperplanes obtained by regression quantiles must interpolate K observations.

By reformulating program (2.8) into a constrained optimization problem over $\mathbb{R}^n$, program (D) further expands the set of admissible exchanges in u assignments, since u is restricted to $[0,1]^n$. Doing this, the problem corresponding to (2.8) becomes the dual regression program (D), where e can take on any real value. It is then natural to take $u_i^* = F_n(e_i^*)$, the empirical cumulative distribution function of the dual regression solution e^*, thereby imposing uniformity to high precision even at small n.

3. Generalization

3.1. **Infeasible generalized dual regression.** The dual regression characterization of location-scale conditional distribution functions via the monotonicity element, the objective, and the independence element, the constraints, can be exploited to characterize more flexible representations. Similarly to the approach introduced in §2, we first analyze the infeasible assignment problem for a general representation of the stochastic structure of Y conditional on X:

$$Y = H(X,\varepsilon) \equiv H_X(\varepsilon), \quad \varepsilon \mid X \sim F, \tag{3.1}$$

where F is a specified cumulative distribution function with support the real line, and for each value x of X, the derivative $H'_x(\varepsilon)$ of $H_x(\varepsilon)$ is strictly positive. Representation (3.1) always exists with H_x defined as the composition of the conditional quantile function of Y given $X = x$ and the distribution function F.

To each monotone function H_x also corresponds a convex function $\widetilde{H}_x$ defined as

$$\widetilde{H}_x(e) \equiv \int_0^e H_x(s)ds \quad (e \in \mathbb{R}).$$

The monotonicity of $H_x(\varepsilon)$ guarantees the convexity of $\widetilde{H}_x(\varepsilon)$. The slope of this function gives the value of Y corresponding to a value e of ε at $X = x$. Thus $F_{Y|X}(Y \mid X)$ corresponds to a collection of convex functions, with one element of this collection for each value of X, together with a single random variable whose distribution is common to all the convex functions.

Equipped with $\widetilde{H}_X$, suppose we are tasked with assigning a value e_i to each of the n realizations $\{(y_i, x_i)\}_{i=1}^n$. Then, for $S_n = \sum_{i=1}^n \widetilde{H}_{x_i}(\varepsilon_i)$, solving the infeasible problem

$$\text{(IGD)} \quad \max_{e \in \mathbb{R}^n} \left\{ y^{\mathrm{T}} e : \sum_{i=1}^n \widetilde{H}_{x_i}(e_i) = S_n \right\},$$

generates the correct $y - e$ assignment: writing the Lagrangian

$$\mathscr{L} = y^{\mathrm{T}} e - \Lambda \left\{ \sum_{i=1}^n \widetilde{H}_{x_i}(e_i) - S_n \right\},$$

the n associated first-order conditions are

$$\frac{\partial \mathscr{L}}{\partial e_i} = y_i - \Lambda H_{x_i}(e_i) = 0 \quad (i = 1, \ldots, n), \tag{3.2}$$

and convexity of $\widetilde{H}_{x_i}$ then guarantees that (3.2) is uniquely satisfied by $(\Lambda, e) = (1, \varepsilon_o)$, with $\varepsilon_o = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$. This demonstrates that maximizing $y^{\mathrm{T}}e$ generally suffices to match e's to y's, regardless of the form of H_X in (3.1).

Theorem 2. *Suppose that (3.1) holds with $H_{x_i} : \mathbb{R} \to \mathbb{R}$ a continuously differentiable function that satisfies $\inf_{e\in\mathbb{R}} H'_{x_i}(e) \geq \tau$ for each x_i and some constant $\tau > 0$. Then the infeasible generalized dual regression problem (IGD) with $S_n = \sum_{i=1}^n \widetilde{H}_{x_i}(\varepsilon_i)$ generates the correct $y-e$ assignment, i.e., the pair $(\Lambda, e) = (1, \varepsilon_o)$ uniquely solves first-order conditions (3.2).*

Theorem 2 shows that problem (IGD) fully characterizes the $y-e$ assignment problem: given $\widetilde{H}_{x_i}$ and S_n, solving (IGD) assigns a value e_i to each sample point (y_i, x_i), and this value is the corresponding value $F_{Y|X}(y_i \mid x_i)$ up to a specified transformation F. If F is specified to be a known distribution, the n values $F_{Y|X}(y_i \mid x_i)$ are then also known. If F is specified to be an unknown distribution, as in our application below, the empirical distribution of ε_o then provides an asymptotically valid estimator for F. Knowledge of $\widetilde{H}_{x_i}$ and S_n can thus be incorporated into a mathematical programming problem which delivers the values of $F_{Y|X}$ at the n sample points.

3.2. **Generalized dual regression representations: definition and characterization.** Problem (IGD) is infeasible because neither $\widetilde{H}_{x_i}$ nor S_n is known. However, Theorem 2 motivates a feasible approach once H_X and F are specified. Denote the components of X without the intercept by $\widetilde{X}$, so that $X = (1, \widetilde{X})^{\mathrm{T}}$. Without loss of generality, let $\widetilde{X}$ be centered, denoted $\widetilde{X}^c$, and let $X^c = (1, \widetilde{X}^c)^{\mathrm{T}}$. With $h_1(\varepsilon) = 1$ and $h_2(\varepsilon) = \varepsilon$, we specify H_X by a linear combination of J basis functions $h(\varepsilon) = \{h_1(\varepsilon), \ldots, h_J(\varepsilon)\}^{\mathrm{T}}$, the coefficients of which depend on X:

$$H_X(\varepsilon) = \sum_{j=1}^{J} \beta_j(X) h_j(\varepsilon), \tag{3.3}$$

and we assume that H_X is linear in X and set:

$$\beta_j(X) = \alpha_j + \beta_j^{\mathrm{T}} \widetilde{X}^c \quad (j = 1, \ldots, J). \tag{3.4}$$

Finally, we specify a zero mean and unit variance distribution for ε by imposing $E(\varepsilon) = 0$ and $E(\varepsilon^2 - 1)/2 = 0$, and setting $\alpha_j = 0$ for $j = 3, \ldots, J$, in (3.4).

With $\alpha_2 + \beta_2^{\mathrm{T}} \widetilde{X}^c > 0$, our normalization and (3.3)–(3.4) together yield the augmented, generalized dual regression model

$$Y = \alpha_1 + \alpha_2\varepsilon + \beta_1^{\mathrm{T}} \widetilde{X}^c + (\beta_2^{\mathrm{T}} \widetilde{X}^c)\varepsilon + \sum_{j=3}^{J} (\beta_j^{\mathrm{T}} \widetilde{X}^c) h_j(\varepsilon) \equiv H_X(\varepsilon; \alpha, \beta), \quad \varepsilon \mid X \sim F. \tag{3.5}$$

Equation (3.5) admits of the following interpretation. When $\widetilde{X}^c = 0$, $Y = \alpha_1 + \alpha_2\varepsilon$ and $\varepsilon = (Y - \alpha_1)/\alpha_2$, so that ε is just a re-scaled version of the distribution of Y at $\widetilde{X}^c = 0$. Since ε is independent of X, transformations of this shape of ε must suffice to produce Y at other values of X. The first two transformations, $\beta_1^{\mathrm{T}} \widetilde{X}^c$ and $(\beta_2^{\mathrm{T}} \widetilde{X}^c)\varepsilon$, are translations of

location and scale which do not essentially affect the shape of Y's response to changes in ε at all. The additional terms $(\beta_j^{\mathrm{T}} \widetilde{X}^c) h_j(\varepsilon)$ achieve that end.

Suppose that we observe a sample of n identically and independently distributed realizations $\{(y_i, x_i)\}_{i=1}^n$ generated according to model (3.5). Define $x_{ij}^c = x_i^c$ for $j = 1, 2$, and $x_{ij}^c = \widetilde{x}_i^c$ for $j = 3, \ldots, J$, and let $(\gamma, \lambda) \in \mathbb{R}^{2+J(K-1)}$ and $e_o \in \mathbb{R}^n$ satisfy the system of n equations and $2n$ inequality constraints

$$y_i = H_{x_i}(e_{oi}; \gamma, \lambda), \quad \gamma_2 + \lambda_2^{\mathrm{T}} \widetilde{x}_i^c > 0, \quad H'_{x_i}(e_{oi}; \gamma, \lambda) > 0 \quad (i = 1, \ldots, n), \tag{3.6}$$

where e_o further satisfies $\sum_{i=1}^n x_{ij}^c \widetilde{h}_j(e_{oi}) = 0$ $(j = 1, \ldots, J)$, with $\widetilde{h}_1(e_{oi}) = e_{oi}$, $\widetilde{h}_2(e_{oi}) = (e_{oi}^2 - 1)/2$, and $\widetilde{h}_j(e_{oi}) = \int_0^{e_{oi}} h_j(s) ds$ $(j = 3, \ldots, J)$. These relations reduce to the linear heteroscedastic representation of §2 for $J = 2$, and impose that e_0 be a zero mean and unit variance vector satisfying the augmented set of orthogonality conditions $\sum_{i=1}^n \widetilde{x}_i^c \widetilde{h}_j(e_{oi}) = 0$ $(j = 1, \ldots, J)$. The sequence of vectors e_o that satisfies the generalized dual regression representation (3.6) as well as the associated orthogonality constraints for each n then provides an asymptotically valid characterization of the data-generating process (3.5).

Each element of this sequence is characterized by the assignment problem

$$\max_{e \in \mathbb{R}^n} \left\{ y^{\mathrm{T}} e : \sum_{i=1}^n \widetilde{H}_{x_i}(e_i; \theta) = \sum_{i=1}^n \widetilde{H}_{x_i}(e_{oi}; \theta) \right\}, \tag{3.7}$$

where $\widetilde{H}_{x_i}(e_i; \theta) = \int_0^{e_i} H_{x_i}(s; \theta) ds$, and $\theta = (\theta_1, \ldots, \theta_J)^{\mathrm{T}}$, with $\theta_j = (\gamma_j, \lambda_j)^{\mathrm{T}} \in \mathbb{R}^K$ for $j = 1, 2$, and $\theta_j = \lambda_j \in \mathbb{R}^{K-1}$ for $j = 3, \ldots, J$. Since e_0 and θ are unknown, problem (3.7) is infeasible; we thus formulate an equivalent, feasible implementation of problem (IGD):

$$\text{(GD)} \quad \max_{e \in \mathbb{R}^n} \left\{ y^{\mathrm{T}} e : \sum_{i=1}^n x_{ij}^c \widetilde{h}_j(e_i) = 0 \quad (j = 1, \ldots, J) \right\},$$

the generalized dual regression program. (GD) then uniquely characterizes representation (3.6).

In order to state the properties of (GD) formally, we define the parameter space Θ_n, which specifies parameter values compatible with monotone representations:

$$\Theta_n = \left\{ \theta \in \Theta_{0,n} : \text{there exists } e \in \mathbb{R}^n : y_i = H_{x_i}(e_i; \theta) \text{ and } \inf_{e \in \mathbb{R}} H'_{x_i}(e; \theta) > 0 \quad (i = 1, \ldots, n) \right\},$$

with $\Theta_{0,n} = \{\theta \in \mathbb{R}^{2+J(K-1)} : \inf_{i \le n} \theta_2^{\mathrm{T}} x_i^c > 0\}$. For $\theta \in \Theta_n$, let $e(y_i, x_i, \theta)$ denote the inverse function of $H_{x_i}(e_i; \theta)$, which is well-defined for each x_i. We assume that the basis functions h and the pair (θ, e_0) satisfy the following conditions.

Condition 4. *There exists a finite constant C_h such that $\max_{j=3,\ldots,J} \sup_{e \in \mathbb{R}} \{|h_j(e)| + |\widetilde{h}_j(e)|\} \le C_h$, and the matrix $E[h\{e(Y, X, \theta)\} h\{e(Y, X, \theta)\}^T \mid X = x_i]$ is finite and non-singular for each x_i and all $\theta \in \Theta_n$.*

Condition 5. *There exists* $(\theta, e_o) \in \Theta_n \times \mathbb{R}^n$ *such that* $y_i = H_{x_i}(e_{oi}; \theta)$ *and* $\inf_{e \in \mathbb{R}} H'_{x_i}(e; \theta) \geq \tau$*, for* $i = 1, \ldots, n$ *and some constant* $\tau > 0$*, and* e_o *satisfies* $\sum_{i=1}^n x_{ij}^c \widetilde{h}_j(e_{oi}) = 0$*, for* $j = 1, \ldots, J$*.*

Let $\phi(\theta) = [H'_{x_1}\{e(y_1, x_1, \theta); \theta\}, \ldots, H'_{x_n}\{e(y_n, x_n, \theta); \theta\}]^{\mathrm{T}}$. Theorem 3 summarizes our finite-sample analysis of generalized dual regression.

Theorem 3. *If Conditions 1, 2, 4 and 5 hold with* $\omega = \phi(\theta)$*, for all* $\theta \in \Theta_n$*, then problem (IGD) admits the equivalent feasible formulation (GD), with solution and multipliers* e^* *and* θ^**, respectively. Moreover, for program (GD) the following holds:*

(i) Primal problem: the dual of (GD) is

$$\textit{(GP)} \quad \min_{\theta \in \Theta_n} \sum_{i=1}^n \sum_{j=2}^J (\theta_j^T x_{ij}^c) \left[h_j \left\{ e(y_i, x_i, \theta) \right\} e(y_i, x_i, \theta) - \widetilde{h}_j \left\{ e(y_i, x_i, \theta) \right\} \right],$$

the primal generalized dual regression problem, with solution θ_n*.*

(ii) First-order conditions: program (GD) admits the method-of-moments representation

$$\sum_{i=1}^n x_{ij}^c \widetilde{h}_j \left\{ e(y_i, x_i, \theta) \right\} = 0 \quad (j = 1, \ldots, J), \tag{3.8}$$

the first-order conditions of (GP).

(iii) With probability 1: (a) uniqueness: the pair (θ_n, e^*) *is the unique optimal solution to (GP) and (GD), and* $\theta_n = \theta^*$*; (b) strong duality: the value of (GD) equals the value of (GP).*

Problem (GD) augments the set of orthogonality constraints in (D) and generates increasingly flexible representations of the form (3.6). For each element of this sequence, (GD) then provides a feasible formulation of the generalized $y - e$ assignment problem (IGD) with optimality condition $-H'_{x_i}(e_i^*; \theta^*) < 0$ equivalent to monotonicity. Theorem 3 also states the form of the corresponding primal problem, whose convexity guarantees that (GD) and (GP) uniquely and equivalently characterize representation (3.5). Uniqueness further implies that if e_o satisfies independence, then the orthogonality conditions in (3.8) are both necessary and sufficient for the dual solution e^* to satisfy independence as well. Theorem 3 thus characterizes and establishes the duality between specification of orthogonality constraints on e and specification of a globally monotone representation for Y conditional on X.

Formally, (GP) is the restriction of the dual of (GD) to Θ_n. The existence Condition 5 and the form of (GD) optimality conditions together ensure that (GD) does not admit a global maximum with associated multipliers outside Θ_n. Implementing (GP) thus requires the imposition of inequality constraints with e_i only implicitly defined in the specification of (GP) for $J > 2$, and problem (GD) therefore provides a greatly simplified dual implementation. The special case of dual regression corresponds to $J = 2$, and imposing $\sum_{i=1}^n \widetilde{h}_j(e_i) = 0$, for $j = 1, 2$, is a normalization. The simple basis $\{e_i, (e_i^2 - 1)/2\}$ is obviously impoverished for

the space of all convex functions, although quite practical for many applications once the flexibility in the distribution of e is taken into account.

