Instrumental Variable Quantile Regression Using Artificial Neural Networks

Sami Abduraman

May 16, 2025

Abstract

In this paper, we extend the instrumental variable quantile regression model to a semiparametric framework. Using a partially linear structural quantile model and a class of artificial neural networks achieving a convergence rate of $o(n^{-1/4})$, we obtain a heterogeneous quantile treatment effect estimator that is root-n asymptotic normal. We also propose a stochastic gradient descent (with momentum) approach to estimating this treatment effect, which possesses several attractive features relative to the grid search method typically used in this setting. We apply the semiparametric procedure to empirically re-investigate the impact of Medicaid coverage on the savings of households across a range of wealth groups.

Author Affiliation: Toronto Metropolitan University, Department of Economics

Email: sami.abdurahman@torontomu.ca

Computations were performed on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by Innovation, Science and Economic Development Canada; the Digital Research Alliance of Canada; the Ontario Research Fund: Research Excellence; and the University of Toronto.

Full code is available at: https://github.com/samiabd8/IVQRNN

Contents

1 Introduction

Estimating heterogeneous quantile treatment effects (QTEs) is central to evaluating how policies impact populations across outcome distributions. The instrumental variable quantile

regression (IVQR) model of Chernozhukov and Hansen (2006) has emerged as a powerful tool for addressing endogeneity in this context, enabling valid causal inference in the presence of endogenous treatment variables. However, the parametric assumptions of this model limits its ability to capture complex nonlinear covariate effects, which is especially problematic when analyzing outcomes shaped by interactive socioeconomic factors.

Despite the flexibility of artificial neural networks (aNN) in modeling nonlinear relationships, their implementation in semiparametric IVQR frameworks remains underexplored. This paper bridges these gaps in the literature by proposing a semiparametric IVQR model that combines a partially linear structural quantile specification with a class of single hidden layer[1] aNN. Our IVQRNN framework applies the sieve estimation approach of Chen and White (1999) to achieve a $o(n^{-1/4})$ convergence rate of the aNN, which is fast enough to ensure root-n asymptotic normality of the linear QTE estimator; making it the first aNN-based QR model that allows for endogeneity.

We also introduce a stochastic gradient descent (SGD) with momentum algorithm to address the shortcomings of the standard grid search method used in the IVQR model. We find that this approach reduces computation time by over 50% without sacrificing the robustness to local minima of grid search. In practice, the IVQRNN model equips researchers with a lightweight, data-driven tool for estimating nonlinear covariate effects without sacrificing the interpretability of the QTE. These advantages are demonstrated through investigating the effect Medicaid has on private household savings, revisting the findings of Maynard and Qiu (2009).

Our key results are as follows: the IVQRNN estimator achieves asymptotic normality under standard sieve estimation conditions, Monte Carlo simulations show reduced bias relative to the linear IVQR model (and shorter computation time under the SGD method), and we uncover statistically significant effects of Medicaid on savings for the poorest households; a result that was masked by parametric assumptions in prior work. In the next section, we

[1] Also referred to as "shallow" networks.

expand on the foundations of IVQR and aNN sieve estimators. Section 3 introduces the IVQRNN model and establishes theoretical guarantees, in addition to outlining the SGD with momentum alternative to grid search. Results of simulations and the aforementioned empirical application are covered in Section 4.

2 Model Framework

This section provides some background for each key component of the proposed original IVQRNN model. The first is the linear instrumental variable quantile regression (IVQR) model from Chernozhukov and Hansen (2005, 2006, 2008), which is modified in the IVQRNN model so that the impact of the covariates on the outcome is estimated using a neural network. In the context of neural networks based on the method of sieves used for estimation in quantile regression, we discuss the "QRNN" model in White (1992) and relevant works building on this approach. The resulting IVQRNN model extends the framework of White (1992) for settings involving endogeneity, the first to do so. Lastly, a class of single hidden layer sieve networks that converges quickly enough to deliver a root-n asymptotic normal plug-in estimator proposed by Chen and White (1999) are introduced. In addition to providing sufficiently fast convergence rates for our linear QTE estimator, this network structure offers a starting point for bridging the literature on sieve estimation with deep neural networks in the econometrics literature.

2.1 The instrumental variable quantile regression model

Given an outcome variable of interest Y and a vector (D, X), where D is the treatment variable and X is a vector of covariates, the structural equation of the linear QR model under endogeneity can be written as: $Y = h(D, X, U)$. Here, U is an unobserved (scalar) random variable[2] and $\tau \to h(D, X, U)$ is strictly increasing in τ by assumption for almost all

[2]Typically referred to as the "rank" variable and captures unobserved heterogeneity, such as latent ability.

(D, X). The case Chernozhukov and Hansen (2006, 2008) consider is the linear-in-parameters structural QR model, where $h(D, X, U) = D'\alpha(U) + X'\beta(U)$. With a set of (excluded) instruments given by Z, the instrumental variable quantile regression (IVQR) model from Chernozhukov and Hansen (2008) can be written as follows:

$$Y = D'\alpha(U) + X'\beta(U), \quad U \mid X, Z \sim Unif(0,1) \tag{1}$$

$$D = \delta(X, Z, V) \; for \; an \; unknown \; function \; \delta, \; random \; vector \; V \tag{2}$$

$$\tau \to D'\alpha(\tau) + X'\beta(\tau) \; is \; strictly \; increasing \; in \; \tau \tag{3}$$

where the stochastic process (Y, D, X, Z) is assumed to be i.i.d and defined on a (complete) probability measure space (Ω, F, P). It should be noted that there is no restriction between the rank variable U and V, which is the potential source of endogeneity in D and also unobserved. In the returns to education example presented by the authors, $d \in \mathcal{D} = \{0, 1, \ldots, \bar{d}\}$ is the level of schooling the individual selects, $Y \equiv Y_d$ is the potential earning outcome and the rank variable is determined by factors which include ability. Given the CDF of Y given X, the conditional quantile function is defined as $Q_Y(\tau|d, x) = \inf\{y : F(y|d, x) \geq \tau\}$ for $\tau \in (0, 1)$. The QTE is obtained by taking the difference $Q_Y(\tau|d_1, x) - Q_Y(\tau|d_0, x)$. In the classic linear QR setting where $Q_Y(\cdot)$ depends only on explanatory variables absent of endogeneity, we have $Q_Y(\tau|x) = X'\beta(\tau)$, typically written in terms of a function $q(\tau, d, x) \equiv Q_Y(\tau|d, x)$.

Chernozhukov and Hansen (2008) make the important distinction between the conditional quantile function $Q_Y(\tau|d, x)$ and the structural quantile function (SQF): $S_Y(\tau|d, x) = d'\alpha(\tau) + x'\beta(\tau)$. The latter involves fixing $D = d$ in the quantile function of the latent outcome variable[3] and sampling the disturbance term U, conditional on X. Since this is generally not feasible in practice as U is typically unobserved (e.g. when the rank variable represents proneness, latent ability, etc.), the second equation defining D in terms of V be-

[3] $Y_d = d'\alpha(U) + x'\beta(U)$

comes necessary. Identification can be obtained under a set of regularity conditions[4] through the conditional moment equation:

$$P[Y \leq S_Y(\tau|D,X) \mid Z,X] = \tau \tag{4}$$

for any $\tau \in (0,1)$. The resulting instrumental variable quantile regression (or alternatively, the"inverse" QR) parameter estimates $\hat{\theta}(\tau) = \big(\hat{\alpha}(\tau), \hat{\beta}(\tau)\big)$ solve the moment equation above. This differs from the ordinary QR estimating equation $P[Y \leq Q_Y(\tau|D,X) = \tau]$ introduced by Koenker and Bassett (1978). To see this, consider a class of measurable functions of (D,X) denoted by $\mathcal{F}_{D,X}$ and write the τ-th conditional quantile of Y in terms of (D,X) as:

$$Q_Y(\tau|W) = \underset{f \in \mathcal{F}_W}{\operatorname{argmin}} E[\rho_\tau(Y - f(D,X))]$$

where $\rho_\tau(u) = (\tau - 1(u<0))u$ is the asymmetric least absolute deviation loss function[5]. The moment equation in ordinary QR entails finding the best predictor of Y given W under $\rho_\tau(\cdot)$, while the case involving endogeneity in the IVQR model defined on $S_Y(\tau|W) \equiv S_Y(\tau|D,X)$ is not as straightforward. Given a class of measurable function of (Z,X) denoted by $\mathcal{F}_{(Z,X)}$, we have:

$$0 = \underset{f \in \mathcal{F}_{(Z,X)}}{\operatorname{argmin}} E[\rho_\tau(Y - S_Y(\tau|D,X) - f(Z,X)]$$

In this setting, the problem now involves finding an $S_Y(\tau|D,X)$ such that 0 is the solution to the QR of $Y - S_Y(\tau|D,X)$ on (Z,X). In settings where $f(\cdot)$ is assumed to be parametric[6], this is equivalent to finding the value of $\alpha(\tau)$ that drives the coefficient of the instrument as close to zero as possible and selecting the corresponding $\beta(\alpha(\tau),\tau)$ value. Thus, finding the parameters $\theta(\tau) = (\alpha(\tau), \beta(\tau))$ that solve the equation above can be viewed as the inverse

[4]Originally outlined in Chernozhukov and Hansen (2005) and included in the Appendix.
[5]Also referred to as the check function or the pinball loss function.
[6]Which is typically the case in applications.

of the ordinary QR problem. As Chernozhukov and Hansen (2006) note, the term "inverse" also applies to the nature of ill-posed problems addressed by Tikhonov and Arsenin (1977) in addition to this relation to ordinary QR. For the structural equation taking on the form $S_Y(\tau|D,X) = D'\alpha(\tau) + X'\beta(\tau)$ and under suitable assumptions, identification of $\theta(\tau)$ arises from the unconditional moment condition given by:

$$E[m_i(\theta)] = E\big[\big(\tau - \mathbf{1}(Y_i \leq D'\alpha(\tau) + X'\beta(\tau))\big)\Psi(X,Z)\big] = 0 \tag{5}$$

where $\Psi(X,Z)$ is a transformation of instruments. Rather than basing estimation on this moment condition directly, the inverse quantile regression estimator $\hat{\theta}_{IVQR}(\tau) = (\hat{\alpha}(\tau), \hat{\beta}(\tau))$ is obtained using the weighted QR objective function $Q_n(\tau,\alpha,\beta,\gamma) = \frac{1}{n}\sum_{t=1}^{n} \rho_\tau(Y_i - S_Y(\tau|D,X) - f(Z,X))V_i$ as follows:

$$\big(\hat{\beta}_\tau(\alpha), \hat{\gamma}_\tau(\alpha)\big) \equiv \underset{\big((\beta,\gamma)\in\mathcal{B}\times\mathcal{G}\big)}{\arg\inf} \frac{1}{n}\sum_{t=1}^{n} \rho_\tau(Y_i - D_i'\alpha - X_i'\beta - \phi(X_i,Z_i)'\gamma)V_i \tag{6}$$

$$\hat{\alpha}(\tau) \equiv \underset{a\in\mathcal{A}}{\arg\inf}\, \hat{\gamma_\tau}(\alpha)' A \hat{\gamma_\tau}(\alpha) \tag{7}$$

where V_i is a positive weight matrix and it is assumed that $dim(\Psi_i) \geq dim(\alpha)$. In practice, one recommended starting point is letting $\Psi_i(X_i,Z_i)$ take on the fitted values obtained by regressing D_i on X_i and Z_i and setting $V_i = 1$. The linear IVQR estimation procedure is as follows:

Linear IVQR Estimation Procedure:

1. Define a grid of values $\{\alpha_j, j = 1,\ldots,J\}$
2. Run the ordinary τ-QR of $Y - D_i'\alpha_j$ on X and Z to obtain $\hat{\beta}(\alpha_j,\tau)$ and $\hat{\gamma}(\alpha_j,\tau)$
3. Choose $\hat{\alpha}(\tau)$ from grid that makes $\|\hat{\gamma}(\alpha_j,\tau)\|$ closest to zero
4. Select corresponding value $\hat{\beta}(\tau) = \hat{\beta}(\hat{\alpha}(\tau),\tau)$

Due to the exhaustive nature of the grid search method used to obtain $\alpha(\tau)$, it is generally computationally infeasible to implement this procedure in settings involving endogenous treatment variables with large dimensions. However in most applications (involving $dim(\alpha) \leq 2$, for example), this approach is an effective way of handling the non-convex optimization problem that involves potentially many local minima.

