Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrics & Economics

Improving the forecasts for the distribution of GDP growth

Daan Lemmen

445526dl

Abstract

Instead of focusing on providing the best point forecast for GDP growth, this paper focuses on providing the best forecast for the full distribution of GDP growth. Adrian, Boyarchenko, and Giannone (2019) propose a two-step method to do this. First, estimates for quantiles of the distribution are made. Secondly, a parametric distribution is fit. This paper extends this two-step method in the following ways: Quantile Regression is used in combination with variable selection techniques and Principal Component Analysis. Alternatively, quantiles are estimated using Quantile Regression Forest. Besides the skewed t-distribution, also a Skew Generalized Secant Hyperbolic distribution is fit on the estimated quantiles. All of these extensions are tested for in-sample and out-of-sample predictions, on an extended dataset with high-dimensional predictor variables, and compared to the original method. I found that Quantile Regression Forest can significantly improve the forecasting performance.

Supervisor:

Prof.dr. D.J.C. van Dijk[1]

Second assessor:

P.A. Opschoor, MSc

July 4, 2020

[1] Thank you for your guidance, and useful feedback.

Contents

1 Introduction 1

2 Data 3

3 Methodology 4

3.1 Quantile Regression 5

3.2 Quantile Regression Forest 6

3.3 Random Forest 7

3.4 Skewed t-distribution 7

3.5 Skew Generalized Secant Hyperbolic (SGSH) distribution 8

3.6 Performance measures 9

3.7 Predictor variable pre-selection for Quantile Regression Forest 10

3.8 Selection of most significant predictor variables for Quantile Regression 10

3.9 Principal Component Analysis for Quantile Regression 11

4 Results 12

4.1 Quantile Regression 12

4.2 Quantile Regression Forest 13

4.3 Quantile Regression - most significant predictor variables 15

4.4 Quantile Regression - Principal Component Analysis 17

4.5 Quantile Regression Forest - predictor variable pre-selection 19

5 Conclusion 20

6 Discussion & Further research 21

References 23

7 Appendix 25

7.1 Implementation details 25

7.1.1 Quantile Regression 25

7.1.2 Quantile Regression Forest 25

7.1.3 Skewed t-distribution 25

7.1.4 SGSH distribution 25

7.1.5 Principal Component Analysis 25

7.2 Leave-one-out estimation 26

1 Introduction

GDP growth is considered one of the most important numerical characteristics of a macroeconomy. Therefore, being able to provide an accurate forecast for the GDP growth has been a popular subject for both economists and econometricians. However, most research that has been done focuses on providing the best point forecast. In this paper, the main focus is to provide the most accurate forecast for the distribution of GDP growth.

Having a prediction for the full distribution of GDP growth has several advantages over a point forecast. First and foremost, it provides more insight into the (asymmetrical) risk of predicted GDP growth. The user of the forecast can thus quantify the likelihood of risk scenarios. Secondly, the full distribution provides multiple point forecasts (e.g. mean or median). Thus, the user can choose which best fits its purpose.

To estimate an empirical distribution a fully parametric or non-parametric method can be used. Fully parametric methods have the downside of assuming linearity over the full distribution, or imposing a specific parametric structure. Fully non-parametric methods are very flexible and therefore tend to overfit the data (Adrian et al., 2019). Adrian et al. (2019) introduced a semi-parametric two-step method. First, four quantiles are estimated using Quantile Regression. In this Quantile Regression, GDP growth is the dependent variable, and the lagged GDP growth and National Financial Conditions Index (NFCI) are used as predictor variables. Secondly, a skewed t-distribution as developed by Azzalini and Capitanio (2003) is fitted on the estimated quantiles. Resulting in a forecast for the full distribution of GDP growth.

The publication of Adrian et al. (2019) got a lot of attention and triggered the IMF to develop the Growth at Risk framework which is based on that paper (Prasad et al., 2019). The IMF extends the work by incorporating more predictor variables from three groups: financial conditions, macrofinancial vulnerabilities, and other factors. Hence, the framework as developed by the IMF can incorporate more information than the original method of Adrian et al. (2019). For all of these three groups, many potential variables can be found. The IMF provides explanation and guidance on choosing predictor variables. However, no exact procedure is provided, and thus the choice is left up to the user. Instead of choosing the predictor variables by hand, I apply and compare multiple methods to estimate quantiles using a high-dimensional dataset without the need for the user to select predictor variables themselves.

To obtain many variables for all of these three groups, I use a high-dimensional dataset by the Federal Reserve Bank of St. Louis. This dataset is specifically developed for macroeconomic big data analysis. It contains many variables that are all related to the macroeconomy, and these variables are transformed such that unit root issues are solved.

Standard Quantile Regression has problems directly handling high-dimensional predictor variables due to multicollinearity and a lack of degrees of freedom. Thus, it is impossible to directly use a high-dimensional dataset in combination with standard Quantile Regression. Therefore, three alternatives are proposed. First, Quantile Regression is performed on the set of most significant predictor variables. These predictor variables are selected through a procedure which does not need any user input. Secondly, Quantile Regression is performed on a number of principal components of the full set of predictor variables. By using Principal

Component Analysis a high proportion of the total variance can be explained by a relatively small number of variables. Lastly, Quantile Regression Forest is used to estimate the quantiles. Quantile Regression Forest is a special kind of Quantile Regression that can directly handle high-dimensional predictor variables. Moreover, the performance of Quantile Regression Forest is best when the number of predictor variables is high.

As GDP growth has time-varying volatility and skewness (Kent & Phan, 2019), it is important to fit a distribution that is flexible in terms of skewness. As can be seen in Figure 1 the unconditional distribution also shows excess kurtosis. The conditional distribution also shows excess kurtosis and skewness (Adrian et al., 2019). When fitting a distribution for the GDP growth it is therefore important to consider a distribution that can incorporate this. The skewed t-distribution is flexible in terms of skewness and kurtosis, while it can equal a normal t-distribution and therefore also converge to a Gaussian distribution (Azzalini & Capitanio, 2003). An alternative is the Skew Generalized Secant Hyperbolic (SGSH) distribution. The SGSH distribution is a generalized version of the Hyperbolic Secant distribution, which has many of the same properties as the standard normal distribution. Like, the skewed t-distribution the SGSH distribution is also flexible in terms of skewness and excess kurtosis. Fischer (2004) fits the SGSH distribution both conditionally and unconditionally on asset returns. The fit provided by the SGSH distribution is described as excellent. Asset returns, both conditionally and unconditionally, also have a distribution with excess kurtosis and skewness. Therefore, I consider two parametric distributions: the skewed t-distribution and the SGSH distribution.

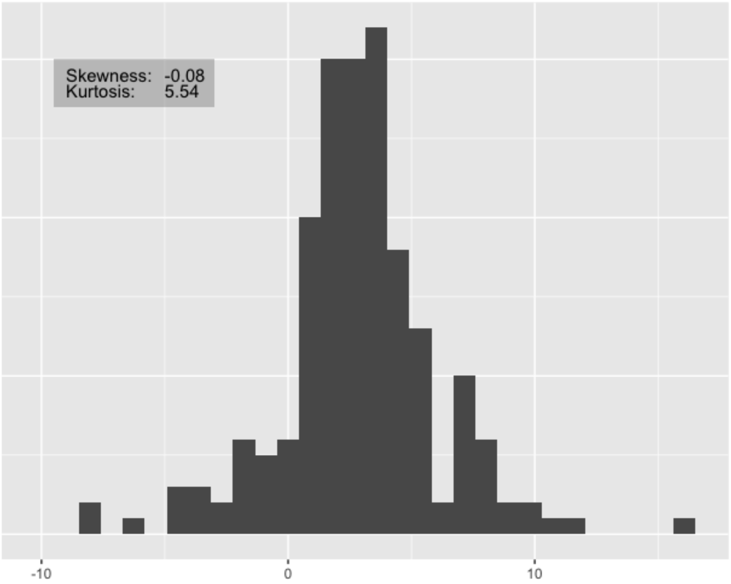

Figure 1: Unconditional empirical GDP growth distribution

The main focus of this research is to investigate if and how the two-step method of Adrian et al. (2019) can be improved. This is done by applying methods to this forecasting problem that have not been applied in the current literature. In short, this research contributes to the current literature by applying and comparing the performance of the following methods: using Quantile Regression Forest for the quantile estimates, using Quantile Regression with variable selection techniques for the quantile estimates, and fitting a Skew Generalized Secant Hyperbolic distribution on the estimated quantiles.