3.3. **Connection with optimal transport formulation of quantile regression.** An alternative approach is to specify F to a known distribution, and alter representation (3.5) and the corresponding problem accordingly. If F is specified to be the standard uniform distribution, then (2.8) in §2.4 can be generalized as

$$\max_{u\in[0,1]^n}\left\{y^{\mathrm{T}}u : \frac{1}{j}\sum_{i=1}^{n} x_i\left(u_i^j - \frac{1}{j+1}\right) = 0 \quad (j=1,\ldots,J)\right\}. \tag{3.9}$$

For $u_i \in [0,1]$, let $m^J(u_i) = \{m_{J1}(u_i),\ldots,m_{JJ}(u_i)\}^{\mathrm{T}}$, with $m_{Jj}(u_i) = j^{-1}\{u_i^j - (j+1)^{-1}\}$. With $\otimes$ denoting the Kronecker product, the large-sample form of program (3.9) is

$$\max_{U_J\in(0,1)}\left\{E(YU_J) : E\{X\otimes m^J(U_J)\} = 0\right\}. \tag{3.10}$$

Letting J increase, both the distributional and the orthogonality constraints get strengthened. Because X includes an intercept, the distribution of U_J approaches uniformity, while simultaneously satisfying an increasing sequence of orthogonality constraints. In the limit, a uniformly distributed random variable U satisfying the full set of orthogonality constraints is thus specified. Since $E\{X\otimes m^J(U)\} = 0$ for all J is equivalent to the mean-independence property $E(X \mid U) = E(X)$ and the uniformity constraint $U \sim U(0,1)$, in the large J limit program (3.10) coincides with the scalar quantile regression problem proposed in independent work by Carlier et al. (2016) (cf., equation 19, p. 1180)

$$\max\left\{E(YU) : U\sim U(0,1),\ E(X\mid U) = E(X)\right\}, \tag{3.11}$$

which provides an optimal transport formulation of quantile regression (we are grateful to an anonymous referee for highlighting this connection). Program (3.11) is directly amenable to a linear programming implementation which maintains and exploits the full specification of the marginal distribution of U to a known distribution, whereas (3.10) provides a sequential nonlinear programming characterization of U which relaxes uniformity for finite n and J.

For $e_i \in \mathbb{R}$, let $\widetilde{h}^J(e_i) = \{\widetilde{h}_1(e_i),\ldots,\widetilde{h}_J(e_i)\}^{\mathrm{T}}$. The large-sample form of program (GD) is

$$\max_{e_J\in\mathbb{R}}\left\{E(Ye_J) : E(e_J) = 0,\ E(e_J^2-1) = 0,\ E\{\widetilde{X}^c\otimes\widetilde{h}^J(e_J)\} = 0\right\}. \tag{3.12}$$

Program (3.12) relaxes the support constraint in (3.9) and only specifies first and second moments of e_J, while the centering of X ensures that this is sufficient for e_J to be uniquely determined. The empirical distribution of solutions of the finite-sample analog (GD) of (3.12) then provides an asymptotically valid characterization of the distribution of e_J.

Letting J increase, orthogonality constraints in (3.12) are strengthened, and e_J gets closer and closer to satisfying the mean-independence property $E(\widetilde{X}^c \mid e_J) = 0$. It follows that for J large enough, (3.12) is equivalent to

$$\max_{e\in\mathbb{R}}\left\{E(Ye) : E(e) = 0,\ E(e^2-1) = 0,\ E(\widetilde{X}^c \mid e) = 0\right\}, \tag{3.13}$$

the limiting generalized dual regression problem. Theorem 4 summarizes this discussion.

Theorem 4. *Assume that $E(||X||^2) < \infty$. (i) Suppose that for any $a(U)$ with $E\{a(U)^2\} < \infty$ there are $J \times 1$ vectors ψ_J such that as $J \to \infty$, $E[\{a(U) - m^J(U)^T\psi_J\}^2] \to 0$. Then programs (3.10) and (3.11) are equivalent in the large J limit. (ii) Suppose that for any $b(e)$ with $E\{b(e)^2\} < \infty$ there are $J \times 1$ vectors ψ_J such that as $J \to \infty$, $E[\{b(e) - \widetilde{h}^J(e)^T\psi_J\}^2] \to 0$. Then programs (3.12) and (3.13) are equivalent for J large enough.*

4. Asymptotic Properties

We apply our framework to the estimation of a J–term generalized dual regression model of the form (3.5). Denote the support of X by $\mathcal{X}$, and, for some finite constant C_θ, define $\Theta_0 = \{\theta \in \mathbb{R}^{2+J(K-1)} : ||\theta|| \leq C_\theta \text{ and } \inf_{x\in\mathcal{X}} \theta_2^{\mathrm{T}} x^c > 0\}$. Letting $\mathcal{C}^1(\mathbb{R})$ denote the space of continuously differentiable functions on $\mathbb{R}$, define the space of strictly increasing functions indexed by X values, $\mathcal{M}(X) = \{e_X \in \mathcal{C}^1(\mathbb{R}) : \inf_{y\in\mathbb{R}} e_x'(y) > 0 \text{ for all } x \in \mathcal{X}\}$. The large-sample analog of Θ_n is then the space of vectors in Θ_0 such that there exists a corresponding optimal generalized dual regression representation:

$$\Theta = \{\theta \in \Theta_0 : \text{there exists } e_X \in \mathcal{M}(X) \text{ with } \Pr[Y = H_X\{e_X(Y); \theta\}] = 1\}.$$

For any $\theta \in \Theta$, denote e_X in $\mathcal{M}(X)$ such that $\Pr[Y = H_X\{e_X(Y); \theta\}] = 1$ by $e(Y, X, \theta)$.

Condition 6. *For some $\theta_0 \in \Theta$ and some mean zero and unit variance cumulative distribution function F, the representation $Y = H_X(\varepsilon; \theta_0)$ holds with probability one, with $\varepsilon \sim F$ and $E\{\widetilde{X}h_j(\varepsilon)\} = 0$, for $j = 1, \ldots, J$, and $\inf_{e\in\mathbb{R}} H_X'(e; \theta_0) \geq \tau$ for some constant $\tau > 0$.*

Condition 7. *The matrix M_n defined in Condition 2 satisfies $\lim n^{-1}M_n = M$, a positive definite matrix of rank K, and for all $\theta \in \Theta$ the matrix $E[h\{e(Y, X, \theta)\}h\{e(Y, X, \theta)\}^T \mid X]$ is nonsingular.*

Condition 8. *(i) Let $E(Y^2) < \infty$, $E(\|X\|^4) < \infty$ and $E(Y^2 \|X\|^2) < \infty$; (ii) let $E(Y^4) < \infty$, $E(\|X\|^6) < \infty$ and $E(Y^4 \|X\|^2) < \infty$.*

These conditions are used to establish existence and consistency of dual regression solutions, and Condition 8(ii) is needed for asymptotic normality of estimates of θ_0. In view of uniqueness stated in part (iii) of Theorem 3, these properties are shared by θ_n and θ^*, which we denote by $\widehat{\theta}$ for notational simplicity. We also denote both e_i^* and indirect estimates $e(y_i, x_i, \theta_n)$, constructed after solving (GP), by $\hat{e}_i$, with empirical distribution function $F_n(e) = n^{-1}\sum_{i=1}^n 1(\hat{e}_i \leq e)$, $e \in \mathbb{R}$. Furthermore, part (ii) of Theorem 3 shows that while the solution e^* is obtained directly by solving the mathematical program (GD), knowledge that the solution obeys representation (3.6) can be exploited to write estimating equations for $\hat{\theta}$ in the form of system (3.8). The computation of the asymptotic distribution of $\hat{\theta}$ follows from this characterization.

Theorem 5. *If $\{(y_i, x_i)\}_{i=1}^n$ are identically and independently distributed, and Conditions 1, 2, 4, and 6–8 hold with $\omega = \phi(\theta)$, for all $\theta \in \Theta_n$, then (i) there exists $\hat{\theta}$ in Θ with probability approaching one, (ii) $\hat{\theta}$ converges in probability to θ_0, and (iii) $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to $N(0, \Sigma)$, with Σ defined in (5.6) in the Supplementary Material.*

Knowledge of the statistical properties of $\hat{\theta}$ can be used to establish the limiting behaviour of the empirical distribution of $\hat{e}$. Define the empirical dual regression process

$$\mathbb{U}_n(e) = n^{1/2}\{F_n(e) - F(e)\} \quad (e \in \mathbb{R}).$$

Theorem 6 establishes weak convergence of the empirical distribution of $\hat{e}$ and the limiting behaviour of $\mathbb{U}_n$, accounting for its dependence on the distribution of $n^{1/2}(\hat{\theta} - \theta_0)$.

Theorem 6. *If the conditions of Theorem 5 hold, and, uniformly in x over $\mathcal{X}$, $f_{Y|X}(y \mid x)$ is uniformly continuous in y, bounded and, for some finite constant C_f and all $\theta \in \Theta_n$, satisfies $\sup_{e\in\mathbb{R}} e^2 f_{Y|X}\{H_x(e;\theta) \mid x\} \leq C_f$, then (i) $\sup_{e\in\mathbb{R}} |F_n(e) - F(e)|$ converges in probability to zero, and (ii) $\mathbb{U}_n$ converges weakly to a zero-mean Gaussian process $\mathbb{U}$ with covariance function defined in (5.11) in the Supplementary Material.*

Theorems 5 and 6 together establish that the pair $(\hat{\theta}, \hat{e})$ provides an asymptotically valid characterization of the generalized dual regression representation specified in Condition 6. When ε is independent of X, Theorem 6 further implies that the empirical distribution of $\hat{e}$ provides an asymptotically valid estimator of the conditional distribution of Y given X. For $u \in (0, 1)$, estimates of the X coefficients in quantile regression form can then be constructed as $\sum_{j=1}^J \hat{\lambda}_j h_j\{F_n^{-1}(u)\}$, exploiting the structure of the conditional quantile function of Y given X implied by representation (3.5). Theorem 6 also establishes asymptotic normality of the empirical dual regression process. The form of the covariance function of $\mathbb{U}$ reflects the influence of imposing sample orthogonality constraints in (GD) on the empirical distribution of e^*, or equivalently, of sample variability of parameter estimates θ_n on the empirical distribution of $e(y_i, x_i, \theta_n)$, as expected from the classical result of Durbin (1973).

Theorem 6 can be applied to perform pointwise inference on the conditional distribution function of Y conditional on X. However, simultaneous inference over regions of the joint support of Y and X is typically of interest in practice. Several approaches for uniform inference in the presence of non-pivotal limit processes have been considered in the literature (e.g., Koenker & Xiao, 2002, and Parker, 2013), including simulation methods (Chernozhukov et al., 2013). Extension of existing results to dual regression is beyond the scope of this paper but they provide a natural direction for future study of uniform inference on the empirical dual regression process.

5. Engel's Data Revisited

5.1. **Empirical analysis.** The classical dataset collected by Engel consists of food expenditure and income measurements for 235 households, and has been studied by means of quantile

regression methods (Koenker, 2005). We illustrate dual regression methods by estimating the statistical relationship between food expenditure and income, with household income as a single regressor and food expenditure as the outcome of interest.

We specify the vector of basis functions by means of trigonometric series. Alternative choices such as splines and shape-preserving wavelets (e.g., DeVore, 1977, and Cosma et al., 2007). In order to choose J, we first implement program (GD) for $J = 2$, which we then augment sequentially adding one pair of cosine and sine basis at a time, up to a representation of order $J = 8$. At each step, we compute a Schwarz Information Criterion (Schwarz, 1978) applied to the primal generalized dual regression problem, exploiting the strong duality result of Theorem 3 in order to compute its value as $y^{\mathrm{T}}e^* + \{2 + J(K-1)\}\log n$. Our procedure selects the location-scale representation $J = 2$. In the Supplementary Material, we describe the procedure and report results from the augmented specifications, which show that our results are robust to the number of terms included. In order to test for the validity of the selected model, a complementary procedure that should be explored in future research is to test for independence of dual regression solutions and explanatory variables. The test for multivariate independence proposed by Genest et al. (2007) constitutes an interesting starting point for such a development.

All computational procedures can be implemented in the software R (R Development Core Team, 2017) using open source software packages for nonlinear optimization such as Ipopt or Nlopt, and their R interface Ipoptr and Nloptr developed by Jelmer Ypma. Quantile regression procedures in the package quantreg have been used to carry our comparisons.

Figure 5.1 illustrates our results and plots the estimated distribution of food expenditure conditional on household income. Estimates $\{u_i^*\}_{i=1}^n$, where $u_i^* = F_n(e_i^*)$, are used in order to plot each observation in the xyu-space with predicted coordinates (x_i, y_i, u_i^*), and the solid lines give the u-level sets for a grid of values $\{0 \cdot 1, \ldots, 0 \cdot 9\}$. Although nonstandard, this representation relates to standard quantile regression plots since the levels of the distribution function give the conditional quantiles of food expenditure for each value of income. These are the plotted shadow solid lines corresponding for each u to dual regression estimates of conditional quantile functions of food expenditure given household income.

Figure 5.1 shows that the predicted conditional distribution function obtained by dual regression is indeed endowed with all desired properties. Of particular interest is the fact that the estimated function is monotone in food expenditure. Also, our estimates satisfy some basic smoothness requirements across probability levels, in the food expenditure values. This feature does not typically characterize estimates of the conditional quantile process by quantile regression methods, as conditional quantile functions are then estimated sequentially and independently of each other. The decreasing slope of the distribution function across values of income provides evidence that the data indeed follow a heteroscedastic generating process. This is the distributional counterpart of quantile functions having increasing slope across probability levels, a feature characterizing the conditional quantile functions on the

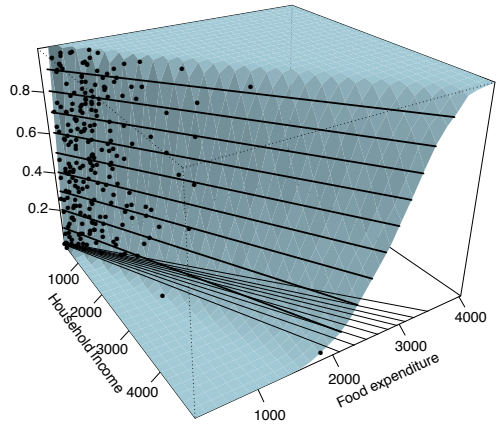


FIGURE 5.1. Dual regression estimate of the distribution of food expenditure conditional on income. Level sets (solid lines) are plotted for a grid of values ranging from $0 \cdot 1$ to $0 \cdot 9$. The projected shadow level sets yield the respective conditional quantile functions appearing on the xy-plane.

xy plane and signalling increasing dispersion in food expenditure across household income values.

Figure 5.2 gives the more familiar quantile regression plots. The plots presented show scatterplots of Engel's data as well as conditional quantile functions obtained by dual and quantile regression methods. The rescaled plots in the right panels of Fig. 5.2 highlight some features of the two procedures. The fitted lines obtained from dual regression are not subject to crossing in this example, whereas several of the fitted quantile regression lines actually cross for small values of household income. Last, the more evenly spread dual regression conditional quantile functions illustrate the effect of specifying a functional form for the quantile regression coefficients, while preserving asymmetry in the conditional distribution of food expenditure.