2.2 Quantile regression using neural networks

In one of the earliest significant contributions to the nonparametric QR literature, White (1992) applies artificial neural networks (aNN) to obtain consistent conditional quantile estimators. Based on a sieve estimation approach where the complexity of the aNN grows as the sample size does, consistency of the nonparametric estimator is established. The most simple form of feedforward aNN, the single hidden layer feedforward network (SLFN) model is considered. White (1990) shows these networks can obtain nonparametric estimators for any conditional expectation function as long as it is square integrable. Furthermore, a SLFN with a smooth activation function is capable of approximating any function and its derivatives arbitrarily well, as shown by Hornik et al. (1989). In other words, these feedforward networks possess unlimited expressiveness without requiring specification a priori as a result of the universal approximation theorem.

The quantile τ of outcome Y given features X is defined in terms of the conditional quantile function $Q_Y : \mathbb{R}^r \to \mathbb{R}$ such that $P[Y_i \leq Q_Y(\tau|x)] = \tau$, as discussed earlier. For the SLFN used to approximate $Q_Y(\cdot)$, the output is given by: $g^r(X_i, \omega^r) = v_0 + \sum_{j=1}^{r} v_j \psi(X_i' w_j)$, where $\psi : \mathbb{R} \to \mathbb{R}$ is the selected nonlinear activation function, $r \in \mathbb{N}$ is the total number of hidden units and $\omega^r \equiv \omega^r(\tau)$ is a vector of network parameters fitted for each value of $\tau \in (0, 1)$ evaluated; where $\omega^r = (v', w')$. In the literature on aNN sieve estimators, ω^r is also referred to as the vector of network "connection strengths".

For illustrative purposes, we consider an (exogenous) treatment variable D and a modified response variable $\tilde{Y}_i \coloneqq Y_i - D_i'\alpha$, noting that the following objective function in White (1992)

corresponds to $\tilde{Y}_i = Y_i$. The nonparametric analog of the objective function from Koenker and Bassett (1978) can be written directly in terms of connectionist sieves for a sample size n as follows:

$$\min_{\omega^{r_n} \in \Omega_n} \frac{1}{n} \sum_{i:\tilde{Y}_i \geq g^{r_n}(\cdot)} \tau \left| \tilde{Y}_i - g^{r_n}(X_i, \omega^{r_n}) \right| + \sum_{i:\tilde{Y}_i < g^{r_n}(\cdot)} (1-\tau) \left| \tilde{Y}_i - g^{r_n}(X_i, \omega^{r_n}) \right| \tag{8}$$

$$\text{where } \Omega_n \equiv \left\{ \omega^{r_n} : \sum_{j=0}^{r_n} |v_j| \leq \Delta_n, \quad \sum_{j=1}^{r_n} \sum_{i=0}^{r} |w_{ji}| \leq \Delta_n \right\}$$

Here, one can apply grid search using $\{\alpha_j, j = 1, \ldots, J\}$ similar to the estimation approach of the linear IVQR model, iteratively fitting the aNN and selecting $\hat{\alpha}(\tau)$ appropriately. This approach benefits from the shallow nature of the aNN and avoids imposing (partial) linearity on the SLFN structure. For the sequence of connectionist sieves with a single hidden layer[7], the sequence of increasingly flexible networks is obtained by letting $r_n \to \infty$ and $\Delta_n \to \infty$. In White (1990), the consistency of connectionist sieve estimators where network complexity grows as $n \to \infty$ is shown through the use of cross-validation. In addition to establishing the asymptotic consistency of conditional quantile estimators SLFNs, White (1992) demonstrates how approximation based on pre-specified accuracy is sufficient to learn conditional quantiles; as opposed to exact network optimization.

In this general "QRNN" setup, Taylor (2000) introduced regularization parameters in order to control for overfitting by directly penalizing model complexity in the objective function. The optimal values for these parameters, as well as for the number of units in the hidden layer (r_n) are selected through the use of cross-validation. The presence of regularization parameters allows for the QRNN objective function to be written directly in terms of network weights, which helps simplify the notation. However (r_n, Δ_n) is nonetheless highly relevant to the class of SLFNs considered in this setting. Cannon (2011) proposes an implementation of the QRNN model involving a smooth approximation of the error/loss function and Cannon (2018) demonstrates how monotonicity contraints on the aNN weights

[7]Defined as $\Theta_n(\psi)$ in White (1992) and $\theta(x) = g^{r_n}(x, \omega^{r_n})$ when $\theta \in \Theta_n(\Psi)$.

can be used to obtain non-crossing quantiles.

2.3 Improved convergence rates for aNN sieves

Chen and White (1999) derive a class of SLFNs using the method of sieves that achieves root-mean-square convergence rates of $o_P(n^{-1/4})$. This is precisely the rate required for future extensions involving double/debiased machine learning (Chernozhukov et al., 2018), as well as many other semiparametric applications used to obtain asymptotic normal linear plug-in estimators. In the current literature, studies establishing convergence properties of neural networks tend to derive results based on nonasymptotic probability bounds, as seen with the deep ReLU networks setting of Yarotsky (2017, 2018).

Let $\underline{d}$ represent the dimension of the target function domain. The authors first define $\mathcal{B}_{\underline{d}}^m$ as the (weighted) Sobolev space of all functions on $\mathbb{R}^{\underline{d}}$ with continuous and uniformly bounded partial derivatives up to order m. The target function $f : \mathbb{R}^{\underline{d}} \to \mathbb{R}$ is assumed to have a Fourier representation such that it belongs to a class[8] of target functions $\mathcal{F}_{\underline{d}}^{m+1}$. This imposes a general smoothness condition[9] required by the class of aNN to be implemented in this setting, $\mathcal{G}_n \equiv G_{\underline{d}}^m(\psi, r_n, \Delta_n)$, defined as follows:

$$\mathcal{G}_n = \Bigg\{ g : g(x) = \sum_{j=1}^{r_n} v_j l(a_j)^{-m} \psi(a_j' x + w_j), \\ a_j \in \mathbb{R}^{\underline{d}},\ (w_j, v_j) \in \mathbb{R},\ \sum_{j=1}^{r_n} |v_j| \leq \Delta_n \Bigg\} \tag{9}$$

Where $\psi \in \mathcal{B}_1^m$ is an appropriately selected activation function that is k-finite for some $k \geq m$, $\omega_j \equiv (a_j, w_j, v_j)$ is the vector of network weights and r is the number of units in the hidden layer as in the QRNN model setting. Given a weight vector a_j, the function $l(a_j) \equiv max\{(a_j' a_j)^{\frac{1}{2}}, 1\}$ performs normalization on weights applied in the input layer and is applied to ensure that the target function is sufficiently smooth.

[8] See the Appendix.

[9] Ensuring the Fourier transform has a bounded first moment.

In order to achieve the desired approximation rate, the number of hidden units are set to increase with the sample size on the order of $O(n)$ (denoted here by r_n) and a Hölder condition is imposed on ψ. This permits the use of activation functions that are not necessarily sigmoidal, which was the standard in the existing aNN sieve literature at the time.

3 IVQRNN

We are interested in the value $\theta_0 = \Upsilon(P_0)$ belonging to the parameter space $\Theta \equiv A \times \mathcal{F}_d^{m+1}$, where A is a non-empty, compact subset of the finite-dimensional space of the linear QTE parameter. In the IVQRNN model we have:

$$\theta_0 = \big(\alpha_0(\tau), g_0(\tau)\big) \text{ and } h(D, X, U) = D'\alpha(U) + g(X, U)$$

where $g(\cdot)$ is a nonlinear control function[10] of X and θ_0 can be interpreted in terms of maximizing the expected quasi-likelihood function $l(Y_i, D_i, X_i, Z_i, \theta)$. For $\hat{\theta}_n \in A \times \mathcal{G}_n \equiv \Theta_n$, we define $\pi_n\theta_0 \equiv \inf_{\theta\in\Theta_n} ||\theta - \theta_0||$ to characterize the approximation error of the sieve-based estimator. By Corollary 2.2 in Chen and White (1999), we have $\pi_n\theta_0 = (\alpha_0, \pi_n g_0)$ where $\pi_n g_0 \in \mathcal{G}_n$, which establishes the direct link between the parameter spaces Θ and Θ_n.

As for the requirement under $\mathcal{G}_n$ that the number of hidden units grows on the order $O(n)$, define $r_n = k_n \times n$. Here, $n \equiv n^{train}$ is the number of observations in the training dataset used to fit the model and $k_n \in (0, 1]$ is a scaling constant that can be selected using cross-validation. The significance of r_n when it comes to SLFN structures, as highlighted by the superscripts in the exposition on QRNN models earlier, cannot be overstated. In general, shallow neural networks are more sensitive to the choice of this hyperparameter relative to deep aNN. Restricting the flexibility of the SLFN by selecting an insufficient number of hidden nodes for a given training sample size may lead to overfitting that cannot be addressed effectively by regularization methods. We also define a unique activation function that

[10]Replacing $X'\beta(\cdot)$ in the structural relationship outlined earlier in the context of linear IVQR.

satisfies the requirements outlined in Chen and White (1999), which serves as a starting point for future research that seeks to bridge the gap between aNN sieve extremum estimators and the evolving econometrics literature on deep ReLU networks that is grounded in asymptotic theory. This choice of activation function to be applied in the aNN sieve structure must satisfy the following Hölder condition:

Assumption H [Chen and White (1999)]: For the activation function $\psi \in \mathcal{B}_1^m$ and weights $a_j \in \mathcal{R}^{\underline{d}},\ w_j \in \mathcal{R}$, there exists an $\xi \in (0,1]$ such that

$$||\psi_{a_j,w_j} - \psi_{a_{j,1},w_{j,1}}||_{\mathcal{B}_1^m} \leq c \times \left[\left((a_j - a_{j,1})'(a_j - a_{j,1})\right)^{\frac{1}{2}} + |w_j - w_{j,1}|\right]^{\xi} \tag{10}$$

for some constant $c > 0$. This condition ensures that the aNN approximation error under activation function $\psi \in \mathcal{B}_1^m$ converges to zero as $r \to \infty$.

