Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Jiayu Sui April 10, 2023

Generalized Quantile Random Forest with Smoothed Estimating Equations

By

Jiayu Sui

Seunghwa Rho, Ph.D.
Advisor

Mathematics

Seunghwa Rho, Ph.D.
Advisor

Zhiyun Gong, Ph.D.
Committee Member

Ruoxuan Xiong, Ph.D.
Committee Member

Yuanzhe Xi, Ph.D.
Committee Member

2023

Generalized Quantile Random Forest with Smoothed Estimating Equations

By

Jiayu Sui

Seunghwa Rho, Ph.D.

Advisor

An abstract of
A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Bachelor of Science with Honors

Mathematics

2023

Abstract

Generalized Quantile Random Forest with Smoothed Estimating Equations
By Jiayu Sui

Quantile regression establishes the relationship between one or more independent variable(s) and specific quantiles or percentiles of a dependent variable. It has been a handy supplement to the Least Squares Regression in the analysis of real-life applications. There are two random forest based implementation of the quantile regression, the quantile regression forest (quantregForest) by Meinshausen [9] and the quantile regression in the Generalized Random Forest (GRF) framework by Athey et al. [1]. The latter achieves a better performance by the redesign of the splitting rule using the quantile regression moment condition. However, the moment condition used contains an indicator function, which makes it non-smooth and non-differentiable.

By applying Kaplan and Sun [6]'s smoothed estimating equation (SEE) to the quantile regression moment condition, we managed to approximate the original moment condition with a smoothed moment condition with flexible bandwidth to adjust for the bias-variance tradeoff. Using self-constructed Python implementation of the GRF framework, we were able to insert the smoothed new moment condition and observe how such modification affect the performance of quantile estimation.

It was observed that on the random forest level, testing of the quantile GRF with SEE on the simulated data did not receive positive effect on the estimation accuracy. However, further inspection on the decision tree level quantile estimation reveals that with increased training sample, the quantile estimation should be able to approach satisfactory accuracy. Hence, we plan to further improve the program's run-time and produce better approach for hyperparameter tuning, such that the program can be executed with increased training sample and larger tree number in reasonable run-time. This would allow us to develop more understanding of the performance on the random forest level given sufficiently large training samples and tree numbers.

Generalized Quantile Random Forest with Smoothed Estimating Equations

By

Jiayu Sui

Seunghwa Rho, Ph.D.

Advisor

A thesis submitted to the Faculty of the
Emory College of Arts and Sciences of Emory University
in partial fulfillment of the requirements for the degree of
Bachelor of Science with Honors

Mathematics

2023

Acknowledgments

First, I would like to express my sincere gratitude to Dr. Seunghwa Rho, my thesis advisor, for her invaluable patience and feedback. For the past one year, I have been working under Dr. Rho's guidance on this research project. Despite the difficulty caused by remote communication and time difference, she has always been very approachable for answering all my questions and providing me helpful directions. I could not have undertaken this journey without her.

I am also deeply grateful to my defense committee, Dr. Zhiyun Gong, Dr. Ruoxuan Xiong, and Dr. Yuanzhe Xi. They have generously provided their insights and expertise to foster the improvement of my work. Their encouraging comments motivate me to further enhance my results in the future.

Lastly, I would like to offer my special thanks to my family and my friends for their emotional support. Their belief in me has kept my spirits and motivation high during this process. Many thanks to Alex Li and Ruby Wu for helping me to rehearse my defense. Their support are deeply meaningful to me.

Contents

List of Figures

Chapter 1

Introduction

In statistical modeling, the most conventional and widely applied regression approach is the least squares regression. With its moment condition minimizing the sum of residual squared, the least squares regression estimates the conditional mean of a dependent variable given one or more independent variables. Although not as commonly used as the least squares regression, quantile regression can be a great supplement to many practical situations, where they may not only require the conditional mean, but also need the conditional median or other conditional percentile to contribute to the process of decision making and resource allocation. This is especially essential for studying problems regarding inequality. Least squares regression assumes homoscedasticity to make many inferences. In cases where the conditional distribution of the dependent variable changes across different values of the independent variables, least squares regression only allows analysis of the changes in the mean due to independent variables but misses the bigger picture of changes in conditional variance. Quantile regression is capable of completing the missing picture by examining how various quantiles or percentiles changes due to the independent variable. Hence, quantile regression is proved to be significant for many social science research including studys in educational inequality, income inequality, income-related inequality in

health, etc.

To obtain the quantile estimation, the quantile regression forest by Meinshausen [9] adopts an approach very similar to Breiman [2]'s random forest for mean estimation. The major modification is instead of computing the mean of terminal node observations, it is the desired quantile of terminal node observations that is returned. However, the splitting rule of the decision trees remains to be maximizing the variance reduction from the parent node to child nodes. Hence, it would be expected that a splitting rule tailored for quantile regression would improve the performance of the estimation.