Figure 5.3 compares our estimates of intercept and income coefficients in quantile regression form, with estimates obtained by quantile regression. For interpretational purposes, we follow Koenker (2005) and estimate the functional coefficients after having recentered household income. This avoids having to interpret the intercept as food expenditure for households with zero income. After centering, the intercept coefficient can be interpreted as the u-th quantile of food expenditure for households with mean income. Fig. 5.3 shows the estimated quantile regression coefficients as a function of u. It illustrates the fact that the flexible structure imposed by dual regression yields estimates that are indeed smoother than their

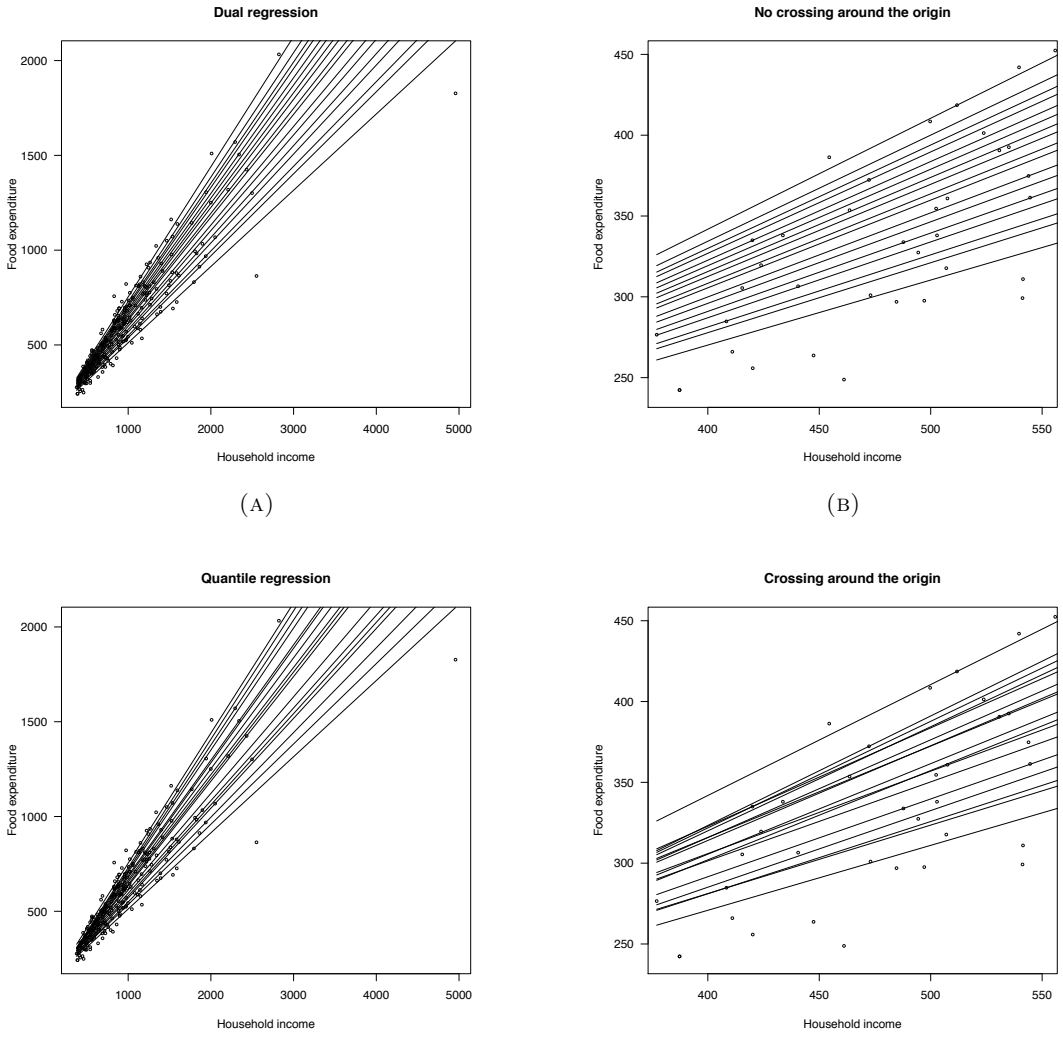


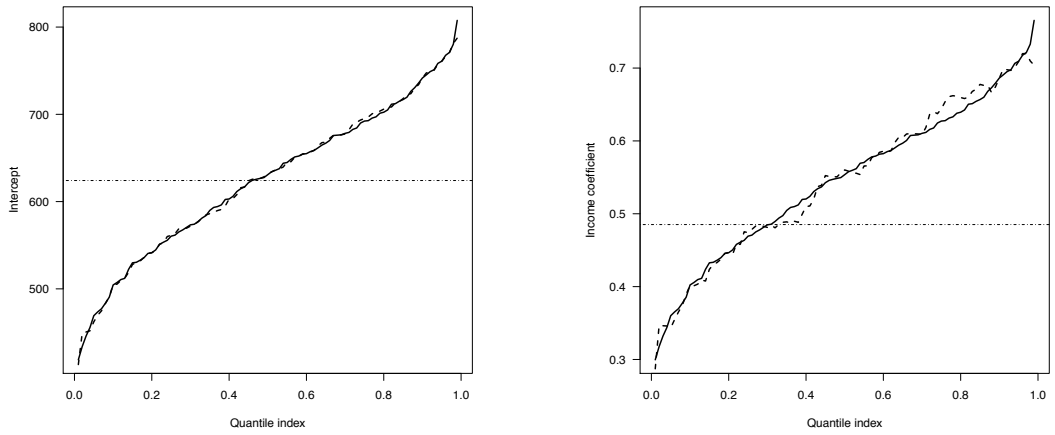
FIGURE 5.2. Scatterplots and dual (a) and quantile (c) regression estimates of the conditional $\{0 \cdot 1, 0 \cdot 15, \ldots, 0 \cdot 9\}$ quantile functions (solid lines) for Engel's data, and their rescaled counterparts ((b),(d)).

quantile regression counterpart, the latter having a somewhat erratic behaviour around our estimates.

5.2. **Simulations.** We give a brief summary of the results of a Monte Carlo simulation in order to assess the finite-sample properties dual regression. The data-generating process is

$$y_i = \alpha_1 + \beta_1 \widetilde{x}_i + (\alpha_2 + \beta_2 \widetilde{x}_i)\varepsilon_i, \quad \varepsilon_i \sim N(0,1), \tag{5.1}$$

with parameter values calibrated to the empirical application, from which 4999 samples are simulated. As a benchmark, we compare generalized dual regression estimates of the values $F_{Y|X}(y_i \mid x_i)$ $(i = 1, \ldots, n)$, to those obtained by applying the inversion procedure



(A) (B)

FIGURE 5.3. Engel coefficient plots revisited. Dual (solid) and quantile (dashes) regression estimates of the intercept (a) and income (b) coefficients as a function of the quantile index. Least squares estimates are also shown (dot-dash).

of Chernozhukov et al. (2010) to the quantile regression process. For each simulation, the estimation and selection procedures are identical to those implemented in the empirical application.

Table 1 reports a first set of results regarding the accuracy of conditional distribution function estimates. We report average estimation errors across simulations of dual regression and quantile regression estimators, respectively, and their ratio in percentage terms. Estimation errors are measured in L^p norms $\|\cdot\|_p$, for $p = 1, 2$, and ∞, where for $f : \mathbb{R} \mapsto [0, 1]$, $\|f\|_p = \left\{\int_{\mathbb{R}} |f(s)|^p \, ds\right\}^{1/p}$, and are computed with e^* the solutions to the selected generalized dual regression program. Correct model selection ranges from 75% of the simulations for $n = 100$ to 90% for $n = 1000$, providing encouraging evidence about the validity of the proposed criterion. The results show that for this setup our estimates systematically outperform quantile regression-based estimates, with the spread in performance increasing with sample size. Whereas the reduction in average estimation error is between 8% and 17%, depending on the norm, for $n = 235$, estimation error is reduced up to 30% when $n = 1000$. The larger reduction in average errors in L^∞ norm reflects the higher accuracy in estimation of extreme parts of the distribution.

In the Supplementary Material, we describe the experiment in detail, and report results on estimation of quantile regression coefficients and the distribution of selected models across simulations. We also include additional simulations that illustrate the empirical performance of dual regression with multiple covariates and show that it performs well relative to the noncrossing quantile regression method proposed by Bondell et al. (2010).

Sample size	L^1_{GDR}	L^1_{GDR}/L^1_{QR}	L^2_{GDR}	L^2_{GDR}/L^2_{QR}	L^∞_{GDR}	$L^\infty_{GDR}/L^\infty_{QR}$
$n = 100$	$4 \cdot 11$	$93 \cdot 34$	$5 \cdot 63$	$92 \cdot 59$	$21 \cdot 89$	$89 \cdot 89$
$n = 235$	$2 \cdot 70$	$91 \cdot 64$	$3 \cdot 73$	$89 \cdot 73$	$17 \cdot 15$	$82 \cdot 27$
$n = 500$	$1 \cdot 85$	$90 \cdot 68$	$2 \cdot 56$	$87 \cdot 98$	$12 \cdot 97$	$75 \cdot 03$
$n = 1000$	$1 \cdot 31$	$90 \cdot 18$	$1 \cdot 83$	$87 \cdot 03$	$9 \cdot 94$	$70 \cdot 12$

TABLE 1. L^p estimation errors ($\times 100$) of generalized dual (L^p_{GDR}) and quantile regression (L^p_{QR}) estimates of $\{F_{Y|X}(y_i \mid x_i)\}_{i=1}^n$, and their ratios ($\times 100$), for $p = 1, 2$ and ∞.

6. Discussion

If we designate problems such as (D) and (GD) as already dual, then their solutions reveal a corresponding primal. Typically, the Lagrange multipliers of the dual appear as parameters in the primal, and the primal has an interpretation as a data-generating process. So perhaps not surprisingly the constraints on the construction of the stochastic elements have shadow values that are parameters of a data-generating representation. In this way the relation between identification and estimation is made perspicuous: a parameter of the data-generating process is the Lagrange multiplier of a specific constraint on the construction of the stochastic element, so to specify that some parameters are non-zero and others are zero is to say that some constraints are in the large-sample limit binding and others are not.

Another way of expressing this is to say that when a primal corresponds to the data-generating process, additional moment conditions are superfluous: they will in the limit attract Lagrange multiplier values of zero and consequently not affect the value of the program nor the solution. In a sense, this is obvious: the parameters of the primal can typically be identified and estimated through an M–estimation problem that will generate K equations to be solved for the K unknown parameters. Nonetheless, the recognition that the only moment conditions that contribute to enforcing the independence requirement are those whose imposition simultaneously reduces the objective function while providing multipliers that are coefficients in the stochastic representation of Y suggests the futility of portmanteau approaches (e.g., those based on characteristic functions) to imposing independence. The dual formulation reveals that to specify the binding moment conditions is to specify an approximating data-generating process representation, which then can be extrapolated to provide estimates of objects of interest beyond the n explicitly estimated values of ε_i that characterize the sample and the definition of the mathematical program.

As is well understood in mathematical programming, dual solutions provide lower bounds on the values obtained by primal problems. In the generic form of the problems we have considered here there is no gap between the primal and dual values; hence in econometrics these problems are said to display point identification. We conjecture that the problems without point identification do have gaps between their dual and primal values, and that this characterization will enhance our understanding.

References

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

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Carlier, G., Chernozhukov, V., & Galichon, A. (2016). Vector quantile regression: an optimal transport approach. *Annals of Statistics* **44**, 1165–1192.

Cosma, A., Scaillet, O., & Von Sachs, R. (2007). Multivariate wavelet-based shape-preserving estimation for dependent observations. *Bernouilli* **13**, 301–329.

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

Chernozhukov, V., Fernandez-Val, I. & Melly, B. (2013). Inference on Counterfactual Distributions. *Econometrica* **81**, 2205–2268.

De Vore, R. (1977). Monotone approximation by splines. *SIAM Journal on Mathematical Analysis* **8**, 891–905.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics*, **1**, 279–290.

Genest, C., Quessy, J.-F., & Remillard, B. (2007). Asymptotic local efficiency of Cramer–von Mises tests for multivariate independence. *Annals of Statistics* **35**, 166–191.

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

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

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

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

Koenker, R. & Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica* **70**, 1583–1612.

Koenker, R. & Zhao, Q. (1994). L-estimation for linear heteroscedastic models. *Nonparametric Statistics* **3**, 223–235.

Owen, A. (2001). *Empirical Likelihood*. Chapman&Hall/CRC, Boca Raton, USA.

Parker, T. (2013). A comparison of alternative approaches to supremum-norm goodness-of-fit tests with estimated parameters. *Econometric Theory* **29**, 969–1008.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

SUPPLEMENTARY MATERIAL FOR "DUAL REGRESSION"

RICHARD H. SPADY[†] AND SAMI STOULI[§]

Abstract. This Supplementary Material contains proofs of all formal results in the main text, as well as additional simulations.

1. Main conditions

For clarity of exposition we recall Conditions 1–8 stated in the main text.

Condition 1. *The random variable Y is continuously distributed conditional on X, with conditional density $f_{Y|X}(y \mid X)$ bounded away from 0.*

Condition 2. *For a specified vector $\omega \in \mathbb{R}^n$, the matrix $diag(\omega_i)$ is nonsingular and the matrix $\sum_{i=1}^n \omega_i^{-1} x_i x_i^T = M_n$ is finite, positive definite, and has rank K.*

Condition 3. *There exists $(\lambda, e_o) \in \Lambda_0 \times \mathbb{R}^n$ such that $y_i = \lambda_1^T x_i + (\lambda_2^T x_i) e_{oi}$ with $\inf_{i \le n} \lambda_2^T x_i \ge \tau$ for some constant $\tau > 0$, and $\sum_{i=1}^n x_i e_{oi} = 0$ and $\sum_{i=1}^n x_i (e_{oi}^2 - 1) = 0$.*

Condition 4. *There exists a finite constant C_h such that $\max_{j=3,\dots,J} \sup_{e \in \mathbb{R}} \{|h_j(e)| + |\widetilde{h}_j(e)|\} \le C_h$, and the matrix $E[h\{e(Y,X,\theta)\} h\{e(Y,X,\theta)\}^T \mid X = x_i]$ is finite and nonsingular for each x_i and all $\theta \in \Theta_n$.*

Condition 5. *There exists $(\theta, e_o) \in \Theta_n \times \mathbb{R}^n$ such that $y_i = H_{x_i}(e_{oi}; \theta)$ and $\inf_{e \in \mathbb{R}} H'_{x_i}(e; \theta) \ge \tau$, for $i = 1, \dots, n$ and some constant $\tau > 0$, and e_o satisfies $\sum_{i=1}^n x_{ij}^c \widetilde{h}_j(e_{oi}) = 0$, for $j = 1, \dots, J$.*

Condition 6. *For some $\theta_0 \in \Theta$ and some mean zero and unit variance cumulative distribution function F, the representation $Y = H_X(\varepsilon; \theta_0)$ holds with probability one, with $\varepsilon \sim F$ and $E\{\widetilde{X} h_j(\varepsilon)\} = 0$, for $j = 1, \dots, J$, and $\inf_{e \in \mathbb{R}} H'_X(e; \theta_0) \ge \tau$ for some constant $\tau > 0$.*

Condition 7. *The matrix M_n defined in Condition 2 satisfies $\lim n^{-1} M_n = M$, a positive definite matrix of rank K, and for all $\theta \in \Theta$ the matrix $E[h\{e(Y,X,\theta)\} h\{e(Y,X,\theta)\}^T \mid X]$ is nonsingular.*

Condition 8. *(i) Let $E(Y^2) < \infty$, $E(\|X\|^4) < \infty$ and $E(Y^2 \|X\|^2) < \infty$; (ii) let $E(Y^4) < \infty$, $E(\|X\|^6) < \infty$ and $E(Y^4 \|X\|^2) < \infty$.*

Date: November 13, 2017.

[†] Department of Economics, Johns Hopkins University, rspady@jhu.edu.

§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

2. Proof of Theorem 2

For $\Lambda \in \mathbb{R}$, the Lagrangian of the infeasible problem (IGD) is

$$\mathscr{L}^{IGD}(e, \Lambda) = y^{\mathrm{T}} e - \Lambda \left\{ \sum_{i=1}^{n} \widetilde{H}_{x_i}(e_i) - S_n \right\} \quad (e \in \mathbb{R}^n).$$

By definition of $\widetilde{H}_{x_i}$ and continuous differentiability of H_{x_i}, for each x_i, the Fundamental Theorem of Calculus implies the n first-order conditions

$$\nabla_{e_i} \mathscr{L}^{IGD} = y_i - \Lambda H_{x_i}(e_i) = 0, \qquad (i = 1, \ldots, n), \tag{2.1}$$

The n second-order conditions

$$\nabla_{e_i e_i} \mathscr{L}^{IGD} = -\Lambda H'_{x_i}(e_i) < 0, \qquad (i = 1, \ldots, n), \tag{2.2}$$

are satisfied if and only if $\Lambda > 0$, since H'_{x_i} is strictly positive for each x_i. Since $y = 0$ cannot hold under Condition 1, with probability one, (2.1) rules out $\Lambda = 0$. Moreover, for $\Lambda < 0$, (2.2) implies that the map $e \mapsto \mathscr{L}^{IGD}(e, \Lambda)$ is strictly convex over the real line, since $\inf_{e \in \mathbb{R}} H'_{x_i}(e) \geq \tau > 0$ for each x_i, and hence unbounded above. Therefore we only need to show that the pair $(1, \varepsilon_o)$ is the unique pair in $(0, \infty) \times \mathbb{R}^n$ that satisfies (2.1).

By strict monotonicity of H_{x_i}, the inverse function $H_{x_i}^{-1}$ is well-defined, for each x_i, and a solution to (2.1) is $e_i = H_{x_i}^{-1}(y_i/\Lambda)$ $(i = 1, \ldots, n)$. Substituting into the constraint of (IGD) yields

$$\sum_{i=1}^{n} \widetilde{H}_{x_i} \left\{ H_{x_i}^{-1} \left(\frac{y_i}{\Lambda} \right) \right\} - S_n = 0. \tag{2.3}$$

By Lemma 1 below, $\Lambda = 1$ is the unique solution to (2.3) such that $\Lambda > 0$. Since $H_{x_i}^{-1}(y_i/\Lambda) = \varepsilon_i$ $(i = 1, \ldots, n)$ for $\Lambda = 1$, strict concavity of $\mathscr{L}^{IGD}$ for $\Lambda > 0$ implied by (2.2) shows that $(1, \varepsilon_o)$ is the unique pair in $(0, \infty) \times \mathbb{R}^n$ that satisfies (2.1).

