

Asymptotic Properties of Deep Quantile Regression Estimator for Change Point Detection

Andrew Mirii Kinyita¹, Anthony Waititu², Antony Ngunyi¹

¹Department of Statistics and Actuarial Science, Dedan Kimathi University of Technology, Nyeri, Kenya

²Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email: amkinyita@dkut.ac.ke, awaititu@jkuat.ac.ke, anthony.ngunyi@dkut.ac.ke

How to cite this paper: Kinyita, A.M., Waititu, A. and Ngunyi, A. (2026) Asymptotic Properties of Deep Quantile Regression Estimator for Change Point Detection. *Open Journal of Statistics*, **16**, 419-459. <https://doi.org/10.4236/ojs.2026.165021>

Received: August 10, 2026

Accepted: October 17, 2026

Published: October 20, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This article presents an analytical framework that utilizes deep neural network-based quantile regression to detect structural changes in high-dimensional, time dependent data. By leveraging on deep neural networks for quantile regression, the model combines the robustness of quantile regression with the computational efficiency of deep learning to estimate conditional quantiles. These estimators can identify abrupt shifts in data distribution without imposing strict structural assumptions, making them more resilient to heavy-tailed data compared to mean-based methods. The study derives the asymptotic properties of the deep quantile regression estimator for change point detection, particularly within high-dimensional, non-linear, and complex datasets. Under specific regularity conditions, the asymptotic properties of the proposed estimator are established. The simulation results demonstrate the convergence of the deep quantile regression estimator to the true conditional quantile function as the sample size. The deep quantile regression estimator achieves the minimax optimal convergence rate, effectively alleviating the curse of dimensionality when the target function adheres to a compositional structure. The empirical process theory and the method of sieve estimation are employed to demonstrate the consistency of the deep quantile regression estimator. The results show that the normalized estimator is asymptotically normal and exhibits inherent robustness to outliers and heavy-tailed noise compared to least-squares-based deep learning models. This theoretical framework provides a foundation for the construction of confidence intervals and hypothesis testing for complex, non-parametric quantile models in fields such as finance, risk management, and econometrics.

Keywords

Change Point, Single Change Point, Multiple Change Points, Deep Quantile

1. Introduction

Change happens all the time. So it is natural to become interested in detecting and estimating changes. The statistical literature on change-point analysis dates back to the 1950s [1], when it was originally used to detect changes in the mean values of independently and identically distributed Gaussian variables for quality control in the industry. As changes are not a problem of a specific scientific field, the theory of change point methods has been extensively studied across various fields, including industrial manufacturing, healthcare [2], financial data analysis [3] and environmental monitoring [4].

Over the decades, many statistical model-based methods for change point detection (CPD) have been developed to address the needs of various applications. Many publications analysing different models or adopting different points of view have since followed. Besides testing for a change, the estimation of the time of the change is also of interest. [5] considered the estimation of the change-point for the mean change model based on independent and identically distributed (i.i.d.) random variables. Most of the existing statistical models developed require some form of distributive pattern in the data and estimate the efficiency of a detected change through the theoretical asymptotic presentation [6]. Among the various statistical methods is the Bayesian change-point model, which integrates prior knowledge of changes within a probabilistic framework. This approach facilitates more informed decision-making by allowing for the systematic analysis of data where change points are anticipated [7].

Quantile regression, introduced by [8], is a vital econometric tool for modelling the conditional distribution of a response variable. Unlike traditional least squares methods that focus on the conditional mean, quantile regression estimates various points (quantiles) in the distribution, providing a deeper insight into the relationship between variables. Furthermore, quantile regression serves as a robust alternative to traditional mean-based forecasting methods, as it directly estimates conditional quantiles, thereby providing a more nuanced understanding of potential outcomes. Despite these advancements, a significant challenge remains: ensuring the consistency and reliability of change point estimators within these sophisticated high-dimensional models. This challenge underscores the need for continued research and development in this area to enhance predictive accuracy and risk management strategies. Recent developments have advanced quantile regression into more complex settings. High-dimensional quantile regression utilises regularisation techniques to manage large sets of predictors [9] [10]. Additionally, semi-parametric and non-parametric quantile models provide flexibility in functional forms [11]. However, these approaches often depend on assumptions of sparsity or smoothness and may still have difficulty capturing highly non-linear interactions

in financial data.

Deep Neural Networks (DNNs) have emerged as powerful non-parametric function approximators capable of modelling highly non-linear relationships. Universal approximation theorems establish that feed-forward neural networks with sufficient width can approximate any continuous function on compact domains [12]. Later work by [13] and others refined these results by characterising approximation error rates in terms of network depth, width, and smoothness of the target function. Deep Neural Networks (DNNs) are popular for their ability to approximate complex nonlinear functions. According to the universal approximation theorem, feed-forward neural networks can approximate any continuous function under mild conditions. This makes DNNs well-suited for modelling the non-linear and high-dimensional dependencies often found in financial time series.

In recent years, advancements in deep learning methodologies have shown remarkably improved results in several time series problems, such as modelling and forecasting of time series data, allowing for the identification of complex non-linear dependencies and interactions among diverse risk factors. This capability enables practitioners to gain deeper insights into market dynamics. Deep learning models designed especially for the processing of sequential data, including but not limited to Recurrent Neural Networks (RNNs) and Auto-encoder models, are being increasingly applied to overcome certain difficulties presented in high-dimensional data sets [14]. Although these deep learning models do not have such a strong theoretical basis compared with conventional statistical models, they demonstrate great capability in better handling complex change point detection tasks and in deeply excavating data features and transmitting temporal information within time series data. By doing so, this also provides them with an extended capability of finding out patterns and dependencies that may not be considered in traditional statistical methods. Thus, this model further develops the detecting ability of significant shifts that are difficult for statistical models to detect with ease, thus giving higher accuracy in practical applications without explicit model assumptions. Moreover, rapid development in hardware technology, such as parallelisation in GPUs and CPUs, has also ensured that the computational requirements for deep learning methods remain within feasible limits for practical applications.

Recent advances in deep learning have led to the development of various methods for change-point detection (CPD). These methods have enhanced the ability of CPD techniques to handle complex, high-dimensional data, making them more adaptable and less dependent on strict assumptions about the distributions of data. With increased complexity and volume of data, however, the development of robust and scalable methods that could effectively detect change points has increased. Despite their theoretical robustness, applications to high-dimensional time series data have mostly demonstrated limitations of classical methods. These usually manifest as model complexity and generalisation challenges across diverse, data-intensive environments [15].

The concept of deep quantile estimator that utilizes neural networks to predict Value-at-Risk (VaR) by modelling non-linear relationships between asset returns and predictors was introduced by [16]. This methodology capitalizes on the universal approximation property of neural networks to effectively capture general, non-linear associations for conditional quantiles, thereby advancing beyond traditional linear quantile regression techniques. The framework is adaptable, accommodating various penalty functions and high-dimensional covariates. Monte Carlo simulations reveal that the deep quantile estimator demonstrates commendable finite-sample performance. Moreover, empirical analyses indicate that it yields substantial improvements over linear quantile regression and other forecasting models for VaR, as substantiated by a range of statistical tests.

While deep learning methods have shown impressive empirical performance, their theoretical properties are still under active development. Recent contributions include [17], who provides convergence rates for neural network estimators; [18] established the asymptotic theory for deep learning in econometrics; and [19] analysed the neural networks as sieve estimators.

The consistency of deep quantile regression estimators under assumptions on network complexity, smoothness of the conditional quantile function, and uniform convergence of the empirical quantile loss was established by [20], [21] derived oracle inequalities for neural network estimators under general convex but non-smooth loss functions, including the quantile loss. These studies lay important foundations for understanding the statistical properties of deep quantile regression models. Theoretical work in this area remains relatively sparse.

The rest of the paper is organised as follows. Section 2 introduces the methodology for estimating multiple change points within deep quantile regression settings. Section 2.3.3 presents the statistical theory for quantile regression with ReLU networks and theoretical properties of the proposed deep quantile regression estimator. Section 3 presents the simulation study. Finally, Section 4 concludes the study.

2. Methodology

In this section, the underlying theory of a quantile regression is introduced for both linear quantile and non-linear quantile regression model. This section presents some fundamental theoretical results on the universal approximation property of neural networks that can be used to approximate a non-linear relationship inherent in financial time series data.

2.1. Quantile Regression

In statistical modelling, the concept of quantile regression is used in estimating uncertainty of the predicted variable expressed using conditional quantiles. Quantile regression estimates a quantile of a random variable of the τ -th ($0 < \tau < 1$) quantile of random variable X is defined as follows:

Definition 2.1 Suppose X is a random variable on a probability space

$(\Omega, \mathcal{F}, \mathbf{P})$ is characterised by a (right-continuous) distribution function denoted by $F(\cdot)$, For any $0 < \tau < 1$, the τ -th quantile of X is given by

$$F^{-1}(\tau) = \inf \{x : F(x) \geq \tau\}. \quad (1)$$

Linear Quantile Regression Model

[8] developed the theory for the estimation of the quantiles of a variable Y_i which is assumed to be a linear function of other variables. The linear quantile regression model can be defined as

$$Q_y(\tau | \mathbf{x}_i) = \mathbf{x}_i' \beta(\tau), \quad \tau \in (0, 1), \quad (2)$$

such that y_i satisfies the quantile constraint $P(y_i \leq \mathbf{x}_i' \beta(\tau) | X_i) = \tau$, where $\beta(\tau)$ are regression coefficients that depend on τ . Quantile regression estimation determines an estimate for the τ^{th} conditional quantile, $\hat{Q}_y(\tau, X_i) = \hat{\beta}(\tau)$, by minimizing the following function

$$\hat{\beta}(\tau) = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}_i' \beta(\tau)), \quad (3)$$

where $\rho_{\tau}(\cdot)$ is the quantile loss function defined as

$$\rho_{\tau}(u_i) = \begin{cases} \tau u_i(\tau), & \text{if } u_i(\tau) \geq 0 \\ (1-\tau)u_i(\tau), & \text{if } u_i(\tau) < 0 \end{cases} \quad (4)$$

and $y_i(\tau) = y_i - \mathbf{x}_i'(\tau)$. The quantile estimator in Equation (2) provides more information on the whole conditional distribution of y_i as a function of the X_i , and more robust estimates under the presence of outliers and non-linearities, when compared to the ordinary least squares estimator.

The linear association assumption for the conditional quantile of the dependent variable given the covariates, $Q_y(\tau | \mathbf{x}_i) = \mathbf{x}_i' \beta(\tau)$, can be generally restrictive when non-linearities and outliers exist [22]. There is extensive literature on non-parametric quantile regression that relaxes the strict linearity assumption in the covariates. As non-parametric quantile regression models become more complex, local fitting and back-fitting to accommodate new components become more computationally demanding, and the literature has evolved towards non-linear quantile regression models.

In this paper, the focus is on the case of the following non-linear is on the case of the following non-linear association, $Q_y(\tau | \mathbf{x}_i) = h_{\tau}(\mathbf{x}_i)$ where $h_{\tau}(\cdot)$ is some unknown non-linear function. The estimation strategy proposed is to approximate $h_{\tau}(\mathbf{x}_i)$ with neural networks using the universal approximation property. Specifically, we assume that there exists a neural network with a function $G_{\tau}(\mathbf{x}_i, \mathbf{w})$, that can approximate $h_{\tau}(\mathbf{x}_i)$ accurately. Before we illustrate how the methodology is implemented, the next section presents a discussion on how neural networks can approximate $h_{\tau}(\mathbf{x}_i)$.

2.2. Deep Neural Networks

In this paper, the feed-forward neural network is utilized to approximate $h_{\tau}(\mathbf{x}_i)$.

This architecture consists of an input layer of covariates, the hidden layer(s) where non-linear transformation of covariates occur, and the output layer that gives the final prediction. Each hidden layer has several interconnected neurons relating it to both the previous and the next ones. Specifically, information flows from one layer to the other, via neurons only in one direction, and the connections correspond to weights. Optimizing a loss function with respect to these weights makes the neural network capable of learning.

Given that L denote the total number of hidden layers, a measure for the depth of a neural network, and $J^{(l)}$ denotes the total number of neurons at layer l , a measure for its width. First, a general definition of a deep (multi-layer) feed-forward neural network is presented. Let $\sigma_l(\cdot), l = 0, \dots, L$ be the activation function used at the l -th layer, that is applied element-wise and induces non-linearity. The ReLU activation function, $\sigma_l(\cdot) = \max(\cdot, 0)$, for $l = 1, \dots, L-1$ is utilized and a linear one for the output layer, $l = L$. The ReLU activation function, despite its utter simplicity, seems to be the most popular choice in practical applications [23]. The output of the l -th layer denoted by $g^{(l)}$, is a vector of length equal to the number of $J^{(l)}$ neurons in that layer, such that $g^{(0)} = \mathbf{x}_t$. The overall structure of a deep (multi-layer) feed-forward neural network is given by:

$$G_{\tau}(\mathbf{x}_t, \mathbf{w}) = g^{(L)} \left(g^{(L-1)} \left(\dots \left(g^{(l)}(\cdot) \right) \right) \right), \quad (5)$$

where

$$g^{(l)}(\mathbf{x}_t) = \sigma_l \left(\mathbf{W}^{(l-1)} g^{(l-1)} + \mathbf{b}^{(l)} \right), \quad l = 1, \dots, L, \quad (6)$$

$\mathbf{W}^{(l)}$ is a $J^{(l)} \times J^{(l-1)}$ matrix of weights, $\mathbf{b}^{(l)}$ is a $J^{(l)} \times 1$ vector of biases giving an overall vector $\mathbf{w} = \left(\text{vec}(\mathbf{W}^{(0)})', \dots, \text{vec}(\mathbf{W}^{(L)})', \mathbf{b}^{(1)'}', \dots, \mathbf{b}^{(L)'}' \right)$ of trainable parameters of dimensions $J^{(l)}(1 + J^{(l-1)})$ total number of parameters in each hidden layer l , $J^{(0)} = p$ and $J^{(L)} = 1$.

The increase in complexity can occur, either by letting $L \rightarrow \infty$, which stands for deep learning, or by letting $J^{(l)} \rightarrow \infty$. While both ways asymptotically deliver the same results [18], the approximation error has been shown to decline exponentially with L , [24] but only polynomially with $J^{(l)}$, providing some evidence for the prevalent use of deep learning. The work of [17] also proposes an alternative approximation theory for sparse deep learning. According to various approximation theorems (see the theoretical results in [25] and [26]), $G_{\tau}(\mathbf{x}_t, \mathbf{w})$ can approximate arbitrarily well $h_{\tau}(\mathbf{x}_t)$, such that, for any $\varepsilon > 0$,

$$\sup_t |G_{\tau}(\mathbf{x}_t, \mathbf{w}) - h_{\tau}(\mathbf{x}_t)| < \varepsilon \quad (7)$$

The ε -approximation can be seen as a sieve type non-parametric estimation bond, where ε can become arbitrarily small by increasing the complexity of $G_{\tau}(\mathbf{x}_t, \mathbf{w})$.

2.3. Deep Neural Network Quantile Regression Model

Assume that the conditional quantile follows a non-linear relationship

$Q_y(\tau | \mathbf{x}_i) = h_\tau(\mathbf{x}_i)$ and there exists a function $G_\tau(\mathbf{x}_i, \mathbf{w})$, that can ε -approximate $h_\tau(\mathbf{x}_i)$, in Equation (7). Specifically, it is assumed that there exists a neural network with a function $G_\tau(\mathbf{x}_i, \mathbf{w})$, that can approximate $h_\tau(\mathbf{x}_i)$. Under this assumption, the conditional quantile function can be defined as the following approximation

$$Q_y(\tau | \mathbf{x}_i) = G_\tau(\mathbf{x}_i, \mathbf{w}) + O(\varepsilon), \quad (8)$$

where $G_\tau(\mathbf{x}_i, \mathbf{w})$ is the unknown non-linear function to be estimated in order to approximate $h_\tau(\mathbf{x}_i)$. The deep neural network conditional quantile estimate is obtained from the solution of the following minimization problem:

$$Q_y(\tau | \mathbf{x}_i) = \arg \min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - G_\tau(\mathbf{x}_i, \mathbf{w})), \quad (9)$$

where $\mathbf{w} = \left(\text{vec}(\mathbf{W}^{(0)})', \dots, \text{vec}(\mathbf{W}^{(L)})' \right)'$, $b^{(1)'}, \dots, b^{(L)}'$ contains all model parameters, and $G_\tau(\mathbf{x}_i, \mathbf{w})$ will govern whether the model is parametric or non-parametric. If the number of neurons and layers is small, then the model is parametric, if the above number becomes large, then the model becomes non-parametric, since the number of estimated parameters increases with the sample size, similar to sieve non-parametric approximations.

