Research Article

Applied Stochastic Models in Business and Industry

Received 21 January 2010, Revised 18 March 2012, Accepted 21 March 2012 Published online 5 July 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.1927

Statistical inference in massive data sets

Runze Li, Dennis K. J. Lin*† and Bing Li

Analysis of massive data sets is challenging owing to limitations of computer primary memory. In this paper, we propose an approach to estimate population parameters from a massive data set. The proposed approach significantly reduces the required amount of primary memory, and the resulting estimate will be as efficient if the entire data set was analyzed simultaneously. Asymptotic properties of the resulting estimate are studied, and the asymptotic normality of the resulting estimator is established. The standard error formula for the resulting estimate is proposed and empirically tested; thus, statistical inference for parameters of interest can be performed. The effectiveness of the proposed approach is illustrated using simulation studies and an Internet traffic data example. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Internet traffic data; kernel density estimation; remedian

1. Introduction

In the past decade, we have witnessed a revolution in information technology. Consequently, routine collection of systematically generated data is now in commonplace. Databases with hundreds of fields, billions of records, and terabytes of information are not unusual. For example, Barclaycard (UK) carries out 350 million transactions a year, Wal-Mart makes over 7 billion transactions a year, and AT&T carries over 70 billion long distance calls annually (see [1] for more details). It is very challenging to extract useful features from a massive data set because many statistics are difficult to compute by standard algorithms or statistical packages when the data set is too large to be stored in primary memory.

Unlike observations resulting from designed experiments, massive data sets sometimes become available without predefined purposes or only with rather vague purposes. Typically, it is desirable to find some interesting features in the data sets that will provide valuable information to support decision making. Primary tasks in analyzing massive data sets include data processing, classification, detection of abnormal patterns, summarization, visualization, and association/correlation analysis.

To obtain a summary and preliminary analysis of a massive data set, some basic statistics are of general interest. For example, to construct a box plot for a massive data set, we need sample quartiles. This is not a trivial task on a massive data set. Consider the problem of percentile estimation. Suppose, given independent observations $x_1, x_2, \ldots, x_n$ from an unknown distribution function F, that we want to find its 100αth percentile, that is, the number ξ_α such that $F(\xi_\alpha) = \alpha$. This is similar to the problem of finding the kth smallest of n observations; an estimate of the 100αth population percentile provides an approximation to the $[\alpha n]$ largest observation. This seems to be a straightforward problem once all observations have been sorted. A major difficulty arises, however, when the available computer memory is much smaller than n. Then sorting $x_1, x_2, \ldots, x_n$ becomes impossible. To overcome the difficulty, Rousseeuw and Bassett [2] proposed the remedian method, and Hurley and Modarres [3] proposed a low-storage quantile estimation method. Chao and Lin [4] studied the asymptotic behaviors of the remedian approach and found that the resulting estimator does not possess asymptotic normality. To obtain quartile estimation accurately and efficiently, many efforts have been made. Mahmoud [5] gave a systematic study on this topic. Unfortunately, many existing methods focus only on the estimation of population medians, quartiles, or percentiles (e.g., [6]). To make statistical inferences for them, one has to know their estimator variation with finite samples. For many parameters, such as medians and percentiles, their asymptotic standard errors depend on other unknown parameters [7, p. 91]. This imposes more challenge to draw statistical inferences on these parameters.

Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111, USA
**Correspondence to: Dennis K. J. Lin, Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111, USA.*
†*E-mail: dkl5@psu.edu*

This work was motivated by the analysis of several massive real-life data sets. One particularly interesting subject here is the streaming data. Streaming data are becoming widely available from a variety of sources (e.g., [6]). In Section 5, a special type of streaming data—an Internet traffic data set—will be analyzed using our proposed method. Internet engineering and management depend on an understanding of the characteristics of network traffic. Much work has been performed in developing various statistical models for Internet data. In this paper, we will focus on developing computational and inferential tools for massive data including Internet traffic data and electronic commerce data. A general approach is proposed to deal with statistical inferences on data sets with very large sample sizes.

Our approach can significantly reduce the need of large primary memory and makes statistical inferences on massive data sets feasible. This is very important in analyzing such data sets, because the number of observations that can be stored in primary memory is often restricted. Furthermore, many computing environments also limit the allowed maximum array size. This can be rather smaller than necessary and independent of the available memory.

One of the advantages of our approach is its generality. The proposed approach is widely applicable and capable of making statistical inferences for any parameter $\theta(F)$ of a population F. By parameter, we mean any unknown quantity associated with the unknown population F. The parameter $\theta(F)$ can be the population percentile, the unknown regression coefficient, or the unknown density function. It will be shown that under some mild conditions, the resulting estimator is consistent and has a normal limiting distribution. Furthermore, we will show that in many situations, the resulting estimate is as efficient if all data were simultaneously used to compute the estimate (Remark 3).

On the basis of the asymptotic normality, we further propose a standard error formula for the resulting estimate. The proposed standard error formula does not involve other unknown quantities of the unknown population. The standard error formula is empirically tested by Monte Carlo simulation and is accurate. Thus, one can directly make statistical inferences for the parameter of interest. This is another advantage of our approach compared with the existing methods for computing sample quartiles.

The paper is organized as follows. Section 2 provides the basic idea of the proposed method and the theoretical justifications. Section 3 discusses the problem of point estimation from the massive data set. Section 4 discusses the problem of density estimation. Section 5 visits the popular Internet traffic data from AT&T. Final conclusions are given in Section 6. For simplicity of presentation, all proofs are given in Appendix A.