Lemma 1. *Under the conditions of Theorem 2,* $\Lambda = 1$ *is the unique solution to the equation*

$$\sum_{i=1}^{n} \widetilde{H}_{x_i} \left\{ H_{x_i}^{-1} \left(\frac{y_i}{\Lambda} \right) \right\} - S_n = 0 \tag{2.4}$$

such that $\Lambda > 0$.

Proof. We first show that Equation (2.4) is the first-order condition of the infeasible generalized dual regression primal problem $\min_{\Lambda > 0} Q_n^{IGD}(\Lambda)$, where

$$Q_n^{IGD}(\Lambda) = \sum_{i=1}^{n} y_i H_{x_i}^{-1} \left(\frac{y_i}{\Lambda} \right) - \Lambda \left[\sum_{i=1}^{n} \widetilde{H}_{x_i} \left\{ H_{x_i}^{-1} \left(\frac{y_i}{\Lambda} \right) \right\} - S_n \right] \quad (\Lambda > 0),$$

and then show that $Q_n^{IGD}(\Lambda)$ admits $\Lambda = 1$ as its unique minimum.

Step 1. Define the Lagrange dual function (Boyd & Vandenberghe, 2004, Chapter 5) $Q_n^{IGD}(\mathit{\Lambda}) \equiv \sup_{e \in \mathbb{R}^n} \mathscr{L}^{IGD}(e, \mathit{\Lambda})$, for $\mathit{\Lambda} > 0$. In order to derive $Q_n^{IGD}(\mathit{\Lambda})$, we show that

the maximum of the map $e \mapsto \mathscr{L}^{IGD}(e, \mathit{\Lambda})$ is attained and is unique, and evaluate $e \mapsto \mathscr{L}^{IGD}(e, \mathit{\Lambda})$ at this value.

For $\mathit{\Lambda} > 0$ and $c \in \mathbb{R}$, consider the level sets $\mathcal{B}_c(\mathit{\Lambda}) = \{e \in \mathbb{R}^n : -\mathscr{L}^{IGD}(e, \mathit{\Lambda}) \leq c\}$ of $-\mathscr{L}^{IGD}$. These sets are compact. Given $e_1, e_2 \in \mathcal{B}_c(\mathit{\Lambda})$, let $t = ||e_1 - e_2||$ and $u = \frac{e_1 - e_2}{||e_1 - e_2||}$, so that $||u|| = 1$ and $e_1 = e_2 + tu$. Thus, by definition of e_1, a second-order Taylor expansion of $t \mapsto -\mathscr{L}^{IGD}(e_2 + tu, \mathit{\Lambda})$ around $t = 0$ yields, for some $\bar{e}$ on the line connecting e_1 and e_2,

$$\begin{aligned} c \geq -\mathscr{L}^{IGD}(e_1, \mathit{\Lambda}) &= -\mathscr{L}^{IGD}(e_2 + tu, \mathit{\Lambda}) \\ &= -\mathscr{L}^{IGD}(e_2, \mathit{\Lambda}) - t\nabla_e \mathscr{L}^{IGD}(e_2, \mathit{\Lambda})^T u - \frac{t^2}{2} u^T \nabla_{ee} \mathscr{L}^{IGD}(\bar{e}, \mathit{\Lambda}) u \\ &\geq -\mathscr{L}^{IGD}(e_2, \mathit{\Lambda}) - t||\nabla_e \mathscr{L}^{IGD}(e_2, \mathit{\Lambda})^T|| + \Lambda\tau\frac{t^2}{2}, \end{aligned}$$

where the last inequality follows from $-\nabla_{ee}\mathscr{L}^{IGD}(\bar{e}, \mathit{\Lambda}) = \Lambda \text{diag}\{H'_{x_i}(\bar{e}_i)\}$ and the uniform lower bound on H'_{x_i} for each x_i which implies that $-\nabla_{ee}\mathscr{L}^{IGD}(\bar{e}, \mathit{\Lambda})$ is positive definite. For $e_2 \in \mathcal{B}_c(\mathit{\Lambda})$, the above inequality implies that t is bounded and therefore $\mathcal{B}_c(\mathit{\Lambda})$ is bounded. Since $e \mapsto -\mathscr{L}(e, \mathit{\Lambda})$ is continuous over $\mathbb{R}^n$, $\mathcal{B}_c(\mathit{\Lambda})$ is also closed. It then follows from the Weierstrass theorem that there exists $e(\mathit{\Lambda}) \in \arg\min_{e \in \mathbb{R}^n}\{-\mathscr{L}^{IGD}(e, \mathit{\Lambda})\} = \arg\max_{e \in \mathbb{R}^n} \mathscr{L}^{IGD}(e, \mathit{\Lambda})$.

Since the Hessian matrix of the map $e \mapsto \mathscr{L}^{IGD}(e, \mathit{\Lambda})$ is negative definite for all $\mathit{\Lambda} > 0$, $e \mapsto \mathscr{L}^{IGD}(e, \mathit{\Lambda})$ is strictly concave with unique maximum $e(\mathit{\Lambda})$, for all $\mathit{\Lambda} > 0$. Upon using first-order conditions (2.1), direct substitution yields $\mathscr{L}^{IGD}\{e(\mathit{\Lambda}), \Lambda\} = Q_n^{IGD}(\Lambda)$, the maximum of the map $e \mapsto \mathscr{L}^{IGD}(e, \Lambda)$, for all $\Lambda > 0$.

Step 2. The function $Q_n^{IGD}(\Lambda)$ is strictly convex for $\mathit{\Lambda} > 0$: since H_{x_i} is continuously differentiable for each x_i by assumption, by the inverse function theorem $H_{x_i}^{-1}$ is continuously differentiable for each x_i, and there are the following derivatives:

$$\nabla_\Lambda H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right) = -\frac{1}{H'_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}\frac{y_i}{\Lambda^2} \tag{2.5}$$

$$\nabla_\Lambda \widetilde{H}_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\} = -\frac{y_i}{\Lambda}\frac{1}{H'_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}\frac{y_i}{\Lambda^2}, \tag{2.6}$$

for every x_i, y_i and $\mathit{\Lambda} > 0$. Upon using (2.5) and (2.6), $Q_n^{IGD}(\Lambda)$ has first-order conditions (2.4), and the second-order conditions

$$\nabla_{\Lambda\Lambda} Q_n^{IGD} = \frac{1}{\Lambda}\sum_{i=1}^n \frac{1}{H'_{x_i}\left\{H_{x_i}^{-1}\left(\frac{y_i}{\Lambda}\right)\right\}}\left(\frac{y_i}{\Lambda}\right)^2 > 0$$

are satisfied for all $\mathit{\Lambda} > 0$ since $H'_{x_i} > 0$ for each x_i. Therefore, $Q_n^{IGD}(\Lambda)$ is strictly convex for all $\mathit{\Lambda} > 0$ and admits at most one minimum. Since $H_{x_i}^{-1}(y_i/\Lambda) = \varepsilon_{oi}$ $(i = 1, \ldots, n)$ for $\Lambda = 1$, and $S_n = \sum_{i=1}^n \widetilde{H}_{x_i}(\varepsilon_{oi})$ by definition, $\Lambda = 1$ is also feasible. The result follows. $\square$

3. Proofs of Theorems 1 and 3

3.1. **Proof of Theorem 1.** Theorem 1 is a corollary of Theorem 3, upon substituting x_i to x_i^c and setting $J = 2$.

3.2. **Preliminary lemmas.** We establish the equivalence (IGD)–(GD) and convexity of (GP).

Lemma 2. *If Conditions 1, 2, 4 and 5 hold with $\omega = \phi(\theta)$, for all $\theta \in \Theta_n$, then the infeasible problem (IGD) admits the equivalent formulation (GD).*

Proof. Letting $\widetilde{H}_{x_i}(e_{oi};\theta) = \int_0^{e_{oi}} H_{x_i}(s;\theta)ds$, the corresponding expression is

$$\widetilde{H}_{x_i}(e_{oi};\theta) = \sum_{j=1}^{J}(\theta_j^{\mathrm{T}} x_{ij}^c)\widetilde{h}_j(e_{oi}) \quad (e_{oi} \in \mathbb{R}). \tag{3.1}$$

Given the form of $\widetilde{H}_{x_i}(\cdot;\theta)$, the constraint $\sum_{i=1}^n \widetilde{H}_{x_i}(e_i;\theta) = S_n$ in (IGD) can be simplified using

$$S_n = \sum_{i=1}^{n}\widetilde{H}_{x_i}(e_{oi};\theta) = \sum_{i=1}^{n}\sum_{j=1}^{J}(\theta_j^{\mathrm{T}} x_{ij}^c)\widetilde{h}_j(e_{oi}) = \sum_{j=1}^{J}\theta_j^{\mathrm{T}}\left\{\sum_{i=1}^{n} x_{ij}^c\widetilde{h}_j(e_{oi})\right\} = 0,$$

by definition of S_n in Theorem 2, expansion (3.1), and the properties of e_{oi} assumed in Condition 5. Therefore, the infeasible problem (IGD) becomes

$$\max_{e\in\mathbb{R}^n}\left\{y^{\mathrm{T}}e : \sum_{j=1}^{J}\sum_{i=1}^{n}(\theta_j^{\mathrm{T}} x_{ij}^c)\widetilde{h}_j(e_i) = 0\right\},$$

with Lagrangian

$$\mathscr{L}(e,\Lambda) = y^{\mathrm{T}}e - \Lambda\left\{\sum_{i=1}^{n}\widetilde{H}_{x_i}(e_i;\theta) - S_n\right\} = y^{\mathrm{T}}e - \Lambda\sum_{j=1}^{J}\sum_{i=1}^{n}(\theta_j^{\mathrm{T}} x_{ij}^c)\widetilde{h}_j(e_i) \quad (e\in\mathbb{R}^n),$$

for $\Lambda \in \mathbb{R}$. For all $\theta \in \Theta_n$, the map $e_i \mapsto H_{x_i}(e_i;\theta)$ satisfies the conditions of Theorem 2, which implies that $\Lambda = 1$ and $e = e_o$, by application of Theorem 2 upon substituting $H_{x_i}(\cdot;\theta)$ for $H_{x_i}(\cdot)$ and e_{oi} for ε_i $(i = 1,\ldots,n)$.

Adding θ to the choice variables of the optimization problem, we obtain the $\dim(\theta)$ additional constraints

$$\nabla_{\theta_j}\mathscr{L} = -\sum_{i=1}^{n} x_{ij}^c\widetilde{h}_j(e_i) \;\; = \;\; 0 \quad (j = 1,\ldots,J). \tag{3.2}$$

Equation (3.2) can be directly appended to the objective $\max_{e\in\mathbb{R}^n}\ y^{\mathrm{T}}e$ to obtain the optimization problem (GD) in which the Lagrange multiplier is θ. By part (iii) of Theorem 3, problem (GD) admits a unique optimal solution e^* over $\mathbb{R}^n$. Since e_o is a feasible solution by Condition 5, $e^* = e_o$. It follows that (GD) and (IGD) are equivalent. □

Lemma 3. *If Conditions 1, 2 and 4 hold with $\omega = \phi(\theta)$, for all $\theta \in \Theta_n$, then, the first-order conditions of (GP) are $\sum_{i=1}^n x_{ij}^c \widetilde{h}_j\{e(y_i, x_i, \theta)\} = 0$ $(j = 1, \ldots, J)$, and the Hessian matrix of the objective function of (GP) is positive definite for all $\theta \in \Theta_n$.*

Proof. For $\theta \in \Theta_n$, define $Q_n(\theta) = \sum_{i=1}^n L(x_i, y_i, \theta)$, with $L(x_i, y_i, \theta)$ defined as

$$L(x_i, y_i, \theta) = \sum_{j=2}^{J} (\theta_j^{\mathrm{T}} x_{ij}^c) \left[h_j\{e(y_i, x_i, \theta)\} e(y_i, x_i, \theta) - \widetilde{h}_j\{e(y_i, x_i, \theta)\} \right], \tag{3.3}$$

and let $e_i = e(y_i, x_i, \theta)$, $\eta_j(e_i) = h_j(e_i)e_i - \widetilde{h}_j(e_i)$, $(i = 1, \ldots, n; j = 1, \ldots, J)$. For $j = 1, \ldots, J$ and $\theta \in \Theta_n$, the derivative of Q_n with respect to θ_j satisfies

$$\nabla_{\theta_j} Q_n(\theta) = \sum_{i=1}^n \sum_{l=2}^J (\theta_l^{\mathrm{T}} x_{il}^c) \eta_l'(e_i) \nabla_{\theta_j} e_i + \sum_{i=1}^n x_{ij}^c \eta_j(e_i) = \sum_{i=1}^n H_{x_i}'(e_i; \theta) e_i \nabla_{\theta_j} e_i + \sum_{i=1}^n x_{ij}^c \eta_j(e_i),$$

upon substituting $\eta_l' = h_l'(e_i)e_i$ and by definition of $H_{x_i}'(e_i; \theta)$. Using

$$\nabla_{\theta_j} e_i = -\nabla_{\theta_j} H_{x_i}(e_i; \theta)\{H_{x_i}'(e_i; \theta)\}^{-1} = -x_{ij}^c h_j(e_i)\{H_{x_i}'(e_i; \theta)\}^{-1},$$

and the definition of η_j, for all $\theta \in \Theta_n$ we obtain

$$\nabla_{\theta_j} Q_n(\theta) = -\sum_{i=1}^n x_{ij}^c h_j(e_i) e_i + \sum_{i=1}^n x_{ij}^c \{h_j(e_i) e_i - \widetilde{h}_j(e_i)\} = -\sum_{i=1}^n x_{ij}^c \widetilde{h}_j(e_i), \quad (j = 1, \ldots, J). \tag{3.4}$$

Letting $p_i = (1, e_i)^{\mathrm{T}}$ and $q_i = \{h_3(e_i), \ldots, h_J(e_i)\}^{\mathrm{T}}$, upon using (3.4) the Hessian matrix is

$$H_n = \sum_{i=1}^n \begin{bmatrix} \frac{x_i^c x_i^{c\mathrm{T}}}{H_{x_i}'(e_i;\theta)} \otimes p_i p_i^{\mathrm{T}} & \frac{x_i^c \widetilde{x}_i^{c\mathrm{T}}}{H_{x_i}'(e_i;\theta)} \otimes p_i q_i^{\mathrm{T}} \\ \frac{\widetilde{x}_i^c x_i^{c\mathrm{T}}}{H_{x_i}'(e_i;\theta)} \otimes q_i p_i^{\mathrm{T}} & \frac{\widetilde{x}_i^c \widetilde{x}_i^{c\mathrm{T}}}{H_{x_i}'(e_i;\theta)} \otimes q_i q_i^{\mathrm{T}} \end{bmatrix} \equiv \begin{bmatrix} H_{11,n} & H_{12,n} \\ H_{21,n} & H_{22,n} \end{bmatrix}.$$

Suppose $H_{11,n}$ is positive definite for all $\theta \in \Theta_n$. Positive definiteness of $H_{11,n}$ then implies that H_n is positive definite for all $\theta \in \Theta_n$ if and only if the Schur complement of $H_{11,n}$ in H_n is positive definite (Boyd & Vandenberghe, 2004, Appendix A.5.5) for all $\theta \in \Theta_n$, i.e., if and only if the determinant of $D_n = H_{22,n} - H_{21,n} H_{11,n}^{-1} H_{12,n}$ is strictly positive, for all $\theta \in \Theta_n$. Letting $\Xi_n = H_{21,n} H_{11,n}^{-1}$ for all $\theta \in \Theta_n$, D_n is equal to

$$\sum_{i=1}^n \left[\left\{ \frac{\widetilde{x}_i^c \otimes q_i}{\{H_{x_i}'(e_i;\theta)\}^{1/2}} - \Xi_n \frac{x_i^c \otimes p_i}{\{H_{x_i}'(e_i;\theta)\}^{1/2}} \right\} \left\{ \frac{\widetilde{x}_i^c \otimes q_i}{\{H_{x_i}'(e_i;\theta)\}^{1/2}} - \Xi_n \frac{x_i^c \otimes p_i}{\{H_{x_i}'(e_i;\theta)\}^{1/2}} \right\}^{\mathrm{T}} \right], \tag{3.5}$$

a positive semidefinite matrix, and equal to zero if and only if

$$\widetilde{x}_i^c \otimes q_i = \Xi_n (x_i^c \otimes p_i) \quad (i = 1, \ldots, n); \tag{3.6}$$

this is an application of the Cauchy-Schwarz inequality for matrices stated in Tripathi (1999). Under Condition 4, system (3.6) cannot hold, with probability 1, for all $\theta \in \Theta_n$.