The paper will continue as follows: in section 2 the data and its sources are described, in section 3 the methodology is explained, in section 4 the results are described, in section 5 the final comparison is described and conclusions are drawn, and in section 6 the limitations are discussed and suggestions for further research are done.

2 Data

All data necessary for this research can be obtained from the Federal Reserve Economic Data (FRED) database of the Federal Reserve Bank of St. Louis. Most data is available monthly or at a higher frequency, and has historical data from 1960 onwards. Since GDP growth is only published quarterly, all data which is available in a higher frequency will be converted into quarterly data.[2] The publication of historical data for the NFCI dates back to the first week of 1971. The investigation period in this paper is therefore chosen to be 1971Q1-2019Q4.

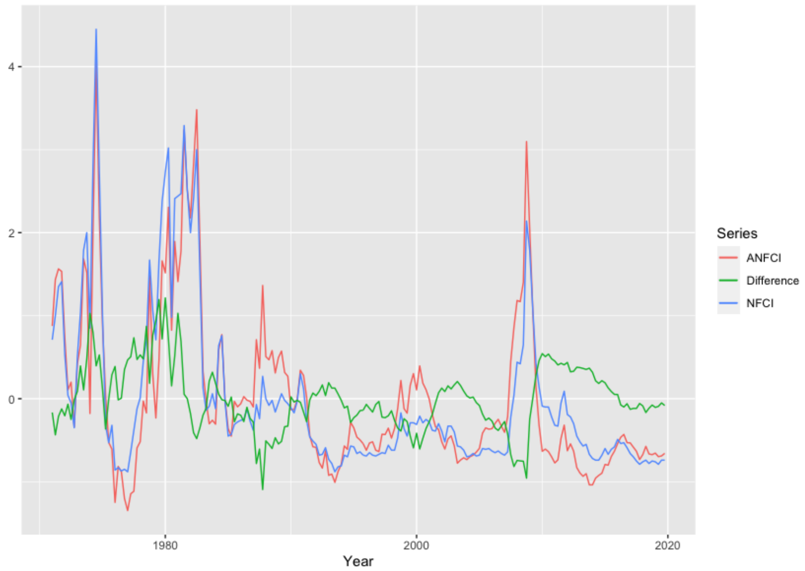

Figure 2: Comparison of the NFCI and ANFCI

[2] The higher frequency data is averaged within each quarter. For attributions of overlapping periods, the conversion standard of the Federal Reserve Economic Data is followed. Which states that attributions of higher frequency data count towards the period of the lower frequency data in which the higher frequency data ends. For example, data of week 40 2019 which begins on Monday, September 30 and ends on Friday, October 4 counts towards the fourth quarter of 2019.

The NFCI is a weekly estimate provided by the Federal Reserve Bank of Chicago for the US financial conditions in money markets, debt and equity markets, and the traditional and 'shadow' banking systems (*National Financial Conditions Index (NFCI)*, 2020). Since economic and financial conditions are correlated, the Adjusted National Financial Conditions Index (ANFCI), which has been revised in 2017, will also be considered (Brave & Kelly, 2017). The ANFCI isolates the component of financial conditions from the economic conditions to provide more insights into the current state of the financial conditions. Figure 2 shows a comparison between the NFCI and ANFCI. Notably, the ANFCI is above the NFCI in the two most recent crashes (dot-com bubble and financial crisis 2007). Since both the NFCI and ANFCI measure how tight financial conditions are a lower value for both indicates more relaxed financial conditions.

The full set of predictor variables includes the factors of the Leading Economic Index (*Leading Economic Index*, 2020), the NFCI, the ANFCI, and the variables available in the FRED-QD database (McCracken & Ng, 2020).

The FRED-QD database is a macroeconomic database containing 248 variables designed for big data analysis. This database is maintained and published by the Federal Reserve Bank of St. Louis. Quarterly, the database is revised and extended to include the most recent data. The 248 variables of the FRED-QD database fall into 14 groups: National Income and Product Accounts; Industrial Production; Employment and Unemployment; Housing; Inventories, Orders, and Sales; Prices; Earnings and Productivity; Interest Rates; Money and Credit; Household Balance Sheets; Exchange Rates; Other; Stock Markets; and Non-Household Balance Sheets. All these variables are related to the US macroeconomy. The publisher performs unit root tests on the variables contained in the database, and the variables are transformed accordingly. The appendix of McCracken and Ng (2020) contains a list of all variables included in the database and their transformation codes.

The full set of predictor variables is cleaned up by removing all duplicate predictor variables, which are identified by a 100% correlation. To ensure consistency over the full sample period, predictor variables that have missing values in the chosen sample period, are removed. The date variable that is contained in the FRED-QD will not be used as a predictor variable. This reduces the set of predictor variables to 242.

3 Methodology

In this paper, several methods of forecasting the full distribution of GDP growth are compared. As a base case, the method of Adrian et al. (2019) is performed on the extended and revised dataset. Then this method is extended in the following ways. First, I use more and different predictor variables in Quantile Regression. Secondly, I use three sets of quantiles the skewed t-distribution is fit. Thirdly, I also fit a SGSH distribution on the estimated quantiles. And lastly, I use Quantile Regression Forest instead of Quantile Regression to estimate the quantiles.

Out-of-sample predictions are made using an expanding window. The first out-of-sample prediction is for 1995Q1, which is predicted using the observations 1972Q1-1994Q4. For each observation after this, the window is expanded, one quarter at the time. At each iteration, the quantiles are estimated using Quantile Regression or Quantile Regression Forest, and a skewed t-distribution and SGSH distribution are fitted on the estimated quantiles. After the last iteration, the performance of the forecast is evaluated using the performance measures described in section 3.6.

For each estimation method, I also perform an in-sample distribution estimate. This is done using the full sample, i.e. 1972Q1-2019Q4. After the quantiles are estimated, both a skewed t-distribution and SGSH distribution are fitted on these quantiles, and the performance of the forecast is evaluated.

All forecasts, both in-sample and out-of-sample, made and evaluated in this research are 1-step ahead forecasts. The dependent variable, GDP growth, is published quarterly. Thus all forecasts are one quarter ahead, i.e. the values of the predictor variables for a given quarter are used to predict the value of the dependent variable for the next quarter.

Besides the two forecasting methods described above, I also performed leave-one-out out-of-sample estimations. The details and results are described in Appendix section 7.2.

3.1 Quantile Regression

Quantile Regression is a technique that makes it possible to estimate conditional quantiles by a simple minimization problem. The minimization problem is given in equation 1. In this equation, the absolute errors are minimized instead of the squared errors that are minimized by Ordinary Least Squares. Quantile Regression is therefore considered more robust to outliers. The prediction of the τ quantile for y_{t+1} conditional on x_t is then given by equation 2.

Quantile Regression is performed on the full data set for the 5%, 15%, 25%, 50%, 75%, 85% and 95% quantiles. Initially, the predictor variables (x_t) are the lagged NFCI and GDP growth, like Adrian et al. (2019) used in their method. Besides this set of predictor variables, I use two methods to obtain other predictor variables. First, I use the method described in section 3.8 to select the five and seven most significant, not (nearly) perfectly correlated predictor variables out of many predictor variables. Secondly, I use a number of principal components of the full set of predictor variables as predictor variables in Quantile Regression. This method is described in detail in section 3.9.

$$\widehat{\beta}_\tau = \underset{\beta_\tau \in R^k}{\arg\min} \sum_{t=1}^{T-1} \mathbf{1}_{(y_{t+1} \geq x_t\beta_\tau)} \cdot \tau \cdot |y_{t+1} - x_t\beta_\tau| + \mathbf{1}_{(y_{t+1} < x_t\beta_\tau)} \cdot (1-\tau) \cdot |y_{t+1} - x_t\beta_\tau| \tag{1}$$

$$\widehat{Q}_\tau(x_t) = x_t\widehat{\beta_\tau} \tag{2}$$

where

τ : quantile

$\widehat{Q}_\tau(x_t)$: estimated τ quantile for y_{t+1} conditional on x_t

$\widehat{\beta}_\tau$: estimated coefficients for τ quantile

y_t : dependent variable

x_t : vector of predictor variables

3.2 Quantile Regression Forest

An alternative to Quantile Regression, named Quantile Regression Forest, is developed by Meinshausen (2006). Quantile Regression Forest is a machine learning algorithm which extends the concept of Random Forest (Breiman, 2001). The concept of Random Forest is explained in short in section 3.3. For more details, I refer to Breiman (2001).