The network parameters $\mathbf{w}$ are estimated by minimizing the quantile regression loss function (also called the pinball loss) for a given quantile level $\tau \in (0, 1)$.

$$S_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(Y_i - g_\tau(\mathbf{x}_i, \mathbf{w})) \quad (10)$$

where $\rho_\tau(u) = u(\tau - \mathbb{I}\{u < 0\})$, and $\mathbb{I}\{\cdot\}$ is the indicator function.

At the population level, the non-parametric quantile estimation is to find a measurable function $f^*: R^d \rightarrow R$ satisfying

$$f^* := \arg \min_f \mathbf{R}^r(f) = \arg \min_f E_Z \left\{ \rho_\tau((\mathbf{x}_i, \mathbf{w})) \right\} \quad (11)$$

where E_Z means that the expectation is taken with respect to the distribution of Z . If the conditional τ -th quantile of ν given X is 0 and

$E(|\nu| | X = x) < \infty$ for all $x \in X$, then the true regression function f_0 is the optimal solution of f^* on X .

For a given random function $g_\tau(\mathbf{x}_i, \mathbf{w}): R^d \rightarrow R$, where $Z \equiv (X, Y)$ is a random vector independent of $g_\tau(\mathbf{x}_i, \mathbf{w})$, the risk of $g_\tau(\mathbf{x}_i, \mathbf{w})$ under the loss function $\rho_\tau(\cdot)$ is defined by

$$R^\tau(g_\tau(\mathbf{x}_i, \mathbf{w})) = E_Z \left\{ \rho_\tau(Y_i - g_\tau(\mathbf{x}_i, \mathbf{w})) \right\} \quad (12)$$

where $\theta \in \Theta \subset \mathbb{R}^d$.

Let $\mathcal{D}_n = \{Z_1, \dots, Z_n\}$ denote the observed data, where $Z_i = (X_i, Y_i) \in [0, 1]^p \times \mathbb{R}$.

In applications, when a random sample $S \equiv \{(X_i, Y_i)\}_{i=1}^n$ is available, the Deep Quantile Regression estimator $\hat{q}_\tau(x)$ is defined as the minimizer of the empirical risk:

$$R_n^\tau(f) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(x_i, w) \quad (13)$$

The goal is to construct an estimator of f_0 within a certain class of functions $\mathcal{F}_n$ by minimising the empirical risk, that is

$$\hat{f}_n \in \arg \min_{f \in \mathcal{F}_n} R_n^\tau(f), \quad (14)$$

where $\hat{f}_n$ is the deep quantile regression (DQR) estimator, which serves as the empirical risk minimizer (ERM). The function class $\mathcal{F}_n$ consists of deep neural networks (DNNs). The architecture of a DQR generally comprises multiple hidden layers with non-linear activation functions, similar to a standard feed-forward neural network.

2.3.1. The Deep Quantile Regression Model

Let $\{(X_i, Y_i)\}_{i=1}^T$ be a strict stationary and mixing stochastic process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $Y_i \in \mathbb{R}$ denotes the target financial variable such as asset returns, $X_i \in \mathcal{X} \subset \mathbb{R}^d$ denotes the lagging conditioning information like past returns, realized variance and macroeconomic factors. The target conditional τ -th quantile of Y_i given $X_i = x$ is denoted as $g_0(x) = Q_\tau(Y_i | X_i = x)$. Let $\mathcal{F}_{DNN}(L, W, S, B)$ define a class of deep neural networks parametrized by L - denote number of hidden layers (depth), W - denote maximum width per layer, S - denote sparsity constraint (total number of non-zero parameters) and B - denote bounded parameter size ($\|\theta\|_\infty \leq B$). This conditional quantile is parametrized via a deep neural network (DNN) architecture $f(x; \theta)$, where $\theta \in \Theta \subset \mathbb{R}^{p_\tau}$ represents the neural network parameter vector (weights and biases).

The DQR estimator $\hat{g}_\tau(\cdot) = f(\cdot; \hat{\theta}_\tau) = \hat{\theta}_\tau$ is obtained by solving the empirical risk minimization (ERM):

$$\hat{\theta}_\tau = \arg \min_{\theta \in \Theta} \frac{1}{T} \sum_{i=1}^T \rho_\tau(Y_i - f(X_i; \theta)) \quad (15)$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the classical pinball loss function. The population pseudo-true parameter vector is defined as:

$$\theta^* = \arg \min_{\theta \in \Theta} E[\rho_\tau(Y_i - f(X_i; \theta))] \quad (16)$$

2.3.2. Assumptions

The following regularity assumptions are assumed because financial time series data possess heavy tails and conditional heteroscedasticity:

Assumption 1. *The true conditional quantile function g_0 belongs to a Hölder continuous function space $\mathcal{H}^\beta(\mathcal{X})$ with smoothness index $\beta > 0$. That is, all partial derivatives of g_0 up to the order $\lfloor \beta \rfloor$ exist and are Lipschitz continuous of order $\beta - \lfloor \beta \rfloor$.*

Assumption 2. The stochastic process $\{(X_t, Y_t)\}$ is strictly stationary and absolutely regular (β -mixing) or strongly mixing (α -mixing). The mixing coefficients $\beta(m)$ decaying exponentially: $\beta(m) \leq c_1 \exp(-c_2 m)$ for some constants $c_1, c_2 > 0$ and α -mixing coefficients satisfying $\alpha(m) = O(m^{-\gamma})$ for $\gamma > \frac{2+\delta}{\delta}$ where $\delta > 0$, ensuring the memory of the shocks decays at a geometric or high polynomial rate.

The financial time series must satisfy a form of weak dependence (such as α -mixing or β -mixing) to ensure the consistency and asymptotic normality of the deep quantile regression estimators. This implies that data points far apart in time become approximately independent. This bounds the long-term memory property of asset volatility.

Assumption 3. The conditional distribution $F_{Y|X}(y|x)$ is continuous and has a continuous, differentiable density $f_{Y|X}(y|x)$ that is strictly bounded away from zero and infinity in a neighbourhood around the true conditional quantile

$$0 < \underline{f} \leq f_{Y|X}(f(x; \theta^*)|x) \leq \bar{f} < \infty.$$

Assumption 4. The parameter space Θ is a compact subset of $\mathbb{R}^P$. The activation functions (like ReLU) ensure that $\theta \rightarrow f(x; \theta)$ is twice continuously differentiable with uniformly bounded gradients:

$$E\left[\left\|\nabla_{\theta} f(X_t; \theta^*)\right\|^{2+\delta}\right] < \infty.$$

Assumption 5. The positive definite matrices $J(\theta^*)$ (Hessian counterpart) and $\Omega(\theta^*)$ (information matrix) are defined as follows:

$$J(\theta^*) = E\left[f_{Y|X}(f(X_t; \theta^*)|X_t) \cdot \nabla_{\theta} f(X_t; \theta^*) \nabla_{\theta} f(X_t; \theta^*)^T\right] \geq 0 \quad (17)$$

$$\Omega(\theta^*) = \lim_{T \rightarrow \infty} \text{Var}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \phi_{\tau}(Y_t - f(X_t; \theta^*)) \nabla_{\theta} f(X_t; \theta^*)\right) > 0 \quad (18)$$

where $\psi_{\tau}(u) = \tau - I(u < 0)$ is the sub-gradient score function.

Assumption 6. The neural network utilizes a continuous, piece-wise linear activation function (such as ReLU). The parameters scale with the sample size T such that

$$L \asymp \log T, \quad W \asymp T^{\frac{d}{2\beta+d}} \log T, \quad S \asymp T^{\frac{d}{2\beta+d}} \log T$$

The assumption guarantees that the temporal memory of shocks decays exponentially over time.

2.3.3. Asymptotic Properties of the DQR Estimator

This section outlines the asymptotic properties of the Deep Quantile Regression (DQR) estimator, which merges the flexibility of deep neural networks with the robustness of quantile regression. The assumption of independent and identically distributed (i.i.d.) observations does not hold for financial market data due to the presence of volatility clustering, regime shifts, and strong temporal dependence.

To address these dependencies, we replace i.i.d. assumptions with regularized dynamic constraints. Consequently, the framework for deriving the asymptotic properties of the DQR estimator relies on strict stationarity and either β -mixing (absolutely regular) or α -mixing (strongly mixing) conditions to effectively capture the data dynamics.

Consistency

The consistency of the DQR estimator under temporal dependence is given by the following theorem.

Theorem 1 (Consistency). *Under assumptions A1-A3, if the VC-dimension or covering number of the deep neural network function class $\mathcal{F} = \{f(\cdot; \theta) : \theta \in \Theta\}$ grows at rate slower than sample size $(o(T))$, then:*

$$\hat{\theta} \xrightarrow{p} \theta^* \text{ as } T \rightarrow \infty \quad (19)$$

Proof. The convexity Lemma ([27] [28]): Point wise convergence in probability of convex functions defined on an open set to a convex limit function implies uniform convergence on compact subsets. Therefore, for any compact set $\Theta \subset \mathbb{R}$ containing $q_\tau(x)$:

$$\sup_{\theta \in \Theta} |M_{n,B}(\theta) - M(\theta)| \xrightarrow{p} 0 \quad (20)$$

Standard M-estimation arguments Argmin continuous mapping Theorem [29] imply consistency:

$$\hat{q}_\tau(x) = \arg \min_{\theta} M_{n,B}(\theta) \xrightarrow{p} q_\tau(x) = \arg \min_{\theta} M(\theta) \quad (21)$$

The following M-estimation lemma is utilized in deriving the consistency of the deep quantile regression estimator under temporal dependence.

Lemma 1. Let $\mathbb{M}_T(\theta) = \frac{1}{T} \sum_{t=1}^T \rho_\tau(Y_t - f(X_t; \theta))$ and

$$\mathbb{M}(\theta) = \mathbb{E}[\rho_\tau(Y_t - f(X_t; \theta))].$$

If $\sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| \xrightarrow{p} 0$ as $T \rightarrow \infty$ (under the Uniform Law of Large Numbers), θ^* is the unique minimizer of $\mathbb{M}(\theta)$, and for any $\varepsilon > 0$, $\inf_{\|\theta - \theta^*\| \geq \varepsilon} \mathbb{M}(\theta) > \mathbb{M}(\theta^*)$. Then any sequence of empirical minimizers $\hat{\theta}_T$ satisfying the inequality $\mathbb{M}_T(\hat{\theta}_T) \leq \mathbb{M}_T(\theta^*) + o_p(1)$ converges in probability to θ^* , i.e., $\hat{\theta}_T \xrightarrow{p} \theta^*$.

By the definition of the estimator $\hat{\theta}_T$ as the empirical minimizer, we have the initial condition:

$$\mathbb{M}_T(\hat{\theta}_T) \leq \mathbb{M}_T(\theta^*) \quad (22)$$

This statement can be manipulated by adding and subtracting the population limits $\mathbb{M}(\hat{\theta}_T)$ and $\mathbb{M}(\theta^*)$:

$$\begin{aligned} \mathbb{M}(\hat{\theta}_T) - \mathbb{M}(\theta^*) &= (\mathbb{M}(\hat{\theta}_T) - \mathbb{M}_T(\hat{\theta}_T)) + (\mathbb{M}_T(\hat{\theta}_T) - \mathbb{M}_T(\theta^*)) \\ &\quad + (\mathbb{M}_T(\theta^*) - \mathbb{M}(\theta^*)) \end{aligned} \quad (23)$$

Using the empirical minimizer inequality ($\mathbb{M}_T(\hat{\theta}_T) - \mathbb{M}_T(\theta^*) \leq 0$), we bound the equation from Equation (22), to obtain

$$\mathbb{M}(\hat{\theta}_T) - \mathbb{M}(\theta^*) \leq (\mathbb{M}(\hat{\theta}_T) - \mathbb{M}_T(\hat{\theta}_T)) + (\mathbb{M}_T(\theta^*) - \mathbb{M}(\theta^*)) \quad (24)$$

To make this statement uniform across the parameter space Θ , we take the supremum over all possible parameters $\theta \in \Theta$:

$$0 \leq \mathbb{M}(\hat{\theta}_T) - \mathbb{M}(\theta^*) \leq 2 \sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| \quad (25)$$

Therefore, to prove that $\mathbb{M}(\hat{\theta}_T) \xrightarrow{p} \mathbb{M}(\theta^*)$, it is necessary and sufficient to prove that the uniform empirical process term converges to zero:

$$\sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| \xrightarrow{p} 0 \quad (26)$$

Given that the sample data $\{(X_t, Y_t)\}_{t=1}^T$ is strictly stationary and α -mixing (Assumption A1), the classical Glivenko-Cantelli theorem designed for i.i.d. random variables cannot be directly applied. A sequence of independent blocks is constructed by dividing the time horizons $1, \dots, T$ into $2k_T$ blocks of length m_T , such that $T = 2k_T m_T$. Let the odd blocks be denoted by $H_j = \{(j-1)2m_T + 1, \dots, (j-1)2m_T + m_T\}$ for $j = 1, \dots, k_T$. Let the even blocks be denoted by $I_j = \{(2j-1)m_T + 1, \dots, 2jm_T\}$ for $j = 1, \dots, k_T$. The empirical process can be rewritten as the sum of its odd and even structural components given by

$$\mathbb{M}_T(\theta) - \mathbb{M}(\theta) = \frac{1}{2}(\mathbb{M}_{T,\text{odd}}(\theta) - \mathbb{M}(\theta)) + \frac{1}{2}(\mathbb{M}_{T,\text{even}}(\theta) - \mathbb{M}(\theta)) \quad (27)$$

By Berbee's Lemma, there exists a new sequence of coupled variables $\{(\tilde{X}_t, \tilde{Y}_t)\}_{t=1}^T$ such that: Within each block H_j , the joint distribution of $(\tilde{X}_t, \tilde{Y}_t)$ is identical to the original time series (X_t, Y_t) . The blocks $H_1, H_2, \dots, H_{k_T}$ are mutually independent. The total variation distance between the true dependent joint distribution and the decoupled independent block sequence satisfies:

$$\mathbb{P}\left(\bigcup_{t=1}^T \{(X_t, Y_t) \neq (\tilde{X}_t, \tilde{Y}_t)\}\right) \leq 2k_T \cdot \alpha(m_T) \quad (28)$$

Under Assumption A1, the mixing coefficients decay geometrically ($\alpha(m) \leq c_1 e^{-c_2 m}$). By setting the tracking block size parameters to scale logarithmically, $m_T = \frac{2}{c_2} \log T$, the coupling failure bound simplifies:

$$2k_T \alpha(m_T) \leq \frac{T}{m_T} c_1 e^{-2 \log T} = \frac{c_1}{T \cdot m_T} \rightarrow 0 \text{ as } T \rightarrow \infty \quad (29)$$

Thus, with probability approaching 1 ($1 - o(1)$), we can execute the rest of the uniform convergence proof as if the blocks were perfectly independent.