2. The proposed estimation procedure

To estimate a parameter $\theta(F)$ of a population F, such as a percentile or the density of the population, it is frequently required to store the entire data set in primary memory to obtain an efficient estimate. One way to overcome the aforementioned difficulty in memory is to use subsampling techniques. This approach is useful for preliminary analysis, but the estimator is less efficient as it only uses information in parts of the data.

For efficiency, an estimator should be derived on the basis of the whole data set rather than on its parts, which is not feasible for massive data sets. Intuitively, we may sequentially read and store the data in primary memory, block by block, and analyze the data in each block separately. As long as the size of the block is small, one can easily implement this estimation procedure within each block under various computing environments. A question that arises here is on how to make a final conclusion based on the results obtained from each block.

Suppose that there is an independent and identically distributed sample with large sample size n, and we are interested in finding an estimate of the population median. To find the sample median, one needs at least n storage elements. When n is large (e.g., 10,000,000), standard algorithms for computing the sample median may exceed the available memory and thus may fail. However, it is easy to compute a sample median of 10,000 in many statistical packages, such as S-plus and SAS. We may sequentially read in the data block by block (each having, say, 10,000 samples) and then compute the sample median of each block, which leads to an independent and identically distributed set of sample medians. It has been shown under some mild conditions that these sample medians are independent and asymptotically distributed as normal with mean equal to the population median. Thus, a natural estimate for the population median is then the average of these 1000 sample medians. In summary, to estimate parameter $\theta(F)$ based on a massive data set, we may employ a two-stage procedure. First, read in the whole data set sequentially block by block, each having a manageable sample size, and compute an estimate of $\theta(F)$ within each block. Second, take the average of the resulting estimates obtained from each block as an estimate for $\theta(F)$. Note that the second stage can be updated as soon as a new block is processed by the first stage and hence does not require additional memory.

2.1. Sampling properties

Suppose that $x_1, \ldots, x_n$ is an independent and identically distributed sample from population F, where x_i can be either a random variable or a random vector. We are interested in estimating parameter $\theta(F)$ of the population. To formulate our estimation procedure, we rewrite the sample as

$$\begin{array}{cccc} x_{11}, & x_{12}, & \ldots, & x_{1\alpha_n} \\ x_{21}, & x_{22}, & \ldots, & x_{2\alpha_n} \\ \vdots & \vdots & \ldots, & \vdots \\ x_{\beta_n 1}, & x_{\beta_n 2}, & \ldots, & x_{\beta_n \alpha_n}, \end{array}$$

where $x_{ij} = x_{(i-1)*\alpha_n+j}$. For $i = 1, \ldots, \alpha_n$ and $j = 1, \ldots, \beta_n$, α_n is the block size and β_n is the total number of blocks. Note that $n = \alpha_n \beta_n$. The block size α_n is chosen so that the estimation of θ can be easily handled within a block. The choice of α_n will be discussed later. It is shown in Sections 3 and 4 that the resulting estimate is robust to the choice of α_n. We use the same estimator for each block. Denote by $\hat{\theta}_{in}$ the resulting estimate based on the subsample in the ith block $x_{i1}, \ldots, x_{i\alpha_n}$. We estimate $\theta(F)$ by averaging of $\hat{\theta}_{in}$; that is,

$$\bar{\theta} = \frac{1}{\beta_n} \sum_{i=1}^{\beta_n} \hat{\theta}_{in}. \tag{1}$$

Now we investigate the sampling properties of the estimate $\bar{\theta}$. Denote $\mu_n = \mathrm{E}(\hat{\theta}_{in})$ and $\sigma_n^2 = \mathrm{var}(\hat{\theta}_{in})$. We have the following proposition.

Proposition 1
Suppose that $x_1, \ldots, x_n$ are independent and identically distributed, $b_n = \mu_n - \theta \to 0$, and $\sigma_n^2/\beta_n \to 0$ as $n \to \infty$. Then $E|\bar{\theta} - \theta|^2 \to 0$.

Under the conditions of Proposition 1, it follows from Chebyshev's inequality that $\bar{\theta}$ tends to θ in probability. Thus, $\bar{\theta}$ is a consistent estimator for θ.

To establish the asymptotic normality of $\bar{\theta}$, we need one of the following two conditions.

Condition (a): α_n is a constant independent of n and $\sigma_n^2 < \infty$.
Condition (b): $\alpha_n \to \infty$ and $\beta_n \to \infty$ as $n \to \infty$, and

$$\frac{E|\hat{\theta}_{in} - \mu_n|^{2+\delta}}{\beta_n^{\delta/2} \sigma_n^{2+\delta}} \to 0 \tag{2}$$

as $n \to \infty$ for some $\delta > 0$.

Theorem 1
Suppose that $x_1, \ldots, x_n$ are independent and identically distributed. If either Condition (a) or (b) holds, then

$$\sqrt{\beta_n} \left(\frac{\bar{\theta} - \mu_n}{\sigma_n} \right) \to N(0, 1) \tag{3}$$

in distribution as $n \to \infty$.