The generalized random forest (GRF) of Athey et al. [1] generalizes random forest of Breiman [2] such that it can be broadly applied to estimation problems that are defined through the moment conditions. While still inheriting the stability and ease of use of random forest, GRF is able to extend the use-case from the conditional mean estimation to various problems. This is possible because GRF adopts the adaptive neighborhood weight through random forest but otherwise still solves the moment condition locally using these weights. When obtaining these weights, GRF uses a splitting rule which directly takes into account of the parameters of interest and the moment conditions which identifies those. This splitting rule is better suited to estimate the heterogeneity in the parameter estimates which a researcher is interested in when performing the local estimation. In addition, through the sample splitting, a researcher can perform hypothesis test in addition to prediction.

The conditional quantile regression (QR) of Koenker and Bassett [8] minimizes a quantile loss function which leads to well-defined moment condition for quantile regression

$$E\left[x_i\left(\tau - \mathbb{1}\left[x_i'\hat{\beta}_\tau - y_i > 0\right]\right)\middle| x_i = c\right] = 0.$$

where τ is the quantile value ranging from 0 to 1, and $\hat{\beta}_\tau$ is the estimated coefficient. The indicator function $\mathbb{1}\left[x_i'\hat{\beta}_\tau - y_i > 0\right]$ is a step function, which makes the moment condition discontinuous. By this moment condition, quantile regression is also incorporated into the GRF framework and hence can be estimated using the GRF.

In the current implementation of quantile regression in GRF, the quantile loss function and the moment function are not smooth and nondifferentiable due to the indicator function. Using Kaplan and Sun [6]'s method, however, either the loss function or the moment condition can be smoothed by replacing the indicator function with a smoothed estimating equation (SEE).

Since the GRF provides a generalized framework which allows users to apply their own tailored moment conditions when fitting the forest, it would be the interest of this paper to find out how the performance changes when the smoothed moment conditions is applied to GRF to obtain quantile estimation.

Chapter 2

Background and Related Work

2.1 Quantile Regression Forest

The quantile regression forest of Meinshausen [9] is implemented in the R package quantregForest. It is similar to the random forest of Breiman [2] in terms of the decision tree splitting rule. Hence, decision trees are grown in a manner similar to the random forest implementation for mean estimation. However, in the prediction stage, instead of obtaining the mean of each terminal node, all observations in the terminal node is retained for estimation.

For a given $X = x$, the observation is passed down through all decision trees $b = 1, ..., B$ to find its corresponding terminal nodes. The neighborhood weights $w_i(x, \theta_t)$ are computed for each tree by assigning positive weight to x's corresponding terminal node entries, and zero weight to entries contained in other terminal nodes. Then the aggregated weight $w_i(x)$ is computed by averaging over all neighborhood weights for all decision trees. The procedure could be summarized as the equation

$$w_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}[x_i \in N_x^b]}{\#N_x^b}. \tag{2.1}$$

where N_x^b denotes the terminal node of b^{th} tree where x belong to and $\#N_x^b$ is the total number of observations in the this node.

Instead of taking the weighted mean for conditional mean estimation at $X = x$, Meinshausen [9] utilizes the conditional distribution function of Y for $X = x$,

$$F\left(y|\, X = x\right) = Pr(Y \leq y|\, X = x) = E(\mathbb{1}\left[Y \leq y\right]|X = x). \tag{2.2}$$

Then, to obtain the τ-quantile of Y would be to find the y such that $Pr(Y \leq y|\, X = x) = \tau$.

Then analogous to how the conditional mean $E(Y|X = x)$ is approximated by a weighted mean of Y, $E(\mathbb{1}\left[Y \leq y\right]|X = x)$ is defined to be approximated by a weighted mean of $\mathbb{1}[Y_i \leq y]$. Hence, the approximated distribution would be expressed in the following way:

$$\hat{F}\left[y|X = x\right] = \sum_{i=1}^{n} w_i(x)\mathbb{1}[Y_i \leq y] \tag{2.3}$$

Therefore, the approximation of τ-quantile of Y would be obtained by

$$\hat{Q}_{Y|X}(\tau) = min\{y : \hat{F}\left[y|X = x\right] \geq \tau\} = min\{y : \sum_{i=1}^{n} w_i(x)\mathbb{1}[Y_i \leq y] \geq \tau\}. \tag{2.4}$$

This would then give the quantile estimation for the quantile regression forest method.

2.2 Generalized Quantile Random Forest

The GRF of Athey et al. [1] is implemented in the R package grf. It also uses random forest to obtain the neighborhood weight when estimating the moment condition locally but with a modified splitting rule. The splitting rule of the GRF seeks to maximize the heterogeneity in parameter estimates. For the τ-quantile estimation,

the GRF obtains the neighborhood weight through random forest with a splitting rule that involves quantile regression's moment condition to maximize the heterogeneity in the τ-quantile of the two child nodes. In quantile regression, the relevant moment conditions would be

$$E\left[\tau - \mathbb{1}\left[\hat{\theta}_\tau - y_i > 0\right] \middle| x_i = c\right] = 0. \tag{2.5}$$

However, it would be computationally expensive to solve the above moment condition for the two child nodes for every possible way of splitting the parent node. As a result, the parameter estimate of the child node would be approximated using the parameter estimate and the gradient of its parent node for computational efficiency. This approximation leads to the splitting based on the pseudo-outcomes.