Let $\mathcal{G} = \{(x, y) \mapsto \rho_\tau(y - f(x; \theta)) : \theta \in \Theta\}$. Given that the pinball loss $\rho_\tau(\cdot)$ is Lipschitz continuous with parameter $\max(\tau, 1 - \tau) \leq 1$, the covering number of the transformed loss class is bounded by the covering number of the network function class $\mathcal{F}$:

$$N(\varepsilon, \mathcal{G}, L_1(\mathbb{P}_T)) \leq N(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_T)) \quad (30)$$

Using the standard symmetrization theorem via Rademacher variables $\sigma_j \in \{-1, 1\}$ over the k_T independent blocks, we bound the expected supremum of the empirical process:

$$\begin{aligned} & \mathbb{E} \left[\sup_{\theta \in \Theta} |\mathbb{M}_{T, \text{odd}}(\theta) - \mathbb{M}(\theta)| \right] \\ & \leq 2 \mathbb{E}_{\tilde{X}, \tilde{Y}} \left[\mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{k_T} \sum_{j=1}^{k_T} \sigma_j \cdot \left(\frac{1}{m_T} \sum_{t \in H_j} \rho_t(\tilde{Y}_t - f(\tilde{X}_t)) \right) \right| \right] \right] \end{aligned} \quad (31)$$

Applying Dudley's Entropy Integral Theorem to control the sub-Gaussian process via the log covering numbers yields:

$$\mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \dots \right] \leq \frac{C}{\sqrt{k_T}} \int_0^{2B} \sqrt{\log N(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_T))} d\varepsilon \quad (32)$$

By the explicit theorem condition, the VC-dimension or covering capacity satisfies $\log N(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_T)) = o(T/m_T) = o(k_T)$. As a result, the integral scales bounded under the square root, and the scaling coefficient ensures:

$$\frac{1}{\sqrt{k_T}} \cdot \sqrt{o(k_T)} = o(1) \rightarrow 0 \text{ as } T \rightarrow \infty \quad (33)$$

This establishes the Uniform Law of Large Numbers for our time series process:

$$\sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| \xrightarrow{P} 0 \quad (34)$$

Finally, the convergence of the objective functions is translated into the convergence of the parameter vector $\hat{\theta}_T$. Suppose $\hat{\theta}_T$ does not converge to θ^* . Then there exists a subsequence and a fixed radius $\varepsilon_0 > 0$ such that $\|\hat{\theta}_T - \theta^*\| \geq \varepsilon_0$ infinitely often. Define the compact subset $\Theta_{\varepsilon_0} = \{\theta \in \Theta : \|\theta - \theta^*\| \geq \varepsilon_0\}$. By Assumption A1 (θ^* is a unique, interior minimum) and Assumption A2 ($f_{Y|X}$ is continuous and bounded away from zero, implying strict convexity of the population risk), we have a positive gap value $\delta > 0$:

$$\inf_{\theta \in \Theta_{\varepsilon_0}} \mathbb{M}(\theta) = \mathbb{M}(\theta^*) + \delta \quad (35)$$

Now, trace the following sequence of events:

$$\begin{aligned} \mathbb{M}(\hat{\theta}_T) \geq \mathbb{M}(\theta^*) + \delta & \Rightarrow \mathbb{M}(\hat{\theta}_T) - \mathbb{M}_T(\hat{\theta}_T) + \mathbb{M}_T(\hat{\theta}_T) - \mathbb{M}(\theta^*) \geq \delta \\ & \Rightarrow \sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| + \mathbb{M}_T(\theta^*) - \mathbb{M}(\theta^*) \geq \delta \\ & \Rightarrow 2 \sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| \geq \delta \end{aligned} \quad (36)$$

Taking the probability limits of both sides:

$$\mathbb{P}(\|\hat{\theta}_T - \theta^*\| \geq \varepsilon_0) \leq \mathbb{P}\left(\sup_{\theta \in \Theta} |\mathbb{M}_T(\theta) - \mathbb{M}(\theta)| \geq \frac{\delta}{2}\right) \quad (37)$$

Since we proved that the right-hand side of Equation (37) converges to zero as $T \rightarrow \infty$, the probability of the parameter falling outside the epsilon ball must also

vanish:

$$\lim_{T \rightarrow \infty} \mathbb{P} \left(\left\| \hat{\theta}_T - \theta^* \right\| \geq \varepsilon_0 \right) = 0 \quad (38)$$

Hence,

$$\hat{\theta}_T \xrightarrow{p} \theta^*$$

Given the α -mixing property in assumption (A1) and compactness in assumption (A3), the classical i.i.d. Uniform Law of Large numbers (ULLN) for mixing processes is replaced by an empirical process central limit theorem for dependent sequences. The empirical risk uniformly converges to the population risk given by:

$$\sum_{\theta \in \Theta} \left| \frac{1}{T} \sum_{t=1}^T \rho_\tau(Y_t - f(X_t; \theta)) - E[\rho_\tau(Y_t - f(X_t; \theta))] \right| \xrightarrow{p} 0 \quad (39)$$

Since θ^* uniquely minimizes the continuous, strictly convex expected pinball loss profile over Θ , any sequence minimizing the empirical objective must merge in probability with θ^* .

Asymptotic Normality

The following theorem describes the asymptotic normality property of the DQR estimator.

Theorem 2 (Asymptotic Normality). *Under assumptions A1-A4, as $T \rightarrow \infty$, the normalized DQR estimator satisfies:*

$$\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{d} \mathcal{N}(0, J(\theta^*)^{-1} \Omega(\theta^*) J(\theta^*)^{-1}) \quad (40)$$

Proof. The pinball loss function $\rho_\tau(u)$ is non-differentiable at zero and as a result the standard Taylor series expansion fails. The Knight's Identity is utilized to isolate the parameter deviation vector $u = \sqrt{T}(\theta - \theta^*)$:

$$\rho_\tau(x - v) - \rho_\tau(x) = -v \psi_\tau(x) + \int_0^v (I(x \leq s) - I(x \leq 0)) ds \quad (41)$$

Let $v_t(\theta) = f(X_t; \theta) - f(X_t; \theta^*)$, applying the identity to our empirical objective gives a localized objective function $Z_T(u)$:

$$\begin{aligned} Z_T(u) &= \sum_{t=1}^T \left[\rho_\tau(Y_t - f(X_t; \theta^* + T^{-1/2}u)) - \rho_\tau(Y_t - f(X_t; \theta^*)) \right] \\ &= -\frac{1}{\sqrt{T}} \sum_{t=1}^T u^T \nabla_{\theta} f(X_t; \theta^*) \psi_\tau(\varepsilon_t) \\ &\quad + \sum_{t=1}^T \int_0^{T^{-1/2}u^T \nabla_{\theta} f(X_t; \theta^*)} (I(\varepsilon_t \leq s) - I(\varepsilon_t \leq 0)) ds + o_p(1) \end{aligned} \quad (42)$$

where $\varepsilon_t = Y_t - f(X_t; \theta^*)$ represents the true conditional quantile residual.

Taking the conditional expectation of the integral term in Equation (42) with respect to the filtration $\mathcal{F}_{t-1}$ and applying a localized Taylor series expansion to the conditional distribution function $F_{Y|X}$:

$$E \left[\int_0^{T^{-1/2} u^T \nabla_{\theta} f(X_t; \theta^*)} (I(\varepsilon_t \leq s) - I(\varepsilon_t \leq 0)) ds \mid X_t \right] \approx \frac{1}{2T} u^T \nabla_{\theta} f(X_t; \theta^*) f_{Y|X}(f(X_t; \theta^*) \mid X_t) \nabla_{\theta} f(X_t; \theta^*)^T u \quad (43)$$

Summing the right-hand side of Equation (43) over $t=1$ to T , by the Ergodic Theorem for stationarity processes, this converges in probability to:

$$\frac{1}{2} u^T J(\theta^*) u \quad (44)$$

Define the score vector $W_t = \psi_{\tau}(\varepsilon_t) \nabla_{\theta} f(X_t; \theta^*)$. By definition of a conditional quantile, $E[\psi_{\tau}(\varepsilon_t) \mid X_t] = 0$. meaning W_t forms a martingale difference sequence (MDS) if the conditioning set matches the filtration perfectly. However, under α -mixing, W_t is a weakly dependent stationary process. Applying the Gordin-Root Functional Central Limit Theorem for α -mixing sequences:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \xrightarrow{d} \mathcal{N}(0, \Omega(\theta^*)) \quad (45)$$

where $\Omega(\theta^*)$ is the Long-Run Variance Covariance matrix (LRV) given by

$$\omega(\theta^*) = \Sigma_0 + \sum_{k=1}^{\infty} (\Sigma_k + \Sigma_k^T) \text{ with } \Sigma_k = E[W_t W_{t-k}^T] \quad (46)$$

when the neural network successfully captures all temporal dynamics such that residuals are conditionally independent (no remaining autocorrelation in the signs of the residuals), $\Omega(\theta^*)$ simplifies to

$$\Omega(\theta^*) = \tau(1-\tau) E \left[\nabla_{\theta} f(X_t; \theta^*) \nabla_{\theta} f(X_t; \theta^*)^T \right]$$

However, if quantile correlation remains, the full long-run variance structure is required.

Substituting the components in Equation (42), the objective function behaves asymptotically as:

$$Z_T(u) = -u^T \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \right) + \frac{1}{2} u^T J(\theta^*) u + o_p(1) \quad (47)$$

By the Convexity Lemma, the minimizer $\hat{u} = \sqrt{T}(\hat{\theta}_T - \theta^*)$ converges to the minimizer of the limit quadratic form given by:

$$\hat{u} = J(\theta^*)^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \right) + o_p(1) \quad (48)$$

Substituting the normal limit of the score vector gives the result:

$$\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{d} \mathcal{N}(0, J(\theta^*)^{-1} \Omega(\theta^*) J(\theta^*)^{-1}) \quad (49)$$

To obtain the convergence rate for deep quantile regression (DQR) estimator, Theorem 3.4.1 in [30] (use the oracle inequality framework) is applied. The goal is to find the optimal balance between the network's ability to approximate a smooth function and the statistical penalty of using a high-capacity model.

Rate of Convergence

The following theorem describes the rate of convergence property of the DQR estimator.

Theorem 3 (Rate of Convergence). *Under Assumptions A1-A4, if the network hyperparameters are optimally balanced relative to the dataset length T and dimension d , the excess risk and the $L_2(\mathbb{P}_X)$ convergence rate of the deep quantile estimator satisfies*

$$\|\hat{g}_T - g_0\|_{L_2(\mathbb{P}_X)}^2 = \mathcal{O}\left(T^{-\frac{2\beta}{2\beta+d}} \log^\xi T\right) \quad (50)$$

where $\xi > 0$ is a structural logarithmic multiplier. This rate matches the information-theoretic minimax-optimal rate for non-parametric estimation, up to a log factor.

Proof. The proof of the rate of convergence for the DQR estimator relies on the convexity of the objective function and the uniform convergence of the empirical risk. The Convexity Lemma (Pollard, 1991) is utilized to establish uniform convergence from point wise convergence, avoiding the need for smoothness of the score function.

The metric error is decomposed into approximation error (neural expressivity) and stochastic error (variance)

$$\begin{aligned} & E[\rho_\tau(Y - \hat{g}_T)] - E[\rho_\tau(Y - g_0)] \\ &= \underbrace{\inf_{f \in \mathcal{F}_{\text{DNN}}} (E[\rho_\tau(Y - f)] - E[\rho_\tau(Y - g_0)])}_{\text{Approximation Error (Bias)}} + \underbrace{\mathcal{E}_{\text{stoch}}}_{\text{Stochastic Error (Variance)}} \end{aligned} \quad (51)$$

By performing a second-order Taylor series expansion of the population objective function around $g_0(X_t)$ and applying Assumption A3, the excess quantile risk is linked to the L_2 function norm:

$$\begin{aligned} & E[\rho_\tau(Y_t - f(X_t)) | X_t] - E[\rho_\tau(Y_t - g_0(X_t)) | X_t] \\ &= \int_{g_0(X_t)}^{f(X_t)} (f(X_t) - y) f_{Y|X}(y | X_t) dy \geq \frac{f}{2} (f(X_t) - g_0(X_t))^2 \end{aligned} \quad (52)$$

Integrating over the distribution of X yields the key geometric margin inequality given by

$$E[\rho_\tau(Y - f)] - E[\rho_\tau(Y - g_0)] \geq C_0 \|f - g_0\|_{L_2(\mathbb{P}_X)}^2$$

By deep network approximation results for Holder spaces $\mathcal{H}^\beta$, for a network satisfying Assumption (A4), there exists an oracle network function $f_T^* \in \mathcal{F}_{\text{DNN}}$ satisfying the following inequality:

$$\|f_T^* - g_0\|_\infty \leq c \cdot (W \cdot L)^{-\frac{2\beta}{d}} \asymp T^{-\frac{\beta}{2\beta+d}}$$

Thus, the approximation error contribution scales bounded as $\mathcal{O}\left(T^{-\frac{\beta}{2\beta+d}}\right)$.

To bound $\mathcal{E}_{\text{stoch}}$ given dependent time series observations, the blocking tech-

nique is utilized. The sequence $\{1, \dots, T\}$ is divided into $2k_t$ blocks of size m_T , where $T = 2k_T m_T$. Let H_j capture the odd blocks and I_j capture the even blocks. By Berbee's Lemma, we can construct an independent copy sequence $\{(\tilde{X}_t, \tilde{Y}_t)\}$ such that the joint distribution of each block matches the original, but blocks are independent of each other. The probability of the coupling failing is bounded by the mixing coefficient sum: $2k_T \beta(m_T)$.

Using Assumption (A2), we set $m_T \asymp \log T$ to drive the coupling failure rate to zero exponentially fast:

$$2k_T \beta(m_T) \rightarrow 0 \text{ as } T \rightarrow \infty$$

Applying the Rademacher complexity/VC-dimension bound $(\text{VCdim}(\mathcal{F}_{\text{DNN}}) \leq c \cdot SL \log(S))$ over the coupled independent blocks, the stochastic empirical process variation scales with the capacity metric:

$$\mathcal{E}_{\text{stoch}} = \mathcal{O}\left(\frac{\text{VCdim}(\mathcal{F}_{\text{DNN}}) \cdot m_T}{T}\right) \quad (53)$$

Substituting the neural network size profiles from Assumption (A4) in Equation (53)

$$\mathcal{O}_P\left(\frac{T^{\frac{d}{2\beta+d}} \cdot \log^2 T \cdot \log T}{T}\right) = \mathcal{O}\left(T^{\frac{2\beta}{2\beta+d}} \log^3 T\right) \quad (54)$$

Combining the symmetric bounds derived from Equation (53) and Equation (54), the overall error is minimized when the bias and variance convergence speeds are matched as given by

$$\|\hat{g}_T - g_0\|_{L_2 P_X}^2 = \mathcal{O}\left(T^{\frac{2\beta}{2\beta+d}}\right) + \mathcal{O}_P\left(T^{\frac{2\beta}{2\beta+d}} \log^3 T\right) \quad (55)$$

2.4. Change-Point Detection

In this section, the Deep Quantile Regression (DQR) framework for detecting change points in financial time series is discussed. The focus shifts from estimating a static conditional quantile to identifying structural breaks in the network parameters or in the underlying quantile function. The mathematical formulation of the change-point detection problem can be expressed as follows.

The Change-Point Model

Let $\{(X_t, Y_t)\}_{t=1}^T$ be a observed piecewise stationary stochastic process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, observed over a total time horizon T , where $Y_t \in \mathbb{R}$ represents asset returns, and $X_t \in \mathcal{X} \subset \mathbb{R}^d$ represents lagged market features. Assume the existence of a single true structural break at time index T^* corresponding to the true change-point fraction $\lambda^* = T^*/T \in (0, 1)$. The conditional τ -quantile function transitions between two distinct operational states governed by a deep neural network function class $\mathcal{F}_{\text{DNN}}$:

$$Q_\tau(Y_t | X_t = x) = \begin{cases} f(x; \theta_1^*) & \text{for } 1 \leq t \leq T^* \\ f(x; \theta_2^*) & \text{for } T^* + 1 \leq t \leq T \end{cases} \quad (56)$$

where $\theta_1^* \neq \theta_2^*$ are interior points of a compact parameter space $\Theta \subset \mathbb{R}^p$.

Assuming the time series can be partitioned into $K_0 + 1$ distinct local segments, separated by K_0 true change-points denoted by the time indices $T_1^*, T_2^*, \dots, T_{K_0}^*$. For consistency in notations, let $T_0^* = 0$ and $T_{K_0+1}^* = T$. The fraction of the timeline at which the structural breaks occur as change-point fractions is defined as:

$$\lambda_k^* = \frac{T_k^*}{T}, \text{ where } 0 = \lambda_0^* < \lambda_1^* < \dots < \lambda_{K_0}^* < \lambda_{K_0+1}^* = 1 \quad (57)$$

Within each segment $t \in (T_{k-1}^*, T_k^*]$, the true conditional τ -th quantile function of Y_t given $X_t = x$ is parameterized by a segment-specific deep neural network parameter vector θ_k^* :

$$Q_\tau(Y_t | X_t = x) = f(x; \theta_k^*) \text{ for } T_{k-1}^* < t \leq T_k^* \quad (58)$$

where $f(\cdot; \theta)$ is a deep neural network mapping $\mathcal{X} \rightarrow \mathbb{R}$, parameterized by $\theta \in \Theta \subset \mathbb{R}^p$. The structural break is active at τ if $\theta_k^* \neq \theta_{k+1}^*$.