Remark 1
When α_n is a fixed finite number, μ_n and σ_n^2 do not depend on n and can be denoted by μ and σ^2, and then

$$\sqrt{\beta_n} \left(\frac{\bar{\theta} - \theta}{\sigma} \right) \to N(0, 1)$$

holds if and only if $\hat{\theta}_{in}$ is unbiased estimators of θ. If $\hat{\theta}_{in}$ is a biased estimator and the bias $\mu - \theta$ is a constant, then the resulting estimator is not consistent.

Remark 2
In many situations in which $\alpha_n \to \infty$,

$$\frac{\hat{\theta}_{in} - \mu_n}{\sigma_n} \to N(0, 1)$$

in distribution. This makes Condition (b) a natural assumption.

2.2. *Choice of $\boldsymbol{\alpha_n}$*

Note that

$$\sqrt{\beta_n}\left(\frac{\bar{\theta}-\theta}{\sigma_n}\right)=\sqrt{\beta_n}\left(\frac{\bar{\theta}-\mu_n}{\sigma_n}\right)+\sqrt{\beta_n}\left(\frac{\mu_n-\theta}{\sigma_n}\right).$$

Thus, $\mu_n-\theta=o(\sigma_n/\sqrt{\beta_n})$, which implies that $\sqrt{\beta_n}\left(\frac{\mu_n-\theta}{\sigma_n}\right)=o_P(1)$. By the Slutsky theorem, it follows from Theorem 1 that

$$\sqrt{\beta_n}\left(\frac{\bar{\theta}-\theta}{\sigma_n}\right)\to N(0,1). \tag{4}$$

When $\alpha_n\to\infty$ and the underlying estimator is of the form given in Proposition 1, $\sigma_n=O(1/\sqrt{\alpha_n})$. This implies that if

$$\mu_n-\theta=o(1/\sqrt{n}) \tag{5}$$

holds, then (4) holds. If $\hat{\theta}_{in}$ is unbiased, then $\mu_n-\theta=0$, and hence, (4) holds. For a biased estimator of θ in parametric settings, usually we have $\mu_n-\theta=O(1/\alpha_n)$. Thus, it is necessary that $\alpha_n/\sqrt{n}\to\infty$. Frequently, the σ_n and the bias $\mu_n-\theta$ decrease as α_n decreases. Therefore, we suggest to take $\alpha_n=O(\sqrt{n}\log\log(n))$.

Remark 3
Denote by $\hat{\theta}_n$ the estimator based on all samples $x_1,\ldots,x_n$. Suppose that $\sqrt{n}(\hat{\theta}_n-\theta)\to N\left(0,\sigma_0^2\right)$ in distribution as $n\to\infty$. Thus, $\sqrt{\alpha_n}(\hat{\theta}_{in}-\theta)\to N\left(0,\sigma_0^2\right)$ in distribution as $n\to\infty$. This implies that $\sqrt{\alpha_n}(\mu_n-\theta)\to 0$ and that $\alpha_n\sigma_n^2\to\sigma_0$. Note that

$$\sqrt{n}(\bar{\theta}-\theta)=\sqrt{\alpha_n\sigma_n^2}\sqrt{\beta_n}\left(\frac{\bar{\theta}-\mu_n}{\sigma_n}\right)+\sqrt{n}(\mu_n-\theta).$$

Under (5), it follows from Theorem 1 and the Slutsky theorem that $\sqrt{n}(\bar{\theta}-\theta)\to N(0,\sigma_0^2)$. This implies that the resulting estimator $\bar{\theta}$ is as efficient as $\hat{\theta}$. In other words, the resulting estimate is as efficient if all data were simultaneously used to compute the estimate.

3. Statistical inference

In this section, we first investigate statistical inferences for a single parameter θ when the sample size is large. We will discuss density estimation in next section.

3.1. *Confidence interval and testing hypothesis*

To draw statistical inferences on θ, we need to know its estimator variation with finite samples. In fact, $\hat{\theta}_{1n},\ldots,\hat{\theta}_{\beta_n n}$ provide us much information about the estimator $\bar{\theta}$. The information can be used for constructing a confidence interval for θ and test statistics for some hypotheses concerning θ.

The standard deviation of $\bar{\theta}$ is $\sigma_n/\sqrt{\beta_n}$, and σ_n can be directly estimated from the $\hat{\theta}_{1n},\ldots,\hat{\theta}_{\beta_n n}$; that is,

$$\hat{\sigma}_n=\left\{\frac{1}{\beta_n-1}\sum_{i=1}^{\beta_n}(\hat{\theta}_{in}-\bar{\theta})^2\right\}^{1/2}.$$

Thus, an estimator of the standard error of $\bar{\theta}$ is

$$\widehat{\mathrm{SE}}(\bar{\theta})=\frac{\hat{\sigma}_n}{\sqrt{\beta_n}}. \tag{6}$$

The standard error estimator (6) can be used to construct a confidence interval for θ. It can be further used to construct a t-test for $H_0:\theta=\theta_0$ versus $H_1:\theta\neq\theta_0$. For some parameters, such as percentiles, their standard error depends on the unknown population. However, the estimated standard error formula (6) allows us to avoid estimating the unknown population. This will be tested in our simulation example.

3.2. *Estimation of population percentiles*

The estimation of population percentiles requires sorting the entire sample and therefore will take a large amount of memory when the sample size is large. Some approaches of low-storage quantile estimation have been proposed in the literature (e.g., [2–4, 9]). Compared with their approaches, the proposed approach does not reduce the amount of storage very drastically. On the other hand, the proposed approach can be used to estimate a general population parameter, including medians and percentiles, and can be easily implemented on a personal computer. In this section, we illustrate our approach by a simulation example. All simulations in this paper were conducted using MATLAB (Mathworks Inc. Natick, MA 01760, USA) codes.

