TIME SERIES QUANTILE REGRESSION USING RANDOM FORESTS

HIROSHI SHIRAISHIa, TOMOSHIGE NAKAMURAa AND RYOTATO SHIBUKIA

aDepartment of Mathematics, Keio University, Yokohama, Kanagawa, Japan
bFaculty of Health Data Science, Juntendo University, Bunkyo-ku, Tokyo, Japan

We discuss an application of Generalized Random Forests (GRF) proposed to quantile regression for time series data. We extended the theoretical results of the GRF consistency for i.i.d. data to time series data. In particular, in the main theorem, based only on the general assumptions for time series data and trees, we show that the tsQRF (time series Quantile Regression Forest) estimator is consistent. Compare with existing article, different ideas are used throughout the theoretical proof. In addition, a simulation and real data analysis were conducted. In the simulation, the accuracy of the conditional quantile estimation was evaluated under time series models. In the real data using the Nikkei Stock Average, our estimator is demonstrated to capture volatility more efficiently, thus preventing underestimation of uncertainty.

Received 24 December 2022; Accepted 6 December 2023

Keywords: Quantile regression; random forest; nonlinear autoregressive model.

MOS subject classification: 62G08; 62M05; 62M10.

1. INTRODUCTION

Quantile Regression (QR) proposed by Koenker and Bassett (1978), is regression model that has been applied in various fields. QR analyzes the effects of covariates on outcomes by focusing on quantiles rather than means. Therefore, it can flexibly analyze the effect of covariates on the tail of the conditional distribution, which cannot be captured by regression on the mean. QR is used in a wide range of fields, including economics, medicine, and epidemiology, and is applicable not only to cross-sectional data, but also to panel data, for which the theory is well developed (Koenker, 2005). In the analysis of time series data, many studies have focused on the dynamics of the mean of the series, and on that of the conditional distribution. Therefore, the estimation of conditional quantiles of time series data by QR can capture the local and quantile-specific dynamics of time series and is expected to improve the quality of analysis.

Research on QR for time series data is ongoing. Value-at-Risk (VaR) is one of the criteria used to measure the market risk of an asset in the field of risk management. Because VaR is defined as the quantile of the asset return at time t, it is necessary to estimate the conditional QR function to estimate VaR. There are two methods to estimate the quantile: parametric and nonparametric methods. For parametric QR, Koul and Saleh (1995) proposed a conditional quantile estimator for AR models. Koenker and Zhao (1996) proposed ARCH models, Taylor (1999) introduced a linear VaR model, and Chernozhukov and Umantsev (2001) introduced a quadratic VaR model. The CAViaR model proposed by Engle and Manganelli (2004) is a broader class of models in time series QR for VaR estimation.

In CAViaR, the effect of past VaR on VaR at time t is described as linear, and the effect of the observed time series on VaR at time t is modeled linearly or nonlinearly by the researcher. Parametric QR has good properties in terms of interpretation and ease of implementation. However, parametric models have a serious bias if they are misspecified.

© 2024 The Authors. Journal of Time Series Analysis published by John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
To avoid this problem, Hall et al. (1999), Cai (2002), Wu et al. (2007), and Cai and Wang (2008) proposed a non-parametric method for estimating the conditional distribution function using kernel smoothing. In particular, Cai (2002) showed that the Weighted Nadaraya–Watson (WNW) estimator proposed by Hall et al. (1999) for time series satisfying the $\alpha$-mixing property has consistency and asymptotic normality. Furthermore, Cai and Wang (2008) proposed a Weighted Double-Kernel Local Linear estimator, an extension of the WNW estimator, and showed its consistency and asymptotic normality. However, in a nonparametric QR estimator using smoothing with kernel functions, the accuracy of the estimator highly depends on the choice of kernel function and bandwidth parameter.

As another approach, Jan et al. (2019), Wu et al. (2021) and so on have proposed to estimate QR nonparametrically using machine learning methods such as neural network. In this article, we also propose a non-parametric QR estimation method using random forests which is one of the typical machine learning methods. Random forest is a representative algorithm in machine learning that has been successfully used in various applications since its proposal in Breiman (2001). In recent years, the asymptotic properties of the estimators obtained by random forests have been studied in terms of consistency by Biau et al. (2008), Denil et al. (2013), Scornet et al. (2015), and Wager and Athey (2018). Consequently, random forests are now treated not only as a predictive model but also as a non-parametric statistical model.

Various extensions of random forests have also been proposed. A method for estimating conditional quantiles using random forests for i.i.d. data is QR forests (Meinshausen, 2006). Davis and Nielsen (2020) showed that random forest estimators are consistent with the problem of regression on the conditional mean of time series data with $\alpha$-mixing properties.

Among the recent extensions of random forests, the most notable is the Generalized Random Forests (GRF) by Athey et al. (2019), which estimates a parameter defined as the solution to a local estimating equation. Athey et al. (2019) showed that the estimator obtained by the GRF had consistency and asymptotic normality under i.i.d. data. Thus, from a theoretical perspective, the GRF can be used to estimate parameter.

In this study, we devote QR estimators for time series data using the GRF. Thus, it is necessary to extend the existing studies to the following points: Davis and Nielsen’s results for time series data focus on estimating a regression model for the mean, which is applicable to QR. However, this method assumes that sample size in each terminal node (leaf) is proportional to the sample size. It is difficult to verify this assumption when considering application. The subsample size included in the leaf should not depend on the sample size. In the GRF, the subsample size in the leaf can be fixed to resolve the above problem. The contribution of this study is to propose a time series QR Forest (tsQRF) using the GRF framework for $\alpha$-mixing time series data. Furthermore, we extend the theoretical results of GRF consistency for i.i.d. data to time series data. In particular, in the main theorem, based only on the general assumptions for time series data in Davis and Nielsen (2020) and trees in Athey et al. (2019), we showed that the tsQRF estimator is consistent. We also visualized the convergence of the tsQRF estimator through several simulation settings and compared its conditional quantile estimation accuracy with that of the WNW estimator. Furthermore, we fit the proposed method to the Nikkei Stock Average data and compared it with the WNW estimator to clarify the high sensitivity of tsQRF to time series volatility.

The remainder of this article is organized as follows. Section 2 discusses the properties of the QR estimator using the GRF. Especially, we define the double-sample tree score and GRF score, and then show the consistency of our estimator. Section 3 discusses the asymptotic behavior of the estimator and compares its accuracy with that of the WNW estimator through simulations. Section 4 illustrates the result of applying the proposed method to Nikkei Stock Average data and compares the results with those obtained using the WNW estimator. Finally, Section 5 summarizes the discussion and presents future issues. Concrete proofs of the theoretical results presented in Section 2 are provided in the Appendix A.

1.1. Relationship to Related Work

Davis and Nielsen (2020, hereafter referred to as DN) also discuss the RF (Random Forests) for time series models, specifically for the nonlinear autoregressive (NLAR) model, similar to this article. However, there are two
notable differences. First, in DN, the errors are i.i.d. random variables, while in this article, we accommodate a model with conditional heteroscedasticity. As a result, our model encompasses a class that includes ARCH-type models. Second, DN assumes that the mean function is globally bounded, whereas this article only assumes boundedness on a compact set. This leads our model to encompass a class that includes both AR- and TAR-types models. Under some regularity conditions, we can show that our model is exponentially $a$-mixing as well as DN’s model.

