Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Neural Networks for Partially Linear Quantile Regression

Qixian Zhong & Jane-Ling Wang

To cite this article: Qixian Zhong & Jane-Ling Wang (02 Jun 2023): Neural Networks for Partially Linear Quantile Regression, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2023.2208183

To link to this article: https://doi.org/10.1080/07350015.2023.2208183

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2023, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/07350015.2023.2208183

Neural Networks for Partially Linear Quantile Regression

Qixian Zhong[a] and Jane-Ling Wang[b]

[a]Department of Statistics and Data Science, School of Economics, Wang Yanan Institute for Studies in Economics (WISE), and MOE Key Laboratory of Econometrics, Xiamen University, Xiamen, China; [b]Department of Statistics, University of California, Davis, Davis, CA

ABSTRACT

Deep learning has enjoyed tremendous success in a variety of applications but its application to quantile regression remains scarce. A major advantage of the deep learning approach is its flexibility to model complex data in a more parsimonious way than nonparametric smoothing methods. However, while deep learning brought breakthroughs in prediction, it is not well suited for statistical inference due to its black box nature. In this article, we leverage the advantages of deep learning and apply it to quantile regression where the goal is to produce interpretable results and perform statistical inference. We achieve this by adopting a semiparametric approach based on the partially linear quantile regression model, where covariates of primary interest for statistical inference are modeled linearly and all other covariates are modeled nonparametrically by means of a deep neural network. In addition to the new methodology, we provide theoretical justification for the proposed model by establishing the root-n consistency and asymptotically normality of the parametric coefficient estimator and the minimax optimal convergence rate of the neural nonparametric function estimator. Across several simulated and real data examples, the proposed model empirically produces superior estimates and more accurate predictions than various alternative approaches.

ARTICLE HISTORY
Received February 2022
Accepted April 2023

KEYWORDS
Curse of dimensionality;
Deep learning;
Interpretability;
Semiparametric regression;
Stochastic gradient descent

1. Introduction

With advances in computational power and the availability of large data, deep learning has emerged as a powerful data analysis tool in a wide variety of applications, such as computer vision (Krizhevsky, Sutskever, and Hinton 2012), economics (Heaton, Polson, and Witte 2017) and business (Nolle, Seeliger, and Mühlhäuser 2018). Deep learning estimates a function from data using neural networks which compose multiple (parameterized) nonlinear transformations. These inferred transformations are jointly optimized *end-to-end* in order to produce the optimal overall map rather than independently estimating each transformation in a separate stage.

Roughly speaking, a neural network, which consists of several layers and neurons between the input and output layers, is a composite function [see formula (1)] with a recursive concatenation of an affine linear function and a simple nonlinear map. The success of neural networks is attributed to their powerful capacity to represent unknown functions. For example, Cybenko (1989) and Hornik, Stinchcombe, and White (1989) showed that any continuous functions can be approximated by shallow neural networks to any degree of accuracy. Telgarsky (2016) and Yarotsky (2017) further showed that deep neural networks enjoy a better representational power than their shallow counterparts.

Despite their superior empirical performance, deep learning models, mostly a black box, often lack intepretability and theoretical support. Different approaches have emerged in recent works to examine various aspects of interpretable deep learning models. For instance, the saliency-based (Zeiler and Fergus 2014) method aims at providing post hoc explanations for a certain type of neural networks. Chernozhukov et al. (2017) and Mi et al. (2021) employed neural networks in semiparametric mean regression models to study the causal effect between variables. For additional work on intepretable deep learning models, we refer readers to the recent review paper by Rudin (2019) and references therein.

Unlike the above approaches, this article adopts the statistical model-based approach for interpretability by constructing neural networks for a partially linear quantile regression (PLQR) problem. Specifically, we model the covariates of interest with a linear predictor for interpretability and statistical inference and model the nonparametric component with neural networks. The proposed deep learning method for PLQR is abbreviated as DPLQR. As a semiparametric approach, DPLQR not only offers interpretibility for the parametric component but also allows model flexibility for the nonparametric component. Importantly, it avoids the curse of dimensionality of nonparametric smoothing methods through the strength of neural networks to detect the structure, often low-dimensional, of the data. We further provide mathematical support for the DPLQR, which not only quantifies the uncertainty of the inference but also reveals why deep learning works.

Since the seminal work of Koenker and Bassett (1978), quantile regression has been extensively investigated, including linear

CONTACT Jane-Ling Wang *janelwang@ucdavis.edu* Department of Statistics, University of California, Davis, Davis, CA.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

quantile regression (Koenker and Bassett 1978; Portnoy 1991), nonparametric quantile regression (Samanta 1989; Jones and Hall 1990; Chaudhuri 1991; Li and Racine 2008; Guerre and Sabbah 2012; Kong, Linton, and Xia 2013; Fang, Li, and Yan 2021) and semiparametric quantile regression (He and Shi 1996; Lee 2003; Wu, Yu, and Yu 2010; Cai and Xiao 2012; Kong and Xia 2012; Noh, Ghouch, and Van Keilegom 2015; Fan and Liu 2016; Ma and He 2016; Bhattacharya, Gimenes, and Guerre 2021). For a comprehensive introduction of quantile regression, we refer to the monographs by Koenker (2005) and Koenker et al. (2017). Compared to the least squares regression approach that focuses on the conditional mean of the response, quantile regression offers a more expansive view of the effect of covariates on the response. Moreover, quantile regression is more robust against outliers when the distribution of the response is heavy-tailed or skewed (Koenker 2005).

While linear and nonparametric quantile regression have been well developed, theory and methodology for partially linear quantile regression models are lagging and existing work mainly focuses on the partially linear additive quantile regression (Lian 2012; Hoshino 2014; Sherwood and Wang 2016) and partially linear single-index quantile regression (Wu and Yu 2014; Zhang, Lian, and Yu 2017, 2020). This approach incorporates linear regression effects for some covariates and an additive or single-index model with smooth but unknown regression functions for the remaining covariates. Such approaches alleviate the curse of dimensionality but they are not amenable to interactions or non-single-index structure among covariates. Meanwhile, existing fully nonparametric approaches suffer from a severe curse of dimensionality, so they are only effective for very low-dimensional covariates. In contrast, the proposed DPLQR not only retains the linear component of the model to interpret the effects of primary covariates, such as the effect of a treatment, but also enjoys the flexibility of a fully nonparametric model that is more resilient to the curse of dimensionality.

Applications of deep learning to quantile regression have emerged in recent years, such as in climate prediction (Hatalis et al. 2017) and electricity and power systems (Gan et al. 2018). However, theoretical understanding of quantile regression with neural networks remains scarce and limited to nonparametric quantile regression. Romano, Patterson, and Candes (2019) employed conformal methods to construct prediction intervals for the response but did not address estimation of the conditional quantile function. Jantre, Bhattacharya, and Maiti (2020) developed consistency results for nonparametric quantile function estimators with a single-hidden-layer neural network. However, the implementation of their procedure requires exponential time for the computation, compared to polynomial time for deep neural networks (Rolnick and Tegmark 2017). As we were wrapping up the first version of our research findings, we became aware of a related work that was independently developed by Padilla, Tansey, and Chen (2020). Although this work also explored the convergence rate of the conditional quantile function estimator, it is substantially different from ours. First, they focused on a black-box nonparametric approach to estimate the quantile function, while we are interested in both estimation and interpretability, as well as statistical inference for the model. For instance, it is of interest to study whether maternal education has an effect on birth weight of infants, since low birth weight is associated with subsequent health problems (Badshah et al. 2008). The vanilla version of the deep learning approach in Padilla, Tansey, and Chen (2020) is not geared toward providing statistical inference for the effect of maternal education. In contrast, the proposed model not only addresses the inference issue but also achieves comparable prediction errors for Natality Birth Data as shown in Section 6. Second, the theoretical analysis of their work only holds for continuous covariates while our theory covers both continuous and discrete covariates, and we establish asymptotic normality for the estimates of the linear component. Recently, another related approach was considered by Shen et al. (2021). Clearly, there is rising interest in employing deep learning to quantile regression.

To summarize, the major contributions of our article are 4-fold.

1. We introduce DPLQR to shed new light on an interpretable deep learning model which overcomes the drawback of a black-box deep learning approach. Previous attempts fail to provide uncertainty quantification. In contrast, we develop confidence intervals for the effects of linear covariates, which are of interest to practitioners. Our approach can thus be viewed as a bridge between machine learning and statistical inference.
2. We provide theoretical justification for the deep learning approach by establishing minimax optimal convergence rates (up to a poly-logarithmic factor) of the nonlinear component of the DPLQR. We further establish root-n convergence and asymptotic normality of the regression coefficient estimator for both homoscedastic and heteroscedastic random errors. This substantially distinguishes our theoretical contributions from previous purely nonparametric approaches (Bauer and Kohler 2019; Padilla, Tansey, and Chen 2020; Schmidt-Hieber 2020; Shen et al. 2021). The asymptotic normality involves the derivation of the influence function, which is a nontrivial task that includes the empirical processes of the subgradient of the check loss function and controlling the order of the remainder term.
3. The proposed DPLQR model is flexible and includes a large number of previously-studied quantile regression models. Specifically, DPLQR reduces to linear quantile regression when the nonparametric component is absent and to nonparametric quantile regression in the absence of linear predictors. The DPLQR model also includes the partially linear additive or single-index quantile regression models as special cases.
4. Our methodology is able to identify the underlying intrinsic dimension of the data, which circumvents the curse of dimensionality that greatly limits the applicability of nonparametric smoothing approach. For example, when the true model corresponds to a partially linear additive quantile regression, the resulting neural network estimators automatically detect this and enjoy a one-dimensional nonparametric convergence rate (up to a poly-logarithmic factor).

