Bayesian Quantile Regression Treed

Mauro Bernardi and Paola Stolfi

Abstract Decision trees and their population counterparts are becoming promising alternatives to classical linear regression techniques because of their superior ability to adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their popularity, those methods have been developed for classification and regression, while often the conditional mean would not be enough when data strongly deviates from the Gaussian assumption. The approach proposed in this paper instead considers an ensemble of nonparametric regression trees to model the conditional quantile at level $\tau \in (0,1)$ of the response variable. Specifically, a flexible generalised additive model (GAM) is fitted to each partition of the data that corresponds to a given leaf of the tree, allowing an easy interpretation of the model parameters. Indeed, while the trees structure easily adapts to regions of the data having different shapes and variability, the nonlinear part handles parsimoniously the local nonlinear structural relationship of the quantile with the covariates. Unlike the most popular Bayesian approach (BART) that assumes a sum of regression trees, quantile estimates are obtained by averaging the ensemble trees, thereby reducing their variance. We develop a Bayesian procedure for fitting such models that effectively explores the space of B–Spline functions of different orders that features the functional nonlinear relationship with the covariates. The approach is particularly valuable when skewness, fat–tails, outliers, truncated and censored data, and heteroskedasticity, can shadow the nature of the dependence between the variable of interest and the covariates. We apply our model to a sample of US companies belonging to different sectors of the Standard and Poor's Composite Index and we provide an evaluation of the marginal contribution to the overall risk of each individual institution.

Key words: Quantile regression treed, Bayesian inference, Conditional Value–at–Risk.

1 Introduction

In empirical studies, researchers are often interested in analysing the behaviour of a response variable given the information on a set of covariates. The typical answer is to specify a linear regression model where unknown parameters are estimated by OLS, thereby leading to the approximation of the mean function. Although the mean describes the average response path as a function of the covariates, it provides little o no information about the behaviour of the conditional distribution on the tails. As far as the entire distribution is concerned, quantile regression methods adequately characterise the behaviour of the response variable at different confidence levels providing a complete picture of the relationship with the covariates. Moreover, the quantile analysis is particularly suitable when the conditional distribution strongly deviates from

Mauro Bernardi
Department of Statistical Sciences, University of Padova, Via Cesare Battisti, 241, 35121, Padova, e-mail: mauro.bernardi@unipd.it

Paola Stolfi
Institute for applied mathematics "Mauro Picone" (IAC) - CNR, Rome, Italy, e-mail: p.stolfi@iac.cnr.it

the Gaussian assumption because it displays heterogeneity, asymmetry or fat–tails, see, e.g., [9]. Linear quantile regression models have been extensively applied in different areas, such as, finance, engineering, econometrics and environmetrics, as a direct approach to quantify the level of risk of a given event, social sciences and quantitative marketing to find appropriate and effective solutions to specific segments of customers, and many other related fields see, [10]. However, despite their relevance and widespread application in empirical studies, linear quantile regression models provide only a rough "first order" approximation of the relationship between the τ–level quantile of the response variable and the covariates. Indeed, as first recognised by [9], quantiles are linear functions only within a Gaussian world, thereby stimulating many recent attempts to overcome this limitation. [6], for example, consider the copula–based approach to formalise nonlinear and parametric conditional quantile relationships. Although quite flexible in fitting marginal data, the copula approach forgets to consider nonlinear interactions among the covariates. Classification and regression trees (CART, [4]) and their population counterparts ([3]) extensively use recursive partitioning algorithms to perform nonparametric regression and variable selection. The attractive feature of decision trees methods rely in their ability to partition the covariates space into disjoint hyperrectangles, thereby improving the local fit. Therefore, CART adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their extensive use in a wide variety of fields, those methods have been mainly developed for classification and mean regression. In this paper, we adopt the Bayesian point of view and we extend the Bayesian regression trees approach of [7] to deal with conditional quantiles estimation. Quantile estimation have been previously extended within the related context of random forest by [11]. However, unlike random forests, the Bayesian approach to decision trees learning, being likelihood–based, provides a complete inferential tool for model assessment and selection.

