Statistical Inference on Large Data Sets

Runze Li[a], Dennis K. J. Lin[b] and Bing Li[a]

[a] Department of Statistics, The Pennsylvania State University

University Park, PA 16802-2111

[b]Department of Management Science and Information Systems

The Pennsylvania State University, University Park, PA 16802-1913

Abstract

Analysis of large data sets is challenging due to limitations of computer primary memory. In this paper, we propose an approach to estimating population parameters from a large data set. The proposed approach significantly reduces the required amount of primary memory, and the resulting estimate will be as efficient as if a system would have allowed to analyze the entire data set simultaneously. Asymptotic properties of the resulting estimate are studied, and 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.

Key Words and Phrases: Internet traffic data, kernel density estimation, remedian.

Statistical Inference on Large Data Sets

Abstract

Analysis of large data sets is challenging due to limitations of computer primary memory. In this paper, we propose an approach to estimating population parameters from a large data set. The proposed approach significantly reduces the required amount of primary memory, and the resulting estimate will be as efficient as if a system would have allowed to analyze the entire data set simultaneously. Asymptotic properties of the resulting estimate are studied, and 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.

Key Words and Phrases: Internet traffic data, kernel density estimation, remedian.

1 Introduction

In the past decade, we have witnessed a revolution in information technology. As a consequence, routine collection of systematically generated data is now 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 Hand, Blunt, Kelly and Adams (2000). It becomes very challenging to extract useful features from a large 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, large 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 large data sets include: data processing, classification, detection of abnormal patterns, summarization, visualization, association/correlation analysis.

To obtain a summarization and preliminary analysis of a large data set, some basic statistics are of general interest. For example, to construct a box-plot for a large data set, we need the sample quartiles. These are not a trivial task on a large data set. Consider the problem of percentile estimation. Suppose, given independent observations $x_1, x_2, \cdots, x_n$ from an unknown distribution function F, we want to find its 100α 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 $x_1, x_2, \cdots, x_n$ becomes impossible. To overcome the difficulty, Rousseeuq and Bassett (1990) proposed the remedian method and Hurley and Nodarres (1995) proposed low-storage quantile estimation method. Chao and Lin (1993) studied the asymptotic behaviors of the remedian approach and found that the resulting estimator does not possess asymptotic normality. This arises difficulties in drawing statistical inference.

The memory space in some computing environments can be as large as several terabyte. However, the number of observations that can be stored in primary memory is often restricted. The available memory, though large, is finite. Many computing environments also limit the maximum array size allowed and this can be much smaller and even independent of the available memory. The large data sets present some obvious difficulties, such as complex relationships among variables, and a large physical memory requirement. In general, a simple job for a small data set may be of major difficulty for large data sets.

This work was motivated from analyzing several large real-life data sets. One of them is an *internet traffic data set*, which will be analyzed in Section 5 below. Internet engineering and management depend on an understanding of the characteristics of network traffic. 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. Cleveland and Sun (2000) provided some useful concepts and developments on the topic of internet traffic data analysis.

Much work has been done in developing various statistical models for internet data. In this paper, we will focus on developing computational and inferential tools for large data, including the internet traffic data and electronic commerce data. When a data set is too large to be stored in primary memory, many statistics are difficult to compute by standard algorithms or statistical packages. Indeed, it is very challenging to analyze a large data set in order to extract useful features from the data set. In this paper, we will propose a general approach to deal with statistical inference on data sets with large sample size. Our approach is widely applicable and capable of making statistical inference for any parameter $\theta(F)$ of a population F. It will be shown that under some mild conditions, the resulting estimator is strongly consistent and has a normal limiting distribution. Furthermore, we will show that in many situations the resulting estimate is as efficient as if all data were simultaneously used to compute the estimate. Our method can also be used for obtaining point estimation as well as density estimation.

The paper is organized as follows. Section 2 gives the basic idea of the proposed method and the theoretical justifications. Section 3 discusses the problem of point estimation from the large 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 the simplicity of presentation, proofs are given in the Appendix.