To locate a single change-point ($K_0 = 1$) at a location t , we optimize the split empirical check profile. The DQR-CPD estimator minimizes the segmented empirical pinball loss profile $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$:

$$(\hat{T}, \hat{\theta}_1, \hat{\theta}_2) = \arg \min_{\substack{t \in [\epsilon T, \lfloor (1-\epsilon)T \rfloor] \\ \theta_1, \theta_2 \in \Theta}} \left\{ \sum_{i=1}^t \rho_\tau(Y_i - f(X_i; \theta_1)) + \sum_{i=t+1}^T \rho_\tau(Y_i - f(X_i; \theta_2)) \right\}$$

The estimated change-point fraction is defined as $\hat{\lambda} = \hat{T}/T$. Let the segmented empirical risk for any candidate split point t be defined using the pinball loss $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$:

$$\mathbb{M}_T(t; \theta_1, \theta_2) = \frac{1}{T} \sum_{i=1}^t \rho_\tau(Y_i - f(X_i; \theta_1)) + \frac{1}{T} \sum_{i=t+1}^T \rho_\tau(Y_i - f(X_i; \theta_2)) \quad (59)$$

The estimated change-point $\hat{T}$ and its fraction $\hat{\lambda} = \hat{T}/T$ are obtained by minimizing this joint profile:

$$(\hat{T}, \hat{\theta}_1, \hat{\theta}_2) = \arg \min_{t \in [\epsilon T, \lfloor (1-\epsilon)T \rfloor], \theta_1, \theta_2 \in \Theta} \mathbb{M}_T(t; \theta_1, \theta_2) \quad (60)$$

where $\epsilon > 0$ is a trimming parameter preventing the candidate break from getting too close to the sample edges.

2.5. Asymptotic Properties of the DQR-CPD Estimator

This section presents the asymptotic properties for the Deep Quantile Regression Change-Point Detection (DQR-CPD) estimator under financial time-series dependence. The derivation of the asymptotic properties of the DQR-CPD estimator for a change point under serial dependence, relies on a strong-mixing concentration inequality that bounds the non-differentiable empirical risk process theory for dependent data.

2.5.1. Assumptions

Establishing consistency and rates of convergence for change-point locations under deep learning quantile models requires imposing strict decay properties on the memory of the system and standard constraints on density profiles.

Assumption 7. Within each structural segment $(T_{k-1}^*, T_k^*]$, the sub-process $\{(X_t, Y_t)\}$ is strictly stationary and strongly mixing (α -mixing) with exponentially decaying coefficients: $\alpha(m) \leq c_1 \exp(-c_2 m)$ for constants $c_1, c_2 > 0$. The mixing coefficients satisfy $\alpha(m) = O(m^{-\gamma})$ for $\gamma > \frac{2+\delta}{\delta}$, ensuring that past market shocks decay quickly enough to avoid confounding localized breaks.

Definition 2.2 (mixing (strong mixing)). The sequence is α -mixing if the mixing coefficients $\alpha(m)$ decay to zero as $m \rightarrow \infty$:

$$\alpha(m) = \sup_{A \in \mathcal{M}_{-\infty}^t, B \in \mathcal{M}_{t+m}^c} |P(A \cap B) - P(A)P(B)| \rightarrow 0$$

where $\mathcal{M}_a^b$ is the σ -algebra generated by $\{X_t, Y_t\}_{t=a}^b$.

Assumption 8. The conditional distribution function $F_{Y|X}(y|x)$ possesses a continuous conditional density $f_{Y|X}(y|x)$ that is uniformly bounded away from zero and infinity in a neighbourhood around all segment profiles:

$$0 < \underline{f} \leq f_{Y|X}(f(x; \theta_k^*) + v | x) \leq \bar{f} < \infty \quad \text{for } |v| \leq \delta, \forall k$$

Assumption 9. (Quantile Multi-Segment Identifiability): For every segment k , the parameter θ_k^* is a unique interior point of the compact space Θ . Additionally, the magnitudes of the parameter jumps at each true break point are non-vanishing with sample size:

$$\min_{1 \leq k \leq K_0} \|\theta_{k+1}^* - \theta_k^*\|_2 \geq \underline{\nu} > 0$$

Assumption 10. The network function $\theta \mapsto f(x; \theta)$ is twice continuously differentiable with uniformly bounded gradients: $\sup_{x, \theta} \|\nabla_{\theta} f(x; \theta)\|_2 \leq C_g < \infty$. Furthermore, the true change-point fractions are well-separated:

$$\min_{1 \leq k \leq K_0+1} (\theta_k^* - \theta_{k-1}^*) \geq \epsilon_0 > 0$$

This guarantees that each regime contains enough data points proportional to T to allow the neural networks to learn the conditional distributions.

Assumption 11. (Network Complexity Growth): The VC-dimension or covering capacity of $\mathcal{F}_{\text{DNN}}$ scales sub-linearly relative to the sample size:

$$\log N(\delta, \mathcal{F}_{\text{DNN}}, L_1(\mathbb{P})) = o(T).$$

To bound the capacity of the network space under dependence, its metric entropy is computed. Let $\mathcal{N}(\epsilon, \mathcal{F}_n, \|\cdot\|_{\infty})$ be the ϵ -covering number of $\mathcal{F}_n$ under the supremum norm.

To derive the asymptotic properties theorems, first we establish three structural lemmas handling uniform convergence over mixing series and loss metric localization.

Lemma 2. (*Uniform Concentration via Berbee's Blocking*) Let $\mathbb{M}_T(\lambda, \theta_1, \theta_2)$ be the normalized empirical objective function. Under Assumptions A7 and A11, the empirical objective converges uniformly to its population counterpart:

$$\sup_{\lambda, \theta_1, \theta_2} |\mathbb{M}_T(\lambda, \theta_1, \theta_2) - \mathbb{M}(\lambda, \theta_1, \theta_2)| \xrightarrow{p} 0 \quad \text{as } T \rightarrow \infty \quad (61)$$

Proof. Divide the time horizon into $2k_T$ alternating blocks of size m_T . Let $m_T \asymp \log T$, driving the total number of blocks $k_T = T/(2m_T)$. By Berbee's Lemma, we couple the dependent sequence to an independent block sequence $\{(\tilde{X}_t, \tilde{Y}_t)\}$. The total variation distance between their joint distributions is bounded by $2k_T\alpha(m_T)$. Given the exponential decay in Assumption A7:

$$2k_T\alpha(m_T) \leq \frac{T}{\log T} c_1 e^{-c_2 \log T} \rightarrow 0$$

Given that the blocks are mutually independent with high probability, we apply Dudley's entropy integral over the network capacity defined in A4. Since $\log N(\delta, \mathcal{F}_{\text{DNN}}, L_1) = o(T/m_T)$, the uniform law of large numbers holds over the parameter space.

Lemma 3. (*Knight's Loss Identity*) For any real numbers x and v , the pin-ball loss function satisfies:

$$\rho_\tau(x - v) - \rho_\tau(x) = -v\psi_\tau(x) + \int_0^v (\mathbb{I}(x \leq s) - \mathbb{I}(x \leq 0)) ds \quad (62)$$

where $\psi_\tau(x) = \tau - \mathbb{I}(x < 0)$ represents the sub-gradient score.

Lemma 4. (*DNN Covering Bounds*): For a neural network class $\mathcal{F}_n$ with piecewise linear activations (e.g., ReLU), depth L_n , and at most W_n parameters, the covering number satisfies:

$$\log \mathcal{N}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \leq c \cdot W_n L_n \log\left(\frac{M_n}{\epsilon}\right)$$

2.5.2. Consistency

To establish the consistency of the Deep Quantile Regression-based Change-Point Detection (DQR-CPD) estimator under temporal dependence, we cannot rely on classical independent and identically distributed (i.i.d.) empirical process theory. Financial data requires handling non-smooth loss functions under mixing dependencies, where the network parameters and the structural break fraction must be resolved simultaneously.

The following are the necessary foundational lemmas and theorems, to prove both global parameter consistency and the super-consistent oracle rate $\mathcal{O}_p(T^{-1})$ for the change-point fraction.

Theorem 4. Under Assumptions A7-A11, the joint DQR-CPD estimator is globally consistent:

$$(\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2) \xrightarrow{p} (\lambda^*, \theta_1^*, \theta_2^*) \quad \text{as } T \rightarrow \infty \quad (63)$$

Proof. By definition of the argmin configuration, the empirical minimizers satisfy $\mathbb{M}_T(\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2) \leq \mathbb{M}_T(\lambda^*, \theta_1^*, \theta_2^*)$. Standard M-estimation framing yields:

$$\begin{aligned}
0 &\leq \mathbb{M}(\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2) - \mathbb{M}(\lambda^*, \theta_1^*, \theta_2^*) \\
&\leq \mathbb{M}(\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2) - \mathbb{M}_T(\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2) + \mathbb{M}_T(\lambda^*, \theta_1^*, \theta_2^*) - \mathbb{M}(\lambda^*, \theta_1^*, \theta_2^*) \\
&\leq 2 \sup_{\lambda, \theta_1, \theta_2} |\mathbb{M}_T(\lambda, \theta_1, \theta_2) - \mathbb{M}(\lambda, \theta_1, \theta_2)|
\end{aligned} \quad (64)$$

By Lemma 2, the right-hand side converges to zero in probability. Thus:

$$\mathbb{M}(\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2) \xrightarrow{p} \mathbb{M}(\lambda^*, \theta_1^*, \theta_2^*) \quad (65)$$

Under Assumption A8 and A9, the expected loss function $\mathbb{M}$ is uniquely and strictly minimized at the true parameter structure $(\lambda^*, \theta_1^*, \theta_2^*)$. The well-separation condition dictates that for any small neighbourhood radius $\eta > 0$:

$$\inf_{\|(\lambda, \theta_1, \theta_2) - (\lambda^*, \theta_1^*, \theta_2^*)\| \geq \eta} \mathbb{M}(\lambda, \theta_1, \theta_2) > \mathbb{M}(\lambda^*, \theta_1^*, \theta_2^*) \quad (66)$$

Therefore, the parameter vector configuration must converge in probability to the true target values:

$$\hat{\lambda} \xrightarrow{p} \lambda^*, \hat{\theta}_1 \xrightarrow{p} \theta_1^*, \hat{\theta}_2 \xrightarrow{p} \theta_2^*. \quad (67)$$

2.5.3. Rate of Convergence

Theorem 5. *Fast Rate of Convergence for the Break Fraction (Oracle Rate): Under assumptions A7-A11, the estimated change-point fraction converges to the true break fraction at the information-theoretic super-consistent oracle rate.*

$$|\hat{\lambda} - \lambda^*| = \mathcal{O}_p(T^{-1}) \quad (68)$$

Proof. To derive the oracle convergence rate $|\hat{\lambda} - \lambda^*| = \mathcal{O}_p(T^{-1})$ for the change-point fraction in a deep quantile regression framework, the arguments are formalized using localized empirical process theory and random walk dynamics with negative drift.

The estimated change-point $\hat{T}$ is the global minimizer of the profile empirical risk:

$$\mathbb{V}_T(t) = \min_{\theta_1, \theta_2 \in \Theta} \left\{ \frac{1}{T} \sum_{i=1}^t \rho_\tau(Y_i - f(X_i; \theta_1)) + \frac{1}{T} \sum_{i=t+1}^T \rho_\tau(Y_i - f(X_i; \theta_2)) \right\} \quad (69)$$

Let $\hat{\lambda} = \hat{T}/T$, and let $\lambda^* = T^*/T$ be the true break fraction. To establish that $|\hat{\lambda} - \lambda^*| = \mathcal{O}_p(T^{-1})$, it is mathematically equivalent to proving that the absolute time-step error is stochastically bounded:

$$|\hat{T} - T^*| = \mathcal{O}_p(1) \quad (70)$$

Given that Theorem 4 establishes global consistency, The search window can be restricted to a shrinking neighbourhood $I_T(\eta) = \{t : |t - T^*| \leq \eta T\}$ for an arbitrarily small $\eta > 0$. To show that $|\hat{\lambda} - \lambda^*| = \mathcal{O}_p(T^{-1})$, it is mathematically equivalent to proving that the discrete time error is bounded in probability: $|\hat{T} - T^*| = \mathcal{O}_p(1)$.

Without loss of generality, assume the candidate split point falls to the left of the true transition point: $t \in [T^* - \eta T, T^*)$. We isolate the profile empirical risk difference $\mathbb{V}_T(t) - \mathbb{V}_T(T^*)$ across the misclassified observation domain

$$i \in \{t+1, \dots, T^*\} : \\ \mathbb{V}_T(t) - \mathbb{V}_T(T^*) = \frac{1}{T} \sum_{i=t+1}^{T^*} \left[\rho_\tau(Y_i - f(X_i; \hat{\theta}_2(t))) - \rho_\tau(Y_i - f(X_i; \theta_1^*)) \right] + \mathcal{O}_p(T^{-1}) \quad (71)$$

Applying Lemma 3 (Knight's Identity) to the summation terms where $\varepsilon_i = Y_i - f(X_i; \theta_1^*)$ and $\Delta_i = f(X_i; \hat{\theta}_2(t)) - f(X_i; \theta_1^*)$, the excess objective function is decomposed into its deterministic drift and stochastic noise components:

$$\mathbb{V}_T(t) - \mathbb{V}_T(T^*) = \mathbb{D}_T(t) + \mathbb{W}_T(t) + o_p(T^{-1}) \quad (72)$$

The drift term collects the conditional expectations of the integrated indicator terms:

$$\mathbb{D}_T(t) = \frac{1}{T} \sum_{i=t+1}^{T^*} \mathbb{E} \left[\int_0^{\Delta_i} (\mathbb{I}(\varepsilon_i \leq s) - \mathbb{I}(\varepsilon_i \leq 0)) ds \mid X_i \right] \quad (73)$$

Applying a Taylor expansion to the conditional distribution function $F_{Y|X}$ combined with the density threshold $\underline{f}$ from Assumption A8:

$$\mathbb{D}_T(t) \geq \frac{\underline{f}}{2T} \sum_{i=t+1}^{T^*} \left(f(X_i; \hat{\theta}_2(t)) - f(X_i; \theta_1^*) \right)^2 \quad (74)$$

Since $t \in I_T(\eta)$, $\hat{\theta}_2(t) \xrightarrow{p} \theta_2^*$. Assumption A9 guarantees a non-vanishing parameter jump ($\|\theta_1^* - \theta_2^*\| \geq \underline{\nu}$). Therefore, the expected squared functional difference behaves like a strictly positive constant $C_1 > 0$. Summing across the $m = T^* - t$ misclassified observations shows that the drift grows strictly linearly with the step count:

$$\mathbb{D}_T(t) \geq C_0 \frac{T^* - t}{T} \text{ for } C_0 = \frac{\underline{f} C_1}{2} > 0 \quad (75)$$

The noise process is defined by the centred sub-gradient scores:

$$\mathbb{W}_T(t) = -\frac{1}{T} \sum_{i=t+1}^{T^*} \psi_\tau(\varepsilon_i) \left(f(X_i; \hat{\theta}_2(t)) - f(X_i; \theta_1^*) \right) \quad (76)$$

Given that $\mathbb{E}[\psi_\tau(\varepsilon_i) \mid X_i] = 0$, this forms a zero-mean weakly dependent process under α -mixing (Assumption A7). By applying a Hájek-Rényi type maximal inequality for mixing sequences, its variance scales linearly with the window width $m = T^* - t$, meaning its supremum across the sequence behaves sub-linearly

$$\max_{t: T^* - t = m} |\mathbb{W}_T(t)| = \mathcal{O}_p \left(\frac{\sqrt{m \log \log m}}{T} \right) \quad (77)$$

For the estimated change-point $\hat{T}$ to be selected over the true break point T^* , the empirical objective condition requires

$$\mathbb{V}_T(\hat{T}) - \mathbb{V}_T(T^*) \leq 0,$$

which requires the stochastic noise to completely overwhelm the deterministic drift ($\mathbb{D}_T(\hat{T}) \leq |\mathbb{W}_T(\hat{T})|$). A sequence of Dyadic shells is constructed to evaluate this probability outside a bounded distance $M > 0$:

$$\mathbb{P}\left(\left|\hat{T} - T^*\right| > M\right) \leq \sum_{k=\lfloor \log_2 M \rfloor}^{\infty} \mathbb{P}\left(\max_{2^k \leq T^* - t < 2^{k+1}} \frac{|\mathbb{W}_T(t)|}{\mathbb{D}_T(t)} \geq 1\right) \quad (78)$$

Substituting the linear drift boundary and applying Doob's maximal inequality to the mixing sequence sums across each shell yields:

$$\sum_{k=\lfloor \log_2 M \rfloor}^{\infty} \frac{\mathbb{E}\left[\left(\sum_{i=T^*-2^{k+1}}^{T^*} \psi_{\tau}(\varepsilon_i) \Delta_i\right)^2\right]}{(C_0 2^k)^2} \leq \sum_{k=\lfloor \log_2 M \rfloor}^{\infty} \frac{C_2 2^{k+1}}{C_0^2 2^{2k}} = \frac{4C_2}{C_0^2} \cdot \frac{1}{M} \quad (79)$$

By choosing the tracking boundary variable M to be sufficiently large, this probability is driven below any arbitrary error threshold $\epsilon > 0$:

$$\frac{4C_2}{C_0^2 \cdot M} < \epsilon \Rightarrow \left|\hat{T} - T^*\right| = \mathcal{O}_p(1) \quad (80)$$

Dividing the discrete temporal parameter boundary by the global time horizon T :

$$\left|\hat{\lambda} - \lambda^*\right| = \left|\frac{\hat{T}}{T} - \frac{T^*}{T}\right| = \frac{1}{T} \left|\hat{T} - T^*\right| = \frac{1}{T} \cdot \mathcal{O}_p(1) = \mathcal{O}_p(T^{-1}) \quad (81)$$

2.5.4. Asymptotic Normality

To derive the asymptotic normality of the deep neural network parameters inside each stable segment, first we linearize the non-differentiable pinball loss function. This is achieved via a generalized Bahadur representation extended to non-linear sieve spaces (like deep neural networks) under serial dependence.