2 Quantile regression treed

The linear quantile regression framework for independent and identically distributed data models the conditional τ–level quantile of the response variable Y, with $\tau \in (0,1)$, as a linear function of the vector of dimension $(q \times 1)$ of exogenous covariates $\mathbf{X}$, i.e., $\mathscr{Q}_\tau(Y \mid \mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, thereby avoiding any explicit assumptions about the conditional distribution of $Y \mid \mathbf{X} = \mathbf{x}$. This is equivalent to assume an additive stochastic error term ε for the conditional regression function $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ to be independent and identically distributed with zero τ–th quantile, i.e, $\mathscr{Q}_\tau(\varepsilon \mid \mathbf{x}) = 0$, and constant variance. Following [12] and [2], the previous condition is implicitly satisfied by assuming that the conditional distribution of the response variable Y follows an Asymmetric Laplace (AL) distribution located at the true regression function $\mu(\mathbf{x})$, with constant scale $\sigma > 0$ and shape parameter τ, i.e., $\varepsilon \sim \mathsf{AL}(\tau, \mu(\mathbf{x}), \sigma)$, with probability density function

$$\mathsf{AL}(Y \mid \mathbf{X} = \mathbf{x}, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\frac{1}{\sigma} \rho_\tau(Y - \mu_\tau(\mathbf{x})) \right\} \mathbb{1}_{(-\infty,\infty)}(Y), \tag{1}$$

where $\mu_\tau(\mathbf{x})$ is the quantile regression function and $\rho_\tau(u) = u\left(\tau - \mathbb{1}_{(-\infty,0)}(u)\right)$ denotes the quantile check function at level τ. The quantile regression model postulated in equation (1) assumes the AL distribution as a misspecified working likelihood that correctly identify the conditional quantile function.

Unlike the Bayesian Additive Regression Tree (BART) approach of [8] which considers a sum–of–trees regression where each tree explains only a small portion of the total variance of the dependent variable Y, we model the conditional quantile of the response variable as a function of the covariates as the average of an ensemble of $m \in \mathbb{N}_+$ regression treeds

$$\mathscr{Q}(Y \mid \mathbf{X} = \mathbf{x}) = \mu_\tau(\mathbf{x}) \tag{2}$$

$$\approx \frac{1}{m} \sum_{l=1}^{m} \mathscr{T}_l^{\mathscr{M}_l}(\mathbf{x}), \tag{3}$$

where $\mathscr{T}_j^{\mathscr{M}_j}$ denotes the j–th treed of the ensemble, for $j = 1, 2, \ldots, m$. In equation (3) each regression trees is composed by a tree structure, denoted by $\mathscr{T}$, and the parameters of the terminal nodes (also called

leaves), denoted by $\mathscr{M}$. Therefore, the j–th tree for $j = 1,2,\ldots,m$, denoted by $\mathscr{T}_j^{\mathscr{M}}$, represents a specific combination of tree structure $\mathscr{T}_j$ and tree parameters $\mathscr{M}_j$, i.e., the regression parameters associated to its terminal nodes.

3 Conditional Value–at–Risk estimation

In this section we apply the methodology introduced in the previously section to analyse the tail co–movements between a financial institution j and the whole financial system k. To this aim we consider the Conditional Value–at–Risk (CoVaR) recently introduced by [1], which is defined as the overall VaR of an institution, conditional on another institution being in distress. To be more specific, let (Y_j, Y_k) be the bivariate random variable describing the return of institutions j and k, for $k \neq j$ and assume (Y_j, Y_k) depend on a vector of exogenous covariates $\mathbf{X} = (X_1, X_2, \ldots, X_q)$, then the Conditional Value–at–Risk $\left(\text{CoVaR}_{k|j}^{\mathbf{x},\tau}\right)$ is the Value–at–Risk of institution k conditional on $Y_j = \text{VaR}_j^{\mathbf{x},\tau}$ at the level $\tau \in (0,1)$, i.e., $\text{CoVaR}_{k|j}^{\mathbf{x},\tau}$ satisfies the following equation

$$\mathbb{P}\left(Y_k \leq \text{CoVaR}_{k|j}^{\mathbf{x},\tau} \mid \mathbf{X} = \mathbf{x}, Y_j = \text{VaR}_j^{\mathbf{x},\tau}\right) = \tau, \tag{1}$$

where $\text{VaR}_j^{\mathbf{x},\tau}$ denotes the Value–at–Risk, $\text{VaR}_j^{\mathbf{x},\tau}$ of institution j, i.e., the τ–th level conditional quantile of the random variable $Y_j \mid \mathbf{X} = \mathbf{x}$, defined as

$$\mathbb{P}\left(Y_j \leq \text{VaR}_j^{\mathbf{x},\tau} \mid \mathbf{X} = \mathbf{x}\right) = \tau. \tag{2}$$