2 The Proposed Estimation Procedure

It is a challenge task to compute some basic statistics, such as sample quartiles, from a large data set because of the limitation of primary memory and the maximum array size in computing environments. In this section, we propose an estimation procedure that overcomes the memory difficulties.

To estimate a parameter $\theta(F)$ of a population F, such as quantitles or the density of the population, it is frequently required to store entire data set in primary memory in order to obtain an efficient estimate. For example, calculation of the quartiles requires the data first to be sorted and counted. One way to overcome the mentioned difficulty in memory is to use the 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 based on the whole data set rather than its parts, which is not feasible for large data sets. Intuitively we may sequentially read the data and store in primary memory block by block, and analyze the data in each block separately. As long as the size of block is small, one can easily implement this estimation procedure within each block under various computing environments. A question arises here is 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, such as 10,000,000, standard algorithms for computing sample median may exceed the available memory and fail. However, it is easy to compute a sample median of 10,000 samples in many statistical packages, such as S-plus and SAS. We may sequentially read in the data block by block, each having, says, 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 1,000 sample medians. In summary, to estimate a parameter $\theta(F)$ based on a large data set, we may employ a two-stage procedure: first read in the whole data set sequentially block by block, each having manageable sample size, and compute an estimate of $\theta(F)$ within each block.

Then take the average of the resulting estimates obtained from each block as a estimate for $\theta(F)$. Note that the second stage can be updated as soon as new block is processed by the first stage, and hence does not require additional memory.

2.1 Sampling Properties

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

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

where $x_{ij} = x_{(i-1)*\alpha_n + j}$, for $i = 1, \cdots, \alpha_n$ and $j = 1, \cdots, \beta_n$, 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. Using the same estimator for each block, and denoting by $\widehat{\theta}_{in}$ the resulting estimate based on the sub-sample in the i-th block $x_{i1}, \cdots, x_{i\alpha_n}$. We estimate $\theta(F)$ by averaging of $\widehat{\theta}_{in}$'s, i.e.,

$$\bar{\theta} = \frac{1}{\beta_n} \sum_{i=1}^{\beta_n} \widehat{\theta}_{in}. \tag{2.1}$$

Now we investigate the sampling properties of the estimate $\bar{\theta}$. It is easy to show that the following propositions hold.

Proposition 1. For any positive integer values of α_n and β_n,

(a) if $\widehat{\theta}_{in}$ is affine equivariant, then so is $\bar{\theta}$;

(b) if $\widehat{\theta}_{in}$ is an unbiased estimator of θ, the so is $\bar{\theta}$.

Proposition 2. Suppose that $x_1, \cdots, x_n$ are independent and identically distributed, and $\alpha_n \to \infty$ and $\beta_n \to \infty$ as $n \to \infty$.

(a) If $\widehat{\theta}_{in}$'s converge weakly to the true value θ, then so does $\bar{\theta}$.

(b) If $\widehat{\theta}_{in}$'s converge in L_2 to the true value θ, then so does $\bar{\theta}$.

(c) If $\widehat{\theta}_{in}$'s converge strongly to the true value θ, then so does $\bar{\theta}$.

Denote $\mu_n = \mathrm{E}(\widehat{\theta}_{in})$ and $\sigma_n^2 = \mathrm{var}(\widehat{\theta}_{in})$. To establish the asymptotic normality of $\bar{\theta}$, we need 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

$$\frac{E|\widehat{\theta}_{in} - \mu_n|^{2+\delta}}{\beta_n^{\delta/2}\sigma_n^{2+\delta}} \to 0 \tag{2.2}$$

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

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

$$\sqrt{\beta_n}\left(\frac{\bar{\theta} - \mu_n}{\sigma_n}\right) \to N(0,1) \tag{2.3}$$

in distribution as $n \to \infty$.

The proof of Theorem 1 is given in the Appendix.

