Mathematical Biosciences and Engineering

MBE, 18(1): 214–230.
DOI: 10.3934/mbe.2021011
Received: 09 September 2020
Accepted: 17 November 2020
Published: 26 November 2020

http://www.aimspress.com/journal/MBE

Research article

A distributed quantile estimation algorithm of heavy-tailed distribution with massive datasets

Xiaoyue Xie[1,2,*] and Jian Shi[1,2]

[1] Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China
[2] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

* **Correspondence:** Email: xiexiaoyue16@mails.ucas.edu.cn; Tel: +8615611535356; Fax: +861082541972.

Abstract: Quantile estimation with big data is still a challenging problem in statistics. In this paper we introduce a distributed algorithm for estimating high quantiles of heavy-tailed distributions with massive datasets. The key idea of the algorithm is to apply the alternating direction method of multipliers in parameter estimation of the generalized pareto distribution in a distributed structure and compute high quantiles based on parameter estimation by the Peak Over Threshold method. This paper proves that the proposed algorithm converges to a stationary solution when the step size is properly chosen. The numerical study and real data analysis also shows that the algorithm is feasible and efficient for estimating high quantiles of heavy-tailed distribution with massive datasets and there is a clear-cut winner for the extreme quantiles.

Keywords: distributed algorithm; big data; high quantile estimation; heavy-tailed distribution; Peak Over Threshold method

Abbreviations: POT: Peak Over Threshold; GPD: Generalized pareto distribution; EDF: Empirical distribution function; WNLS: Nonlinear weighted least squares method; ADMM: The alternating direction method of multipliers; ARB: The absolute relative bias; RMSE: The root of mean square errors

1. Introduction

Changes in the magnitude and frequency of extremes of nature phenomena, such as extreme

precipitation, streamflow, floods, temperature, or wind speed, have a serious negative effect on human society (loss of life, damage to buildings and infrastucture) and environment. Knowledge of high quantiles of the probability distribution is of great importance for planning and design engineering. For example, high quantile of the maximal water level provides a risk measure for dike design in water resources structures [1], extreme wind patterns associated with speed and direction provide some guidance in structural projects related to onshore and offshore activities, wind farms, and oil and gas exploitation [2]. In addition, the generalized Pareto (GPD) and generalized extreme-value (GEV) distribution have been used in the modeling natural extreme events in hydrology, climatology, meteorology, and other areas [3–5]. So it is very necessary to provide a good estimation of high quantiles for the heavy-tailed distributions.

Many of the statisticians who have been interested in this problem have used extreme value theory as their major tool. Among the several methods considered, the Peak Over Threshold method with generalized Pareto distribution has been quite successful and very widely used. To estimate high quantiles, the estimators of parameters of the generalized Pareto distribution are required. For this parameter estimation problem, many methods have been proposed in the literature, see [6–12]. The Peak Over Threshold method has been commonly used to estimate extreme quantiles when the datasets are not massive. Recently, as datasets are becoming increasingly large, some researchers have proposed some new methods for applying the Peak Over Threshold method to high quantile estimation with massive data. Using a nonlinear least square, Song and Song [13] proposed a new parameter estimation method with generalized Pareto distribution for massive datasets that minimizes the sum of squared deviations between the empirical distribution function and the theoretical generalized Pareto distribution, and used this parameter method to estimate high quantiles by the conventional Peak Over Threshold. Sequently, Park and Kim [14] pointed out a drawback of Song's method that their actual formation was only applicable when the underlying model's tail fitted the generalized Pareto distribution unconditionally and it was also not fair in comparison with the other estimation methods as the alternative method was under the Peak Over Threshold framework whereas the Song's nonlinear least square method was not. To this end, they revised the nonlinear least square method and proposed a new version the nonlinear weighted least squares method. The performance of this modified method was highly competitive in estimating high quantiles for heavy-tailed distributions. Meanwhile, Kang and Song [15] also pointed out that this method outperformed other methods in estimating more extreme values such as the 99.9th and 99.99th quantiles with a small number of observations. However, the above-mentioned methods are conducted with full sample in a single machine. At present, the exceedingly large size of data often makes it impossible to store all of them on a single machine and many applications have individual agents collecting data independently. Communication between agents is expensive due to the limited bandwidth, and direct data sharing has also raised privacy and lost of ownership concerns. These constraints often make it prohibitive to direct application of the existing methods. Therefore, this paper attempts to develop a distributed algorithm for estimating high quantiles of heavy-tailed distributions.

An approach for the distributed and parallel computation is based on the alternating direction method of multipliers (ADMM), which is proposed by [16]. And this algorithm has been extensively used in various areas such as model fitting [17], sparse inverse covariance selection [18], constrainted sparse regression [19]. In the distributed alternating direction method of multipliers, the original problem is partitioned into N subproblems, and each subproblem contains a subset of

samples or learning parameters. At each iteration, the workers solve the subproblems and send the up-to-date variable information to the master, who summarizes this information and broadcasts the result to the workers. In addition, convergence properties of the distributed ADMM have been extensively studied, see Ref [16,20–22]. Further, the synchronous distributed method has also been extended to the asynchronous setting to speed up the computation, see Ref [23,24]. This paper proposes a distributed algorithm for high quantile estimation based on the synchronous distributed method because each worker uses same estimation method to handle the samples of the same size so that each worker's computation time is almost the same. Specially, we adopt the nonlinear weighted least squares method to estimate the local parameter of generalized Pareto distribution in each processor, conduct the distributed alternating direction method of multipliers to compute the population parameter, and then use the Peak Over Threshold method to estimate high quantiles in a distributed structure.

The remainder of this paper is organized as follows. Section 2 introduces the Peak Over Threshold method for high quantile estimation, the nonlinear weighted least squares method for parameter estimation and the distributed alternating direction method of multipliers for consensus problem. Section 3 proposes a distributed algorithm based on alternating direction method of multipliers and the nonlinear weighted least squares method to estimate high quantiles under the Peak Over Threshold framework and discusses the algorithm convergence. Section 4 conducts simulation to discuss the estimation performance of the proposed distributed method compared with the WNLS method based on full sample and the mean WNLS method. Section 5 applies our method to a real example. Section 6 concludes the paper.