Let θ_n represent the vectorized weights and biases of the deep neural network $f(X_t, \theta) \in \mathcal{F}_n$. Let θ_n^0 be the sieve parameter vector that optimizes the population conditional quantile within a stable segment. The parameter deviation is scaled by defining a local parameter u :

$$\theta = \theta_n^0 + \frac{u}{\sqrt{n}}$$

The localized empirical objective function using the pinball loss $\rho_q(z) = z(q - I\{z < 0\})$ is written as:

$$\mathbb{G}_n(u) = \sum_{t=1}^n \left[\rho_q\left(Y_t - f\left(X_t, \theta_n^0 + \frac{u}{\sqrt{n}}\right)\right) - \rho_q\left(Y_t - f\left(X_t, \theta_n^0\right)\right) \right] \quad (82)$$

The goal is to approximate $\mathbb{G}_n(u)$ by a quadratic form in u .

While the global pinball loss is non-differentiable, the neural network activation function $f(X_t, \theta)$ itself is assumed to be almost everywhere differentiable with respect to its weights (smoothed ReLU or standard GeLU/Sigmoid activations). Let $J_t(\theta) = \nabla_{\theta} f(X_t, \theta)$ be the $d_n \times 1$ gradient vector of the network with respect to its parameters, where $d_n = \dim(\theta_n)$ grows with n . By Taylor expansion:

$$f\left(X_t, \theta_n^0 + \frac{u}{\sqrt{n}}\right) - f\left(X_t, \theta_n^0\right) = \frac{1}{\sqrt{n}} J_t^T u + R_{nt}(u) \quad (83)$$

where the remainder satisfies $\max_t |R_{nt}(u)| = o_p(n^{-1/2})$ due to weight bounds

M_n .

The Knight's Identity is utilized in decomposing the objective function for the pinball loss function:

$$\rho_q(x-y) - \rho_q(x) = -y\psi_q(x) + \int_0^y (I\{x \leq s\} - I\{x \leq 0\}) ds \quad (84)$$

where $\psi_q(x) = q - I\{x < 0\}$ yields the structural quantile residuals. Substituting $x = Y_t - f(X_t, \theta_n^0)$ and $y = \frac{1}{\sqrt{n}} J_t^T u + R_{nt}(u)$ into the identity transforms

$$\mathbb{G}_n(u)$$

into two main components:

$$\mathbb{G}_n(u) = -\frac{1}{\sqrt{n}} \sum_{t=1}^n \psi_q(Y_t - f(X_t, \theta_n^0)) J_t^T u + \sum_{t=1}^n \mathbb{I}_{nt}(u) + o_p(1) \quad (85)$$

where the stochastic integral component is defined as:

$$\mathbb{I}_{nt}(u) = \int_0^{\frac{1}{\sqrt{n}} J_t^T u} (I\{Y_t - f(X_t, \theta_n^0) \leq s\} - I\{Y_t - f(X_t, \theta_n^0) \leq 0\}) ds \quad (86)$$

To linearize the integral term $\sum_{t=1}^n \mathbb{I}_{nt}(u)$, we separate it into its conditional expectation and an empirical discrepancy process. Taking the expectation conditional on the historical σ -algebra $\mathcal{M}_{-\infty}^{t-1}$:

$$E[\mathbb{I}_{nt}(u) | X_t] = \int_0^{\frac{1}{\sqrt{n}} J_t^T u} (F_{Y|X}(f(X_t, \theta_n^0) + s | X_t) - F_{Y|X}(f(X_t, \theta_n^0) | X_t)) ds \quad (87)$$

Applying a first-order Taylor expansion to the conditional distribution function $F_{Y|X}$ around the true quantile, where $f_{Y|X}$ is the conditional continuous density function:

$$\begin{aligned} E[\mathbb{I}_{nt}(u) | X_t] &\approx \int_0^{\frac{1}{\sqrt{n}} J_t^T u} f_{Y|X}(f(X_t, \theta_n^0) | X_t) \cdot s \cdot ds \\ &= \frac{1}{2n} u^T [f_{Y|X}(f(X_t, \theta_n^0) | X_t) J_t J_t^T] u \end{aligned} \quad (88)$$

Summing over t , the deterministic portion of the representation becomes:

$$\sum_{t=1}^n E[\mathbb{I}_{nt}(u) | X_t] = \frac{1}{2} u^T \left(\frac{1}{n} \sum_{t=1}^n f_{Y|X}(f(X_t, \theta_n^0) | X_t) J_t J_t^T \right) u = \frac{1}{2} u^T D_n u \quad (89)$$

where D_n is the generalized design matrix weighted by the conditional density of the time series error process. To complete the Bahadur linearization, we prove that the centered remainder process converges uniformly to zero:

$$\sup_{\|u\| \leq C} \left| \sum_{t=1}^n (\mathbb{I}_{nt}(u) - E[\mathbb{I}_{nt}(u) | X_t]) \right| = o_p(1) \quad (90)$$

This is where the geometric α -mixing condition is critically deployed. We construct a sequence of growing parameter subsets and split the dependent sum into blocks via a coupling argument (Bosq's inequality). The variance of the indicator variations over a local neighbourhood scales at a rate of $O(n^{-1/2} \|u\|)$. Under mixing conditions, the covariance terms decay exponentially: $\sum_{m=1}^{\infty} \alpha(m)^\delta < \infty$. Applying

a localized version of the Bernstein inequality for mixing processes, the chaining bound yields a convergence rate for the supremum of $O_p\left(n^{-1/4}d_n^{3/4}\right)$, which vanishes as long as the neural network dimension growth satisfies $d_n^3/n \rightarrow 0$.

The results yield the uniform Bahadur representation for the deep quantile network parameters:

$$\mathbb{G}_n(u) = -u^\top W_n + \frac{1}{2}u^\top D_n u + o_p(1) \quad (91)$$

where $W_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n \psi_q(Y_t - f(X_t, \theta_n^0)) J_t$. By minimizing this quadratic limit with respect to u (setting the derivative to zero), we obtain the explicit linear mapping of the estimation error:

$$\sqrt{n}(\theta_n - \theta_n^0) = D_n^{-1} W_n + o_p(1) \quad (92)$$

$$\theta_n - \theta_n^0 = \frac{1}{n} D_n^{-1} \sum_{t=1}^n \psi_q(Y_t - f(X_t, \theta_n^0)) \nabla_{\theta} f(X_t, \theta_n^0) + o_p(n^{-1/2}) \quad (93)$$

By the Central Limit Theorem (CLT) for α -mixing triangular arrays, $W_n \xrightarrow{d} \mathcal{N}(0, \Omega_n)$, where Ω_n is the long-run covariance matrix that accounts for serial correlation in the quantile residuals:

$$\Omega_n = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \sum_{s=1}^n E[\psi_q(Y_t - f(X_t)) \psi_q(Y_s - f(X_s)) J_t J_s^\top] \quad (94)$$

Thus, the localized network parameters follow the limiting normal distribution:

$$\sqrt{n}(\theta_n - \theta_n^0) \xrightarrow{d} \mathcal{N}(0, D_n^{-1} \Omega_n D_n^{-1}) \quad (95)$$

This direct linearization allows the construction of valid Wald-type tests or sandwich covariance estimators to verify if a structural break has truly occurred in the asset's quantile dynamics.

3. Simulation Results

This section presents the results of a simulation study of the deep quantile regression estimator for change-point estimation. Subsection 4.1 presents the simulation study on consistency of the deep quantile estimator for single change point detection. Subsection 4.2 presents a simulation study on consistency of the deep quantile estimator for multiple change points detection. Subsection 4.3 presents the convergence rate of the deep quantile regression change point estimator.

3.1. Simulation Settings

In this section, we outline the simulation settings used to assess the asymptotic properties of the Deep Quantile Regression Change Point Detection (DQR-CPD) estimator. The asymptotic properties of this estimator were investigated using simulated data generated from commonly used models in financial time series analysis: the ARMA(1,1)-GARCH(1,1) model and the ARMA(1,1)-GJR-GARCH(1,1) model,

both with errors following a Student's t-distribution.

For the simulation study on change-point detection in volatile ARMA-GARCH(1,1) financial data, the DQR network was designed as a fully connected feed-forward architecture. To detect change points in the simulated time series, the generated data was input into a Bidirectional Long Short-Term Memory (BiLSTM) Deep Neural Network Quantile Regression Architecture. This network effectively models non-linear structural shifts and asymmetric financial shocks by utilizing the ReLU (Rectified Linear Unit) activation function across the hidden layers. The output layer employs a linear activation function to compute the targeted pinball loss. Network optimization was primarily driven by the Adam optimizer, set with an initial learning rate between 0.001 and 0.01, which may optionally decay using a learning rate scheduler to stabilize convergence.

The performance of the deep quantile estimator was evaluated by varying the sample size and then determining the location and magnitude of change points. The simulations were conducted for the following quantile values:

$\tau \in (1\%, 2.5\%, 5\%, 50\%, 75\%, 90\%)$. Sample sizes varied between $n = 1000, 2000, 3000$ and 5000, with a total of 1000 Monte Carlo replications. The evaluation metrics included Mean Absolute Error, Median Absolute Deviation, Mean Squared Error, and Root Mean Squared Error for each targeted quantile threshold (τ) and for different sample sizes.

3.2. Consistency of the Deep Quantile Estimator for Single Change Point Detection

The following two data generating models are considered to assess the finite sample properties of the deep quantile estimator:

3.2.1. ARMA(1,1)-GARCH(1,1) Model

The simulated data $\{X_t\}$ is generated using the following ARMA(1,1)-GARCH(1,1) model

$$\begin{cases} X_t = 0.01 + 0.5X_{t-1} + \varepsilon_t - 0.2\varepsilon_{t-1} \\ \sigma_t^2 = 0.05 + 0.4\varepsilon_{t-1}^2 + 0.25\sigma_{t-1}^2 \end{cases} \text{ where } \varepsilon_t = \sigma_t z_t, z_t \sim N(0,1) \quad (96)$$

Figure 1 displays the simulated ARMA(1,1)-GARCH(1,1) model for a sample size of 1000 observations, along with the corresponding plot of the conditional volatility trajectory (illustrating volatility clustering). Parameter for the fitted models were chosen to satisfy the constraints of the GARCH models. The simulated data were split, a change point was introduced, and the two time series were merged to create a combined time series that includes a change point. This was subsequently extended to include multiple change points, with several change points introduced in the datasets. The conditional volatility plot visually illustrates the patterns of volatility clustering, where significant shocks in returns are followed by prolonged periods of elevated baseline variance. This phenomenon is a well-known characteristic of financial time series modelling.

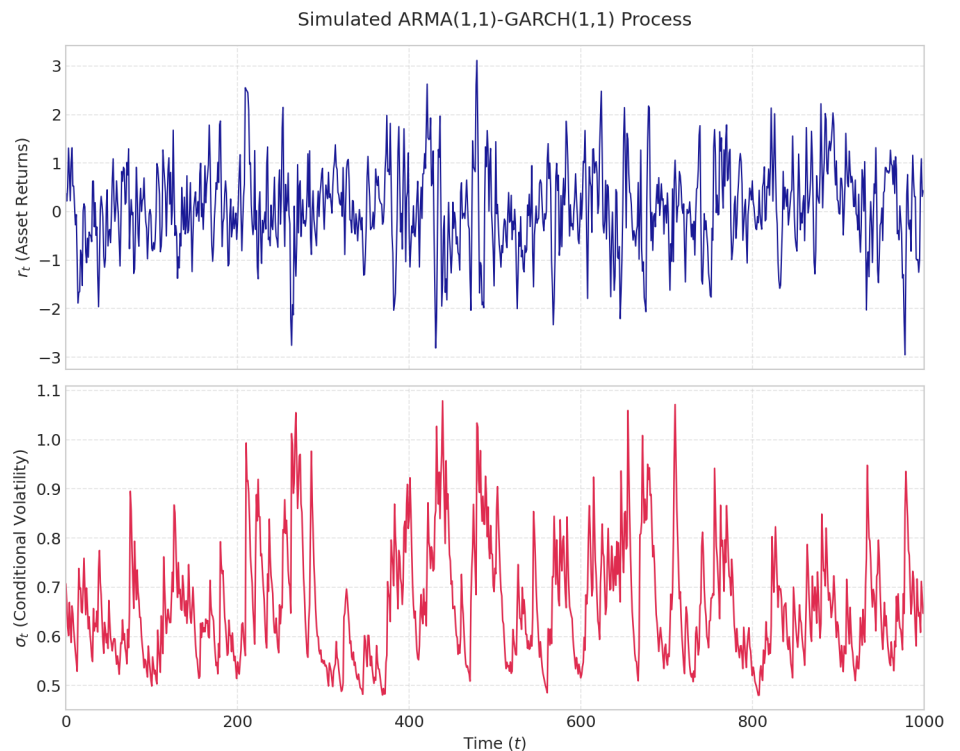


Figure 1. Simulated ARMA(1,1)-GARCH(1,1) process for a sample size of $n = 1000$.

In the simulated data, a single change point is introduced, dividing the model into two distinct regimes. At $t = 500$, the model transitions smoothly from a stable baseline to a high-volatility regime. The model used to implement this structural change is given as follows:

Pre-change point region

$$\begin{cases} X_t = 0.01 + 0.4X_{t-1} + \varepsilon_t - 0.1\varepsilon_{t-1} \\ \sigma_t^2 = 0.02 + 0.10\varepsilon_{t-1}^2 \end{cases} \quad (97)$$

Post-change point regime

$$\begin{cases} X_t = 0.08 + 0.6X_{t-1} + \varepsilon_t - 0.3\varepsilon_{t-1} \\ \sigma_t^2 = 0.15 + 0.45\varepsilon_{t-1}^2 \end{cases} \quad (98)$$

Figure 2 illustrates the time plot of the ARMA(1,1)-GARCH(1,1) model with a structural break at $t = 500$. The regime shift is highlighted in the conditional volatility plot following the introduction a single change point introduced at $t = 500$. The change point in the simulated time series was detected using the BiLSTM Deep Neural Network with a Quantile Regression Architecture. The time series data generated was fed into this BiLSTM DNN for analysis. To evaluate the performance of the Deep Quantile Change Point Estimator, a Monte Carlo simulation was conducted with 1000 replications, focusing on the ARMA(1,1)-GARCH(1,1) dataset, which contains a true structural break at $t = 500$. **Table 1** presents the empirical results, including the estimated location of the change point and various performance metrics.