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 $\widehat{\theta}_{in}$'s are unbiased estimators of θ. If $\widehat{\theta}_{in}$ is a biased estimator, the resulting estimator is not consistent, because the bias $\mu - \theta$ is a constant.

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

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

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

2.2 Choice of α_n

When $\alpha_n \to \infty$, it can be shown that

$$\sqrt{\beta_n}\left(\frac{\bar{\theta} - \theta}{\sigma_n}\right) \to N(0,1) \tag{2.4}$$

holds if and only if $\mu_n - \theta = o(\sigma_n/\sqrt{\beta_n})$, In most cases, $\sigma_n = O(1/\sqrt{\alpha_n})$. This implies that if

$$\mu_n - \theta = o(1/\sqrt{n}) \tag{2.5}$$

holds, then (2.4) holds. If $\widehat{\theta}_{in}$ is unbiased, then $\mu_n - \theta = 0$, and hence (2.4) 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. In practice, we suggest to take $\alpha_n = O(\sqrt{n}\log\log(n))$.

3 Statistical Inference

In this section we investigate statistical inference 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

As proposed in the last section, an estimator for parameter θ is

$$\bar{\theta} = \frac{1}{\beta_n}\sum_{i=1}^{\beta_n} \widehat{\theta}_{in}. \tag{3.1}$$

To draw statistical inference on θ, we have to know its estimator variation with finite samples. In fact $\widehat{\theta}_{1n}, \cdots, \widehat{\theta}_{\beta_n n}$ 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 $\widehat{\theta}_{1n}, \cdots, \widehat{\theta}_{\beta_n n}$. Namely

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

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

$$\widehat{\mathrm{SE}}(\bar{\theta}) = \frac{\widehat{\sigma}_n}{\sqrt{\beta_n}}. \tag{3.2}$$

For some parameters, such as percentiles, their standard error depends on the unknown population. However, the estimated standard error formula (3.2) allows us to avoid estimating the unknown population. This formula will be tested in our simulation example.

If the asymptotic normality (2.4) holds, then a $100(1-\alpha)\%$ confidence interval is approximately $\bar{\theta} \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}_n/\sqrt{\beta_n}$, where $\Phi^{-1}(1-\alpha/2)$ is the $100(1-\alpha)\%$ percentile of the standard normal distribution.

Similarly, for testing the hypothesis:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0. \tag{3.3}$$

The test statistic is given by

$$T = \frac{\bar{\theta} - \theta_0}{\widehat{\sigma}_n / \sqrt{\beta_n}} \tag{3.4}$$

and the rejection region is $\{|T| > \Phi^{-1}(1 - \alpha/2)\}$ with significance level α.

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 literature. See, for example, Rousseeuw and Bassett (1990), Chao and Lin (1993) and Hurley and Modarres (1995). Compared with their approaches, the proposed approach does not reduce the amount of storage so drastically. On the other hand, the proposed approach can be used to estimate a general population parameter, including median 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 codes.

Example 3.1 In this example, we generated 1000 data sets, each consisting of 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 1. In Table 1, the $\widehat{\mathrm{SE}}$ and std($\widehat{\mathrm{SE}}$)are the average and sample standard deviation of the 1000 estimated standard errors defined in (3.2), respectively, and the standard deviation of the 1000 estimated percentiles, regarded as the true standard deviation of the corresponding estimators and denoted by $\mathrm{SE}_{\mathrm{true}}$, is also displayed in Table 1. Comparing the last two columns in Table 1, it can be found that the estimated standard error formula works surprising well. In order to obtain the standard error of estimated percentiles, the proposed standard error formula allows us avoiding to estimate the density of population. This is different from the traditional approaches, which require estimation of some parameters depending on the unknown density of population.

Table 1: Estimated percentiles ($\alpha_n = 8000$)

p	True value	Estimate	$\text{SE}_{\text{true}}(10^{-4})$	$\widehat{\text{SE}}\,(\text{std}(\widehat{\text{SE}}))\,(10^{-4})$
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)

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 32000. The results are described in Table 2. Comparing the results based on the three different choices of α_n, it can be seen from Tables 1 and 2 that the choice of α_n is insensitive to the results, although the results using $\alpha_n = 8000$ and 32000 seem to work slightly better than those using $\alpha_n = 2000$.