Note that both the VaR and the CoVaR corresponds to the τ–th quantiles of the conditional distribution of $Y_j \mid \{\mathbf{X} = \mathbf{x}\}$ and $Y_k \mid \{\mathbf{X} = \mathbf{x}, Y_j = \text{VaR}_j^{\mathbf{x},\tau}\}$, respectively. Therefore, both the VaR and CoVaR equations can be estimated using the Bayesian quantile regression treed models introduced in the previous section.

The financial data we utilise are taken from the Standard and Poor's Composite Index (k) for the U.S market, where different sectors (j) are included. For both the institutions and for the whole system, we consider microeconomics and macroeconomics variables, in order to take into account for individual information and for global economic conditions respectively. Our empirical analysis is based on publicly traded U.S. companies belonging to different sectors of the Standard and Poor's Composite Index (S&P500) listed in Table 1. The sectors considered are: Financials, Consumer Goods, Energy, Industrials, Technologies and Utilities. Financials consists of banks, diversified financial services and consumer financial services. Daily equity price data are converted to weekly log–returns (in percentage points) for the sample period from January 2, 2004 to December 28, 2012, covering the recent global financial crisis. To control for the general economic conditions we use observations of the following macroeconomic regressors as suggested by [1] and [5]: the VIX index (VIX), measuring the model-free implied stock market volatility as evaluated by the Chicago Board Options Exchange (CBOE), a short term liquidity spread (LIQSPR), computed as the difference between the 3-month collateral repo rate and the 3-months Treasury Bill rate, the weekly change in the three-month Treasury Bill rate (3MTB) the change in the slope of the yield curve (TERMSPR), measured by the difference of the 10-years Treasury rate and the 3-month Treasury Bill rate, the change in the credit spread (CREDSPR) between 10-years BAA rated bonds and the 10-years Treasury rate and the weekly return of the Dow Jones US Real Estate Index (DJUSRE). To capture the individual firms' characteristics, we include observations from the following microeconomic regressors: leverage (LEV), calculated as the value of total assets divided by total equity (both measured in book values), the market to book value (MK2BK), defined as the ratio of the market value to the book value of total equity, the size (SIZE), defined by the logarithmic transformation of the market value of total assets, and the maturity mismatch (MM), calculated as short term debt net of cash divided by the total liabilities. To have a complete picture of the contributes from individual and systemic risk we plot the estimated VaR and CoVaR for some of the assets listed in Table 1 in Figure 1. Looking at individual risk assessment, it is clear that the VaR profiles are relatively similar across institutions, displaying strong negative downside effects upon the occurrence of the recent financial crises of 2008 and 2010 and the sovereign debt crisis of 2012. However, the analysis of the time

Name	Ticker Symbol	Sector
CITIGROUP INC.	C	Financial
BANK OF AMERICA CORP.	BAC	Financial
COMERICA INC.	CMA	Financial
JPMORGAN CHASE & CO.	JPM	Financial
KEYCORP	KEY	Financial
GOLDMAN SACHS GROUP INC.	GS	Financial
MORGAN STANLEY	MS	Financial
MOODY'S CORP.	MCO	Financial
AMERICAN EXPRESS CO.	AXP	Financial
MCDONALD'S CORP.	MCD	Consumer
NIKE INC.	NKE	Consumer
CHEVRON CORP.	CVX	Energy
EXXON MOBIL CORP.	XOM	Energy
BOEING CO.	BA	Industrial
GENERAL ELECTRIC CO.	GE	Industrial
INTEL CORP.	INTC	Technology
ORACLE CORP.	ORCL	Technology
AMEREN CORPORATION.	AEE	Utilities
PUBLIC SERVICE ENTERPRISE INC.	PEG	Utilities

TABLE 1: List of companies included in empirical analysis. All listed companies belonged to the Standard and Poor's Composite Index (S&P500) at the start of the trading day of February 15, 2013.