The most significant difference between Quantile Regression Forest and Random Forest is that, at each leaf, Quantile Regression Forest takes note of all observations at that leaf instead of only the conditional mean at that leaf.

Quantile Regression Forest estimates the empirical cdf conditional on x, $F(y|X = x)$, by dropping x down each regression tree and calculating the proportion of observations at the selected leafs that are below y, see equation 4. The quantiles are then inferred from this empirical cdf by taking the infimum, see equation 3.

$$\widehat{Q}_\tau(x) = \inf\{y : \widehat{F}(y \mid X = x) \geq \tau\} \tag{3}$$

$$\widehat{F}(y \mid X = x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}} \tag{4}$$

$$w_i(x) = \frac{1}{k} \sum_{t=1}^{k} w_i(x, \theta_t) \tag{5}$$

where

n : number of observations

k : number of trees

Y_i : value of the dependent variable of observation i

$w_i(x, \theta_t)$: weight vector as in Random Forest, which is a positive constant if observation i is part of the same leaf as an observation with $X = x$, and zero otherwise.

Quantile Regression Forest is performed to estimate the following quantiles: 5%, 15%, 25%, 50%, 75%, 85% and 95%. Since regression trees can make very specific leaves, I exclude the observation that is predicted from the training set. This causes that for every observation a Quantile Regression Forest is estimated.

Quantile Regression Forest has a number of hyperparameters that can be tuned. The two main hyperparameters are `mtry` and `nodesize`. `mtry` corresponds to the number of predictor variables that are considered at each split point. `mtry` tends to work well for a large range of values (Meinshausen, 2006). Therefore, the standard value of $p/3$ is used, and this hyperparameter is not tuned because of limited computing resources. The hyperparameter `nodesize` corresponds to the minimum number of observations in a leaf (terminal node). Setting a low `nodesize` grows larger trees, i.e. more leaves, and has a better in-sample fit. However, a low `nodesize` might have a worse out-of-sample fit due to overfit to the sample data. The standard `nodesize` is 5. I will test the following values for this hyperparameter: 5, 10, 20, 25, 30, 35, and 40.

Because the dataset contains a large number of predictor variables I will test if Quantile Regression Forest can be improved by doing a pre-selection of the predictor variables. The method used for this pre-selection is described in section 3.7.

3.3 Random Forest

In Random Forest[3] a large number of regression trees are grown, each with a bootstrapped sample of the data. The prediction of the Random Forest is the average prediction of each regression tree.

A regression tree starts off with one leaf which contains all observations. Leaves are then split until a split can no longer improve the prediction, or the number of observations at the leaf does not exceed the hyperparameter `nodesize`. The prediction of a single leaf is the conditional mean of all observations at that leaf.

Splitting a leaf works as follows: A random subset of size `mtry` is selected from the set of predictor variables. The variable of that random subset that can best split the data at that leaf is then chosen as the split variable. The error that is minimized is the sum of squared residuals.

3.4 Skewed t-distribution

The skewed t-distribution is a distribution that is flexible in terms of excess kurtosis and skewness. Its probability density function (pdf) is given by equation 6. The skewed t-distribution has four parameters: the location parameter μ, the scale parameters σ, the slant (skewness) parameter α and the degrees of freedom ν.

$$f(y;\mu,\sigma,\alpha,\nu)=\frac{2}{\sigma}t\left(\frac{y-\mu}{\sigma};\nu\right)T\left(\alpha\frac{y-\mu}{\sigma}\sqrt{\frac{\nu+1}{\nu+\left(\frac{y-\mu}{\sigma}\right)^2}};\nu+1\right) \tag{6}$$

where

$t(y;\nu)$: probability density function (pdf) of the standard t-distribution

$T(y;\nu)$: cumulative density function (cdf) of the standard t-distribution

A skewed t-distribution is fitted on three subsets of the estimated quantiles, such that for each quantile estimate three skewed t-distributions are fitted. The three subsets of the estimated quantiles are given in

[3] Random Forest as discussed in this paper is the regression version of Random Forest. There is also a classification version of Random Forest which is not applicable to this paper.

Table 1. Note that the set of four quantiles is the set of quantiles as used by Adrian et al. (2019). From now on I refer to these sets of quantiles by their cardinality. The set of 4 quantiles gives the most weight to the tails of the distribution. Where the set of 5 quantiles gives more weight to the median of the distribution. The set of 7 quantiles gives most weight to the tails of the distribution, while also giving weight to the median.

# quantiles	Quantiles
4	5%, 25%, 75%, 95%
5	5%, 25%, 50%, 75%, 95%
7	5%, 15%, 25%, 50%, 75%, 85%, 95%

Table 1: Sets of quantiles

Fitting the skewed t-distribution is done by minimizing the sum of squared differences between the parametric distribution quantiles and the estimated quantiles. This minimization problem is described mathematically in equation 7. As the distribution has four parameters, the parameters are exactly identified when fit using the set of 4 quantiles. When the distribution is fit on 5 or 7 quantiles, the parameters are over-identified.

$$\{\widehat{\mu}_{t+h}, \widehat{\sigma}_{t+h}, \widehat{\alpha}_{t+h}, \widehat{\nu}_{t+h}\} = \underset{\mu,\sigma,\alpha,\nu}{\arg\min} \sum_{\tau} \left(\widehat{Q}_{\tau}(x_t) - F^{-1}(\tau; \mu, \sigma, \alpha, \nu) \right)^2 \tag{7}$$

where

F^{-1} : the inverse cumulative distribution function (CDF) of the skewed t-distribution

3.5 Skew Generalized Secant Hyperbolic (SGSH) distribution

As an alternative to the skewed t-distribution, the SGSH distribution is fit. Its probability density function (pdf) is given by equation 8. The SGSH distribution has four parameters: the location parameter μ, the scale parameter σ, the slant (skewness) parameter s and the kurtosis parameter t.

$$f_{SGSH}(x; \mu, \sigma, s, t) = \frac{2c_1}{s + \frac{1}{s}} \left(\frac{\exp(c_2 \frac{x-\mu}{\sigma s}) \mathrm{I}^{-}(\frac{x-\mu}{\sigma})}{\exp(2c_2 \frac{x-\mu}{\sigma s}) + 2a \exp(c_2 \frac{x-\mu}{\sigma s}) + 1} + \frac{\exp(c_2 \frac{x-\mu}{\sigma} s) \mathrm{I}^{+}(\frac{x-\mu}{\sigma})}{\exp(2c_2 \frac{x-\mu}{\sigma} s) + 2a \exp(c_2 \frac{x-\mu}{\sigma} s) + 1} \right) \tag{8}$$

$$a(t) = \cos(t), \qquad c_2(t) = \sqrt{(pi^2 - t^2)/3},\ c_1(t) = c_2(t)\sin(t)/t, \qquad \text{for } -\pi < t \leq 0$$

$$a(t) = \cosh(t), \qquad c_2(t) = \sqrt{(pi^2 + t^2)/3},\ c_1(t) = c_2(t)\sinh(t)/t, \qquad \text{for } t > 0$$

where

$\mathrm{I}^{-}(x)$: indicator function for x on $\mathbb{R}^{-}$

$\mathrm{I}^{+}(x)$: indicator function for x on $\mathbb{R}^{+}$

The estimated parameters for the distribution for each observation are obtained by the minimization problem described in equation 9. In this minimization problem the sum of squared differences between the quantiles of the parametric distribution and the estimated distribution are minimized. This minimization is done for three subsets of the estimated quantiles, see Table 1. The SGSH distribution has four parameters. Therefore, the set of 4 quantiles exactly identifies the distribution, and the sets of 5 and 7 quantiles over-identify the distribution.