3.1 IVQRNN structure

We define the quasi-likelihood function for the IVQRNN model as follows:

$$l(Y_i, D_i, X_i, Z_i, \theta) = -\Big|Y_i - S_Y(\tau|D_i, X_i) - f(Z_i, X_i)\Big| \times \tag{11}$$
$$\Big[\tau \mathbf{1}\big(Y_i \geq S_Y(\tau|D_i, X_i) + f(Z_i, X_i)\big) + (1-\tau)\mathbf{1}\big(Y_i < S_Y(\tau|D_i, X_i) + f(Z_i, X_i)\big)\Big]$$

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, D_i, X_i, Z_i, \theta) \tag{12}$$

where $S_Y(\tau|D_i, X_i) = D_i'\alpha(\tau) + g(X_i, \omega(\tau))$, $f(Z_i, X_i) = \phi(X_i, Z_i)'\gamma(\tau)$ and $L_n(\theta)$ is referred to as the training objective function. Without loss of generality, let $\phi(X_i, Z_i) \equiv \phi_i$ take on the value of a first-stage regression of D_i on Z_i and X_i rather than the instrument Z_i itself. We note that the IVQR moment condition can also be used to formulate a GMM objective function, such as in Chernozhukov and Hong (2003) where $L_n(\theta) = -\frac{1}{2}\big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta)\big)' W_n(\theta)\big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_i(\theta)\big)$ and $W_n(\theta)$ is a suitable positive definite weight-

ing matrix[11] for achieving efficiency. Implementation of the IVQRNN model based on this GMM approach is left to future research.

The expression $\omega(\tau)$ highlights the fact that we obtain a new set of network weights permitted by $A \times \mathcal{G}_n \equiv \Theta_n$ for each quantile evaluated. By saving and storing these weights across each iteration of τ, we are able to employ a variety of inference-based applications without incurring the additional computational costs of retraining the SLFN. In the context of quasi-maximum likelihood estimation, the sieve extremum estimator $\hat{\theta}_n$ involves finding the set of weights under Θ_n that maximize L_n.

Following Chernozhukov and Hansen (2008), define $\epsilon_i(\tau) \equiv Y_i - D_i'\alpha(\tau) - g(X_i, \omega(\tau))$ and $V^*(\tau) = f_{\epsilon(\tau)}(0|X, Z)$, the latter of which is an alternative weight function to be applied in simulations[12]. We assume the following boundedness condition on the conditional density of Y so that the SLFN $\hat{g}_n \in \mathcal{G}_n$ satisfies Corollary 3.2 from Chen and White (1999), given a suitable $\psi \in \mathcal{B}_1^m$ and r_n.

IVQRNN model assumptions:

R1: $\{Y_i, D_i, X_i, Z_i\}$ are iid or m-dependent ($m \in \mathbf{N}$), defined on the probability space (Ω, F, P_0) with compact support.

R2: Density $f_Y(Y|X, D, Z)$ is bounded by a constant $\bar{f}$ a.s.

R3(a): For the given τ, $\theta_0 \equiv (\alpha_0(\tau), g_0(\cdot; \omega(\tau)))$ belongs to the interior of the parameter space $\Theta \equiv A \times \mathcal{F}_d^{m+1}$ equipped with the pseudometric $||\cdot||$ and $\hat{\theta}_n \in \Theta_n \equiv A \times \mathcal{G}_n$, where A is a compact subset of a finite-dimensional parameter space with a non-empty interior and $\omega(\tau)$ is the set of network weights permitted by Θ_n.

R3(b): For all $\theta \in \Theta$, $E[l(Y_i, D_i, X_i, Z_i, \theta_0)] \geq E[l(Y_i, D_i, X_i, Z_i, \theta)]$

R3(c): There exists a measurable $\hat{\theta}_n$ such that for all $\theta \in \Theta_n$,

$$L_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta_n} L_n(\theta) - O(\epsilon_n^{2\xi}), \quad \epsilon_n \to 0 \text{ as } n \to \infty$$

[11] And $m_i(\theta) = [\tau - \mathbf{1}(Y_i \leq D_i'\alpha(\tau) + g(X_i, \omega(\tau))]\Psi(Z_i, X_i)$ for the nonparametric estimation of the covariates in the IVQRNN setting, where $\Psi(Z_i, X_i) = (\phi(X_i, Z_i)', X_i')'$.

[12] In addition to the baseline case involving $V(\tau) = 1$.

In order to characterize the rank condition of the IVQRNN model, we make use of the pathwise derivative outlined in Newey (1994) and Chen et al. (2003). Given a sequence δ_n such that $||\hat{\theta}_n - \theta_0|| = o_P(\delta_n)$, let $\Theta_{\delta_n} \equiv \{\theta \in \Theta_n : ||\theta - \theta_0|| \leq \delta_n\}$. For any $\alpha \in \Theta_{\delta_n}$, $l(Y_i, W_i, \theta)$ is pathwise differentiable at $g \in \Theta_{\delta_n}$ in the direction $[\bar{g} - g]$ if $\{g + \iota(\bar{g} - g) : \iota \in [0, 1]\} \subset \mathcal{G}_n$ and $\lim_{\iota \to 0}[l(\alpha, h(\cdot, \alpha) + \iota(\bar{h}(\cdot, \alpha) - h(\cdot, \alpha))) - l(\alpha, h(\cdot, \alpha))]/\iota$ exists. This limit represents the pathwise derivative and is denoted by $\Gamma_2(\alpha, g)[\bar{g} - g]$.

IVQRNN model assumptions (continued):

R4(a): For all $\alpha \in A$, the matrix

$$\Gamma_1(\theta) = E[\nabla_\alpha l(Y_i, D_i, X_i, Z_i, \theta)] = E[f_{\epsilon_i(\tau)}(0|X_i, Z_i)\Psi(Z_i, X_i)]$$

exists, is continuous and is of full rank at θ_0.

R4(b): $\mathcal{I}(\theta) = -E\big[\frac{\partial^2 l(Y_i, D_i, X_i, Z_i, \theta)}{\partial \alpha \partial \alpha'}\big]$ exists, is nonsingular and is continuous at θ_0

R5: The pathwise derivative

$$\Gamma_2(\alpha, g)[\bar{g} - g] = E[f_Y(D_i'\alpha + g(X_i) + \phi_i'\gamma\big|X) log f_Y(D_i'\alpha + g(X_i) + \phi_i'\gamma\big|X)(\bar{g}(X_i) - g(X_i))]$$

exists in all directions of $[g - g_0] \in \Theta$ and for all $\theta \in \Theta_{\delta_n}$

R6: The image of Θ_n under $(\alpha, g) \mapsto E[\{\tau - \mathbf{1}(Y < D'\alpha + g(X)\}V]$ is simply connected.

R7: Activation function $\psi \in \mathcal{B}_1^m$ is k-finite for some $k \geq m$ and satisfies a Hölder condition (**Assumption H**) for smoothness parameter $\xi = 1$.

We now consider the smooth activation functions required for our SLFN to deliver $o_P(n^{-1/4})$ convergence rates. Due to advances in aNN activation functions over the past two decades, B-spline activation functions have fallen out of use in general network structures. This leads us to consider newer activation functions that are both suitable for this modeling environment and leverage strengths of commonly used activations. As our baseline activation function, we consider a B-spline activation function that satisfies *Assumption H*

from Chen and White (1999) for QR applications[13]. For knots $(t_i, t_{i+1}, \ldots t_{i+b+1})$ and letting b denote the B-spline degree, we have:

$$\psi_{i,b}(x) = \frac{x - t_i}{t_{i+b} - t_i}\psi_{i,b-1}(x) + \frac{t_i - x}{t_{i+b+1} - t_{i+1}}\psi_{i+1,b-1}(x)$$

Similar to the various other network hyperparameters, the number of knots can be selected through cross-validation or tuned using a validation dataset obtained by splitting the training data. We also consider two alternative smooth activation functions, the softplus function and an original hybrid between two commonly used activation functions, before verifying that they satisfy *Assumption H*. The former, which we define below as $\psi_S(x)$, is a smooth approximation[14] to the rectified linear unit (ReLU) activation function, while the latter $\psi_H(x)$ combines ReLU with the sigmoidal activation function:

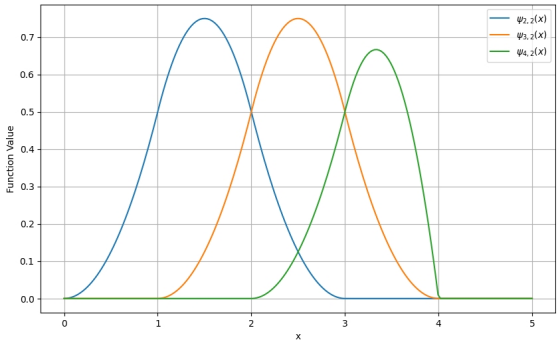


Figure 1: B-Spine Activation (Degree 2)

$$\psi_S(x) = \frac{1}{s}[log(1 + e^{sx})]$$

$$\psi_H(x) = x\sigma(x) = \frac{x}{1 + e^{-x}}$$

The ReLU activation function performs especially well in semiparametric aNN models[15] while application of the sigmoid function ensures smoothness. By leveraging the strengths of these two types of activation functions, $\psi_H(x)$ possesses some robustness to both the vanishing gradient and exploding gradient problems[16]. The smoothness provided by the sigmoid function in general improves computational efficiency while the ReLU activation

[13]See Example 3.5 in Chen and White (1999).
[14]Where for most applications including this one, $s = 1$.
[15]e.g. Zhong and Wang (2024), Farrell et al. (2021).
[16]see Goodfellow et al. (2016)