First, the parameter estimate $\tilde{\theta}_C$ of the child node is approximated with

$$\tilde{\theta}_C = \hat{\theta}_p - \frac{1}{|\{i : X_i \in C\}|} \sum_{\{i:X_i \in C\}} \xi' A_p^{-1} \psi_{\hat{\theta}_p}(O_i) \tag{2.6}$$

where $\hat{\theta}_p$ is the parameter estimate of its parent node through solving the moment condition at the parent node, $A_p = \frac{1}{|\{i:X_i \in P\}|} \sum_{\{i:X_i \in P\}} \nabla \psi_{\hat{\theta}_p}(O_i)$ is the gradient of the parent node, and $\psi_{\hat{\theta}_p}(O_i)$ is the scoring function within the moment condition.

To maximize the heterogeneity of $\tilde{\theta}_C$ in the two child nodes, we want to maximize the criterion

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{n_j} \left(\sum_{i:x_i \in C_j} \left(-\xi' A_p^{-1} \psi_{\hat{\theta}_p}(O_i) \right) \right)^2 \tag{2.7}$$

Thus, the pseudo-outcome is defined as

$$\rho_i = -\xi' A_p^{-1} \psi_{\hat{\theta}_p}(O_i) \tag{2.8}$$

so that after calculating the pseudo-outcome, we can calculate the $\tilde{\Delta}(C_1, C_2)$ criterion for splitting without the need of estimating child node parameter explicitly.

Based on the splitting rule defined, the GRF would be able to grow decision trees and obtain neighborhood weights for each $X = c$ in the same way as Meinshausen [9]'s quantile regression forest. The neighborhood weight $w(x_i; c)$ defines the local area of c where the parameter of interest should be estimated. This weight would be larger when x_i lies closer to c. If more trees define x_i to be neighborhood of c, then the weight would be larger.

Once the weight is obtained, the problem would boil down to solving the following weighted moment condition:

$$\frac{1}{n}\sum_{t=1}^{T} w(x_i; c)\left(\tau - \mathbb{1}\left[\theta_\tau - y_i > 0\right]\right) = 0. \tag{2.9}$$

Additionally, the GRF applied the honesty technique which separates the training set into splitting and estimation sets. The splitting set would be used to split the trees, and the estimation set would be used to populate the terminal nodes once the splitting process is fully complete. As a result, only the estimation set will be used to obtain the weights of the neighbors within each tree. By separating the splitting and estimation sets, the estimates obtained from GRF have asymptotic normality and hence statistical significance can be obtained.

Chapter 3

Method

3.1 Smoothed Moment Condition

In the GRF implementation of the quantile regression forest, the relevant moment conditions is

$$\sum_{i=1}^{n} x_i \left(\tau - \mathbb{1}[x_i^{'}\hat{\beta}_\tau - y_i > 0]\right) = 0$$

Due to the nondifferentiable indicator function, in the splitting process, the pseudo-outcome can not be obtained with (2.8) since computing A_p requires a differentiable moment condition scoring function. Hence, to obtain the parameter estimate $\hat{\beta}_\tau$, we apply Kaplan and Sun [6]'s smoothed estimating equation to smoothes out the moment condition by replacing the indicator function $\mathbb{1}[x_i^{'}\hat{\beta}_\tau - y_i > 0]$ with $G\left(\frac{x_i'\hat{\beta}_\tau - y_i}{h}\right)$. The $G()$ function is defined as

$$G(\nu) = \begin{cases} 0 & \nu \leq -1 \\ 0.5 + \frac{105}{64}\left(\nu - \frac{5}{3}\nu^3 + \frac{7}{5}\nu^5 - \frac{3}{7}\nu^7\right) & \nu \in [-1, 1] \\ 1 & \nu \geq 1 \end{cases}$$

which is plotted below. By its definition, it is not difficult to find that $G()$ is not only continuous, but also differentiable on the first order. How smooth would be defined by the bandwidth h. The larger bandwidth h would lead to smoother function which would in turn lead to larger bias and smaller variance.

