Data Engineering and Applications

Research Paper

Efficient Sampling Algorithms for Approximate Computation of High-Dimensional Data in Scientific Computing

Bronisław Lucjan Latocha[1,*], Celina Hajdukowa[1] and Czesława Latocha[2]

[1] Faculty of Information Technology, Casimir Pulaski University of Radom, Radom, 26-600, Poland
[2] Faculty of Informatics, Jan Kochanowski University, Kielce, 25-020, Poland
*Corresponding author: bronislaw.latocha@uniwersytetradom.pl

Abstract. As the dimensionality of high-dimensional data increases, scientific computing faces many challenges. There is an urgent need for high-quality approximation methods. To address the high-dimensional data approximation problems frequently encountered in fields such as environmental monitoring, genomics, and remote sensing, this paper introduces a new class of efficient sampling algorithms. In order to improve the speed and reliability of data-based statistics, the aforementioned two methods were chosen. A large number of experiments have shown that the above methods perform well in synthesizing Gaussian fields, hyperspectral images, and single-cell gene expression matrices. According to the results, the proposed algorithm achieved an average approximation error of less than 0.018, even with a sampling ratio of only 5%. It outperforms traditional importance sampling (0.043) and uniform random sampling (0.071), and exhibits linear scalability in terms of computation time and memory. To demonstrate that reliable and accurate estimators can still be used under adaptive weighting and feedback mechanisms, ablation experiments and sensitivity analyzes were also conducted. These findings provide new standards for the sample economy and accuracy of data-intensive scientific processes. The study provides the scientific community with a reproducible and open workflow to help accelerate research and gain deeper insights into high-dimensional data.

Keywords: *High-Dimensional Data, Efficient Sampling, Approximate Computation, Scientific Computing, Statistical Analysis*

Received on 16 October 2024, Accepted on 19 April 2025, Published on 25 April 2025

Introduction

Simulation technology, sensor networks, and large-scale experimental facilities in the new era of research now all use high-dimensional data. Fields such as genomics and climate modeling have recently generated large amounts of data containing many variables. Many researchers need to handle datasets containing thousands or even millions of variables. As the volume of data increases, the opportunities for scientific discovery also increase, leading to many other practical issues that require efficient large-scale data processing techniques to extract knowledge from complex datasets [1]. It has been proven that the ability to accurately approximate, summarize, and interpret high-dimensional data is a key capability for more effective research and application of high-dimensional data in fundamental research and engineering [2]. Due to the increasing scale and difficulty of scientific problems, a general method is needed to handle these high-dimensional inputs [3].

Despite recent progress, most methods used for approximating and reducing high-dimensional data still face significant issues. Traditional sampling and dimensionality reduction methods include Principal Component Analysis (PCA), random projection, simple random sampling, and stratified sampling. For a large number of variables or complex data, these methods are not accurate enough or efficient [4]. These methods are usually sensitive to data sparsity, noise, or complex dependency structures. Therefore, they are unreliable in large-scale heterogeneous data, and their usage is also very costly [5]. Many popular methods lack the ability to capture

https://doi.org/10.64972/dea.2025.v4i2.2005d:57-71

local data features and the adaptability required to meet the diverse analytical needs in practical scientific work [6]. Due to the rapid development of scientific computing, current algorithms cannot meet the demands for speed and large-scale operations. Lack of strong theoretical support for approximation accuracy and interpretability [7]. The demand for overcoming the curse of dimensionality, providing stable performance, and effectively extending to the next generation of sampling and approximation frameworks for various scientific problems is gradually being met [8]. The integration of machine learning, uncertainty quantification, and automated modeling pipelines in computational science has exacerbated this long-standing issue [9]. Addressing the aforementioned issues is one of the key obstacles to the large-scale application of data-driven discoveries across various fields [10].

To address the aforementioned issues, this paper develops an efficient sampling algorithm for high-dimensional data in scientific computing. In addition, based on the latest research in adaptive sampling and approximation theory, the aforementioned methods can also be developed, which are both computationally feasible and highly accurate across a wide range of problems. Through multiple experiments on various scientific datasets, it was found that the proposed algorithm is more accurate and requires less computational effort than traditional baselines. The above content will provide a platform for scientific data analysis and offer new directions for the next generation of data-driven research.