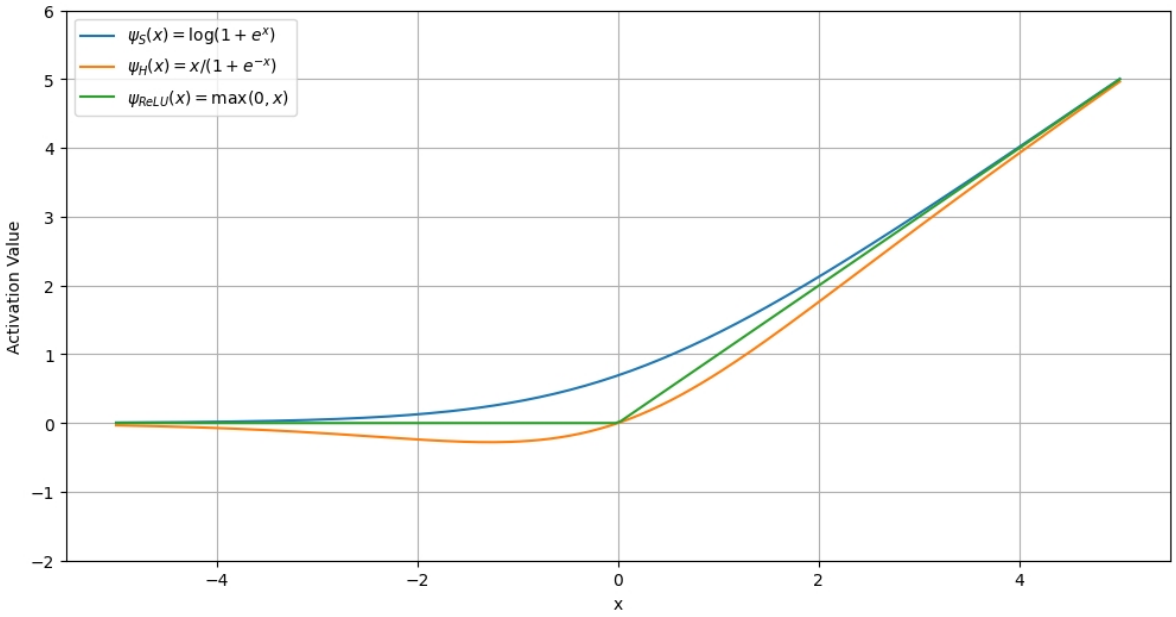


Figure 2: Softplus and Hybrid Activations

function does not suffer from vanishing gradients when the input is positive. We note that in the absence of thresholds restricting the domain of inputs (x), all three activation functions are especially volatile for larger inputs; highlighting the importance of data standardization and the use of suitable weight initialization techniques.

Theorem 1: For $\xi = 1$, the activation functions $\psi_S \in \mathcal{B}_1^m$ and $\psi_H \in \mathcal{B}_1^m$ with the Hilbert space metric $L_2(\cdot)$ satisfy **Assumption H**.

See the Appendix for the proof of Theorem 1. Although these activation function allow for m-dependent processes, we only consider the i.i.d case in order to simplify the analysis. The flexibility provided by the ReLU-based activation function is relevant to original research that seeks to extend the method of sieves approach to deep aNN (i.e., sieve networks with multiple hidden layers) architectures. Such an extension requires the application of empirical process theory in a manner similar to Bartlett et al. (2019) for deep aNN structures consistent with the sieve estimation approach[17] and is left to future work. An alternative activation function based on the Morlet wavelet with compact support that satisfies the same requirements (see

[17]Namely, with sufficiently smooth activation functions.

the Appendix) was also considered, however this is left out of the analysis due to unstable results under the larger sample sizes evaluated in the simulations.

We obtain the result for $\hat{g}_n^{NN}(\cdot, \tau) \in \mathcal{G}_n$ using $l(Y_i, D_i, X_i, Z_i, \theta)$, which follows from *Example 3.5* in Chen and White (1999):

$$\begin{aligned}&\left[\int \left[\hat{g}_n(x, \omega(\tau)) - g_0(x, \omega(\tau))\right]^2 d\mu(x)\right]^{1/2} \\ &\quad = O_P\left([n/log(n)]^{-(1+(2/(d+1)))/[4(1+(1/(d+1)))]}\right) = o_P(n^{-1/4}) \end{aligned} \tag{13}$$

3.2 Root-n asymptotic normal QTE estimator

We also seek to perform inference on the resulting linear QTE estimator $\hat{\alpha}_{NN}(\tau)$. The resulting Wald statistic provides a method to directly compare $\hat{\alpha}_{IVQR}(\tau)$ against $\hat{\alpha}_{NN}(\tau)$ and derive their coverage regions. For a known functional $\Gamma : \Theta^{NN} \to \mathbb{R}$ where $\Gamma_0 = \Gamma(\theta_0^{NN})$ and the corresponding plug-in estimator is $\Gamma(\hat{\theta}_0^{NN})$. Assumption 3 follows from Chen and White (1999), where $\hat{\theta}^{NN} \in A \times \mathcal{G}_n$ as before. These assumptions involve imposing a smoothness condition on Γ around Γ_0 in order to achieve the root-n asymptotic normality result we seek.

Assumption 3.1: For any $\theta \in \Theta$:

$$|\Gamma(\theta) - \Gamma(\theta_0) - \Gamma'_{\theta_0}[\theta - \theta_0]| \leq O\big(\|\theta - \theta_0\|^2\big), \quad \text{as } \|\theta - \theta_0\| \to 0$$

Assumption 3.1 ensures that $\Gamma(\cdot)$ is well-behaved locally and is able to be approximated using a first-order Taylor expansion. This is of significance with respect to the implementation of alternative methods that are less exhaustive than grid search to estimate $\alpha(\tau)$, due to the non-convex nature of the optimization problem. Assumption 3.2 ensures that the variance of the linear plug-in estimator is well-defined and bounded, while Assumption 3.3 imposes regularity conditions on the remainder term. All three of these assumptions allow for small perturbations in the neighborhood of the true parameter θ_0 and can be weak-

ened by utilizing Neyman orthogonal moment conditions[18] introduced by Chernozhukov et al. (2018). This task, which is also relevant to extending aNN sieve estimators to deep networks, is left to future research.

Assumption 3.2: The variance term σ_*^2 is positive and finite, where

$$\sigma_*^2 \equiv \lim_{n\to\infty} \frac{1}{n} Var\left(\sum_{t=1}^{n} l'_{\theta_0}[v^*, W_i] \right)$$

Assumption 3.3: $\kappa(\theta, W_i)$ (defined in the Appendix) satisfies Assumption A.1 $\forall \theta \in \{\theta \in \Theta_n : \|\theta - \theta_0\| \leq \delta_n\}$, where $\|\theta - \theta_0\| = o_p(\delta_n)$.

In the following, $\psi_i(\cdot)$ corresponds to any activation function satisfying *Assumption H*, including the three discussed earlier. We note that *Theorem 2* can be proven by an argument identical to a key result (Theorem 2) from Chen and Shen (1998).

Theorem 2: Given $\psi_i(\cdot)$, Assumption 1 and Assumption 3.1-3.3, we have: $n^{1/2}(\Gamma(\hat{\theta}_n) - \Gamma(\theta_0) \to N(0, \sigma_*^2)$

By **Theorem 2**, we obtain the result:

$$n^{1/2}(\hat{\alpha}_{NN} - \alpha_{NN}) \to N(0, \Gamma_1^{-1}\Omega\Gamma_1^{-1}) \tag{14}$$

where $\Gamma_1 \equiv \Gamma_1(\theta_0)$ and $\Omega \equiv \tau(1-\tau)E[\Psi(Z,X)\Psi(Z,X)']$ is a positive semi-definite matrix.

The corresponding Wald statistic (also evaluated for $\hat{\alpha}_{IVQR}$, which is asymptotic normal) is computed for a given $\tau \in (0,1)$ as follows:

$$W_n(\alpha) = \frac{(\hat{\alpha}_{NN} - \alpha_0)^2}{\hat{Var}(\hat{\alpha}^{NN})}, \quad W \xrightarrow{d} \chi^2(1) \tag{15}$$

We also construct a Wald statistic for directly testing $\gamma(\alpha(\tau), \tau) = 0$ based on the inference procedure from CH (2008). This serves as a method for conducting inference on

[18] Which are insensitive to perturbations in the nuisance parameters/functions, in this case $g(\cdot)$.

$\alpha(\tau)$ itself in the original linear model (addressed in the discussion on the contrast between $S_Y(\tau|W)$ and $Q_Y(\tau|W)$ earlier) as one can construct a confidence region for this parameter of interest, which can be compared to the estimate produced by the IVQRNN model. In CH (2008), this Wald statistic is defined as:

$$W_n(\gamma(\alpha)) := n[\hat{\gamma}(\alpha,\tau)\hat{A}(\alpha)\hat{\gamma}(\alpha,\tau)] \tag{16}$$

$$\hat{\alpha}(\tau) = arg \inf_{\alpha\in\mathcal{A}}[W_n(\alpha)] \tag{17}$$

and is referred to as the "dual" inference approach, in contrast to the "direct" inference originally introduced in CH (2006).

3.3 SGD-based approach

Although grid search is an effective algorithm for solving a non-convex optimization problem with potentially many local minima, it less than optimal for most applications involving treatment variables with a dimension greater than two. Furthermore, in the context of IVQR when X is estimated by a nonparametric control function, the exhaustive nature of grid search is especially inefficient when the computational costs of using an estimation method such as aNN or sieve estimators in general.

To address this, we propose implementing stochastic gradient descent (SGD) with momentum in order to select the optimal alpha value[19]. This involves two different versions, one that makes use of a grid of alpha values in the same setup and only re-fits the SLFN for values in promising regions of the grid. Given a suitable momentum hyperparameter, the SGD algorithm is able to escape local minima which is of particular importance for this approach that is effectively a hybrid between grid search and SGD. The latter version is a two-level optimization approach, where alpha values are updated in the gradient descent as the SLFN is simultaneously being trained. Although both versions offer a significant im-

[19]That drives the instrument coefficient, γ, closest to zero as before.

provement over grid search for IVQRNN, the latter has a significantly reduced computation cost due to the fact $g(X, \omega(\tau))$ is only fit once. This comes at the cost of asymptotic normal linear QTE estimates, even if the SGD algorithm achieves a linear convergence rate, due to the complexity that arises from the two-level optimization procedure. Nonetheless, this offers a useful baseline to compare results of the IVQRNN model and for testing various (fixed) network hyperparameters at minimal computational cost.

For iterations $t \in \{1, \ldots, T\}$, the algorithm for SGD with momentum can be written in terms of two main hyperparameters, velocity (v_t) and momentum (η):

$$\alpha_{t+1} = \alpha_t + v_{t+1} = \alpha_t + [\eta v_t - \eta \nabla_\alpha L(\alpha_t)] \tag{18}$$

$$L(\alpha_t) = \|\gamma(\alpha_t, \tau)\|^2 \tag{19}$$

We impose several assumptions regarding this SGD with momentum method that helps to avoid fitting the aNN for values of the grid far from the global minimum. The most important of which is **Assumption PL**, a local Polyak-Lojasiewicz condition that imposes convexity only upon the neighborhood of the true value α_0 for a given quantile. This assumption is significantly weaker than a global PL condition while ensuring that the gradient of $L(\alpha)$ has sufficient curvature to deliver a linear convergence rate.