Finally, a similar argument shows that, under Condition 2, $H_{11,n}$ is positive definite for all $\theta \in \Theta_n$ if and only if

$$x_i^c e_i = \Upsilon_n x_i^c \quad (i = 1, \ldots, n), \tag{3.7}$$

where

$$\Upsilon_n = \left[\sum_{i=1}^n \frac{x_i^c x_i^{c\mathrm{T}}}{\{H'_{x_i}(e_i;\theta)\}^{1/2}} e_i\right] \left[\sum_{i=1}^n \frac{x_i^c x_i^{c\mathrm{T}}}{\{H'_{x_i}(e_i;\theta)\}^{1/2}}\right]^{-1}.$$

In particular, with $\Upsilon_{n,1}$ denoting the first row of Υ, since x_i^c includes an intercept system (3.7) implies $e_i = \Upsilon_{n,1} x_i^c$ $(i = 1, \ldots, n)$, for all $\theta \in \Theta_n$, which cannot hold under Condition 1, with probability 1. □

3.3. **Proof of Theorem 3.** The equivalence result follows by Lemma 2. For $\theta \in \mathbb{R}^{2+J(K-1)}$, define the Lagrangian for (GD) as

$$\mathscr{L}(e,\theta) = \sum_{i=1}^n \{y_i - \theta_1^{\mathrm{T}} x_i^c\} e_i - \frac{1}{2}\sum_{i=1}^n (\theta_2^{\mathrm{T}} x_i^c)(e_i^2 - 1) - \sum_{i=1}^n \sum_{j=3}^J (\theta_j^{\mathrm{T}} \widetilde{x}_i^c)\widetilde{h}_j(e_i) \quad (e \in \mathbb{R}^n),$$

with n first-order conditions

$$y_i = H_{x_i}(e_i;\theta) \quad (i = 1, \ldots, n), \tag{3.8}$$

and denote any vector in $\mathbb{R}^n$ satisfying (3.8) by $e(\theta)$, and the ith element of $e(\theta)$ by $e(y_i, x_i, \theta)$.

Proof of part (i). Define the Lagrange dual function (Boyd & Vandenberghe, 2004, Chapter 5) $Q_n(\theta) \equiv \sup_{e\in\mathbb{R}^n} \mathscr{L}(e,\theta)$ for $\theta \in \Theta_n$. In order to derive $Q_n(\theta)$, we show that the maximum of the mapping $e \mapsto \mathscr{L}(e,\theta)$ is attained and is unique, and evaluate $e \mapsto \mathscr{L}(e,\theta)$ at this value.

Step 1. We show that the map $e \mapsto \mathscr{L}(e,\theta)$ admits at least one maximum in $\mathbb{R}^n$ for all $\theta \in \Theta_{0,n}$. Since $\Theta_n \subseteq \Theta_{0,n}$, existence of a maximum then holds for all $\theta \in \Theta_n$. For $\theta \in \Theta_{0,n}$ and $c \in \mathbb{R}$, consider the level sets $\mathcal{B}_c(\theta) = \{e \in \mathbb{R}^n : -\mathscr{L}(e,\theta) \le c\}$ of $-\mathscr{L}$. These sets are compact. Consider a sequence $(e_{(m)})$ in $\mathbb{R}^n$ such that $||e_{(m)}|| \to \infty$ as $m \to \infty$. Let $z_{(m)} = \frac{e_{(m)}}{||e_{(m)}||}$, a bounded sequence with unit norm. By the Bolzano-Weierstrass theorem there exists a convergent subsequence $z_{(m_l)}$, $m_l \to \infty$ as $l \to \infty$, with limit z_o, say. Then, using that $\theta_2^{\mathrm{T}} x_i > 0$ $(i = 1, \ldots, n)$, and $\max_{j=3,\ldots,J} ||\widetilde{h}_j||_\infty$ is bounded, for $\theta \in \Theta_{0,n}$

$$\begin{aligned} -\mathscr{L}(e_{(m_l)},\theta) = &-||e_{(m_l)}|| \sum_{i=1}^n \{y_i - \theta_1^{\mathrm{T}} x_i\} z_{i,(m_l)} + ||e_{(m_l)}||^2 \sum_{i=1}^n \frac{1}{2}(\theta_2^{\mathrm{T}} x_i) z_{i,(m_l)}^2 \\ &+ \frac{1}{2}\sum_{i=1}^n (\theta_2^{\mathrm{T}} x_i) + \sum_{i=1}^n \sum_{j=3}^J (\theta_j^{\mathrm{T}} \widetilde{x}_i)\widetilde{h}_j(||e_{(m_l)}|| z_{i,(m_l)}) \to \infty \end{aligned}$$

as $l \to \infty$, since $||e_{(m_l)}||^2 \frac{1}{2}\sum_{i=1}^n (\theta_2^{\mathrm{T}} x_i) z_{i,o}^2 \to \infty$ as $l \to \infty$. Therefore $-\mathscr{L}(e,\theta)$ grows unboundedly as $||e|| \to \infty$, and $\mathcal{B}_c(\theta)$ is bounded. Since $e \mapsto -\mathscr{L}(e,\theta)$ is continuous over $\mathbb{R}^n$ for $\theta \in \Theta_n$, $\mathcal{B}_c(\theta)$ is also closed. It then follows from the Weierstrass theorem and

$\Theta_n \subseteq \Theta_{0,n}$ that there exists $e(\theta) \in \arg\min_{e\in\mathbb{R}^n}\{-\mathscr{L}(e,\theta)\} = \arg\max_{e\in\mathbb{R}^n}\mathscr{L}(e,\theta)$, for all $\theta \in \Theta_n$.

Step 2. The Hessian matrix of the map $e \mapsto \mathscr{L}(e,\theta)$ is $-\text{diag}[H'_{x_i}\{e(y_i,x_i,\theta);\theta\}]$, and is thus negative definite for all $\theta \in \Theta_n$. Therefore, $e \mapsto \mathscr{L}(e,\theta)$ is strictly concave with unique maximum $e(\theta)$ for all $\theta \in \Theta_n$. With $L(x_i,y_i,\theta)$ defined in (3.3), and upon using first-order conditions (3.8), direct substitution yields $\mathscr{L}\{e(\theta),\theta\} = \sum_{i=1}^n L(x_i,y_i,\theta)$, the maximum of the map $e \mapsto \mathscr{L}(e,\theta)$ for all $\theta \in \Theta_n$, and the primal objective function.

Proof of part (ii). By Lemma 3, the first-order conditions of (GP) implied by (3.4) coincide with the system

$$\sum_{i=1}^n x_{ij}^c \widetilde{h}_j(e_i) = 0 \quad (j=1,\ldots,J), \quad y_i = \sum_{j=1}^J (\theta_j^{\mathrm{T}} x_{ij}^c) h_j(e_i) \quad (i=1,\ldots,n). \tag{3.9}$$

Moreover, the n first-order conditions (3.8) and the constraints of (GD) together yield the method-of-moments representation of (GD).

Proof of part (iii). (a) By Condition 5, there exists $\theta \in \Theta_n$ such that first-order conditions (3.9) are satisfied. By Lemma 3, $Q_n(\theta)$ is strictly convex over Θ_n. Therefore, θ_n is the unique minimum of $Q_n(\theta)$ and uniquely solves (3.9).

By definition, a solution e^* to (GD) with Lagrange multiplier θ^* satisfies first-order conditions (3.8). Suppose $\theta^* \in \Theta_n$. By Step 2 in part (i), the map $e \mapsto \mathscr{L}(e,\theta)$ admits a unique maximizer $e(\theta)$, for all $\theta \in \Theta_n$: each pair $\{\theta, e(\theta)\}$ is well-defined and satisfies first-order conditions (3.8). Since θ_n uniquely solves (3.9) over Θ_n, the pair $\{\theta_n, e(\theta_n)\}$ is the unique pair satisfying system (3.9) in $\Theta_n \times \mathcal{E}_n$, where $\mathcal{E}_n = \{e \in \mathbb{R}^n : y_i = H_{x_i}(e_i;\theta) \quad (i = 1,\ldots,n), \text{ for some } \theta \in \Theta_n\}$ is the set of admissible optimal solutions to (GD). It follows that the pair $(\theta^*, e^*) = \{\theta_n, e(\theta_n)\}$ is the unique pair satisfying system (3.9) in $\Theta_n \times \mathcal{E}_n$. Therefore, the pair (θ_n, e^*) uniquely solves (GP) and (GD) over $\Theta_n \times \mathcal{E}_n$.

Suppose $\theta^* \notin \Theta_n$. For $\theta \notin \Theta_n$, a pair $\{\theta, e(\theta)\}$ (not necessarily unique) does not satisfy the second-order conditions of (GD). Thus a solution to (GD) with Lagrange multipliers $\theta^* \notin \Theta_n$ is not a global maximum of (GD) over $\mathbb{R}^n$. Thus there is no solution to (GD) such that both $\theta^* \notin \Theta_n$ and the value of (GD) is equal to or exceeds the optimal value of (GD) at $e^* = e(\theta_n)$. Therefore, the pair $\{\theta_n, e^*\}$ is the unique optimal solution to (GP) and (GD) over $\Theta_n \times \mathbb{R}^n$.

(b) By direct substitution and using that $\sum_{i=1}^n x_i^c e_i^* = 0$, at a solution the value of (GD) is $\sum_{i=1}^n y_i e_i^* = \sum_{i=1}^n \sum_{j=2}^J (\theta_j^{*\mathrm{T}} x_{ij}^c) h_j(e_i^*) e_i^*$. Using that $\sum_{i=1}^n (\theta_{jn}^{\mathrm{T}} x_{ij}^c)\widetilde{h}_j\{e(y_i,x_i,\theta_n)\} = 0$ $(j = 1,\ldots,J)$, at a solution the value of (GP) is $\sum_{i=1}^n \sum_{j=2}^J (\theta_{jn}^{\mathrm{T}} x_{ij}^c) h_j\{e(y_i,x_i,\theta_n)\} e(y_i,x_i,\theta_n)$. Strong duality then follows from $\theta_n = \theta^*$ established in (a).

4. Proof of Theorem 4

Proof. Proof of part (i). Let U satisfy $U \sim U(0,1)$ and $E(X \mid U) = E(X)$. Then:

$$E\{X \otimes m^J(U)\} = E\{E(X \mid U) \otimes m^J(U)\} = E\{E(X) \otimes m^J(U)\} = E(X) \otimes E\{m^J(U)\} = 0$$

for all J. The first equality holds by iterated expectations, the second by mean independence, the third by linearity of the expectation, and the last by uniformity of U and definition of m^J.

In order to show the converse statement, suppose that $E(X \mid U) = E(X)$ does not hold. Following steps similar to the proof of Lemma 2.1 in Donald et al. (2003), and letting $\varphi(U) = E(X \mid U) - E(X)$, for Ψ_J such that $E[||\varphi(U) - \Psi_J m^J(U)||^2] \to 0$,

$$E\left[m^J(U)^{\mathrm{T}} \Psi_J^{\mathrm{T}} \{X - E(X)\}\right] = E\left[m^J(U)^{\mathrm{T}} \Psi_J^{\mathrm{T}} \{E(X \mid U) - E(X)\}\right] \to E[||\varphi(U)||^2] > 0,$$

as $J \to \infty$, which implies $E[X \otimes m^J(U)] \neq 0$ for all J large enough, since

$$E\left[m^J(U)^{\mathrm{T}} \Psi_J^{\mathrm{T}} \{X - E(X)\}\right] = E[\{X - E(X)\} \otimes m^J(U)]\mathrm{vec}(\Psi_J) = E\{X \otimes m^J(U)\}\mathrm{vec}(\Psi_J).$$

Now suppose that $U \sim U(0,1)$ does not hold. Because X includes an intercept, any random variable $\widetilde{U}$ such that $E\{X \otimes m^J(\widetilde{U})\} = 0$ for all J must also satisfy $E\{m^J(\widetilde{U})\} = 0$ for all J, and therefore $\widetilde{U} \sim U(0,1)$. It follows that $E\{X \otimes m^J(U)\} \neq 0$ in the large J limit.

Therefore, $E\{X \otimes m^J(U)\} = 0$ for all J if and only if $E(X \mid U) = E(X)$ and $U \sim U(0,1)$, and the result follows.

Proof of part (ii). Let e be a random variable with mean 0 and variance 1 satisfying $E(\widetilde{X} \mid e) = 0$. Then $E\{\widetilde{X}^c \otimes \widetilde{h}^J(e)\} = E\{E(\widetilde{X}^c \mid e) \otimes \widetilde{h}^J(e)\} = 0$, for all J, by iterated expectations and mean independence.

In order to show the converse statement, suppose that $E(\widetilde{X}^c \mid e) \neq 0$. Letting $\varphi(e) = E(\widetilde{X}^c \mid e)$ and Ψ_J such that $E\{||\varphi(e) - \Psi_J \widetilde{h}^J(e)||^2\} \to 0$, and following steps similar to the proof of Lemma 2.1 in Donald et al. (2003),

$$E\{\widetilde{X}^c \otimes \widetilde{h}^J(e)\}\mathrm{vec}(\Psi_J) = E\{\widetilde{h}^J(e)^{\mathrm{T}} \Psi_J^{\mathrm{T}} \widetilde{X}^c\} = E\{\widetilde{h}^J(e)^{\mathrm{T}} \Psi_J^T E(\widetilde{X}^c \mid e)\} \to E\{||\varphi(e)||^2\} > 0,$$

as $J \to \infty$, which implies $E\{\widetilde{X}^c \otimes \widetilde{h}^J(e)\} \neq 0$ as $J \to \infty$.

Therefore, a random variable e with mean 0 and variance 1 satisfies $E\{\widetilde{X}^c \otimes \widetilde{h}^J(e)\} = 0$ for all J large enough if and only if $E(\widetilde{X}^c \mid e) = 0$, and the result follows. □

5. Asymptotic Theory

In this Section, C denotes a generic constant whose value may vary from place to place.

5.1. **Proof of Theorem 5.** Letting $e = e(Y, X, \theta)$ for $\theta \in \Theta$, by definition (3.3), $L(X, Y, \theta)$ can be decomposed as

$$L(X,Y,\theta) = \frac{1}{2}(\theta_2^{\mathrm{T}} X^c)(e^2+1)+\sum_{j=3}^{J}(\theta_j^{\mathrm{T}} \widetilde{X}^c)\{h_j(e)e-\widetilde{h}_j(e)\} \equiv L_1(X,Y,\theta)+L_2(X,Y,\theta). \tag{5.1}$$

Define $Q_0(\theta) = E\{L(X, Y, \theta)\}$, the population objective of the generalized primal problem.

Both existence and consistency of $\hat{\theta}$ result from strict convexity of $Q_0(\theta)$, and pointwise convergence of $Q_n(\theta)$ to $Q_0(\theta)$, since strict convexity and pointwise convergence together imply uniform convergence, as in, for instance, Theorem 2.7 in Newey & Mc Fadden (1994). The asymptotic distribution of $\hat{\theta}$ follows from the method-of-moments characterization of the estimates given in part (ii) of Theorem 3, and Theorem 3.4 in Newey & Mc Fadden (1994).

*Proof of parts (i) and (ii).*We verify the conditions of Theorem 2.7 in Newey & Mc Fadden (1994). We first show that θ_0 is the unique minimizer of $Q_0(\theta)$ in Θ, using the next result.

Lemma 4. *Suppose that Conditions 1, 2, 4, 7 and 8(i) hold. Then, $Q_0(\theta)$ is continuously differentiable, $E\{|L(X, Y, \theta)|\} < \infty$ and $\nabla_\theta E\{L(X, Y, \theta)\} = E\{\nabla_\theta L(X, Y, \theta)\}$ for $\theta \in \Theta$.*

Proof. We first show that $E\{|L(X, Y, \theta)|\} < \infty$ for all $\theta \in \Theta$. $|L_1|$ in (5.1) satisfies

$$\left|\frac{1}{2}(\theta_2^{\mathrm{T}} X^c)(e^2+1)\right| \leq \frac{1}{2}||\theta_2||\,||X^c||\,(e^2+1), \tag{5.2}$$

which has finite expectation if $E(||X^c||e^2)$ and $E(||X^c||)$ are bounded. Since $\{h_j\}_{j=3,\ldots,J}$ and $\{\widetilde{h}_j\}_{j=3,\ldots,J}$ are bounded, $|L_2|$ in (5.1) satisfies

$$\left|\sum_{j=3}^{J}(\theta_j^{\mathrm{T}} \widetilde{X}^c)\{h_j(e)e-\widetilde{h}_j(e)\}\right| \leq C\sum_{j=3}^{J}||\theta_j||(||X^c||\,|e|+||X^c||), \tag{5.3}$$

which has finite expectation if $E(||X^c||\,|e|)$ and $E(||X^c||)$ are bounded. It follows that $|L(X, Y, \theta)|$ has finite expectation if $E(||X^c||e^2) < \infty$.