2. Preliminaries

2.1. High quantile estimation —POT method

The distribution function of GPD is defined as

$$G_{\xi,\sigma}(x)=\begin{cases}1-(1+\xi x/\sigma)^{-1/\xi}, & if\ \xi\neq 0,\\ 1-\exp(-x/\sigma), & if\ \xi=0,\end{cases} \tag{2.1}$$

where ξ and σ are the shape and scale parameters, respectively. $x>0$ if $\xi>0$ and $0<x<\sigma/\xi$ if $\xi<0$. This GPD family can be also extended to define a GPD with a location parameter μ, as $G_{\xi,\mu,\sigma}(x)=G_{\xi,\sigma}(x-\mu)$. In this paper, we consider the heavy-tailed distributions only, that is, the shape parameter $\xi>0$. See [23] for details on GPD.

Pickands [25] and Balkema and de Haan [26] showed that the distribution of excesses can be approximated by the GPD if the distribution is in the maximum domain of attraction. As seen in Ref [27], most of the common continuous distributions are in the maximum domain of attraction. We now introduce the POT method to estimate high quantiles of the unknown continuous distribution. Specifically, we let $F(x)$ be the distribution function of an arbitrary continuous distribution, and define the exceedance of loss events over u by

$$F_u(y)=P(X-u\le y\mid X>u)=\frac{F(u+y)-F(u)}{1-F(u)}, \tag{2.2}$$

which can be fitted by GPD (ξ, σ). Thus we can rewrite $F(x)$ as

$$\begin{aligned} F(x) &= P(X \le x) = (1-F(u))F_u(x-u) + F(u) \\ &= (1-F(u))G_{\xi,\sigma}(x-u) + F(u) \\ &= (1-F(u))G_{\xi,u,\sigma}(x) + F(u), \end{aligned} \tag{2.3}$$

using the three-parameter GPD, and estimate it by

$$\hat{F}(x) = (1-F_n(u))G_{\hat{\xi},u,\hat{\sigma}}(x) + F_n(u), \tag{2.4}$$

where $F_n(x)$ is the EDF and $G_{\hat{\xi},u,\hat{\sigma}}(x)$ is the theoretical distribution function of GPD fitted with observations over u. Compute the *p*-th quantile by inverting Eq (2.4) as

$$x_p = G^{-1}_{\hat{\xi},u,\hat{\sigma}}(1-\frac{1-p}{1-F_n(u)}) = u + \frac{\hat{\sigma}}{\hat{\xi}}[(\frac{1-p}{1-F_n(u)})^{-\hat{\xi}} - 1]. \tag{2.5}$$

Therefore, for high quantile estimation of heavy-tailed distributions, we only focus on the GPD parameter estimation. For more theoretical details on POT, see [26].

2.2. GPD parameter estimation method —WNLS

Here, we review the WNLS method proposed by Park and Kim [14] for estimating GPD parameters. Suppose that we have a sample $x_1, \cdots, x_n$ of size n, and $n_u < n$ observations that are greater than the selected GPD threshold u. Without loss of generality, it is assumed that $x_1 > x_2 > \cdots > x_n$. The WNLS method is divided into two steps. The first step finds the interim estimate $(\hat{\xi}_1, \hat{\sigma}_1)$ using a nonlinear minimization:

$$\begin{aligned} (\hat{\xi}_1, \hat{\sigma}_1) &= \arg\min \sum_{i=1}^{n_u} [\log(1-F_n(x_i)) - \log(1-F(x_i))]^2 \\ &= \arg\min \sum_{i=1}^{n_u} [\log\frac{1-F_n(x_i)}{1-F_n(u)} - \log(1-G_{\xi,u,\sigma}(x_i))]^2. \end{aligned} \tag{2.6}$$

In the second step, because the distribution of $F(X_i)$ is that of $U_{n-i+1;n}$ that the $(n-i+1)$th order statistic of the uniform random variable, of which the distribution is known to be Beta(n-i+1,i) from the standard distribution theory. $[Var(U_{n-i+1;n})]^{-1}$ was used as the weight for each square deviance term, which yields the weight nonlinear minimization. The weighted nonlinear optimization was given as follow:

$$\begin{aligned} (\hat{\xi}_2, \hat{\sigma}_2) &= \arg\min \sum_{i=1}^{n_u} [Var(U_{n-i+1;n})]^{-1} [F_n(x_i) - F(x_i)]^2 \\ &= \arg\min \sum_{i=1}^{n_u} [\frac{i(n-i+1)}{(n+1)^2(n+2)}]^{-1} [\frac{1-F_n(x_i)}{1-F_n(u)} - G_{\xi,u,\sigma}(x_i)]^2. \end{aligned} \tag{2.7}$$

Combined with the first step, the estimated parameters $(\hat{\xi}_2, \hat{\sigma}_2)$ called the WNLS estimator under the POT framework.

2.3. Distributed ADMM for consensus optimization

Consider minimization of a function $g(\theta)$ in a distributed computing environment. Assume

that this function can be decomposed into K components as

$$g(\theta) = \sum_{k=1}^{K} g_k(\theta), \tag{2.8}$$

where $\theta \in R^n$ and each g_k is a local objective on node k. It is useful to solve this problem either when there are many samples that it is inconvenient or impossible to process them on a single machine or when the data is naturally collected or stored in a distributed fashion. This problem is common in various areas such as machine learning, signal processing and wireless communication [28]. For example, in regularized risk minimization, θ is the model parameter to be estimated, and g_k is the regularized risk functional defined on node k.