The rest of the article proceeds as follows. In Section 2, we briefly introduce the fundamental concept of neural networks and quantile regression. Asymptotic properties of the estimators are presented in Section 3. The implementation of the proposed approach is discussed in Section 4, along with the calculation of

the asymptotic covariance matrix for the vector parameter. Sections 5 and 6 provide simulation studies and a data application comparing the proposed method with linear quantile regression and partially linear additive quantile regression. Section 7 discusses some potential extensions. The online supplementary material provides mathematical proofs and additional numerical results.

2. Preliminaries

2.1. Neural Networks

We first briefly present the relevant background on deep neural networks. For some integer $L \geq 2$, let $\boldsymbol{q} = (q_0, q_1, \ldots, q_L)^\top \in \mathbb{N}^{L+1}$. An L-layer neural network with input dimension q_0 and output dimension q_L is a function $m : \mathbb{R}^{q_0} \to \mathbb{R}^{q_L}$ that satisfies the following recursive relation:

$$\begin{aligned} m(z) &= \widetilde{W}_L m_{L-1}(z) + \tilde{b}_L, \\ m_{L-1}(z) &= \xi(\widetilde{W}_{L-1} m_{L-2}(z) + \tilde{b}_{L-1}), \ldots, \\ m_1(z) &= \xi(\widetilde{W}_1 z + \tilde{b}_1), \end{aligned} \tag{1}$$

where $\widetilde{W}_k$ and $\tilde{b}_k$ are a $q_k \times q_{k-1}$ matrix and q_k-dimensional column vector, respectively, and ξ is a prior deterministic function which operates component-wise on vectors, that is, $\xi(\boldsymbol{v}) = (\xi(v_1), \ldots, \xi(v_k))^\top$, for a vector $\boldsymbol{v} = (v_1, \ldots, v_k)^\top$. We call L the *depth* of the neural network; m_k, for $1 \leq k \leq L-1$, the kth *hidden layer*; and $\xi : \mathbb{R} \to \mathbb{R}$ the *activation function*. A two-layer ($L = 2$) neural network is often called a *shallow neural network*. At the kth hidden layer, there are q_k neurons, or nodes, and q_k is called the *width* of the kth layer. The activation function ξ links adjacent layers and is often set to be a simple nonlinear function. In this article, we consider the *rectified linear unit* (ReLU) activation function $\xi(z) = \max(z, 0)$ since it is computationally efficient and often achieves best performance in practice (Krizhevsky, Sutskever, and Hinton 2012). The matrices $\widetilde{W}_k$ and vectors $\tilde{b}_k$ are often referred to as the "weight" and "bias", respectively in the machine learning literature, but we avoid using these terms here to prevent confusion. We write $W_k = (\widetilde{W}_k, \tilde{b}_k) \in \mathbb{R}^{q_k \times (q_{k-1}+1)}$. Then the neural network in (1) can be succinctly expressed as

$$m(z) = W_L \tilde{\xi} \circ \cdots \circ W_2 \tilde{\xi}(W_1 \tilde{z}), \tag{2}$$

where $\tilde{\xi}(v) = (\xi(v)^\top, 1)^\top$ and $\tilde{z} = (z^\top, 1)^\top$. Figure 1 illustrates a three layers neural network with $\boldsymbol{q} = (4, 5, 5, 1)^\top$. For an overview of the structure of neural networks, see the recent papers by Yuan et al. (2020), Fan, Ma, and Zhong (2021) and the monograph by Goodfellow, Bengio, and Courville (2016).

Note that the total number of parameters in (2) is $\sum_{k=1}^{L} q_k(q_{k-1} + 1)$, which can be very large and may lead to overfitting. Han et al. (2015), Bauer and Kohler (2019), and Schmidt-Hieber (2020) mitigated against this by deactivating some of the links of neurons between the adjacent hidden layers. Following this strategy, for $s \in \mathbb{N}$, $L \geq 2$, $A > 0$ and $\boldsymbol{q} = (q_0, q_1, \ldots, q_L)^\top \in \mathbb{N}^{L+1}$, we consider a sparsely connected

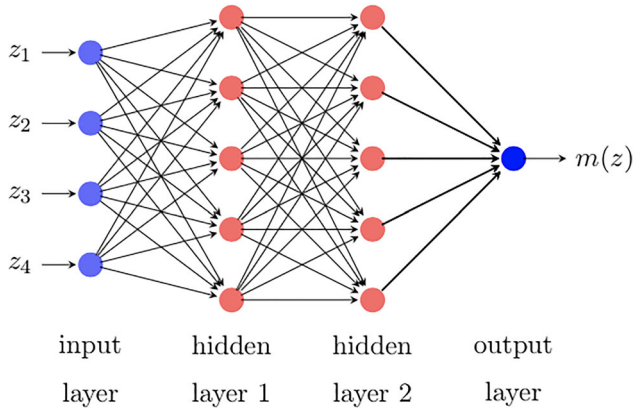

Figure 1. A three-layer neural network with four input variables and one output.

neural network class

$$\begin{aligned} &\mathcal{M}(s, L, \boldsymbol{q}, A) \\ &= \Big\{ m(z) = W_L \tilde{\xi} \circ \cdots \circ W_2 \tilde{\xi}(W_1 \tilde{z}) \mid W_k \in \mathbb{R}^{q_k \times (q_{k-1}+1)}, \\ &\qquad \|W_k\|_\infty \leq 1 \text{ for } k = 1, \ldots, L, \ \sum_{k=1}^{L} \|W_k\|_0 \leq s \text{ and} \\ &\qquad \|m\|_\infty \leq A \Big\}, \end{aligned} \tag{3}$$

where $\|\cdot\|_\infty$ is the sup-norm of a matrix or function and $\|\cdot\|_0$ is the number of nonzero elements of a matrix.

2.2. Partially Linear Quantile Regression Model and Estimation

Consider a univariate continuous random variable Y and a multivariate random variable $U = (X, Z) \in \mathbb{R}^p \times \mathbb{R}^q$, where X may include treatment (indicator) variables and continuous covariates of interest. Let $F_{Y|U}(\cdot|u)$ be the conditional distribution function of Y given $U = u$. For some $0 < \tau < 1$, the τth conditional quantile of Y given $U = u$ is defined as

$$\hbar_\tau(u) = \inf_{y \in \mathbb{R}} \{y \mid F_{Y|U}(y|u) \geq \tau\}.$$

In this article, we assume $\hbar_\tau(X, Z) = X^\top \theta_\tau + m_\tau(Z)$, which leads to the following partially linear quantile regression model:

$$Y = X^\top \theta_\tau + m_\tau(Z) + \epsilon, P(\epsilon \leq 0 | U) = \tau, \tag{4}$$

where $\theta_\tau \in \mathbb{R}^p$ is an unspecified parameter without an intercept term, $m_\tau : \mathbb{R}^q \to \mathbb{R}$ is an unknown function and the error ϵ may be heteroscedastic by allowing it to vary with $u = (x, z)$.

Let $\{(X_i, Z_i, Y_i) : i = 1, \ldots, n\}$ denote independent and identically distributed realizations of (X, Z, Y). For simplicity, we use the notation $\mathcal{M}_A$ to denote the neural network class $\mathcal{M}(s, L, \boldsymbol{q}, A)$ in (3) with $q_0 = q$, $q_L = 1$ and some large enough A. To estimate the vector θ_τ and the function m_τ, we minimize the loss function:

$$(\hat{\theta}_\tau, \hat{m}_\tau) = \underset{(\theta, m) \in \mathbb{R}_A^p \times \mathcal{M}_A}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top \theta - m(Z_i)), \tag{5}$$

where $\rho_\tau(t) = t\{\tau - I(t < 0)\}$ is called the check function with $I(\cdot)$ being an indicator function and $\mathbb{R}_A^p = \{\theta \in \mathbb{R}^p : \|\theta\|_\infty <$

$A\}$. This loss function becomes the absolute value L^1-loss when $\tau = 0.5$ which leads to the median estimators. For brevity, we suppress the subscript τ and write

$$(\theta_0, m_0) = (\theta_\tau, m_\tau) \text{ and } (\hat{\theta}, \hat{m}) = (\hat{\theta}_\tau, \hat{m}_\tau). \tag{6}$$

3. Theory

In this section, we establish the theoretical properties of the estimators $\hat{\theta}$ and $\hat{m}$. We first introduce a class of smooth functions in which m_0 resides.

Let γ and B be two positive constants and $\lfloor\gamma\rfloor$ denote the largest integer strictly less than γ. We call a function $h : \mathbb{T} \subset \mathbb{R}^q \to \mathbb{R}$ a (γ, B)-Hölder smooth function if it satisfies

$$\sup_{z\in\mathbb{T}} \left| \frac{\partial^{|\boldsymbol{\alpha}|} h}{\partial^{\alpha_1} z_1 \dots \partial^{\alpha_q} z_q}(z) \right| \le B, \quad \text{for all } \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)^\top \in \mathbb{N}^q$$

$$\text{and } |\boldsymbol{\alpha}| = \sum_{i=1}^{q} \alpha_i \le \lfloor\gamma\rfloor,$$

and

$$\sup_{z,z^*\in\mathbb{T}} \left| \frac{\partial^{|\boldsymbol{\alpha}|} h}{\partial^{\alpha_1} z_1 \dots \partial^{\alpha_q} z_q}(z) - \frac{\partial^{|\boldsymbol{\alpha}|} h}{\partial^{\alpha_1} z_1 \dots \partial^{\alpha_q} z_q}(z^*) \right| \le B\|z - z^*\|_2^{\gamma - \lfloor\gamma\rfloor}, \quad \text{for all } |\boldsymbol{\alpha}| = \lfloor\gamma\rfloor.$$