As for the scheme of Random Forests, we do not assume assumption (A-3) of DN which is a condition for the minimum subsample size falling in each leaf. Instead, we introduce PNN (Potentical Nearest Neighbor) $k$-set in the splitting rule following Wager and Athey (2018) and Athey et al. (2019) (see A-3). DN use assumption (A-3) to control the number of terminal leaves in each tree and derive a concentration inequality. This can be interpreted as an extension of the proof method of Wager and Walther (2015) to the time series setting. On the other hand, Wager and Athey (2018) and Athey et al. (2019) introduce the notion of ‘honest’ and introduce a double sample to construct the honest forests. While ‘non-honest forests’ such as Wager and Walther (2015) and Davis and Nielsen (2020) require the condition ‘with larger leaves’ to prove consistency. However, the double sample (i.e., honest forests) in this article makes it possible to prove consistency even under smaller leaves.

DN and Wager and Walther (2015), among others, discuss the estimation or prediction problem of ‘conditional mean’ by using RFs. In contrast, Meinshausen (2006) discusses that of ‘conditional quantile’ in the same way. This article also takes the conditional quantile as the parameter of interest, but discusses it in the framework of the GRF, which defines the parameter of interest as the solution to the local estimating equation, because the conditional quantile can also be defined as the solution to the local estimating equation while the score function is appropriately defined (cf. p. 12 of Athey et al. (2019)).

2. TIME SERIES QR FORESTS

In the classical nonparametric (and nonlinear) regression setting, we suppose that $(X_1, Y_1), \ldots, (X_T, Y_T)$ are i.i.d. observations from the model

$$Y = g(X) + \varepsilon,$$

where $g$ is some nonlinear function. In this article, we extend the model to time series. Time series models are also supposed in DN, this article further supposes conditional heteroscedasticity models.

Suppose that a sequence of random variables $\{Y_t\}_{t \geq 1}$ given initial value $X_1 := (Y_0, Y_{-1}, \ldots, Y_{-p})$ is the $p$th order stationary nonlinear autoregressive (NLAR($p$)) model

$$Y_t = g_1(X_t) + g_2(X_t)\varepsilon_t, \quad X_t = (Y_{t-1}, \ldots, Y_{t-p}), \quad t \geq 1. \quad (2.1)$$

where $g_1, g_2$ are unknown function, $\{\varepsilon_t\}_{t \geq 1}$ is a sequence of i.i.d. random variables with $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] < \infty$, and $p \geq 1$ is a fixed integer.

For this model, we consider estimation problem of the conditional time series QR function using GRFs (tsQRF) defined below, under several assumptions described in later section. Let $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X} = \mathcal{Y}^p = \mathbb{R}^p$ be sets of the spaces taken value of $Y_t$ and $X_t$. Let $Q \subset \ell^\infty (\mathcal{X})$ be a set of function $Q := \{q : \mathcal{X} \to \mathcal{Y}\}$ where $\ell^\infty (\mathcal{X})$ is a set of all uniformly bounded real functions on $\mathcal{X}$. Under some fixed $\tau \in (0, 1)$, we are interested in estimation or prediction of $\tau$-(conditional) quantile of $Y_t$ given by $X_t = x$, that is, $q_0 := (q_0(x))_{x \in \mathcal{X}} \in Q$ defined by

$$q_0(x) := F^{-1}_{Y|X}(\tau|x), \quad \forall x \in \mathcal{X}, \quad (2.2)$$

where $F^{-1}_{Y|X}$ is an inverse function of $F_{Y|X}(y|x) := P(Y_t \leq y | X_t = x)$ (i.e., $F_{Y|X}(q_0(x)|x = \tau)$.

As the empirical version of the $F_{Y|X}$ in (2), we introduce the GRF score $\hat{F}_{Y|X}$, which will be defined below.
2.1. Double-Sample Tree Score and GRF Score

GRF introduced by Athey et al. (2019) estimates a parameter defined as the solutions to a local estimating equation for i.i.d. data \((X_t, Y_t)\). Since we suppose that time series data \(\{Y_t\}_{t=1}^T\) is observed, the following adjustment is needed. Given a vector \(X_t\) of initial data independent of \(\{\epsilon_t\}_{t=1}^T\), we suppose that \(T\) observations \(Y_1, \ldots, Y_T\) from the model (1) are available and we group them in input–output pairs,

\[
D_T = \{(X_1, Y_1), \ldots, (X_T, Y_T)\}.
\]

To define RF based on \(D_T\), we first introduce a map of subsamples from \(D_T\) by \(D_T \times A \mapsto (I_s, J_s)\) where \(A\) is an index subset of \(\{1, \ldots, T\}\) defined below.

**Definition 1.** (Double-sample) Let \(s = s(T)\) be subsample size and let an index set be \(A \subset \{1, 2, \ldots, T\}\) with \(|A| = s\). A family of the index set \(A\) denoted by \(A_s\) is defined as follows:

\[
A_s := \left\{ A = (A^T, A^J) : A^T, A^J \subset \{1, 2, \ldots, T\} \mid |A^T| \cap |A^J| = \emptyset, |A^T| = \left\lfloor \frac{s}{2} \right\rfloor, |A^J| = \left\lceil \frac{s}{2} \right\rceil \right\},
\]

where the elements of \(A_s\) are different from each other. In addition, for any \(A = (A^T, A^J) \in A_s\), subsamples \(I_s\) and \(J_s\) of \(D_T\) are defined by \(I_s = D_{A^T}\) and \(J_s = D_{A^J}\) with \(D_A = \{(X_t, Y_t)\}_{t \in A}\) respectively.

In the double-sample tree defined below, we achieve ‘honesty’ by dividing its training subsamples into two halves \(I_s\) and \(J_s\). Then, \(J_s\)-sample is used to place the splits, while \(I_s\)-sample is used to do within-leaf estimation falling in each leaf. In what follows, we define the splitting rule using only \(J_s\)-sample.

**Definition 2.** (Splitting rule) Given \(J_s\)-sample, we define a sequence of partitions \(P_0, P_1, \ldots\) by starting from \(P_1 = \{X\}\) and then, for each \(n \geq 1\), construct \(P_n+1\) from \(P_n\) by replacing one set (parent node) \(P \in P_n\) by (child node) \(C_1 := \{x = (x_1, \ldots, x_p) \in P \mid x_\xi \leq \xi\} \) and \(C_2 := \{x = (x_1, \ldots, x_p) \in P \mid x_\xi > \xi\}\), where the split direction \(\xi \in \{1, \ldots, p\}\) is randomly chosen, and the split position \(\xi = \zeta(\xi) \in \{Y_{t-1}, \ldots, Y_{t-p}\} \subset P\) is chosen to maximize a criterion \(\Delta(C_1, C_2)\).

Furthermore, we impose the following assumptions for the splitting rule.

**Assumption 1.**

(A-1) (\(\omega\)-Regular) Every split puts at least a fraction \(\omega\) of the observations in the parent node into each child node, with \(\omega \in (0, 0.2]\).

(A-2) (Random Split) At every split, the probability that the tree splits on the \(j\)th feature (i.e., \(X_j = Y_{t-j}\)) is bounded from below by some \(\pi > 0\), for all \(j = 1, \ldots, p\).

(A-3) (PNN (Potential Nearest Neighbor) k-set) There are between \(k\) and \(2k - 1\) observations in each terminal node.

(A-4) (Subsample Size) Subsample size \(s\) scales \(s = T^\beta\) for some \(\beta \in (\beta_{\min}, 1)\) with

\[
\beta_{\min} = 1 - \left(1 + \frac{1}{p} \log(\omega^{-1}) \right) \left(\log(1 - \omega^{-1}) \right)^{-1}.
\]

**Remark 1.** In this article, the criterion \(\Delta(C_1, C_2)\) is the same as that of Athey et al. (2019). In practice, the optimal direction is chosen from ‘mtry’ directions at each step of the division where \(mtry \sim \min\{\max\{\text{Poisson}(m), 1\}, p\}\) for some \(m \in \mathbb{N}\). The assumption on \(\beta\) is the same as (13) of Athey et al. (2019), in which \(s/T \rightarrow 0\) and \(s \rightarrow \infty\) as \(T \rightarrow \infty\) are satisfied.

