Frontiers of Computer Science CN 10-1014/TP
https://doi.org/10.1007/s11704-023-2356-x ISSN 2095-2228
RESEARCH ARTICLE

Density estimation-based method to determine sample size for random sample partition of big data

Yulin HE[1,2], Jiaqi CHEN[2,1], Jiaxing SHEN[3], Philippe FOURNIER-VIGER[2], Joshua Zhexue HUANG[1,2]

1 Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518107, China
2 College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
3 Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, New Territories 999077, Hong Kong, China

Front. Comput. Sci., **Just Accepted Manuscript** • 10.1007/s11704-023-2356-x
https://journal.hep.com.cn on April 28, 2023

Front.Comput.Sci.,
https://doi.org/10.1007/s11704-023-2356-x

RESEARCH ARTICLE

Density estimation-based method to determine sample size for random sample partition of big data

Yulin HE[1,2], Jiaqi CHEN[2,1], Jiaxing SHEN[3], Philippe FOURNIER-VIGER[2] and Joshua Zhexue HUANG(✉)[1,2]

1 Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518107, China
2 College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
3 Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, New Territories 999077, Hong Kong, China

Abstract Random sample partition (RSP) is a newly developed big data representation and management model to deal with big data approximate computation problems. Academic research and practical applications have confirmed that RSP is an efficient solution for big data processing and analysis. However, a challenge for implementing RSP is determining an appropriate sample size for RSP data blocks. While a large sample size increases the burden of big data computation, a small size will lead to insufficient distribution information for RSP data blocks. To address this problem, this paper presents a novel density estimation-based method (DEM) to determine the optimal sample size for RSP data blocks. First, a theoretical sample size is calculated based on the multivariate Dvoretzky-Kiefer-Wolfowitz (DKW) inequality by using the fixed-point iteration (FPI) method. Second, a practical sample size is determined by minimizing the validation error of a kernel density estimator (KDE) constructed on RSP data blocks for an increasing sample size. Finally, a series of persuasive experiments are conducted to validate the feasibility, rationality, and effectiveness of DEM. Experimental results show that (1) the iteration function of the FPI method is convergent for calculating the theoretical sample size from the multivariate DKW inequality; (2) the KDE constructed on RSP data blocks with sample size determined by DEM can yield a good approximation of the probability density function (*p.d.f.*); and (3) DEM provides more accurate sample sizes than the existing sample size determination methods from the perspective of *p.d.f.* estimation. This demonstrates that DEM is a viable approach to deal with the sample size determination problem for big data RSP implementation.

Keywords random sample partition, big data, sample size, Dvoretzky-Kiefer-Wolfowitz inequality, kernel density estimator, probability density function

1 Introduction

A popular algorithmic strategy to handle big data computation problems is divide-and-conquer [1]. According to that paradigm, the big data is first partitioned into several subsets. Then, each subset is processed, and local results from all subsets are combined to obtain the global results. *MapReduce* [2] is a widespread programming model to implement the divide-and-conquer paradigm to perform big data computation tasks using distributed file systems such as HDFS (Hadoop Distributed File System) [3]. The default size of a data subset or data block in HDFS is 64 MB or 128 MB, depending on the specific hardware support of the distributed computing environment. The HDFS data block size influences the efficiency for processing big data [4]. If the block size is too small, there will be too many HDFS data blocks and thus too much "metadata" will be stored. And if the block size is too large, the time to transfer data from the disk can be significantly longer than the time to seek the start of a block. Thus, it is very important to determine an appropriate block size for handling big data in a distributed computing environment.

An intrinsic issue to data partitioning when using HDFS data blocks to handle big data computation tasks is that the probability distribution of each HDFS data block is often inconsistent with that of the whole big data. This can be a problem especially

Received June 14, 2022; Accepted April 28, 2023

E-mail: zx.huang@szu.edu.cn

when analyzing big data containing numerical attribute values. To obtain data blocks that have a consistent distribution, a novel big data representation model was introduced in 2019, named random sample partition (RSP) [5]. According to that model, big data is partitioned into a series of RSP data blocks using sampling such that each block has a consistent probability distribution with the whole big data for a given significance level. RSP data blocks can be generated by transforming HDFS data blocks using a two-stage data processing algorithm [6].

The RSP model is receiving more and more attention from academia and the industry, as it solves a key problem of HDFS. However, the issue of determining an appropriate block size remains. From a statistical perspective, the key issue to select an adequate block size for RSP is to determine an effective sample size. Current sample size determination methods, e.g., the Slovin formula [7], population estimation method (PEM) [8], and population mean estimation method (PMEM) [9] have the following limitations for big data. They were designed for sampling problems with univariate random variables and the sample size is determined based on the assumption of a normal distribution. Moreover, these methods were proposed to deal with small and medium sized data sets. Given the aforementioned limitations, these methods cannot be directly used to handle big data since big data can be multivariate and have complex probability distributions.

To solve the problem of determining an optimal sample size for RSP data block, this paper proposes a novel density estimation-based method (DEM). The main contributions of this paper are summarized as follows.

- A theoretical sample size that is used to model the *extrinsic empirical distribution* is first calculated based on the multivariate Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [10] by utilizing the fixed-point iteration (FPI) method.
- On the basis of theoretical sample size, a practical sample size is then determined to characterize the *intrinsic probability density* by minimizing the validation error of the kernel density estimator (KDE) constructed for an RSP data block.
- Finally, extensive experiments are reported, which were conducted to validate the feasibility (i.e., convergence of the iteration function of the FPI method), rationality (i.e., the probability density function is well estimated), and effectiveness (i.e., the sample size is appropriate for estimating the probability density function) of the DEM. Results demonstrate that the DEM is a viable approach to determine the sample size for implementing the big data RSP model.

The remainder of this paper is organized as follows. Related work is reviewed in Section 2. Section 3 introduces the basic concepts of RSP and KDE. Section 4 presents the proposed DEM method to determine the sample size of RSP data blocks. Section 5 describes experiments and analyzes the results. Finally, Section 6 draws a conclusion and discusses future work.

2 Related works

Representative methods for sample size determination are briefly reviewed. The typical Slovin formula [7] calculates the sample size as

$$\mathcal{M}_{\text{Slo}} = \frac{\mathcal{N}}{1 + \mathcal{N}E^2}, \tag{1}$$

where $\mathcal{N}$ is the population size and E is a margin of error (MOE). The sample size determined by the Slovin formula strongly depends on the selection of the MOE. If $E = 0$, the sample size is the population size, and if $E \to 1$, the sample size is close to 1. The PEM [8] and PMEM [9] are based on the assumption that the population from which samples are drawn follows a normal distribution. They respectively determine sample sizes as

$$\mathcal{M}_{\text{PEM}} = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 P\left(1 - P\right)}{E^2} \tag{2}$$

and

$$\mathcal{M}_{\text{PMEM}} = \left(\frac{z_{\frac{\alpha}{2}} \times \sigma}{E}\right)^2, \tag{3}$$

where P is the sample proportion, σ is the population standard deviation, and $z_{\frac{\alpha}{2}}$ is the bilateral quantile of standard normal distribution corresponding to the significance level α. These two methods are only suitable to determine sample sizes for univariate random variables.

Kleiner et al. proposed the bag of little bootstraps (BLB) method [11] to provide a robust and efficient mean of assessing the quality of estimators, where each BLB resample contains at most

$$\mathcal{M}_{\text{BLB}} = \mathcal{N}^{\nu} \tag{4}$$

distinct sample points and $\nu \in [0.5, 1.0]$ is a scale factor. In fact, the BLB sample size was firstly used in Reshef et al.'s work [12], where a maximal information coefficient was designed to measure the dependence between two variables. Sengupta et al. [13] presented a subsampled double bootstrap method for massive data analysis. Similarly, the BLB sample size was used to fix the subset size for the design of a new resampling method.

Other sample size determination methods can be found in the literature [14–17], and are tailored to specific application scenarios.

3 Preliminaries

Let there be a $\mathcal{D}$-dimensional big data set

$$\begin{aligned}\mathbb{X} = \{\mathrm{x}_n \,|\mathrm{x}_n = (x_{n1}, x_{n2}, \cdots, x_{n\mathcal{D}}),\\ x_{nd} \in \mathfrak{R}, d = 1, 2, \cdots, \mathcal{D}, n = 1, 2, \cdots, \mathcal{N}\}\end{aligned} \tag{5}$$

having $\mathcal{N}$ distinct sample points of a random variable $\mathcal{X}$ with probability distribution function (PDF) $F(\mathrm{x})$ and probability density function (*p.d.f.*) $f(\mathrm{x})$. The random sample partition (RSP) and estimated *p.d.f.* $\hat{f}(\mathrm{x})$ are introduced as follows. Without loss of generality, we assume that each x_{nd} is a numerical attribute value because symbolic attribute values can be easily transformed into numerical attribute values using existing encoding techniques such as one-hot encoding [18] and deep encoding [19].

3.1 Random sample partition of big data

An RSP $\mathrm{T} = \{\mathbb{X}_1, \mathbb{X}_2, \cdots, \mathbb{X}_\mathcal{K}\}$ contains $\mathcal{K}$ RSP data blocks and is a sample point partition of a big data set $\mathbb{X}$, which satisfies the following conditions:

1) $\bigcup\limits_{k=1}^{\mathcal{K}} \mathbb{X}_k = \mathbb{X}$ and $\sum\limits_{k=1}^{\mathcal{K}} \mathcal{N}_k = \mathcal{N}$, where $|\mathbb{X}_k| = \mathcal{N}_k$;

2) for $\forall i \neq j \in \{1, 2, \cdots, \mathcal{K}\}$, $\mathbb{X}_i \cap \mathbb{X}_j = \emptyset$;

3) for a given significance level $\alpha \in (0, 1)$ and error threshold $\varepsilon \in (0, 1)$, the probability

$$\mathrm{P}(\sup_{\mathrm{x}\in\mathfrak{R}^\mathcal{D}} |\hat{F}_{\mathcal{N}_k}^{(k)}(\mathrm{x}) - F(\mathrm{x})| > \varepsilon) \leq \alpha \tag{6}$$

holds for the empirical distribution function (EDF) $\hat{F}_{\mathcal{N}_k}^{(k)}(\mathrm{x})$ estimated based on the RSP data block $\mathbb{X}_k$, $k = 1, 2, \cdots, \mathcal{K}$.