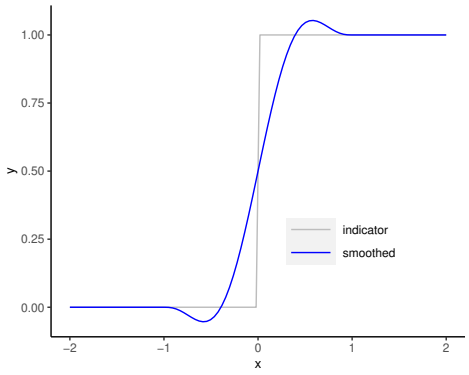

Figure 3.1: $G\left(\frac{x}{h}\right)$ function when bandwidth $h = 1$

After replacing the indicator function with the SEE, we are able to follow the GRF procedure in computing the pseudo-outcome. The new moment condition is

$$\sum_{i=1}^{n} x_i \left(\tau - G\left(\frac{x_i'\hat{\beta}_\tau - y_i}{h}\right)\right) = 0 \tag{3.1}$$

3.2 Splitting Rule Modification

For a given $X = x$, we are still following (2.1) in obtaining the weights for each terminal node entry. Therefore, within each decision tree, within the terminal node where x belong to, we assign equal weights to the each entry. This suggests that locally in the area defined through the terminal node entries, we are estimating the τ-quantile through direct computation of the τ-quantile of Y value, and the X values of the terminal nodes do not participate in the computation. As a result, we can

simplify the moment condition (3.1) to

$$\sum_{i=1}^{n} \tau - G\left(\frac{\hat{\theta}_\tau - y_i}{h}\right) = 0 \tag{3.2}$$

From this moment condition, we identify that

$$\psi_{\hat{\theta}_p}(y_i) = \tau - G\left(\frac{\theta - y_i}{h}\right) \tag{3.3}$$

Since this scoring function is differentiable, following (2.8), we were able to find the gradient of the parent node through differentiating on θ

$$\nabla_\theta \psi(y_i) = -G'\left(\frac{\theta - y_i}{h}\right)\frac{1}{h} \tag{3.4}$$

$$A_p = -\frac{1}{h} \cdot \frac{1}{n_p} \sum_{i \in P} G'\left(\frac{\theta - y_i}{h}\right) \tag{3.5}$$

Finally, the pseudo-outcome can be computed with

$$\rho_i = -A_p^{-1}\psi_{\hat{\theta}_p}(y_i) = \left(\frac{1}{h}\frac{1}{n_p}\sum_{j \in P} G'\left(\frac{\theta - y_j}{h}\right)\right)^{-1}\left(\tau - G\left(\frac{\theta - y_i}{h}\right)\right) \tag{3.6}$$

Following the pseudo-outcome ρ_i, we would only need to find $\tilde{\Delta}(C_1, C_2)$ by equation (2.7) to replace the original splitting criteria.

The newly defined splitting criteria would allow each parent node to split into two child nodes using the optimal splitting feature and feature value. To enable child nodes to recursively perform the splitting, it is required to "relabel" those child nodes by calculating their θ estimate explicitly. This is performed by solving (3.2) via numerical estimation functions.

The same standard CART regression split procedure is used on the newly defined splitting criteria, to grow decision trees using the splitting set of the training data. After the tree splitting process is complete, the estimation set is used to populate the

terminal nodes. This completes the fitting step of the random forest.

3.3 Estimation Stage

In the estimation stage, for each $X = c$ we obtain neighborhood weights by passing c through decision trees grown with modified splitting criterion. Then in the same manner as Meinshausen [9]'s quantile regression forest, we are able to aggregate neighborhood weights of each decision tree into a final weight for the random forest $w(x_i; c)$. With the smoothed moment condition, the new weighted moment condition would be

$$\frac{1}{n}\sum_{t=1}^{T} w(x_i; c)\left(\tau - G\left(\frac{\theta_\tau - y_i}{h}\right)\right) = 0. \tag{3.7}$$

Therefore, the τ-quantile of Y given $X = c$ can be expressed as

$$\hat{\theta}_\tau \in \arg\min_{\theta_\tau} \frac{1}{n}\sum_{t=1}^{T} w(x_i; c)\left(\tau - G\left(\frac{\theta_\tau - y_i}{h}\right)\right) = 0. \tag{3.8}$$

which can be solved through numerical estimation functions.

Chapter 4

Result

The design described in the Method section is implemented in Python. The following simulated data is generated to test the performance of the implemented GRF quantile forest with SEE:

Mean shift simulation data: Contains n = 1000 independent and identically distributed observations. There are 5 independent variables X_i. Each independent variable is set to be uniformly distributed over $[-1, 1]$. The dependent variable Y_i is only dependent on X_1. The other 4 dependent variables are noise. When $X_1 < 0$, Y_i is distributed normally with mean of 0 and standard deviation of 1. When $X_1 > 0$, Y_i is distributed normally with mean of 0.8 and standard deviation of 1.

Scale shift simulation data: Contains n = 1000 independent and identically distributed observations. There are 5 independent variables X_i. Each independent variable is set to be uniformly distributed over $[-1, 1]$. The dependent variable Y_i is only dependent on X_1. The other 4 dependent variables are noise. When $X_1 < 0$, Y_i is distributed normally with mean of 0 and standard deviation of 1. When $X_1 > 0$, Y_i is distributed normally with mean of 0 and standard deviation of 2.