Under this splitting rule, we denote a given partition of the feature space \(\mathcal{X}\) by \(\Lambda\), and the subspace (leaf) of rectangular type created by the partitioning by \(L_\ell\) (\(\ell = 1, \ldots, |\Lambda|\)). Then,

\[
\Lambda = \Lambda(J_s; \xi) = \{L_1, \ldots, L_{|\Lambda|}\}, \quad \mathcal{X} = \bigotimes_{\ell \neq \ell'} L_\ell, \quad L_\ell \cap L_{\ell'} = \emptyset (\ell \neq \ell').
\]
where $\xi = (\xi_1, \ldots, \xi_{|\Lambda|})$ with $\xi_{\ell'} (\ell' = 1, \ldots, |\Lambda|)$ being the split direction for $L_{\ell'}$ satisfying A-2, independently, each other.

By $\Lambda = \Lambda(J; \xi)$, the double-sample tree score is defined as follows.

**Definition 3.** (Double-sample tree score) For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the double-sample tree score $T$ is defined by

$$T(x, y; I_s, J_s, \xi) = \sum_{s \in A} \mathbb{1}_{\{x \in L_s(x)\}} \mathbb{1}_{\{y \leq x\}},$$

where $L(x) \in \Lambda(J; \xi)$ is a leaf containing the test point $x$, and $\#L(x) = |\{t \in A^d | X_t \in L(x)\}|$.

Gathering the double-sample tree scores, we introduce the GRF score.

**Definition 4.** (GRF score) The GRF score $\hat{F}_{y|x}$ is defined by

$$\hat{F}_{y|x}(y|x) = \frac{1}{|A|^2} \sum_{s \in A} T(x, y; I_s, J_s, \xi) = \sum_{s \in A} a_s(x) \mathbb{1}_{\{y \leq x\}},$$

where

$$a_s(x) = \frac{1}{|A|} \sum_{s \in A} a_{s,s}(x), \quad a_{s,s}(x) = \mathbb{1}_{\{s \in A^d \}} \mathbb{1}_{\{x \in L_s(x)\}} \mathbb{1}_{\{y \leq x\}},$$

and

$$|A| = \left( \begin{array}{c} T \\ \lceil s/2 \rceil \end{array} \right) = \left( \begin{array}{c} s \\ \lfloor s/2 \rfloor \end{array} \right) = \frac{T!}{[s/2]! [s/2]! (T - s)!}.$$ 

### 2.2. tsQRF Estimation

Based on $\hat{F}_{y|x}$ defined by Definition 4, we introduce a time series QR forest (tsQRF) estimator $\hat{q}_T$ as follows.

**Definition 5.** tsQRF estimator $\hat{q}_T = (\hat{q}_T(x))_{x \in \mathcal{X}}$ is defined by

$$\hat{q}_T(x) = \inf \{y \in \mathcal{Y} | \hat{F}_{y|x}(y|x) \geq \tau \}. \quad (2.3)$$

Then, we have our main result, that is, uniformly consistency of $\hat{q}_T$.

**Theorem 1.** Under Assumptions 1–3, tsQRF estimator $\hat{q}_T$ converges to the true conditional quantile function $q_0$, in probability, that is,

$$\|\hat{q}_T - q_0\|_X \stackrel{p}{\rightarrow} 0 \quad \text{as} \quad T \to \infty.$$
Definition 6. (GRF score) Define $A_s^b = \{A_s^{(0)}, A_s^{(1)} \} \in A_s$. For each subsamples $\{I_s^{(b)}, J_s^{(b)}\}$ determined by $D_T$ and $A_s^{(b)} \in A_s^b$, we generate the double-sample tree score $T$ by Definition 3. Then, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the GRF score $\hat{F}_{Y|x}$ is defined by

$$\hat{F}_{Y|x}(y|x) = \frac{1}{B} \sum_{b=1}^{B} T(x, y; T_s^{(b)}, J_s^{(b)}, \xi) = \sum_{t=1}^{T} \alpha_t(x) 1_{\{y \leq y_t\}},$$

where

$$\alpha_t(x) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{A_s^{(b)}, x}(x), \quad \alpha_{A_s^{(b)}, x}(x) = \mathbf{1}_{\{x \in A_s^{(b)}\}} \frac{1_{\{|y \in L_s^{(b)}(x)\}}}{\sharp L_s^{(b)}(x)},$$

and $L_s^{(b)}(x) \in \Lambda(J_s^{(b)}, \xi)$ is a leaf containing the test point $x$; $\sharp L_s^{(b)}(x) = \{|t \in A_s^{(b)}| X_t \in L_s^{(b)}(x)\}$.

Based on $\hat{F}_{Y|x}$ defined by Definition 6, we introduce another tsQRF estimator $\hat{q}_T^b = (\hat{q}_T^b(x))_{x \in \mathcal{X}}$ as follows.

Definition 7. Another tsQRF estimator $\hat{q}_T^b$ is defined by

$$\hat{q}_T^b(x) = \inf \{y \in \mathcal{Y} | \hat{F}_{Y|x}(y|x) \geq \tau \}.$$

If $B$ is sufficiently large, $\hat{q}_T^b$ is also uniformly consistent.

Theorem 2. Under Assumptions 1–3, and $B \equiv B(T)$ with $\lim_{T \to \infty} \frac{1}{B} = 0$, $\hat{q}_T^b$ converges to $q_0$, in probability, that is,

$$\|\hat{q}_T^b - q_0\|_{\mathcal{X}}^p \to 0 \quad \text{as } T \to \infty.$$

2.3. Example

Suppose that $\{Y_t\}$ follows the AR(1) model: $Y_t = 0.7Y_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, 1)$. Generate a time series of length $T = 1000$ from this model, where $p, g_1, g_2$ in (1) are $p = 1, g_1(x) = 0.7x, g_2(x) = 1$, which satisfies the condition (A-6). Based on the observation $\{y_1, \ldots, y_{1000}\}$ and the initial value $x_1 = 0$, we put the input-output pairs $D_{1000} = \{(x_t, y_t)\}_{t=1, \ldots, 1000}$ with $x_t = y_{t-1}$. Under $\omega = 0.05, k = 5$, subsample size $s = T/2 = 500$ and $B = 2000$, we calculated the true conditional quantile $\{(q_0(x; \tau))\}$ and tsQRF $\{(\hat{q}_T^b(x; \tau))\}$ for some $\tau$. Figure 1 shows $\{x_{\tau}\}$, $\{(q_0(x_{\tau}); \tau), q_0(x_{\tau}); \tau\}$ and $\{(\hat{q}_T^b(x; \tau), \hat{q}_T^b(x; \tau))\}$ for $(\tau, \tau) = (0.01, 0.99), (0.05, 0.95), (0.10, 0.90), (0.25, 0.75), (0.40, 0.60)$.

There are some papers discussed about ‘crossing problem’, that is, a non-monotonicity problem as percentiles increase (cf., Yu and Jones (1998)). With our proposed method, this problem is ignored, but if $F_{Y|x}(y|x') < F_{Y|x'}(y|x')$ for some $y \in \mathcal{Y}, x, x' \in \mathcal{X}, F_{Y|x}(y|x') < F_{Y|x'}(y|x')$ is asymptotically satisfied with high probability from Lemma 5, which implies this problem is vanished asymptotically. For the initial value $X_1$, the distribution can be chosen such that $\{Y_1\}_{t=1}$ is strictly stationary which is assumed throughout the article. In practice, the initial value is fixed, but since the model is stationary, ergodic, and $p$ is fixed, its effect is negligible (cf., Anderson and Hsiao (1981)).
3. SIMULATION

Here, we verify the consistency shown in Theorem 2 by fitting the proposed tsGRF to data under various data-generating processes. Estimating time series using Random Forest can be seen as a non-parametric regression using \(a_t(x)\) as a kernel, given the way \(a_t(x)\) is constructed. Therefore, we compare the predictive performance of the tsGRF estimator with the conventional kernel-weighted estimator, the WNW estimator (Cai, 2002).