Assumption PL: There exists some $\mu > 0$ such that for all α in a neighborhood $\mathcal{N}(\alpha_0)$:

$$\frac{1}{2}\|\nabla_\alpha L(\alpha)\|^2 \geq \mu\big(L(\alpha) - L(\alpha_0)\big) \tag{20}$$

Under this assumption in addition to standard assumptions regarding the loss function $L(\alpha)$ and SGD hyperparameters (see **Assumption A.1-3** in the Appendix), we are able to guarantee global convergence of the procedure at a linear rate as long as the grid is sufficiently dense. The neighborhood satisfying **Assumption PL** can be determined using computational methods and we still achieve global convergence in cases where the condition fails to hold, albeit at a sublinear rate.

Theorem 3: Under **Assumption PL** and **Assumption A.1-3**, the grid-based SGD with momentum algorithm converges to α_0 at a linear rate:

$$E[L(\alpha_t) - L(\alpha_0)] \leq (1 - \eta\mu)^t [L(\alpha_{init}) - L(\alpha_0)] + \frac{\eta\sigma^2}{2\mu}$$

By nature of the non-convex optimization problem with potentially many local minima, more adaptive and/or forward-looking gradient descent methods (e.g. Nesterov's accelerated gradient) may provide an even larger improvement in terms of computation time/cost. This is left to future research, as we limit complexity of the SGD with momentum algorithm in order to more closely observe the model dynamics as it relates to the SLFN.

3.4 IVQRNN estimation

The estimation procedure for the grid-based IVQRNN model follows the linear version outlined earlier closely, only replacing $X_i'\beta(\tau)$ with the SLFN $g(X, \omega(\tau))$ and selecting the optimal set of network weights based on the $\alpha(\tau)$ that drives the instrument coefficient closest to zero (rather than $\beta(\alpha(\tau), \tau)$ previously). The procedure for the SGD with momentum approach that makes use of requires several modifications:

IVQRNN (SGD) Estimation Algorithm:

1. Define a grid of values $\{\alpha_j, j = 1, \ldots, J\}$ using the same initial guess $(\alpha_{init}(\tau))$ the IVQR grid is defined upon for a given τ.

2. Train SLFN for a given α_j, estimate $\gamma(\alpha_j, \tau)$ and use $\hat{g}(\cdot; \tau)$ to predict $Y - \alpha_j(\tau)D$

3. Momentum accelerates around regions with promising α candidates and for poor candidates, takes backward steps to prevent overshooting

4. Apply early stopping to speed up computation of clearly suboptimal candidates, sequentially evaluate values surrounding promising candidates until convergence

Similar to the SLFN, optimal hyperparameters for the SGD algorithm such as the learning rate can be determined using cross-validation. This requires a third partition of the data beyond the training and test datasets to create the validation data used in this process (without directly impacting model fitting and evaluation). However since we seek to obtain a linear convergence rate for the estimation procedure involving SGD, the learning rate and momentum hyperparameters are fixed within suitable ranges and only the SLFN training makes use of the validation set.

The quasi-likelihood IVQRNN objective function can be written more conveniently as below, recall the discussion about solving the semiparametric moment condition entailing we find the set of weights permitted by Θ_n that maximizes $L_n(\theta)$:

$$\max_{\alpha,\omega} L_n(\alpha, g(\cdot,\omega))V_i - \lambda R(\omega) \tag{21}$$

where $R(\omega)$ is a regularization penalty used to control for overfitting and λ is the corresponding penalty term. For example, in the case of L2 regularization we have $\lambda R(\omega) = \lambda\left(\frac{1}{2}\sum_{j=1}^{r_n}\|\omega_j\|_2^2\right)$ and when this penalty is directly implemented in the updating steps of the aNN fitting across iterations, it is commonly referred to as the "weight decay" method. As discussed earlier, avoiding the use of excessively large inputs in the B-spline activation function is essential and implementing even a minuscule penalty when updating weights is another way to help ensure this. For a full discussion on weight decay regularization, see Bishop (1995). In the context of sparse aNN sieve models, one may want to employ a method similar to the L1 regularized approach for QR from Belloni and Chernozhukov (2011). However, the main method used in the simulations and empirical example to help avoid overfitting the SLFN is early stopping, in order to avoid impacting convergence rates by applying L1/L2 regularization in the objective function.

4 Results

4.1 Monte Carlo Simulations

We define the data generating process (DGP) to be implemented in the Monte Carlo simulations in terms of binary treatment variable D and binary instrument Z, along with set of covariates X. Although the IVQRNN model setup allows for overidentification, we only consider the case where $dim(D) = dim(Z) = 1$ to simplify the analysis. In many settings involving treatment effect analysis, binary variables typically make up a large fraction of the covariates. However in order to highlight the performance of the models in the presence of nonlinearity in the DGP, we only consider continuous covariates generated under a normal distribution. The sequence of quantiles evaluated is given by $\tau \in [0.1, 0.2, \ldots, 0.9]$ and results are aggregated across 1000 total simulations. Due to increased stability under minimal regularization penalties in small scale simulations, $\phi_H(\cdot)$ is the activation function considered for the SLFN under grid search and the SGD approach used to obtain $\hat{\alpha}(\tau)$.

$$
\begin{aligned}
U \sim Unif(0,1), \quad V \sim N(0,1) \\
X_j \sim N(0,1), \quad Z \sim Bernoulli(0.7) \\
D = Z \times \mathbf{1}(U > 0.5V)
\end{aligned}
$$

Due to the nature of the threshold-based selection in $D = Z \times \mathbf{1}(U > 0.5V)$, one would expect relatively lower bias in general for the middle quantiles due to reduced noise. Since IVQRNN performs especially well when many of the covariates are binary, we only generate standard normal regressors to highlight efficiency gains arising from the model. We consider two main DGP equations, a linear and one with a component that is nonlinear in the covariates

through a sine transformation:

$$Y = \alpha_0(U)D + X'\beta_0 + \phi^{-1}(U) \tag{22}$$

$$Y = \alpha_0(U)D + X'\beta_0 + \sin(X) + \phi^{-1}(U) \tag{23}$$

For each sample, we generate $10,000$ observations and split the data according to a $0.7 : 0.2 : 0.1$ ratio to obtain our training, test and validation sets (respectively). As discussed earlier, the training set is used to fit the model while the excluded test set is used to evaluate the model performance when faced with unseen data. The validation set is used to determine the optimal network hyperparameters[20], in addition to applying early stopping by monitoring performance with respect to this data excluded in training. This helps determine the optimal point in the training process that balances between model accuracy and the ability of the aNN to generalize to new data, through an implicit form of model selection. As discussed earlier, we do not apply L1/L2 regularization in this setting. Rather, weight decay using an extremely small penalty (of size 1e-7) is applied to promote stability across simulations by preventing potential situations with uncharacteristically large inputs in the activation function. For simplicity, we refer to the grid-based semiparameteric model as IVQRNN, while

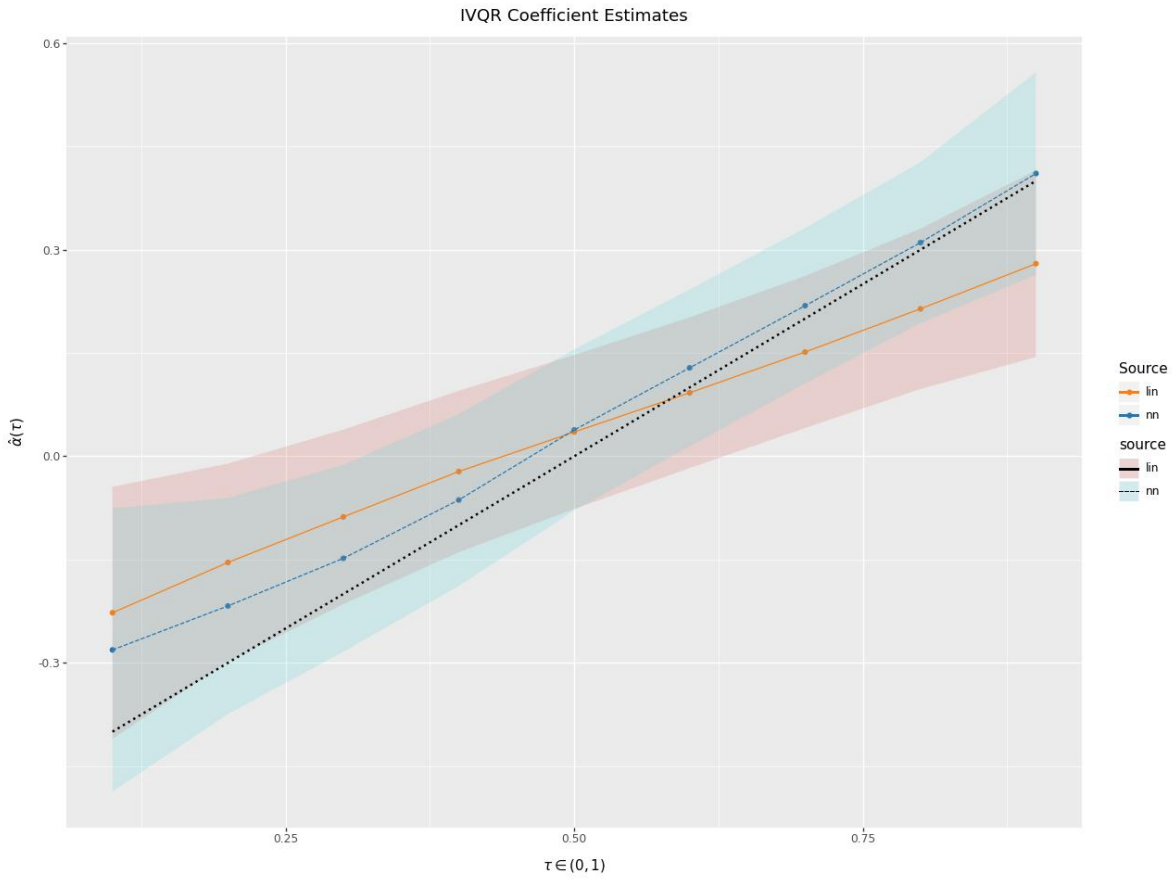
Table 1: Parameters for Simulations

Parameter	**Description**	**Simulation 1**	**Simulation 2**
n	Number of observations	10000	10000
α_0	DGP parameter 1	$\tau - 0.5$	$\tau - 0.5$
β_0	DGP parameter 2	$\in (0.2, 0.8)$	$\in (0.2, 0.8)$
K	Number of covariates	10	25

the version of the model that uses SGD with momentum in order to obtain the parameter of interest is referred to as IVQRNN-SGD henceforth. Discussion of the simulation results is centered around the nonlinear DGP in simulation #2, as each model performs similarly in the linear setting.