Example 1

In this example, we generated 1000 data sets, each consisting of $n = 8$ million independent and identically distributed random samples from a chi-square distribution with 1 degree of freedom. We illustrate our approach via estimating various percentiles of the population. In our simulation, we let $\alpha_n = 8000$, which is approximately equal to $\sqrt{n}\log\log(n)$. The simulation results are summarized in Table I. Denote the resulting estimate $\bar{\theta}_k$ in (1) and its standard error $\widehat{\mathrm{SE}}_k(\bar{\theta})$ in (6) based on the kth simulation data set for $k = 1, \ldots, 1000$. The second column of Table I reports the sample average of $\bar{\theta}_k$ (i.e., $\bar{\bar{\theta}} = \frac{1}{1000}\sum_{k=1}^{1000}\bar{\theta}_k$) over 1000 simulations. The third column reports the sample standard deviation of $\bar{\theta}_k$ (i.e., $\sqrt{\frac{1}{999}\sum_{k=1}^{1000}(\bar{\theta}_k - \bar{\bar{\theta}})^2}$) over 1000 simulations. Similarly, we report the sample average and the sample standard deviation of $\widehat{\mathrm{SE}}_k(\bar{\theta})$ in the fourth column. Specifically, $\widehat{\mathrm{SE}} = \frac{1}{1000}\sum_{k=1}^{1000}\widehat{\mathrm{SE}}_k(\bar{\theta})$ and $\mathrm{std}(\widehat{\mathrm{SE}}) = \left[\frac{1}{999}\sum_{k=1}^{1000}(\widehat{\mathrm{SE}}_k(\bar{\theta}) - \widehat{\mathrm{SE}})^2\right]^{1/2}$ in the fourth column. Comparing the last two columns, it can be found that the estimated standard error formula works surprisingly well. To obtain the standard error of the estimated percentiles, the proposed standard error formula allows us to avoid estimating the density of population. This is different from traditional approaches, which require estimation of some parameters, depending on the unknown density of population.

It is of interest to investigate how sensitive the results are on the choice of α_n. To this end, we took $\alpha_n = 2000$ and 32,000. Note that $n\log\log(n)$ is about 8000. To examine the effect of choice of α_n, we set $\alpha_n = 8000/4 = 2000$ and $8000 \times 4 = 32,000$. The results are described in Table II. Comparing the results based on the three different choices of α_n, we can see from Tables I and II that the choice of α_n is insensitive to the results, although the results using $\alpha_n = 8000$ and 32,000 seem to work slightly better than those using $\alpha_n = 2000$.

Table I. Estimated percentiles ($\alpha_n = 8000$).

p	True value	Estimate	$\mathrm{SE}_{\mathrm{true}}(10^{-4})$	$\widehat{\mathrm{SE}}$ (std($\widehat{\mathrm{SE}}$)) (10^{-4})
0.01	1.5709×10^{-4}	1.5902×10^{-4}	0.0114	0.0112 (0.0003)
0.05	3.9321×10^{-3}	3.9414×10^{-3}	0.1243	0.1219 (0.0028)
0.15	3.5766×10^{-2}	3.5794×10^{-2}	0.5962	0.6093 (0.0137)
0.25	0.1015	0.1016	1.3157	1.2855 (0.0296)
0.35	0.2059	0.2060	2.1628	2.1269 (0.0480)
0.45	0.3573	0.3574	3.2412	3.1513 (0.0714)
0.50	0.4549	0.4551	3.8397	3.7506 (0.0841)
0.55	0.5707	0.5708	4.5618	4.4294 (0.1002)
0.65	0.8735	0.8736	6.1789	6.1121 (0.1341)
0.75	1.3233	1.3236	8.8288	8.5601 (0.1939)
0.85	2.0723	2.0728	13.3976	12.8548 (0.2914)
0.95	3.8415	3.8436	26.5211	25.8665 (0.6072)
0.99	6.6349	6.6460	63.3566	62.7383 (1.4758)

Table II. Comparison between different choices of α_n.

	$\alpha_n = 2000$			$\alpha_n = 32000$		
Quartiles	$\hat{q}_i$	$\widehat{\mathrm{SE}}$	$\mathrm{SE}_{\mathrm{true}}$	$\hat{q}_i$	$\widehat{\mathrm{SE}}$	$\mathrm{SE}_{\mathrm{true}}$
$q_1 = 0.1015$	0.1017	1.287×10^{-4}	1.301×10^{-4}	0.1015	1.283×10^{-4}	1.291×10^{-4}
$q_2 = 0.4549$	0.4554	3.751×10^{-4}	3.830×10^{-4}	0.4550	3.750×10^{-4}	3.837×10^{-4}
$q_3 = 1.3233$	1.3246	8.577×10^{-4}	8.878×10^{-4}	1.3234	8.560×10^{-4}	8.827×10^{-4}

4. Nonparametric kernel density estimation

As argued in Remark 3, the proposed estimate can be as efficient as the estimator computed by simultaneously using all samples provided that $\mu_n - \theta = o(1/\sqrt{n})$, which is valid for unbiased estimator and can be valid for root n consistent but biased estimator by setting an appropriate α_n. Estimators in nonparametric kernel estimator and nonparametric regression are typically not of root n consistent. In this section, we discuss how to apply the proposed procedures in the context of estimation of density.