Denote the class of all such (γ, B)-Hölder smooth functions as $\mathcal{H}_q^\gamma(\mathbb{T}, B)$. Let $J \in \mathbb{N}$, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_J)^\top \in \mathbb{R}_+^J$, $\boldsymbol{d} = (q, d_1, \dots, d_J)^\top \in \mathbb{N}^{J+1}$ and $\bar{\boldsymbol{d}} = (\bar{d}_1, \dots, \bar{d}_J)^\top \in \mathbb{N}^J$ with $\bar{d}_1 \le q$ and $\bar{d}_k \le d_{k-1}, k = 2, \dots, J$. We further define a composite function class:

$$\begin{aligned}\mathcal{H}(J, \boldsymbol{\gamma}, \boldsymbol{d}, \bar{\boldsymbol{d}}, B) = \Big\{ & h = h_J \circ \dots \circ h_1 : \mathbb{T} \to \mathbb{R} \mid h_k = (h_{k1}, \dots, h_{kd_k})^\top \text{ and} \\ & h_{kj} \in \mathcal{H}_{\bar{d}_k}^{\gamma_k}([a_k, b_k]^{\bar{d}_k}, B) \quad \text{for some } |a_k|, |b_k| \le B \Big\}.\end{aligned} \tag{7}$$

This class of functions, first proposed by Schmidt-Hieber (2020), contains two kinds of dimension $\boldsymbol{d}$ and $\bar{\boldsymbol{d}}$. We call $\bar{\boldsymbol{d}}$ the *intrinsic dimension* of the function h in $\mathcal{H}(J, \boldsymbol{\gamma}, \boldsymbol{d}, \bar{\boldsymbol{d}}, B)$. Its *effective smoothness* is defined as $\bar{\boldsymbol{\gamma}} = (\bar{\gamma}_1, \dots, \bar{\gamma}_J)^\top$ with $\bar{\gamma}_k = \gamma_k \prod_{i=k+1}^{J} (\gamma_i \wedge 1), k = 1, \dots, J-1$ and $\bar{\gamma}_J = \gamma_J$. To establish the convergence rate with sample size n in Theorem 3.1 below, we denote

$$r_n = \max_{k\in\{1,\dots,J\}} n^{-\frac{\bar{\gamma}_k}{2\bar{\gamma}_k + \bar{d}_k}}. \tag{8}$$

As an illustration, consider the function

$$h(z) = h_{31}(h_{21}(h_{11}(z_1, z_2), h_{12}(z_3, z_4)), h_{22}(h_{13}(z_5, z_6), h_{14}(z_7))), \tag{9}$$

where all h_{ij} are $(\gamma, 1)$-Hölder smooth with $\gamma \ge 1$. It is clear that $h \in \mathcal{H}(J, \boldsymbol{\gamma}, \boldsymbol{d}, \bar{\boldsymbol{d}}, B)$ with $J = 3, \boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}} = (\gamma, \gamma, \gamma)^\top, \boldsymbol{d} = (7, 4, 2, 1)^\top, \bar{\boldsymbol{d}} = (2, 2, 2)^\top, B = 1$ and $r_n = n^{-\gamma/(2\gamma+2)}$.

With different choices of $J, \boldsymbol{\gamma}, \boldsymbol{d}$ and $\bar{\boldsymbol{d}}$, $\mathcal{H}(J, \boldsymbol{\gamma}, \boldsymbol{d}, \bar{\boldsymbol{d}}, B)$ includes a large number of function classes that have been considered in the statistical and economics literature. Below we provide two examples to illustrate the ubiquity of such function classes. We say a function h is (∞, B)-Hölder smooth if it is (γ, B)-Hölder smooth for all $\gamma > 0$.

Example 3.1 (Additive functions). A function $h : \mathbb{R}^q \to \mathbb{R}$ is additive if it can be represented a sum of univariate functions of each component (Stone 1985), that is, for $z = (z_1, \dots, z_q)^\top$,

$$h(z) = h_1(z_1) + \dots + h_q(z_q), \tag{10}$$

where $h_k, k = 1, \dots, q$ are univariate (γ, B)-Hölder smooth functions with $\gamma \ge 1$. Here $J = 2, \boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}} = (\gamma, \infty)^\top, \boldsymbol{d} = (q, q, 1)^\top, \bar{\boldsymbol{d}} = (1, q)^\top, r_n = n^{-1/(2\gamma+1)}, h_{1k}(z) = h_k(z_k), k = 1, \dots, q$, and $h_{21}(y) = y_1 + \dots + y_q$, where $y = (y_1, \dots, y_q)^\top$.

Example 3.2 (Single/multiple index functions). A single index function, first introduced by Ichimura (1993) and later extended to multiple indices by Hristache et al. (2001), is given by

$$h(z) = h_1(z^\top \alpha_1, \dots, z^\top \alpha_K), \tag{11}$$

where $\alpha_k, k = 1, \dots, K$ are unknown parameters and $z^\top \alpha_j$ are the index functions. It is easy to see that $h_{1k}(z) = z^\top \alpha_k, k = 1, \dots, K$ and $h_{21}(y) = h_1(y)$. Thus, if h_1 is (γ, B)-Hölder smooth with $\gamma \ge 1, \boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}} = (\infty, \gamma)^\top, \boldsymbol{d} = (q, K, 1)^\top, \bar{\boldsymbol{d}} = (\bar{d}_1, K)^\top$ with $\bar{d}_1 = \max_k\{\|\alpha_k\|_0\}$, and $r_n = n^{-1/(2\gamma+1)}$.

For the covariate $X = (X_1, \dots, X_p)^\top$, we define

$$\varphi_k^* = \arg\min_{\varphi\in L^2(P_Z)} \mathbb{E}[f(0|U)\{X_k - \varphi(Z)\}^2], k = 1, \dots, p, \tag{12}$$

where $L^2(P_Z) = \{\varphi \mid \mathbb{E}\varphi^2(Z) < \infty\}$ and $f(\cdot|U)$ is the conditional density of error ϵ in (4). Let $\boldsymbol{\varphi}^*(Z) = (\varphi_1^*(Z), \dots, \varphi_p^*(Z))^\top$, and

$$\begin{aligned}\Sigma_1 &= \mathbb{E}[\tau(1-\tau)\{X - \boldsymbol{\varphi}^*(Z)\}\{X - \boldsymbol{\varphi}^*(Z)\}^\top], \\ \Sigma_2 &= \mathbb{E}[f(0|U)\{X - \boldsymbol{\varphi}^*(Z)\}\{X - \boldsymbol{\varphi}^*(Z)\}^\top].\end{aligned} \tag{13}$$

It is easy to show that $\boldsymbol{\varphi}^* = \mathbb{E}(X|Z)$ if the conditional error density $f(\cdot|U)$ is independent of U at zero, see also Lian (2012) and Hoshino (2014) for partially linear additive regression.

Next, we state the assumptions for the deep partially linear quantile regression model.

Assumption 1. The true vector parameter θ_0 belongs to a compact subset $\Theta \subset \mathbb{R}^p$ bounded by B, the true nonparametric function m_0 in (6) belongs to $\mathcal{H} = \mathcal{H}(J, \boldsymbol{\gamma}, \boldsymbol{d}, \bar{\boldsymbol{d}}, B)$, and A in (5) satisfies $B < A$.

Assumption 2. The covariates (X, Z) take values in a compact subset of $\mathbb{R}^{p+q}$ that without loss of generality will be assumed to be $[0, 1]^{p+q}$. In addition, the probability density function (PDF) of Z is bounded away from zero and from infinity.

Assumption 3. The conditional PDF $f(\cdot|u)$ of the random error ϵ given the covariate $U = u$ has continuous derivative $f'(\cdot|u)$, and there exist positive constants b_0 and c_0 such that $1/c_0 < f(t|u) < c_0$ and $|f'(t|u)| < c_0$ for all $|t| \le b_0, u \in [0, 1]^{p+q}$.

Assumption 4. The depth L, width vector $\boldsymbol{q} = (q_0, q_1, \dots, q_L)^\top$ and number of nonzero parameters s of the neural network class (3) satisfy $L = O(\log n), s = O(nr_n^2 \log n)$ and $nr_n^2 \lesssim \min_{k=1,\dots,L}\{q_k\} \le \max_{k=1,\dots,L}\{q_k\} \lesssim n$, where r_n is defined in (8).

Assumption 5. The matrices Σ_1 and Σ_2 in (13) are both positive definite.

Assumption 6. $\max_{k=1,\ldots,p}(\mathbb{E}|X_k|^4) < \infty$ and $\bar{\gamma}_{\bar{k}} > \bar{d}_{\bar{k}}/2$, where $\bar{k}$ is the index to achieve the maximum value in (8).

The boundedness of both the parameters and covariate spaces in Assumptions 1 and 2 are standard for semiparametric/nonparametric regression. In Assumption 3 we postulate that the PDF of the error and its derivative are bounded to guarantee that the true parameter (θ_0, m_0) is a well-separated point of the minimum of the expected check loss function. In Assumption 4, we assume that the size of neural networks $\mathcal{M}$ used in (5) grows with the sample size n at a certain rate to balance the approximation and estimation errors of the estimators. Assumptions 5 and 6 are common conditions for asymptotic normality of the vector estimator $\hat{\theta}$ in semiparametric regression (He and Shi 1996; Wang, Zhu, and Zhou 2009; Sherwood and Wang 2016), where Assumption 5 is used to develop the asymptotic variance, while Assumption 6 guarantees $\sqrt{n}$-consistency.

We are now ready to state the convergence rate of the estimators.

Theorem 3.1. Under Assumptions 1–5, we have

$$\lim_{C\to\infty}\lim_{n\to\infty}\sup_{m_0\in\mathcal{H}}\mathbb{P}(\|\hat{m}-m_0\|_{L^2([0,1]^q)}\geq Cr_n\log^2 n)=0.$$

From the proof of Theorem 3.1 one can see that the convergence rate is the result of a tradeoff between estimation error and approximation error. Here the approximation error is defined as the distance between the true parameter m_0 and the neural network set $\mathcal{M}$, that is, $\min_{m\in\mathcal{M}}\|m-m_0\|_{L^2([0,1]^q)}$. It is known that a more complex neural network structure is more flexible and thus leads to a smaller approximation error (Anthony and Bartlett 1999; Yarotsky 2017; Bauer and Kohler 2019; Schmidt-Hieber 2020). However, too many parameters, for example the number of nonzero weights s in (3), will lead to high variance. Hence, there is an implicit "bias-variance" tradeoff that is reflected in the growth of neural networks.