3.1. Data Generation Process

Four data generating models were used in this simulation. These models satisfy the assumption (A-6) without model (c) for function \(g_1, g_2\) where \(g_2 \equiv 1\). In model (c) nonlinear term is included. Although it does not satisfy assumption (A-6), we investigate the behavior of tsGRF estimator under such setting.

(a) First-order Markov chain model (Davis and Nielsen, 2020, equation (4.1))

\[
Y_t = \cos(5Y_{t-1})e^{-Y_{t-1}^2} + \varepsilon_t.
\]
The average bias is 

\[ \text{Bias} = \frac{1}{R} \sum_{r=1}^{R} \text{Bias}^{(r)}. \]

\[ \text{SDBias} = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} (\text{Bias}^{(r)} - \text{MBias})^2}. \]

\[ \text{MSE} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{T} \sum_{t=1}^{T} \text{Bias}^{(r)}_t^2. \]

We estimated the quantiles for \( \tau = 1\% , 10\% , 50\% , 90\% , \) and 99\% for each scenario. We utilized the R package \texttt{grf} to estimate the target quantiles, setting the GRF parameters as follows: subsample size \( s = \frac{T}{2} \), number of trees \( B = 2000 \), and the parameter \( \omega = 0.05 \) for the ratio of splits. These are the default values from the \texttt{grf} package.

In view of Theorem 1, \( B \) should be set to \( |A_r| \). However, as both \( T \) and \( s \) increase, \( |A_r| \) can become prohibitively large, leading to high computational costs. As a result, we restrict the number of trees to a value much less than \( |A_r| \), where \( B \) elements of \( A \) from \( A_r \) are chosen randomly. If \( B \) approaches the sample size, the approximation error remains sufficiently small, as noted by Theorem 2 and Wagner and Athey (2018). For each data-generating
Table I. MBias (training data)

| Model | $T$  | 1%  | 10% | 50% | 90%  | 99%  |
|-------|-----|-----|-----|-----|-----|-----|
| (a)   | 1000| 0.297| 0.021| 0.004| −0.012| −0.288|
|       | 5000| 0.297| 0.018| 0.000| −0.017| −0.294|
| (b)   | 1000| −0.007| −0.072| −0.011| 0.048| −0.005|
|       | 5000| 0.078| −0.014| −0.000| 0.014| −0.074|
| (c)   | 1000| 0.012| −0.039| −0.003| 0.046| 0.057|
|       | 5000| 0.077| −0.007| 0.001| 0.009| −0.064|
| (d)   | 1000| −0.281| −0.163| 0.000| 0.161| 0.263|
|       | 5000| −0.170| −0.108| −0.002| 0.107| 0.168|

Table II. SDBias (training data)

| Model | $T$  | 1%  | 10% | 50% | 90%  | 99%  |
|-------|-----|-----|-----|-----|-----|-----|
| (a)   | 1000| 0.068| 0.046| 0.037| 0.054| 0.081|
|       | 5000| 0.036| 0.024| 0.015| 0.025| 0.038|
| (b)   | 1000| 0.084| 0.045| 0.031| 0.048| 0.083|
|       | 5000| 0.041| 0.019| 0.014| 0.022| 0.039|
| (c)   | 1000| 0.093| 0.047| 0.038| 0.054| 0.085|
|       | 5000| 0.041| 0.020| 0.015| 0.021| 0.037|
| (d)   | 1000| 0.109| 0.054| 0.034| 0.048| 0.105|
|       | 5000| 0.044| 0.023| 0.015| 0.025| 0.049|

Table III. MSE (training data)

| Model | $T$  | 1%  | 10% | 50% | 90%  | 99%  |
|-------|-----|-----|-----|-----|-----|-----|
| (a)   | 1000| 0.316| 0.118| 0.062| 0.117| 0.304|
|       | 5000| 0.309| 0.116| 0.063| 0.116| 0.310|
| (b)   | 1000| 0.182| 0.091| 0.055| 0.090| 0.189|
|       | 5000| 0.156| 0.067| 0.038| 0.066| 0.158|
| (c)   | 1000| 0.368| 0.150| 0.054| 0.130| 0.492|
|       | 5000| 0.207| 0.075| 0.037| 0.071| 0.225|
| (d)   | 1000| 0.335| 0.189| 0.145| 0.188| 0.321|
|       | 5000| 0.183| 0.108| 0.084| 0.108| 0.179|

model from (a) to (d), we performed the simulations using time series lengths of $T = 1000$ and $T = 5000$. Results, including MBias, SDBias, and MSE for each scenario, are summarized in Tables I–III.

We discuss the results of the simulation in detail. In scenario (a), when comparing MBias for $T = 1000$ and $T = 5000$, there is not much difference across quantiles. However, for SDBias, it is evident that $T = 5000$ has smaller values across all quantiles, indicating that the estimator converges to a certain value. The MSE doesn’t show much difference between $T = 1000$ and $T = 5000$ for any quantile. In scenario (b), comparing MBias for $T = 1000$ and $T = 5000$, there isn’t much difference at the 50% quantile. However, at the 10% and 90% quantiles, as the sample size increases, the bias of the estimator decreases. At the extreme 1% and 99% quantiles, a larger sample size leads to a larger bias. SDBias shows smaller values for $T = 5000$ across all quantiles. Hence, this confirms that the estimator converges to a certain value. The MSE is smaller for $T = 5000$ across all quantiles compared to $T = 1000$. For scenario (c), because the indicator function is included in the expected value of $Y_t$, so the assumption A-5 is violated. However, MBias, SDBias, and MSE show a similar trend to scenario (b). This result may show that tsGRF performs well if the continuous assumption for conditional regression function is violated. Lastly, in scenario (d), the series order is $p = 5$. By comparing with scenario (b) which has $p = 2$, we
can consider the impact on order estimation. Observing the MBias table, it shows similar trends at the 10%, 50%,
and 90% quantiles as scenario (b). There is not much difference at the 50% quantile, but for the 10% and 90%
quantiles, \( T = 5000 \) shows a smaller MBias. Moreover, in scenario (d), at the 1% and 99% quantiles, \( T = 5000 \)
shows a smaller MBias compared to \( T = 1000 \). The other two metrics, SDBias and MSE, show smaller values
for \( T = 5000 \). The opposite result of MBias at the 1% and 99% quantiles compared to scenario (b) is due to
the estimation accuracy of random forests depending on sample size and feature dimensionality. In general, for
random forests in a regression context, if the input dimension increases, the estimation accuracy decreases unless
the sample size is increased. In this case, the decrease in estimation accuracy due to increased feature dimensionality
is mitigated by the larger sample size.
To summarize these results, for the 50% quantile, both $T = 1000$ and $T = 5000$ have an MBias close to 0, and the SDBias decreases as the time series lengthens, confirming consistency. Similarly, for quantiles like 10% and 90%, MBias shows a smaller value with a larger sample size, indicating the bias is nearing 0. SDBias also decreases with larger sample sizes, confirming consistency for these quantiles. On the other hand, for quantiles like 1% and 99%, even with a sufficiently long time series, there are cases where the bias doesn’t decrease significantly. However, since the SDBias decreases, it is hard to say that it converges to the true value with a moderate sample size. These results indicate that the estimation of extreme quantiles by GRF is unstable and that the estimation results are not very reliable.