With the use of the subsample $x_{i1}, \ldots, x_{i\alpha_n}$ in the ith block, a kernel density estimator is as follows:

$$\hat{f}_h(x) = \frac{1}{\alpha_n} \sum_{j=1}^{\alpha_n} K_h(x_{ij} - x), \tag{7}$$

where $K_h(z) = \frac{1}{h} K(z/h)$, $K(z)$ is a kernel density function, and h is a selected bandwidth that controls the smoothness of the estimated density curve. The choice of kernel function is not crucial, but the choice of bandwidth h is critical. It is well known that

$$E\hat{f}_h(x) = f(x) + \frac{1}{2} h^2 \mu_2(K) f''(x) + o(h^2)$$

and

$$\mathrm{var}\left\{\hat{f}_h(x)\right\} = (\alpha_n h)^{-1} R(K) f(x) + o\left\{(nh)^{-1}\right\},$$

where $\mu_2(K) = \int z^2 K(z)\,\mathrm{d}x$ and $R(K) = \int K^2(z)\,\mathrm{d}z$ (e.g., [10, p. 20–21]). Thus, the mean square error is

$$\begin{aligned}
\mathrm{MSE}(\alpha_n) &= \frac{R(K) f(x)}{\alpha_n h \beta_n} + \frac{h^4}{4} \mu_2^2(K) \{f''(x)\}^2 + o\left\{h^4 + (\alpha_n \beta_n h)^{-1}\right\} \\
&= \frac{R(K) f(x)}{nh} + \frac{h^4}{4} \mu_2^2(K) \{f''(x)\}^2 + o\left\{h^4 + (nh)^{-1}\right\},
\end{aligned}$$

which is the same as that when we simultaneously use all the n samples to estimate the density. Integrating this expression leads to the mean integrated square error

$$\mathrm{MISE} = \mathrm{AMISE}\left\{\hat{f}_h(\cdot)\right\} + o\left\{h^4 + (nh)^{-1}\right\},$$

where the asymptotic mean integrated square error is

$$\mathrm{AMISE} = \frac{R(K)}{nh} + \frac{h^4}{4} \mu_2^2(K) \int \{f''(x)\}^2\,\mathrm{d}x.$$

Thus, the optimal bandwidth minimizing the asymptotic mean integrated square error is

$$h_{\mathrm{opt}} = \frac{R(K)}{\mu_2(K)^2 \int f''(x)\,\mathrm{d}x} n^{-1/5}, \tag{8}$$

which does not depend on α_n. It is well known that the optimal bandwidth of a kernel estimator for each block is of the order $O(\alpha_n^{-1/5})$ as the sample size of each block is α_n and that the bias of the resulting kernel estimator is of the order $\alpha_n^{-2/5}$. That is, $\mu_n - \theta = O(\alpha_n^{-2/5})$. Then, $\sqrt{n}(\mu_n - \theta) = O(\alpha_n^{1/10} \beta_n^{1/2})$, which may not tend to zero as $n \to \infty$. The optimal bandwidth in (8) implies that the estimated density curve in the first step should be undersmoothing to reduce bias, compared with the bandwidth $h = O(\alpha_n^{-1/5})$. Undersmoothing means that the bandwidth h_{opt} tends to zero faster than the usual $h = O(\alpha_n^{-1/5})$. Because the first multiplier in the h_{opt} only depends on the kernel function and on the curvature of the density function, many methods of bandwidth selection in the literature can be easily modified for our purpose. Denote by h^* the optimal bandwidth by using the data $x_{i,1}, \ldots, x_{i,\alpha_n}$ under some criterion. We may take

$$h_{\mathrm{opt}} = \left(\frac{\alpha_n}{n}\right)^{1/5} h^*$$

as our bandwidth. With this bandwidth, the resulting density estimation will be as good if we used the entire sample simultaneously.

Example 2
In this example, one million independent and identically distributed random samples were generated from the mixture normal distribution

$$0.5N(-2,1)+0.5N(2,1).$$

We want to estimate its density based on the random sample. In this example, $\alpha_n = 1000$, Gaussian kernel was used, and the bandwidth was selected using the rule of thumb (ROT) given in [11]. That is, the bandwidth used to estimate the density is $h_{\text{rot}} = 0.9 \times 1.06 \times \sigma \times n^{-1/5}$, where σ is the population standard deviation. Factor 1.06 is due to the Gaussian kernel, and multiplying by factor 0.9 is carried out to adjust for oversmoothing as noted in [11]. In our simulation, σ is substituted by its robust estimate, the mean of absolute deviation of the subsample $x_{i1}, \ldots, x_{i\alpha_n}$ within each block. We also investigated how sensitive the resulting estimate is to the choice of α_n. In this example, we took $\alpha_n = 500$, 1000, and 5000. The resulting estimated density curves are plotted in Figure 1. All of them are very close to the true density curve visually. To assess the performance for different α_n's, we define the root average squared error (RASE) as

$$\text{RASE} = \left\{ \frac{1}{n_{\text{grid}}} \sum_{j=1}^{n_{\text{grid}}} (\hat{f}(x_j) - f(x_j))^2 \right\}^{1/2},$$

where x_j are the grid points at which the density were computed, and $n_{\text{grid}} = 400$ here and throughout this paper. The RASEs are 11×10^{-4}, 9.12×10^{-4}, and 5.94×10^{-4} for $\alpha_n = 500$, 1000, and 5000, respectively. It is seen that the performance becomes better as α_n increases as it should. On the other hand, note that the RASEs are in the same order of magnitude but that the α_n's are very different. This suggests that the performance of the estimator is insensitive to the choice of α_n.