$$\left\{\widehat{\mu}_{t+h}, \widehat{\sigma}_{t+h}, \widehat{s}_{t+h}, \widehat{t}_{t+h}\right\} = \underset{\mu,\sigma,s,t}{\arg\min} \sum_{\tau} \left(\widehat{Q}_{\tau}(x_t) - F^{-1}(\tau; \mu, \sigma, s, t)\right)^2 \tag{9}$$

where

F^{-1} : the inverse cumulative distribution function (CDF) of the SGSH distribution

3.6 Performance measures

To compare the forecasting accuracy of different distributions the test suggested by Mitchell and Hall (2005) will be used. This test can formally test which of two distribution forecasts has a better accuracy and if this difference is significant. For a given distribution forecast, $\widehat{f}$, the log-likelihood for each observation $t = 1, ..., T-1$ is calculated by equation 10.

$$s^{\text{CL}}(\widehat{f}_t; y_{t+1}) = \ln(\widehat{f}_t(y_{t+1})) \tag{10}$$

$$S^{\text{CL}}(\widehat{f}) = \frac{1}{T-1} \sum_{t=1}^{T-1} s^{\text{CL}}(\widehat{f}_t; y_{t+1}) \tag{11}$$

To compare forecasts, the mean log-likelihood for each distribution forecast $S^{\text{CL}}(\widehat{f})$ is compared. A higher $S^{\text{CL}}(\widehat{f})$ indicates a better performing forecast. To formally test this a Diebold-Mariano like test is performed using the log-likelihoods.

The procedure described above evaluates the forecast over the full distribution. For some purposes, the forecasting performance of negative GDP growth is most important. Therefore, I also evaluate the forecasting accuracy of solely negative GDP growth. Diks, Panchenko, and Van Dijk (2011) present a method to evaluate forecasting accuracy in a specific region. For a given estimate of the distribution $\widehat{f}$, the log-likelihood for each observation $t = 1, ..., T-1$ is calculated by equation 12. The comparison and formal test based on log-likelihoods are similar to those for the full distribution.

$$s^{\text{cl}}_{\mathbb{R}^-}(\widehat{f}_t; y_{t+1}) = \text{I}(y_{t+1} \in \mathbb{R}^-) \ln\left(\frac{\widehat{f}_t(y_{t+1})}{\int_{\mathbb{R}^-} \widehat{f}_t(s) ds}\right) \tag{12}$$

$$S^{\text{cl}}_{\mathbb{R}^-}(\widehat{f}) = \frac{1}{T-1} \sum_{t=1}^{T-1} s^{\text{cl}}_{\mathbb{R}^-}(\widehat{f}_t; y_{t+1}) \tag{13}$$

where

I$(\cdot)$: indicator function

3.7 Predictor variable pre-selection for Quantile Regression Forest

Since the FRED-QD dataset contains many predictor variables, it might be favorable to do a pre-selection of the predictor variables to exclude variables with a low predictive power. Quantile Regression Forest splits on a random subset of the predictor variables. Therefore, it could be possible that this random subset only contains predictor variables with no or low predictive power. This could be overcome by excluding these predictor variables beforehand.

To determine the marginal predictive power of each predictor variable, I use a method similar to the method of Bai and Ng (2008). When the marginal predictive power of each predictor variable is determined, Quantile Regression Forest is performed on various subsets of predictor variables.

1. Regress y_t on a constant, y_{t-1} and $x_{i,t-1}$ for $\forall x_i \in X$
2. Order X on the marginal predictive power, i.e. $|t_1|, |t_2|, ..., |t_N|$ in descending order

where

X : $T \times N$ predictor matrix

x_i : $T \times 1$ column vector of X

y_t : value of the dependent variable of observation t

t_i : t-statistic of x_i in the regression

The regression in step 1 is an Ordinary Least Squares (OLS) regression, and the standard errors used to calculate the t-statistic are the standard OLS standard errors.

For in-sample estimation, the marginal predictive power is determined using the full sample. For out-of-sample estimation, the marginal predictive power is determined at each iteration using the in-sample data, which is expanded by one quarter each iteration. Therefore, the selected predictor variables might differ at each iteration.

3.8 Selection of most significant predictor variables for Quantile Regression

Quantile Regression is limited in the number of predictor variables it can handle, due to multicollinearity and degrees of freedom. It is therefore impossible to use all predictor variables of the FRED-QD dataset with Quantile Regression. I therefore use a method to select the most significant variables of the FRED-QD dataset that are not (nearly) perfectly correlated. The following algorithm is used to select the k most significant predictor variables that are not (nearly) perfectly correlated.

1. Perform Quantile Regression of y_t on a constant, y_{t-1} and $x_{i,t-1}$ for $\forall x_i \in X$ and $\forall \tau \in T$
2. Order X on the marginal predictive power averaged over the quantiles, i.e. $|t_1|, |t_2|, ..., |t_N|$ in descending order
3. Add x_1 to `predictors`
4. Loop over $i \in [2, N]$
    (a) Calculate the maximum correlation between x_i and all variables in `predictors`
    (b) If the maximum correlation is below 80%; add x_i to `predictors`
    (c) If the length of `predictors` equals k; terminate

where
T : the set of quantiles
X : $T \times N$ predictor matrix
x_i : $T \times 1$ column vector of X
y_t : value of the dependent variable of observation t
$t_{i,t}$: t-statistic of x_i in the regression for $\tau = t$
t_i : t-statistic of x_i averaged over t, i.e. $t_i = \frac{1}{|T|} \sum_{t \in T} t_{i,t}$

The standard errors used in step 2 to calculate the t-statistic are the bootstrapped standard errors. Bootstrapping is done by the xy-pair method, which maintains the pairs of dependent variables and predictor variables. For each regression 200 bootstrap repetitions are performed.

3.9 Principal Component Analysis for Quantile Regression

An alternative to selecting a limited number of variables for Quantile Regression is Principal Component Analysis (PCA). PCA is a technique to reduce the dimensionality of a dataset while maintaining most of the variation (Ringnér, 2008). After performing PCA on the dataset containing the predictor variables, I will perform Quantile Regression with the first three, five, seven, ten, and fifteen components as predictor variables. The idea behind this technique is to use as much information as possible in the Quantile Regression.

4 Results

4.1 Quantile Regression

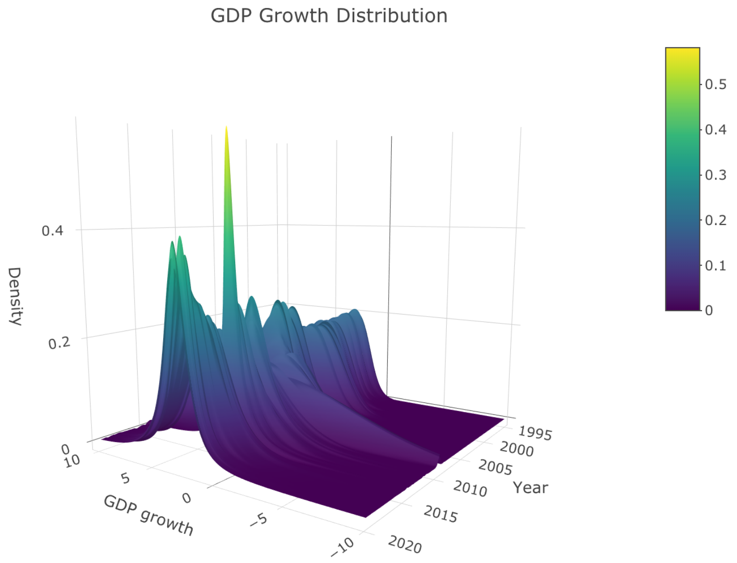

Figure 3: Quantile Regression, skewed t-distribution, 4 quantiles

First, quantiles are estimated using Quantile Regression with the same predictor variables as Adrian et al. (2019), i.e. lagged GDP and NFCI. Figure 3 shows the out-of-sample estimated skewed t-distribution fitted on 4 quantiles. The mode of most distributions is around 3. Interestingly, the variance decreases as the window expands, this can be seen in the figure by the increased density peak. As expected, the distribution has higher variance and is shifted to the left in the period corresponding to the 2007-2008 financial crisis.