In fact, the big data set $\mathbb{X}$ is composed of $\mathcal{N}$ sample points corresponding to the simple random sample (SRS) $\mathcal{X}_1$, $\mathcal{X}_2$, $\cdots$, $\mathcal{X}_\mathcal{N}$ which are $\mathcal{N}$ independent and identically distributed random variables with PDF $F(\mathrm{x})$. An RSP data block

$$\begin{aligned}\mathbb{X}_k = \{\mathrm{x}_n^{(k)} \,|\mathrm{x}_n^{(k)} = (x_{n1}^{(k)}, x_{n2}^{(k)}, \cdots, x_{n\mathcal{D}}^{(k)}),\\ x_{nd}^{(k)} \in \mathfrak{R}, d = 1, 2, \cdots, \mathcal{D}, n = 1, 2, \cdots, \mathcal{N}_k\}\end{aligned} \tag{7}$$

is composed of $\mathcal{N}_k$ sample points that are sampled randomly from the big data set $\mathbb{X}$, where there exists a unique $n' \in \{1, 2, \cdots, \mathcal{N}\}$ such that $\mathrm{x}_n^{(k)} = \mathrm{x}_{n'}$. Because the EDF of an RSP data block can approximate the PDF of the big data for a given significance level and error threshold, combining the results obtained by processing RSP data blocks can be used to approximate the results for the whole big data. Extensive experiments [5] have demonstrated that the RSP model performs well for the approximate computation of big data.

3.2 Kernel density estimator

Because the true PDF and *p.d.f.* of a random variable $\mathcal{X}$ are unknown, the estimated *p.d.f.* can be constructed as

$$\begin{aligned}\hat{f}(\mathrm{x}) &= \frac{1}{\mathcal{N}\prod\limits_{d=1}^{\mathcal{D}} h_d^{(k)}} \sum_{n=1}^{\mathcal{N}} \mathrm{K}(\frac{x_1 - x_{n1}^{(k)}}{h_1^{(k)}}, \cdots, \frac{x_\mathcal{D} - x_{n\mathcal{D}}^{(k)}}{h_\mathcal{D}^{(k)}})\\ &= \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \prod_{d=1}^{\mathcal{D}} \frac{1}{\sqrt{2\pi}h_d^{(k)}} \exp[-\frac{1}{2}(\frac{x_d - x_{nd}^{(k)}}{h_d^{(k)}})^2]\end{aligned} \tag{8}$$

using a KDE [20, 21] based on the RSP data block $\mathbb{X}_k$, where

$$\mathrm{K}(\mathrm{u}) = \mathrm{K}(u_1, \cdots, u_\mathcal{D}) = \frac{1}{(\sqrt{2\pi})^\mathcal{D}} \prod_{d=1}^{\mathcal{D}} \exp(-\frac{1}{2}u_d^2) \tag{9}$$

is the $\mathcal{D}$-dimensional kernel function and $h_d^{(k)}$ is the kernel width or bandwidth parameter that is a function of $\mathcal{N}_k$ and satisfies the conditions

$$\begin{cases}\lim\limits_{\mathcal{N}_k\to+\infty} h_d^{(k)} = 0\\ \lim\limits_{\mathcal{N}_k\to+\infty} \mathcal{N}_k h_d^{(k)} = +\infty\end{cases}. \tag{10}$$

The estimation quality of the *p.d.f.* depends on the selection of the kernel width for the given kernel function. A small bandwidth results in an under-smoothed *p.d.f.* estimation, while a large bandwidth leads to an over-smoothed *p.d.f.* estimation.

4 Density estimation-based method for sample size determination

A challenge to apply the RSP model in a distributed environment is to select a suitable sample size for RSP data blocks. A large sample size will increase the burden of data computation, while a small sample size will increase that of data scheduling. A new density estimation-based method (DEM) is presented in this section to determine the optimal sample size for RSP data blocks, which includes the calculation of the theoretical sample size and the determination of the practical sample size.

4.1 Calculation of theoretical sample size

Eq. (2) shows that an RSP data block has a consistent probability distribution with the big data set for the given significance level and error threshold. The big data set $\mathbb{X}$ can be deemed as the population of random variable $\mathcal{X}$. We need to randomly draw $\mathcal{M}$ sample points from $\mathbb{X}$ based on the sampling scheme without replacement so that they can be used to empirically determine the probability distribution function. Recently, the multivariate

Algorithm 1 Theoretical sample size determination

```
Input: The dimension $\mathcal{D}$ of the data set, a significance level $\alpha \in (0,1)$, and an error threshold $\varepsilon \in (0,1)$;
Output: The theoretical sample size $\mathcal{M}$;
Initialize the theoretical sample size $\mathcal{M}^{(0)}$;
Initialize the iteration number $\mathcal{I} = -1$;
repeat
    $\mathcal{I} \leftarrow \mathcal{I} + 1$;
    $\mathcal{M}^{(\mathcal{I}+1)} \leftarrow \mathrm{I}(\lceil \mathcal{M}^{(\mathcal{I})} \rceil)$, where $\lceil \bullet \rceil$ is the ceiling function;
until $|\lceil \mathcal{M}^{(\mathcal{I}+1)} \rceil - \lceil \mathcal{M}^{(\mathcal{I})} \rceil| = 0$
Obtain the theoretical sample size $\mathcal{M} = \lceil \mathcal{M}^{(\mathcal{I})} \rceil$.
```

Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [10] provided a bound for the approximation between EDF and PDF as

$$\mathrm{P}(\sup_{\mathrm{x} \in \Re^{\mathcal{D}}} \left| \hat{F}_{\mathcal{M}}(\mathrm{x}) - F(\mathrm{x}) \right| > \varepsilon) \leq \mathcal{D}(\mathcal{M}+1)e^{-2\mathcal{M}\varepsilon^2}, \tag{11}$$

where $\hat{F}_{\mathcal{M}}(\mathrm{x})$ is the EDF constructed based on the $\mathcal{M}$ random sample points and $\mathcal{M}$ is the theoretical sample size. We let

$$\mathcal{D}(\mathcal{M}+1)e^{-2\mathcal{M}\varepsilon^2} = \alpha \tag{12}$$

and can derive the iterative expression as

$$\mathcal{M} = \frac{1}{2\varepsilon^2} \ln \frac{\mathcal{D}(\mathcal{M}+1)}{\alpha}. \tag{13}$$

It is difficult to calculate the analytical solution of $\mathcal{M}$ from Eq. (13). Thus, the fixed-point iteration (FPI) method described in Algorithm 1 is used to find an approximation of $\mathcal{M}$. Let

$$\mathrm{I}(\mathcal{M}) = \frac{1}{2\varepsilon^2} \ln \frac{\mathcal{D}(\mathcal{M}+1)}{\alpha} \tag{14}$$

express the iteration function of the FPI method, which is a monotonically increasing concave function of $\mathcal{M}$ and satisfies the following two conditions:

1) for $\mathcal{M} \in [1, \mathcal{N}]$, $\mathrm{I}(\mathcal{M}) \in [1, \mathcal{N}]$ holds;

2) there exists $\gamma \in (0,1)$ such that $|\mathrm{I}'(\mathcal{M})| \leq \gamma$ holds for any $\mathcal{M} \in [1, \mathcal{N}]$.

The theoretical analysis of the convergence of the iteration function $\mathrm{I}(\mathcal{M})$ is provided in Appendix A.

Because $\mathcal{M}$ is the sample size, the original design intent was for its value to be smaller than the population size $\mathcal{N}$. Due to the iteration function's convergence, the theoretical sample size satisfies the condition $|\mathrm{I}(\mathcal{M})| \leq \mathcal{N}$ when $\mathcal{M} \leq \mathcal{N}$. For example, when $\mathcal{N}$=1,000,000 (assume that this is the population size of a big data set), the maximum $\mathcal{M}$ of 2982 for $\varepsilon = 0.05$, $\alpha = 0.05$, and $\mathcal{D} = 50$ by using the FPI method with an initial $\mathcal{M}$=500 and the minimum of $\mathcal{M}$ as 2982 with initial $\mathcal{M}$=10,000 can be calculated respectively. We can see that both M and its convergent value $\mathrm{I}(\mathcal{M})$ are smaller than $\mathcal{N}$. In fact, the main reason why $\mathcal{M}$ or $\mathrm{I}(\mathcal{M})$ is smaller than $\mathcal{N}$ is that the iteration function is monotonically increasing concave. In addition, because $\mathrm{I}(\mathcal{M})$ is a monotonically increasing concave function of $\mathcal{M}$, the first derivative of $\mathrm{I}(\mathcal{M})$ decreases gradually until its value is smaller than a given threshold γ.