This problem can be reformulated as the following global variable consensus optimization problem.

$$\begin{aligned} &\min \sum_{k=1}^{K} g_k(\eta_k) \\ &\text{subject to } \eta_k - \theta = 0, \text{ for } k = 1, 2, \cdots, K, \end{aligned} \tag{2.9}$$

where θ is called the consensus variable, and $\eta_k \in R^n$ is node k's local copy of the parameter. In a distributed computing environment, this problem can be efficiently solved by the ADMM algorithm, and the iteration procedure is

$$\begin{aligned} \eta_k^{t+1} &:= \arg\min_x g_k(\eta) + < y_k^t, \eta > + \frac{\rho}{2} \| \eta - \theta^t \|_2^2, \\ \theta^{t+1} &:= \frac{1}{K} \sum_{k=1}^{K} \eta_k^{t+1}, \\ y_k^{t+1} &:= y_k^t + \rho(\eta_k^{t+1} - \theta^{t+1}), \end{aligned} \tag{2.10}$$

where y_k is the Lagrange dual variable. This distributed ADMM algorithm can be easily implemented in a distributed system with one master and K workers. The master is responsible for updating the consensus variable θ and each worker minimizes its local objective g_k (in parallel) based on its own data subset, and sends the updated local copy η_k to the master. The master, in turn, updates θ by driving the η_k into consensus, and then distributes the updated value back to the workers, and the process reiterates. Therefore, the distributed algorithm can be described as a computation network with a star topology as shown in Figure 1, in which a master node coordinates the computation of a set of distributed workers.

In addition, the primal and dual residuals of this problem are

$$r^t = (\eta_1^t - \theta^t, \ldots, \eta_K^t - \theta^t), \qquad s^t = \rho(\theta^t - \theta^{t-1}, \ldots, \theta^t - \theta^{t-1}), \tag{2.11}$$

and their squared norms are

Figure 1. A star computer cluster with master and workers.

$$\| r^t \|_2^2 = \sum_{k=1}^{K} \| \eta_k^t - \theta^t \|_2^2, \qquad \| s^t \|_2^2 = K\rho^2 \| \eta^t - \theta^{t-1} \|_2^2, \tag{2.12}$$

then a reasonable termination criterion of this algorithm is that the primal and dual residuals must be small

$$\| r^t \|_2 \le \varepsilon^{pri}, \qquad \| s^t \|_2 \le \varepsilon^{dual}, \tag{2.13}$$

where $\varepsilon^{pri} > 0$ and $\varepsilon^{dual} > 0$ are feasibility tolerances for the primal and dual feasibility conditions, respectively.

3. A distributed quantile estimation algorithm for heavy-tailed distributions

In this section, we propose a distributed algorithm to estimate high quantiles for heavy-tailed distributions. This algorithm combines the WNLS method and the distributed ADMM to estimate GPD parameters, and then use parameter estimators to compute high quantiles by the conventional POT. Next, we will introduce the algorithm procedures and its some properties in detail.

3.1. Algorithm deveploment

Supposed that we have N observations $x_1, \ldots, x_N$ from a heavy-tailed distribution F and these observation are stored on K nodes and each node has m observations. These observations can be written as $m \times K$ matrix, *i.e.*

$$X = (X_1, X_2, \ldots, X_K) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mK} \end{pmatrix}.$$

For each node observations, we choose the q-quantile as a threshold value and pick out observations larger than the threshold value. More precisely, let u_k be each node threshold value

for $k=1, 2, \ldots, K,$ then the observations selected from all node can be written as

$$Z=(Z_1, Z_2, \ldots, Z_K)=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1K} \\ z_{21} & z_{22} & \cdots & z_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nK} \end{pmatrix},$$

where $n=m-[mq],$ i.e., there are n observations which are greater than u_k for each node. It is assumed that $u_k < z_{nk} < \cdots < z_{1k}$ without lost of generality for any $1 \le k \le K,$ For fixed k and any $1 \le i \le n,$ let w_{ik} be the variance of $F(z_{ik})$, which is the distribution of the $(m-i+1)$th order statistic of the uniform random variable. $(F_m(z_{ik})-F(z_{ik}))^2$ is the squared deviation between the empirical distribution function and the theoretical distribution function. The optimization objective of GPD parameter estimation is

$$\min\ g(\theta)=\sum_{k=1}^{K} g_k(\theta), \tag{3.1}$$

where $\theta=(\xi,\sigma)$ and $g_k(\theta)=\sum_{i=1}^{n}\frac{i(m-i+1)}{(m+1)^2(m+2)}[\frac{F_m(z_{ik})-F_m(u_k)}{1-F_m(u_k)}-G(z_{ik}-u_k \mid \theta)]^2$ is the local objective function of the *k*-th node.

The minimization of $g(\theta)$ can be reformulated as the following global variable consensus optimization problem

$$\begin{aligned} &\min\ \sum_{k=1}^{K} g_k(\eta_k) \\ &\text{subject to } \eta_k=\theta, \ \text{ for } k=1, 2, \cdots, K, \end{aligned} \tag{3.2}$$

where θ is the so-called consensus parameter, and η_k is the *k*-th local copy of the parameter. And the augmented Lagrangian is given by

$$L(\{\eta_k\},\theta)=\sum_{k=1}^{K} g_k(\eta_k)+\sum_{k=1}^{K}\langle y_k, \eta_k-\theta\rangle+\sum_{k=1}^{K}\frac{\rho_k}{2}\|\eta_k-\theta\|_2^2, \tag{3.3}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product.

Algorithm 1 Distributed ADMM for Eq (3.2).

(1): Given initial variable θ^0 and y^0, set $t=0.$

(2): repeat

(3): update

$$\eta_k^{t+1} := \arg\min_{\eta}\ g_k(\eta)+<y_k^t, \eta>+\frac{\rho}{2}\|\eta_k-\theta^t\|_2^2, \tag{3.4}$$

$$\theta^{t+1} := \frac{1}{K}\sum_{k=1}^{K}\eta_k^{t+1}, \tag{3.5}$$

$$y_k^{t+1} := y_k^t+\rho(\eta_k^{t+1}-\theta^{t+1}). \tag{3.6}$$

(4): set $t \leftarrow t+1.$

(5): until a predefined stopping criterion is satisfied.