For observation 2014Q2 a higher peak in the distribution is observed. For this observation the quantiles are close to each-other, and thus the variance in the distribution is low. The reason why the estimated quantiles are so close to each-other seems to be unknown.

Using the same estimated quantiles, a skewed t-distribution and a SGSH distribution are also fit on the three subsets of the estimated quantiles. The mean log scores are reported in Table 2. For the out-of-sample performance, the differences are small and not significant. Even though the differences are small, the skewed t-distribution performs better for all quantiles. For both distributions, I see that if the distribution is fit on a higher number of quantiles the performance for the negative growth forecast is better.

	In-sample				Out-of-sample			
	Skewed t		SGSH		Skewed t		SGSH	
# quantiles	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$
4	-2.33	-1.93	-2.33	-1.90	-2.19	-1.85	-2.21	-1.99
5	-2.33	-1.92	-2.33	-1.89	-2.19	-1.82	-2.21	-1.96
7	-2.33	-1.89	-2.36	-1.78	-2.19	-1.81	-2.21	-1.91

Table 2: Quantile Regression with lagged GDP and NFCI

4.2 Quantile Regression Forest

	In-sample				Out-of-sample			
	Skewed t		SGSH		Skewed t		SGSH	
`nodesize`	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}^-}$
5	**-0.90**	**-1.19**	**-0.89**	**-1.13**	-2.31	-2.37	-2.37	-3.79
10	-1.33	-1.43	-1.41	-1.49	-2.17	-2.54	-2.27	-3.47
20	-1.84	-1.69	-1.83	-1.62	-2.15	-3.33	-2.19	-2.48
25	-1.96	-1.76	-1.96	-1.70	-2.13	-1.97	-2.16	-2.69
30	-2.04	-1.79	-2.03	-1.77	-2.14	-2.03	-2.16	-2.30
35	-2.06	-1.80	-2.05	-1.75	**-2.13**	**-1.82**	-2.14	**-1.94**
40	-2.11	-1.83	-2.09	-1.78	-2.13	-2.04	**-2.13**	-2.41

Table 3: Hypertuned Quantile Regression Forest

Alternatively, quantiles are estimated using Quantile Regression Forest. First, I tune the hyperparameter `nodesize`. All variables are included and the distributions are fit on the set of five quantiles. The performances are compared for `nodesize` $= 5, 10, 20, 25, 30, 35, 40$. The results are shown in Table 3.

For out-of-sample estimation `nodesize` $= 35$ performs best for the skewed t-distribution, and `nodesize` $=$ 40 performs best for the SGSH distribution. However, only the difference between `nodesize` $= 35$ and 5 is significant. That these high values for `nodesize` perform the best is probably because Quantile Regression Forest tends to overfit to the in-sample (trainings) data, when the value for the hyperparameter `nodesize` is low.

For in-sample estimation `nodesize` $= 5$ performs significantly better than all other values for `nodesize`. This is in line with the thinking that Quantile Regression Forest tends to overfit to the in-sample (trainings) data, good performance for in-sample forecasting is then expected.

	In-sample				Out-of-sample			
	Skewed t		SGSH		Skewed t		SGSH	
# quantiles	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$
4	-2.06	-1.83	-2.05	-1.79	-2.17	-2.11	-2.18	-2.10
5	-2.06	-1.80	-2.05	-1.75	-2.13	-1.82	-2.14	-1.94
7	-2.06	-1.80	-2.04	-1.75	-2.13	-1.90	-2.13	-1.91

Table 4: Quantile Regression Forest, `nodesize` $= 35$

As my goal is to provide the best out-of-sample forecast, I further investigate `nodesize` $= 35$. To see whether the set of quantiles has an impact on the performance, I fit both a skewed t-distribution and SGSH distribution on the three sets of quantiles. The results are shown in Table 4. The differences in performance are small and not significant. This indicates that the number of quantiles does not impact the performance by much.

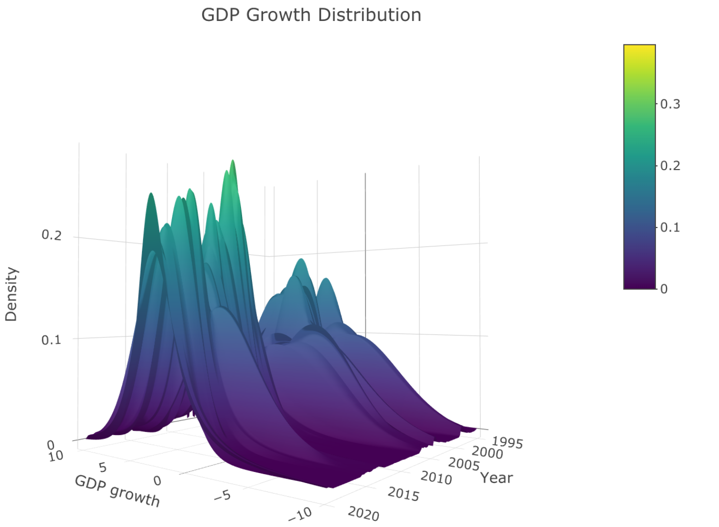

Figure 4: Quantile Regression Forest, skewed t-distribution, 5 quantiles

Figure 4 shows the estimated distribution by Quantile Regression Forest in combination with a skewed t-distribution on five quantiles. Like the estimated distributions by Quantile Regression, the variance decreases as the window is expanded. However, the modes of the distributions of Quantile Regression Forest are lower. Another interesting fact is that not only the distributions around the 2007-2008 financial crisis have higher variance and are shifted to the left. The estimates by Quantile Regression Forest show the same characteristics for the estimates around 2000. This period corresponds to the dot-com bubble burst.

4.3 Quantile Regression - most significant predictor variables

Once again, quantiles are estimated again by Quantile Regression. However, now using the five and seven most significant predictor variables. For the in-sample fit the following predictor variables are chosen by the procedure described in section 3.8:

`GDP`	GDP growth
`CPF3MTB3Mx`	3-Month Commercial Paper Minus 3-Month Treasury Bill
`T5YFFM`	5-Year Treasury Constant Maturity Minus Federal Funds Rate
`A014RE1Q156NBEA`	Change in private inventories (Gross private domestic investment)
`PERMITW`	New Private Housing Units Authorized by Building Permits
`UMCSENTx`	Consumer Sentiment
`USALOLITONOSTSAM`	Leading Indicators OECD, normalised for the United States

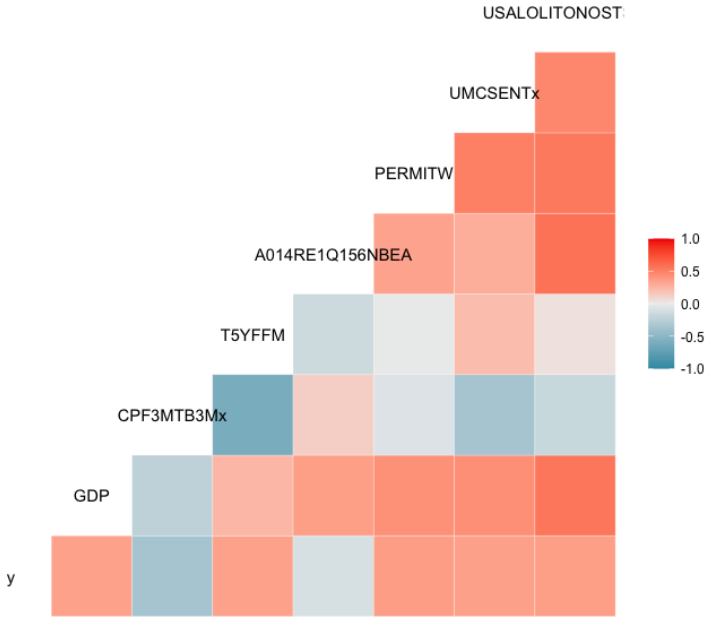

Figure 5: Correlation of seven most significant variables

Figure 5 shows the correlation between the dependent variable, GDP growth, and these lagged predictor variables. Notably, neither the NFCI or ANFCI is selected by the procedure. This is due to the high correlation of 82% and 88% between 3-Month Commercial Paper Minus 3-Month Treasury Bill and NFCI and ANFCI respectfully.