For the two sets of simulated data, we want to estimate the quantiles at $\tau = 0.1$, 0.5, 0.9. The result is illustrated in the following figures:

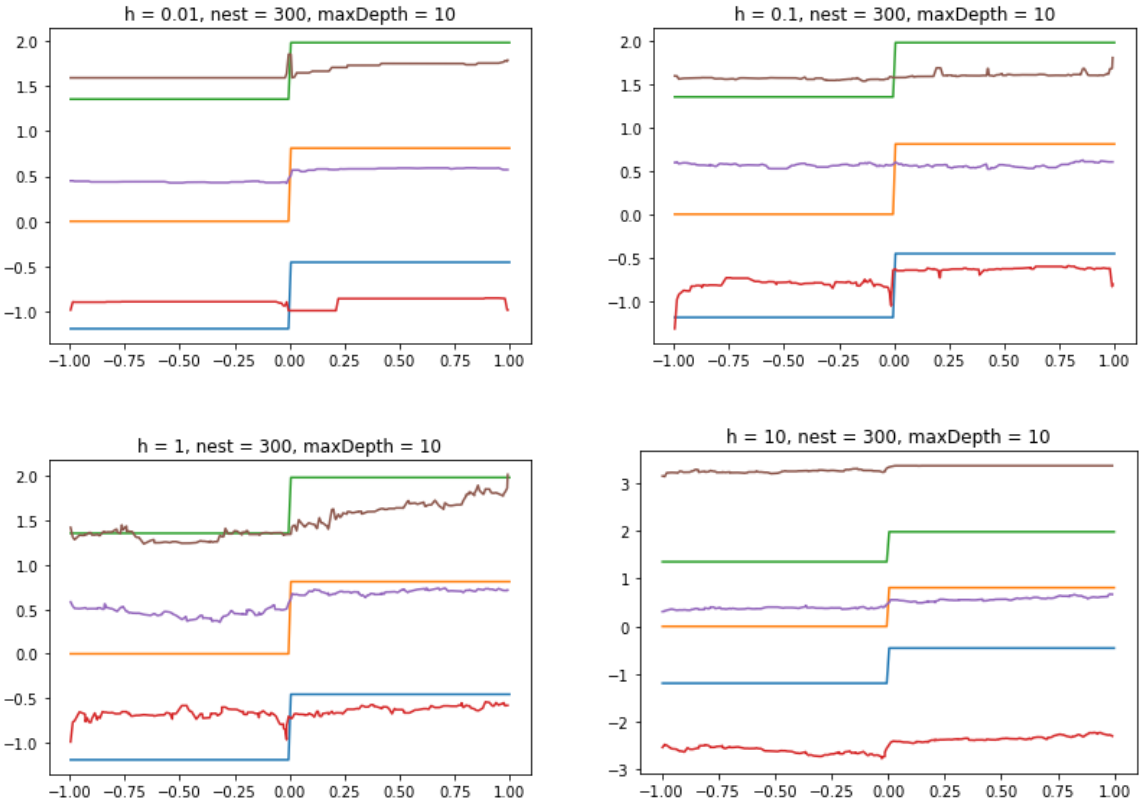

Figure 4.1: Quantile estimation over the mean shift data with setting: number of trees nest = 300, maximum depth of trees maxDepth = 10, minimum number of samples in terminal node minLeafSample = 5, The number of features to consider when looking for the best split maxFeat = 5. The bandwidth of the SEE h is adjusted to equal to 0.01, 0.1, 1, 10.

From the definition of the $G()$ function, we would expect that with smaller bandwidth h, the function is going to approximate the step function. As h increases, the function is smoother, incurring larger biases for the estimation. It can be observed that for $h = 10$, the 0.1 and 0.9 quantile estimations are indeed severely biased from the truth value. However, even for smaller h values, the quantile estimations do not meet the expected performance. It can be observed that the method is able to detect a shift of mean at $X_1 = 0$. However, the method is consistently overestimating at $X_1 < 0$ and underestimating at $X1 > 0$.

Similar performance is found for the scale shift simulation data. It is more doubtful about whether the method is able to detect the shift of standard deviation at $X_1 = 0$.