Table 2: Comparison between different choices of α_n

	$\alpha_n = 2000$			$\alpha_n = 32000$		
Quartiles	$\widehat{q}_i$	$\widehat{\text{SE}}$	SE_{true}	$\widehat{q}_i$	$\widehat{\text{SE}}$	SE_{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}

3.3 Estimation of population mean and variance

Estimation of population mean and variance does not require large-storage in primary memory. In general, however, it is of interest to compare the resulting estimators of the proposed approach with the traditional ones. It is easy to show that the proposed estimator for population mean indeed equals to the sample mean when we use the sample mean as an estimator for the population mean for each sub-sample $x_{i1}, \cdots, x_{i\alpha_n}$. When we use the sample variance to estimate the population variance for each sub-sample, the resulting estimator for population variance is

$$\bar{\sigma}^2 = \frac{1}{\beta_n(\alpha_n - 1)} \sum_{ij} (x_{ij} - \bar{x}_{i\cdot})^2,$$

where $\bar{x}_{i\cdot} = \frac{1}{\alpha_n} \sum_j x_{ij}$. This is the mean squared error in one-way ANOVA model. Although it is an unbiased estimator, the resulting estimator will lose $(\beta_n - 1)$ degrees of freedom. This implies that it is not as efficient as the traditional one, the sample variance based on all samples:

$$s_n^2 = \frac{1}{n-1} \sum_{ij} (x_{ij} - \bar{x})^2.$$

In fact, under the normality assumption, it follows that

$$\begin{aligned} \mathrm{var}(\bar{\sigma}^2) &= \frac{2\sigma^4}{n - \beta_n}, \\ \mathrm{var}(s_n^2) &= \frac{2\sigma^4}{n-1}. \end{aligned}$$

As such $\mathrm{var}(\bar{\sigma}^2)$ reaches its minimum at $\beta_n = 1$. However, if we take $\alpha_n = O(\sqrt{n} \log\log(n))$, the loss is very small in terms of efficiency because $\beta_n/n = O(\alpha_n^{-1}) \to 0$ rapidly.

4 Nonparametric kernel density estimation

The proposed procedure in Section 2 can be directly applied to the context of estimation of density. Here we use nonparametric kernel smoothing, although other approaches, such as smoothing spline (See Wahba, 1990), are also applicable. Using the subsample $x_{i1}, \cdots, x_{i\alpha_n}$ in the i-th block, a kernel density estimator is as follows:

$$\widehat{f}_h(x) = \frac{1}{\alpha_n} \sum_{j=1}^{\alpha_n} K_h(x_{ij} - x), \tag{4.2}$$

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 estimated density curve. The choice of kernel function is not crucial, but the choice of bandwidth h is critical. It is well-known that

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

and

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

where $\mu_2(K) = \int z^2 K(z)\,dx$ and $R(K) = \int K^2(z)\,dz$, (See, for example, Wand and Jones (1995) page 20-21). Thus

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

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

$$\text{MISE} = \text{AMISE}\{\widehat{f}_h(\cdot)\} + o\{h^4 + (nh)^{-1}\}$$

where

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

Thus the optimal bandwidth minimizing the AMISE is

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

which does not dependent on α_n. This also implies that the estimated density curve in the first step should be undersmoothed in order to reduce bias, compared to the bandwidth $h = O(\alpha_n^{-1/5})$, because the first multiplier in the h_{opt} only depends on the kernel function and 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 using the data $x_{i,1}, \cdots, x_{i,\alpha_n}$, under some criterion. We may take

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

as our bandwidth. Using this bandwidth, the resulting density estimation will be as good as if we used entire samples simultaneously, at least from a theoretical point of view.