In a distributed computing environment, this problem can be efficiently solved by the ADMM algorithm. The ADMM algorithm for solving Eq (3.2) is presented in Algorithm 1. This Algorithm can be easily implemented in a distributed system with one master and K workers. Each worker is responsible for updating its using Eqs (3.4) and (3.6), and the updated η_k is then sent to the master, which is responsible for updating the consensus parameter θ and distributing its updated value back to the workers. Note that, as the (η_k, y_k) is local to each worker, their updates can be performed by all the workers in parallel. In addition, at the beginning of the algorithm, each worker can solve the optimization problem Eq (3.7) as parameter initial value for each data subset.

$$\eta_k^0 = \arg\min_{\eta} \sum_{i=1}^{n} [\log \frac{1-F_m(z_{ik})}{1-F_m(u_k)} - \log(1-G(z_{ik}-u_k \mid \eta))]^2,\ k=1,\ 2,\ \ldots,\ K, \tag{3.7}$$

Therefore, let $\hat{\theta} = (\hat{\xi}, \hat{\sigma})$ be the GPD parameter estimates with the above distributed algorithm, then a distributed quantile estimation under the POT framework is

$$x_p = u + \frac{\hat{\sigma}}{\hat{\xi}}[(\frac{1-p}{1-q})^{-\hat{\xi}} - 1], \tag{3.8}$$

where $u = \frac{1}{K}\sum_{k=1}^{K} u_k$.

Note: This algorithm aims to compute the quantiles of heavy-tailed distributions in a distributed storage architecture. The number of nodes K is generally fixed in this scenario. Of course, it is also feasible to deal with exceedingly large size of data on a single machine. But it should be noted that in this scenario higher accuracy generally requires more observations when estimating the tail index of the distribution (such as high quantiles). Therefore, we suggest the number of blocks should be as small as possible when each block sample size is manageable when our algorithm is applied to this scenario. We only focus on the distributed situation in this paper.

3.2. Algorithm convergence and termination criterion.

Since the objective $g(\theta)$ is a continous but might be non-convex function in Eq (3.1), we analyze the convergence of the proposed distributed algorithm by means of the convergence idea of the ADMM algorithm for non-convex problems. At first, we make the following assumption.

Assumption 1. (i) For all k, $\eta_k \in \Theta$, where Θ is a closed convex sets on R^2. (ii) For all k, the step size ρ_k is chosen large enough such that

$$\rho_k(\lambda_{\min}(G_k) + \rho_k) > 2\lambda_{\max}^2(G_k), \tag{3.9}$$

where

$$\begin{aligned} \lambda_{\min}(G_k) &= \min_{\eta} \lambda_{\min}(G_k(\eta)), \\ \lambda_{\max}(G_k) &= \max_{\eta} \lambda_{\max}(G_k(\eta)), \end{aligned} \tag{3.10}$$

where G_k is the Hessian matrix of g_k, $\lambda_{\min}(G_k(\eta))$ and $\lambda_{\max}(G_k(\eta))$ are respectively the maximum and minimum eigenvalues of G_k for $\eta_k \in \Theta$.

Theorem 3.1. If the Assumption 1 is satisfied, the proposed distributed algorithm converges to the set of stationary solutions of problem Eq (3.1).

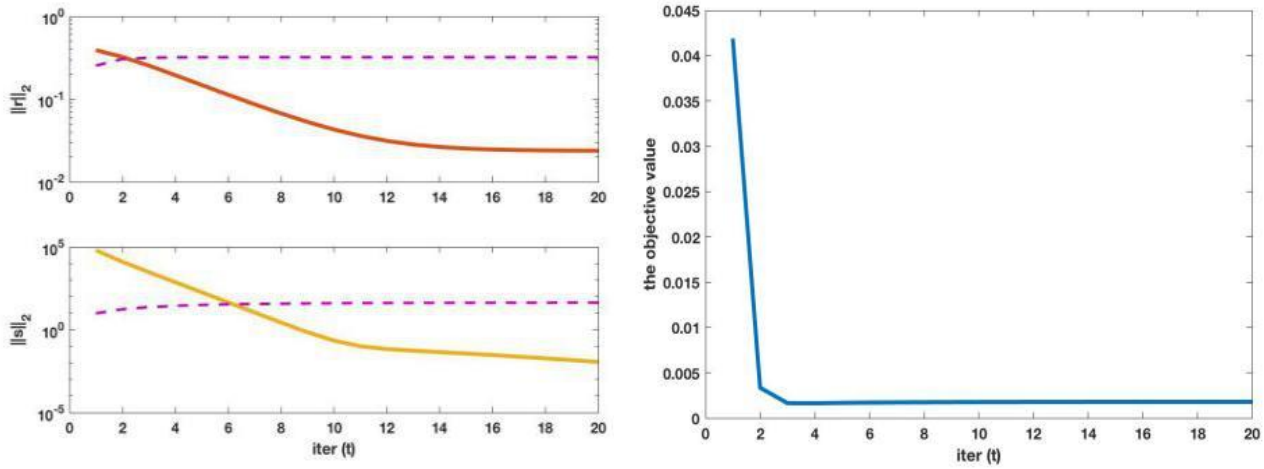
In addition, the termination criterion for this distributed algorithm is that the primal and dual residual must be small, that is

$$\| r^t \|_2 = \sqrt{\sum_{k=1}^{K} \| \eta_k^t - \theta^t \|_2^2} \le \varepsilon^{pri}, \qquad \| s^t \|_2 = \sqrt{K}\rho \| \theta^t - \theta^{t-1} \|_2 \le \varepsilon^{dual}, \tag{3.11}$$

where $\rho = \max_k \{\rho_k\}$, and

$$\begin{aligned} \varepsilon^{pri} &= \sqrt{2K}\varepsilon^{abs} + \varepsilon^{rel} \max(\| \eta^t \|_2, \sqrt{K} \| \theta^t \|_2), \\ \varepsilon^{dual} &= \sqrt{2K}\varepsilon^{abs} + \varepsilon^{rel} \max(\| \eta^t \|_2, \| y^t \|_2), \end{aligned} \tag{3.12}$$

where $\varepsilon^{abs} > 0$ and $\varepsilon^{rel} > 0$ are respectively an absolute and a relative tolerance. The choice of ε^{abs} depends on the scale of typical variable value and a reasonable value of ε^{rel} might be 10^{-3} or 10^{-4} depending on the application.