The identity $Y = \theta_1^{\mathrm{T}} X^c + (\theta_2^{\mathrm{T}} X^c)e + \sum_{j=3}^{J}(\theta_j^{\mathrm{T}} \widetilde{X}^c)h_j(e)$ holds with probability one for $\theta \in \Theta$, and $\{h_j\}_{j=3,\ldots,J}$ bounded thus implies

$$\begin{aligned} |e^2| &= |(\theta_2^{\mathrm{T}} X^c)^{-2}\{Y-\theta_1^{\mathrm{T}} X^c-\sum_{j=3}^{J}(\theta_j^{\mathrm{T}} \widetilde{X}^c)h_j(e)\}^2| \\ &\leq C\{\inf_{x\in\mathcal{X}}(\theta_2^{\mathrm{T}} x^c)\}^{-2}\{2|Y|^2+2(||\theta_1||^2\,||X^c||^2+\sum_{j=3}^{J}||\theta_j||^2\,||X^c||^2)\}. \end{aligned} \tag{5.4}$$

Therefore,

$$E(||X^c||\,|e^2|) \leq C\{\inf_{x\in\mathcal{X}}(\theta_2^{\mathrm{T}} x^c)\}^{-2}E(||X^c||\,|Y|^2+||\theta_1||^2\,||X^c||^3+\sum_{j=3}^{J}||\theta_j||^2\,||X^c||^3) < \infty.$$

Bounds (5.2) and (5.3) now imply $E\{|L_1(X,Y,\theta)|\} < \infty$ and $E\{|L_2(X,Y,\theta)|\} < \infty$, for all $\theta \in \Theta$ since Θ is bounded. Hence $E\{|L(X,Y,\theta)|\} < \infty$ for all $\theta \in \Theta$.

Bound (5.4) implies that $E\{\sup_{\theta\in\Theta} ||\nabla_\theta L(X,Y,\theta)||\} < \infty$. By Lemma 3, $\nabla_{\theta_1} L(X,Y,\theta) = -X^c e$ and $\nabla_{\theta_2} L(X,Y,\theta) = -X^c(e^2-1)/2$, and $\nabla_{\theta_j} L(X,Y,\theta) = -\widetilde{X}^c \widetilde{h}_j(e)$. Bound (5.4) together with $\{\widetilde{h}_j\}_{j=3,\ldots,J}$ bounded, boundedness of Θ and Holder's inequality thus imply that $E\{\sup_{\theta\in\Theta} ||\nabla_\theta L(X,Y,\theta)||\} < \infty$ under Condition 8(i). Lemma 3.6 in Newey & Mc Fadden (1994) then implies that $Q_0(\theta)$ is continuously differentiable and that the order of differentiation and integration can be interchanged for $\theta \in \Theta$. □

By Lemma 4, $Q_0(\theta)$ is continuously differentiable and the order of differentiation and integration can be interchanged, for $\theta \in \Theta$. Moreover, $\nabla_\theta Q_0(\theta)$ is differentiable for $\theta \in \Theta$. Letting $P = (1, e)^{\mathrm{T}}$ and $Q = \{h_3(e), \ldots, h_J(e)\}^{\mathrm{T}}$, from the proof of Lemma 3,

$$\nabla_{\theta\theta} L(X,Y,\theta) = \begin{bmatrix} \frac{X^c X^{c\mathrm{T}}}{H'_X(e;\theta)} \otimes PP^{\mathrm{T}} & \frac{X^c \widetilde{X}^{c\mathrm{T}}}{H'_X(e;\theta)} \otimes PQ^{\mathrm{T}} \\ \frac{\widetilde{X}^c X^{c\mathrm{T}}}{H'_X(e;\theta)} \otimes QP^{\mathrm{T}} & \frac{\widetilde{X}^c \widetilde{X}^{c\mathrm{T}}}{H'_X(e;\theta)} \otimes QQ^{\mathrm{T}} \end{bmatrix}.$$

Applying steps similar to those leading to the bound (5.4) in the proof of Lemma 4 and using that $\inf_{e\in\mathbb{R}} H'_X(e;\theta) > 0$ for all $\theta \in \Theta$ shows that $||\{X^c X^{c\mathrm{T}}/H'_X(e;\theta)\}e^2||$ has finite expectation for all $\theta \in \Theta$ under Condition 8(i). Therefore, boundedness of $\{h_j\}_{j=3,\ldots,J}$ and Θ, and Holder's inequality imply that $E\{\sup_{\theta\in\Theta} ||\nabla_{\theta\theta} L(X,Y,\theta)||\} < \infty$. It then follows from Lemma 3.6 in Newey & Mc Fadden (1994) that $\nabla_\theta Q_0(\theta)$ is continuously differentiable, and that the Hessian matrix of $Q_0(\theta)$ is $H(\theta) = E\{\nabla_{\theta\theta} L(X,Y,\theta)\}$, which is a finite positive definite matrix under the assumed conditions, by application of Lemma 3. Therefore, $Q_0(\theta)$ is strictly convex and θ_0 is the unique minimizer of $Q_0(\theta)$ in Θ, and Condition (i) of Newey and McFadden's Theorem 2.7 is verified.

By Condition 6, θ_0 is in the interior of Θ, which is convex, and $Q_n(\theta)$ is convex with probability 1 by Lemma 3, and their Condition (ii) is verified. Finally, since the sample is independently and identically distributed by assumption, pointwise convergence of $Q_n(\theta)$ to $Q_0(\theta)$ follows from boundedness of $Q_0(\theta)$, established in the proof of Lemma 4, and application of Khinchine's law of large numbers. All conditions of Newey and McFadden's Theorem 2.7 are therefore satisfied, and there exists $\hat{\theta} \in \Theta$ with probability approaching one and $\hat{\theta}$ converges in probability to θ_0.

Proof of part (iii). Define $m_j(Y,X,\theta) = X^c \widetilde{h}_j\{e(Y,X,\theta)\}$ for $j = 1, 2$, and $m_j(Y,X,\theta) = \widetilde{X}^c \widetilde{h}_j\{e(Y,X,\theta)\}$ for $j = 3, \ldots, J$, and let $m(Y,X,\theta) = \{m_1(Y,X,\theta), \ldots, m_J(Y,X,\theta)\}^{\mathrm{T}}$, $G = E\{\nabla_\theta m(Y,X,\theta)\}|_{\theta=\theta_0}$ and $S = E\{m(Y,X,\theta_0)m(Y,X,\theta_0)^{\mathrm{T}}\}$. By part (ii) of Theorem 3, the Lagrange multiplier vector $\hat{\theta}$ solves the $2 + J(K-1)$ equations system

$$\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \theta) = 0. \tag{5.5}$$

System (5.5) can be equivalently viewed as minimizing

$$\mathcal{Q}_n^{MM}(\theta) \;=\; \left\{\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \theta)\right\}^{\mathrm{T}} \left\{\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \theta).\right\}$$

Asymptotic normality of the method-of-moments estimator then follows after verifying conditions of Theorem 3.4 in Newey & Mc Fadden (1994).

From the proof of Lemma 3, the derivative of e with respect to θ_j is $\nabla_{\theta_j} e = -X^c h_j(e)\{H'_X(e;\theta)\}^{-1}$, for $j = 1, 2$, and $\nabla_{\theta_j} e = -\widetilde{X}^c h_j(e)\{H'_X(e;\theta)\}^{-1}$, for $j = 3, \ldots, J$, which is continuous for all $\theta \in \Theta$ with probability one by definition of e and Θ. Thus the mapping $\theta \mapsto m(Y, X, \theta)$ is continuously differentiable in $\theta \in \Theta$ with probability one, and Newey and McFadden's Condition (ii) is satisfied. By definition $\varepsilon = e(Y, X, \theta_0)$ is independent of X which implies that $E\{m(Y, X, \theta_0)\} = 0$, and the first part of their Condition (iii) is satisfied. In addition, steps similar to the proof of Lemma 4 show that $E\{||m(Y, X, \theta_0)||^2\}$ and $E\{\sup_{\theta\in\Theta} ||\nabla_\theta m(Y, X, \theta)||\}$ are finite under Conditions 8, and their Conditions (iii)–(iv) are thus verified. Finally, their full rank condition on $G = E\{\nabla_\theta m(Y, X, \theta)\}|_{\theta=\theta_0}$ is satisfied under our conditions since $G = H(\theta_0)$ is then positive definite. Therefore, $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to $N(0, \Sigma)$ with

$$\Sigma = G^{-1} S (G^{-1})^{\mathrm{T}}. \tag{5.6}$$

5.2. **Proof of Theorem 6.** *Proof of part (i).* Denote the cumulative distribution function of $\hat{e} = e(Y, X, \hat{\theta})$, by $\hat{F}(e) = E\{1(\hat{e} \le e)\}$ for all $e \in \mathbb{R}$, and consider the decomposition

$$F_n(e) - F(e) = \{F_n(e) - \hat{F}(e)\} + \{\hat{F}(e) - F(e)\} \qquad (e \in \mathbb{R}). \tag{5.7}$$

For the first term, convergence in probability of $\sup_e |\hat{F}(e) - F(e)|$ to 0 is implied by Glivenko-Cantelli (e.g., Theorem 19.1 in van der Vaart, 1998). For the second term, upon using that the events $\{e(Y, X, \theta) \le e\}$ and $\{Y \le H_X(e;\theta)\}$ are equivalent conditional on X for $\theta \in \Theta$, and in particular for $\hat{\theta}, \theta_0 \in \Theta$, applying iterated expectations, a change of variable and a mean-value expansion, yields

$$\begin{aligned}
\hat{F}(e) - F(e) &= E[E\{1(\hat{e} \le e) \mid X\} - E\{1(\varepsilon \le e) \mid X\}] \\
&= E[F_{Y|X}\{H_X(e;\hat{\theta}) \mid X\} - F_{Y|X}\{H_X(e;\theta_0)\} \mid X] \\
&= (\hat{\theta} - \theta_0)^{\mathrm{T}} E[f_{Y|X}\{H_X(e;\bar{\theta}) \mid X\} m\{H_X(e;\theta), X, \theta\}],
\end{aligned}$$

where $\bar{\theta}$ is on the line connecting $\hat{\theta}$ and θ_0. Since $\sup_e e^2 f_{Y|X}\{H_X(e;\theta) \mid X\}$ and $\{\widetilde{h}_j\}_{j=3,\ldots,J}$ are uniformly bounded, it follows that

$$\sup_{e\in\mathbb{R}} |\hat{F}(e) - F_\varepsilon(e)| \le C||\hat{\theta} - \theta_0||E(||X^c||).$$

Consistency of $\hat{\theta}$ and $E(||X^c||)$ finite then imply convergence in probability of $\sup_e |\hat{F}(e) - F(e)|$ to 0. The result follows from combining the two uniform convergence results.

Proof of part (ii). For $D = (Y, X)$, let $\mathbb{E}_n f = \mathbb{E}_n f(d_i) = n^{-1}\sum_{i=1}^n f(d_i)$ and $\mathbb{G}_n f = \mathbb{G}_n\{f(d_i)\} = n^{-1/2}\sum_{i=1}^n [f(d_i) - E\{f(d_i)\}]$, and define the class of functions $\mathcal{F} = \{1\{e(Y, X, \theta) \leq e\}, e \in \mathbb{R}, \theta \in \Theta\}$. Following van der Vaart & Wellner (2007), the empirical dual regression process $\mathbb{U}_n(e) = n^{1/2}(\mathbb{E}_n f_{e,\hat{\theta}} - E f_{e,\theta_0})$ admits the following decomposition:

$$n^{1/2}(\mathbb{E}_n f_{e,\hat{\theta}} - E f_{e,\theta_0}) = \mathbb{G}_n(f_{e,\hat{\theta}} - f_{e,\theta_0}) + \mathbb{G}_n f_{e,\theta_0} + \sqrt{n} E(f_{e,\hat{\theta}} - f_{e,\theta_0}). \tag{5.8}$$

The proof thus proceeds by (i) establishing that the first term on the right in (5.8) converges in probability to zero, (ii) using the fact that the second term converges in distribution to a mean zero Gaussian process, and (iii) expanding the last term uniformly in $e \in \mathbb{R}$.

Step 1. Stochastic equicontinuity. By Theorem 2.1 in van der Vaart & Wellner (2007), since $\Pr(\hat{\theta} \in \Theta) \to 1$ by part (i) of Theorem 5, $\sup_{e\in\mathbb{R}} |\mathbb{G}_n(f_{e,\hat{\theta}} - f_{e,\theta_0})|$ converges in probability to 0 holds if the class of functions $\mathcal{F}$ is Donsker and if the pseudometric $\rho\{(e', \theta'), (e'', \theta'')\}^2 \equiv E[\{f_{e',\theta'}(D) - f_{e'',\theta''}(D)\}^2]$ satisfies $\delta_n \equiv \sup_{e\in\mathbb{R}} \rho\{(e, \hat{\theta}), (e, \theta_0)\}^2$ converges in probability to 0.

We first show that the class of functions $\mathcal{F}$ is Donsker. Define the parametric class of functions $\widetilde{\mathcal{F}} = \{e(Y, X, \theta), \theta \in \Theta\}$. For all $\theta', \theta'' \in \Theta$, a mean-value expansion and Cauchy-Schwarz inequality yield

$$|e(Y, X, \theta') - e(Y, X, \theta'')| \leq ||\nabla_\theta e(Y, X, \theta)|_{\theta=\bar{\theta}}|| \, ||\theta' - \theta''||,$$

where $\bar{\theta}$ is on the line joining θ' and θ''. Steps similar to those in the proof of Theorem 5 show that $E\{||\nabla_\theta e(Y, X, \theta)|_{\theta=\bar{\theta}}||^2\}$ is bounded under Condition 8, so that $\widetilde{\mathcal{F}}$ is Donsker by Example 19.7 in van der Vaart (1998). Therefore, $\mathcal{F}$ is Donsker, by monotonicity of the indicator function, with unit envelope.

We now show that δ_n converges in probability to 0. Since the events $\{e(Y, X, \theta) \leq e\}$ and $\{Y \leq H_X(e; \theta)\}$ are equivalent conditional on X for $\theta \in \Theta$, the law of iterated expectations, a mean-value expansion and Cauchy-Schwarz inequality yield:

$$\begin{aligned}
\sup_{e\in\mathbb{R}} \rho\{(e, \hat{\theta}), (e, \theta_0)\}^2 &= \sup_{e\in\mathbb{R}} E[|1\{e(Y, X, \hat{\theta}) \leq e\} - 1\{e(Y, X, \theta_0) \leq e\}|] \\
&= \sup_{e\in\mathbb{R}} E(|(\hat{\theta} - \theta_0)^{\mathrm{T}}[f_{Y|X}\{H_X(e; \bar{\theta}) \mid X\} m\{H_X(e; \theta), X, \theta\}]|) \\
&\leq C||\hat{\theta} - \theta_0|| E(||X^c||),
\end{aligned}$$

where $\bar{\theta}$ is on the line joining $\hat{\theta}$ and θ_0. Convergence in probability of δ_n to zero now follows from $E(||X^c||)$ finite and consistency of $\hat{\theta}$.