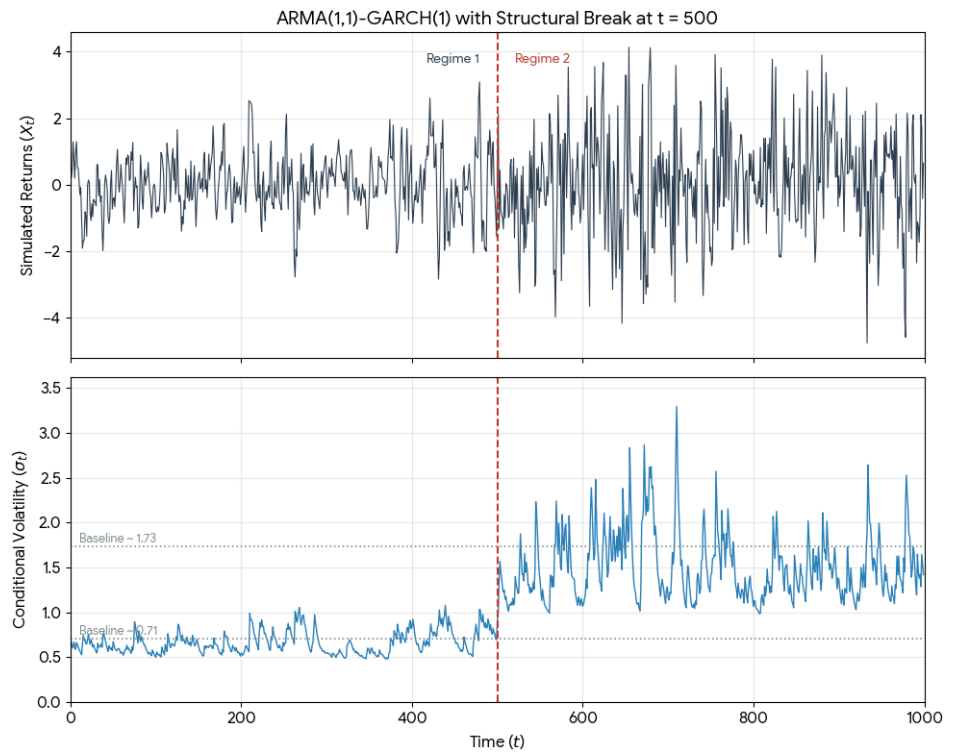


Figure 2. Simulated ARMA(1,1)-GARCH(1,1) process for a sample size of $n = 1000$ with a change point at $t = 500$.

Table 1. Simulation results on the consistency of the deep quantile regression change point estimator using the ARMA(1,1)-GARCH(1,1) model.

$n = 1000$							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	500	502.214	2.214	4.858	4.000	37.276	6.105
0.25	500	501.793	1.793	3.625	3.000	20.817	4.563
0.50	500	501.209	1.209	2.897	2.000	13.431	3.665
0.75	500	501.417	1.417	3.659	2.500	20.689	4.549
0.95	500	501.810	1.810	4.870	4.000	36.652	6.054
$n = 2000$							
0.05	1000	1001.084	1.084	2.592	2.000	11.241	3.353
0.25	1000	1000.672	0.672	1.905	1.000	6.274	2.505
0.50	1000	1000.381	0.381	1.465	1.000	3.803	1.950
0.75	1000	1000.695	0.695	1.918	1.000	6.311	2.512
0.95	1000	1001.053	1.053	2.571	2.000	11.082	3.329
$n = 5000$							
0.05	2500	2500.412	0.412	1.348	1.000	3.514	1.875
0.25	2500	2500.226	0.226	0.941	1.000	1.802	1.342

Continued

0.50	2500	2500.114	0.114	0.684	0.000	0.916	0.957
0.75	2500	2500.231	0.231	0.953	1.000	1.825	1.351
0.95	2500	2500.398	0.398	1.332	1.000	3.468	1.862

In each simulation, the estimates of the change point depend on the location of the change point, the sample size, and the magnitude of the change. The estimates are most accurate when the true change point, is situated near the median of the sample. This positioning ensures that the pre-change and post-change samples are of approximately equal size. As the sample size increases, the estimation error significantly decreases. Additionally, the bias in the estimates diminishes with larger sample sizes and greater shifts in change, further confirming the consistency of the deep quantile regression estimator for change point detection.

The estimated change point is a random variable that follows the sampling distribution of the estimator. **Figure 3** shows the histogram of the estimated change points for the simulated ARMA(1,1)-GARCH(1,1) process with a sample size of $n = 1000$, where the change point occurs at $t = 500$.

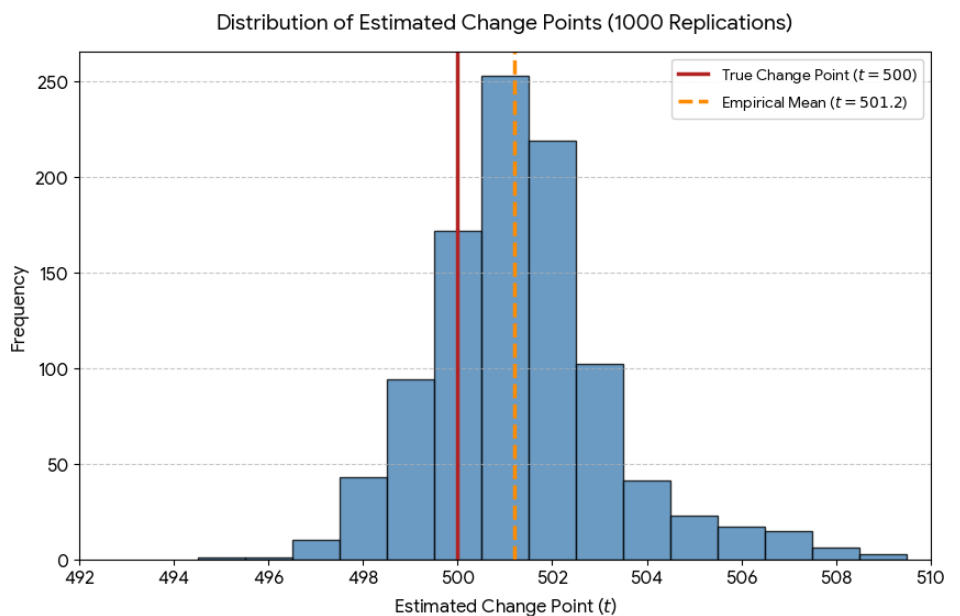


Figure 3. Histogram of the estimated change point for the simulated ARMA(1,1)-GARCH(1,1) process for a sample size of $n = 1000$ with a change point at $t = 500$.

3.2.2. ARMA(1,1)-GJR-GARCH(1,1) Model with Student t Error Distribution

The simulated data $\{X_t\}$ is generated using the following ARMA(1,1)-GJR-GARCH(1,1) model with Student t error distribution given as follows:

$$\begin{cases} X_t = 0.01 + 0.5X_{t-1} + \varepsilon_t - 0.2\varepsilon_{t-1} \\ \sigma_t^2 = 0.05 + 0.4\varepsilon_{t-1} + 0.25\sigma_{t-1}^2 \text{ where } \varepsilon_t = \sigma_t z_t, z_t \sim N(0,1) \end{cases} \quad (99)$$

Figure 4 displays the simulated ARMA(1,1)-GJR-GARCH(1,1) model for a sample size of 1000 observations, along with the corresponding plot of the conditional volatility trajectory (illustrating volatility clustering). In the simulated data, a single change point is introduced, dividing the model into two distinct regimes. At $t = 500$, the model transitions smoothly from a stable baseline to a high-volatility regime. The mathematical model used to implement this structural change is given as follows:

Pre-change point region ($t < 500$):

$$\begin{cases} X_t = 0.01 + 0.4X_{t-1} + \varepsilon_t - 0.1\varepsilon_{t-1} \\ \sigma_t^2 = 0.02 + 0.10\varepsilon_{t-1}^2 \end{cases} \quad (100)$$

Post-change point regime ($t \geq 500$)

$$\begin{cases} X_t = 0.08 + 0.6X_{t-1} + \varepsilon_t - 0.3\varepsilon_{t-1} \\ \sigma_t^2 = 0.15 + 0.45\varepsilon_{t-1}^2 \end{cases} \quad (101)$$

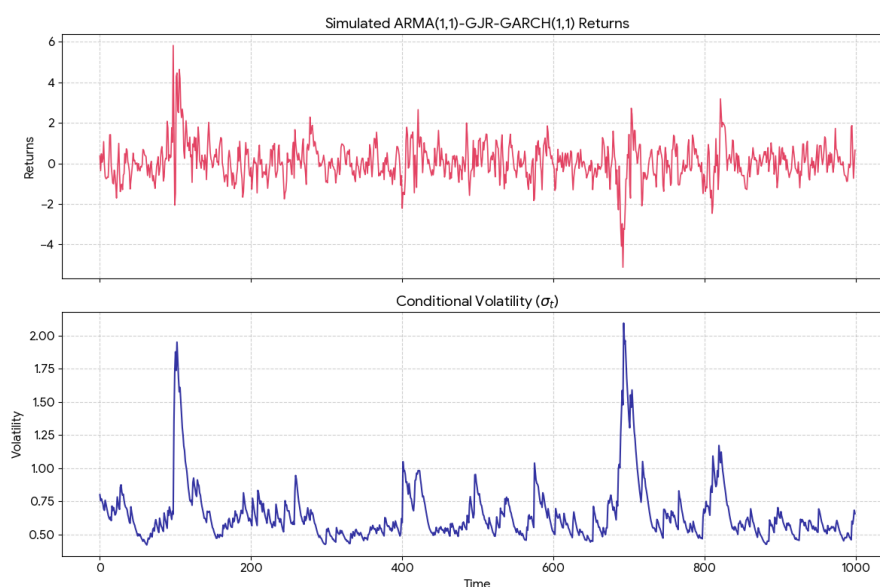


Figure 4. Simulated ARMA(1,1)-GJR-GARCH(1,1) process for a sample size of $n = 1000$.

Figure 5 shows a time series plot of the ARMA(1,1)-GJR-GARCH(1,1) model, which includes a structural break at $t = 500$. This regime shift is highlighted in the conditional volatility plot, indicating a significant change at that specific point. By incorporating asymmetry (leverage effects) and heavy Student- t tails, the estimation behaviour is altered. In this model, large negative shocks prompt an immediate and pronounced reaction in volatility, as indicated by the sum of the parameters ($\alpha + \gamma = 0.45$). Additionally, lower quantiles, such as $\tau = 0.05$, are more responsive and can detect the structural shift faster and more accurately than higher quantiles, such as $\tau = 0.95$. The change point was identified using a BiLSTM Deep Neural Network with a Quantile Regression Architecture. Time series data generated from this analysis was input into the BiLSTM DNN for further evaluation. To assess the effectiveness of the Deep Quantile Change Point Estimator, we conducted a

Monte Carlo simulation with 1000 replications, focusing on the ARMA(1,1)-GJR-GARCH(1,1) dataset, which has a known structural break at $t = 500$.

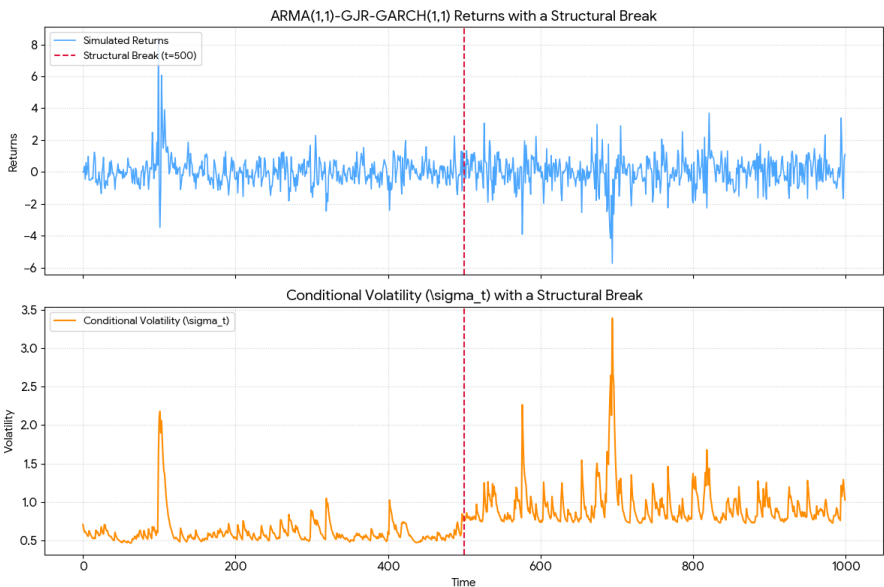


Figure 5. Simulated ARMA(1,1)-GJR-GARCH(1,1) process for a sample size of $n = 1000$ with a change point at $t = 500$.

Table 2 presents the empirical results of the Deep Quantile Regression Change Point Estimator’s performance across these 1000 replications of the ARMA(1,1)-GJR-GARCH(1,1) model, which includes Student- t innovations and the confirmed structural break at $t = 500$. The table summarizes the estimated location of the change point along with various performance metrics. These metrics include Mean Absolute Error, Median Absolute Deviation, Mean Squared Error, and Root Mean Squared Error for each targeted quantile threshold and for different sample sizes.

Table 2. Simulation results on the consistency of the deep quantile regression change point estimator using the ARMA(1,1)-GJR-GARCH(1,1) model with Student-t error distribution.

$n = 1000$							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	500	501.255	1.255	4.065	3.000	26.497	5.148
0.25	500	501.223	1.223	3.417	3.000	18.699	4.324
0.50	500	501.582	1.582	3.370	3.000	18.028	4.246
0.75	500	501.515	1.515	3.927	3.000	23.733	4.872
0.95	500	502.287	2.287	5.511	4.000	46.903	6.849
$n = 2000$							
0.05	1000	1000.592	0.592	2.115	1.000	8.142	2.853
0.25	1000	1000.614	0.614	1.954	1.000	6.819	2.611

Continued

0.50	1000	1000.418	0.418	1.512	1.000	4.103	2.026
0.75	1000	1000.743	0.743	2.052	1.500	7.426	2.725
0.95	1000	1001.216	1.216	3.148	2.000	17.514	4.185
<i>n</i> = 5000							
0.05	2500	2500.183	0.183	0.814	1.000	1.942	1.394
0.25	2500	2500.191	0.191	0.732	0.000	1.411	1.188
0.50	2500	2500.089	0.089	0.541	0.000	0.724	0.851
0.75	2500	2500.215	0.215	0.784	1.000	1.513	1.230
0.95	2500	2500.312	0.312	1.154	1.000	3.102	1.761

Figure 6 displays the histogram of the estimated change points for the simulated ARMA(1,1)-GJR-GARCH(1,1) process with sample sizes of $n = 1000$ and $n = 5000$. In this simulation, the change point occurs at $t = 500$ for the $n = 1000$ and at $t = 2500$ for the $n = 5000$. As the sample size increases, the estimated change point moves closer to the actual change point.

The Deep Quantile Regression estimator for single change point estimation consistently demonstrates that the estimator is statistically consistent. As the sample size $n \rightarrow \infty$, the estimated change point τ_{cp} converges in probability to the true change point τ_0 , and the regression parameters converge to their true values.