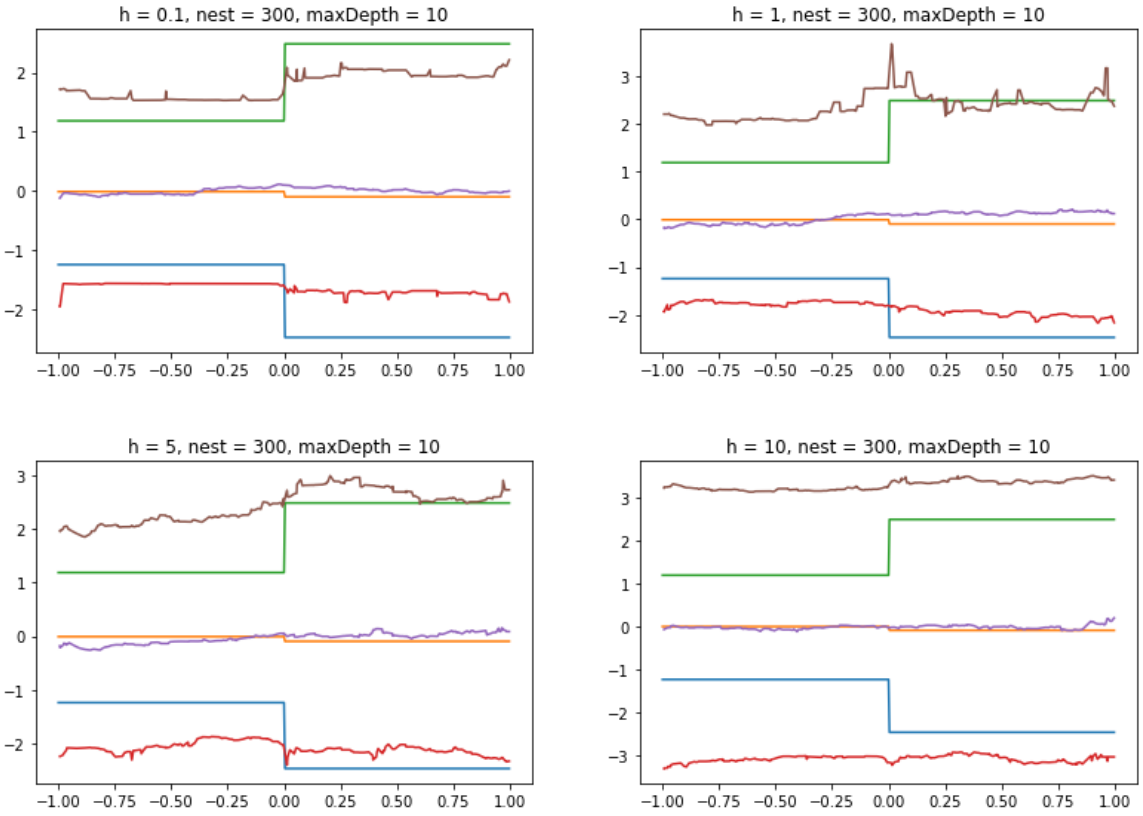

Figure 4.2: Quantile estimation over the scale shift data with setting: number of trees nest = 300, maximum depth of trees maxDepth = 10, minimum number of samples in terminal node minLeafSample = 5, The number of features to consider when looking for the best split maxFeat = 5. The bandwidth of the SEE h is adjusted to equal to 0.1, 1, 5, 10.

Since no positive impact is observed on the random forest level estimation, we would like to further inspect the mechanism with estimations from a single decision tree. This can be easily achieved by limiting the number of trees in the random forest $nest = 1$.

However, we need to be cautious that a single decision tree is subject to considerably larger variance as compared to the random forest. As a result, we need sufficiently large training and testing set to ensure that the simulated data is asymptotically following the designed distribution. With a much smaller tree number, the program is able to be executed with expanded training and testing sample within reasonable run-time.