The partial derivatives of $\mathrm{I}(\mathcal{M})$ with respect to $\mathcal{D}$, α, and ε are calculated as

$$\frac{\partial \mathrm{I}}{\partial \mathcal{D}} = \frac{1}{2\mathcal{D}\varepsilon^2}, \tag{15}$$

$$\frac{\partial \mathrm{I}}{\partial \alpha} = -\frac{1}{2\alpha\varepsilon^2}, \tag{16}$$

and

$$\frac{\partial \mathrm{I}}{\partial \varepsilon} = -\frac{1}{\varepsilon^3} \ln \frac{\mathcal{D}(\mathcal{M}+1)}{\alpha}. \tag{17}$$

We can derive the results of

$$|\frac{\partial \mathrm{I}}{\partial \varepsilon}| > |\frac{\partial \mathrm{I}}{\partial \mathcal{D}}| \tag{18}$$

and

$$|\frac{\partial \mathrm{I}}{\partial \varepsilon}| > |\frac{\partial \mathrm{I}}{\partial \alpha}| \tag{19}$$

for

$$\mathcal{D} > \max\{\frac{\alpha e}{\mathcal{M}+1}, \frac{\alpha e^{\frac{\varepsilon}{2\alpha}}}{\mathcal{M}+1}\}. \tag{20}$$

Thus, Eq. (18) and Eq. (19) always hold for the positive integers $\mathcal{D}$ and $\mathcal{M}$ and indicate that ε has a more significant impact on $\mathcal{M}$ than $\mathcal{D}$ and α. Table 1 quantitatively illustrates the influence of $\mathcal{D}$, α, and ε on the sample size $\mathcal{M}$. It can be observed that $\mathcal{M}$ is more sensitive to the error threshold ε than to the data dimension $\mathcal{D}$ and significance level α. For some given $\mathcal{D}$ and α, we can determine the theoretical sample size $\mathcal{M}_{\min}$ with a smaller ε (e.g., 0.05) and the incremental sample size $\mathcal{S}$ with a larger ε (e.g., 0.10) by calculating the practical sample size as explained next.

4.2 Determination of practical sample size

The selection of the theoretical sample size $\mathcal{M}$ depends on the data dimension $\mathcal{D}$, significance level α, and error threshold ε and does not take the probability distribution into account. Thus, we try to use the density estimation strategy to further optimize the theoretical sample size. An initial training data set with $\mathcal{M}_{\min}$ sample points and a validation data set are prepared

$$\begin{aligned} \mathbb{X}_{\mathrm{V}} = \{\mathrm{v}_l | \mathrm{v}_l = (v_{l1}, v_{l2}, \cdots, v_{l\mathcal{D}}), v_{ld} \in \Re, \\ d = 1, 2, \cdots, \mathcal{D}, l = 1, 2, \cdots, \mathcal{L}\} \end{aligned} \tag{21}$$

by randomly choosing the sample points from the big data set $\mathbb{X}$, where $\mathcal{L}$ is the number of sample points of the validation data set. The practical sample size $\mathcal{P}$ can be determined using Algorithm 2. Below, we provide a discussion of that algorithm.

Table 1: Influence of $\mathcal{D}$, α, and ε on sample size $\mathcal{M}$

α=0.05, $\mathcal{D}$=10		ε=0.05, $\mathcal{D}$=10		α=0.05, ε=0.05	
ε	$\mathcal{M}$	α	$\mathcal{M}$	$\mathcal{D}$	$\mathcal{M}$
0.01	83132	0.01	2981	2	2284
0.02	18933	0.02	2832	4	2436
0.03	7931	0.03	2745	6	2524
0.04	4267	0.04	2683	8	2586
0.05	2634	0.05	2634	10	2634
0.06	1774	0.06	2595	12	2674
0.07	1269	0.07	2562	14	2707
0.08	949	0.08	2533	16	2736
0.09	734	0.09	2507	18	2761
0.10	583	0.10	2484	20	2784

For the training data set

$$\begin{aligned}\mathbb{X}_{\mathrm{T}}^{(i)} = \{&\mathrm{a}_p^{(i)} \Big| \mathrm{a}_p^{(i)} = (a_{p1}^{(i)}, a_{p2}^{(i)}, \cdots, a_{p\mathcal{D}}^{(i)}),\\ &a_{pd}^{(i)} \in \Re, d = 1, 2, \cdots, \mathcal{D}, p = 1, 2, \cdots, \mathcal{P}^{(i)}\}\end{aligned} \tag{23}$$

corresponding to the i-th ($i = 0, 1, \cdots, \mathcal{I} + 1$) iteration, the estimated *p.d.f.*

$$\hat{f}^{(i)}(\mathrm{x}) = \frac{1}{\mathcal{P}^{(i)}} \sum_{p=1}^{\mathcal{P}^{(i)}} \prod_{d=1}^{\mathcal{D}} \frac{1}{\sqrt{2\pi} h_d^{(i)}} \exp[-\frac{1}{2}(\frac{x_d - a_{pd}^{(i)}}{h_d^{(i)}})^2] \tag{24}$$

can be constructed according to Eq. (4), where the rule-of-thumb [22] bandwidth parameters $h_1^{(i)}, h_2^{(i)}, \cdots, h_{\mathcal{D}}^{(i)}$ are determined as

$$\begin{aligned} h_d^{(i)} &= 1.06\sigma_d^{(i)}[\mathcal{P}^{(i)}]^{-\frac{1}{5}}\\ &= 1.06\sqrt{\frac{1}{\mathcal{P}^{(i)} - 1}\sum_{p=1}^{\mathcal{P}^{(i)}}(a_{pd}^{(i)} - \frac{\sum_{q=1}^{\mathcal{P}^{(i)}} a_{qd}^{(i)}}{\mathcal{P}^{(i)}})^2}\ [\mathcal{P}^{(i)}]^{-\frac{1}{5}}\end{aligned} \tag{25}$$

The estimated error in Algorithm 2 is calculated as

$$\begin{aligned}&\left\|\log_2 \mathrm{p}^{(i+1)} - \log_2 \mathrm{p}^{(i)}\right\|_2^2\\ &\qquad = \sum_{l=1}^{\mathcal{L}} [\log_2 \hat{f}^{(i+1)}(\mathrm{v}_l) - \log_2 \hat{f}^{(i)}(\mathrm{v}_l)]^2\end{aligned} \tag{26}$$

The incremental data set $\Delta\mathbb{X}_{\mathrm{T}}^{(j)}$, $j = 1, 2, \cdots, \mathcal{I} + 1$ is composed of $\mathcal{S}$ sample points that are randomly drawn from the big data set $\mathbb{X}$. In the implementation for experiments presented in this paper, the big data set is partitioned into three parts, i.e., the initial training data set with $\mathcal{M}_{\min}$ sample points, the validation data set with $\mathcal{L}$ sample points, and the incremental data set with Q RSP data blocks, where

$$Q = \left\lfloor \frac{\mathcal{N} - \mathcal{M}_{\min} - \mathcal{L}}{\mathcal{S}} \right\rfloor \gg \mathcal{I} + 1. \tag{27}$$

Then, the data block sampling scheme can be used to gradually increase the size of the training data set rather than the data point sampling scheme.

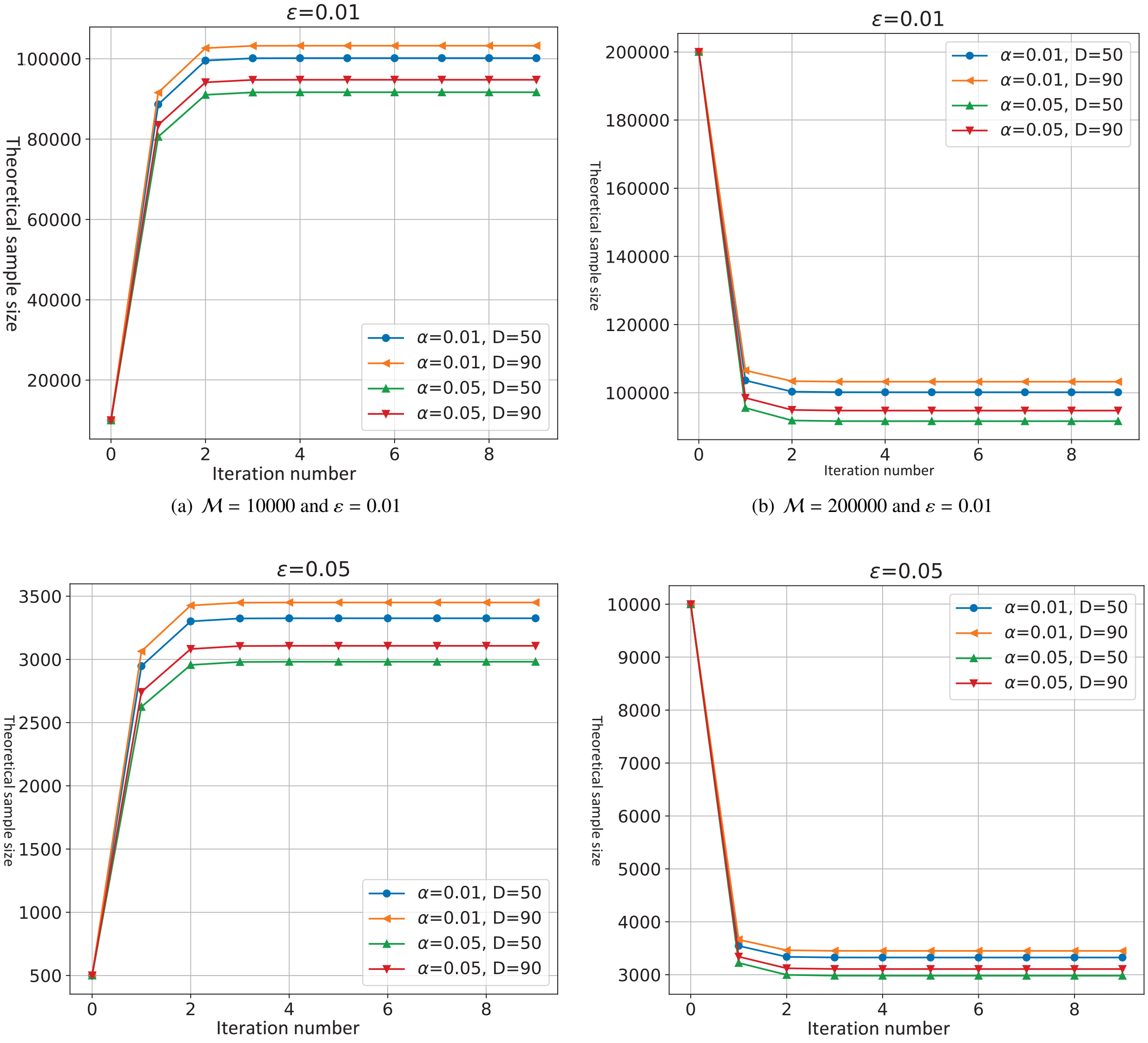
Algorithm 2 Practical sample size determination

```
Input: The incremental sample size $\mathcal{S}$, an initial training data set $\mathbb{X}_{\mathrm{T}}^{(0)}$, a validation data set $\mathbb{X}_{\mathrm{V}}$, and a stopping threshold $\xi > 0$;
Output: The practical sample size $\mathcal{P}$;
Calculate the theoretical sample size $\mathcal{M}_{\min}$ using Algorithm 1;
Initialize the iteration number $\mathcal{I} = -1$;
Estimate the p.d.f. $\hat{f}^{(0)}(\bullet)$ based on the training data set $\mathbb{X}_{\mathrm{T}}^{(0)}$;
Calculate the predicted output vector $\mathrm{p}^{(0)} = (\hat{f}^{(0)}(\mathrm{v}_1), \hat{f}^{(0)}(\mathrm{v}_2), \cdots, \hat{f}^{(0)}(\mathrm{v}_{\mathcal{L}}))$ of $\mathbb{X}_{\mathrm{V}}$;
repeat
    $\mathcal{I} \leftarrow \mathcal{I} + 1$;
    Update the training data set $\mathbb{X}_{\mathrm{T}}^{(\mathcal{I}+1)} \leftarrow \mathbb{X}_{\mathrm{T}}^{(\mathcal{I})} \cup \Delta\mathbb{X}_{\mathrm{T}}^{(\mathcal{I}+1)}$, where $\Delta\mathbb{X}_{\mathrm{T}}^{(\mathcal{I}+1)}$ is an incremental training data set with $\mathcal{S}$ sample points;
    Update the practical sample size
        $\mathcal{P}^{(\mathcal{I}+1)} = \mathcal{M}_{\min} + (\mathcal{I} + 1) \times \mathcal{S}$.    (22)
    Estimate the p.d.f. $\hat{f}^{(\mathcal{I}+1)}(\bullet)$ based on the training data set $\mathbb{X}_{\mathrm{T}}^{(\mathcal{I}+1)}$;
    Calculate the predicted output vector $\mathrm{p}^{(\mathcal{I}+1)} = (\hat{f}^{(\mathcal{I}+1)}(\mathrm{v}_1), \hat{f}^{(\mathcal{I}+1)}(\mathrm{v}_2), \cdots, \hat{f}^{(\mathcal{I}+1)}(\mathrm{v}_{\mathcal{L}}))$;
until $\left\|\log_2 \mathrm{p}^{(\mathcal{I}+1)} - \log_2 \mathrm{p}^{(\mathcal{I})}\right\|_2^2 \leq \xi$
Obtain the practical sample size $\mathcal{P} = \mathcal{P}^{(\mathcal{I})}$.
```