(a) The primal and dual residual (b) The objective function

Figure 2. Convergence rate of the distributed algorithm for parameter estimation. (a) The primal and dual residual as a function of the number of iterations iter (t), where the dotted lines respectively denote the upper bound of error of the primal and dual residual in each iteration. (b) The objective function as a function of the number of iterations iter (t).

Furthermore, in order to have a more intuitive understanding of the convergence rate of this distributed algorithm, we take $\xi = 1$, $\sigma = 10$ to generate 10^6 observations from GPD (10,1) via simulation. These observations are stored into 10 nodes, we select 95% as threshold for each data subset and choose $\rho = \max_k \{\rho_k\} = 2500$ where ρ_k is computed by Eq (3.9). The distributed algorithm is applied to estimate the parameter, and the algorithm convergence is showed in Figure 2. Figure 2(a) shows that the primal and dual residual are decreasing as the number of iterations increases, and this algorithm only iterates a few steps to reach the given termination condition. Figure 2(b) shows

that the objective function decreases rapidly and then stabilizes as the number of iterations increases. Therefore, these results can be seen that our algorithm convergence rate is fast. For simplicity, the distributed algorithm is called ADMM-WNLS.

4. Results and discussions

In this section, we will discuss performances of the distributed ADMM-WNLS method on estimation accuracy through simulation. Here, we firstly describe the simulation procedure as follow.

(a) Generate N random observations from the given distribution, and these observations are stored into K node and the sample size of each node is m .

(b) Fix $0 << q < p < 1,$ where q and p are used for setting threshold value and targeted quantile, respectively.

(c) For each node observations, select the $100q$-th quantile as the threshold u_k , $k = 1, 2, \ldots, K$.

(d) Fit the GPD using the ADMM-WNLS method to estimate (ξ, σ) with all the observations above u_k, for $k = 1, 2, \ldots, K,$ and then estimate the quantile $\hat{x}_p$

$$\hat{x}_p = \frac{1}{K}\sum_{k=1}^{K} u_k + \frac{\hat{\sigma}}{\hat{\xi}}[(\frac{1-p}{1-q})^{-\hat{\xi}} - 1]. \tag{4.1}$$

(e) Repeat the above steps 1000 times to compute the absolute relative bias and the root of mean square errors of quantile.

4.1. Comparison between the ADMM-WNLS and WNLS methods

In this part, we discuss performances of the distributed ADMM-WNLS method on quantile estimation accuracy compared with the WNLS method with full sample.

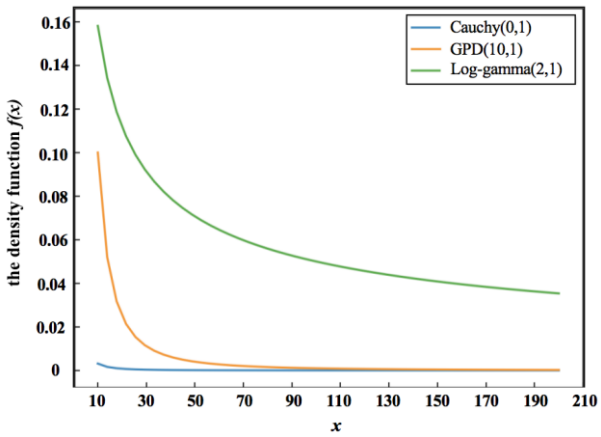


Figure 3. The tail figure of three heavy-tailed distributions.

The simulated samples are generated from the three common heavy-tailed distributions Log-gamma (2,1), GPD (10,1) and Cauchy (0,1). The degree of heavy tail of these three distributions is Log-gamma (2,1) > GPD (10,1) > Cauchy (0,1) as shown in Figure 3. We try sample sizes of 10^6 and 10^7, choose 90%, 93% and 95% sample quantiles for each node data as each node thresholds,

and node number $K=10$. The estimated quantiles p are 95%, 99%, 99.9% and 99.99%. The ARB and RMSE are computed to compare the estimation performances of the two methods. The ARB is defined as $|\hat{\theta}-\theta|/\theta$ where $\hat{\theta}$ and θ are the estimated and true quantiles, respectively. In this paper, we only present the results of the case of sample size of 10^6, the results of sample size 10^7 are similar. The simulated results are presented in Table 1.

Table 1. Quantile estimation results under Log-gamma (2,1), GPD (10,1) and Cauchy (0,1): RMSE and ARB (in parenthesis).