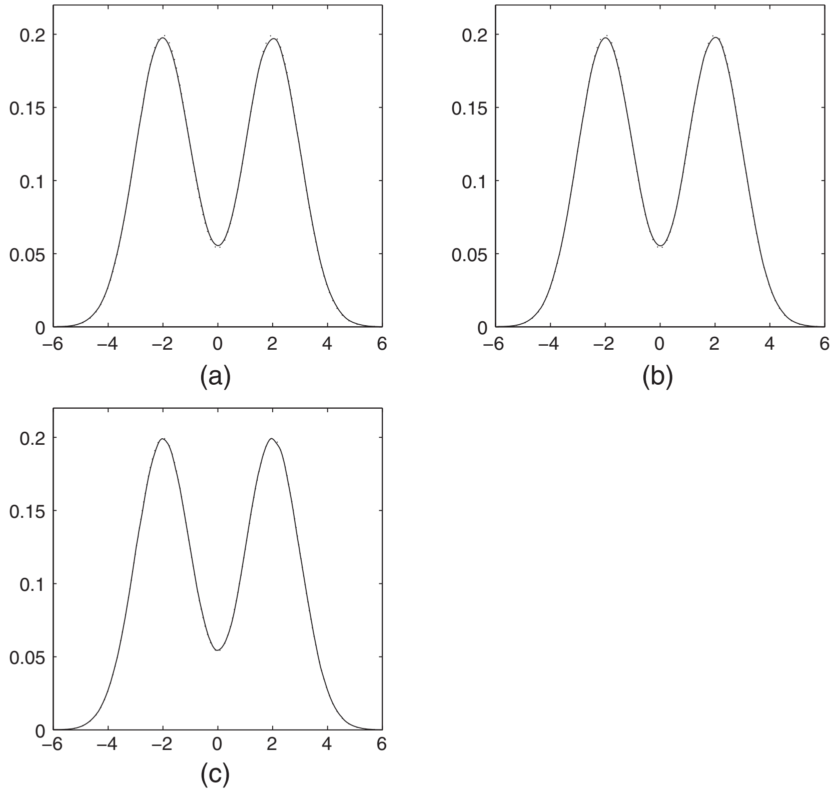

Figure 1. Plot of the estimated density curve of mixture normal. The solid curves are the estimated density curves, and the dotted curves are the true density curves. The estimated density curves using (a) $\alpha_n = 500$, (b) $\alpha_n = 1000$, and (c) $\alpha_n = 5000$.

Example 3

In this example, we compare the performance of the proposed approach and the kernel estimation based on various entire data sets. We will investigate how bandwidth selection affects the performance of the proposed estimation procedure. Data are generated from the normal mixture distribution

$$0.425N(0.35, 0.0144) + 0.425N(0.575, 0.0144) + 0.15N(0.8, 0.0009).$$

Chaudhuri and Marron [12] used this mixture normal distribution to illustrate their feature detection approach.

Figure 2 depicts the estimated density curves. The simulation results are summarized in Table III. In Table III, New/ROT refers to the newly proposed approach using the h_{rot} bandwidth, New/SJPI refers to the newly proposed approach using the Sheather–Jones plug-in (SJPI) [13] bandwidth selector, and Kernel refers to the kernel estimation using whole data sets. In this example, all simulations were conducted on a PC Pentium III 800 mHz. From Table III, the SJPI bandwidth selector