5 Experimental settings and results

This section presents a series of experiments that were conducted to validate the feasibility, rationality, and effectiveness of DEM for determining the sample size of RSP data blocks. Ten multiple-mode-and-multiple-dimension *p.d.f.*s were used to generate synthetic data sets. The number of *p.d.f.* modes are 10 and 20 and the number of *p.d.f.* dimensions are 10, 30, 50, 70, and 90. Each *p.d.f.* is expressed with the multivariate Gaussian mixture model, where the necessary parameters including component weights, mean vectors, and covariance matrices can be obtained from the *DEM Data Sets* folder of our BaiDuPan online storage space [1] using the extraction code “fbbp”. In addi-

[1] https://pan.baidu.com/s/1Vj0jHdSl3q99ygqnFWKt3A



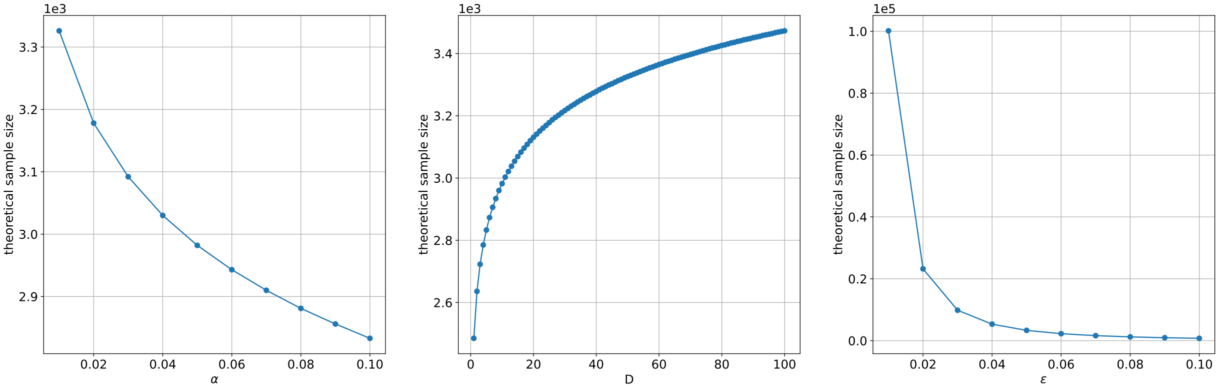
(a) $\mathcal{M}$ = 10000 and ε = 0.01

(b) $\mathcal{M}$ = 200000 and ε = 0.01

(c) $\mathcal{M}$ = 500 and ε = 0.05

(d) $\mathcal{M}$ = 10000 and ε = 0.05

Fig. 1: Convergence of Algorithm 1 for two error thresholds (ε = 0.01 and ε = 0.05)

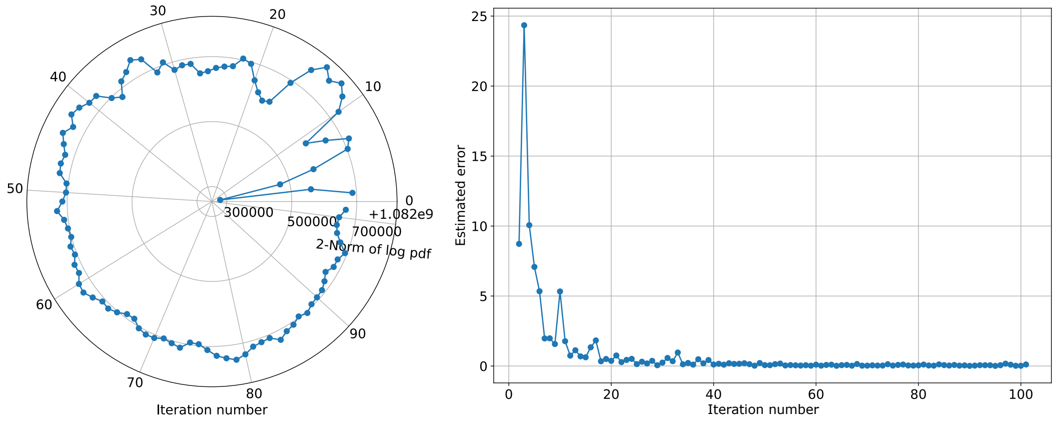


(a) Impact of significance level α

(b) Impact of data dimension $\mathcal{D}$

(c) Impact of error threshold ε

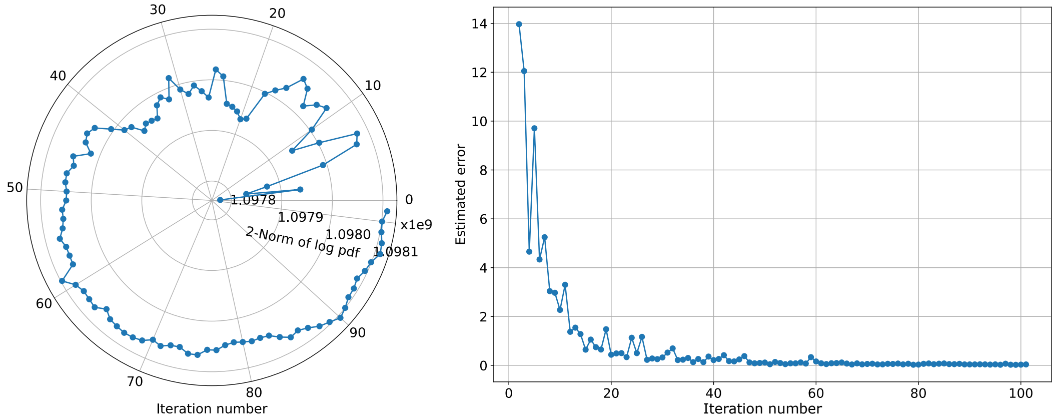
Fig. 2: Influence of α, $\mathcal{D}$, and ε for determining the optimal theoretical sample size



(a) 2-norm of log(*p.d.f.*)