Threshold	Method	0.95	0.99	0.999	0.9999
Log-gamma (2,1)					
q = 0.90	WNLS	0.5607 (0.0040)	7.4622 (0.0077)	772.77 (0.0705)	27555 (0.2063)
	ADMM-WNLS	0.5586 (0.0039)	7.4361 (0.0077)	746.61 (0.0679)	26932 (0.2014)
q = 0.93	WNLS	0.6060 (0.0042)	8.5094 (0.0087)	550.51 (0.0477)	21016 (0.1535)
	ADMM-WNLS	0.6060 (0.0042)	8.8412 (0.0091)	520.52 (0.0477)	20247 (0.1473)
q = 0.95	WNLS	\	8.6016 (0.0088)	446.01 (0.0358)	17158 (0.1167)
	ADMM-WNLS	\	8.9735 (0.0092)	415.18 (0.0328)	16187 (0.1087)
\	EMP	0.6223 (0.0041)	8.7913 (0.0089)	339.36 (0.0278)	15220 (0.0984)
GPD (10,1)					
q = 0.90	WNLS	0.7534 (0.0031)	7.8255 (0.0064)	213.69 (0.0170)	3548.48 (0.0288)
	ADMM-WNLS	0.7504 (0.0031)	7.9157 (0.0065)	207.93 (0.0168)	3529.43 (0.0285)
q = 0.93	WNLS	0.8835 (0.0037)	8.6265 (0.0072)	248.64 (0.0205)	4539.23 (0.0379)
	ADMM-WNLS	0.8807 (0.0037)	8.6518 (0.0073)	248.68 (0.0205)	4401.53 (0.0368)
q = 0.95	WNLS	\	6.9982 (0.0057)	228.37 (0.0182)	4328.99 (0.0344)
	ADMM-WNLS	\	7.1027 (0.0057)	224.87 (0.0177)	4265.73 (0.0337)
\	EMP	0.8404 (0.0036)	8.6347 (0.0068)	340.08 (0.0288)	6684.76 (0.0803)
Cauchy (0,1)					
q = 0.90	WNLS	0.0299 (0.0038)	0.2544 (0.0063)	9.6805 (0.0252)	230.2367 (0.0636)
	ADMM-WNLS	0.0299 (0.0083)	0.2579 (0.0064)	10.2597 (0.0273)	242.5223 (0.0679)
q = 0.93	WNLS	0.0275 (0.0034)	0.2894 (0.0071)	8.2730 (0.0212)	137.6747 (0.0357)
	ADMM-WNLS	0.0274 (0.0034)	0.2911 (0.0072)	8.6283 (0.0221)	142.0018 (0.0369)
q = 0.95	WNLS	\	0.2574 (0.0065)	8.0840 (0.0203)	149.6838 (0.0381)
	ADMM-WNLS	\	0.2578 (0.0065)	8.2916 (0.0210)	149.1795 (0.0386)
\	EMP	0.0282 (0.0034)	0.3142 (0.0081)	10.2107 (0.0260)	324.8123 (0.0791)

Compared with the WNLS method, the distributed ADMM-WNLS method has a relatively good estimator for all cases and it performs better for 99.99% quantile in some cases. For Log-gamma (2,1), the ADMM-WNLS performs better for all cases. For GPD (10,1), the ADMM-WNLS performs relatively poor for the 95% and 99% quantiles, however, it gives greatly better results for the extreme quantiles (99.9% and 99.99%) with larger threshold values. For Cauchy (0,1), the two methods perform almost the same for the 95% quantile, and the ADMM-WNLS performs relatively poor for other higher quantiles. These results are shown that the estimation performance of the proposed distributed algorithm is better when the tail of distribution is thicker and the targeted quantiles is higher. It can been found from Table 1 that the different thresholds have little effect on the estimation accuracy, which illustrates the proposed ADMM-WNLS method is less

sensitive to the selection of the threshold.

In additoion, the (non-parametric) empirical quantiles are usually very good estimators for quantiles when sample size is large. For this, we also compute the corresponding empirical quantiles (EMP) with full sample (this result is shown in the last column of each example, Table 1). Overall, the distributed ADMM-WNLS performs better than the EMP. This result is consistent with Song and Song [13] who found that the nonlinear least square method performs better than the EMP with 10^4 observations. Furthermore, we suggest to choose the larger quantile as threshold for the heaviest-tailed distribution such as Log-gamma (2,1) because it is difficult to accurately describe the tail characteristics of this type of distribution if the selected threshold is small.

4.2. Comparison between the ADMM-WNLS and mean-WNLS methods

Next, we will compare performances of the proposed ADMM-WNLS method and another distributed method on parameter estimation and high quantile estimation of heavy-tailed distributions via simulation.

Table 2. The RMSE and BIASa of GPD parameter estimation with 95% threshold value.

sample size	parameter	RMSE		BIAS[a]	
		mean-WNLS	ADMM-WNLS	mean-WNLS	ADMM-WNLS
10^4	$\sigma=1$	0.2571	0.1546	0.1853	0.0647
	$\xi=1$	0.0656	0.0529	0.0385	0.0231
	$\sigma=3$	0.6915	0.3654	0.5774	0.2938
	$\xi=1$	0.0728	0.0819	0.0545	0.0397
	$\sigma=10$	2.4036	1.8136	1.8046	1.0228
	$\xi=1$	0.0700	0.0707	0.0451	0.0465
	$\sigma=3$	2.8364	1.7676	2.3902	1.2315
	$\xi=2$	0.1473	0.1446	0.1157	0.1077
10^5	$\sigma=1$	0.0490	0.0480	0.0257	0.0254
	$\xi=1$	0.0173	0.0173	0.0081	0.0084
	$\sigma=3$	0.1393	0.1277	0.0711	0.0493
	$\xi=1$	0.0173	0.0173	0.0081	0.0082
	$\sigma=10$	0.2895	0.2740	0.1585	0.0909
	$\xi=1$	0.0173	0.0173	0.0081	0.0081
	$\sigma=3$	0.3369	0.2744	0.2038	0.1013
	$\xi=2$	0.0316	0.0316	0.0096	0.0065
10^6	$\sigma=1$	0.0135	0.0134	0.0023	0.0016
	$\xi=1$	0.0056	0.0056	0.0008	0.0008
	$\sigma=3$	0.0382	0.0381	0.0038	0.0016
	$\xi=1$	0.0055	0.0054	0.0005	0.0005
	$\sigma=10$	0.1273	0.1273	0.0128	0.0040
	$\xi=1$	0.0057	0.0056	0.0005	0.0005
	$\sigma=3$	0.0860	0.0849	0.0152	0.0044
	$\xi=2$	0.0100	0.0100	0.0010	0.0010

BIAS[a]: the bias of each estimator.

4.2.1. Parameter estimation

The distributed mean-WNLS method averages the estimators of each node. To compare the performance of the ADMM-WNLS and mean-WNLS methods, we generate the GPD random variables with the shape parameter $\xi = (1,2)$ and the scale parameter $\sigma = (1,3,10)$. We try sample sizes of 10^4, 10^5 and 10^6, node number $K = 10$ and the 90%, 95% sample quantiles as the threshold values. In this paper, we only present the results for the case of the 95% threshold value, the results for the 90% threshold value are similar. The simulation results of some parameters are listed in Table 2. As shown in Table 2, both the RMSE and BIAS values decrease as the sample size increases for the same parameter, and the overall performances of ADMM-WNLS estimators are better for all cases. Moreover, the performance of ADMM-WNLS becomes greatly better as the scale parameter increases for a fixed shape parameter and fixed sample size. Therefore, the parameter estimation accuracy of the ADMM-WNLS method is more higher as the tail of distribution is more thicker.