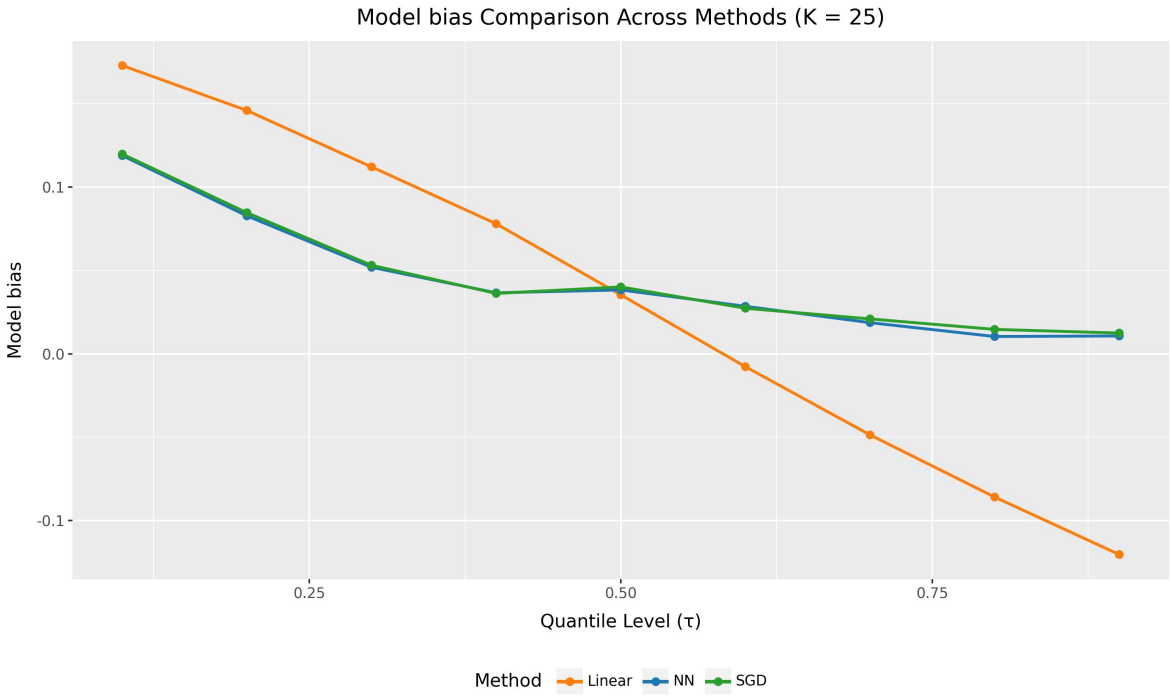
[20] A process that would otherwise require cross-validation when training the aNN.

Figure 3: Parameter estimate, linear vs. SLFN (Grid), sim #2 (Nonlinear)



Across both simulations, the true parameter value ($\alpha_0(\tau) = \tau - 0.5$) is contained within the confidence intervals (10% significance level) of the IVQRNN estimates under both simulation DGPs, which is not quite the case for the IVQR model. This can be seen in Figure 3, where the linear model confidence interval almost entirely fails to encompass the true parameter value for the first three quantiles. In addition to an overall poor performance relative to semiparametric methods, one would expect misspecified linear QR models to especially struggle at the extremes of the distribution in settings involving nonlinear dynamics/processes, which is the case here.

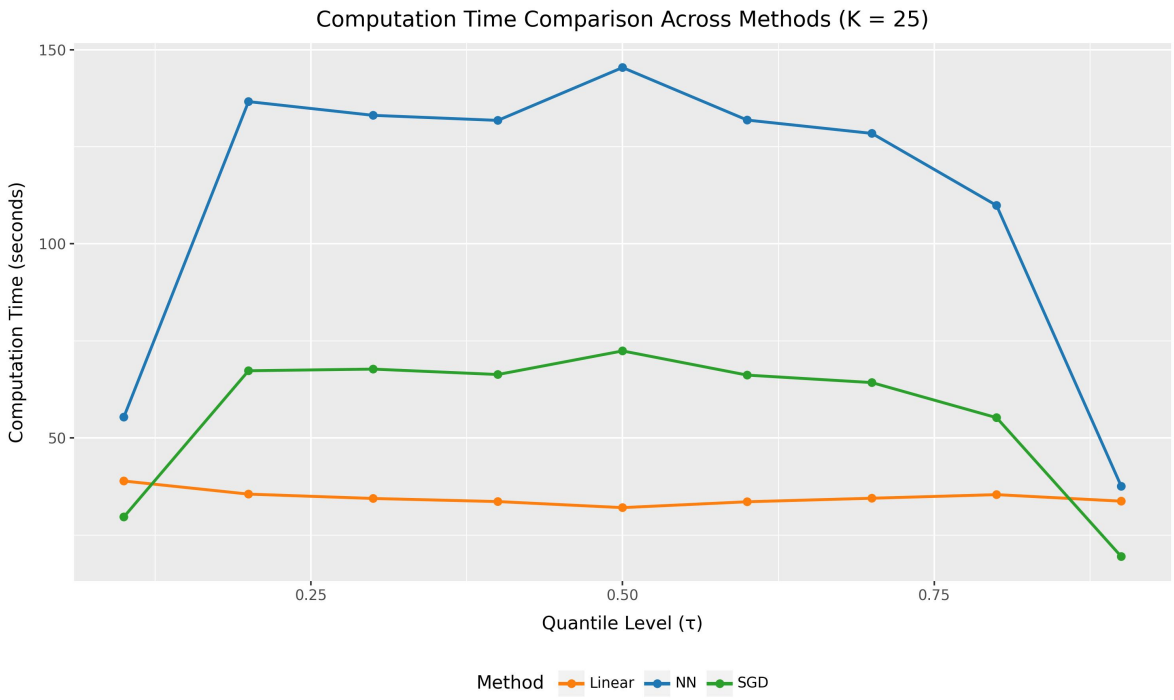
Figure 4: Bias, linear vs. SLFN (Grid, SGD) bias, sim #2 (Nonlinear)



The IVQRNN-SGD model produces estimates of the QTE that are identical to the grid method across the 1000 simulations, with a slight increase in bias and variance arising from the SGD approximation error. This is despite how the computation time decreases by over 50% on average when moving from the grid search method to the SGD approach, and in the case of simulations with $K = 25$ covariates, the latter runs faster than the linear IVQR model for the extreme quantiles. This is a very promising result not only in the context of grid search in semiparametric models[21] but for situations where potential aNN hyperparameters are required to be evaluated sequentially.

[21]Such as the DML-IVQR model from Chen et al. (2021).

Figure 5: Computation time, linear vs. SLFN (Grid, SGD), sim #2 (Nonlinear)



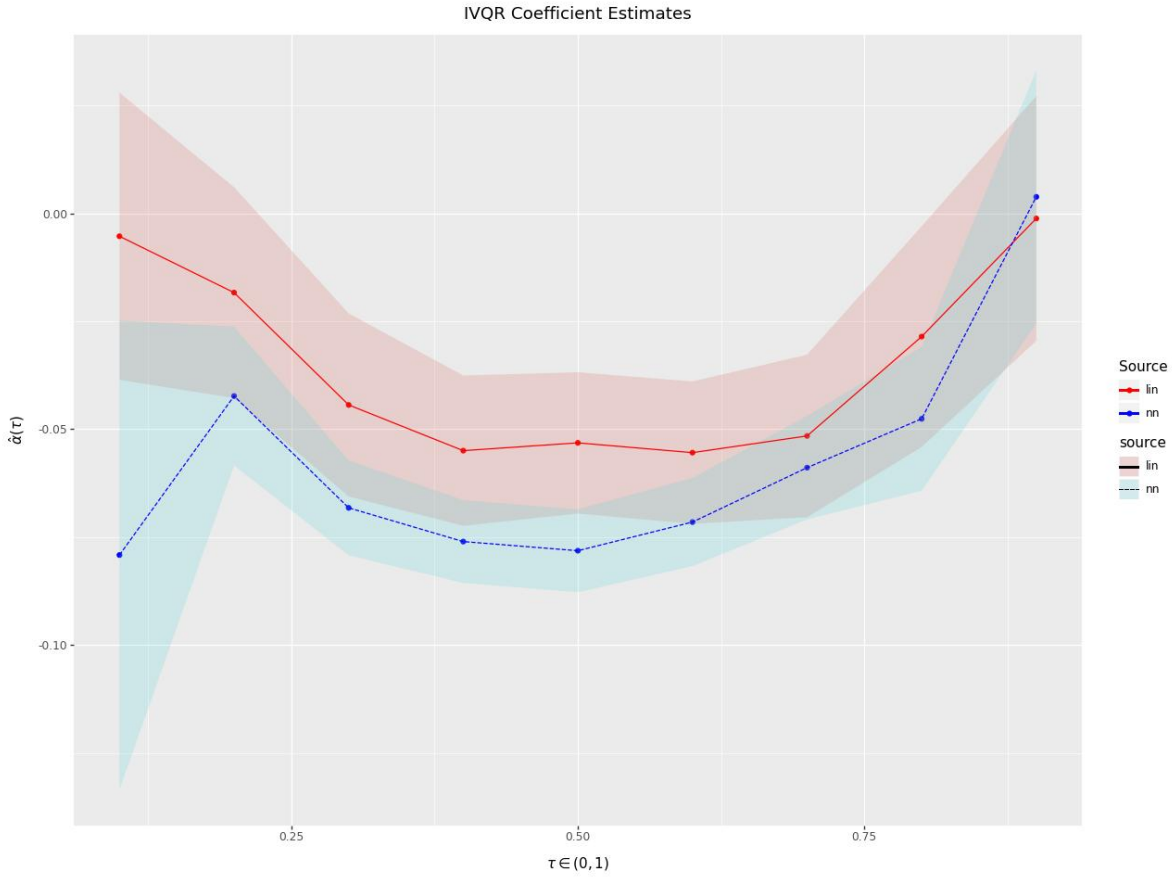
4.2 Empirical Application

Maynard and Qiu (2009) analyzed the impact of Medicaid on the savings of households belonging to various wealth groups. Using data from the Survey of Income and Program Participation (SIPP). This survey collects data from the same household every four months over a total of 24-32 months and the data spans from 1984-1993. In the (pooled) cross-section, there is a total of $52,706$ households and net worth is defined in terms of financial assets in addition to household equity net of debt. The instrument for Medicaid eligible dollars is given by a simulated measure that isolates the exogenous component of the treatment variable by conditioning on factors exogenous to savings decisions.

For this application, we implement a $0.8 : 0.1 : 0.1$ split of the original data and evaluate the same quantiles as Maynard and Qiu (2009), indexed by $\tau \in [0.1, 0.2, \ldots, 0.9]$. Although one could apply some form of cross-validation (e.g. splitting the sample using a $0.5 : 0.25 : 0.25$ ratio twice and averaging parameter estimates across both training datasets), we find

that the linear IVQR replicates the results from Maynard and Qiu (2009) sufficiently well due to the random sampling nature used to obtain the training dataset. Of the $K = 341$ covariates used in the model, the vast majority are binary variables and we end up with a total sample size of $n = 40442$ observations.