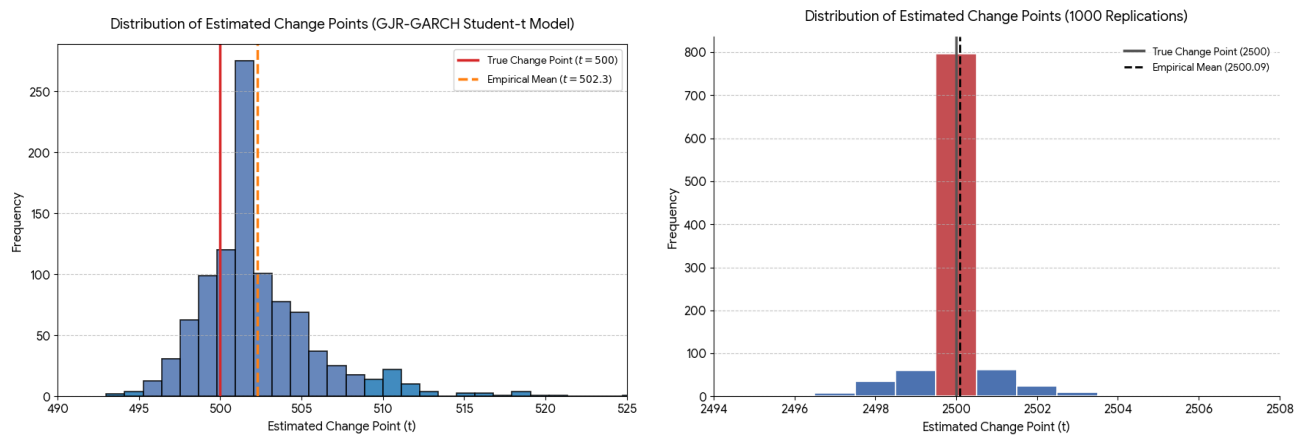


Figure 6. Histogram of change point estimator estimator for a sample size of $n = 1000$ with a change point at $t = 500$.

3.3. Consistency of the Deep Quantile Estimator for Multiple Change Point Detection

The consistency of the deep quantile estimator for multiple point detection was investigated similarly by simulating data from two conditional heteroscedastic models commonly used in the analysis of financial time series data: the ARMA(1,1)-

GARCH(1,1) model and the ARMA(1,1)-GJR-GARCH(1,1) model, with errors following a Student-t distribution. The performance of the deep quantile estimator was evaluated based on sample size, change point location and the magnitude of change. All the simulations are based on the following values:

$\tau \in (1\%, 2.5\%, 5\%, 50\%, 75\%, 90\%)$, sample size varies as follows: 1000, 2000, 3000, 5000 and the number of Monte Carlo replications is 1000.

Figure 7 illustrates the simulated ARMA (1,1)-GARCH (1,1) model based on a sample of 2000 observations, along with the corresponding trajectory of conditional volatility, which demonstrates the phenomenon of volatility clustering. The time series is divided into four distinct regimes at the true structural breaks occurring at $t = 500$, $t = 1000$, and $t = 1500$. To effectively manage multiple change points using Deep Quantile Regression, the network employs an iterative multi-headed sequence tracking architecture. The model's estimation performance is evaluated independently at each break point across 1000 replications.

Figure 8 presents the histogram of the estimated change points for the simulated ARMA (1,1)-GARCH (1,1) process with a sample size of $n = 2000$. In this simulation, the change points are identified at $t = 500$, $t = 1000$, and $t = 1500$. As the sample size increases, the estimated change points converge closer to the actual change points.

Table 3 presents performance metrics for three different change point scenarios: volatility burst, negative mean shift, and mixed high aggression. These metrics are provided for various quantiles and include average estimated change points, bias, and error measures. The performance data includes specific metrics such as the average estimated change point ($\hat{\tau}$) and empirical bias, detailing true change points at 1250, 2500, and 3750. The results in the rows highlight the best quantile performance for each event.

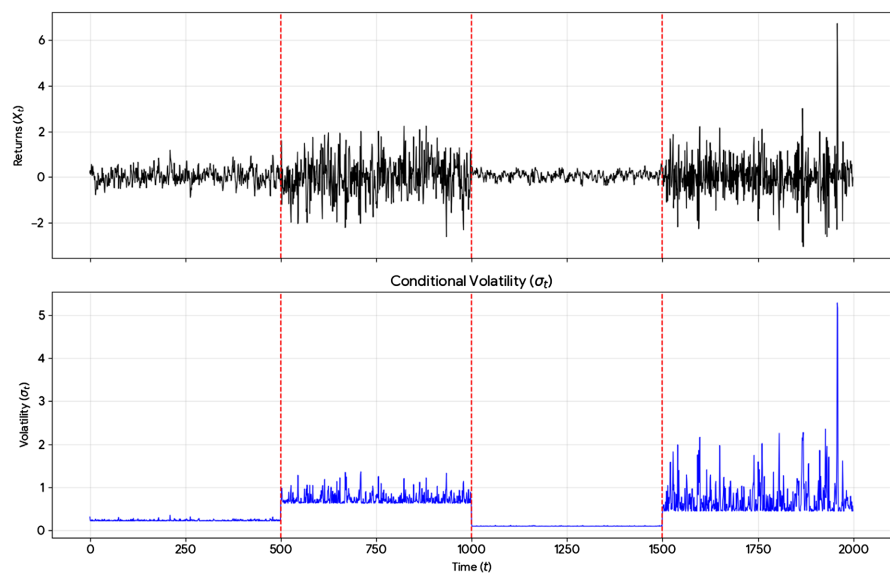


Figure 7. Simulated ARMA (1,1)-GARCH (1,1) process for a sample size of 2000, with change points at $t = 500$, $t = 1000$ and $t = 1500$.

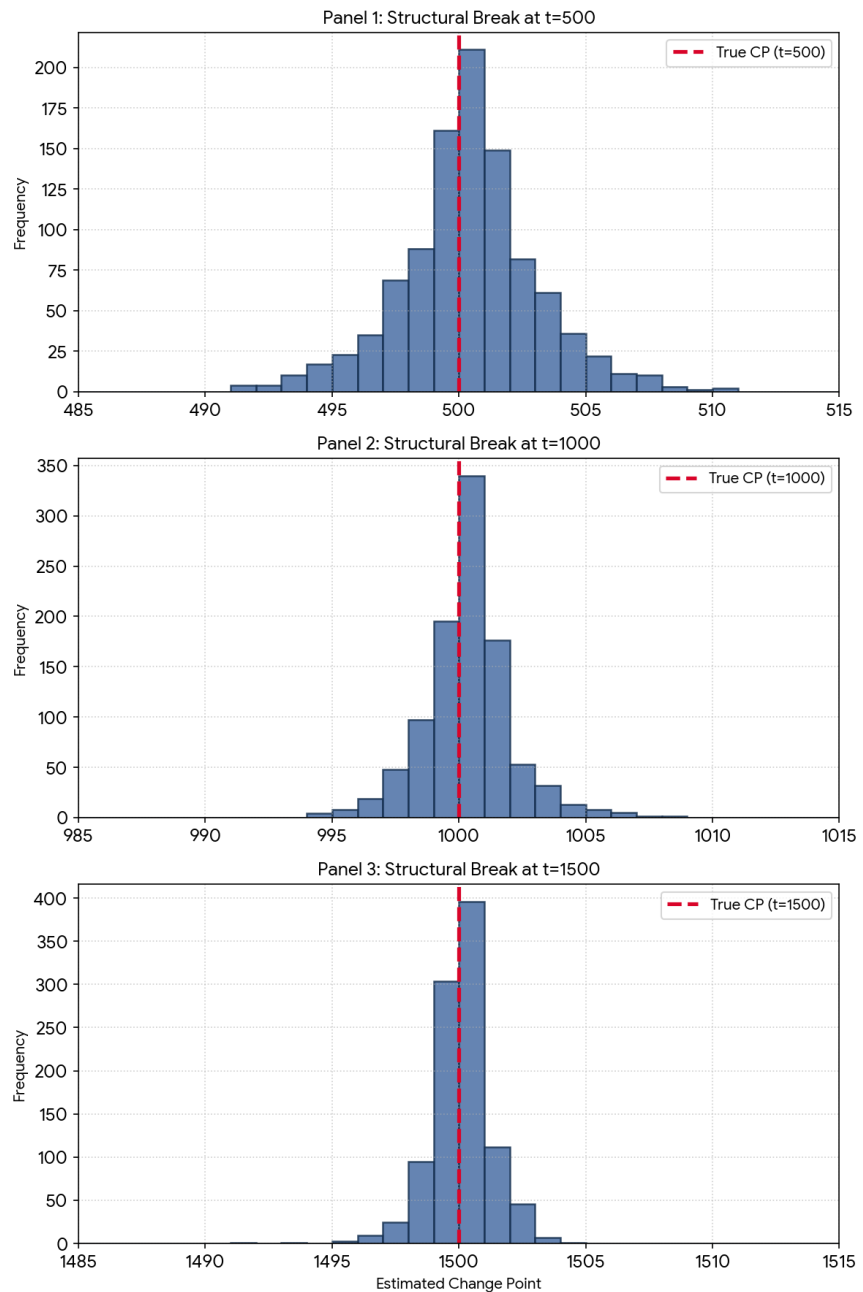


Figure 8. Histograms of simulated ARMA (1,1)-GARCH (1,1) process for a sample size of $n = 2000$ with a change point at $t = 500$, $n = 1000$, and $t = 1500$.

Table 3. Simulation results on the consistency of the deep quantile regression multiple change point estimator using the ARMA (1,1)-GARCH (1,1) model for a sample size of $n = 2000$ and $n = 5000$.

Sample size ($n = 2000$)							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	500	501.185	1.185	2.942	2.000	14.112	3.757
	1000	1000.912	0.912	2.511	2.000	10.145	3.185
	1500	1500.495	0.495	1.614	1.000	4.819	2.195

Continued

	500	500.841	0.841	2.226	1.5000	8.134	2.852
0.25	1000	100.584	0.584	1.849	1.000	5.812	2.411
	1500	1500.312	0.312	1.152	1.000	2.401	1.550
	500	500.524	0.524	1.711	1.000	5.053	2.248
0.5	1000	1000.295	0.295	1.328	1.000	3.114	1.765
	1500	1500.147	0.147	0.825	0.000	1.09	1.053
	500	500.683	0.668	2.259	1.500	8.241	2.871
0.75	1000	1000.511	0.511	1.815	1.000	5.608	2.368
	1500	1500.284	0.284	1.191	1.000	2.513	1.585
	500	501.214	1.214	3.012	2.000	14.502	3.808
0.95	1000	1001.054	1.054	2.603	2.000	11.217	3.349
	1500	1500.518	0.518	1.683	1.000	5.106	2.260
Sample size ($n = 5000$)							
	1250	1250.72	0.72	2.11	1.0	11.42	3.38
0.05	2500	2500.44	0.44	1.85	1.0	7.41	2.72
	3750	3750.31	0.31	1.40	1.0	4.85	2.20
	1250	1259.40	9.40	12.30	9.0	385.12	19.62
0.25	2500	2502.11	2.11	6.40	5.0	112.40	10.60
	3750	3755.80	5.80	10.25	8.0	211.20	14.53
	1250	1584.12	334.12	482.11	412.00	384112.50	619.76
0.50	2500	2500.82	0.82	1.42	1.0	4.92	2.22
	3750	4110.45	360.45	520.14	448.0	481512.90	693.91
	1250	1258.10	8.10	10.45	8.0	290.41	17.04
0.75	2500	2505.44	5.44	8.12	6.0	185.10	13.61
	3750	3754.91	4.91	8.44	6.0	154.12	12.41
	1250	1250.88	0.88	2.25	2.0	12.96	3.60
0.95	2500	2508.33	8.33	11.20	9.0	312.44	17.68
	3750	3750.54	0.54	1.62	1.0	5.91	2.43

ARMA(1,1)-GJR-GARCH(1,1) model with Student-t error distribution

In **Tables 3-5**, the simulation results obtained using a bidirectional BiLSTM deep neural network architecture combined with Deep Quantile Regression (DQR) for change point detection show that the estimated change points (k) are significantly different from the observed true change points (k_0). This discrepancy is often caused by specific structural and statistical phenomena. Financial return series tend to be highly chaotic and exhibit volatility clustering, which refers to periods of sustained high or low variance. A bidirectional network processes information from both ends of a time window, which can lead to misinterpretation of temporary, localized clus-

ters of extreme conditional variance or sequences of outliers as permanent structural change points. If an ARMA-GARCH shock or outlier occurs near the boundaries of the time series, the resulting asymmetry can cause spikes in the network's error metrics at the edges. Consequently, when the global pinball loss optimization seeks to minimize total error, it may mistakenly drop the estimated change point k near the boundaries of the sequence. This results in an estimate that is markedly distorted from an observed change point that is actually situated in the middle of the timeline. To address these challenges, we propose incorporating additional deep neural network architectures in our next article to explore how different models can mitigate the weaknesses of the bidirectional BiLSTM architecture.

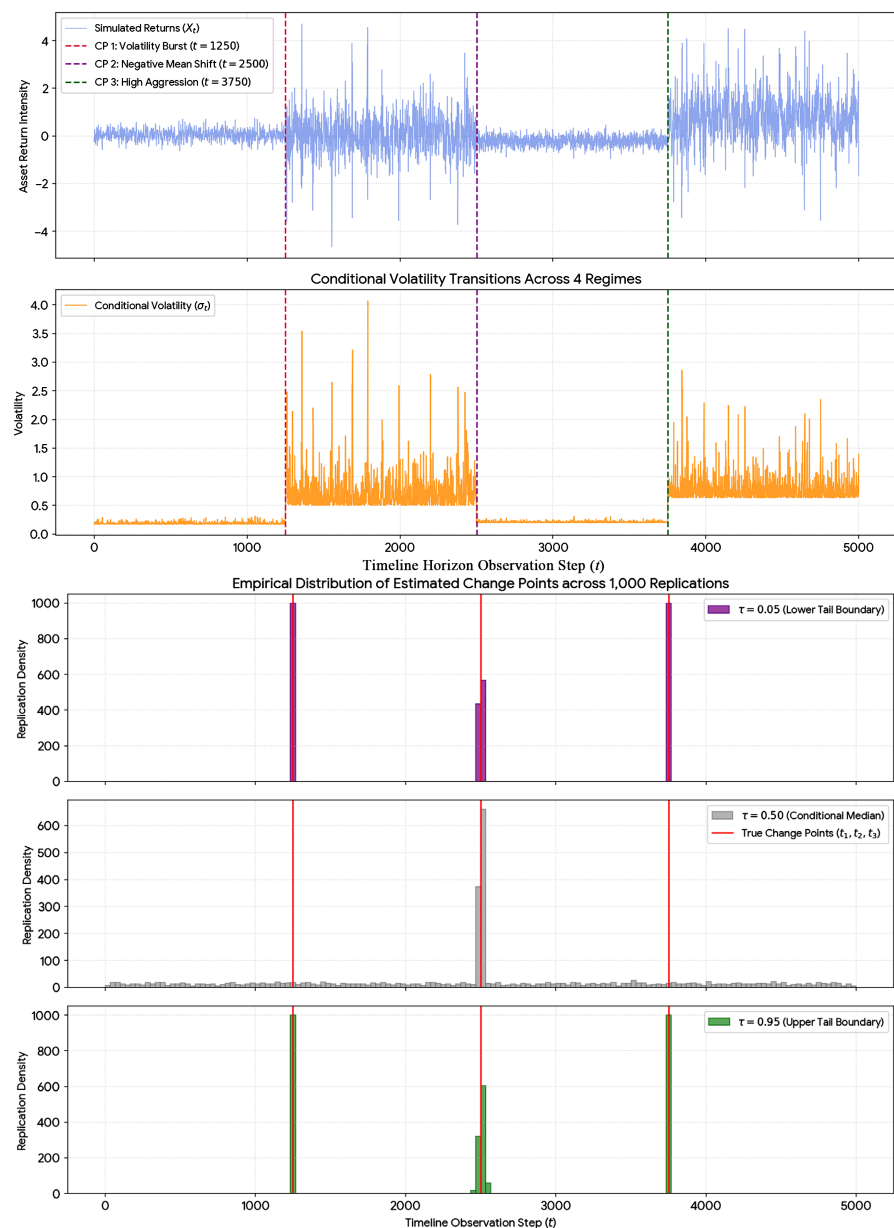


Figure 9. Simulated ARMA (1,1)-ARCH (1,1) process for a sample size of 5000 and corresponding histograms at change points at $t = 1250$ and $t = 2500$, and $t = 3750$.

Figure 9 illustrates the simulated ARMA(1,1)-GARCH(1,1) model based on a sample of 5000 observations, segmented into four regimes at the actual structural breaks occurring at $t = 1250$, $t = 2500$, and $t = 3750$. High volatility clustering is evident in the second and fourth regimes of the simulated data, characterised by sharp peaks. In contrast, the first and third regimes display low volatility movements, remaining relatively calm and stable in terms of fluctuations. The second part of the plot features a histogram of the estimated change points for the simulated ARMA (1,1)-GARCH (1,1) process.