Step 2. Expansion. Letting $g(e) = E[f_{Y|X}\{H_X(e; \theta_0) \mid x_i\} m\{H_X(e; \theta_0), X, \theta_0\}]$, $e \in \mathbb{R}$, we show that the following expansion is valid uniformly in $e \in \mathbb{R}$:

$$E\{f_{e,\hat{\theta}}(D) - f_{e,\theta_0}(D)\} = (\hat{\theta} - \theta_0)^{\mathrm{T}}\{g(e) + o_P(1)\}. \tag{5.9}$$

Steps similar to above yield:

$$E\{f_{e,\hat{\theta}}(D) - f_{e,\theta_0}(D)\} = (\hat{\theta} - \theta_0)^{\mathrm{T}} E[\nabla_\theta F_{Y|X}\{H_{x_i}(e;\theta) \mid x_i\}|_{\theta=\bar{\theta}}],$$

where $\bar{\theta}$ is on the line joining $\hat{\theta}$ and θ_0. We obtain

$$E[f_{Y|X}\{H_X(e;\bar{\theta}) \mid x_i\} m\{H_X(e;\theta), X, \theta\}] = E[f_{Y|X}\{H_X(e;\theta_0) \mid X\} m\{H_X(e;\theta), X, \theta\}] + o_p(1),$$

uniformly in $e \in \mathbb{R}$, by uniform continuity of the mapping $y \mapsto f_{Y|X}(y \mid x)$, uniformly in x over $\mathcal{X}$, consistency of $\hat{\theta}$, and since $\sup_{e\in\mathbb{R}} e^2 f_{Y|X}\{H_X(e;\theta_0) \mid X\}$, $\max_{j=3,\ldots,J}\{\widetilde{h}_j\}$ are bounded and $E(||X^c||)$ is finite. Hence (5.9) holds by definition of $g(e)$, uniformly in $e \in \mathbb{R}$.

Finally, the method-of-moments representation of dual regression implies that the dual regression estimator $\hat{\theta}$ is asymptotically linear with influence function

$$\psi(Y, X, \theta_0) = -G^{-1} m(Y, X, \theta_0). \tag{5.10}$$

Thus (5.8)–(5.10) together imply that uniformly in $e \in \mathbb{R}$

$$\begin{aligned}
\mathbb{U}_n(e) &= \mathbb{G}_n(f_{e,\hat{\theta}} - f_{e,\theta_0}) + \mathbb{G}_n f_{e,\theta_0} + n^{1/2}(\hat{\theta} - \theta_0)^{\mathrm{T}}\{g(e) + o_P(1)\} \\
&= o_P(1) + \mathbb{G}_n f_{e,\theta_0} + g(e)^{\mathrm{T}} n^{-1/2} \sum_{i=1}^{n} \psi(y_i, x_i, \theta_0) + o_P(1) \\
&= n^{-1/2} \sum_{i=1}^{n} \varphi_e(y_i, x_i, \theta_0) + o_P(1),
\end{aligned}$$

where

$$\varphi_e(y_i, x_i, \theta_0) = 1\{e(y_i, x_i, \theta_0) \le e\} - F(e) - g(e)^{\mathrm{T}} G^{-1} m(y_i, x_i, \theta_0).$$

Therefore, the empirical dual regression process $\mathbb{U}_n$ weakly converges to the zero-mean Gaussian process $\mathbb{U}$, where $\mathbb{U}$ has covariance function

$$E\{\varphi_e(Y, X, \theta_0)\varphi_{e'}(Y, X, \theta_0)\}. \tag{5.11}$$

6. Numerical Illustrations

6.1. **Implementation of generalized dual regression.** Define the criterion

$$\mathrm{SIC}(J, n, \theta) = 2 \sum_{i=1}^{n} \sum_{j=2}^{J} (\theta_j^{\mathrm{T}} x_{ij}^c)\{h_j(e_i) e_i - \widetilde{h}_j(e_i)\} + \{2 + J(K-1)\} \log n,$$

where e_i solves $y_i = \sum_{j=1}^{J} (\theta_j^{\mathrm{T}} x_{ij}^c) h_j(e_i)$ $(i = 1, \ldots, n)$, denoted $e(y_i, x_i, \theta)$. For θ^* such that $H'_{x_i}\{e(y_i, x_i, \theta^*); \theta^*\} > 0$ holds for each $i = 1, \ldots, n$, using the strong duality result of Theorem 3, the value of the criterion can be computed as $y^{\mathrm{T}} e^* + \{2 + J(K-1)\} \log n$, with e^* the solution of the corresponding dual problem. We then select an even value of J according to the following alogorithm:

Step 1. For each J in the grid $\{2, 4, 6, 8\}$:

TABLE 1. Distribution of selected models across simulations in percentages.

	$n = 100$	$n = 235$	$n = 500$	$n = 1000$
$J^* = 2$	$74 \cdot 97$	$84 \cdot 14$	$87 \cdot 70$	$90 \cdot 32$
$J^* = 4$	$24 \cdot 00$	$15 \cdot 80$	$12 \cdot 30$	$9 \cdot 68$
$J^* = 6$	$0 \cdot 72$	$0 \cdot 06$	$0 \cdot 00$	$0 \cdot 00$
$J^* = 8$	$0 \cdot 30$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$

Step 1.1. Run program (GD) with basis functions specified as

$$h_j(e) = \begin{cases} \cos\{2\pi(j-2)e\} & j \text{ odd} \\ \sin\{2\pi(j-2)e\} & j \text{ even} \end{cases} \qquad (e \in \mathbb{R}),$$

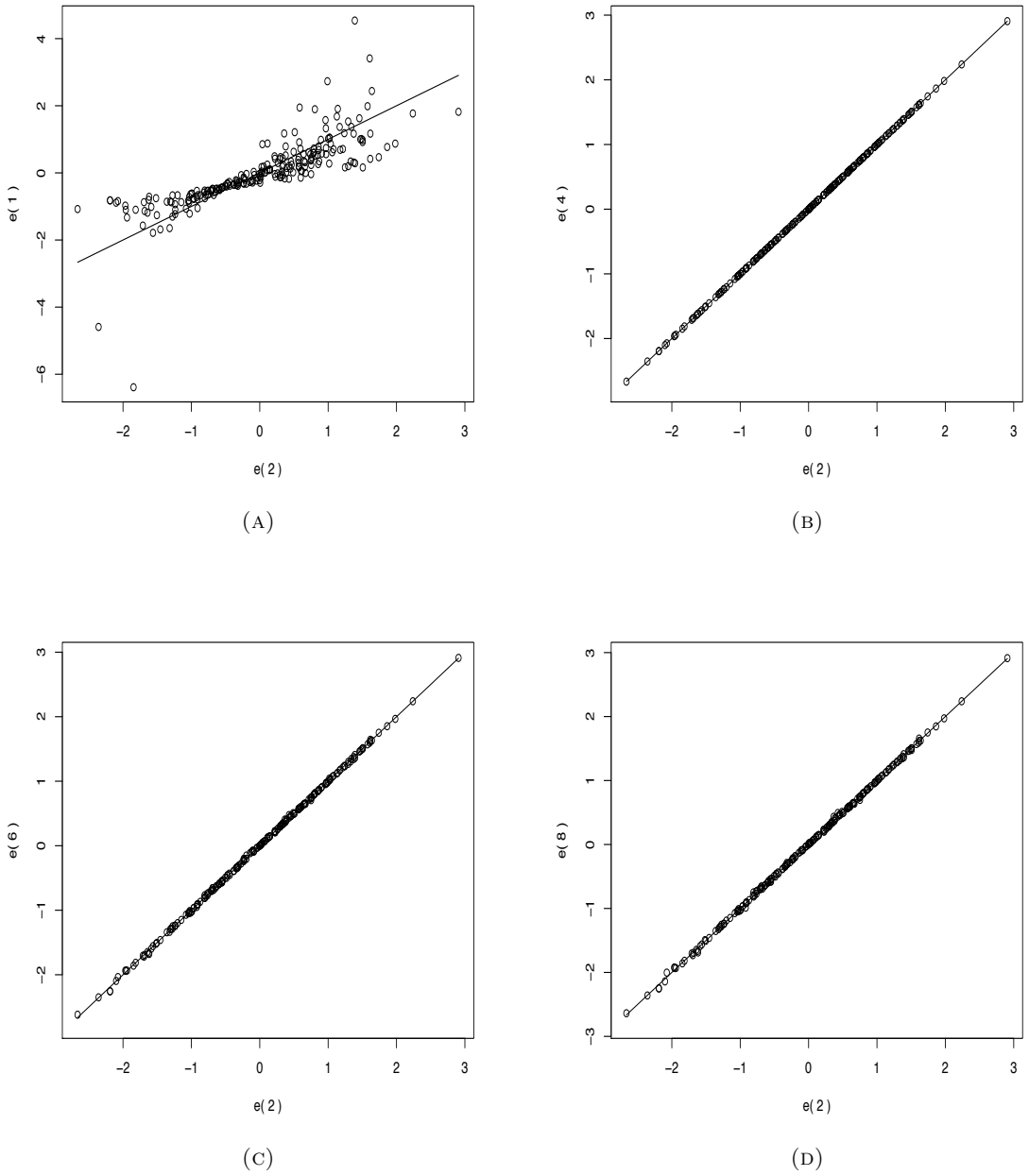
for $j = 3, \ldots, J$, and J even. Denote the solution by $e(J)$, with corresponding multipliers $\theta(J)$.

Step 1.2. Compute $\text{SIC}\{J, n, \theta(J)\} = y^{\mathrm{T}} e(J) + \{2 + J(K-1)\} \log n$.

Step 2. Select the value of J that minimizes $\text{SIC}\{J, n, \theta(J)\}$, denoted J^*.

Results in the empirical application are robust to using a larger grid for J. The grid specified above is also used in all simulations. Although the proposed algorithm provides a convenient semi-automated method for the specification of a generalized dual regression representation, it is also instructive to examine the solutions obtained for J greater than two. Fig. 6.1 plots the solutions $e(4)$, $e(6)$ and $e(8)$ obtained in Step 1.1. against the selected solution $e(2)$ in Step 2. We also plot the solution obtained for a location model, denoted $e(1)$ and obtained as $F_n\{(y_i - \hat{\gamma}_1 - \hat{\lambda}_1 x_i^c)/\hat{\gamma}_2\}$. Visual inspection then confirms that although the dual regression solution $e(2)$ differs significantly from the location solution $e(1)$, our results are robust to the addition of extra terms in the representation.

6.2. **Design and implementation of the numerical simulations.** We generate 4999 datasets of size $n = 100, 235, 500, 1000$ according to the model $y_i = \alpha_1 + \beta_1 \widetilde{x}_i + (\alpha_2 + \beta_2 \widetilde{x}_i)\varepsilon_i$ with $\varepsilon_i \sim N(0,1)$ and $\widetilde{x}_i \sim U\{\min(\text{Income}), \max(\text{Income})\}$, calibrated to Engel's data. The value of β is set to the value of estimates obtained by the method suggested in Koenker & Xiao (2002): for a grid of $R = 235$ quantile indices $\{u_1, \ldots, u_R\}$, $\{\hat{\beta}_{0,QR}(u_r), \hat{\beta}_{1,QR}(u_r)\}^{\mathrm{T}}$ are estimated by quantile regression, and α and β are set equal to the estimates obtained from linear regression of $\{\hat{\beta}_{0,QR}(u_r), \hat{\beta}_{1,QR}(u_r)\}^{\mathrm{T}}$ on $[\{1, \Phi^{-1}(u_r)\} : 1, \ldots, R]^{\mathrm{T}}$, where Φ^{-1} is the inverse standard normal distribution. We set $\alpha = (86 \cdot 56, 0 \cdot 55)^{\mathrm{T}}$ and $\beta = (-22 \cdot 17, 0 \cdot 12)^{\mathrm{T}}$. Thus the quantile regression parameters are $\beta_0(u) = \alpha_1 + \alpha_2 \Phi^{-1}(u)$ and $\beta_1(u) = \beta_1 + \beta_2 \Phi^{-1}(u)$, and $F_{Y|X}(y \mid x) = \Phi\{(y - \alpha_1 - \beta_1 \widetilde{x})/(\alpha_2 + \beta_2 \widetilde{x})\}$. As a benchmark, $F_{Y|X}(y \mid x)$ is also estimated by applying the inversion procedure of Chernozhukov et al. (2010) to the quantile regression process, as $\hat{u}_i^{QR} = \epsilon + \int_\epsilon^{1-\epsilon} 1\{\hat{\beta}_{0,QR}(u) + \hat{\beta}_{1,QR}(u)\widetilde{x}_i \leq y_i\} du$, with $\epsilon = 0 \cdot 01$. Dual regression multipliers yield functional coefficients estimates $\beta_0^*(u) = (\gamma_1^* - \lambda_1^* \bar{x}) + (\gamma_2^* - \lambda_2^* \bar{x}) F_n^{-1}(u)$ and $\beta_1^*(u) = \sum_{j=1}^J \lambda_j^* h_j\{F_n^{-1}(u)\}$, where F_n^{-1} is the



(A) (B)

(C) (D)

FIGURE 6.1. $e(1)$, $e(4)$, $e(6)$ and $e(8)$ plotted against dual regression solutions $e(2)$.

empirical quantile function of e^* and $\bar{x} = n^{-1}\sum_{i=1}^{n} \widetilde{x}_i$, and with the transformed intercept coefficients accounting for the centering of $\widetilde{x}_i$ in the implementation of (GD).

Table 1 shows the distribution of selected models across simulations. The 2 and 4 terms representations are selected in most simulations, wth the proportion of incorrect selections decreasing from 25% to 10% as sample size increases. For completeness, Table 2 reports

TABLE 2. L^p estimation errors ($\times 100$) and ratios of L^p estimation errors ($\times 100$) of J term generalized dual ($L^p_{GDR(J)}$) and quantile regression (L^p_{QR}) estimates of $F_{Y|X}(y_i \mid x_i)$ $(i = 1, \ldots, n)$, for $p = 1, 2, \infty$ and $J = 4, 6, 8$.

Sample size	$L^1_{GDR(4)}$	$L^1_{GDR(4)}/L^1_{QR}$	$L^2_{GDR(4)}$	$L^2_{GDR(4)}/L^2_{QR}$	$L^\infty_{GDR(4)}$	$L^\infty_{GDR(4)}/L^\infty_{QR}$
$n = 100$	$4 \cdot 09$	$92 \cdot 89$	$5 \cdot 59$	$91 \cdot 87$	$21 \cdot 47$	$88 \cdot 16$
$n = 235$	$2 \cdot 69$	$91 \cdot 34$	$3 \cdot 71$	$89 \cdot 19$	$16 \cdot 75$	$80 \cdot 32$
$n = 500$	$1 \cdot 85$	$90 \cdot 42$	$2 \cdot 55$	$87 \cdot 52$	$12 \cdot 63$	$73 \cdot 05$
$n = 1000$	$1 \cdot 31$	$89 \cdot 95$	$1 \cdot 82$	$86 \cdot 61$	$9 \cdot 68$	$68 \cdot 31$

Sample size	$L^1_{GDR(6)}$	$L^1_{GDR(6)}/L^1_{QR}$	$L^2_{GDR(6)}$	$L^2_{GDR(6)}/L^2_{QR}$	$L^\infty_{GDR(6)}$	$L^\infty_{GDR(6)}/L^\infty_{QR}$
$n = 100$	$4 \cdot 12$	$93 \cdot 59$	$5 \cdot 66$	$93 \cdot 07$	$22 \cdot 30$	$91 \cdot 56$
$n = 235$	$2 \cdot 71$	$92 \cdot 03$	$3 \cdot 76$	$90 \cdot 43$	$17 \cdot 75$	$85 \cdot 11$
$n = 500$	$1 \cdot 86$	$91 \cdot 10$	$2 \cdot 59$	$88 \cdot 78$	$13 \cdot 69$	$79 \cdot 19$
$n = 1000$	$1 \cdot 32$	$90 \cdot 64$	$1 \cdot 85$	$87 \cdot 90$	$10 \cdot 66$	$75 \cdot 23$

Sample size	$L^1_{GDR(8)}$	$L^1_{GDR(8)}/L^1_{QR}$	$L^2_{GDR(8)}$	$L^2_{GDR(8)}/L^2_{QR}$	$L^\infty_{GDR(8)}$	$L^\infty_{GDR(8)}/L^\infty_{QR}$
$n = 100$	$4 \cdot 13$	$93 \cdot 81$	$5 \cdot 68$	$93 \cdot 38$	$22 \cdot 44$	$92 \cdot 14$
$n = 235$	$2 \cdot 71$	$92 \cdot 18$	$3 \cdot 77$	$90 \cdot 71$	$17 \cdot 92$	$85. \cdot 96$
$n = 500$	$1 \cdot 86$	$91 \cdot 26$	$2 \cdot 60$	$89 \cdot 08$	$13 \cdot 91$	$80 \cdot 46$
$n = 1000$	$1 \cdot 32$	$90 \cdot 81$	$1 \cdot 86$	$88 \cdot 22$	$10 \cdot 92$	$77 \cdot 01$

average estimation errors of conditional distribution function estimates across simulations for $J = 4, 6$ and 8 terms generalized dual regression and quantile regression-based estimators, respectively, and their ratio in percentage terms. The performance of dual regression estimates in the simulations is robust to incorrect choice of J, with only a small loss in accuracy caused by misspecification. For the 8 terms representation, the gains over quantile regression-based estimates remain significant, ranging from 6% to 23% depending on the norm and sample size.