Figure 6: IVQRNN model, linear parameter estimate, MQ (2009)



We find that the estimates $\hat{\alpha}(\tau)$ are statistically significant at the 1% critical level for each τ except for the 90-th percentile. The insignificance of this QTE estimate at $\tau = 0.9$ is consistent with the findings of Maynard and Qiu (2009)[22] and the rationale is straightforward: those in the top of the household wealth distribution are the least likely to be impacted by policy changes regarding public insurance due to alternatives available to them. Furthermore, the stronger treatment effect observed for each decile excluding $\tau = 0.9$ can be attributed

[22] Albeit for the 5% critical level instead of the 1% level here.

to the aNN preventing contamination of the QTE in its role as a nonparametric control function.

Table 2: MQ (2009) IVQRNN Results

τ	$\hat{\alpha}_{NN}(\tau)$	$SE(\hat{\alpha}_{NN}(\tau))$	p_value	significant_1%
0.1	-0.0791	0.0211	0.0002	True
0.2	-0.0423	0.0062	0.0000	True
0.3	-0.0682	0.0043	0.0000	True
0.4	-0.0760	0.0037	0.0000	True
0.5	-0.0781	0.0037	0.0000	True
0.6	-0.0715	0.0040	0.0000	True
0.7	-0.0589	0.0047	0.0000	True
0.8	-0.0476	0.0065	0.0000	True
0.9	0.0038	0.0114	0.7368	False

The biggest implication of this is in the bottom two deciles, where the estimated QTE in Maynard and Qiu (2009) is insignificant, which is also the case for the linear IVQR model fitted using our training data. Using the IVQRNN model, we find a savings effect in the bottom decile of a magnitude similar to the middle deciles that is significant at the 1% level despite producing a larger standard error. If nonlinear dynamics surrounding saving decisions has a greater impact on households at the extremes of the wealth distribution (e.g. through eligibility probabilities) as one might expect, the aNN facilitates sharper identification of this treatment effect. However in the case of households in the bottom decile, we note that this effect likely has more to do with relatively high time preferences[23] as opposed to savings decisions influenced by asset testing, unlike the middle deciles.

5 Conclusion

In this paper, we demonstrated how a shallow aNN improves estimation of a linear treatment effect in a semiparametric IVQR model. By guaranteeing a sufficiently fast convergence rate under standard sieve estimation assumptions, we obtain a linear QTE that is asymptotic

[23] As observed in Lawrance (1991).

Table 3: MQ (2009) Linear IVQR Results

τ	$\hat{\alpha}(\tau)$	$SE(\hat{\alpha}(\tau))$	p_value	significant_1%
0.1	-0.0053	0.0129	0.6815	False
0.2	-0.0184	0.0095	0.0527	False
0.3	-0.0444	0.0082	0.0000	True
0.4	-0.0550	0.0068	0.0000	True
0.5	-0.0532	0.0064	0.0000	True
0.6	-0.0554	0.0064	0.0000	True
0.7	-0.0516	0.0073	0.0000	True
0.8	-0.0285	0.0099	0.0040	True
0.9	-0.0012	0.0110	0.9148	False

normal. We extend the inference procedure introduced in Chernozhukov and Hansen (2008) for the semiparametric framework of the IVQRNN model. Due to the computationally costly nature of a grid search approach to obtain the optimal QTE in this setting, we introduce a new SGD with momentum algorithm that guarantees global convergence at a linear rate. This is demonstrated in simulations where we observe a reduction in computation time by approximately 50% on average at a minimal cost in efficiency. This is a promising result not only in the context of econometric machine learning tools (e.g. DML) involving nuisance parameters that require grid search, but also for hyperparameter selection in aNN structures; in particular, with deep networks where one is restricted in the range of hyperparameter values they can evaluate due to significant computational cost.

One shortcoming of the proposed IVQRNN model relative to other nonparametric estimation methods (e.g. spline estimation) is the black-box nature of the aNN. However in settings with many covariates such as our empirical application, this is a (partial) interpretability tradeoff that a researcher is likely willing to accept in general as it improves estimation of the parameter of interest. Furthermore, the aNN architecture is grounded in a data-adaptive foundation in practice, where the key network parameters are tied to the sample size and dimension of the inputs. The ease of application is further enhanced by the two activation functions we propose (ψ_S and ψ_H), as these functions bypass the parameter selection required for the B-spline activation. In addition to the SGD with momentum ap-

proach we introduced, this offers an attractive and computationally lightweight alternative for researchers considering an IVQR model in empirical studies.

References

[1] Bartlett, P. L., Harvey, N., Liaw, C., & Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63), 1-17.

[2] Belloni, A., & Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39, 82–130.

[3] Cannon, A. J. (2011). Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers & Geosciences*, 37(9), 1277-1284.

[4] Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32(11), 3207-3225.

[5] Chen, J. E., Huang, C. H., & Tien, J. J. (2021). Debiased/double machine learning for instrumental variable quantile regressions. *Econometrics*, 9(2), 15.

[6] Chen, X., Linton, O., Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5), 1591-1608.

[7] Chen, X., & White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2), 682-691.

[8] Chen, Y. (2019). IVQR: Instrumental Variables Quantile Regression (Version 0.1.0). *GitHub*. Retrieved from https://github.com/yuchang0321/IVQR

[9] Chernozhukov, V., & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1), 245-261.

[10] Chernozhukov, V., & Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2), 491-525.

[11] Chernozhukov, V., & Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1), 379-398.

[12] Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181-213.

[13] Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366.

[14] Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5), 551-560.

[15] Lawrance, E. C. (1991). Poverty and the Rate of Time Preference: Evidence from Panel Data. *Journal of Political Economy*, 99(1), 54–77.

[16] Loken, C., Gruner, D., Groer, L., Peltier, R., Bunn, N., Craig, M., ... & Van Zon, R. (2010). SciNet: lessons learned from building a power-efficient top-20 system and data centre. In *Journal of Physics: Conference Series* 256(1), 12-26.

[17] Maynard, A., & Qiu, J. (2009). Public insurance and private savings: who is affected and by how much?. *Journal of Applied Econometrics*, 24(2), 282-308.

[18] Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 1349-1382.

[19] Ponce, M., Van Zon, R., Northrup, S., Gruner, D., Chen, J., Ertinaz, F., ... & Peltier, W. R. (2019). Deploying a top-100 supercomputer for large parallel workloads: The Niagara supercomputer. In *Practice and Experience in Advanced Research Computing 2019: Rise of the Machines (learning)*, 1-8.

[20] Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, 19(4), 299-311.

[21] White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3(5), 535-549.

[22] White H (1992). Nonparametric estimation of conditional quantiles using neural networks. In: Page C, R. LePage (Eds.). (1992) *Computing Science and Statistics*. Springer, 190–199. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press.

[23] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 103-114.

[24] Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on Learning Theory*, (639-649).

[25] Zhong, Q., & Wang, J. L. (2024). Neural networks for partially linear quantile regression. *Journal of Business & Economic Statistics*, 42(2), 603-614.

6 Appendix

6.1 Additional Definitions and Proofs

Assumption H [Chen and White (1999)] : For the activation function $\psi \in \mathcal{B}_1^m$ and weights $a_j \in \mathcal{R}^{\underline{d}},\ w_j \in \mathcal{R}$, there exists an $\xi \in (0, 1]$ such that

$$||\psi_{a_j,w_j} - \psi_{a_{j,1},w_{j,1}}||_{\mathcal{B}_1^m} \leq c \times \left[\left((a_j - a_{j,1})'(a_j - a_{j,1})\right)^{\frac{1}{2}} + |w_j - w_{j,1}|\right]^{\xi} \tag{24}$$

for some constant $c > 0$. This condition ensures that the aNN approximation error under activation function $\psi \in \mathcal{B}_1^m$ converges to zero as $r \to \infty$.

Softplus and Hybrid ReLU activation functions:

$$\psi_S(x) = \frac{1}{s}[log(1 + e^{sx})]$$
$$\psi_H(x) = x\sigma(x) = \frac{x}{1 + e^{-x}}$$

Theorem 1: For $\xi = 1$, the activation functions $\psi_S \in \mathcal{B}_1^m$ (where $s = 1$) and $\psi_H \in \mathcal{B}_1^m$ with the Hilbert space $L_2(\cdot)$ satisfy **Assumption H**

Proof:

Assume ψ_S is compactly supported and consider the first two derivatives of ψ_S:

$$\psi_S'(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \tag{25}$$
$$\psi_S''(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{e^x}{(1 + e^x)^2} \tag{26}$$

where $\psi_S'(x)$ is the sigmoid function $\sigma(x)$ defined earlier and is Lipschitz continuous for $L = \frac{1}{4}$ (i.e. the derivatives are bounded: $0 < \psi_S'(x) < 1$ and $0 < \psi_S''(x) < \frac{1}{4}$ under compact support). Similarly, the k-th derivative ($k \leq m$) of $\psi_S(x)$ is bounded: $|D^k\psi_S(x)| \leq c_k e^{-x}$ for some constant $c_k > 0$. Let $z_j(x) \equiv a_j'x + w_j$ and $z_{j,1}(x) \equiv a_{j,1}'x + w_{j,1}$. By the mean value theorem, there exists some intermediate point $z_{j,t} = tz_j(x) + (1 - t)z_{j,1}(x)$ for $t \in (0, 1)$ such

that we have the following (suppressing S for convenience):

$$|\psi(z_j(x)) - \psi(z_{j,1}(x))| = |\psi'(z_{j,t}(x))| \times |z_j(x) - z_{j,1}(x)| \tag{27}$$

$$|\psi(z_j(x)) - \psi(z_{j,1}(x))| \leq |z_j(x) - z_{j,1}(x)| \tag{28}$$

where the inequality in (5) follows from the boundedness of the first derivative ($\psi'_S(x) < 1$).

By definition of $z_j(x)$, we have: $|z_j(x) - z_{j,1}(x)| = |(a_j - a_{j,1})'x + (w_j - w_{j,1})|$. Applying the triangle inequality and Cauchy-Schwarz inequality (respectively), we obtain:

$$|z_j(x) - z_{j,1}(x)| \leq |(a_j - a_{j,1})'x| + |(w_j - w_{j,1})| \tag{29}$$

$$|(a_j - a_{j,1})'x| \leq ||(a_j - a_{j,1})|| \times ||x|| \tag{30}$$

By asssumption (R.1) of the IVQRNN model, x belongs to a bounded domain. Let $||x|| \leq c < \infty,\ \forall x \in \mathbf{X} \subset \mathcal{R}^d$. Then, we have:

$$|z_j(x) - z_{j,1}(x)| \leq c \times ||(a_j - a_{j,1})|| + |(w_j - w_{j,1})| \tag{31}$$

$$|z_j(x) - z_{j,1}(x)| \leq c \times \left[\left((a_j - a_{j,1})'(a_j - a_{j,1})\right)^{\frac{1}{2}}\right] + |(w_j - w_{j,1})| \tag{32}$$

Combining (27) and (31), the proof is complete. The proof for ψ_H follows from an identical argument.