Background

High-Dimensional Scientific Data

In recent years, new scientific research has generated high-dimensional data and developed new analytical and computational methods. In fields such as climate science, genomics, neuroscience, and advanced materials characterization, data often generate thousands or even millions of features and variables, each of which may have different scales and complex dependencies. Environmental monitoring arrays collect large amounts of temporal and spatial data from multiple channels. With the increase in speed and resolution, next-generation sequencing technology has begun to generate multidimensional maps of biological systems [11]. A large amount of data will be used to study new problems in science, discover subtle changes in the data, and build accurate predictive models. Moreover, high-dimensional data cannot be processed independently. Due to the curse of dimensionality, statistical noise and sparsity increase. Traditional geometric and statistical assumptions are no longer applicable [12]. Due to the fact that the distance between data points decreases as the dimensionality increases, the performance of most algorithms declines, making it more difficult to find relevant structures [13]. Storing, visualizing, and interpreting are becoming increasingly difficult because the feature dimensions exceed the availability of traditional computational resources [14]. To fully utilize the high-dimensional data of current scientific problems, methodological improvements and computational optimizations are needed [15].

Sampling and Approximate Computing Methods

To address the aforementioned issues, many data reduction and approximation methods have been proposed. In this context, importance sampling, stratified sampling, and random sampling are often used to obtain representative datasets for subsequent analysis [16]. Dimensionality reduction techniques can uncover the underlying structure of high-dimensional data, such as Principal Component Analysis (PCA), Non-negative Matrix Factorization, and many subspace learning algorithms [17]. Sampling-based strategies have also garnered attention; the development of sparse coding and compressed sensing techniques allows signals to be reconstructed with only a few measurements [18]. These methods are not suitable for large-scale problems in modern scientific computing, despite their many applications. Random sampling may result in the inability to obtain important outlier data or rare scientific discoveries [19]. Stratified and uniform sampling can be used for certain distributions, but they usually perform poorly when the data has heterogeneity or nonlinear structures [20]. Although approximate calculations can reduce computational load, some fundamental statistical or physical characteristics of the data, such as uncertainty quantification constraints or conservation laws, are often overlooked [21]. These obstacles require a reassessment of the origins and applications of current methods [22].

https://doi.org/10.64972/dea.2025.v4i2.2005d:57-71

Key Limitations and Motivation

A review of existing methods indicates that when faced with increasing data volumes and scientific demands, general issues such as accuracy, speed, and generalization ability have not yet been resolved. Due to their lower approximation accuracy or excessive resource requirements, classical algorithms are generally not suitable for studying local features or specific needs [23]. Few of these methods can provide theoretical support for risk-sensitive settings while considering the performance in both average and worst-case scenarios. Due to the slow development of hardware and computing facilities, many advanced sampling algorithms have not yet been developed to address the speed issue of high-dimensional scientific data accumulation [24]. Therefore, a method for intelligently selecting representative samples is needed to maintain the fundamental structure of the original data and perform well in high-dimensional environments. Recently, new effective and adaptive sampling methods have been developed to provide robust mathematical support. It is expected that these methods will help extract more value from scientific computations in the future and drive new advancements [25].

Proposed Sampling Techniques

Problem Formulation and Notation

A typical case is a high-dimensional scientific dataset that contains a large number of samples and many features for each sample, such as computational physics, environmental monitoring, and genomics data. First, obtaining a relatively short and accurate representative sample of the data is achieved by selecting a small number of samples and assigning different weights; otherwise, only basic statistical and structural features are obtained. Use an approximate method to reconstruct the target scientific quantity, such as an aggregate response function or transformation. This method strictly constrains the expected error between the estimated results and the actual potential values under a specific loss function.

An aggregation function can be used to represent the initial scientific measure of the entire dataset, such as the empirical mean or extended linear and nonlinear operators, to formalize this task. The approximation is constructed by means of a weighted sum over a much smaller, carefully chosen subset of samples:

$$\text{Estimate} = \sum_{i=1}^{k} \omega_i q\left(x_{s_i}\right) \qquad \text{Eq.(1)}$$

where $q\left(x_{s_i}\right)$ indicates the measured quantity from the selected sample, and ω_i is its associated adaptive weight.