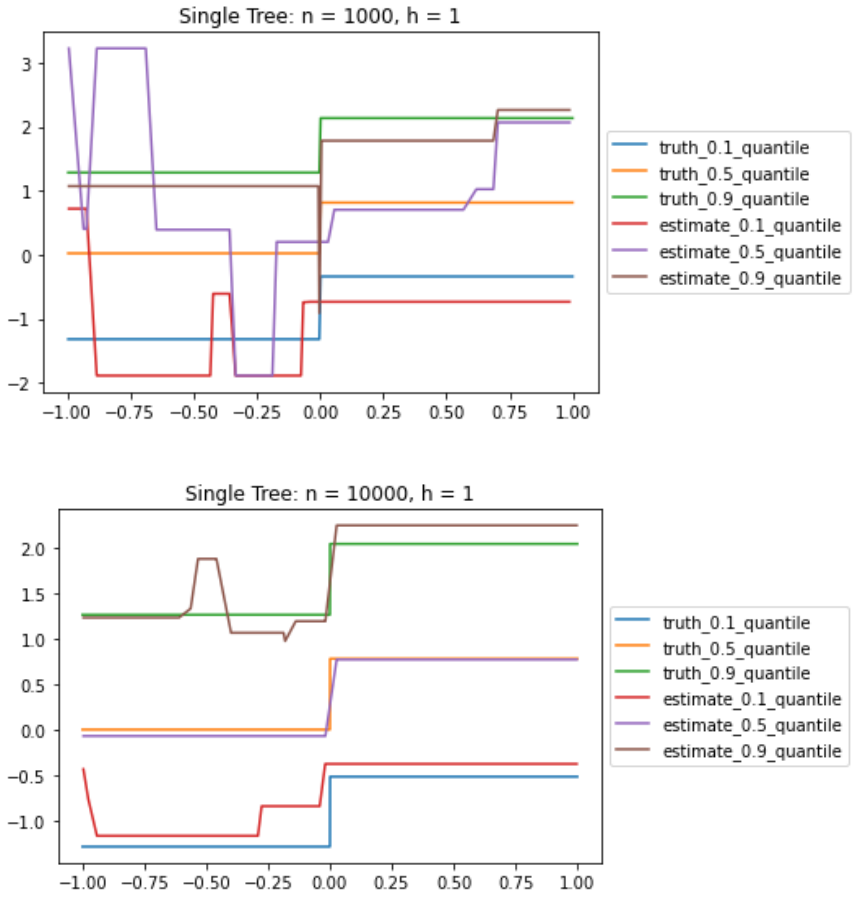

Figure 4.3: Single tree quantile estimation over the mean shift data with setting: number of trees nest = 1, maximum depth of trees maxDepth = 10, minimum number of samples in terminal node minLeafSample = 5, The number of features to consider when looking for the best split maxFeat = 5, bandwidth if the SEE $h = 1$. The size of the training data is adjusted to equal to 1000, 10000.

The settings described in Figure 4.3 are applied multiple times on simulated mean shift data to ensure that randomness involved in the data generation process does not produce exceptionally well-performed or ill-performed estimations by chance. It is evident from Figure 4.3 that increasing the training set size by 10 times effectively causes the estimation to conform better with the ground-truth quantile values in simulation design. However, certain deviations from the ground-truth values still exist, including the overall over-estimation of the 0.1 quantile that occurred constantly in all trials under such setting.

Similar improvement of performance by increasing size of training set also presents for the scale shift simulation data.

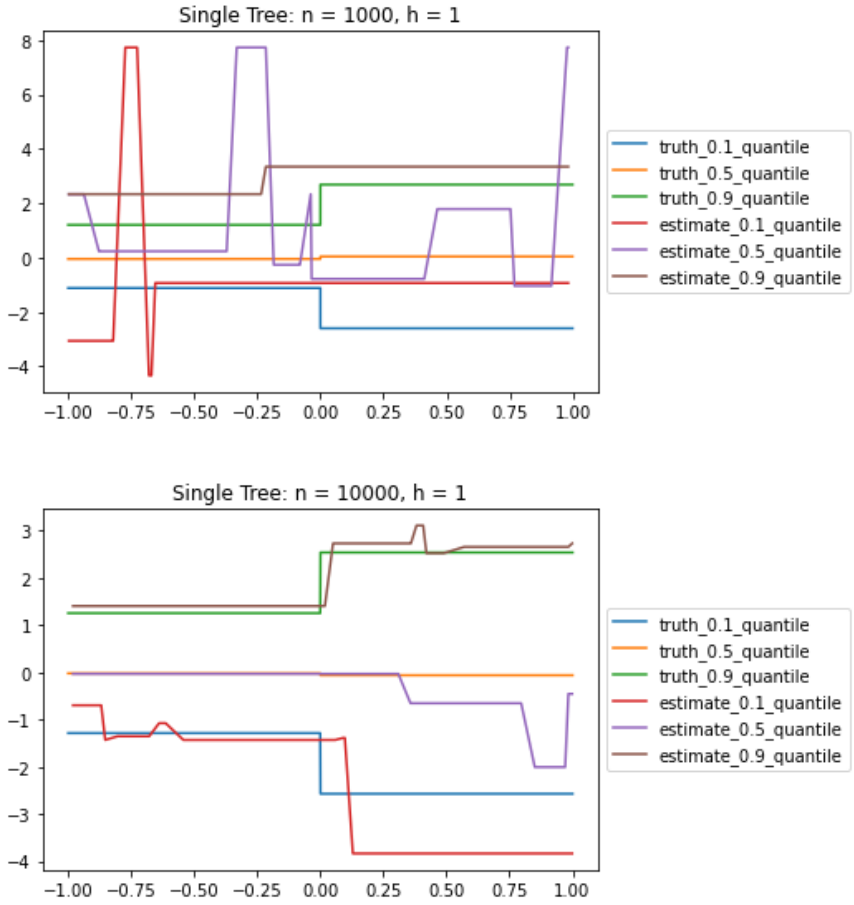

Figure 4.4: Single tree quantile estimation over the scale shift data with setting: number of trees nest = 1, maximum depth of trees maxDepth = 10, minimum number of samples in terminal node minLeafSample = 5, The number of features to consider when looking for the best split maxFeat = 5, bandwidth if the SEE $h = 1$. The size of the training data is adjusted to equal to 1000, 10000.

Since the bandwidth h can be adjusted to control the bias-variance tradeoff, a smaller training set would require h to be relatively large to restrict the variance of the estimations. Therefore, as we increase the training set, it is expected that by further adjusting the bandwidth h via hyperparameter tuning, the optimal bandwidth h should decrease, and the program should be able to yield more accurate estimations.