SGD Algorithm:

Assumption A.1: Loss function $L(\alpha)$ has Lipschitz continuous gradients (with constant $\bar{L}$).

Assumption A.2: The variance of the gradient $\delta(\alpha)$ is bounded: $E\big[\|\delta(\alpha)-\nabla_\alpha L(\alpha)\|^2\big] \leq \sigma^2$, where $E[\delta(\alpha)] = \nabla_\alpha L(\alpha)$ and σ^2 decays as the number of iterations increases.

Assumption A.3: The momentum hyperparameter $\eta \in [0, 1)$ and learning rate (β) satisfy

$$\beta \leq \frac{1-\beta}{\bar{L}(1+\eta)}$$

Proof of **Theorem 3**:

Define a grid $\{\alpha_j, j = 1, \ldots, J\}$ with spacing $\Delta \leq \bar{r}/2$, where $\bar{r}$ is the radius of $\mathcal{N}(\alpha_0)$ defined in Assumption PL. Then there exists an $\alpha_k \in \{\alpha_j\}$ such that $\|\alpha_k - \alpha_0\| \leq \frac{\bar{r}}{2}$. By Assumption A.1, we have:

$$L(\alpha_k) - L(\alpha_0) \leq \frac{\bar{L}}{2}\|\alpha_k - \alpha_0\|^2 \leq \bar{L}\frac{\bar{r}^2}{8}$$

Given the SGD update rule (equation 18 in the main text), under Assumption PL and A.1:

$$E[L(\alpha_{t+1}) - L(\alpha_0)] \leq \big(1 - \beta\mu(1-\eta)\big)E[L(\alpha_t - L(\alpha_0)] + \bar{L}\frac{\beta^2\sigma^2}{2}$$

Setting the learning rate as $\beta \leq \frac{1-\beta}{\bar{L}(1+\eta)}$ and assuming σ^2 decays as the number of iterations increases, we are left with $E[L(\alpha_{t+1}) - L(\alpha_0)] \leq \big(1 - \beta\mu(1-\eta)\big)^t E[L(\alpha_t - L(\alpha_0)]$, ensuring linear convergence at rate $\rho = 1 - \beta\mu(1-\eta)$.

Additional Assumption for Asymptotic Normality :

Let $W_i \equiv (Y_i, X_i')'$ and define the following

$$R[\theta - \theta_0, w] \equiv l(\theta, w) - l(\theta_0, w) - l_{\theta_0}[\theta - \theta_0, w] \tag{33}$$

$$\kappa(\theta, W_i) \equiv R[\theta - \theta_0, W_i] - R[\pi_n\zeta^*(\theta, \epsilon_n) - \theta_0, W_i] \tag{34}$$

where $\zeta^*(\theta, \epsilon_n) = (1 - \epsilon_n)\theta + \epsilon_n(u^* + \theta_0)$ and $u^* = v^*$ or $u^* = -v^*$ for the Reisz representer $v^* \in \bar{V}$. We note that the assumption regarding these terms (**Assumption 3.3**) is a primitive version of the IVQRNN condition regarding the pathwise derivative. The term $\kappa(\theta, W_i)$ is assumed to satisfy a similar condition (Assumption 3.4, Chen and White (1999)) as the aNN sieve estimator in the absence of the plug-in estimator.

6.2 Linear IVQR

Linear IVQR model assumptions [Chernozhukov and Hansen (2008)]:

R1: Y_i, D_i, X_i, Z_i are iid defined on the probability space (Ω, F, P) with compact support.

R2: For the given $\tau, (\alpha(\tau), \beta(\tau))$ is in the interior of the specified set Θ.

R3: Density $f_Y(Y \mid X, D, Z)$ is bounded by a constant $\bar{f}$ a.s.

R4: $\partial E[\mathbf{1}(Y < D'\alpha + X_i'\beta + Z'\gamma)\Psi]\partial(\beta', \gamma')$ at $(\beta, \gamma) = (\beta(\alpha, \tau), \gamma(\alpha, \tau))$ has full rank for each $\alpha \in \mathcal{A}$, for $\Psi = V_i(Z_i', X_i')'$

R5': $\partial E[\mathbf{1}(Y < D'\alpha + X'\beta)\Psi]\partial(\alpha', \beta')$ has full rank at each $(\alpha, \beta) \in \Theta$

R6': The image of Θ under $(\alpha, \beta) \mapsto E[\{\tau - \mathbf{1}(Y < D'\alpha + X'\beta)\}\Psi]$ is simply connected.

6.3 Additional Results

Figure 7: IVQRNN model, linear vs. SLFN (Grid) parameter estimate, sim #1 (Nonlinear)

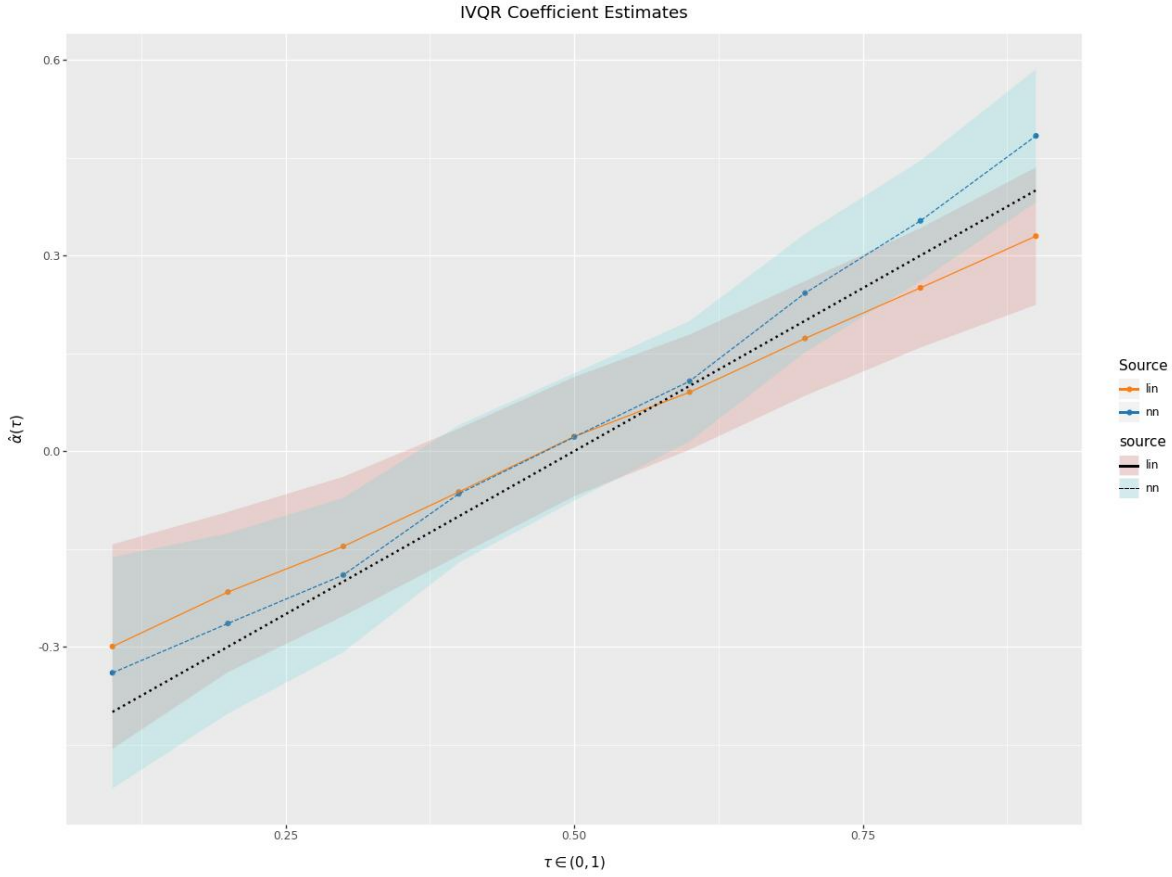


Figure 8: IVQRNN model, linear vs. SLFN (Grid, SGD) bias, sim #1 (Nonlinear)

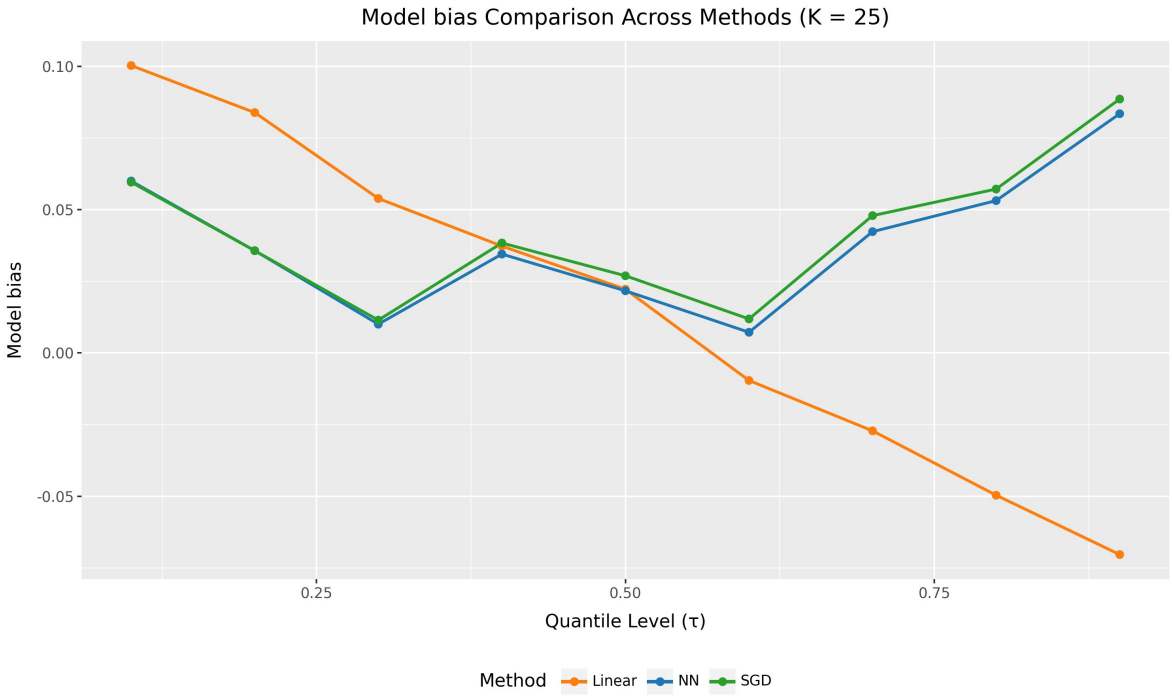


Figure 9: IVQRNN model, linear vs. SLFN (Grid) computation time, sim #1 (Nonlinear)