Note that the convergence rate of the estimator $\hat{m}$ is determined by both the intrinsic dimension $\bar{\boldsymbol{d}} = (\bar{d}_1,\ldots,\bar{d}_J)^\top$ and the effective smoothness $\bar{\boldsymbol{\gamma}} = (\bar{\gamma}_1,\ldots,\bar{\gamma}_J)^\top$ of the true function $m_0 \in \mathcal{H}(J,\boldsymbol{\gamma},\boldsymbol{d},\bar{\boldsymbol{d}},B)$ in (7), rather than the dimension q of the covariate Z. For example, if m_0 has the composite structure in (9), the convergence rate for the proposed method is $n^{-\gamma/(2\gamma+2)}\log^2 n$. In contrast, the convergence rate for a nonparametric method, such as kernel or spline smoothing is of the order $n^{-\gamma/(2\gamma+7)}$ (Simonoff 2012). This shows that our method is able to detect the low dimensional structure of the data and thus circumvents the curse of dimensionality.

In particular, when m_0 reduces to an additive or a single index function, as shown in Examples 3.1 and 3.2, respectively, the resulting estimators have one-dimensional nonparametric rates of convergence (up to a poly-logarithmic factor). This is similar to results of Stone (1985) and Ichimura (1993) for nonparametric regression.

The next theorem establishes the minimax lower bound for estimating m_0, which implies that the resulting estimator $\hat{m}$ in Theorem 3.1 is rate-optimal.

Theorem 3.2. Let $\mathcal{F}$ be the class of probability density functions that satisfy Assumption 3. Then we have

$$\lim_{C\to\infty}\lim_{n\to\infty}\inf_{\hat{m}}\sup_{(\theta_0,m_0,f)\in\mathbb{R}^p\times\mathcal{H}\times\mathcal{F}}\mathbb{P}_{(\theta_0,m_0,f)}\left(\|\hat{m}-m_0\|_{L^2([0,1]^q)}\geq Cr_n\right)=1,$$

where the infimum is taken over all possible predictors $\hat{m}$ based on the observed data.

Below we show that the estimator $\hat{\theta}$ for the parameter vector is asymptotically normal at the $\sqrt{n}$ rate.

Theorem 3.3. Under Assumptions 1–6, we have

$$\sqrt{n}(\hat{\theta}-\theta_0)\to N(0,\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}),$$

where Σ_1 and Σ_2 are two p-by-p matrices defined in (13).

When $f(0|U)$ is a constant function, the solution of (12) would be $\boldsymbol{\varphi}^*(Z) = \mathbb{E}(X|Z)$, which leads to $\Sigma_1 = \tau(1-\tau)\text{var}\{X-\mathbb{E}(X|Z)\}$, $\Sigma_2 = f(0)\text{var}\{X-\mathbb{E}(X|Z)\}$ and more generally, the following corollary.

Corollary 3.1. Under the assumptions of Theorem 3.3 and when $f(0|U)$ is a constant function, we have

$$\sqrt{n}(\hat{\theta}-\theta_0)\to N(0,\Sigma),$$

where $\Sigma = \tau(1-\tau)[\text{var}\{X-\mathbb{E}(X|Z)\}]^{-1}/f^2(0)$.

For partially linear quantile regression with homoscedastic error, the random error ϵ is independent of the covariate U, which implies that $f(0|U=u) = f(0)$, for all $u\in[0,1]^{p+q}$, hence, Corollary 3.1 holds.

4. Implementation and Asymptotic Covariance

Estimation of $\hat{\theta}$ and $\hat{m}$: Since the check function in (5) is not differentiable at the origin, the Newton-Raphson algorithm and its variants cannot be directly used to find the solution for linear quantile regression. Koenker and Ng (2005) proposed several algorithms, such as the interior point algorithm for linear programming, to solve this optimization problem. However, with the layer-by-layer structure of the neural network and the large number of parameters involved, this approach is infeasible for our purpose. We resort to the *Adam* algorithm (Kingma and Ba 2014), a variant of the *stochastic gradient descent* (Robbins and Monro 1951), in the Python package *PyTorch* (Paszke et al. 2019) to solve the optimization problem (5). This algorithm is widely used in the deep learning field due to its computational and memory efficiency. For our purpose, since we have a parametric and a nonparametric component, we wrap the linear predictor $\theta^\top X$ and $m(Z)$ together and iteratively estimate the corresponding parameters simultaneously. That is, with the neural network m in (2), we use Adam to update the parameters $\{\theta, W_1,\ldots,W_L\}$. Here we use the default values in PyTorch for the initial values $\theta^{(0)}$ and $W_k^{(0)}, k=1,\ldots,L$.

The algorithm also requires the specification of tuning parameters, such as the depth L, width $\boldsymbol{q}$, step size, minibatch size, the number of iterations, early stopping, the constraint of sparsity parameter s and boundedness of weights $W_k, k =$

$1, \ldots, L$ of neural network m for the class $\mathcal{M}(s, L, \boldsymbol{q}, A)$ in (3). Here the minibatch size is defined as the subsample size used to calculate the gradient of the objective function for each iteration, and early stopping prevents overfitting by specifying the number of iterations to continue when the model does not improve any more on a hold-out validation dataset. We enforce a certain proportion of elements in each row of the weight matrix W_k to be zero and keep them at zero during the training. We set the proportion to be 50% since a different choice leads to similar results as shown in Table 12 of the Supplementary material. After the training, we set the elements of W_k, whose absolute values are greater than 1, to equal to 1 or -1, according to the sign of the elements.

In the simulation study, we randomly split the data into a training and validation set in a 80:20 ratio, where the tuning parameters were selected based on the 20% validation set. In the data analysis, where the ground truth is not available, we randomly reserved 20% of the data as testing set and then randomly selected 20% of the remaining data as a validation set to select the tuning parameter for the training set, which is 64% of the original data. Based on minimizing the check loss on the validation set, we used grid search to choose the tuning parameters among a set of candidates for both the simulation and data application. The choice of the grid points are provided in Tables 13 and 14 of the supplementary material.

Asymptotic Covariance Estimation: To conduct inference for the parameter θ_0, we need to estimate the asymptotic covariance matrix of $\hat{\theta}$ in Theorem 3.3 or Corollary 3.1. For simplicity, we demonstrate how to estimate the asymptotic covariance matrix for the case of homoscedastic random errors. The first step is to obtain a density estimate for $\hat{f}(0)$ from the residuals $\{\hat{\epsilon}_i = Y_i - \hat{Y}_i \mid \hat{Y}_i = X_i^\top \hat{\theta} + \hat{m}(Z_i), i = 1, \ldots, n\}$, for which we use the function *density* in the R package *stats*. Then, we employ the deep neural network to estimate the projections $\varphi_k^*, k = 1, \ldots, p$ empirically, that is,

$$\hat{\varphi}_k^* = \arg\min_{\varphi \in \mathcal{M}_1} \frac{1}{n} \sum_{i=1}^{n} \{X_{ik} - \varphi(Z_i)\}^2,$$

where X_{ik} is the kth component of covariates X_k and $\mathcal{M}_1$ is a class of neural networks. Let $\hat{\boldsymbol{\varphi}}^* = (\hat{\varphi}_1^*, \ldots, \hat{\varphi}_p^*)^\top$, $V_i = X_i - \hat{\boldsymbol{\varphi}}^*(Z_i)$, $\bar{V} = (\sum_{i=1}^n V_i)/n$ and $\hat{\Omega} = \{\sum_{i=1}^n (V_i - \bar{V})(V_i - \bar{V})^\top\}/(n-1)$. We estimate the asymptotic covariance matrix by

$$\hat{\Sigma} = \frac{\tau(1-\tau)\hat{\Omega}^{-1}}{\hat{f}^2(0)}. \tag{14}$$

For heteroscedastic random errors, we can estimate the corresponding asymptotic covariance matrix by a bootstrap method, see Feng, He, and Hu (2011) and Wang, Van Keilegom, and Maidman (2018) for details.

5. Simulations

In this section, we demonstrate the numerical performance of the proposed deep quantile regression method and compare it with linear quantile regression and partially linear additive quantile regression, abbreviated as LQR and PLAQR, respectively. The code is available at *https://github.com/qxzhong/dplqr*.

5.1. Simulation I: Homoscedastic Errors

We first generated $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_{10})^\top$ where $(\tilde{Z}_1, \ldots, \tilde{Z}_{10})^\top$ is from a Gaussian copula on $[0, 2]$ with correlation parameter 0.5. Marginally, each coordinate of $\tilde{Z}$ is a uniform distribution on $[0, 2]$. We then set $Z = (\tilde{Z}_1, \ldots, \tilde{Z}_8)^\top$ and $X = (X_1, X_2)^\top$ with $X_1 = I(\tilde{Z}_9 > 1)$ and $X_2 = \tilde{Z}_{10}$ as covariates. The response Y was generated from

$$Y = \theta^\top X + m(Z) + \epsilon, \tag{15}$$

where $\theta = (\theta_1, \theta_2) = (1, -1)^\top$, and the error ϵ, independent of (X, Z), is a Student's t-distribution with zero mean and 3 degrees of freedom. Three choices of m were implemented:

Case 1 (linear): $m(z) = 0.56 \times \sum_{k=1}^{8} z_k$;
Case 2 (additive): $m(z) = 0.82 \times \{(z_1 - 1)^2 - z_2^2 + 3|z_3 - 1| + 0.6 \sin(\pi z_4) + \log(z_5 + 0.5) + \sqrt{z_6 + 0.5} + 3\cos(0.1\pi z_7) + 3(z_8 - 1 + |z_8 - 1|)\}$
Case 3 (deep): $m(z) = 0.61 \times [\exp\{z_1(1 + z_2 - \pi z_3 z_4)/2\}(z_5 + 0.2) + z_5(z_4 - 0.3)/(|2z_4 - 1| + 1) + 2\sin(z_5)|z_5 z_6 - 0.6| + \log(z_6 + z_7 z_8)]$

The first two cases correspond to, respectively, the LQR and PLAQR model, and the third case is designed for DPLQR. The factors 0.56, 0.82, and 0.61 in each case were scaled to attain a signal-to-noise ratio around 5.