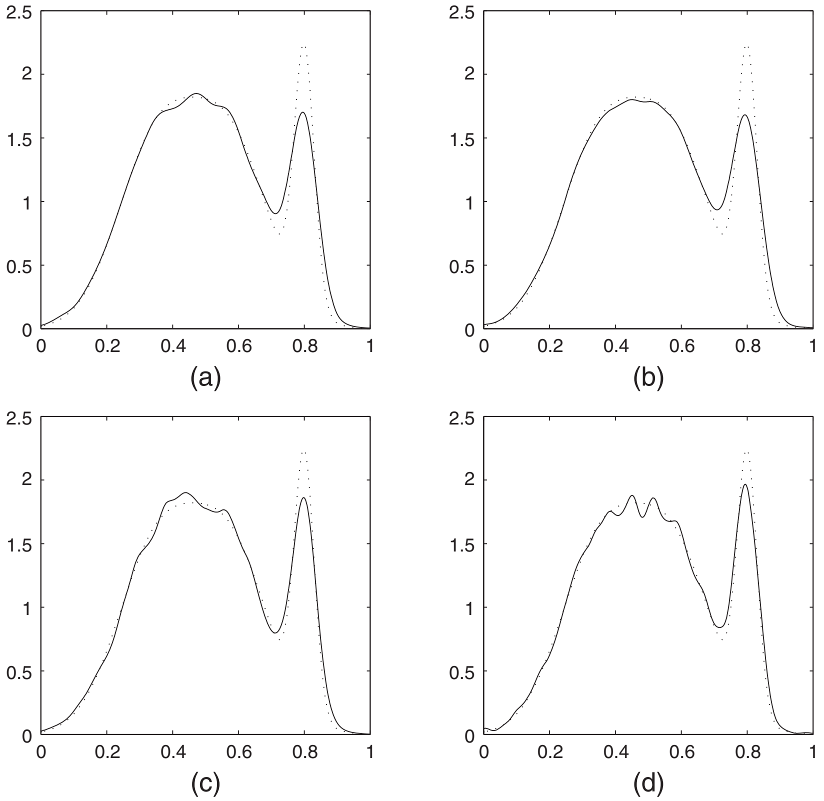

Figure 2. Plot of the estimated density curve of mixture normal in Example 2 when the sample size is equal to 10,000. The solid curves are the estimated density curves, and the dotted curves are the true density curves. ((a) and (b)) The bandwidth was selected using the rule of thumb; ((c) and (d)) The bandwidth was selected using the Sheather–Jones plug-in approach. ((a) and (c)) The proposed estimation procedure; ((b) and (d)) the kernel estimate using whole data sets at the same time.

Table III. Simulation results in Example 3.

Method	n (α_n)	RASE	Time (s)
New/ROT	10^4 (100)	0.1173	3.08
Kernel/ROT	10^4	0.1282	15.6
New/SJPI	10^4 (100)	0.0838	11.10
Kernel/SJPI	10^4	0.0714	0.86
New/SJPI	10^5 (1000)	0.0207	17.5
Kernel/SJPI	10^5	0.0247	3.9

RASE, root average squared error; ROT, rule of thumb; SJPI, Sheather–Jones plug-in.

may reduce the RASEs much, compared with the results of the ROT, but it needs more time for computing an SJPI bandwidth in the newly proposed estimation procedure. Note that the new procedure with the SJPI bandwidth selector requires to select a bandwidth by using the plug-in bandwidth selector proposed by Sheather and Jones [13] for each block. Thus, it requires to calculate the bandwidth β_n times. This is the reason why New/SJPI needs much more computing time than Kernel/SJPI, which requires to estimate the bandwidth once for the whole data set. From Table III, it can be seen that the newly proposed estimate is as efficient as the kernel estimate using the whole data set. This implies that our approach does not lose efficiency and is easily implemented in various computing environments.

5. Internet traffic data

Internet traffic data are exciting because they measure an intricate, fast-growing network connecting the world and transforming culture, politics, and business. A deep understanding of Internet traffic can contribute substantially to network performance monitoring, equipment planning, quality of service, security, and the engineering of Internet communications technology. Each Internet communication consists of a transfer of information from one computer to another; examples are the downloading of a Web page and the sending of an email message. When a file is transferred, it is not sent across the Internet as a continuous block of bits. Rather, the file is broken up into pieces, and each piece is sent individually. A detailed description of Internet traffic data can be found in [6]. As mentioned in [8], the success of analyzing Internet traffic data depends heavily on an ability to analyze the traffic Very Large Database (VLDB) in great details. One needs to explore the raw data in its full complexity; relying only on summaries is inadequate. They further commented that 'As for most database, visualization tools are vital for analyzing a VLDB'.

The original data file includes three fields: (i) time of the packet (in seconds), (ii) direction of the packet, and (iii) size of the packet. The variable under study is throughput, defined by (size of packet in bytes)/(time between two packets).

This data set consists of 8.1 million nonzero throughputs (packet size per second). We took $\alpha_n = 8000$ (approximately equal to $\sqrt{n}\log\log(n)$). First, we estimated the various percentiles of the population. The results are listed in Table IV. We also estimated the density of the population, which is shown in Figure 3. From Table IV, it can be seen that the standard error for the sample median is larger than that for the first and third quartiles as the value of density at the median is smaller than that of the first and third quartiles (see Figure 3 for details). Figure 3 shows that there are three typical values of throughput, one is close to 0 and the other two have a large size of throughput. If we multiply the first, second, and third quartiles in Table IV by eight, the bits per second throughput becomes 1.8, 5, and 8.3 mbps, respectively. Such a finding, among others, is currently discussed by researchers in Bell Labs for its potential implications.

6. Discussion and conclusion

In this paper, we have proposed an estimation procedure for a parameter $\theta(F)$ based on a massive data set. The proposed procedure significantly reduces the required amount of computing memory without loss of efficiency in many situations. It is readily applicable to both point estimation and density estimation. Asymptotic properties of the resulting estimators have been studied, and the asymptotic normality has been established. A standard error formula for the resulting estimate

Table IV. Estimated quartiles of Internet traffic data.

p	$\hat{\pi}_p(10^6)$	$\widehat{\text{SE}}(10^3)$
0.01	0.0015	0.1308
0.05	0.0219	1.3730
0.15	0.1120	4.2115
0.25	0.2372	7.2303
0.35	0.3836	9.5022
0.45	0.5415	10.3241
0.50	0.6300	10.2400
0.55	0.7226	9.8707
0.65	0.9033	8.1293
0.75	1.0476	5.1797
0.85	1.1329	2.3094
0.95	1.1787	0.8707
0.99	1.1858	0.1689

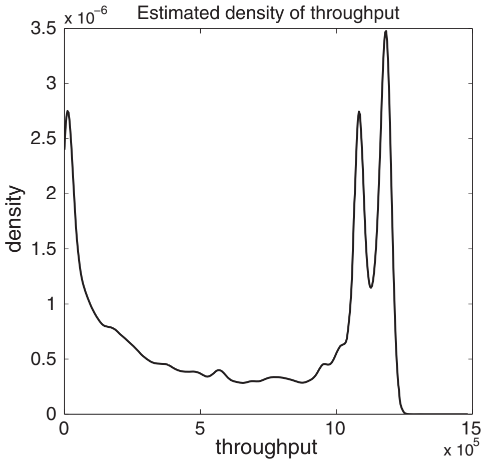

Figure 3. Plot of the estimated density curve of Internet traffic data.

has been proposed and empirically tested; thus, statistical inference for $\theta(F)$ can be performed. Simulation studies and an Internet data example have been used to illustrate the usefulness of the proposed approaches.