4.2.2. High quantile estimation

Next, we compare the performance of the ADMM-WNLS and mean-WNLS in estimating high quantiles of heavy-tailed distributions. Sample sizes of 10^4, 10^5 and 10^6 are randomly generated from Cauchy (0,1), GPD (10,1), Log-gamma (2,1) and node number $K = 10$. We choose the 95% sample quantile as threshold value. The estimated quantiles p are 99%, 99.9% and 99.99%. The ARB_{wnls} is defined as $|\hat{\theta} - \theta_{wnls}|/\theta_{wnls}$, where $\hat{\theta}$ is the estimated quantile and θ_{wnls} is the WNLS estimator with full sample. The simulation results are listed in Table 3.

Table 3. Quantile estimation results for Cauchy (0,1) GPD (10,1) and Log-gamma (2,1) with 95% threshold value.

Distribution	Sample size	Method	ARB_{wnls}		
			0.99	0.999	0.9999
Cauchy (0,1)	10^4	ADMM-WNLS	0.0896	0.2186	0.3517
		mean-WNLS	0.0522	0.2208	0.3860
	10^5	ADMM-WNLS	0.0081	0.0364	0.0630
		mean-WNLS	0.0076	0.0367	0.0643
	10^6	ADMM-WNLS	0.0009	0.0037	0.0073
		mean-WNLS	0.0008	0.0038	0.0073
GPD (10,1)	10^4	ADMM-WNLS	0.0863	0.0787	0.1287
		mean-WNLS	0.0538	0.2180	0.3697
	10^5	ADMM-WNLS	0.0115	0.0354	0.0590
		mean-WNLS	0.0074	0.0370	0.0679
	10^6	ADMM-WNLS	0.0011	0.0040	0.0068
		mean-WNLS	0.0008	0.0041	0.0076
	10^4	ADMM-WNLS	0.0998	0.1525	0.1799
		mean-WNLS	0.0582	0.2362	0.3894

Continued on next page

Distribution	Sample size	Method	ARB $_{wnls}$		
			0.99	0.999	0.9999
Log-gamma (2,1)	10^5	ADMM-WNLS	0.0128	0.0333	0.0518
		mean-WNLS	0.0079	0.0406	0.0718
	10^6	ADMM-WNLS	0.0013	0.0043	0.0075
		mean-WNLS	0.0009	0.0044	0.0081

For three heavy-tailed distributions, the ADMM-WNLS performs better for the extreme high quantiles such as 99.9% and 99.99% in many cases when sample sizes are 10^4 and 10^5. When sample size is 10^6, the ADMM-WNLS performs better for the extreme high quantiles such as 99.9% and 99.99% of the Log-gamma distribution and the GPD while the ADMM-WNLS doesn't perform worse for the high quantiles of the Cauchy distribution. These results can be seen that the ADMM-WNLS is a clear-cut winner for the extreme high quantile levels of the heaviest-tailed distribution.

5. Real data applications

The dataset used in this section is the SOA group medical insurance large claims data in 1991, which is available on the Internet at http: // lstat.kuleuven.be/Wiley/. There are 75,789 observations and this dataset has a long tail as shown in Figure 4.

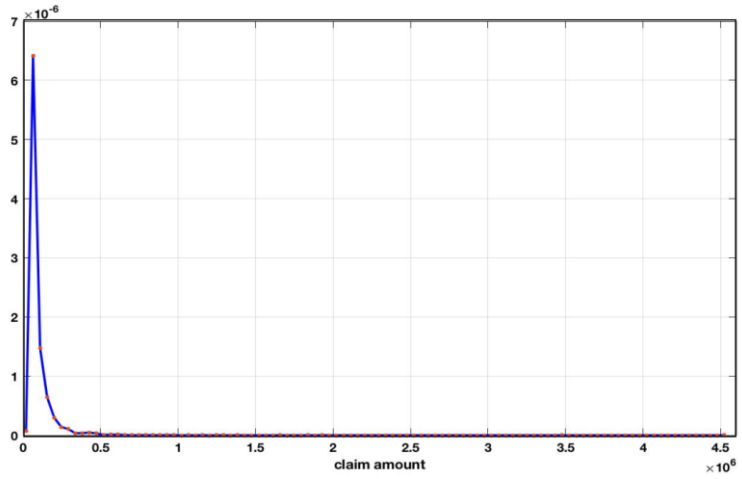


Figure 4. The estimated density curve of observations.

Next, we will apply the ADMM-WNLS and WNLS methods to estimate the extreme quantiles of this dataset. For simplicity, we take a sample of size 70,000 as our targeted data, select 90%, 93% and 95% sample quantiles as threshold values, and this data is stored into 10 nodes. Table 4 shows the ARB of high quantile estimates for this data. It can seen from Table 4 that the 99%, 99.9% quantiles are estimated well for the two methods and the choice of threshold has little effect on estimation accuracy of a fixed quantile. However, since the amount of this data itself is not very large, sample size of each node is relatively small after partitioning, which leads to a bit large bias of the distributed ADMM-WNLS method than the WNLS method with full sample for the 99.99% extremely high quantile. Therefore, our distributed method is applicable for the quantile estimation

of heavy-tailed distribution in real application.

Table 4. Quantile estimation results: ARB for SOA data.