The results are visually represented in Figure 2. This figure aggregates the calculated biases into histograms for each quantile when estimating the 1%, 2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99% quantiles using tsGRF for scenario (c) over 100 repetitions. The purple histogram illustrates the biases of quantile estimates conducted at $T = 1000$, while the blue histogram represents those conducted at $T = 5000$. As discerned from the figure, the estimation of the 50% quantile, whether at $T = 1000$ or $T = 5000$, is characterized by small biases. Similarly, for quantiles such as 5%, 10%, 25%, 75%, 90%, and 95% which are not at the extreme tails of the distribution, although some biases can be observed at $T = 1000$, extending the series length to $T = 5000$ significantly reduces these biases, suggesting consistent behavior. However, for the extreme quantiles of 1% and 99%, the rate at which the estimation bias approaches zero is outpaced by the speed at which the estimation variance approaches zero. These findings confirm that tsGRF exhibits consistent behavior in conditional quantile estimation, with the exception of extremely peripheral quantiles. This insight is crucial for practical applications, offering a lesson that tsGRF might not be the optimal choice for estimating extreme quantiles of conditional distribution.

3.3. Comparison of tsQRF and WNW Estimator

Next, we compare the predictive accuracy of the quantile estimation model using the WNW method proposed by Cai (2002) to that of our proposed tsGRF-based quantile estimation model. In this simulation, we generate a time series of length 1500. We train models for the 10%, 50%, and 90% conditional quantiles using the initial series of length 1000 and evaluate the predictive accuracy by applying the trained model to the subsequent series of length 1000. We then conduct at $T = 1000$, while the blue histogram represents those conducted at $T = 5000$. As discerned from the figure, the estimation of the 50% quantile, whether at $T = 1000$ or $T = 5000$, is characterized by small biases. Similarly, for quantiles such as 5%, 10%, 25%, 75%, 90%, and 95% which are not at the extreme tails of the distribution, although some biases can be observed at $T = 1000$, extending the series length to $T = 5000$ significantly reduces these biases, suggesting consistent behavior. However, for the extreme quantiles of 1% and 99%, the rate at which the estimation bias approaches zero is outpaced by the speed at which the estimation variance approaches zero. These findings confirm that tsGRF exhibits consistent behavior in conditional quantile estimation, with the exception of extremely peripheral quantiles. This insight is crucial for practical applications, offering a lesson that tsGRF might not be the optimal choice for estimating extreme quantiles of conditional distribution.

The evaluation of prediction accuracy was performed by MBias, SDBias, and MSE, similar to Section 3.2. However, while Section 3.2 utilized training data to evaluate these metrics for the purpose of demonstrating consistency, this simulation trains the model on the training data and evaluates it on the test data. Therefore, the evaluation expression’s $\{ (x_t^{(r)}, y_t^{(r)}) \}_{t=1}^T$ is replaced with $\{ (x_t^{(r)}, y_t^{(r)}) \}_{t=T+1}^{T+T}$ except for computing function $q_r$, and these are denoted as MBias, SDBias, and MSE. Tables IV present the MBias, SDBias, and MSE of tsQRF and kernel QR (WNW) for data generation models (a)–(d) respectively.
Table IV. Comparison of bias for estimator on test data

| Models | Method | 10%     | 50%     | 90%     | 10%     | 50%     | 90%     | 10%     | 50%     | 90%     |
|--------|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| (a)    | WNW    | -0.104  | 0.003   | 0.111   | 0.052   | 0.035   | 0.058   | 0.059   | 0.030   | 0.059   |
|        | tsQRF  | 0.019   | 0.004   | -0.008  | 0.051   | 0.039   | 0.058   | 0.119   | 0.062   | 0.117   |
| (b)    | WNW    | -0.112  | -0.008  | 0.090   | 0.063   | 0.055   | 0.071   | 0.078   | 0.054   | 0.077   |
|        | tsQRF  | -0.078  | -0.005  | 0.067   | 0.063   | 0.052   | 0.066   | 0.119   | 0.062   | 0.117   |
| (c)    | WNW    | -0.136  | -0.002  | 0.138   | 0.054   | 0.046   | 0.059   | 0.181   | 0.085   | 0.161   |
|        | tsQRF  | -0.045  | -0.008  | 0.048   | 0.049   | 0.041   | 0.057   | 0.161   | 0.063   | 0.135   |
| (d)    | WNW    | -0.146  | 0.002   | 0.153   | 0.058   | 0.040   | 0.054   | 0.151   | 0.110   | 0.153   |
|        | tsQRF  | -0.180  | 0.003   | 0.184   | 0.062   | 0.042   | 0.057   | 0.201   | 0.160   | 0.203   |

We begin by discussing scenario (a). There is little discernible difference in the estimated bias at the 50% quantile between WNW and tsQRF. However, for the predictions at the 10% and 90% quantiles, tsQRF exhibits a smaller bias compared to the predictions made by WNW. The variance of the bias between the two methods is also comparable, indicating that in predictions, tsQRF displays a reduced bias. Yet, on examining the MSE, we observe that WNW shows a lower value than the predictions made by tsQRF. This suggests that tsQRF’s predictive variance is larger than that of WNW. Therefore, in scenario (a), while tsQRF possesses a smaller bias than WNW, its predictive variance is higher. Turning to scenario (b), we observe a similar result to that of scenario (a). In scenario (c), there isn’t a significant difference in MBias or SDBias compared to scenarios (a) and (b). However, while WNW showed a smaller MSE than tsQRF in scenarios (a) and (b), in scenario (c), tsQRF exhibits a reduced MSE, indicating its superiority in both predictive bias and variance. This can be attributed to the influence of the threshold included in the model. The tsQRF, which is based on random forests, can flexibly estimate the shape of the weight-adjusted kernel for each point, whereas WNW cannot, which is likely the underlying reason. Lastly, in scenario (d), WNW consistently outperforms tsQRF across all metrics. This suggests that as the order of variables included in the model increases, predictions by tsQRF tend to deteriorate compared to WNW. Taking these results into account, when the true model order is low-dimensional, tsQRF compared to WNW, has a smaller predictive bias but a larger predictive variance. On the other hand, when the true model incorporates nonlinear terms, tsQRF can capture these aspects better than WNW. However, as the time series order included in the model grows, it is evident that the predictive accuracy of tsQRF tends to deteriorate compared to WNW.

4. EMPIRICAL RESULTS

Here, we examine the closing prices of the Nikkei Stock Average from January 1, 2014, to December 31, 2019 at Yahoo Finance (https://finance.yahoo.com/), using the tsQRF method and then compare the results with the WNW estimator. The data had gaps for days like weekends and holidays, so we removed these gaps. The data also shows a long-term increasing trend (Figure 3), and we transformed the price \( p_t \) at time \( t \) to \( r_t = \log\frac{p_t}{p_{t-1}} \) (Figure 4). Furthermore, the ADF test yields a \( p \)-value of 0.335, so the null hypothesis that ‘the data follows a unit root process’ cannot be rejected. In this analysis, we used the first four years (length: 976) as training data and the remaining 2 years (length: 489) as test data.