Stock and W Watson (2003) concluded that the term spread was the most useful predictor for US economic output. In line with this, the variable selection procedure used selected `T5YFFM`. `T5YFFM` is the difference between the 5-Year Treasury Constant and the Federal Funds Rate, and thus can be seen as the spread between the interest rate on long-term government bonds and the overnight interest rate. Other useful predictors included the Federal Funds Rate and Real Money Stock (M0, M1, M2, and M3). None of these predictors were selected by the variable selection procedure.

The Conference Board's Leading Economic Index is a combination of ten factors. These factors include measures for average weekly hours, average weekly initial claims for unemployment insurance, new orders, building permits, S&P500, Leading Credit Index, term spread, and consumer expectation. A couple of these factors are also selected by the proposed procedure as useful predictor variables.

For the out-of-sample fit, the predictor variables are re-selected for every forecast, as the window expands one quarter at the time.

		In-sample				Out-of-sample			
		Skewed t		SGSH		Skewed t		SGSH	
# predictors	# quantiles	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}^-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}^-}$
5	4	-2.32	-2.00	-2.33	-1.97	-4.17	-2.69	-4.62	-3.69
5	5	-2.32	-2.00	-2.32	-2.00	-3.68	-2.52	-4.51	-6.88
5	7	-2.33	-1.96	-2.31	-1.93	**-2.74**	**-1.81**	**-3.64**	**-1.86**
7	4	-2.32	-1.96	-2.33	-1.85	-3.50	-1.96	-4.35	-1.93
7	5	**-2.22**	-1.91	-2.41	-2.51	-3.93	-2.57	-6.99	-6.12
7	7	-2.27	**-1.89**	**-2.28**	**-1.83**	-5.05	-5.84	-5.01	-10.24

Table 5: Quantile Regression with most significant predictor variables

Table 5 shows the fit of quantiles estimated by Quantile Regression with the five and seven most significant predictor variables. On these quantiles both a skewed t-distribution and a SGSH distribution are fit. For out-of-sample forecasting, the performance of the model with five predictors (7 quantiles, skewed t-distribution) is significantly better than the performance of the model with seven predictors. However, the mean log scores differ very much between the number of quantiles the distribution is fit on. Also both Quantile Regression with GDP and NFCI, and Quantile Regression Forest significantly outperform the best model of Quantile Regression with variable selection.

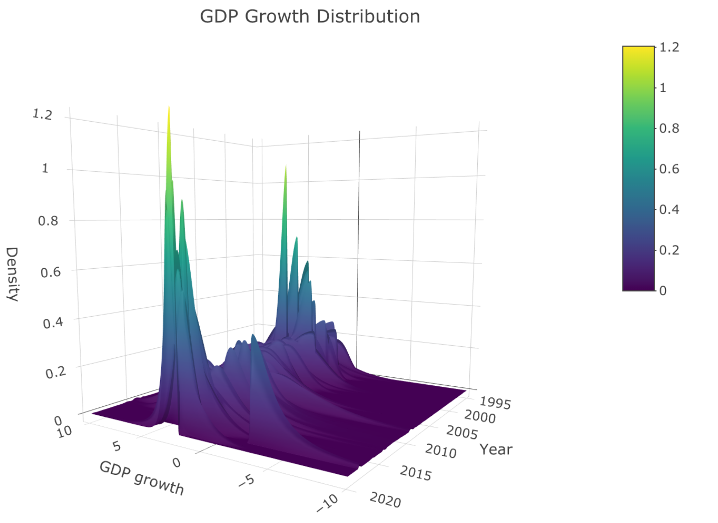

Figure 6: Quantile Regression, five predictors, skewed t-distribution, 7 quantiles

Figure 6 shows the best performing model in this class; five predictors with a skewed t-distribution fit on 7 quantiles. Notably, the variance of some distributions is extremely low. This causes high density peaks for the mode of the distribution. When the actual value of GDP growth is very close to the mode, the distribution provides an extremely good fit. However, in many cases, the actual value is not so close to the mode and thus, the log score is very low. The low variance in these distributions is caused by the estimated quantiles which are very close to each-other for these observations. The estimated quantiles are probably so close to each-other for some observations due to overfitting to the in-sample data.

4.4 Quantile Regression - Principal Component Analysis

Principal Component Analysis is applied to the predictor variables. Figure 7 shows the variance explained by the first 10 principal components. The first principal component explains over 65% of the total variance, the first 5 components explain roughly 90% of the variance, and the first 10 components explain over 95% of the variance. In contrast to the predictor variables, the principal components are orthogonal. Thus, when Quantile Regression is performed on principal components there is no issue of multicollinearity. The marginal variance explained by the 11th principal component is very low (0.5%). Therefore, I performed Quantile Regression with the first 3, 5, 7, and 10 principal components as predictor variables. Table 6 shows the results.

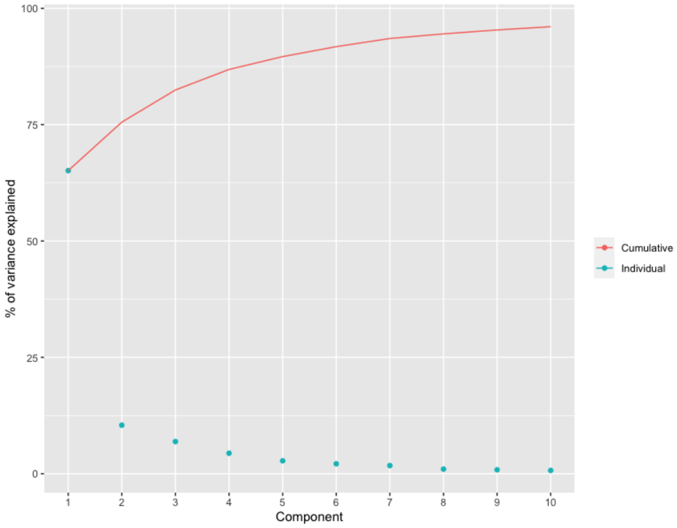

Figure 7: Variance explained by principal components

For in-sample estimation, the performance is better when more components are included. However, for out-of-sample estimation, the opposite is observed. The best performing model is the model with the least principal components. Probably, the model overfits to the in-sample data causing a good fit for the in-sample data but worse performance for out-of-sample forecasting.

For the out-of-sample SGSH-distribution fitted on 7 quantiles estimated using 10 principal components, the mean log score equals -Inf. This is due to observation 2009Q1. The estimated quantiles for this observation are in the range $[-16, -10]$ and therefore the distribution has a peak around 13 and its variance is relatively low. However, the actual value for this observation is -4.4, causing a log score of -Inf. The skewed t-distribution also its mode for this observation around 13, however, the variance is higher.

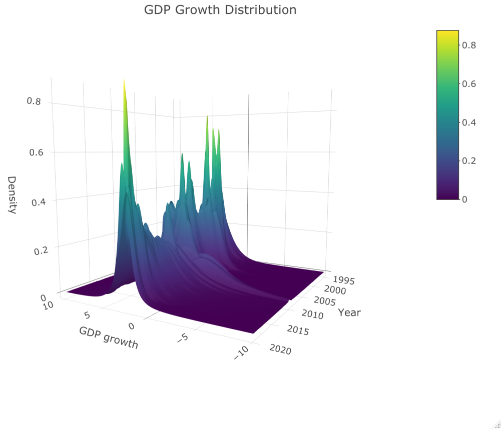

Figure 8: Quantile Regression, 3 principal components, skewed t-distribution, 4 quantiles

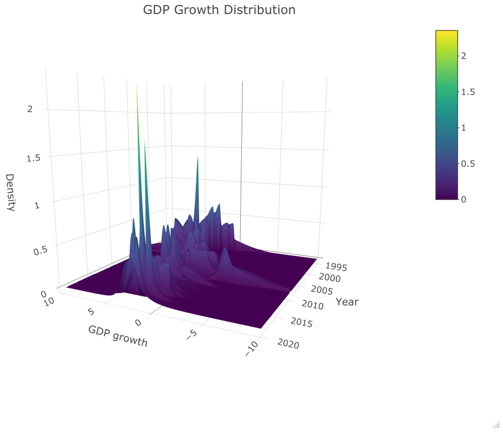

Figure 9: Quantile Regression, 10 principal components, skewed t-distribution, 4 quantiles

Figures 8 and 9 show the out-of-sample skewed t-distributions fitted on 4 quantiles estimated using 3 and 10 principal components respectfully. Notably, the distributions estimated using 10 principal components have a much lower variance than when estimated using 3 principal components. This clearly shows the cause of the worse out-of-sample performance.