(b) Estimated error

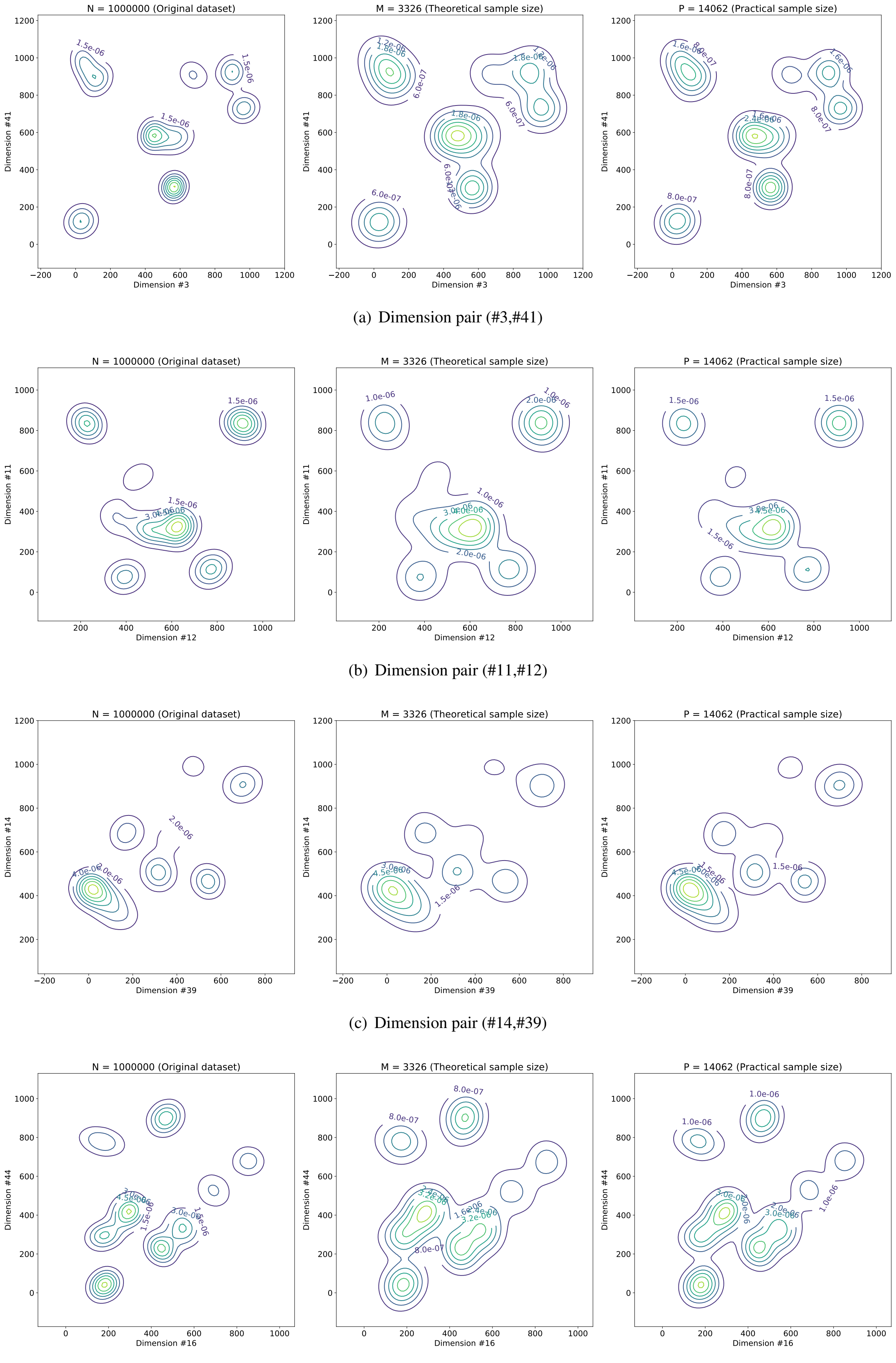
Fig. 3: Convergence of Algorithm 2 on a 10-mode-and-50-dimension synthetic data set (the theoretical sample size is 3326)



(a) 2-norm of log(*p.d.f.*)

(b) Estimated error

Fig. 4: Convergence of Algorithm 2 on a 20-mode-and-50-dimension synthetic data set (the theoretical sample size is 3326)



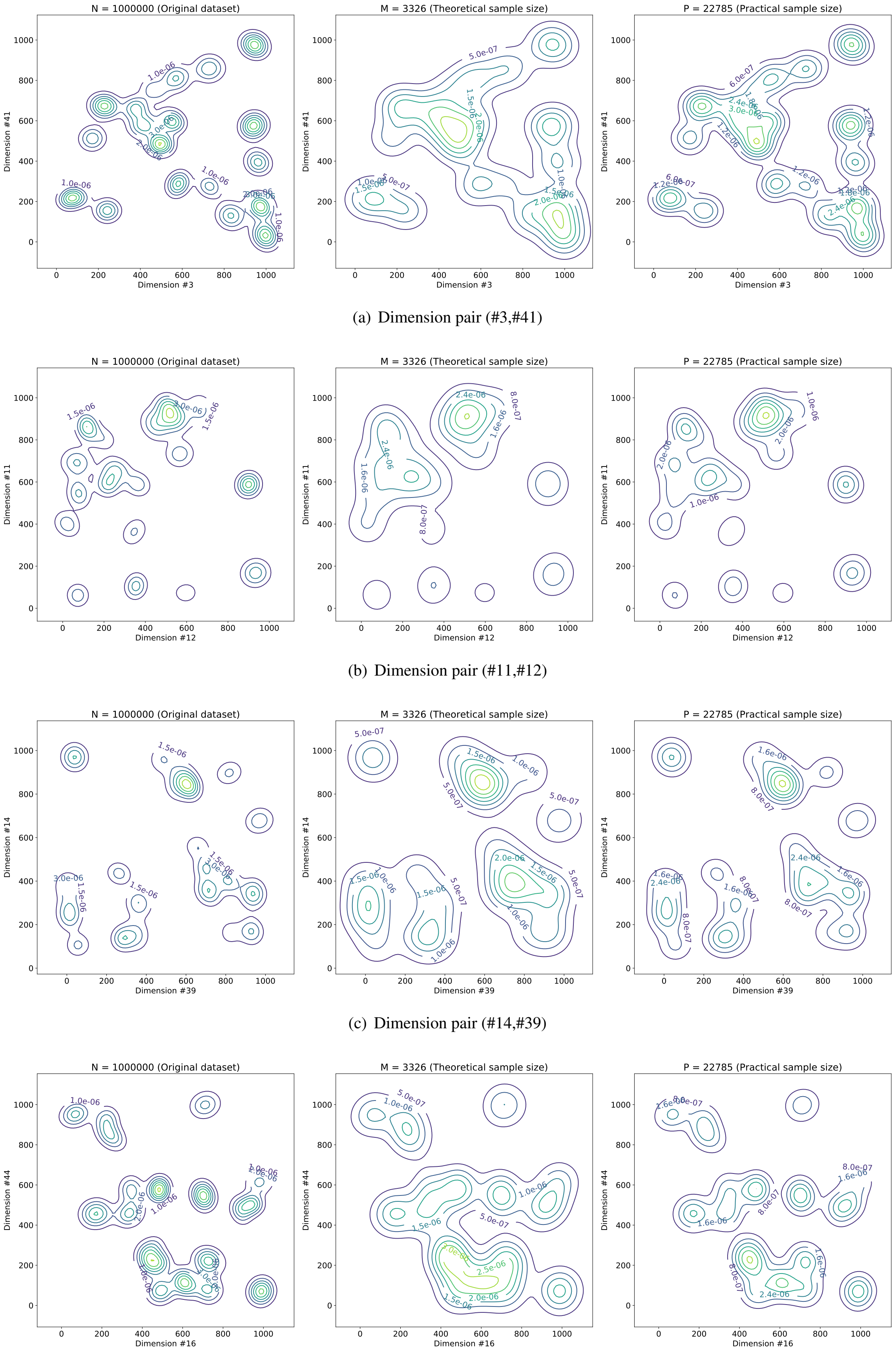
(a) Dimension pair (#3,#41)

(b) Dimension pair (#11,#12)

(c) Dimension pair (#14,#39)

(d) Dimension pair (#16,#44)

Fig. 5: Comparison of *p.d.f.* estimations for the 10-mode-and-50-dimension synthetic data set



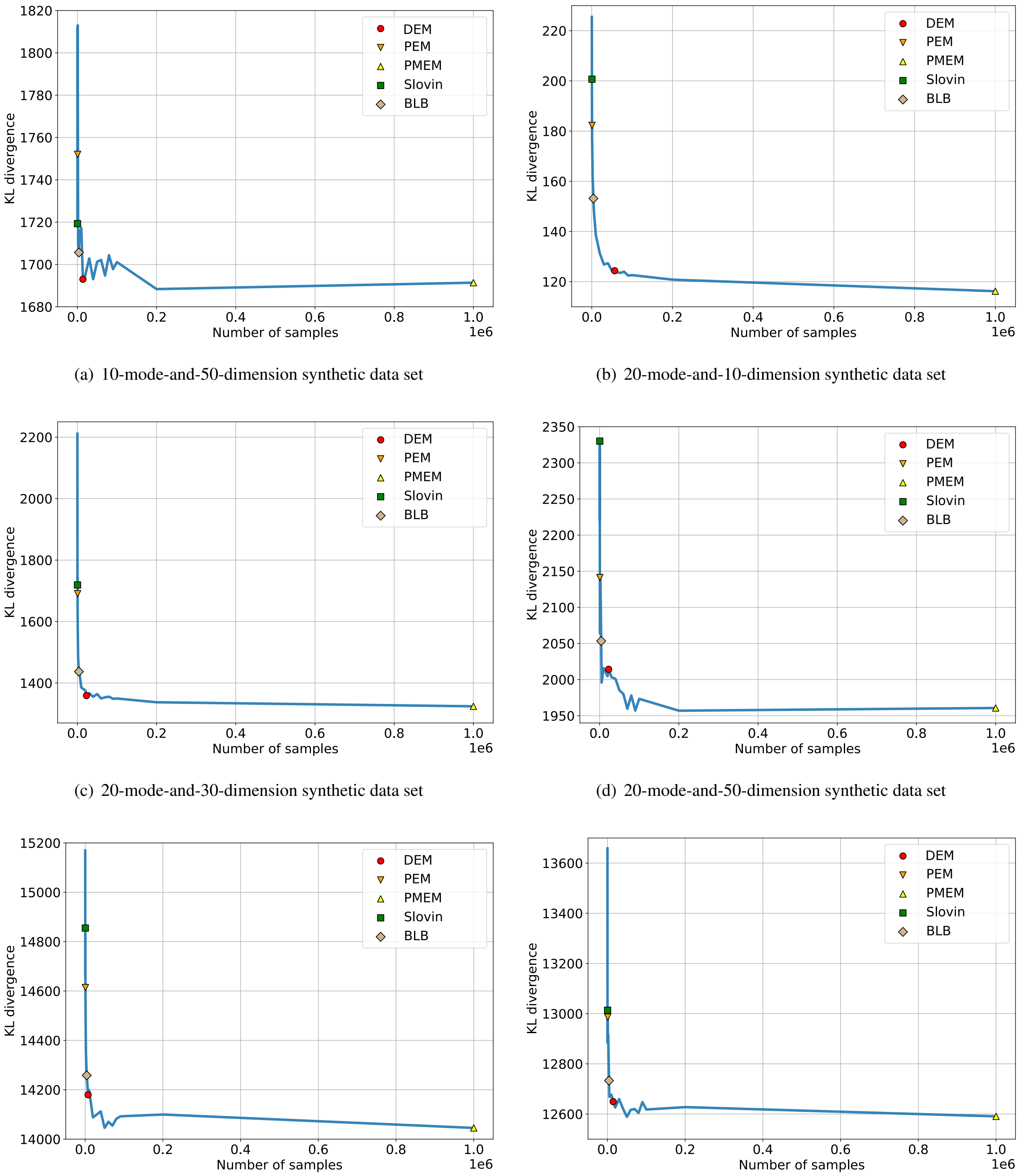
(a) Dimension pair (#3,#41)