Since we cannot observe the true conditional quantile, we assess the quality of the estimation using the empirical coverage rate. The empirical coverage rate is defined for both the training and test data as follows.

\[
\hat{\theta} = \frac{N}{T} \sum_{t=1}^{T} 1_{\{Y_t \leq \hat{q}_T(x_t)\}} \quad \text{(training data)},
\]

\[
\hat{\theta} = \frac{N}{T'} \sum_{t'=1}^{T'} 1_{\{Y_{t'} \leq \hat{q}_T(x_{t'})\}} \quad \text{(test data)}.
\]

wileyonlinelibrary.com/journal/jtsa © 2024 The Authors. J. Time Ser. Anal. 45: 639–659 (2024) Journal of Time Series Analysis published by John Wiley & Sons Ltd. DOI: 10.1111/jtsa.12731
Figure 3. Time series of price $p_t$

Figure 4. Time series of $r_t = \log(p_t/p_{t-1})$
Table V. Empirical coverage rate of $\hat{\theta}$ on training data

| Model | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ | $\hat{\theta}$ |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|       | 2.5% | 10% | 50% | 90% | 97.5% | 2.5% | 10% | 50% | 90% | 97.5% |
| WNW   | 0.027 | 0.102 | 0.501 | 0.891 | 0.969 | 0.024 | 0.097 | 0.496 | 0.894 | 0.974 |
| tsQRF | 0.000 | 0.041 | 0.463 | 0.880 | 0.969 | 0.000 | 0.037 | 0.462 | 0.890 | 0.970 |

Table VI. Empirical coverage rate of $\tilde{\theta}$ on test data

| Model | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ | $\tilde{\theta}$ |
|-------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|       | 2.5% | 10% | 50% | 90% | 97.5% | 2.5% | 10% | 50% | 90% | 97.5% |
| WNW   | 0.012 | 0.084 | 0.497 | 0.926 | 0.988 | 0.016 | 0.086 | 0.487 | 0.926 | 0.982 |
| tsQRF | 0.022 | 0.108 | 0.493 | 0.922 | 0.984 | 0.018 | 0.086 | 0.497 | 0.924 | 0.988 |

We apply the kernel QR (WNW) and tsQRF methods with orders $p = 2$ and $p = 5$, and set their parameters to the values used in Section 3 as specified for the grf and np functions in the R packages. Tables V and VI display the estimated empirical coverage rates for WNW and tsQRF for the training and test data respectively. Values of $\hat{\theta}$ and $\tilde{\theta}$ that are closer to $\tau$ for both WNW and tsQRF are highlighted in red. From Table V, tsQRF provides a more conservative result compared to WNW for the training data, and the accuracy of $\hat{\theta}$ and $\tilde{\theta}$ does not significantly differ between $p = 2$ and $p = 5$. For the test data (Table VI), tsQRF yields better estimated values for $\tau$ than WNW.

Figures 5–8 illustrate the variation in the estimated quantiles of $r_t$ at $\tau = 2.5\%$, 97.5\% using $p = 5$ for both the training and test data. Figures 5 and 6 indicate that tsQRF is more sensitive to $r_t$ fluctuations than WNW. WNW fails to capture the shocks in $r_t$. This trend is also observed for the test data in Figures 7 and 8.
tsQRF for $r_t_{2.5\%}$ and $97.5\%$

Figure 6. tsQRF ($p = 5$)

kernel QR for $r_t_{2.5\%}$ and $97.5\%$

Figure 7. WNW ($p = 5$)
Figures 6 and 8 demonstrate that the quantile function estimated by tsQRF captures the variation of \( r_t \). Even amidst large fluctuations in the time series, \( r_t \) rarely exceeds the estimated 2.5% and 97.5% points. For instance, on September 9, 2015, as global economic concerns eased, stocks in Europe and the USA had risen the previous day. WNW could not capture the shock from this event, while tsQRF captured it accurately. On June 24, 2016, the UK voted to leave the European Union (EU). This led to concerns about a negative global economic impact, causing the Nikkei Stock Average to plummet. While tsQRF captured these fluctuations aptly, WNW did not.

Considering these findings, especially with data exhibiting large volatility fluctuations, tsQRF can capture changes in time series data with high sensitivity, offering a potential advantage in preventing risk underestimation compared to the WNW method.

5. SUMMARY AND FUTURE WORK

We applied the GRFs proposed by Athey et al. (2019) to QR for time series data. Although theoretical confirmation has not been considered for their use in a time series setting, we derived the uniform consistency of the estimated function under mild conditions. Davis and Nielsen (2020) also discussed the estimation problem using random forests (RF) for time series data, but the construction procedure of the RF treated by the GRF was essentially different, and different ideas were used throughout the theoretical proof. In addition, simulations and real data analyses were conducted. In the simulation, the accuracy of the conditional quantile estimation was evaluated under some time series models. In the real data using the Nikkei Stock Average, it was demonstrated that our estimator is more sensitive than the WNW estimator in terms of volatility and can prevent the underestimation of risk.

However many challenges remain. In this study, the order \( p \) is fixed; however, in practice, it should be determined using an information criterion or other methods. This is related to the variable selection problems in the GRF. On the theoretical side, the discussion of asymptotic normality and asymptotic efficiency is the subject of
future research. In particular, the efficiency involves the splitting procedure, and some methods, such as Neyman orthogonalization, are expected to be effective.

ACKNOWLEDGEMENTS

We thank the main editor, a co-editor and two referees for valuable comments which improved the article. This study was supported by JSPS KAKENHI Grant Number JP 21K11793.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Yahoo Finance at https://finance.yahoo.com/.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

REFERENCES

Anderson TW, Hsiao C. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76(375):598–606.

Athey S, Tibshirani J, Wager S. 2019. Generalized random forests. *The Annals of Statistics* 47(2):1148–1178.

Biau G, Devroye L, Lugosi G. 2008. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* 9(2008):2015–2033.

Breiman L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Cai Z. 2002. Regression quantiles for time series. *Econometric Theory* 18(1):169–192.

Cai Z, Wang X. 2008. Nonparametric estimation of conditional var and expected shortfall. *Journal of Econometrics* 147(1):120–130.

Chernozhukov V, Umantsev L. 2001. Conditional value-at-risk: aspects of modeling and estimation. *Empirical Economics* 26(1):271–292.

Davis R, Nielsen M. 2020. Modeling of time series using random forests: theoretical developments. *Electronic Journal of Statistics* 14(2):3644–3671.

Denil M, Matheson D, Freitas M. 2013. Consistency of online random forests. *Proceedings of the 30th International Conference on Machine Learning* 28:1256–1264.

Doukhan P. 2012. *Mixing: properties and examples*, Vol. 85 Springer Science & Business Media, New York.

Engle R, Manganelli S. 2004. Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22(4):367–381.

Hall P, Wolff R, Yao Q. 1999. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* 94(445):154–163.

Jan, G., Konstantinos, B., Yuyang, W., Syama, S. R., David, S., Valentin, F., Tim, J. 2019. Probabilistic forecasting with spline quantile function RNNs. *The 22nd International Conference on Artificial Intelligence and Statistics*; 1901–1910.

Koenker R. 2005. *Quantile Regression*. Econometric Society Monographs Cambridge University Press, New York.

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* 46(1):33–50.

Koenker R, Zhao Q. 1996. Conditional quantile estimation and inference for arch models. *Econometric Theory* 12(5):793–813.

Kosorok M. 2008. *Introduction to Empirical Processes and Semiparametric Inference* Springer, New York.

Koul H, Saleh AK. 1995. Autoregression quantiles and related rank-scores processes. *The Annals of Statistics* 23(2):670–689.

Meinshausen N. 2006. Quantile regression forests. *Journal of Machine Learning Research* 7(2006):983–999.

Scornet E, Biau G, Vert JP. 2015. Consistency of random forests. *The Annals of Statistics* 43(4):1716–1741.

Taylor J. 1999. A quantile regression approach to estimating the distribution of multiperiod returns. *The Journal of Derivatives* 7(1):64–78.

Tjøstheim D. 1990. Non-linear time series and Markov chains. *Advances in Applied Probability* 22(3):587–622.

Wager S, Walthers G. 2015. Adaptive Concentration of Regression Trees, with Application to Random Forests. arXiv preprint arXiv:1503.06388.
Wager S, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**(523):1228–1242.