Threshold	Method	ARB		
		0.99	0.999	0.9999
q = 0.95	ADMM-WNLS	0.0060	0.0241	0.0572
	WNLS	0.0058	0.0226	0.0414
q = 0.93	ADMM-WNLS	0.0057	0.0262	0.0620
	WNLS	0.0060	0.0256	0.0496
q = 0.90	ADMM-WNLS	0.0058	0.0205	0.0418
	WNLS	0.0061	0.0177	0.0311

6. Conclusions

With the amount of data increasing, it is impossible to store all of them on a single machine and the distributed storage architectures have been widely used. In this background, this paper attempt to find an algorithm to compute the quantiles of heavy-tailed distributions in a distributed storage architecture. To this end, we propose the ADMM-WNLS method under the POT framework. The key idea of the ADMM-WNLS method is to combine the distributed optimization algorithm (ADMM) and the efficient quantile estimation method (WNLS). Moreover, we discuss the distributed ADMM-WNLS performances on estimating the high quantiles through simulation study and real data analysis, and it can be seen that the distributed ADMM-WNLS provides a feasible and efficient solution to estimate high quantiles for heavy-tailed distributions in a distributed manner. In addition, Our proposed method is a centralized distributed estimation algorithm, the master node plays an important role in this algorithm procedures. When we conduct this algorithm, we need to ensure that the master node can complete this task normally in advance. If the master node fails, this distributed algorithm will not be applicable. To avoid this failure problem, we will attempt to extend the proposed algorithm to the decentralized distributed algorithm in the future.

Acknowledgments

We would like to thank the editor and reviewers for their valuable comments and suggestion that lead to the significant improvement of the paper.

Conflict of Interest

The authors declare there is no conflict of interest.

References

1. H. Rootzén, R. W. Katz, Design life level: Quantifying risk in a changing climate. *Water Resour. Res.*, **49** (2013), 5964–5972.

2. M. M. de Oliveira, N. F. Ebecken, J. L. de Oliveira, E. Gilleland, Generalized extreme wind speed distributions in south America over the Atlantic Ocean region, *Theor. Appl. Climatol.*, **104** (2011), 377–385.
3. R. Potocky, M. Stehlik, H. Waldl, On sums of claims and their applications in analysis of pension funds and insurance products, *Prague Econ. Pap.*, **23** (2014), 349–370.
4. P. Jordanova, Z. Fabian, P. Hermann, L. Střelec, A. Rivera, S. Girard, et al., Weak properties and robustness of t-hill estimators, *Extremes*, **19** (2016), 591–626.
5. M. Stehlík, L. N. Soza, Z. Fabián, M. Jiřina, P. Jordanova, S. C. Arancibia, et al., On ecological aspects of dynamics for zero slope regression for water pollution in Chile, *Stochastic Anal. Appl.*, **37** (2019), 574–601.
6. J. Pickands, Statistical inference using extreme order statistics, *Ann. Stat.*, **3** (1975), 119–131.
7. J. Hosking, J. Wallis, Parameters and quantile estimation for the generalized pareto distribution, *Technometrics*, **29** (1998), 339–349.
8. S. Juarez, W. Schucany, Robust and efficient estimation for the generalized pareto distribution, *Extremes*, **7** (2004), 237–251.
9. J. Zhang, Likelihood moment estimation for the generalized pareto distribution, *Aust. N. Z. J. Stat.*, **49** (2007), 69–77.
10. J. Zhang, Improving on estimation for the generalized pareto distribution, *Technometrics*, **52** (2010), 335–339.
11. J. Zhang, M. Stephens, A new and efficient estimation method for the generalized pareto distribution, *Technometrics*, **51** (2009), 316–325.
12. J. He, Z. Sheng, B. Wang, K. Yu, Point and exact interval estimation for the generalized Pareto distribution with small samples, *Stats its interface*, **7** (2014), 389-404.
13. J. Song, S. Song, A quantile estimation for massive data with generalized Pareto distribution, *Comput. Stat. Data Anal.*, **56** (2012), 143–150.
14. M. H. Park, J. H. T. Kim, Estimating extreme tail risk measures with generalized Pareto distribution, *Comput. Stat. Data Anal.*, **98** (2016), 91–104.
15. S. Kang, J. Song, Parameter and quantile estimation for the generalized pareto distribution in peaks over threshold framework, *J. Korean Stat. Soc.*, **46** (2017), 487–501.
16. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.*, **3** (2010), 1–122.
17. E. Chu, A. Keshavarz, S. Boyd, A distributed algorithm for fitting generalized additive models, *Optim. Eng.*, **14** (2013), 213–224.
18. X. Yuan, Alternating direction method for covariance selection models, *J. Sci. Comput.*, **51** (2012), 261–273.
19. Y. Gu, J. Fan, L. Kong, S. Ma, H. Zou, ADMM for high-dimensional sparse penalized quantile regression, *Technometrics*, **60** (2018), 319–331,
20. M. Hong, Z. Q. Luo, M. Razaviyayn, Convergence analysis of alternating direction method of multipliers for a family of non-convex problems, *SIAM J. Optim.*, **26** (2014), 3836–3840.
21. B. He, X. Yuan, On the O(1/n) convergence rate of the douglas-rachford alternating direction method, *SIAM J. Numer. Anal.*, **50** (2012), 700–709.
22. W. Deng, W. Yin, On the global and linear convergence of the generalized slternating direction method of multipliers, *J. Sci. Comput.*, **66** (2016), 889–916.

23. J. Liu, S. J. Wright, C. Ré, V. Bittorf, S. Sridhar, An asynchronous parallel stochastic coordinate descent algorithm, *J. Mach. Learn. Res.*, **16** (2013), 285–322.
24. H. R. Feyzmahdavian, A. Aytekin, M. Johansson, An asynchronous mini-batch algorithm for regularized stochastic optimization, *IEEE Trans. Autom. Control*, **61** (2016), 3740–3754.
25. A. McNeil, T. Saladin, The peaks over thresholds method for estimating high quantiles of loss distributions, *Proc. 28th Int. ASTIN Colloq.*, (1997), 23–43.
26. A. A. Balkema, L. de Haan, Residual life time at great age, *Ann. Probab.*, **2** (2004), 792–804.
27. P. Embrechts, C. Kluppelberg, T. Mikosch, Modelling Extremal Events for Insurance and Finance, Springer-Verlag, Berlin Heidelberg, 1997.
28. H. Zhu, A. Cano, G. Giannakis, Distributed consensus-based demodulation: Algorithms and error analysis, *IEEE Trans. Wireless Commun.*, **9** (2010), 2044–2054.

AIMS Press