(b) Dimension pair (#11,#12)

(c) Dimension pair (#14,#39)

(d) Dimension pair (#16,#44)

Fig. 6: Comparison of *p.d.f.* estimations for the 20-mode-and-50-dimension synthetic data set



(a) 10-mode-and-50-dimension synthetic data set

(b) 20-mode-and-10-dimension synthetic data set

(c) 20-mode-and-30-dimension synthetic data set

(d) 20-mode-and-50-dimension synthetic data set

(e) 20-mode-and-70-dimension synthetic data set

(f) 20-mode-and-90-dimension synthetic data set

Fig. 7: Relationship between the sample size and KL divergence

tion, the data used in the experiments can be obtained from this online folder. All methods were implemented with the Python programming language and experiments were run on a workstation equipped with an Intel(R) Xeon(R) CPU E5-2630 v2 running at 2.60 GHz with 12 Core(s), 24 logical processor(s) and 125 GB of main memory.

5.1 Feasibility validation of DEM

A first experiment was done to assess the ability of Algorithm 1 to converge for determining the theoretical sample size. For different parameter value pairs of significance level α and data dimension $\mathcal{D}$, the theoretical sample size ($\mathcal{M}$) was initialized to 10,000 and 200,000 for $\varepsilon = 0.01$, and 500 and 10,000 for $\varepsilon = 0.05$, respectively. The convergence of Algorithm 1 for different error thresholds is illustrated in Fig. 1. It can be observed that as the number of iterations increased, the small theoretical sample size gradually increased while the large theoretical sample size gradually decreased, until convergence. These experimental results confirm that the iteration function shown in Eq. (14) is reasonable and helpful to determine an optimal theoretical sample size.

Then, the influence of the significance level α, data dimension $\mathcal{D}$, and error threshold ε for determining the optimal theoretical sample size was evaluated. The experimental results are provided in Fig. 2, where the parameters were set as $\mathcal{D} = 50$, $\varepsilon = 0.05$, and $\alpha = 0.01, 0.02, \cdots, 0.10$ for Fig. 2(a); $\alpha = 0.01$, $\varepsilon = 0.05$, and $\mathcal{D} = 1, 2, \cdots, 100$ for Fig. 2(b); and $\alpha = 0.01$, $\mathcal{D} = 50$, and $\varepsilon = 0.01, 0.02, \cdots, 0.10$ for Fig. 2(c). These experimental results demonstrate the aforementioned conclusion that the theoretical sample size is more sensitive to the error threshold ε than to the significance level α and data dimension $\mathcal{D}$ because the variation range of theoretical sample sizes corresponding to ε is obviously larger than the ones corresponding to α and $\mathcal{D}$. This experiment provides useful insights on how to determine the sample size in practice.

5.2 Rationality validation of DEM

Another experiment was carried to test the convergence of Algorithm 2 and the *p.d.f.* estimation performance of a data subset having the practical sample size. For the 10-mode-and-50-dimension and 20-mode-and-50-dimension synthetic data sets, we first checked the 2-norm of the *p.d.f.* logarithm and the variation tendencies of the estimated error as shown in Eq. (26), as the number of iterations in Algorithm 2 increased. Then, we compared the estimated *p.d.f.*s with the true *p.d.f.*, where the unknown *p.d.f.* was constructed by applying the *Scipy* API *scipy.stats.gaussian_kde* [2) with the rule-of-thumb bandwidth. The learning parameters were set as follows: initial sample size of $\mathcal{M} = 100,000$, error threshold $\varepsilon = 0.05$ and significance level $\alpha = 0.01$ for determining the theoretical sample size with Algorithm 1, $\varepsilon = 0.1$ and $\alpha = 0.05$ to determine the incremental sample size with Algorithm 2, the validation data size $\mathcal{L} = 10,000$, and stopping threshold $\xi = 10^{-6}$. Figs. 3 and 4 show the convergence ability of Algorithm 2 for the two selected mixture distributions. It can be observed that Algorithm 2 converged as the iteration number increased, i.e., the 2-norm of the log(*p.d.f.*) and estimated error tend to gradually stabilize. This indicates that Algorithm 2 is appropriate for determining the optimal sample size of RSP data blocks.

In addition, we plotted the contour maps of the true *p.d.f.*s, and the estimated *p.d.f.*s based on the RSP data blocks with theoretical sample size and practical sample size. Four dimension pairs (#3,#41), (#11,#12), (#14,#39), and (#16,#44) were selected for this comparison. The comparative results are presented in Figs. 5 and 6. It can be seen that the estimated *p.d.f.*s based on the RSP data blocks with practical sample size are closer to the true *p.d.f.*s than the estimated *p.d.f.*s based on the RSP data blocks with the theoretical sample size. We also calculate numerical values to compare the *p.d.f.*s. The kullback-Leibler (KL) divergence [23,24] is used to measure the difference between two *p.d.f.*s. The model parameters of the true *p.d.f.*s are provided in the aforementioned BaiDuPan. The Python implementation of the KL divergence is available in the source code repository [3). It was found that the *p.d.f.*s estimated based on the RSP data blocks with practical sample sizes have smaller KL divergences than the *p.d.f.*s estimated based on the RSP data blocks with theoretical sample sizes. The KL divergences corresponding to the theoretical sample size and practical sample size are 1,738.030 and 1,693.005 in Fig. 5, while the KL divergences corresponding to the theoretical sample size and practical sample size are 2,104.779 and 2,014.266 in Fig. 6. This indicates that Algorithm 2 is able to determine the proper sample size based on the theoretical sample size found by Algorithm 1.

5.3 Effectiveness validation of DEM

Another experiment was done to compare the sample size of the proposed density estimation-based method (DEM) with those of the Slovin formula, population estimation method (PEM), population mean estimation method (PMEM), and the bag of little bootstraps (BLB) method for ten multiple-mode-and-multiple-dimension probability distributions. The parameters of the

[2) https://docs.scipy.org/doc/scipy/reference/generated

[3) https://gitee.com/nick-stu/dds

Table 2: Details of ten multiple-mode-and-multiple-dimension probability distributions

Probability distribution	Mode $\mathcal{R}$	Dimension $\mathcal{D}$	Sample number $\mathcal{N}$	Data size
10-mode-and-10-dimension *p.d.f.*	10	10	1,000,000	239 MB
10-mode-and-30-dimension *p.d.f.*	10	30	1,000,000	716 MB
10-mode-and-50-dimension *p.d.f.*	10	50	1,000,000	1.2 GB
10-mode-and-70-dimension *p.d.f.*	10	70	1,000,000	1.7 GB
10-mode-and-90-dimension *p.d.f.*	10	90	1,000,000	2.1 GB
20-mode-and-10-dimension *p.d.f.*	20	10	1,000,000	239 MB
20-mode-and-30-dimension *p.d.f.*	20	30	1,000,000	716 MB
20-mode-and-50-dimension *p.d.f.*	20	50	1,000,000	1.2 GB
20-mode-and-70-dimension *p.d.f.*	20	70	1,000,000	1.7 GB
20-mode-and-90-dimension *p.d.f.*	20	90	1,000,000	2.1 GB

DEM, Slovin formula, PEM, PMEM, and BLB method are summarized as follows.

- DEM: initial sample size $\mathcal{M} = 100,000$, error threshold $\varepsilon = 0.05$ and significance level $\alpha = 0.01$ for determining the theoretical sample size using Algorithm 1, $\varepsilon = 0.1$ and $\alpha = 0.05$ for determining the incremental sample size with Algorithm 2, validation data size $\mathcal{L} = 10,000$ and stopping threshold $\xi = 10^{-6}$;
- Slovin: $E = 0.05$;
- PEM: $E = 0.05$, $P = 0.5$, and $\alpha = 0.01$;
- PMEM: $E = 0.05$, $\alpha = 0.01$, and σ^2 is the minimum value of variances corresponding to all data dimensions;
- BLB: $v = 0.6$.

For each distribution, a series of big data sets were randomly generated. The sample number and data size for each dataset are listed in Table 2. For each distribution, 10 dependent data sets including 1,000,000 samples were randomly generated and the KL divergences between the true *p.d.f.* and estimated *p.d.f.*s corresponding to different sample size determination methods were calculated. The average sample sizes and KL divergences are presented in Table 3.

It is found that (1) all methods that yield a smaller sample size than DEM (i.e., Slovin, PEM, or BLB) provide a worse *p.d.f.*, which leads to obviously larger KL divergences than DEM, and (2) the method that generates a slightly better KL divergence than DEM (i.e., PMEM) yield a significantly larger sample size than DEM. For the Slovin formula, PEM, or BLB method, the sample sizes depend on the empirical parameters and are determined without considering information about the specific probability distribution of the big data. Although PMEM considers the variances of the big data set, it is easy to obtain extremely large sample sizes when variances are very large.