For each setting, we generated $Q = 200$ datasets, each with sample sizes $n = 1000$ and 2000, respectively. Throughout the simulation, we split the data into training data and validation data in a 80:20 ratio. That is, 80% of the data were used for estimation and the remaining 20% were used for the selection of tuning parameters as introduced in Section 4. To evaluate the performance of the estimation and prediction, we additionally generated a test data with sample size $N = 5000$ in each simulation. Specifically,

The performance of $\hat{m}_\tau$ was assessed by the relative mean squared error (RMSE):

$$\text{RMSE}(\hat{m}_\tau) = \frac{\frac{1}{N}\sum_{i=1}^{N}\{\hat{m}_\tau(Z_i) - m_\tau(Z_i)\}^2}{\frac{1}{N}\sum_{i=1}^{N}\{m_\tau(Z_i)\}^2}, \tag{16}$$

where $\hat{m}_\tau$ and m_τ are evaluated on the covariates $Z_i, i = 1, \ldots, N$ of the test data ($N = 5000$). Moreover, with the estimates $\hat{\theta}_\tau$ and $\hat{m}_\tau$, we use $\hat{Y}_i = X_i^\top \hat{\theta}_\tau + \hat{m}_\tau(Z_i)$ to predict the τth quantile of Y_i and evaluated its performance through the excess risk error (Van der Vaart and Wellner 1996):

$$\text{ERE}(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\{\rho_\tau(\hat{Y}_i - Y_i) - \rho_\tau(Y_i^* - Y_i)\},$$

where $\rho_\tau(\cdot)$ is the check loss function defined in (5) and $Y_i^* = X_i^\top \theta_\tau + m_\tau(Z_i)$.

Table 1 presents the biases and standard deviations of the estimates, $\hat{\theta}_1$, based on 200 simulation runs at three quantile levels $\tau = 0.25, 0.50, 0.75$. In general, both the bias and variance decrease steadily for all three methods as the sample size increases from 1000 to 2000. As expected, the mean squared error of the resulting estimates are the smallest at the median ($\tau = 0.5$) level. Under Case 1 (linear) and Case 2 (additive), the proposed DPLOR method performed comparably with the

Table 1. Bias and standard deviation (in parentheses) of $\hat{\theta}_1$ for the LQR, PLAQR, and DPLQR methods under homoscedastic random errors.

			$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
θ	Case	n	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR
θ_1	Case 1 (linear)	1000	0.0896 (0.1402)	0.0944 (0.1428)	0.1235 (0.1446)	0.0219 (0.1116)	0.0413 (0.1199)	0.0636 (0.1235)	−0.0721 (0.1389)	0.0836 (0.1411)	0.1203 (0.1439)
		2000	0.0742 (0.0970)	0.0821 (0.0998)	0.1078 (0.1074)	−0.0148 (0.0796)	0.0304 (0.0812)	0.0483 (0.0925)	−0.0679 (0.0988)	−0.0795 (0.1003)	0.0983 (0.1118)
	Case 2 (additive)	1000	0.0764 (0.1384)	−0.0297 (0.1253)	0.0360 (0.1301)	−0.0611 (0.1158)	−0.0124 (0.1008)	0.0338 (0.1116)	−0.0757 (0.1493)	−0.0210 (0.1364)	−0.0307 (0.1417)
		2000	0.0710 (0.1086)	−0.0145 (0.0970)	0.0229 (0.0976)	0.0506 (0.0794)	0.0099 (0.0782)	0.0158 (0.0785)	−0.0727 (0.0943)	0.0112 (0.1060)	−0.0266 (0.0984)
	Case 3 (deep)	1000	0.1244 −0.1060 (0.1787)	0.0762 (0.1647)	0.1068 (0.1394)	−0.0902 (0.1444)	0.0403 (0.1337)	−0.1322 (0.0998)	−0.1095 (0.1718)	0.0604 (0.1691)	(0.1330)
		2000	0.1170 (0.1231)	−0.0935 (0.1146)	0.0556 (0.0951)	−0.0951 (0.1046)	−0.0630 (0.0978)	0.0297 (0.0848)	−0.1212 (0.1165)	−0.0942 (0.1146)	−0.0445 (0.0941)

Table 2. Empirical coverage probability of the 95% confidence interval for θ_1 by the LQR, PLAQR, and DPLQR methods under homoscedastic random errors.

			$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
θ	Case	n	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR
θ_1	Case 1	1000	0.910	0.900	0.895	0.975	0.915	0.905	0.900	0.910	0.905
	(linear)	2000	0.925	0.920	0.915	0.945	0.940	0.935	0.930	0.920	0.925
	Case 2	1000	0.855	0.910	0.895	0.960	0.955	0.945	0.865	0.915	0.910
	(additive)	2000	0.885	0.925	0.915	0.960	0.945	0.955	0.895	0.935	0.930
	Case 3	1000.	0.865	0.885	0.910	0.885	0.875	0.930	0.870	0.895	0.925
	(deep)	2000	0.900	0.925	0.935	0.915	0.920	0.955	0.905	0.920	0.945

Table 3. Relative mean squared error of $\hat{m}$ for the LQR, PLAQR, and DPLQR methods under homoscedastic random errors.

		$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
Case	n	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR
Case 1	1000	0.0071	0.0080	0.0086	0.0064	0.0070	0.0081	0.0074	0.0083	0.0089
(linear)	2000	0.0036	0.0043	0.0047	0.0034	0.0038	0.0040	0.0037	0.0044	0.0049
Case 2	1000	0.0097	0.0072	0.0074	0.0062	0.0053	0.0059	0.0093	0.0074	0.0080
(additive)	2000	0.0059	0.0040	0.0044	0.0034	0.0027	0.0030	0.0051	0.0039	0.0045
Case 3	1000	0.0997	0.0872	0.0390	0.0679	0.0539	0.0209	0.0976	0.0820	0.0342
(deep)	2000	0.0904	0.0812	0.0298	0.0633	0.0508	0.0159	0.0883	0.0774	0.0275

optimal method (LQR and PLAQR, respectively) with slightly larger mean squared errors. However, under Case 3 (deep), the DPLQR method clearly outperforms LQR and PLAQR. We also construct the 95% confidence intervals for θ_1 based on the estimates of the asymptotic variance in Section 4. Table 2 reports the empirical coverage probabilities of the 95% confidence intervals. For all three cases, the empirical coverage probabilities of the proposed method generally approach 95% as n increases. In addition, the proposed method is comparable to the other two methods under Case 1 (linear) and Case 2 (additive), and has more accurate coverage rates under Case 3 (deep). We also compare the empirical overage probabilities of the 95% bootstrap confidence intervals for θ and variance estimate of $\hat{\theta}$ in Table 7 of the supplementary material and the proposed method is seen to outperform the bootstrap method in terms of coverage probability.

The average relative mean squared errors of the estimated nonparametric function $\hat{m}$ over 200 repetitions are given in Table 3. They decline with the increasing sample sizes as expected. When the true model is Case 3 (deep), the proposed method substantially outperforms LQR and PLAQR, while it performs slightly worse under Case 1 (linear) and Case 2 (additive).

Based on the 200 simulation runs, Table 4 shows the mean and standard deviation of the excess risk errors for the prediction at three quantile levels $\tau = 0.25, 0.50, 0.75$. This reveals that the proposed DPLQR is competitive with the optimal procedure (LQR in Case 1 (linear) and PLAQR in Case 2 (additive)) and superior in Case 3 (deep).

5.2. Simulation II: Heteroscedastic Errors

We also studied the performance of the proposed method for heteroscedastic errors. The covariates $U = (X, Z)$, coefficient θ and nonparametric function m are similar to the settings in Section 5.1 but the response Y now comes from the heteroscedastic regression model:

$$Y = X^\top \theta + m(Z) + \sigma(X, Z)\epsilon.$$

Here ϵ follows the Student's t-distribution with zero mean and 3 degrees of freedom. The function $\sigma(X, Z)$ was chosen with three settings:

Case 4 (linear): $\sigma(x, z) = (x_1 + x_1 + \sum_{k=1}^{8} z_k)/10$;
Case 5 (additive): $\sigma(x, z) = \{x_1 + x_1 + \sum_{k=1}^{8}(z_k - 1)^2\}/10$;
Case 6 (deep): $\sigma(x, z) = (x_1 + x_1)/10 + \Phi\big(\sum_{k=1}^{8}(z_k - 1)/8\big)$

Table 4. Mean and standard deviation (in parentheses) of the excess risk to evaluate prediction for the LQR, PLAQR, and DPLQR methods under homoscedastic random errors.