Future work will include statistical inference on massive data sets when records may be correlated. That is, when the observations $x_1, \ldots, x_n$ are m-dependent series (see, e.g., [14] for definition), the θ_{in} becomes 2-dependent series as θ_{in} only depends on $x_{i1}, \ldots, x_{i\alpha_n}$ and $\alpha_n \to \infty$, which implies that $\alpha_n \geqslant m$ eventually. In this situation, $\text{var}(\bar{\theta}) = (\beta_n \sigma_n^2 + (\beta_n - 1)\rho_n)/\beta_n^2$, where ρ_n equals to the correlation coefficient between $\hat{\theta}_{in}$ and $\hat{\theta}_{(i+1)n}$. Furthermore, the asymptotic normality of the resulting estimate $\bar{\theta}$ may be also established under some regularity conditions.

The idea of dividing a massive data set into several small pieces and combining the estimators derived from these pieces is also used in data privacy (e.g., [15]).

APPENDIX A.

Proof of Proposition 1:
Because $\hat{\theta}_{in}$ is independent and identically distributed,

$$E(\bar{\theta} - \theta)^2 = \beta_n^{-1}\sigma_n^2 + \frac{(\beta_n - 1)}{2\beta_n} b_n^2 \to 0$$

by the assumption. This completes the proof of Proposition 1. □

Proof of Theorem 1:
If Condition (a) holds, then the μ_n and σ_n do not depend on n. Note that $\hat{\theta}_{in}$ is independent and identically distributed with finite variance σ^2 and that it does not depend on n. With the use of the central limit theorem, asymptotic normality holds.

When $\alpha_n \to \infty$ as $n \to \infty$, to establish the asymptotic normality for $\bar{\theta}$, it is sufficient to show that Liapounov's condition holds, because $\hat{\theta}_{1,n}, \ldots, \hat{\theta}_{\beta_n,n}$ are independent and identically distributed. Note that

$$\frac{\sum_{i=1}^{\beta_n} \mathrm{E}|\hat{\theta}_{in} - \mu_n|^{2+\delta}}{\left(\sum_{i=1}^{\beta_n} \sigma_n^2\right)^{(2+\delta)/2}} = \frac{1}{\beta_n^{\delta/2}} \mathrm{E}\left|\frac{\hat{\theta}_{1,n} - \mu_n}{\sigma_n}\right|^{2+\delta},$$

which tends to 0 as $n \to \infty$ as $\beta_n \to \infty$ and (2) holds. Thus, Liapounov's condition holds. Therefore,

$$\sqrt{\beta_n}\left(\frac{\bar{\theta} - \mu_n}{\sigma_n}\right) \to N(0, 1).$$

in distribution as $n \to \infty$. □

Acknowledgements

The authors would like to thank Dr D. X. Sun for providing the Internet data set and Dr H. McGrath for the valuable comments and suggestions that greatly improved the presentation of this paper. The MATLAB codes used here to run the proposed method are available from the authors.

References

1. Hand DJ, Blunt G, Kelly MG, Adams NM. Data mining for fun and profit. *Statistical Sciences* 2000; **15**:111–131.
2. Rousseeuw PJ, Bassett Jr GW. The remedian: a robust averaging method for larger data sets. *Journal of the American Statistical Association* 1990; **85**:97–104.
3. Hurley C, Modarres R. Low-storage quantile estimation. *Computational Statistics* 1995; **10**:311–325.
4. Chao MT, Lin GD. The asymptotic distributions of the remedians. *Journal of Statistical Planning and Inference* 1993; **37**:1–11.
5. Mahmoud HM. *Sorting: A Distribution Theory*. Wiley InterScience: New York, 2000.
6. Wegman Edward J, Marchette David J. On some techniques for streaming data: a case study of Internet packet headers. *Journal of Computational and Graphical Statistics* 2003; **12**(4):893–914.
7. Ferguson TS. *A Course in Large Sample Theory*. Chapman & Hall: New York, 1996.
8. Cleveland WS, Sun DX. Internet traffic data. *Journal of the American Statistical Association* 2000; **95**:979–985.
9. Mcdermott J, Lin Dennis K J. Quantile contours and multivariate density estimation for massive datasets via sequential convex hull peeling. *IIE Transactions in Quality and Reliability* 2007; **39**:581–591.
10. Wand MP, Jones MC. *Kernel Smoothing*. Chapman & Hall: London, 1995.
11. Silverman BW. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall: London, 1986.
12. Chaudhuri P, Marron JS. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* 1999; **94**:807–822.
13. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B* 1991; **53**:683–690.
14. Brockwell PJ, Davis RA. *Time Series: Theory and Methods*, 2nd ed. Springer-Verlag: New York, 1991.
15. Dwork C, Smith A. Differential privacy for statistics: what we know and what we want to learn. *Journal of Privacy and Confidentiality* 2009; **1**:135–154.