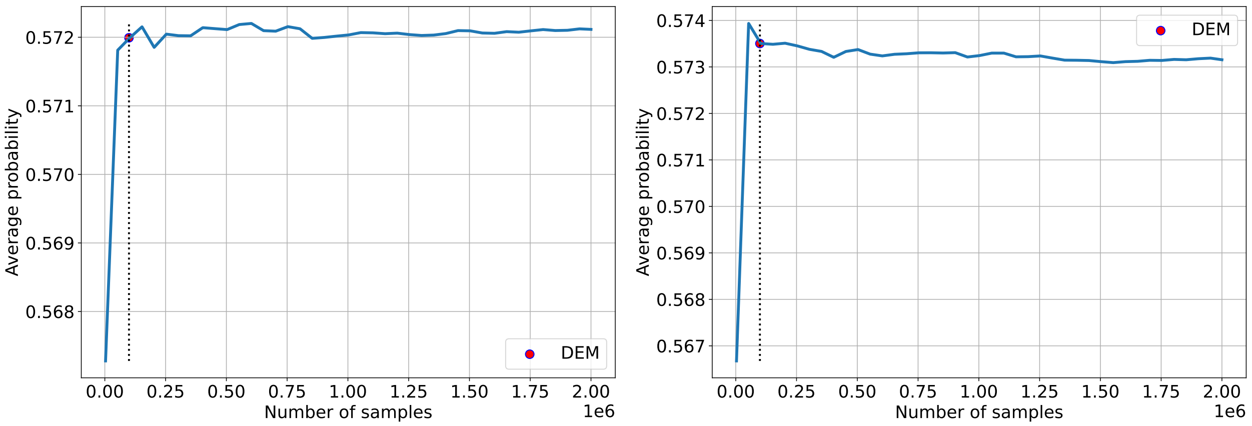
In addition, we analyzed the relationship between the sample size and KL divergence for six representative probability distributions. The sample sizes were varied from 100 to 1,000,000 samples using a step of 100. The experimental results are reported in Fig. 7. It can be observed that the KL divergences between the true *p.d.f.*s and estimated *p.d.f.*s gradually decrease as sample sizes are increased. The samples sizes determined by DEM, the Slovin formula, PEM, PMEM, and the BLB method are also marked on the convergence curves. It can be seen that the sample sizes determined by DEM are basically located at the convergence points of KL divergence curves. This indicates that the proposed DEM is an effective and efficient sample size determination method.

5.4 Application of DEM on real-world data sets

We selected two large-scale real-world data sets, i.e., the 28-dimensional HIGGS data set[4] with 11,000,000 sample points and the 27-dimensional HEPMASS data set [5] having 10,500,000 sample points to further validate the availability of DEM for the case of an unknown *p.d.f.*. Because the true *p.d.f.* of real-world data is unknown, the DEM's performance cannot be evaluated by measuring the KL divergence between the true *p.d.f.* and the estimated *p.d.f.*. Hence, two continuous-valued attributes were selected for each dataset to estimate their empirical distribution functions (EDFs) for demonstrating the rationality of the sample sizes. The theoretical sample size and incremental sample size are determined with parameter pairs $\alpha = 0.01$, $\varepsilon = 0.02$ and $\alpha = 0.05$, $\varepsilon = 0.05$, respectively. The experimental

[4] https://archive.ics.uci.edu/ml/datasets/HIGGS

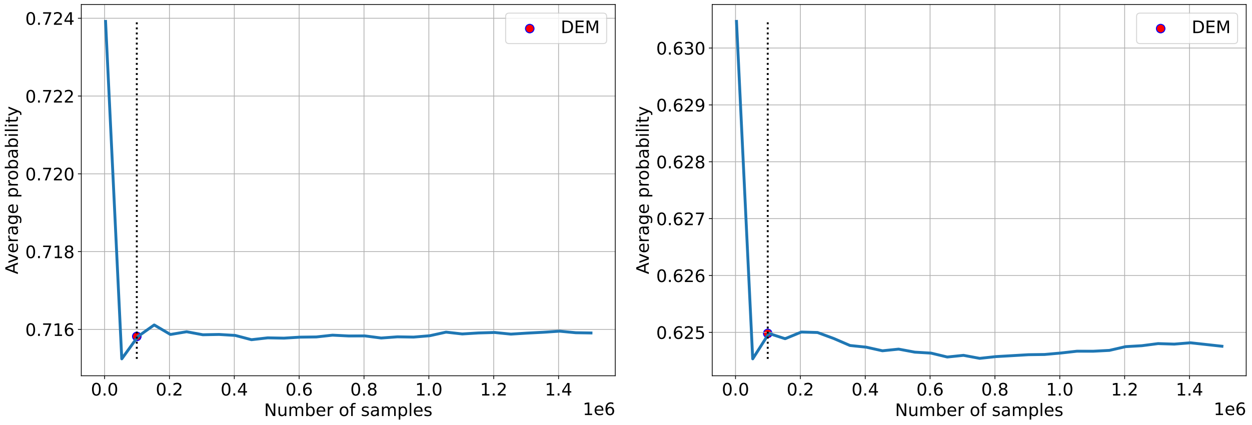
[5] https://archive.ics.uci.edu/ml/datasets/HEPMASS



(a) Convergence of EDF on attribute #7

(b) Convergence of EDF on attribute #11

Fig. 8: Simple size ($\mathcal{P} = 79,606$) determination on the HIGGS data set



(a) Convergence of EDF on attribute #1

(b) Convergence of EDF on attribute #7

Fig. 9: Simple size ($\mathcal{P} = 99,348$) determination on the HEPMASS data set

results are summarized in Fig. 8 and Fig. 9.

For each attribute, we estimated a series of EDFs for increasing sample sizes and checked the change of average probability calculated with the estimated EDF. The average probability is represented with the average value of probabilities corresponding to the sample points selected from a fixed sample point interval. In Fig. 8 and Fig. 9, we can see that (1) the average probability estimated with the EDF is convergent with the increase of sample size and (2) the practical sample size determined by the proposed DEM is consistent with the sample size to guarantee the convergence of the average probability. Table 4 shows the change of average probability when the sample size is increased with a step of 10,000. We can see that the error between two average probabilities calculated with EDFs is smaller than a given threshold of 10^{-4}. For example, the average probabilities estimated with EDFs corresponding to sample sizes $\mathcal{P}$ and $\mathcal{P}$+10,000 are 0.57200 and 0.57207 for attribute #7 of the HIGGS data set. It is worth noting that it is unnecessary to require identical values for two sample sizes (i.e., EDF's sample size and DEM's sample size) when the population size is very large. The small difference between the EDF's sample size and that of the DEM is acceptable for big data approximate computation in the real-world applications. This experiment demonstrates that the proposed DEM is also effective for real-world data sets with unknown *p.d.f.*.

6 Conclusion and future work

This paper proposed a new and effective density estimation method (DEM) to determine the sample size of random sample partition (RSP) data blocks. There are two main components in DEM, i.e., the theoretical sample size and practical sample size. The former is calculated by solving the multivariate

Dvoretzky-Kiefer-Wolfowitz (DKW) inequality and the latter is determined by minimizing the training error of kernel density estimator (KDE) constructed on RSP data blocks as the sample size is increased. The exhaustive experiments on a series of big data sets demonstrated the feasibility, rationality, and effectiveness of DEM.

In future work, we explore the following three research directions. First, we will implement the spatial distribution-based sample size determination method for RSP data block generation for big data. Second, we will derive an upper bound on the number of RSP data blocks based on the sample size determined by DEM for specific real-world applications. Third, we will devise a sample size determination method for large-scale mixed-attribute data sets.

Acknowledgements The authors would like to sincerely thank the editors and three anonymous reviewers whose valuable suggestions considerably helped improve the paper after two rounds of review. This paper was supported by National Natural Science Foundation of China (61972261), Natural Science Foundation of Guangdong Province (2023A1515011667), Key Basic Research Foundation of Shenzhen (JCYJ20220818100205012), and Basic Research Foundation of Shenzhen (JCYJ20210324093609026).

Appendixes

Appendix A. Convergence of $\mathrm{I}(\mathcal{M})$

If the condition 1) holds, we can derive

$$\mathcal{M} < \frac{\alpha \exp\left(2\varepsilon^2 \mathcal{N}\right)}{\mathcal{D}} - 1. \tag{28}$$

Let $\mathrm{B}(\mathcal{M})$ represent the right term of Eq. (28). For the big data set, $\mathrm{B}(\mathcal{N})$ will become very large when the value of $\mathcal{N}$ is large enough. For example, $\mathrm{B}(\mathcal{N}) = 1.404 \times 10^{214}$ for $\alpha = 0.05$, $\varepsilon = 0.05$, $\mathcal{D} = 50$, and $\mathcal{N} = 100000$. The result of $\mathcal{M} < \mathrm{B}(\mathcal{N})$ can be easily derived.

For the condition 2), the derivative of $\mathrm{I}(\mathcal{M})$ with respect to $\mathcal{M}$ is calculated as

$$|\mathrm{I}'(\mathcal{M})| = |\frac{\mathrm{dI}(\mathcal{M})}{\mathrm{d}\mathcal{M}}| = \frac{1}{2\varepsilon^2(\mathcal{M}+1)}. \tag{29}$$

If $\frac{1}{2\varepsilon^2(\mathcal{M}+1)} \leq 1$, we can derive

$$\mathcal{M} \geq \frac{1}{2\varepsilon^2} - 1 \tag{30}$$

which is easy to satisfy. For example, $\frac{1}{2\varepsilon^2} - 1 = 199$ for $\varepsilon = 0.05$. Thus, we can easily find a $\gamma \in (0,1)$ such that $|\mathrm{I}'(\mathcal{M})| \leq \gamma$ holds for any $\mathcal{M} \in [1, \mathcal{N}]$.