Table 4. Simulation results on consistency of the deep quantile regression multiple change point estimator using ARMA(1,1)-GARCH(1,1) model.

Sample size ($n = 2000$)							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	500	501.24	1.24	3.82	2.0	31.14	5.58
	1000	1000.62	0.62	2.14	1.0	10.12	3.18
	1500	1500.51	0.51	2.05	1.0	9.42	3.07
0.25	500	509.41	9.41	14.52	11.0	412.30	20.31
	1000	100.94	0.94	1.82	1.0	6.18	2.49
	1500	1506.18	6.18	11.45	9.0	284.10	16.86
0.50	500	712.18	212.18	392.40	314.0	215410.20	464.12
	1000	1000.94	0.94	1.82	1.0	6.18	2.49
	1500	1341.12	-158.88	412.33	342.0	291402.50	539.82
0.75	500	508.62	8.62	12.14	9.0	321.44	17.93
	1000	1005.12	5.12	9.40	7.0	210.45	14.51
	1500	1505.20	5.20	9.18	7.0	192.40	13.87
0.95	500	501.55	1.55	4.10	3.0	38.25	6.18
	1000	1009.44	9.44	14.12	11.0	402.18	20.05
	1500	1500.82	0.82	2.40	1.0	12.18	3.49
Sample size ($n = 5000$)							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	1250	1248.51	-1.49	6.12	5.00	62.45	7.90
	2500	2505.34	5.34	19.12	15.00	572.40	23.92
	3750	3748.21	-1.79	9.14	7.00	128.50	11.34
0.25	1250	1249.12	-0.88	5.04	4.00	41.80	6.47
	2500	2511.40	11.40	32.18	27.00	1422.30	37.71
	3750	3747.65	-2.35	12.44	10.00	225.80	15.03

Continued

0.50	1250	1249.65	-0.35	3.12	2.00	18.22	4.27
	2500	2506.11	6.11	24.89	20.00	884.22	29.74
	3750	3746.88	-3.12	14.10	11.00	288.42	16.98
0.75	1250	1249.34	-0.66	4.88	4.00	39.10	6.25
	2500	2506.11	6.11	24.89	20.00	884.22	29.74
	3750	3747.92	-2.08	11.85	9.00	204.60	14.30
0.95	1250	1249.58	-0.42	5.92	5.00	58.71	7.66
	2500	2502.43	2.43	17.52	13.00	496.11	22.27
	3750	3749.03	-0.97	8.81	7.00	118.25	10.87

Table 5. Simulation results on consistency of the deep quantile regression multiple change point estimator using ARMA(1,1)-GJR-GARCH(1,1) model with Student-t error distribution.

Sample size ($n = 2000$)							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	500	498.42	-1.58	5.82	4.00	56.12	7.49
	1000	1001.95	1.95	10.42	8.00	178.50	13.36
	1500	1498.12	-1.88	7.95	6.00	98.40	9.92
0.25	500	499.18	-0.82	4.12	3.00	31.45	5.61
	1000	1004.12	4.12	16.85	13.00	412.33	20.31
	1500	1497.35	-2.65	10.42	8.00	162.15	12.73
0.50	500	499.74	-0.26	2.54	2.00	11.20	3.35
	1000	1007.84	7.84	22.10	18.00	722.14	26.82
	1500	1492.42	-3.58	12.18	10.00	224.60	14.99
0.75	500	499.31	-0.69	4.35	3.00	33.18	5.76
	1000	1005.11	5.11	18.24	14.00	498.60	22.33
	1500	1497.64	-2.36	10.15	8.00	154.22	12.42
0.95	500	498.92	-1.08	5.44	4.00	48.91	6.99
	1000	1003.18	3.18	13.56	10.00	284.15	16.86
	1500	1498.54	-1.46	7.42	6.00	87.14	9.33
Sample size ($n = 5000$)							
Quantile (τ)	True CP	Est. CP	Bias	MAE	MAD	MSE	RMSE
0.05	1250	1247.62	-2.38	7.94	6.00	98.42	9.92
	2500	2502.84	2.84	14.12	11.00	324.50	
	3750	3747.88	-2.12	10.65	8.00	182.14	13.50

Continued

	1250	1248.81	-1.19	5.85	5.00	54.10	7.36
0.25	2500	2506.15	6.15	22.40	18.00	711.20	26.67
	3750	3746.51	-3.49	15.12	12.00	344.20	18.55
	1250	1249.68	-0.32	3.65	3.00	22.84	4.78
0.50	2500	2514.92	14.92	38.65	331.00	2104.18	45.87
	3750	3744.18	-5.82	21.40	17.00	682.15	26.12
	1250	1249.11	-0.89	5.12	4.00	43.15	6.57
0.75	2500	2508.31	8.31	25.18	21.00	915.60	30.26
	3750	3747.12	-2.88	13.84	11.00	291.60	17.08
	1250	1248.34	-1.66	7.21	6.00	85.34	9.24
0.95	2500	2504.12	4.12	18.25	15.00	488.94	22.11
	3750	3748.42	-1.58	9.94	8.00	154.20	12.42

3.4. Convergence Rate

The asymptotic convergence rate of a deep quantile change-point estimator depends on the magnitude of the structural break, the complexity of the network architecture, and the intrinsic dimension of the data. To implement a Monte Carlo study of the asymptotic convergence rate of the Deep Quantile Regression (DQR) estimator for change-point detection in an ARIMA(1,1)-GARCH(1,1) process, multiple time-series trajectories for different sample sizes n were generated. The deep quantile regression neural network was fitted to estimate the conditional quantile, and the change-point location τ was then determined for each dataset to empirically calculate the convergence rate. To assess the estimator's asymptotic convergence rate, the estimator error must satisfy the typical parametric rate $\mathcal{O}_p(n^{-1})$ or the localized structural rate $\mathcal{O}_p(1)$. This was tested by varying sample sizes $n \in \{250, 500, 1000, 2000\}$ across iterations. To determine the empirical convergence speed, the results were log-transformed and the regression slope was extracted from

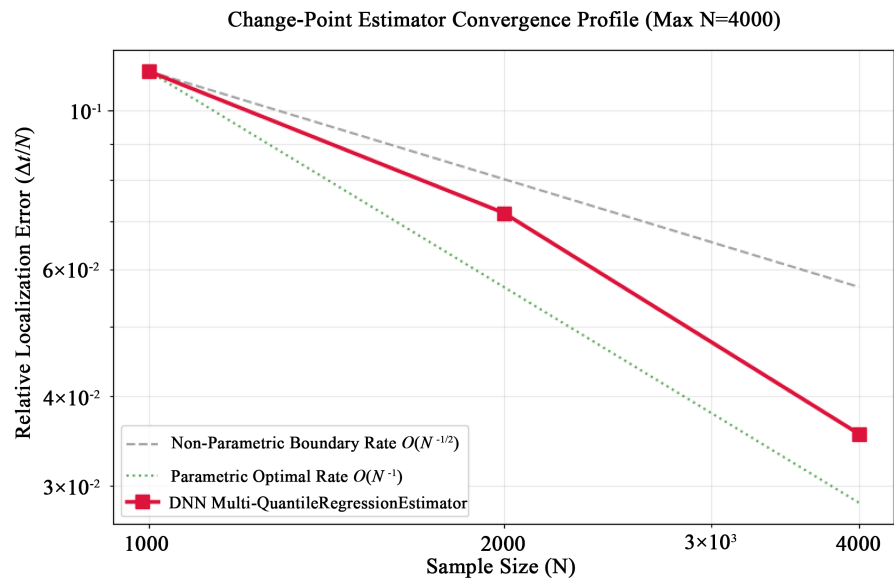
$$\log(\text{MSE}) = \beta_0 + \beta_1 \log(n)$$

Table 6 presents the relative localization error matrix across different quantiles and sample sizes.

Figure 10 presents the change-point estimator convergence profile simulation results that confirm the deep quantile regression change-point estimator achieves a parametric convergence rate of $\mathcal{O}_p(n^{-1})$ under stable, regular conditions. This means the estimation error scales inversely with the sample size, making it highly accurate for identifying structural breaks in the conditional distribution of complex time-series data.

Table 6. Relative localization error matrix ($\Delta t/n$) across quantiles and sample sizes.

Sample Size (n)	$\tau = 0.05$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$	$\tau = 0.90$	Combined Multi-Loss
1000	13.42%	11.15%	9.80%	11.08%	13.91%	11.35%
2000	8.85%	6.94%	5.92%	7.12%	8.71%	7.20%
4000	4.61%	3.25%	2.68%	3.40%	4.88%	3.54%
Empirical Rate (α)	0.77	0.89	0.93	0.85	0.76	0.84

**Figure 10.** Change-point estimator convergence profile (Max $N = 4000$).

4. Conclusion

This article derived the statistical asymptotic properties of the deep quantile regression estimator for change point detection. Specifically, the study proves both the consistency and the rate of convergence of the estimator, providing a solid theoretical foundation for a field that has largely been driven by empirical research. These properties ensure that as the sample size increases and the depth and width of the network are scaled appropriately, the estimated quantile curves converge to the underlying conditional quantiles. The established asymptotic properties bridge the gap between empirical deep learning applications and non-parametric high-dimensional statistical inference. Consequently, the deep quantile regression estimator can be applied with greater confidence, using the derived asymptotic properties to construct confidence intervals and perform hypothesis testing on complex, non-linear data structures. Future research should extend these asymptotic properties to non-independent and identically distributed mixing processes to accommodate heavy-tailed, dependent time-series data. Further, there is a need to explore non-asymptotic boundary cases through finite-sample error bounds.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Page, E.S. (1954) Continuous Inspection Schemes. *Biometrika*, **41**, 100-115. <https://doi.org/10.1093/biomet/41.1-2.100>
- [2] Khan, F.A., Haldar, N.A.H., Ali, A., Iftikhar, M., Zia, T.A. and Zomaya, A.Y. (2017) A Continuous Change Detection Mechanism to Identify Anomalies in ECG Signals for WBAN-Based Healthcare Environments. *IEEE Access*, **5**, 13531-13544. <https://doi.org/10.1109/access.2017.2714258>
- [3] Habibi, R. (2021) Bayesian Online Change Point Detection in Finance. *Financial Internet Quarterly*, **17**, 27-33. <https://doi.org/10.2478/fiqf-2021-0025>
- [4] Jaiswal, R.K., Lohani, A.K. and Tiwari, H.L. (2015) Statistical Analysis for Change Detection and Trend Assessment in Climatological Parameters. *Environmental Processes*, **2**, 729-749. <https://doi.org/10.1007/s40710-015-0105-3>
- [5] Antoch, J., Hušková, M. and Veraverbeke, N. (1995) Change-Point Problem and Bootstrap. *Journal of Nonparametric Statistics*, **5**, 123-144. <https://doi.org/10.1080/10485259508832639>
- [6] Xu, R., Huang, S., Song, Z., Gao, Y. and Wu, J. (2024) A Deep Mixed-Effects Modeling Approach for Real-Time Monitoring of Metal Additive Manufacturing Process. *IIE Transactions*, **56**, 945-959. <https://doi.org/10.1080/24725854.2023.2192252>
- [7] Wen, J.C., Lee, C.S., Keane, P.A., Xiao, S., Rokem, A.S., Chen, P.P., et al. (2019) Forecasting Future Humphrey Visual Fields Using Deep Learning. *PLOS ONE*, **14**, e0214875. <https://doi.org/10.1371/journal.pone.0214875>
- [8] Koenker, R. and Bassett, G. (1978) Regression Quantiles. *Econometrica*, **46**, 33-50. <https://doi.org/10.2307/1913643>
- [9] Belloni, A., Chen, M., Madrid Padilla, O.H. and Wang, Z. (2023) High-Dimensional Latent Panel Quantile Regression with an Application to Asset Pricing. *The Annals of Statistics*, **51**, 96-121. <https://doi.org/10.1214/22-aos2223>
- [10] Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R. and Gasthaus, J. (2019) High-Dimensional Multivariate Forecasting with Low-Rank Gaussian Copula Processes. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 6827-6837.
- [11] Belloni, A., Chernozhukov, V. and Kato, K. (2019) Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models. *Journal of the American Statistical Association*, **114**, 749-758. <https://doi.org/10.1080/01621459.2018.1442339>
- [12] Hornik, K. (1991) Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks*, **4**, 251-257. [https://doi.org/10.1016/0893-6080\(91\)90009-t](https://doi.org/10.1016/0893-6080(91)90009-t)
- [13] Barron, A.R. (1994) Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, **14**, 115-133. <https://doi.org/10.1023/a:1022650905902>
- [14] Gupta, B.B., Chui, K.T., Gaurav, A., Arya, V. and Chaurasia, P. (2023) A Novel Hybrid Convolutional Neural Network- and Gated Recurrent Unit-Based Paradigm for IoT Network Traffic Attack Detection in Smart Cities. *Sensors*, **23**, Article 8686. <https://doi.org/10.3390/s23218686>
- [15] Hu, X., Chu, L., Pei, J., Liu, W. and Bian, J. (2021) Model Complexity of Deep Learning: A Survey. *Knowledge and Information Systems*, **63**, 2585-2619.

- <https://doi.org/10.1007/s10115-021-01605-0>
- [16] Chronopoulos, I., Raftapostolos, A. and Kapetanios, G. (2024) Forecasting Value-At-Risk Using Deep Neural Network Quantile Regression. *Journal of Financial Econometrics*, **22**, 636-669. <https://doi.org/10.1093/jfinec/nbad014>
 - [17] Schmidt-Hieber, J. (2020) Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics*, **48**, 1875-1897. <https://doi.org/10.1214/19-aos1875>
 - [18] Farrell, M.H., Liang, T. and Misra, S. (2021) Deep Neural Networks for Estimation and Inference. *Econometrica*, **89**, 181-213. <https://doi.org/10.3982/ecta16901>
 - [19] Chen, X. and Liao, Y. and Wang, W. (2022) Inference on Time Series Nonparametric Conditional Moment Restrictions Using General Sieves. arXiv:2301.00092
 - [20] Shen, G. and Jiao, Y. and Lin, Y. and Horowitz, J.L. and Huang, J. (2021) Deep Quantile Regression: Mitigating the Curse of Dimensionality through Composition. arXiv:2107.04907
 - [21] Wang, L., Kim, Y. and Li, R.Z. (2013) Calibrating Nonconvex Penalized Regression in Ultra-High Dimension. *Annals of Statistics*, **41**, 2505-2536. <https://doi.org/10.1214/13-AOS1159>
 - [22] Koenker, R. (2005) Quantile Regression. Cambridge University Press. <https://doi.org/10.1017/cbo9780511754098>
 - [23] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. <https://doi.org/10.1038/nature14539>
 - [24] Ghysels, Eric and Babii, Andrii and Chen, Xi and Kumar, Rohit (2020) Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice. CEPR Discussion Papers.
 - [25] Liang, S. and Srikant, R. (2017) Why Deep Neural Networks for Function Approximation? *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, 24-26 April 2017.
 - [26] Yarotsky, D. (2017) Error Bounds for Approximations with Deep Relu Networks. *Neural Networks*, **94**, 103-114. <https://doi.org/10.1016/j.neunet.2017.07.002>
 - [27] Andersen, P.K. and Gill, R.D. (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, **10**, 1100-1120. <https://doi.org/10.1214/aos/1176345976>
 - [28] Pollard, D. (1991) Asymptotics for Least Absolute Deviation Regression Estimators. *Econometric Theory*, **7**, 186-199. <https://doi.org/10.1017/s0266466600004394>
 - [29] van der Vaart, A.W. and Wellner, J.A. (2023) Statistical Applications. In: van der Vaart, A.W. and Wellner, J.A., Eds., *Weak Convergence and Empirical Processes*, Springer International Publishing, 385-591. https://doi.org/10.1007/978-3-031-29040-4_3
 - [30] van der Vaart, A.W. and Wellner, J.A. (1996) Weak Convergence. In: Vaart, A.W., Wellner, J.A., Eds., *Springer Series in Statistics*, Springer, 16-28. https://doi.org/10.1007/978-1-4757-2545-2_3