Example 4.1. In this example, 1 million independent and identically distributed random sample 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 Silverman (1986). That is, the bandwidth used to estimate the density is

$$h_{\text{rot}} = 0.9 \times 1.06 \times \sigma \times n^{-1/5},$$

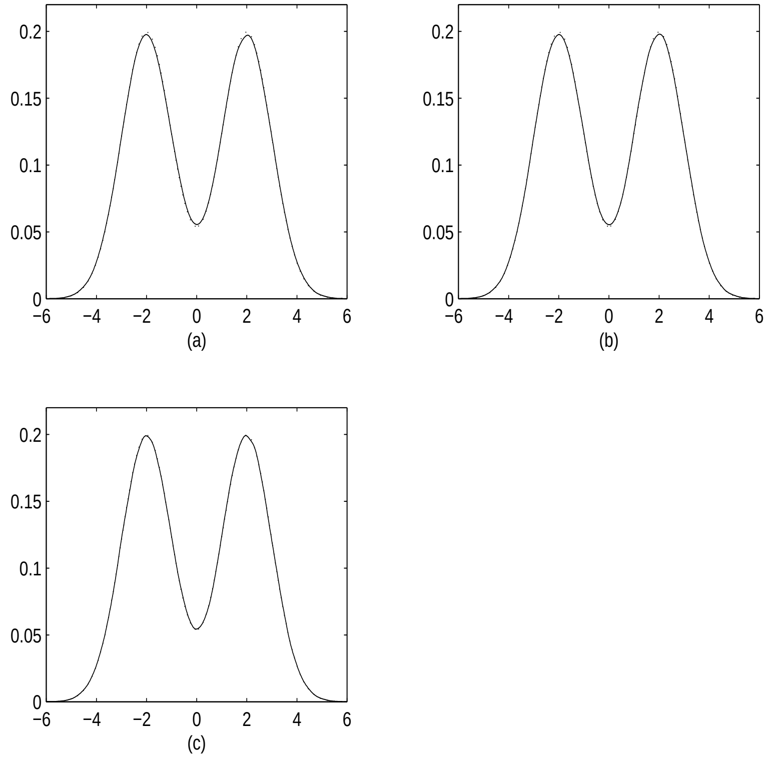

Figure 1: *Plot of estimated density curve of mixture normal. The solid curves are estimated density curve and dotted curve is the true density curve. (a) is the estimated density curve using* $\alpha_n = 500$*, (b) corresponds to* $\alpha_n = 1000$ *and (c) is corresponds to* $\alpha_n = 5000$.

where σ is the population standard deviation. The factor 1.06 is due to the Gaussian kernel, and multiplying by the factor 0.9 is done to adjust for oversmoothing as noted in Silverman (1986). In our simulation, the σ is substituted by its robust estimate, the mean of absolute deviation (MAD) of the sub-sample $x_{i1}, \cdots, x_{i\alpha_n}$ 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, we define the Root of Average of Squared Errors (RASE) as

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

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 the α_n's are very different. This suggests that the performance of the estimator is not very sensitive to the choice of α_n.