		$\tau = 0.25$			$\tau = 0.50$			$\tau = 0.75$		
Case	n	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR	LQR	PLAQR	DPLQR
Case 1 (linear)	1000	0.0053 (0.0026)	0.0061 (0.0036)	0.0067 (0.0038)	0.0054 (0.0029)	0.0065 (0.0040)	0.0069 (0.0045)	0.0049 (0.0026)	0.0053 (0.0033)	0.0056 (0.0040)
	2000	0.0026 (0.0014)	0.0042 (0.0019)	0.0045 (0.0020)	0.0032 (0.0019)	0.0049 (0.0026)	0.0053 (0.0031)	0.0026 (0.0014)	0.0040 (0.0018)	0.0041 (0.0021)
Case 2 (additive)	1000	0.0141 (0.0054)	0.0111 (0.0039)	0.0124 (0.0050)	0.0158 (0.0065)	0.0124 (0.0043)	0.0143 (0.0053)	0.0146 (0.0056)	0.0106 (0.0039)	0.0127 (0.0052)
	2000	0.0105 (0.0041)	0.0081 (0.0024)	0.0098 (0.0039)	0.0129 (0.0048)	0.0096 (0.0032)	0.0113 (0.0040)	0.0102 (0.0040)	0.0076 (0.0025)	0.0100 (0.0037)
Case 3 (deep)	1000	0.0141 (0.0064)	0.0123 (0.0050)	0.0116 (0.0051)	0.0160 (0.0071)	0.0130 (0.0065)	0.0128 (0.0058)	0.0140 (0.0062)	0.0125 (0.0056)	0.0110 (0.0048)
	2000	0.0108 (0.0046)	0.0099 (0.0044)	0.0089 (0.0039)	0.0128 (0.0054)	0.0103 (0.0049)	0.0092 (0.0040)	0.0099 (0.0044)	0.0103 (0.0041)	0.0088 (0.0037)

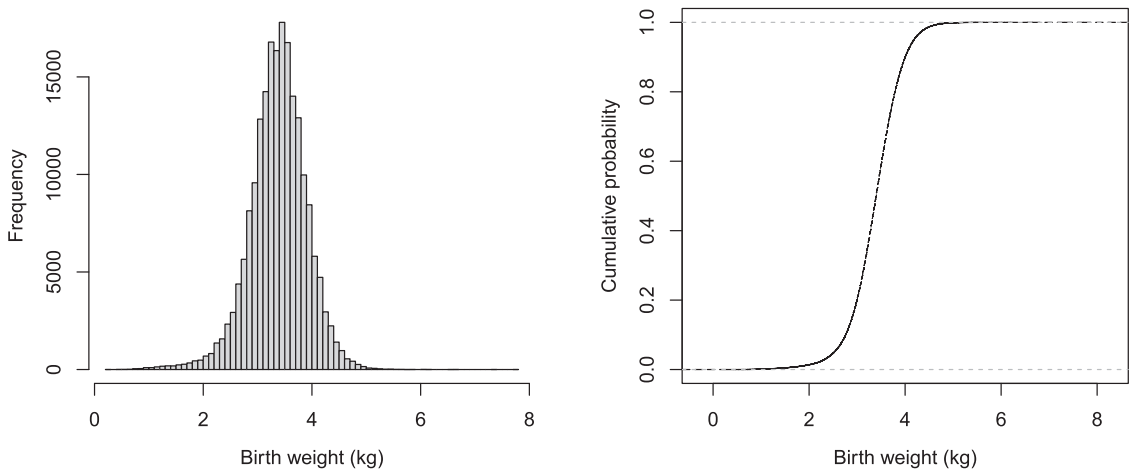

Figure 2. Histogram (left panel) and empirical cumulative distribution function (right panel) of birth weight on 200 thousand randomly selected subjects.

with the cumulative distribution function $\Phi(\cdot)$ of the standard normal distribution.

These lead to $\theta_\tau = \theta + t_\tau\theta^*$ and $m_\tau(z) = m(z) + t_\tau m^*(z)$ where t_τ is the τ quantile of Student's t-distribution with zero mean and degree of freedom 3 and $(\theta^*, m^*(z))$ takes $(0.1, \sum_{k=1}^{10} z_k/10), (0.1, \sum_{k=1}^{8}(z_k - 1)^2/10)$ and $(0.1, \Phi(\sum_{k=1}^{8}(z_k - 1)/8))$ in Case 4 (linear), Case 5 (additive) and Case 6 (deep), respectively.. The simulation results, summarized in Tables 8–11 of the supplementary material, are comparable to those in Simulation I in Section 5.1.

In summary, when the true model is linear or partially linear additive quantile regression, our method is competitive for both the parametric coefficients and nonparametric function estimates, and the coverage probabilities for the parametric coefficients are close to the 95% nominal level as sample sizes increase. Furthermore, the proposed method is superior to the LQR and PLAQR methods when the true model comes from the deep partially linear quantile regression.

6. An application to Natality Birth Data

We apply DPLQR to analyze the relationship between maternal education and the birth weight of infants using the 2020 Natality Birth Data published by National Center for Health Statistics. The data, available at *https://www.nber.org/research/data/vital-statistics-natality-birth-data*, consists of more than 3.6 million births with demographic and health information. It is well documented that low birth weight is associated with several short-term and long-term consequences, such as high risk of mortality (Badshah et al. 2008) and impaired language development (Zerbeto, Cortelo, and C Filho 2015). Scientists are interested in whether maternal education beyond high school is associated with birth weight (Shi et al. 2004; Gage et al. 2013). We focus on singleton births to white mothers who were over 30 as very few continued to get a college degree after 30.

Along with birth weight of the baby in kilograms (Y) and indicator for education beyond high school (X), where X is 0 if the mother's education is high school or less and 1 else, six continuous variables [mother's age (Z_1), mother's body mass index (Z_2), mother's weight gain during pregnancy (Z_3), gestation period (Z_4), number of cigarettes the mother smoked during pregnancy (Z_5), and father's age (Z_6)] and seven categorical variables [gender of infant (Z_7), mother's marital status (Z_8), pre-pregnancy diabetes indicator (Z_9), usage of induction of labor (Z_{10}), usage of antibiostic during labor (Z_{11}), receipt of the Special Supplemental Nutrition Program for Women, Infants, and Children (Z_{12}), indicator for father's education beyond high school (Z_{13})] are included.

After excluding subjects with missing values in any variable, there are about 978,000 subjects left. From these, we randomly select 200,000 subjects in our study to reduce the computational burden. Figure 2 shows the histogram and empirical cumulative

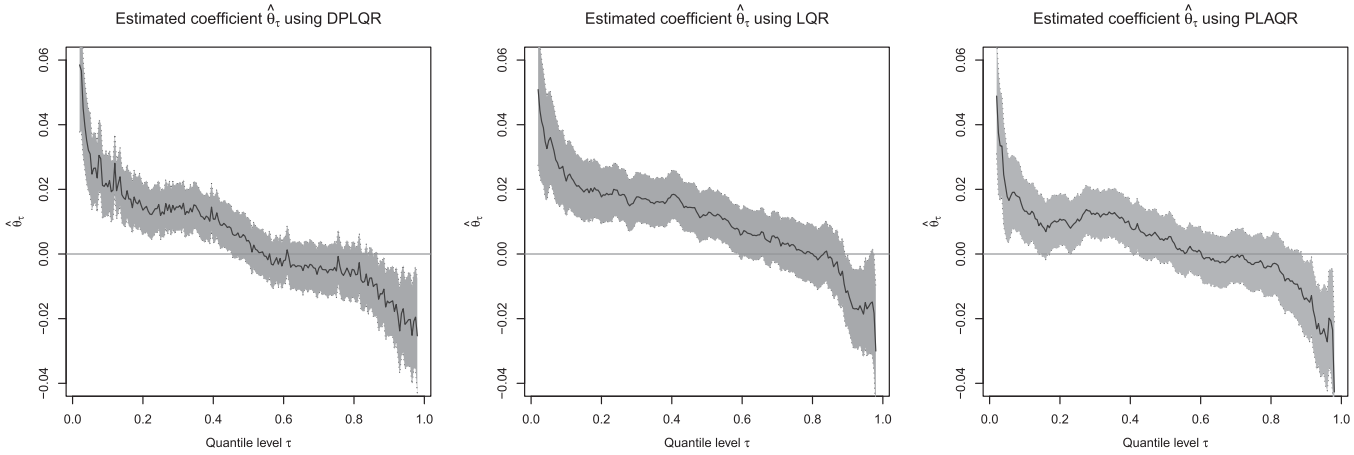

Figure 3. Estimation for θ_τ with $\tau \in [0.02, 0.98]$ and 95% confidence interval using DPLQR (left panel), LQR (middle panel), and PLAQR (right panel) in Natality Birth Data.

distribution function of birth weight for 200,000 infants, for which the mean (standard deviation) and median (median absolute deviation) are 3.38 (0.52) and 3.40 (0.46), and 20.21% of the mothers have no more than a high school degree.

With the above variables we consider the following DPLQR model:

$$Y = \theta_\tau X + m_\tau(Z_1, \ldots, Z_{13}) + \epsilon. \tag{17}$$

We randomly reserve 20% of the data as testing set and then randomly select 20% of the remaining data as a validation set to select the tuning parameters for the training set, which is 64% of the data, to estimate the unknown θ_τ and m_τ for $\tau = 0.020 + 0.005k, k = 0, \ldots, 192$. The left panel of Figure 3 shows the estimated coeffcient θ_τ, where the shaded area is the 95% confidence interval. Under the DPLQR model (17), a positive θ_τ implies that mother with education beyond high school tends to have a heavier infant and vice versa for a negative θ_τ.

Along the quantile levels τ, the estimated θ_τ in the left panel of Figure 3 starts with a largest positive value, but goes down fast before $\tau = 0.15$ and declines relatively slowly to zero at about $\tau = 0.52$. Then $\hat{\theta}_\tau$ flattens out up to $\tau = 0.85$, and finally decreases to -0.028 at $\tau = 0.98$. The 95% confidence intervals (shaded regions) cover zero at $\tau \in [0.45, 0.89]$. The results also show that a lower level of maternal education is significantly associated with lower birth weight for newborns with low birth weights (below $\tau = 0.45$ quantiles). The effect decreases monotonically as the quantile increases from $\tau = 0$ to $\tau = 0.45$. This is consistent with existing knowledge (Shi et al. 2004; Gage et al. 2013) that mothers with less education are more likely to deliver low birth weight infants.