Wu WB, Yu K, Mitra G. 2007. Kernel conditional quantile estimation for stationary processes with application to conditional value-at-risk. *Journal of Financial Econometrics* **6**(2):253–270.

Wu D, Gao L, Chinazzi M, Xiong X, Vespignani A, Ma Y, Yu R. 2021. Quantifying uncertainty in deep spatiotemporal forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ACM, New York; 1841–1851.

Yu K, Jones MC. 1998. Local linear quantile regression. *Journal of the American Statistical Association* **93**(441):228–237.

### APPENDIX A: PROOFS

Here, we present arguments leading up to our main result described in Section 2. Before we begin the proof, we impose the following assumption for the model.

**Assumption 2.**

(A-5) The random variable $\varepsilon_1$ has a density function $f_{\varepsilon}$, which is continuous and positive almost everywhere on $\mathcal{Y}$. For some $c \in (0, \infty)$,

$$
E[|\varepsilon_1|^m] \leq m!c^{m-2}, \quad m = 3, 4, \ldots.
$$

Moreover, the cumulative distribution function $F_{\varepsilon}$ of $\varepsilon_1$ satisfies

$$
\sup_{y \in \mathcal{Y}} \frac{F_{\varepsilon}(y + \lambda)}{F_{\varepsilon}(y)} < \infty
$$

for any $\lambda \in (0, \infty)$.

(A-6) The functions $g_1, g_2 : \mathcal{X} \to \mathcal{Y}$ are bounded and Lipschitz continuous on any compact sets in $\mathcal{X}$, $|g_1(x)| \leq C\|x\|_2$ with $0 \leq C < 1$ for $\|x\|_2 > r$ for some $r > 0$, $\inf_{x \in \mathcal{X}} g_2(x) \geq \nu$ for some $\nu > 0$, and $|g_2(x)|/\|x\|_2 \to 0$ as $\|x\|_2 \to \infty$.

Note that the assumption on the moments of $\varepsilon_1$ is known as Bernstein’s condition and implies that $\varepsilon_1$ is sub-Gaussian to prove concentration inequalities, but also those with a slightly heavier tail such as the Laplace distribution. The assumption of $F_{\varepsilon}$ is also satisfied for both sub-Gaussian and Laplace distributions. Under this assumption, the following lemma is obtained without the proof (For the proofs, see the Data S1). Note that this result is also obtained under the DN’s model.

**Lemma A1.** Under Assumption 2, the process $\{Y_i\}_{i \geq 1}$ is exponentially $\alpha$-mixing.

DN and Wager and Walther (2015) prepare a one-to-one mapping that transfers the feature space $\mathcal{X}$ to $\mathcal{Z} := [0, 1]^p$ to discuss the asymptotic property of the rectangle composed of each tree. In this article, we will prepare a similar mapping. Let $\mathcal{X}^e$ be a compact set in $\mathcal{X}$ with $M := \sup_{x \in \mathcal{X}^e} |g_1(x)| < \infty$. Define a map $t_h : \mathcal{X} \to \mathcal{Z}$, which transforms the input vector $X_t = (Y_{t-1}, \ldots, Y_{t-p})$ by

$$(y_1, \ldots, y_p) \mapsto t_h(y_1, \ldots, y_p) := (F_{\theta}(y_1), \ldots, F_{\theta}(y_p)),
$$

where $F_{\theta}$ is a cumulative distribution function defined by $F_{\theta}(y) = \int_{y \in [-\infty,y]} h(y)dy$ with

$$
h(y) := \frac{1 - \xi^{-1}}{\zeta - \xi^{-1}} f_{\theta} \left( \frac{y + M}{\nu} \right) + \frac{\zeta - 1}{\zeta - \xi^{-1}} f_{\theta} \left( \frac{y - M}{\nu} \right),
$$
\( \tilde{\zeta} := \sup_{x \in \mathcal{Y}} \frac{F_{F(y|M)}}{F_{F(x|M)}} \in (1, \infty) \). Then, \( t_{\mathbf{h}} \) is one-to-one function. Moreover, there exits a bounded probability density function of \( Z_1 = t_{\mathbf{h}}(X_1) \), which is similar to Lemma 5.1 of DN.

**Lemma A2.** Let \( \mathcal{Z}^o \) be a compact subset of \( \mathcal{Z} \) satisfying that for any \( x \in \mathcal{X}^o \), \( t_{\mathbf{h}}(x) \in \mathcal{Z}^o \). Under Assumption 2, it follows that the density \( h_{\mathcal{Z}} : \mathcal{Z} \rightarrow [0, \infty) \) of \( Z_1 = t_{\mathbf{h}}(X_1) \) satisfies

\[
\zeta^{-1} \leq h_{\mathcal{Z}}(z) \leq \zeta, \tag{A.1}
\]

for almost all \( z \in \mathcal{Z}^o \) and some \( \zeta \in (1, \infty) \).

Let \( \tilde{\Lambda} \) be a partition of feature space \( \mathcal{Z} \) in the same manner of \( \Lambda \) with \( J^X := \{ X_x \}_{x \in \mathcal{A}} \) replaced by \( J^Z := \{ Z_x \}_{x \in \mathcal{A}} \). Also, let \( \tilde{L}(z) \in \tilde{\Lambda} \) be the leaf containing the test point \( z := t_{\mathbf{h}}(x) \) transformed from the (original) test point \( x \in \mathcal{X} \) into \( \mathcal{Z} \).

Then, the moment bound of \( \text{diam}(\tilde{L}(z)) \) holds (see Lemma 3). Moreover, the same moment bound of \( \text{diam}(L(x)) \) also holds (see Corollary 1).

From the definition, each leaf \( L \in \tilde{\Lambda} \) can be expressed as \( L := \bigotimes_{j=1}^{p} [r_j, \bar{r}_j] \) based on a sequence \( \{(r_j, \bar{r}_j)\}_{j=1}^{p} \) with \( 0 \leq r_j < \bar{r}_j \leq 1 \) and \( r_j, \bar{r}_j \in J^Z \). Denoting the diameter of \( \tilde{L} \) by \( \text{diam}(\tilde{L}) \), we can write

\[
\text{diam}(L) := \sup_{z, z'' \in L} ||z' - z''|| = \sqrt{\sum_{j=1}^{p} (\bar{r}_j - r_j)^2}.
\]

By using the above results, we have the following moment condition.

**Lemma A3.** Under Assumptions 1, 2, we have

\[
\mathbb{E} \left[ \sup_{z \in \mathcal{Z}} \{ \text{diam}(L(z)) \} \right] = O(s^{-\gamma}),
\]

where \( \gamma = \frac{1}{2} \frac{\log(1 - \mu^{-1})}{\log \mu} \).

From Lemma 3, the upper bound of the moment of \( \text{diam}(L(x)) \) for the original feature space \( \mathcal{X} \) can be derived. For any leaf \( L \in \Lambda \), denoting the diameter of \( L \) by \( \text{diam}(L) \), we define

\[
\text{diam}(L) := \sup_{x', x'' \in L} ||x' - x''||.
\]

**Corollary A1.** Under Assumptions 1, 2, we have

\[
\mathbb{E} \left[ \sup_{x \in \mathcal{X}} \{ \text{diam}(L(x)) \} \right] = O(s^{-\gamma}).
\]

Define

\[
\tilde{F}_{Y|X}(y|x) := \sum_{t=1}^{T} \alpha_t(x) F_{Y|X}(y|X_t), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y},
\]

and for \( x \in \mathcal{X}, y, y' \in \mathcal{Y} \)

\[
\delta(x, y, y') := \tilde{F}_{Y|X}(y|x) - \tilde{F}_{Y|X}(y'|x) - \left\{ \tilde{F}_{Y|X}(y'|x) - \tilde{F}_{Y|X}(y'|x) \right\}.
\]
Then, for the moments of \( \delta(x, y, y') \), we can derive the followings. This result is similar to Lemma 8 of Athey et al. (2019) in i.i.d. case. You can see that there exists a difference between i.i.d. case and dependent case.