Table 3 summarizes the results corresponding to the accuracy of functional intercept and covariate coefficients estimates across simulations. Estimates are based on the selected model in each simulation. For each coefficient, we compute the root mean absolute error of estimates, by computing errors for quantile indices in $\{0 \cdot 5, 0 \cdot 9, 0 \cdot 99\}$ for each replication, and then computing the summary statistic. We also report average root mean absolute error over the grid $\{0 \cdot 01, 0 \cdot 02, \ldots, 0 \cdot 99\}$ of quantile indices. In all cases selected generalized dual regression estimates have lower root mean absolute error, which corroborates results shown in Table 1 in the main text for the conditional distribution function.

6.3. **Additional Simulations.** We provide additional simulations comparing dual regression to the noncrossing quantile regression method introduced by Bondell et al. (2010), replicating the experiments they propose. In their simulation study they consider three

TABLE 3. Summary results for intercept and X coefficients across sample sizes: square root of mean absolute error across simulations (RMAE) for $\{0\cdot5, 0\cdot9, 0\cdot99\}$ quantile indices and Average (Ave.) RMAE over $\{0\cdot01, 0\cdot02, \ldots, 0\cdot99\}$ quantile indices.

		Intercept $\beta_0(u)$			
Sample size	Method	$\tau = 0\cdot5$	$\tau = 0\cdot9$	$\tau = 0\cdot99$	Ave.
$n = 100$	GDR	$8\cdot19$	$9\cdot32$	$10\cdot94$	$9\cdot49$
	QR	$8\cdot22$	$9\cdot38$	$11\cdot46$	$9\cdot69$
$n = 235$	GDR	$6\cdot59$	$7\cdot46$	$9\cdot01$	$7\cdot69$
	QR	$6\cdot63$	$7\cdot50$	$9\cdot28$	$7\cdot81$
$n = 500$	GDR	$5\cdot47$	$6\cdot25$	$7\cdot45$	$6\cdot39$
	QR	$5\cdot50$	$6\cdot28$	$7\cdot71$	$6\cdot50$
$n = 1000$	GDR	$4\cdot58$	$5\cdot23$	$6\cdot33$	$5\cdot38$
	QR	$4\cdot60$	$5\cdot26$	$6\cdot50$	$5\cdot45$

		X coefficient $\beta_1(u)$			
Sample size	Method	$\tau = 0\cdot5$	$\tau = 0\cdot9$	$\tau = 0\cdot99$	Ave.
$n = 100$	GDR	$0\cdot13$	$0\cdot15$	$0\cdot20$	$0\cdot16$
	QR	$0\cdot14$	$0\cdot17$	$0\cdot27$	$0\cdot19$
$n = 235$	GDR	$0\cdot11$	$0\cdot12$	$0\cdot16$	$0\cdot13$
	QR	$0\cdot11$	$0\cdot13$	$0\cdot20$	$0\cdot15$
$n = 500$	GDR	$0\cdot09$	$0\cdot10$	$0\cdot14$	$0\cdot11$
	QR	$0\cdot09$	$0\cdot11$	$0\cdot17$	$0\cdot12$
$n = 1000$	GDR	$0\cdot07$	$0\cdot09$	$0\cdot12$	$0\cdot09$
	QR	$0\cdot08$	$0\cdot09$	$0\cdot14$	$0\cdot10$

examples which are special cases of the linear heteroscedastic model

$$y_i = \alpha_1 + \beta_1^{\mathrm{T}}\widetilde{x}_i + (\alpha_2 + \beta_2^{\mathrm{T}}\widetilde{x}_i)\varepsilon_i,$$

where each component of $\widetilde{x}_i$ satisfies $\widetilde{x}_{ik} \sim U(0,1)$, $\varepsilon_i \sim N(0,1)$, and with $\alpha_1 = \alpha_2 = 1$. Their method imposes noncrossing constraints on the quantile regressions estimated, and they show that it outperforms both linear quantile regression and the method of He (1997) in their proposed experiments. The three examples are:

Example 1. $\dim(\widetilde{x}_i) = 4$, $\beta_1 = (1,1,1,1)^{\mathrm{T}}$, and $\beta_2 = (0\cdot1, 0\cdot1, 0\cdot1, 0\cdot1)^{\mathrm{T}}$.

Example 2. $\dim(\widetilde{x}_i) = 10$, $\beta_1 = (1,1,1,1,0^{\mathrm{T}})^{\mathrm{T}}$, and $\beta_2 = (0\cdot1, 0\cdot1, 0\cdot1, 0\cdot1, 0^{\mathrm{T}})^{\mathrm{T}}$.

Example 3. $\dim(\widetilde{x}_i) = 7$, $\beta_1 = (1,1,1,1,1,1,1)^{\mathrm{T}}$, and $\beta_2 = (1,1,1,0,0,0,0)^{\mathrm{T}}$.

For each example, 500 datasets of size 100, 200 and 500 are simulated. For the method of Bondell et al. (2010), six quantile curves are fitted to the data for each example, $u = \{0\cdot1, 0\cdot3, 0\cdot5, 0\cdot7, 0\cdot9, 0\cdot99\}$. We also implemented the noncrossing quantile regression method by fitting eleven quantile curves for the larger sequence $u = \{0\cdot01, 0\cdot1, 0\cdot2, \ldots, 0\cdot9, 0\cdot99\}$, the results are similar and are thus omitted.

TABLE 4. Replication of Bondell et al. (2010) experiment 1: average root mean integrated squared error (×100) over 500 simulations, with standard error in parentheses. NCRQ: noncrossing quantile regression.

	Example 1		
	$\tau = 0 \cdot 5$	$\tau = 0 \cdot 9$	$\tau = 0 \cdot 99$
		$n = 100$	
GDR	$26 \cdot 09\,(0 \cdot 39)$	$37 \cdot 04\,(0 \cdot 54)$	$57 \cdot 18\,(0 \cdot 85)$
NCRQ	$29 \cdot 91\,(0 \cdot 45)$	$41 \cdot 97\,(0 \cdot 57)$	$72 \cdot 24\,(0 \cdot 88)$
GDR (2): Ratio×100	$87 \cdot 21$	$88 \cdot 25$	$79 \cdot 16$
GDR (4): Ratio×100	$92 \cdot 56$	$92 \cdot 15$	$80 \cdot 39$
GDR (6): Ratio×100	$93 \cdot 24$	$92 \cdot 98$	$81 \cdot 03$
GDR (8): Ratio×100	$94 \cdot 65$	$93 \cdot 48$	$81 \cdot 29$
		$n = 200$	
GDR	$18 \cdot 80\,(0 \cdot 27)$	$25 \cdot 72\,(0 \cdot 39)$	$42 \cdot 16\,(0 \cdot 64)$
NCRQ	$22 \cdot 18\,(0 \cdot 32)$	$30 \cdot 03\,(0 \cdot 46)$	$57 \cdot 04\,(0 \cdot 72)$
GDR (2): Ratio×100	$84 \cdot 76$	$85 \cdot 66$	$73 \cdot 92$
GDR (4): Ratio×100	$91 \cdot 16$	$88 \cdot 31$	$74 \cdot 94$
GDR (6): Ratio×100	$92 \cdot 72$	$89 \cdot 46$	$74 \cdot 96$
GDR (8): Ratio×100	$93 \cdot 27$	$89 \cdot 42$	$75 \cdot 11$
		$n = 500$	
GDR	$12 \cdot 16\,(0 \cdot 17)$	$16 \cdot 36\,(0 \cdot 24)$	$27 \cdot 21\,(0 \cdot 43)$
NCRQ	$14 \cdot 32\,(0 \cdot 20)$	$19 \cdot 50\,(0 \cdot 29)$	$40 \cdot 08\,(0 \cdot 57)$
GDR (2): Ratio×100	$84 \cdot 89$	$83 \cdot 89$	$67 \cdot 89$
GDR (4): Ratio×100	$90 \cdot 29$	$86 \cdot 84$	$69 \cdot 03$
GDR (6): Ratio×100	$91 \cdot 60$	$87 \cdot 45$	$69 \cdot 35$
GDR (8): Ratio×100	$92 \cdot 15$	$87 \cdot 96$	$69 \cdot 38$

Tables 4–6 show the average root mean integrated squared errors over the 500 datasets along with their estimated standard errors, for each sample size, and for each of $u = \{0 \cdot 5, 0 \cdot 9, 0 \cdot 99\}$. For each simulation, the empirical root mean integrated squared error is calculated as RMISE $= [n^{-1} \sum_{i=1}^{n} \{\hat{\beta}(u)^{\mathrm{T}} x_i - \beta(u)^{\mathrm{T}} x_i\}^2]^{1/2}$, where $\hat{\beta}(u)$ and $\beta(u)$ are the estimated and true vector of quantile regression coefficients, respectively. The results for the other quantiles are similar, and are thus omitted.

In all three examples the location-scale structure, $J = 2$, is selected by the Schwartz criterion for each simulation and our proposed estimator significantly outperforms the noncrossing quantiles method for all quantiles and all sample sizes, except for $n = 100$ and $\tau = 0 \cdot 9$ in Example 2. The good relative performance of dual regression results from the selected location-scale structure, which adds further smoothness and stability across quantile curves, beyond the noncrossing constraints imposed by noncrossing quantile regression. This improvement is greater in the tails, as dual regression solutions are estimated globally whereas the local nature of quantile regression affects estimation of extreme quantiles.

TABLE 5. Replication of Bondell et al. (2010) experiment 2: average root mean integrated squared error (×100) over 500 simulations, with standard error in parentheses. NCRQ: noncrossing quantile regression.

	Example 2		
	$\tau = 0 \cdot 5$	$\tau = 0 \cdot 9$	$\tau = 0 \cdot 99$
	$n = 100$		
GDR	$40 \cdot 24\,(0 \cdot 40)$	$56 \cdot 37\,(0 \cdot 57)$	$87 \cdot 62\,(0 \cdot 80)$
NCRQ	$42 \cdot 55\,(0 \cdot 43)$	$53 \cdot 18\,(0 \cdot 49)$	$90 \cdot 30\,(0 \cdot 84)$
GDR (2): Ratio×100	$94 \cdot 58$	$106 \cdot 00$	$97 \cdot 03$
GDR (4): Ratio×100	$105 \cdot 02$	$110 \cdot 83$	$100 \cdot 26$
GDR (6): Ratio×100	$110 \cdot 37$	$113 \cdot 69$	$101 \cdot 87$
GDR (8): Ratio×100	$124 \cdot 84$	$123 \cdot 50$	$105 \cdot 45$
	$n = 200$		
GDR	$28 \cdot 63\,(0 \cdot 28)$	$39 \cdot 03\,(0 \cdot 37)$	$60 \cdot 34\,(0 \cdot 61)$
NCRQ	$31 \cdot 48\,(0 \cdot 31)$	$39 \cdot 99\,(0 \cdot 38)$	$66 \cdot 98\,(0 \cdot 63)$
GDR (2): Ratio×100	$90 \cdot 93$	$97 \cdot 59$	$90 \cdot 10$
GDR (4): Ratio×100	$98 \cdot 25$	$102 \cdot 37$	$91 \cdot 20$
GDR (6): Ratio×100	$100 \cdot 57$	$103 \cdot 43$	$91 \cdot 56$
GDR (8): Ratio×100	$102 \cdot 00$	$104 \cdot 04$	$91 \cdot 73$
	$n = 500$		
GDR	$17 \cdot 78\,(0 \cdot 17)$	$24 \cdot 23\,(0 \cdot 23)$	$37 \cdot 02\,(0 \cdot 39)$
NCRQ	$20 \cdot 87\,(0 \cdot 20)$	$27 \cdot 86\,(0 \cdot 26)$	$47 \cdot 65\,(0 \cdot 43)$
GDR (2): Ratio×100	$85 \cdot 19$	$86 \cdot 98$	$77 \cdot 69$
GDR (4): Ratio×100	$91 \cdot 82$	$90 \cdot 59$	$79 \cdot 60$
GDR (6): Ratio×100	$93 \cdot 49$	$91 \cdot 74$	$79 \cdot 99$
GDR (8): Ratio×100	$94 \cdot 33$	$92 \cdot 28$	$80 \cdot 07$

We also report the relative performance of non-selected dual regression estimates. Apart from Examples 2 and 3 with $n = 100$, the results are similar for all J to the selected model $J = 2$. For $n = 100$, results for Example 2, and to a lesser extent Example 3, show that the relative performance of dual regression deteriorates, especially for $J = 8$. These results are driven by a few simulations where the solver was unable to find an optimal solution, 7 instances for Example 2 and 5 for Example 3. Since for Example 2 and $J = 8$ the number of parameters is $2 + 8 \times 10 = 82$ for 100 observations, this is not unexpected. Compared to the simulations calibrated to the Engel data example, the fact that representations with J greater 2 are never selected for Examples 1–3 suggest that the presence of multiple covariates provides useful information effectively accounted for by the proposed model selection procedure.

TABLE 6. Replication of Bondell et al. (2010) experiment 3: average root mean integrated squared error (×100) over 500 simulations, with standard error in parentheses. NCRQ: noncrossing quantile regression.

	Example 3		
	$\tau = 0 \cdot 5$	$\tau = 0 \cdot 9$	$\tau = 0 \cdot 99$
		$n = 100$	
GDR	$69 \cdot 04\,(0 \cdot 84)$	$95 \cdot 85\,(1 \cdot 14)$	$152 \cdot 77\,(1 \cdot 75)$
NCRQ	$75 \cdot 09\,(0 \cdot 85)$	$97 \cdot 98\,(1 \cdot 20)$	$178 \cdot 10\,(2 \cdot 01)$
GDR (2): Ratio×100	$91 \cdot 94$	$97 \cdot 82$	$85 \cdot 78$
GDR (4): Ratio×100	$99 \cdot 80$	$102 \cdot 94$	$87 \cdot 23$
GDR (6): Ratio×100	$102 \cdot 96$	$104 \cdot 99$	$88 \cdot 09$
GDR (8): Ratio×100	$105 \cdot 49$	$106 \cdot 96$	$88 \cdot 29$
		$n = 200$	
GDR	$49 \cdot 22\,(0 \cdot 56)$	$67 \cdot 26\,(0 \cdot 77)$	$105 \cdot 22\,(1 \cdot 30)$
NCRQ	$55 \cdot 12\,(0 \cdot 62)$	$72 \cdot 82\,(0 \cdot 83)$	$135 \cdot 24\,(1 \cdot 59)$
GDR (2): Ratio×100	$89 \cdot 29$	$92 \cdot 37$	$77 \cdot 80$
GDR (4): Ratio×100	$95 \cdot 78$	$95 \cdot 74$	$79 \cdot 65$
GDR (6): Ratio×100	$97 \cdot 41$	$97 \cdot 06$	$80 \cdot 22$
GDR (8): Ratio×100	$98 \cdot 51$	$97 \cdot 08$	$80 \cdot 24$
		$n = 500$	
GDR	$30 \cdot 84\,(0 \cdot 35)$	$42 \cdot 17\,(0 \cdot 51)$	$66 \cdot 86\,(0 \cdot 89)$
NCRQ	$35 \cdot 77\,(0 \cdot 42)$	$48 \cdot 80\,(0 \cdot 57)$	$94 \cdot 12\,(1 \cdot 19)$
GDR (2): Ratio×100	$86 \cdot 22$	$86 \cdot 40$	$71 \cdot 04$
GDR (4): Ratio×100	$92 \cdot 60$	$90 \cdot 27$	$72 \cdot 66$
GDR (6): Ratio×100	$94 \cdot 38$	$91 \cdot 20$	$72 \cdot 73$
GDR (8): Ratio×100	$94 \cdot 85$	$91 \cdot 27$	$72 \cdot 82$

References

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

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

Chernozhukov, V., Fernandez-Val, I. & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica* **81**, 2205–2268.

Donald, S.G., Imbens, G.W. and Newey W.K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*, **117**, 55–93.

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

Koenker, R. & Xiao, Z. (2002). Inference on the quantile regression process. *Econometrica* **70**, 1583–1612.

Newey, W. & Mc Fadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4, ch. 36, 1st ed., pp. 2111–2245. Amsterdam: Elsevier.

Tripathi, G. (2006). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters* **63**, 1–3.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A.W. and Wellner, J. (2007). *Empirical processes indexed by estimated functions*. Lecture Notes-Monograph Series, 234–252.