We also model the data with the linear quantile regression (LQR) and partially linear additive quantile regression (PLAQR), where the unknown m_τ in (17) is a linear and nonparametric additive function, respectively. The middle and right panels of Figure 3 provide the point estimates and associated 95% confidence intervals. Generally, both estimates $\hat{\theta}_\tau$ also have a downward trend as the quantile levels τ increases from 0.02 to 0.98.

However, PLAQR produces inconsistent results, namely that education is statistically significant for all birth weights before $\tau = 0.40$ quantile levels but not at an interval around $\tau = 0.18$. The 95% confidence intervals of LQR appear significant for $\tau \leq 0.55$, which implies that the education level of a mother also has an effect on normal-birth-weight ($0.45 \leq \tau \leq 0.55$) infants. This is scientifically questionable and not supported by the other two methods.

Intriguingly, the proposed DPLQR method produces narrower confidence intervals than LQR and PLAQR. The areas covered by the 95% confidence intervals (shaded regions) are 0.0186, 0.0184, and 0.0168 for LQR, PLAQR, and DPLQR, respectively. We conclude that DPLQR provides the best fit and interpretation of the maternal education effect for the Natality Birth Data.

We next compare the prediction results of DPLQR with LQR, PLAQR, and the deep nonparametric quantile regression (DNQR) approach in Padilla, Tansey, and Chen (2020) and Shen et al. (2021), which are not amenable to checking the effect of maternal education on birth weight. The evaluation criterion is the check loss (CL) at τ and the average check loss (ACL),

$$\text{CL}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(\hat{Y}_i - Y_i) \text{ and ACL} = \int_0^1 \text{CL}(\tau) d\tau, \tag{18}$$

where $N = 40{,}000$ is the size of the test set, Y_i is the birth weight of the ith infant in the test set, and $\hat{Y}_i$ is the predicted quantile level of this subject. The results in Figure 4 suggest that the proposed DPLQR is superior to LQR and PLAQR, while it is as good as DNQR. This means that DPLQR does not sacrifice much prediction accuracy compared to the larger DNQR model, yet it provides an interpretable model.

Overall, the proposed method yields more stable, accurate and convincing results.

7. Conclusion

We provide an interpretable-yet-flexible deep learning model with partially linear quantile regression, where we leverage neural networks to represent the nonparametric function and the linear predictor to obtain inference. The proposed method is able to detect the parsimonious structure of the data automatically, thereby producing a better convergence rate for the nonparametric estimator $\hat{m}$ than conventional nonparametric

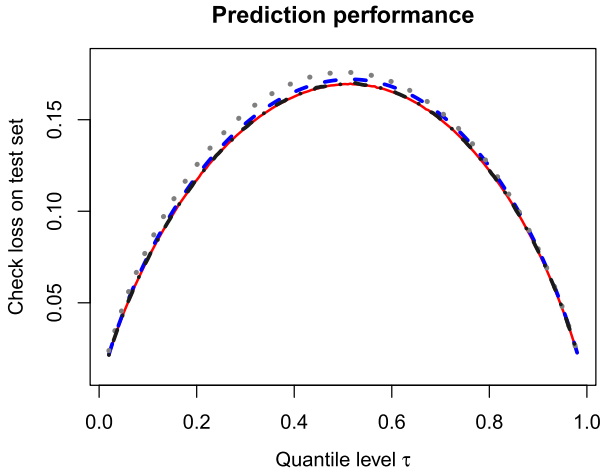

Figure 4. Check loss evaluated on test set with quantile levels $\tau \in [0.02, 0.98]$ under four models: LQR, PLAQR, DPLQR, and DNQR. Dotted line: LQR (ACL=0.1298); dashed line: PLAQR (ACL=0.1265); solid red line: DPLQR (ACL=0.1245); dot-dashed line: DNQR (ACL=0.1244). Solid line and dot-dashed line almost overlap.

smoothing methods. Furthermore, the estimator of the parameter θ_0 attains $\sqrt{n}$-consistency and asymptotic normality. These features substantially distinguish our method from the nonparametric approaches of Padilla, Tansey, and Chen (2020), Schmidt-Hieber (2020), Shen et al. (2021) and opens up a myriad of future research opportunities for semiparametric regression models.

So far, we have developed estimation and statistical inference for the linear parameters of the proposed DPLQR model. Model checking for the linearity assumption is a challenging problem of future interest. An ad hoc lack-of-fit test for linearity could be implemented by randomly spliting the data into a training and test set with 8:2 ratio, where the nonparametric component is estimated on the training set, giving an estimate $\hat{m}_\tau(\cdot)$. Then one can employ methods such as in Zheng (1998), He and Zhu (2003), and Escanciano and Goh (2014) to conduct a lack-of-fit test for linear quantile regression with the "observations" $(X_i, \tilde{Y}_i), i = 1, \ldots, N$ on the test set, where $\tilde{Y}_i = Y_i - \hat{m}_\tau(Z_i)$ and N is the sample size of the test set.

A possible extension is to investigate the quantile regression process instead of fitting quantiles at fixed levels. Chao, Volgushev, and Cheng (2017) and Belloni et al. (2019) studied convergence results uniformly on τ for quantile functions approximated by linear combinations of basis functions, for example, polynomial, Fourier, spline and wavelet bases. However, their approaches cannot easily be extended to the deep learning setting because of the layer structure in a neural network.

In additional, one can use techniques like monotonic constrained regression (Barlow et al. 1972; Bondell, Reich, and Wang 2010) to ensure noncrossing of the estimated quantile functions, that is, $\hat{\theta}_{\tau_1}^\top x + \hat{m}_{\tau_1}(z) \leq \hat{\theta}_{\tau_2}^\top x + \hat{m}_{\tau_2}(z)$ whenever $\tau_1 \leq \tau_2$.

As we focus in this article on a fixed but moderate size of the linear covariates X, future work of interest is to study DPLQR with high-dimensional covariates, where the number of linear covariates may grow at a certain rate with sample size. A special case for PLAQR was studied in Sherwood and Wang (2016), which may shed some light on extending the DPLQR approach. Another future direction is to test whether a particular set of covariates contributes significantly to prediction, which includes testing whether the function m_0 satisfies $m_0(z) = 0$, for all $z \in [0, 1]^q$. Reliable estimates of the unknown parameters are fundamental to the hypothesis test procedures. The proposed methods provide a stepping stone to establish the asymptotic normality of the neural network estimates via Donsker Theorems [see, e.g., chap. 2 in Van der Vaart and Wellner (1996)] or of neural network test statistics via sample splittings (Dai, Shen, and Pan 2021).

Supplementary Materials

Mathematical proofs and additional numerical results are given in the supplementary material file.

Acknowledgments

The authors are grateful to the editor, associate editor and referees for their constructive comments and suggestions that led to numerous improvements of the article.

Disclosure Statement

There are no competing interests to declare.

Funding

Zhong's research was supported by the National Natural Science Foundation of China (12201527 and 11931001) and the Fundamental Research Funds for the Central Universities (20720221032). Wang's research was supported by the US National Science Foundation (DMS19-14917 and 22-10891).

ORCID

Qixian Zhong http://orcid.org/0000-0003-1666-2191

References

Anthony, M., and Bartlett, P. L. (1999), *Neural Network Learning: Theoretical Foundations*, Cambridge: Cambridge University Press. [5]

Badshah, S., Mason, L., McKelvie, K., Payne, R., and Lisboa, P. J. (2008), "Risk Factors for Low Birthweight in the Public-Hospitals at Peshawar, NWFP-Pakistan," *BMC Public Health*, 8, 1–10. [2,8]

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference under Order Restrictions.*, Wiley Series in Probability and Mathematical Statistics, London-New York-Sydney: Wiley. [10]

Bauer, B., and Kohler, M. (2019), "On Deep Learning as a Remedy for the Curse of Dimensionality in Nonparametric Regression," *The Annals of Statistics*, 47, 2261–2285. [2,3,5]

Belloni, A., Chernozhukov, V., Chetverikov, D., and Fernández-Val, I. (2019), "Conditional Quantile Processes based on Series or Many Regressors," *Journal of Econometrics*, 213, 4–29. [10]

Bhattacharya, J., Gimenes, N., and Guerre, E. (2021), "Semiparametric Quantile Models for Ascending Auctions with Asymmetric Bidders," *Journal of Business & Economic Statistics*, 40, 1020–1033. [2]

Bondell, H. D., Reich, B. J., and Wang, H. (2010), "Noncrossing Quantile Regression Curve Estimation," *Biometrika*, 97, 825–838. [10]

Cai, Z., and Xiao, Z. (2012), "Semiparametric Quantile Regression Estimation in Dynamic Models with Partially Varying Coefficients," *Journal of Econometrics*, 167, 413–425. [2]

Chao, S.-K., Volgushev, S., and Cheng, G. (2017), "Quantile Processes for Semi and Nonparametric Regression," *Electronic Journal of Statistics*, 11, 3272–3331. [10]

Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and their Local Bahadur Representation," *The Annals of Statistics*, 19, 760–777. [2]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017), "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, 107, 261–65. [1]

Cybenko, G. (1989), "Approximation by Superpositions of a Sigmoidal Function," *Mathematics of Control, Signals and Systems*, 2, 303–314. [1]

Dai, B., Shen, X., and Pan, W. (2021), "Significance Tests of Feature Relevance for a Blackbox Learner," arXiv preprint arXiv:2103.04985. [10]

Escanciano, J. C., and Goh, S.-C. (2014), "Specification Analysis of Linear Quantile Models," *Journal of Econometrics*, 178, 495–507. [10]