**Lemma A4.** Under Assumptions 1, 2, for any \( x \in \mathcal{X}, y, y' \in \mathcal{Y}, \) we have

(i) \( \mathbb{E} [\delta(x, y, y')] = O \left( \frac{1}{T} \right) \)

(ii) \( \text{Var}(\delta(x, y, y')) = O \left( \frac{\log_2 T}{T} \right) \).

By Lemmas 3 and 4, we can derive the uniform consistency of \( \hat{F}_{Y|X}(y|x) - F_{Y|X}(y|x) \) for any \( (x, y) \in \mathcal{X} \times \mathcal{Y} \).

**Lemma A5.** Under Assumptions 1, 2, we have

\[
\left\| \hat{F}_{Y|X} - F_{Y|X} \right\|_{\mathcal{X} \times \mathcal{Y}} \xrightarrow{p} 0 \quad \text{as } T \to \infty.
\]

By using Lemma 5, we show Theorem 1. Here we impose the following assumption to guarantee the consistency of our estimator.

**Assumption A3.**

(A-7) There exists some constant \( C > 0 \), such that, for almost surely,

\[
\left\| \hat{F}_{Y|X} - \tau \right\|_{\mathcal{X} \times \mathcal{Y}} \leq C \max_{t \in \{1, \ldots, T\}} \| a_t \|_{\mathcal{X}}.
\]

This assumption corresponds to Assumption 5 of Athey et al. (2019).

**Proof of Theorem 1**

From (2),

\[
F_{Y|X}(q_0(x)|x) = \tau,
\]

for any \( x \in \mathcal{X} \).

Suppose that a sequence \( \{q_T\} \in \mathcal{Q} \) satisfies \( \left\| F_{Y|X}(q_T(\cdot)|\cdot) - \tau \right\|_X \to 0 \) as \( T \to \infty \). Then, from (1), we have

\[
\left\| F_{Y|X}(q_T(\cdot)|\cdot) - \tau \right\|_X \to 0.
\]

From the strictly monotonicity of \( F_{\epsilon} \), for \( \tau_T := F_{\epsilon}((q_T(x) - g_1(x))/g_2(x)) \), there exists uniquely the inverse function \( F_{\epsilon}^{-1} \) such that \( F_{\epsilon}^{-1}(\tau_T) = (q_T(x) - g_1(x))/g_2(x) \). Hence, it follows from Assumption 2,

\[
\left\| (q_T - g_1)/g_2 - F_{\epsilon}^{-1}(\tau) \right\|_X \to 0.
\]

Notably, \( q_0(x) = g_1(x) + g_2(x)F_{\epsilon}^{-1}(\tau) \), we have

\[
\left\| q_T - q_0 \right\|_X \to 0 \quad \text{as } T \to \infty
\]

(Identifiability condition).

From Assumption (refA7), there exists \( C > 0 \) such that

\[
\left\| \hat{F}_{Y|X}(\hat{q}_T(\cdot)|\cdot) - \tau \right\|_X \leq \left\| \hat{F}_{Y|X} - \tau \right\|_{\mathcal{X} \times \mathcal{Y}} \leq C \max_{t \in \{1, \ldots, T\}} \| a_t \|_X \leq C \max_{t \in \{1, \ldots, T\}} \sup_{x \in \mathcal{X}, \epsilon \in \mathcal{A}_t} \| a_t \|_X
\]

and by Assumption (A-3),

\[
\mathbb{E} \left[ \alpha_{A_t}(x) \right] \leq \frac{1}{k} \sum_{t \in \mathcal{T}} \mathbb{E} \left[ 1_{\{x_t \in \mathcal{L}(\epsilon)\}} \right] = \frac{1}{k|A_T|} \sum_{t \in \mathcal{T}} \mathbb{E} \left[ 1_{\{x_t \in \mathcal{L}(\epsilon)\}} \right] = O \left( \frac{1}{s^2} \right).
\]
for any $A \in \mathcal{A}$, and $x \in \mathcal{X}$, which implies
\[
\|F_{Y|X}(\hat{q}_T(\cdot)|\cdot) - \tau\|_{X} = O_p(s^{-1}) = o_p(1).
\] (A.5)

From (A.2), (A.3), (A.5) and
\[
\sup_{q \in \mathbb{Q}} \|\hat{F}_{Y|x}(q(\cdot)|\cdot) - F_{Y|x}(q(\cdot)|\cdot)\|_{X} \leq \|\hat{F}_{Y|x} - F_{Y|x}\|_{X \times Y} \to 0,
\]
by Lemma 5, we obtain $\|\hat{q}_T - q_0\|_{X} \to 0$ from Theorem 2.10 (i) of Kosorok (2008).

**Proof of Theorem 2**

Suppose that for a fixed $s$, $T$ and $B$, $W = (W_A)_{A \in \mathcal{A}}$ is a multinomial random $|\mathcal{A}|$-vector taking values on \{0, 1, … , $B$\} with probabilities $1/|\mathcal{A}|$ and number of trials $B$, and is independent of the data $D_T$ and $\xi$. Note that $\sum_{A \in \mathcal{A}} W_A = B$. Then, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we can write
\[
\hat{F}_{Y|x}(y|x) = \frac{1}{B} \sum_{A \in \mathcal{A}} W_A T(x, y; I_x, J_y, \xi),
\]
with
\[
E[W_A] = \frac{B}{|\mathcal{A}|}, \quad \text{Cov}(W_{A_1}, W_{A_2}) = \begin{cases} \frac{B}{|\mathcal{A}|} \left(1 - \frac{1}{|\mathcal{A}|}\right) & \text{if } A_1 = A_2 \\ -\frac{B}{|\mathcal{A}|^2} & \text{if } A_1 \neq A_2. \end{cases}
\]

Let $E_w$ denote taking the expectation over $W = (W_A)_{A \in \mathcal{A}}$. Then, we have
\[
E_w \left[\hat{F}_{Y|x}(y|x)\right] = \hat{F}_{Y|x}(y|x),
\]
\[
E_w \left[\left\{\hat{F}_{Y|x}(y|x)\right\}^2\right] = \left(1 - \frac{1}{B}\right) \left\{\hat{F}_{Y|x}(y|x)\right\}^2 + \frac{|\mathcal{A}| - B + 1}{B|\mathcal{A}|^2} \sum_{A \in \mathcal{A}} \{T(x, y; I_x, J_y, \xi)\}^2,
\]
which implies that
\[
E \left[\left\{\hat{F}_{Y|x}(y|x) - \hat{F}_{Y|x}(y|x)\right\}^2\right] = \frac{1}{B} \left\{\left(1 - \frac{B-1}{|\mathcal{A}|}\right) \sum_{A \in \mathcal{A}} E \left[\left\{T(x, y; I_x, J_y, \xi)\right\}^2\right] - E \left[\left\{\hat{F}_{Y|x}(y|x)\right\}^2\right]\right\}
\]
\[
\quad \quad = O \left(\frac{1}{B}\right).
\]

Therefore, under $B^{-1} = o(1)$, we have $\|\hat{F}_{Y|x} - \hat{F}_{Y|x}\|_{X \times Y} \to 0$, and from Lemma 5, $\|\hat{F}_{Y|x} - \hat{F}_{Y|x}\|_{X \times Y} \to 0$. By the same argument of Theorem 1, we obtain $\|\hat{q}_T - q_0\|_{X} \to 0$ as $T \to \infty$. \qed