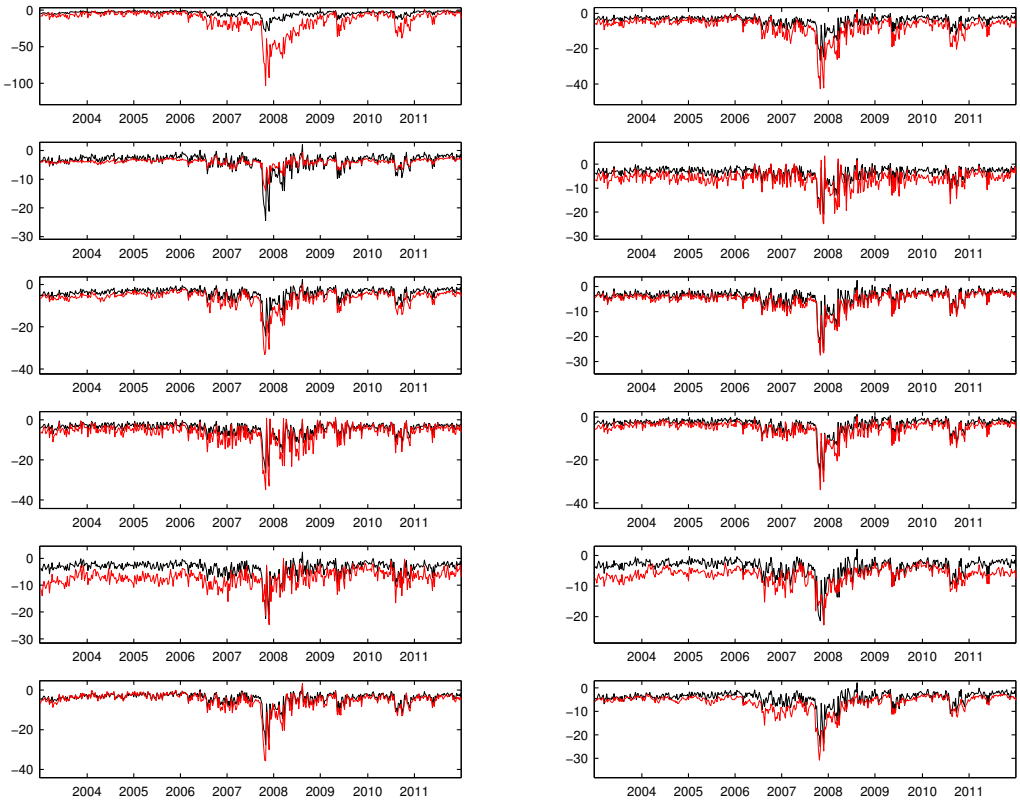

FIG. 1: Time series plot of the $\text{VaR}_j^{\mathbf{x},\tau}$ (red line) and $\text{CoVaR}_{k|j}^{\mathbf{x},\tau}$ (gray line) at the confidence level $\tau = 0.025$, for the following assets: top panel (financial): C (left), GS (right); second panel (consumer): MCD (left) and NKE (right); third panel (energy): CVX (left), XOM (right); fourth panel (industrial): BA (left), GE (right); fifth panel (technology): INTC (left), ORCL (right); bottom panel (utilities): AEE (left), PEG (right).

series evolution of the marginal contribution to the systemic risk, measured by CoVaR, reveals different behaviors for the considered assets. In particular, Citygroup (C), which belongs to the Financials, seems to contribute more to the overall risk than other assets do.

References

1. ADRIAN, T. AND BRUNNERMEIER, M., (2016). CoVaR. *American Economic Review*, 31, 106, 1705-1741.
2. Bernardi, M., Gayraud, G., and Petrella, L. (2015). Bayesian tail risk interdependence using quantile regression. *Bayesian Anal.*, 10(3):553–603.
3. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
4. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.
5. CHAO, S.-K., HÄRDLE, W.F. AND CHANG, W. (2012). Quantile Regression in Risk Calibration. *Handbook for Financial Econometrics and Statistics* in Cheng-Few Lee, ed., Springer Verlag.
6. Chen, X., Koenker, R., and Xiao, Z. (2009). Copula-based nonlinear quantile autoregression. *Econometrics Journal*, 12:S50–S67.
7. Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–948.
8. Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298.
9. KOENKER, P., (2005). *Quantile Regression*. Cambridge University Press, Cambridge.
10. Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). *Handbook of Quantile Regression*. CRC Press.
11. Meinshausen, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999.
12. Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.