Chapter 5

Discussion and Future Works

By testing on simulated data, we were able to observe the variation in estimation performance when adjusting for various hyperparameters and training data. It is notable that increasing training sample size could significantly improve the estimation accuracy on the decision tree level, since larger sample size effectively reduces the variance in the estimation outcome. Meanwhile, it is worth mentioning that the estimation performance also differs across different quantiles. With the 0.5 quantile (median) having more accurate results, the 0.1 and 0.9 quantile estimations constantly produce more deviations from the truth value. Since the desired quantile τ is part of the smoothed moment condition, it is likely that the optimal bandwidth h differs for different input of the desired quantile. As a result, we would attempt to adjust for the optimal bandwidth h for quantile = 0.1, 0.5, and 0.9 separately instead of using the uniform bandwidth for three different quantiles.

Another essential aspect for future improvement is the runtime efficiency. With the current implementation of the GRF framework in Python, it is beyond the ability of our computer hardware to test our method on decision tree with even larger sample size of 50,000 or above. The computation requirement for aggregating hundreds of such decision trees' outcome into the random forest estimation would be more de-

manding. Therefore, a necessary future step to take is to improve the runtime of the current implementation. One direction is to find heuristics to speed up the tree splitting process. Another approach is to implement numerical estimation method specifically suited for solving our particular quantile moment condition. Since recursive tree split and numerical estimation are the two most time-consuming steps in our current implementation, improvement on either aspect should speed up the program execution substanially.

With a more efficiently implemented program, we would also attempt to continue examining in the following aspects:

First, our quantile estimation is still implemented based on the framework of random forest, which allows us to adjust various hyperparameters including not only the SEE bandwidth h, but also the maximum tree depth, minimum number of sample within terminal node, number of features considered in the best split search, and the number of trees. It would be expected that increasing the maximum tree depth would result in a smaller bias. Meanwhile, it might be necessary to increase the tree number simultaneously to avoid the resulting rise in variance due to the bias-variance tradeoff. The optimal approach would be adjusting all hyperparameters at the same time using grid search.

In addition, in the current research we use one realization of the simulated data to test the performance, so the resulting estimation also depends on how well this realization represent the ground-truth data distribution. As a response to this concern, we increased the number of observations to make the simulated data approach the ground-truth data distribution asymptotically. In the future step, we can also apply multiple realizations of the simulated data and obtain the averaged quantile estimations to reduce the estimation variance to a greater extent.

Chapter 6

Conclusion

The original quantile regression implemented in GRF by Athey et al. [1] requires special care in deriving pseudo-outcome for the tree splitting task, because of containing an indicator function which is non-smooth and non-differentiable. Using Kaplan and Sun [6]'s smoothed estimating equation, we were able to replace the indicator function with a smoothed $G()$ function with adjustable bandwidth to approximate the original moment condition. We took advantage of the generalized framework of the GRF to derive the pseudo-outcome for the smoothed new moment condition, with no need to alter the pseudo-outcome calculation process to accommodate non-differentiable moment condition.

Through a self-constructed Python implementation of the quantile GRF with SEE, we experimented on simulated data how a smoothed moment condition would affect the performance of quantile regression. It was observed that on the random forest level, the quantile GRF with SEE produces no positive effect on the estimation accuracy using the simulated data. Adjusting the bandwidth h receives only small improvement on the estimation performance. Therefore, we focus on a single decision tree in the implemented random forest to examine the mechanism more deeply. Inspection on

the decision tree level quantile estimation reveals that with increased training sample, the quantile estimation is able to approach satisfactory accuracy, implying that by increasing training sample size and tree number, the quantile GRF with SEE should also be able to yield much more accurate estimations.

Hence, we plan to further improve the program's run-time and produce better approach for hyperparameter tuning, such that the program can be executed with increased training samples and larger tree numbers in reasonable run-time. This would allow us to develop more understanding of the performance of our quantile GRF on the random forest level given sufficiently large training samples and tree numbers.

Bibliography

[1] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525, 2006.

[4] Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398, 2008.

[5] Antonella Costanzo and Marta Desimoni. Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using invalsi survey data. *Large-scale Assessments in Education*, 5(14), 2017.

[6] David M. Kaplan and Yixiao Sun. Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory*, 33(1):105–157, 2017.

[7] Roselinde Kessels, Anne Hoornweg, Thi Kim Thanh Bui, and Guido Erreygers. A distributional regression approach to income-related inequality of health in australia. *International Journal for Equity in Health*, 19(102), 2020.

[8] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1): 33–50, 1978.

[9] Nicolai Meinshausen. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983—-999, 2006.

[10] Tse-Chuan Yang, Vivian Yi-Ju Chen, Carla Shoff, and Stephen A. Matthews. Using quantile regression to examine the effects of inequality across the mortality distribution in the u.s. counties. *Social Science Medicine*, 74(12):1900 – 1910, 2012.