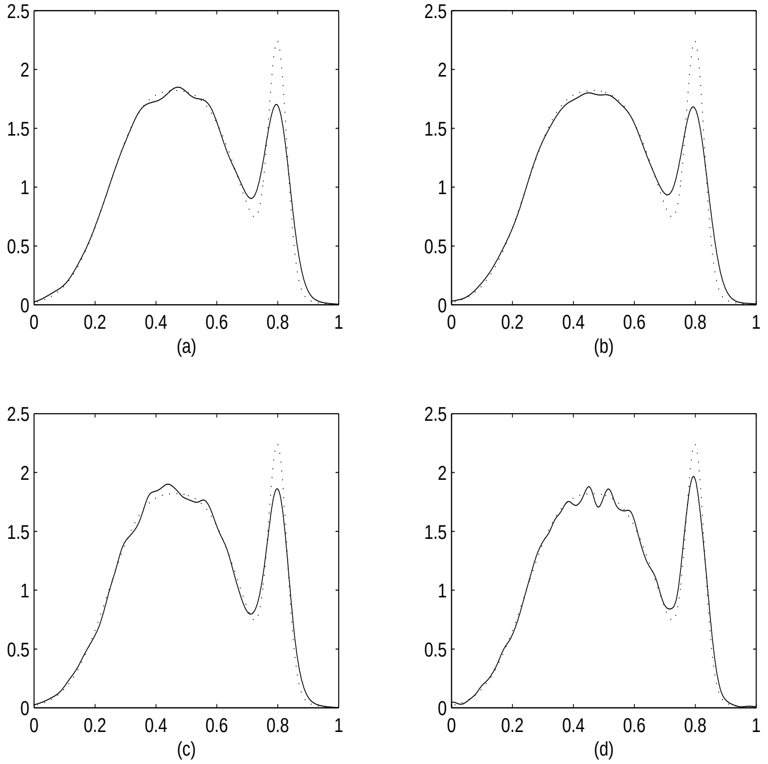

Figure 2: *Plot of estimated density curve of mixture normal in Example 4.2 when the sample size equals to 10000. The solid curves are estimated density curves and dotted curves are the true density curve. In (a) and (b), the bandwidth was selected by the rule of thumb, while in (c) and (d), the bandwidth was selected by the SJPI approach. (a) and (c) are for the proposed estimation procedure, and (b) and (d) are for kernel estimate using whole data sets at the same time.*

Example 4.2 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. The data is 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 (1999) used this mixture normal distribution to illustrate their feature detection approach.

Figures 2 depicts the estimated density curves. The simulation results are summarized in Table 3. In Table 3, New/ROT stands for the newly proposed approach using the h_{rot} bandwidth, New/SJPI for the newly proposed approach using Sheather and Jones (1991) plug-in (SJPI) bandwidth selector, and Kernel for kernel estimation using whole data sets. In this example, all simulations were conducted on a PC Pentium III 800 mHz. From Table 3, SJPI bandwidth selector may reduce the RASEs a lot, compared with the results of the rule of thumb, but it needs more time for computing a SJPI bandwidth in the newly proposed estimation procedure. From Table 3, 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.

Table 3: Simulation Results in Example 4.2

Method	n (α_n)	RASE	Time (Seconds)
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

5 Internet Traffic Data

In this section, we analyze the internet traffic data mentioned in Section 1. The original data file includes three fields: (1) time of the packet (in second), (2) direction of the packet, and (3) 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 4. We also estimate the density of the population, which is shown in Figure 3. From Table 4, it can be seen that the standard error for the sample median is larger than those for the first and third quartiles as the value of density at the median is smaller than those of the first and third quartiles. See Figure 3 for details. Figure 3 shows that