The goal is to determine the selection sequence and weight vector that jointly minimize the expected loss, expressed as

$$\mathcal{L} = \mathbb{E}\left|Q_{\text{true}} - \sum_{i=1}^{k} \omega_i q\left(x_{s_i}\right)\right| \qquad \text{Eq.(2)}$$

where Q_{true} denotes the evaluation over the full dataset. Fundamental to this construction is the tradeoff between computational gain-quantified by the reduction in sample size-and the statistical risk, quantified by the increase in estimation error. This tradeoff can be formalized through a fidelity constraint, such that the probability the estimation error exceeds a threshold is negligible:

$$\text{Prob}(\mathcal{L} > \varepsilon) < \delta \qquad \text{Eq.(3)}$$

To ensure that the chosen subset and corresponding weights together minimize both bias and variance, we further introduce explicit variance decomposition:

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} \omega_i q\left(x_{s_i}\right) - Q_{\text{true}}\right)^2\right] = \frac{V}{k} \qquad \text{Eq.(4)}$$

where V denotes data-driven variance that arises from the adaptive sampling mechanism.

Finally, the challenge is to design a selection and weighting procedure under which this expected risk, and higher moments of error, exhibit rapid decay rates in terms of the sampling budget.

https://doi.org/10.64972/dea.2025.v4i2.2005d:57-71

Algorithmic Design and Theoretical Analysis

According to the aforementioned issues, the algorithm adjusts the sampling method based on the degree of loss sensitivity and feature importance. A reasonable and data-adaptive method was adopted. In each recursive step, the new selection probability for candidates is dynamically set based on feedback to reflect the remaining errors and data structure modifications.

In the j -th step, the adaptive importance score of each unselected candidate is based on a combination of global metrics of feature impact and loss gradient response. The following formula is used to determine the probability score of candidates x:

$$\pi_j(x) = \frac{\alpha\phi\left(\left|\nabla_x R_{j-1}\right|\right) + \beta \cdot \psi(x)}{\sum_{y \in U_j} \alpha\phi\left(\left|\nabla_y R_{j-1}\right|\right) + \beta \cdot \psi(y)} \quad \text{Eq.(5)}$$

where ϕ captures the magnitude of local error not explained by prior samples, ψ expresses the aggregate relevance of the candidate's features, and the coefficients α and β adapt to trends in statistical residuals and global structure.

According to the inverse relationship with the sampling probability, the new selection and its corresponding weight will be immediately determined:

$$\omega_j = \frac{1}{k\pi_j\left(x_{s_j}\right)} \quad \text{Eq.(6)}$$

This ensures the variance and statistical unbiasedness of the overall estimator.

In each iteration, the estimates of the scientific function or statistic are updated as follows:

$$\hat{Q}_j = \hat{Q}_{j-1} + \omega_j\left[q\left(x_{s_j}\right) - \hat{Q}_{j-1}\right] \quad \text{Eq.(7)}$$

Provided an unbiased recursive updating process that maintains similar stability while collecting more information.

The first statistical guaranty is that, as the inverse of the square root of the sample size increases, the expected estimation error will decrease, and

$$\mathbb{E}\left|\hat{Q}_k - Q_{\text{true}}\right| \leq \frac{C}{\sqrt{k}} \quad \text{Eq.(8)}$$

with C dynamically reflecting dataset variance after adaptive refinement. The above indicates that the adaptive priority strategy can quickly and accurately estimate data complexity. This is a relatively general method that can provide timely feedback on local and global areas of the problem at low cost and with high accuracy.

Structural Overview of the Algorithm

Figure 1 shows the process of the new efficient sampling method. The first step in data preprocessing is to determine the baseline feature correlation and initial loss sensitivity. Then comes the creation of the first set of importance parameters. Each time, adaptively select samples and update the selection probability based on the most recent estimator residuals and feature-level evaluations. After selecting an option, check if the estimator and error monitoring module are available. Then, update the estimator and monitor for errors.

The parameter adjustment process is conducted in parallel, using optional out-of-sample validation and empirical error convergence to optimize the adaptive scoring system and stopping threshold. Comprehensive feedback will help the algorithm simultaneously adjust the data variation paths.