		In-sample				Out-osample			
		Skewed t		SGSH		Skewed t		SGSH	
# components	# quantiles	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\text{cl}}_{\mathbb{R}-}$
3	4	-2.39	-2.05	-2.38	-2.02	**-2.38**	-2.05	**-2.31**	-2.13
3	5	-2.39	-2.05	-2.38	-2.01	-2.39	-2.04	-2.34	-2.10
3	7	-2.39	-2.04	-2.38	-2.00	-2.58	-2.01	-2.81	-2.06
5	4	-2.29	-1.96	-2.35	-1.94	-3.00	-1.72	-3.03	-1.78
5	5	-2.29	-1.94	-2.39	-1.92	-2.99	-1.76	-3.02	-1.77
5	7	-2.29	-1.89	-2.30	-2.03	-2.86	**-1.70**	-3.35	**-1.76**
7	4	-2.22	-1.90	-2.20	-1.88	-2.78	-1.78	-3.73	-1.81
7	5	-2.20	-1.90	-2.19	-1.88	-3.08	-1.76	-4.07	-1.79
7	7	-2.21	-1.92	-2.18	-1.89	-3.55	-1.74	-4.15	-1.76
10	4	-2.20	-1.82	-2.24	-1.78	-4.25	-7.85	-4.80	-7.02
10	5	-2.21	**-1.82**	-2.21	-1.82	-4.13	-4.18	-4.03	-6.17
10	7	**-2.19**	-1.83	**-2.17**	**-1.82**	-3.08	-3.69	-Inf	-Inf

Table 6: Quantile Regression in combination with Principal Component Analysis

4.5 Quantile Regression Forest - predictor variable pre-selection

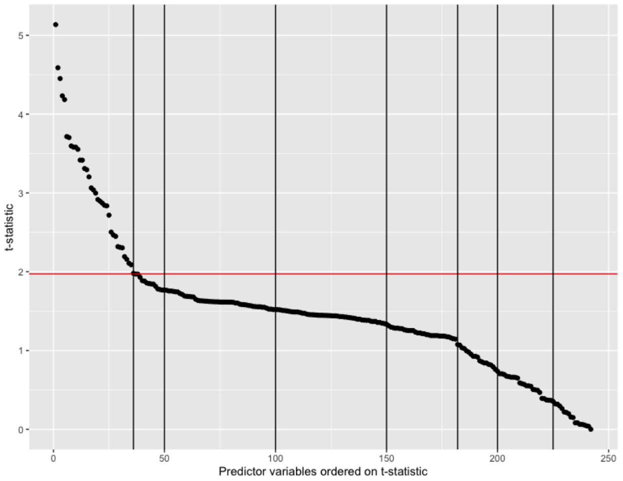

Figure 10: Ordered t-statistic's of predictor variables

Furthermore, Quantile Regression Forest is performed in combination with predictor variable pre-selection. Figure 10 shows the t-statistics of the ordered predictor variables when the marginal predictive power is determined using the full sample. The red horizontal line shows the 5% critical value of the corresponding t-distribution. A total of 36 predictor variables are significant. I choose to test including predictor variables in steps of 50, and only the significant predictor variables, and the right tail points 182 and 225. At the point 182, I see an increasing negative slope, indicating that the predictive power after this point decreases faster. I also choose the point 225 to see whether excluding the variables with the least predictive power increases performances. This is done for both in-sample and out-of-sample estimation. For out-of-sample estimation, the marginal predictive power is determined using only the in-sample data at each iteration recursively.

Thus, Quantile Regression Forest is performed with the 36, 50, 100, 150, 182, 200, and 225 most significant variables. All results are shown in Table 7. For in-sample estimation, including 182 variables performs the best for the full distribution. For the negative part of the distribution, including 100 or 150 variables performs best. When looking at out-of-sample performance, including all variables performs best. This difference in performance between in-sample and out-of-sample suggests that pre-selection of predictor variables tends to overfit to the in-sample data.

	In-sample				Out-of-sample			
	Skewed t		SGSH		Skewed t		SGSH	
Variables	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$
36	-2.06	-1.81	-2.05	-1.77	-2.15	-2.51	-2.16	-3.95
50	-2.05	-1.80	-2.04	-1.75	-2.16	-3.41	-2.22	-2.91
100	-2.06	-1.80	-2.05	**-1.71**	-2.14	-1.82	-2.18	-3.20
150	-2.06	**-1.78**	-2.05	-1.76	-2.14	-2.92	-2.15	-3.56
182	**-2.04**	-1.79	**-2.02**	-1.75	-2.14	-1.97	-2.20	-3.05
200	-2.06	-1.82	-2.05	-1.76	-2.13	-3.46	-2.19	-2.75
225	-2.05	-1.81	-2.05	-1.79	-2.13	**-1.80**	-2.14	-2.64
All	-2.06	-1.80	-2.05	-1.75	**-2.13**	-1.82	**-2.14**	**-1.94**

Table 7: Quantile Regression Forest in combination with various numbers of predictor variables

5 Conclusion

For out-of-sample forecasting, I found that the skewed t-distribution has a significantly better fit (at 10% confidence level) when the quantiles are estimated by Quantile Regression Forest. Making a pre-selection of the predictor variables does not improve the performance of Quantile Regression Forest. When comparing in-sample forecasting performance, I found that Quantile Regression Forest can again significantly improve the estimates. Especially when the hyperparameter `nodesize` is set to a low value. However, it should be taken into account that the lower `nodesize` is set, the more the model is (over)fit to the in-sample data.

When Quantile Regression in combination with variable selection is performed, I found that the in-sample fit can be significantly better than the fit provided by Quantile Regression with two predictors. However, the out-of-sample forecasting performance is significantly worse than Quantile Regression with two predictors. This suggests that Quantile Regression with variable selection overfits to the in-sample data.

Furthermore, I orthogonalized the predictor variables using Principal Component Analysis. This allowed me to include more predictor variables in Quantile Regression without multicollinearity concerns. I found that this technique also provides a good in-sample fit. The in-sample fit is even better than the fit provided by Quantile Regression with variable selection. However, like Quantile Regression with variable selection, the out-of-sample forecasting performance is worse than Quantile Regression with two predictors. Especially when more principal components are included.

For given estimated quantiles, the performance for the fit provided by the skewed t-distribution or SGSH distribution is, in general, not significantly different. Therefore, I conclude that using a skewed t-distribution or SGSH distribution does not make a significant difference.

I also looked at the difference in fit when a distribution is fit on more quantiles. Generally, I found that it makes no significant difference whether a distribution is fit on 4, 5, or 7 quantiles. This indicates that the exact identification of the distribution provides significant information.

In general, there was no significant difference in choice of models when the performance was evaluated only for negative GDP growth, or the full distribution. This indicates that no specific models are necessary when one is only interested in the negative risk.

The Appendix contains results of leave-one-out estimation. Most interestingly, the results are similar to the in-sample estimation results. This is probably due to the (auto)correlation between the value of GDP growth, and the future values of GDP growth and predictor variables.

I conclude that the two-step method of Adrian et al. (2019) can be improved by Quantile Regression Forest and a high-dimensional dataset. Quantile Regression Forest has better forecasting performance both in-sample and out-of-sample.

IMF's Growth at Risk framework already provides an implementation of Quantile Regression with more predictor variables. However, this framework does not provide any specific procedure to select the best predictor variables. By using Quantile Regression Forest no user-selection of predictor variables is necessary if the user has access to a high-dimensional macroeconomy dataset, like the FRED-QD database.

6 Discussion & Further research

The conclusion drawn in this paper comes with some known limitations which will be discussed here.