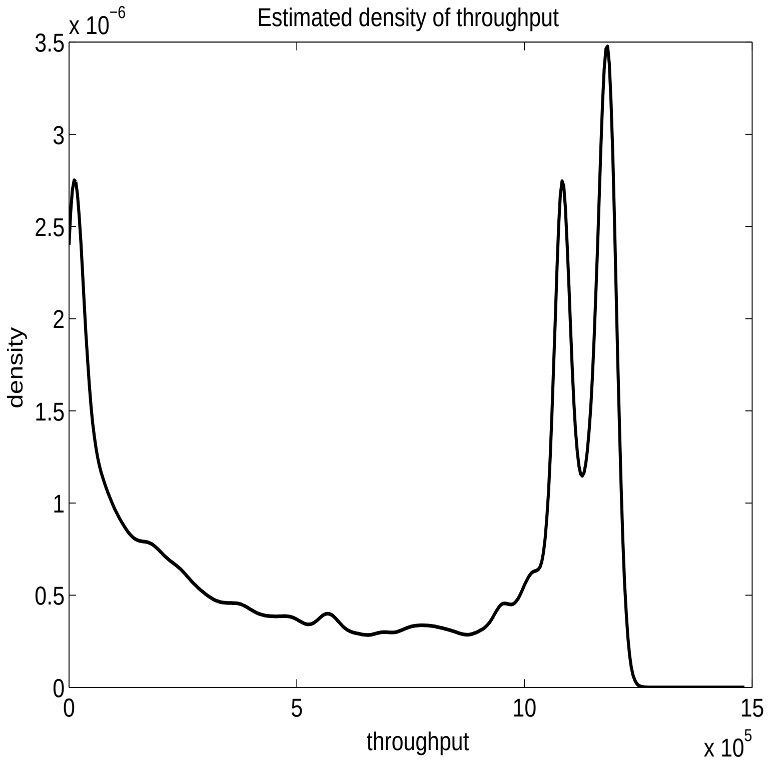

Figure 3: *Plot of estimated density curve of internet traffic data*

Table 4: Estimated Quartiles of Internet Traffic Data

p	$\widehat{\pi}_p(10^6)$	$\widehat{\mathrm{SE}}(10^3)$
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

there are 3 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 4 by 8, the bits per second throughput become 1.8 mbps (mega bits per second), 5 mbps and 8.3 mbps, respectively.

6 Discussion and Conclusion

In this paper, we have proposed an estimation procedure for a parameter $\theta(F)$ based on a large 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 and density estimation. Asymptotic properties of the resulting estimators have been studied and asymptotic normality has been established. A standard error formula for the resulting estimate 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 large data sets when records may be collected. Namely, when the observations $x_1, \cdots, x_n$ are m-dependent series (see, for example, Brockwell and Davis (1991) for definition), the θ_{in} becomes 2-dependent series as θ_{in} only depends on $x_{i1}, \cdots, x_{i\alpha_n}$ and $\alpha_n \to \infty$, which implies that $\alpha_n \geq m$ 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 $\widehat{\theta}_{in}$ and $\widehat{\theta}_{(i+1)n}$. Furthermore asymptotic normality of the resulting estimate $\bar{\theta}$ may be also established under some regularity conditions.

Acknowledgement: The authors would like to thank Dr. D. X. Sun for providing them the internet data set, and Dr. H. McGrath for their valuable comments and suggestions which lead to greatly improve presentation of this paper.

Appendix: Proof

Proof of Theorem 1: If Condition (a) holds, then the μ_n and σ_n do not depend on n. Note that the $\widehat{\theta}_{in}$ are independent and identically distributed with finite variance σ^2, and do not dependent on n. By 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, since the $\widehat{\theta}_{1,n}, \cdots, \widehat{\theta}_{\beta_n,n}$ are independent and identically distributed. Note that

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

which tends 0 as $n \to \infty$ as $\beta_n \to \infty$ and (2.2) holds. Thus the Liapounov condition holds. Therefore

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

in distribution as $n \to \infty$.

References

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods (2nd Edition)*, Springer-Verlag, New York.

Chao, M. T. and Lin, G. D. (1993). The asymptotic distributions of the remedians. *Journal of Statistical Planning and Inference*, **37**, 1-11.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves, *J. Amer. Statist. Assoc.*, **94**, 807-822.

Cleveland, W. S., Lin, D., and Sun, D. X. (2000). Network simulation: modeling the nonstationary and long-range dependence of client TCP connection start times under HTTP. in *Proceedings of ACM SIGMETRICS'00.* To appear.

Cleveland, W. S. and Sun, D. X. (2000). Internet traffic data. *Journal of the American Statistical Association*, **95**, 979-985.

Hand, D.J., Blunt, G., Kelly, M. G. and Adams, N. M. (2000). Data Mining for Fun and Profit. *Statistical Sciences*, **15**, 111-131.

Hurley, C. and Modarres, R. (1995). Low-storage quantile estimation. *Computational Statistics*, **10**, 311-325.

Rousseeuw, P. J. and Bassett, G. W., Jr. (1990). The remedian: A robust averaging method for larger data sets. *Journal of the American Statistical Association*, **85**, 97-104.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B.*, **53**, 683-690.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, Chapman & Hall, London.