References

1. Sookhak M, Yu F R, Zomaya A Y. Auditing big data storage in cloud computing using divide and conquer tables. IEEE Transactions on Parallel and Distributed Systems, 2017, 29(5): 999–1012.
2. Zhao S, Li R, Tian W, Xiao W, Dong X, Liao D, Khan S U, Li K. Divide-and-conquer approach for solving singular value decomposition based on MapReduce. Concurrency and Computation: Practice and Experience, 2016, 28(2): 331–350.
3. Ghazi M R, Gangodkar D. Hadoop, MapReduce and HDFS: a developers perspective. Procedia Computer Science, 2015, 48: 45–50.
4. Hasan M I. Improving HDFS write performance using efficient replica placement. In: Proceedings of 2014 International Conference-Confluence The Next Generation Information Technology Summit. 2014, 36–39.
5. Salloum S, Huang J Z, He Y L. Random sample partition: a distributed data model for big data analysis. IEEE Transactions on Industrial Informatics, 2019, 15(11): 5846–5854.
6. Wei C H, Salloum S, Emara T Z, Zhang X L, Huang J Z, He Y L. A two-stage data processing algorithm to generate random sample partitions for big data analysis. In: Proceedings of 2018 International Conference on Cloud Computing, 2018, 347–364.
7. Yamane T. Statistics: an introductory analysis. Harper & Row, 1967.
8. Cochran W G. Sampling techniques. John Wiley & Sons, 2007.
9. Smith M F. Sampling considerations in evaluating cooperative extension programs, Florida Cooperative Extension Service Bulletin PE-1. Institute of Food and Agricultural Sciences, University of Florida, 1983.
10. Naaman M. On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. Statistics & Probability Letters, 2021, 173: 109088.
11. Kleiner A, Talwalkar A, Sarkar P, Jordan M. A scalable bootstrap for massive data. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 2014, 76(4): 795–816.
12. Reshef D N, Reshef Y A, Finucane H K, Grossman S R, McVean G, Turnbaugh P J, Lander E S, Mitzenmacher M, Sabeti P C. Detecting novel associations in large data sets. Science, 2011, 334(6062): 1518–1524.
13. Sengupta S, Stanislav V, Shao X F. A subsampled double bootstrap for massive data. Journal of the American Statistical Association, 2016, 111(515): 1222–1232.
14. R. H. Browne, On the use of a pilot sample for sample size determination. *Statistics in Medicine*, vol. 14, no. 17, pp. 1933–1940, 1995.
15. Lenth R V. Some practical guidelines for effective sample size determination. The American Statistician, 2002, 55(3): 187–193.
16. Ahmad W M, Amin W A, Aleng N A, Mohamed N. Some practical guidelines for effective sample-size determination in observational studies. Aceh International Journal of Science and Technology, 2012, 1(2): 51–53.
17. Burmeister E, Aitken L M. Sample size: how many is enough? Australian Critical Care, 2012, 25(4): 271–274.
18. Okada S, Ohzeki M, Taguchi S. Efficient partition of integer optimization problems with one-hot encoding. Scientific reports, 2019, 9(1): 1–2.
19. He Y L, Ye X, Huang D F, Fournier-Viger P, Huang J Z. A hybrid method to measure distribution consistency of mixed-attribute data sets. IEEE Trans-

actions on Artificial Intelligence, 2023, 4(1): 182–196.

20. Parzen E. On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 1962, 33(3): 1065–1076.

21. Jiang J, He Y L, Dai D X, Huang J Z. A new kernel density estimator based on the minimum entropy of data set. Information Sciences, 2019, 491: 223–231.

22. Jones M C, Marron J S, Sheather S J. A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association, 1996, 91(433): 401–407.

23. Perez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions. In: Proceedings of 2008 IEEE International Symposium on Information Theory, 2008, 1666–1670.

24. Perez-Cruz F. Estimation of information theoretic measures for continuous random variables. In: Proceedings of the 21st International Conference on Neural Information Processing Systems, 2008, 1257–1264.

Table 3: Comparison of KL divergence for different sample size determination methods

Data set	($\mathcal{R}$, $\mathcal{D}$)	DEM		PEM		PMEM		Slovin		BLB	
		sample size	KL divergence	sample size	KL divergence	sample size	KL divergence	sample size	KL divergence	sample size	KL divergence
Data set #1	(10,10)	22,721±1,801	5.472±0.215	664	27.191±0.724	120,704,033	1.382±0.009	400	33.443±1.048	3,982	12.391±0.160
Data set #2	(20,10)	17,231±1,609	133.085±2.747	664	193.707±7.939	126,073,272	117.481±0.558	400	210.179±7.095	3,982	152.980±4.777
Data set #3	(10,30)	23,309±5,806	390.160±6.047	664	461.030±31.830	88,161,518	387.137±2.550	400	482.872±32.974	3,982	408.454±7.421
Data set #4	(20,30)	18,222±2799	1,371.175±11.426	664	1,518.077±73.759	146,430,907	1,332.028±5.745	400	1,570.067±110.786	3,982	1,418.196±37.204
Data set #5	(10,50)	19,497±1,358	1,692.886±17.667	664	1,766.343±83.468	58,328,135	1,698.777±5.474	400	1,832.928±112.976	3982	1,713.058±49.177
Data set #6	(20,50)	21,644±6,451	2,002.653±18.110	664	2,325.073±105.687	115,185,014	1,954.991±8.062	400	23,11.615±164.011	3,982	2,074.893±45.890
Data set #7	(10,70)	18,577±2,909	3,567.995±36.217	664	3,659.661±130.785	73,566,581	3,563.855±11.920	400	3,714.130±113.911	3,982	3,593.872±27.881
Data set #8	(20,70)	17,473±2,912	14,160.358±42.196	664	14,727.513±168.544	110,208,079	14,070.002±14.483	400	14,649.220±175.349	3982	14,241.059±59.866
Data set #9	(10,90)	18,143±4,488	4,896.886±29.501	664	5,104.065±120.971	69,407,840	4,883.667±7.847	400	5,215.207±83.576	3,982	4,913.157±53.618
Data set #10	(20,90)	18,776±2,827	12,648.245±42.346	664	13,025.018±166.676	123,067,620	12,603.501±16.031	400	13,102.256±172.041	3,982	12,697.893±108.720

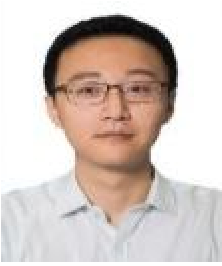
[1] The sample sizes determined by the PEM, PMEM, Slovin, and BLB rely on population size, margin of error, sample proportion, population standard deviation, and significance level.

[2] DEM may determine the different sample sizes for different data sets, even if they have the same population size and obey the same probability distribution.

[3] The KL divergence is used to measure the error between the true *p.d.f.* and estimated *p.d.f.*, which is constructed based on the samples corresponding to the sample size determined with the specific sample size determination method (e.g., DEM, PEM, PMEM, Slovin, and BLB).

Table 4: Average probabilities of EDFs with different sample sizes

Sample size	HIGGS		HEPMASS	
	Attrib #7	Attrib #11	Attrib #1	Attrib #7
$\mathcal{P}$	0.57200	0.57350	0.71586	0.62498
$\mathcal{P}$+10,000	0.57207	0.57365	0.71592	0.62513



Yulin HE was born in 1982. He received the Ph.D. degree from Hebei University, China, in 2014. From 2011 to 2014, he has served as a Research Assistant with the Department of Computing, Hong Kong Polytechnic University, Hong Kong, China. From 2014 to 2017, he worked as a Postdoctoral Fellow in the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is currently a Research Associate with Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China. His main research interests include big data approximate computing technologies, multi-sample statistical analysis theories and methods, and data mining/machine learning algorithms and their applications. He has published over 100+ research papers in ACM Transactions, CAAI Transactions, IEEE Transactions, Elsevier, Springer Journals and PAKDD, IJCNN, CEC, DASFAA conferences. Dr. He is an ACM member, CAAI member, CCF member, IEEE member, and the Editorial Review Board members of several international journals.

Jiaqi CHEN was born in 1999 and received her bachelor degree from Shenzhen University in 2021. She is currently pursuiting her Ph.D. degree in the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her main research interests include big data approximate computing technologies, multi-sample statistical analysis theories and methods, and data mining/machine learning algorithms and their applications.

Jiaxing SHEN an Assistant Professor at the Department of Computing and Decision Sciences, Lingnan University. He obtained B.E. in Software Engineering and Ph.D. in Computer Science from Jilin University in 2014 and The Hong Kong Polytechnic University in 2019, respectively. His central research theme is Human Dynamics which refers to interdisciplinary research of human behavior with an aim to understand human behavior and provide actionable insights. Under the theme, he has several research interests including Mobile Computing, Data Mining, and IoT systems. He has published over 25 papers in top-tier journals and conferences including TMC, TKDE, TOIS, IMWUT, IoT-J, INFOCOM, WWW, ICDM, and ICDCS. He has won Best Paper Award twice including one from INFOCOM 2020. He also served as a Session Chair of MASS 2021.

Philippe Fournier-Viger was born in 1980. He is a distinguished professor at the College of Computer Science and Software Engineering at the Shenzhen University (China). He obtained a title of national talent from the National Science Foundation of China. He has published more than 300 research papers related to data mining, big data, intelligent systems and applications, which have received more than 10,000 citations (H-Index 51). He is editor-in-chief of the Data Science and Pattern Recognition journal and former associate editor-in-chief of the Applied Intelligence journal (SCI, Q1). He is the founder of the SPMF data mining library, offering more than 230 algorithms, used in more than 1,000 research papers. He is co-founder of the UDML, PMDB and MLiSE series workshop held at the ICDM, PKDD, DASFAA and KDD conferences. His interests are data mining, algorithm design, pattern mining, sequence mining, big data, and applications.

Joshua Zhexue Huang was born in 1959. He received the Ph.D. degree from The Royal Institute of Technology, Stockholm, Sweden, in 1993. He is currently a Distinguished Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is also the Director of Big Data Institute, Shenzhen, China, and the Deputy Director of National Engineering Laboratory for Big Data System Computing Technology. He has published over 200 research papers in conferences and journals. His main research interests include big data technology and applications. Prof. Huang received the first PAKDD Most Influential Paper Award in 2006. He is known for his contributions to the development of a series of k-means type clustering algorithms in data mining, such as k-modes, fuzzy k-modes, k-prototypes, and w-k-means that are widely cited and used, and some of which have been included in commercial software. He has extensive industry expertise in business intelligence and data mining and has been involved in numerous consulting projects in many countries.