Fan, J., Ma, C., and Zhong, Y. (2021), "A Selective Overview of Deep Learning," *Statistical Science*, 36, 264–290. [3]

Fan, Y., and Liu, R. (2016), "A Direct Approach to Inference in Nonparametric and Semiparametric Quantile Models," *Journal of Econometrics*, 191, 196–216. [2]

Fang, Z., Li, Q., and Yan, K. (2021), "A Simple Nonparametric Approach for Estimation and Inference of Conditional Quantile Functions," *Econometric Theory*, 39, 290–320. [2]

Feng, X., He, X., and Hu, J. (2011), "Wild Bootstrap for Quantile Regression," *Biometrika*, 98, 995–999. [6]

Gage, T. B., Fang, F., O'Neill, E., and DiRienzo, G. (2013), enquoteMaternal Education, Birth Weight, and Infant Mortality in the United States, *Demography*, 50, 615–635. [8,9]

Gan, D., Wang, Y., Yang, S., and Kang, C. (2018), "Embedding based Quantile Regression Neural Network for Probabilistic Load Forecasting," *Journal of Modern Power Systems and Clean Energy*, 6, 244–254. [2]

Goodfellow, I., Bengio, Y., and Courville, A. (2016), *Deep Learning*, Cambridge, MA: MIT Press. [3]

Guerre, E., and Sabbah, C. (2012), "Uniform Bias Study and Bahadur Representation for Local Polynomial Estimators of the Conditional Quantile Function," *Econometric Theory*, 28, 87–129. [2]

Han, S., Pool, J., Tran, J., and Dally, W. (2015), "Learning both Weights and Connections for Efficient Neural Network," in *Proceedings of Neural Information Processing Systems*, pp. 1135–1143. [3]

Hatalis, K., Lamadrid, A. J., Scheinberg, K., and Kishore, S. (2017), "Smooth Pinball Neural Network for Probabilistic Forecasting of Wind Power," arXiv preprint arXiv:1710.01720. [2]

He, X., and Shi, P. (1996), "Bivariate Tensor-Product B-splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58, 162–181. [2,5]

He, X., and Zhu, L.-X. (2003), "A Lack-of-Fit Test for Quantile Regression," *Journal of the American Statistical Association*, 98, 1013–1022. [10]

Heaton, J. B., Polson, N. G., and Witte, J. H. (2017), "Deep Learning for Finance: Deep Portfolios," *Applied Stochastic Models in Business and Industry*, 33, 3–12. [1]

Hornik, K., Stinchcombe, M., and White, H. (1989), "Multilayer Feedforward Networks are Universal Approximators." *Neural Networks*, 2, 359–366. [1]

Hoshino, T. (2014), "Quantile Regression Estimation of Partially Linear Additive Models," *Journal of Nonparametric Statistics*, 26, 509–536. [2,4]

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics*, 29, 1537–1566. [4]

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120. [4,5]

Jantre, S. R., Bhattacharya, S., and Maiti, T. (2020), "Quantile Regression Neural Networks: A Bayesian Approach," arXiv preprint arXiv:2009.13591. [2]

Jones, M., and Hall, P. (1990), "Mean Squared Error Properties of Kernel Estimates or Regression Quantiles," *Statistics & Probability Letters*, 10, 283–289. [2]

Kingma, D. P., and Ba, J. (2014), "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980. [5]

Koenker, R. (2005), *Quantile Regression*, Cambridge: Cambridge University Press. [2]

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50. [1,2]

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017), *Handbook of Quantile Regression*, Boca Raton, FL: CRC Press. [2]

Koenker, R., and Ng, P. (2005), "Inequality Constrained Quantile Regression," *Sankhyā: The Indian Journal of Statistics*, 67, 418–440. [5]

Kong, E., Linton, O., and Xia, Y. (2013), "Global Bahadur Representation for Nonparametric Censored Regression Quantiles and its Applications," *Econometric Theory*, 29, 941–968. [2]

Kong, E., and Xia, Y. (2012), "A Single-Index Quantile Regression Model and its Estimation," *Econometric Theory*, 730–768. [2]

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012), "Imagenet Classification with Deep Convolutional Neural Networks," in *Proceedings of Neural Information Processing Systems*, pp. 1097–1105. [1,3]

Lee, S. (2003), "Efficient Semiparametric Estimation of a Partially Linear Quantile Regression Model," *Econometric Theory*, 19, 1–31. [2]

Li, Q., and Racine, J. S. (2008), "Nonparametric Estimation of Conditional CDF and Quantile Functions with Mixed Categorical and Continuous Data," *Journal of Business & Economic Statistics*, 26, 423–434. [2]

Lian, H. (2012), "Semiparametric Estimation of Additive Quantile Regression Models by Two-Fold Penalty," *Journal of Business & Economic Statistics*, 30, 337–350. [2,4]

Ma, S., and He, X. (2016), "Inference for Single-Index Quantile Regression Models with Profile Optimization," *The Annals of Statistics*, 44, 1234–1268. [2]

Mi, X., Tighe, P., Zou, F., and Zou, B. (2021), "A Deep Learning Semiparametric Regression for Adjusting Complex Confounding Structures," *The Annals of Applied Statistics*, 15, 1086–1100. [1]

Noh, H., Ghouch, A. E., and Van Keilegom, I. (2015), "Semiparametric Conditional Quantile Estimation through Copula-based Multivariate Models," *Journal of Business & Economic Statistics*, 33, 167–178. [2]

Nolle, T., Seeliger, A., and Mühlhäuser, M. (2018), "BINet: Multivariate Business Process Anomaly Detection Using Deep Learning," in *International Conference on Business Process Management*, pp. 271–287, Springer. [1]

Padilla, O. H. M., Tansey, W., and Chen, Y. (2020), "Quantile Regression with ReLU Networks: Estimators and Minimax Rates," arXiv preprint arXiv:2010.08236. [2,9,10]

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019), "Pytorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems* (Vol. 32). [5]

Portnoy, S. (1991), "Asymptotic Behavior of Regression Quantiles in Nonstationary, Dependent Cases," *Journal of Multivariate Analysis*, 38, 100–113. [2]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *Annals of Mathematical Statistics*, 22, 400–407. [5]

Rolnick, D., and Tegmark, M. (2017), "The Power of Deeper Networks for Expressing Natural Functions," arXiv preprint arXiv:1705.05502. [2]

Romano, Y., Patterson, E., and Candes, E. (2019), "Conformalized Quantile Regression," in *Proceedings of Neural Information Processing Systems*, pp. 3543–3553. [2]

Rudin, C. (2019), "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, 1, 206–215. [1]

Samanta, M. (1989), "Non-parametric Estimation of Conditional Quantiles," *Statistics & Probability Letters*, 7, 407–412. [2]

Schmidt-Hieber, J. (2020), "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function," *The Annals of Statistics*, 48, 1875–1897. [2,3,4,5,10]

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. (2021), "Deep Quantile Regression: Mitigating the Curse of Dimensionality Through Composition," arXiv preprint arXiv:2107.04907. [2,9,10]

Sherwood, B., and Wang, L. (2016), "Partially Linear Additive Quantile Regression in Ultra-High Dimension," *The Annals of Statistics*, 44, 288–317. [2,5,10]

Shi, L., Macinko, J., Starfield, B., Xu, J., Regan, J., Politzer, R., and Wulu, J. (2004), "Primary Care, Infant Mortality, and Low Birth Weight in the States of the USA," *Journal of Epidemiology & Community Health*, 58, 374–380. [8,9]

Simonoff, J. S. (2012), *Smoothing Methods in Statistics*, New York: Springer. [5]

Stone, C. J. (1985), "Additive Regression and other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [4,5]

Telgarsky, M. (2016), "Benefits of Depth in Neural Networks," arXiv preprint arXiv:1602.04485. [1]

Van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [6,10]

Wang, H. J., Zhu, Z., and Zhou, J. (2009), "Quantile Regression in Partially Linear Varying Coefficient Models," *The Annals of Statistics*, 37, 3841–3866. [5]

Wang, L., Van Keilegom, I., and Maidman, A. (2018), "Wild Residual Bootstrap Inference for Penalized Quantile Regression with Heteroscedastic Errors," *Biometrika*, 105, 859–872. [6]

Wu, C., and Yu, Y. (2014), "Partially Linear Modeling of Conditional Quantiles Using Penalized Splines," *Computational Statistics & Data Analysis*, 77, 170–187. [2]

Wu, T. Z., Yu, K., and Yu, Y. (2010), "Single-Index Quantile Regression," *Journal of Multivariate Analysis*, 101, 1607–1621. [2]

Yarotsky, D. (2017), "Error Bounds for Approximations with Deep ReLU Networks," *Neural Networks*, 94, 103–114. [1,5]

Yuan, Y., Deng, Y., Zhang, Y., and Qu, A. (2020), "Deep Learning from a Statistical Perspective," *Stat*, 9, e294. [3]

Zeiler, M. D., and Fergus, R. (2014), "Visualizing and Understanding Convolutional Networks," in *Proceedings of European Conference on Computer Vision*, pp. 818–833. [1]

Zerbeto, A. B., Cortelo, F. M., and C Filho, É. B. (2015), "Association between Gestational Age and Birth Weight on the Language Development of Brazilian Children: A Systematic Review," *Jornal de Pediatria*, 91, 326–332. [8]

Zhang, Y., Lian, H., and Yu, Y. (2017), "Estimation and Variable Selection for Quantile Partially Linear Single-Index Models," *Journal of Multivariate Analysis*, 162, 215–234. [2]

——— (2020), "Ultra-High Dimensional Single-Index Quantile Regression," *Journal of Machine Learning Research*, 21, 1–25. [2]

Zheng, J. X. (1998), "A Consistent Nonparametric Test of Parametric Regression Models under Conditional Quantile Restrictions," *Econometric Theory*, 14, 123–138. [10]