Even though the sample is an extended version of the sample used by Adrian et al. (2019), the sample contains a relatively small number of observations; 196. This is due to the low, quarterly, frequency that the dependent variable, GDP growth is available at. For further research, I would suggest to replace the dependent variable, GDP growth, by a related variable which is available in a higher frequency. One such variable is the Chicago Fed National Activity Index (CFNAI) which can be seen as a coincident indicator for economic activity. The CFNAI is available monthly. With more observations, Quantile Regression can be performed on a higher number of predictor variables which would probably increase the performance. Another interesting area to investigate when more observations are at hand, is the hyper tuning of `nodesize`. A higher value for `nodesize` could be used without making the trees in the forest consist of a too-small number of leaves.

Secondly, the method of Quantile Regression with Principal Component Analysis could be extended. Instead of using standard Principal Component Analysis, weighted Principal Component Analysis could be performed where the marginal predictive power of the predictor variables is used as weight. This method would give more weight to predictor variables with a higher marginal predictive power, which may improve the forecasting performance.

Thirdly, all forecasts in this research are 1-step ahead forecasts. It would be interesting to investigate whether the results differ when the forecast horizon is changed.

Finally, all methods discussed in this paper are tested on US GDP growth. To draw general conclusions on the performance of the discussed methods, the methods should be applied to different countries.

References

Adrian, T., Boyarchenko, N., & Giannone, D. (2019). Vulnerable growth. *American Economic Review*, *109*(4), 1263–89.

Azzalini, A. (2020). The R package `sn`: The skew-normal and related distributions such as the skew-*t* (version 1.6-1). [Computer software manual]. Università di Padova, Italia. Retrieved from `http://azzalini.stat.unipd.it/SN`

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(2), 367–389.

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.

Brave, S. A., & Kelly, D. L. (2017). Introducing the chicago fed's new adjusted national financial conditions index. *Chicago Fed Letter*, *386*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Corporation, M., & Weston, S. (2019). doparallel: Foreach parallel adaptor for the 'parallel' package [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=doParallel` (R package version 1.0.15)

Diks, C., Panchenko, V., & Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*(2), 215–230.

Fischer, M. (2004). Skew generalized secant hyperbolic distributions: Unconditional and conditional fit to asset returns. *Austrian Journal of Statistics*, *33*(3), 293–304.

Hooke, R., & Jeeves, T. A. (1961). "direct search"solution of numerical and statistical problems. *Journal of the ACM (JACM)*, *8*(2), 212–229.

Kent, L., & Phan, T. (2019). Time-varying skewness and real business cycles. *Economic Quarterly*(2Q), 59–103.

Koenker, R. (2020). quantreg: Quantile regression [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=quantreg` (R package version 5.55)

Leading economic index. (2020). The Conference Board. Retrieved from `https://conference-board.org/data/bcicountry.cfm?cid=1`

McCracken, M., & Ng, S. (2020). *Fred-qd: A quarterly database for macroeconomic research* (Tech. Rep.). St. Louis, United States of America: National Bureau of Economic Research.

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

Meinshausen, N. (2017). quantregforest: Quantile regression forests [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=quantregForest` (R package version 1.3-7)

Mitchell, J., & Hall, S. G. (2005). Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr 'fan'charts of inflation. *Oxford bulletin of economics and statistics*, *67*, 995–1033.

National financial conditions index (nfci). (2020, May). Federal Reserve Bank of Chicago. Retrieved from `https://www.chicagofed.org/publications/nfci/index`

Prasad, M. A., Elekdag, S., Jeasakul, M. P., Lafarguette, R., Alter, M. A., Feng, A. X., & Wang, C. (2019). *Growth at risk: Concept and application in imf country surveillance.* International Monetary Fund.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ringnér, M. (2008). What is principal component analysis? *Nature biotechnology*, *26*(3), 303–304.

Stock, J. H., & W Watson, M. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, *41*(3), 788–829.

Varadhan, R., University, J. H., Borchers, H. W., & Research., A. C. (2018). dfoptim: Derivative-free optimization [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=dfoptim` (R package version 2018.2-1)

7 Appendix

7.1 Implementation details

All code used in this paper, is written in R Core Team (2020). Below I further discuss implementation details. To reduce runtimes as much operations as possible are performed asynchronously through multi-threading. The R package `doParallel` offers an easy implementation (Corporation & Weston, 2019).

7.1.1 Quantile Regression

The R implementation of Quantile Regression is provided through the package `quantreg` by Koenker (2020).

7.1.2 Quantile Regression Forest

For the R implementation of Quantile Regression Forests the package `quantregForest` is used (Meinshausen, 2017).

7.1.3 Skewed *t*-distribution

The R package `sn` is used to calculate the pdf and inverse CDF of the skewed t-distribution (Azzalini, 2020).

The minimization problem corresponding to fitting the skewed t-distribution is solved using the bounded version of the Hooke-Jeeves derivative-free minimization algorithm (Hooke & Jeeves, 1961). An implementation of this algorithm is provided by the package `dfoptim` (Varadhan, University, Borchers, & Research., 2018).

The minimization is bounded by the following bounds: $-20 < \mu < -20$; $0 < \sigma < 50$; $-30 < \alpha < 30$; $1 < \nu < 30$. These bounds are introduced by the specifications of the skewed t-distribution and to reduce optimization run times. The following values are used as initial values in the optimization: $\mu = 0$; $\sigma = 1$; $\alpha = 0$; $\nu = 10$.

7.1.4 SGSH distribution

The minimization problem corresponding to fitting the SGSH distribution is also solved by the bounded version of the Hooke-Jeeves derivative-free minimization algorithm.

The following bounds are incorporated: $\sigma > 0$; $-\pi < t < 10$; $0 < s < 10$. The initial values used in the algorithm are: $\mu = 0$; $\sigma = 1$; $s = 1$; $t = 1$.

7.1.5 Principal Component Analysis

In R the function `prcomp` is used which is part of standard R `stats` package (R Core Team, 2020). `prcomp` calculates the principal components by singular value decomposition of the centered and scaled data matrix.

7.2 Leave-one-out estimation

Besides the in-sample and out-of-sample estimations as described in section 3, I performed leave-one-out estimations. For the results below, each method used to estimate quantiles is applied recursively wherein each iteration the full sample except the observation that is predicted is used.

Most interestingly, the results found here differ much from the results of out-of-sample expanding window estimation. Quantile Regression with Principal Component Analysis is found the best performing model. These differences are probably due to the (auto)correlation between the value of GDP growth, and the future values of GDP growth and predictor variables.

	Skewed t		SGSH	
# quantiles	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$
4	-2,33	-2,01	-2,34	-1,98
5	-2,33	-2,00	-2,35	-1,96
7	-2,33	-1,97	-2,32	-1,95

Table 8: Quantile Regression with lagged GDP and NFCI

	Skewed t		SGSH	
# quantiles	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$
4	-2,30	-2,15	-2,35	-2,23
5	-2,29	-2,08	-2,30	-2,10
7	-2,31	-2,10	-2,33	-2,14

Table 9: Quantile Regression Forest with all predictors, `nodesize` $= 35$

		Skewed t		SGSH	
# predictors	# quantiles	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$
5	4	-2,28	-2,02	-2,30	-1,95
5	5	-2,28	-1,98	-2,27	-1,98
5	7	-2,26	-1,96	-2,28	-1,94
7	4	-2,18	-1,96	-2,16	-1,93
7	5	-2,18	-1,94	-2,17	-2,09
7	7	-2,18	-1,98	-2,16	-1,94

Table 10: Quantile Regression with five predictors

		Skewed t		SGSH	
# components	# quantiles	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$	S^{CL}	$S^{\mathrm{cl}}_{\mathbb{R}-}$
3	4	-2,39	-2,14	-2,38	-2,10
3	5	-2,39	-2,14	-2,38	-2,10
3	7	-2,39	-2,13	-2,38	-2,08
5	4	-2,14	-2,38	-2,10	-2,01
5	5	-2,14	-2,38	-2,10	-2,01
5	7	-2,13	-2,38	-2,08	-2,12
7	4	-2,04	-2,35	-2,01	-2,34
7	5	-2,02	-2,33	-2,01	-1,97
7	7	-1,97	-2,30	-2,12	-1,98
10	4	-1,98	-2,26	-2,34	-1,88
10	5	-1,98	-2,21	-1,97	-1,90
10	7	-2,00	-2,18	-1,98	-2,01

Table 11: Quantile Regression in combination with Principal Component Analysis