Preprocessing, dynamic importance allocation, prioritized sample selection, parameter optimization, and weighted estimator updates are all parts of the workflow to ensure the stability and efficiency of bidirectional information flow. Therefore, it is a closed-loop system that can automatically adapt to changes in data structure and reliably achieve scientific and reasonable summaries.

https://doi.org/10.64972/dea.2025.v4i2.2005d:57-71

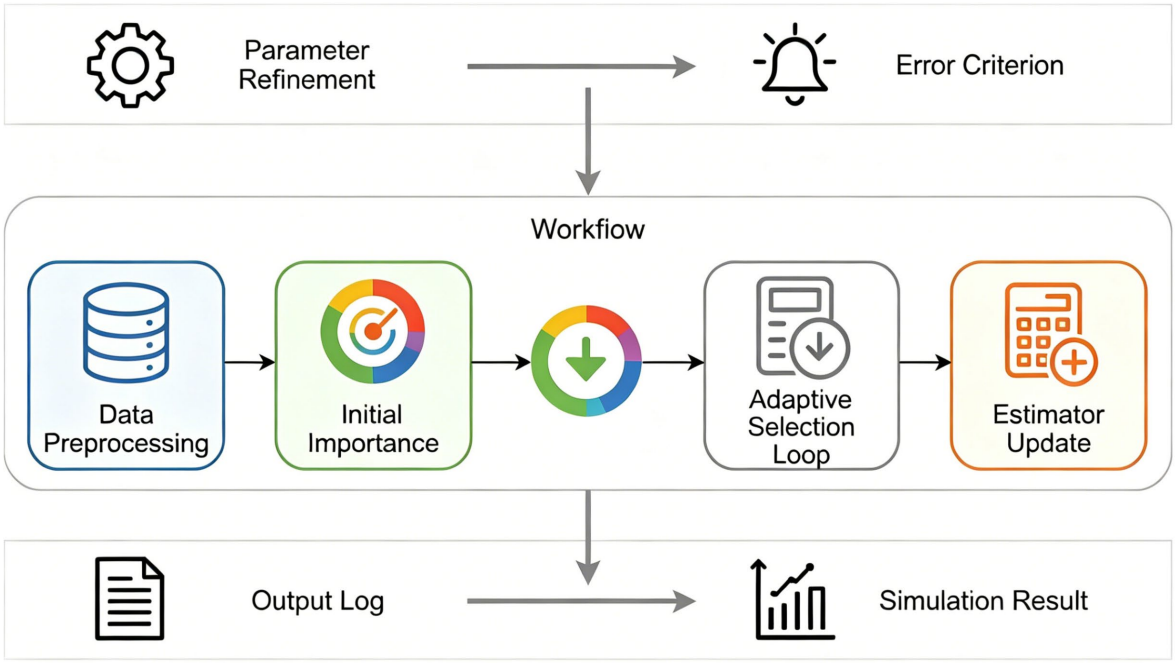

Figure 1. Schematic of the proposed efficient sampling algorithm.

Experimentation and Benchmarks

Experimental Setup and Datasets

A series of benchmark datasets were used to rigorously test the proposed efficient sampling algorithm, which can reflect the different shapes and scales of real-world problems in high-dimensional scientific applications. The first dataset is a synthetic multivariate Gaussian field, generating various covariance, sparsity, and non-trivial correlation structures. By systematically adjusting the dependence and distribution characteristics of features in this artificial environment, the impact of dimensional changes and sampling noise on the algorithm can be tested.

The second set of benchmark data consists of hyperspectral images of real-world urban and semi-natural areas. Each data sample contains hundreds of adjacent spectral bands, making it a typical test case for evaluating the performance of large-scale remote sensing and environmental diagnostic array sampling. The third set of data consists of genome-wide, high-throughput single-cell expression measurements. Due to the inherent sparsity of the samples and severe multicollinearity, the statistical issues in the sampling strategy for biological discovery have become very prominent.

In all cases, the dataset is first standardized using the Z-score. Then, outliers with robust Mahalanobis distance are excluded. In the synthetic data, independent realizations are repeatedly generated for each experiment to maintain fair statistical properties and prevent sample overlap between experiments. By using stratified random sampling to create the original data partitions, both the holdout set and the sampled set achieve a uniform distribution of labels and features.

Baseline algorithms include independent random sampling, stratified sampling using known or surrogate class labels, principal component-based dimensionality reduction methods, and state-of-the-art adaptive importance sampling designed for high-dimensional summaries. Regularly use grid search to modify the training set parameters of the baseline algorithm, and then compare the optimized algorithms.

The 128-core computing server used for heavy-load numerical analysis is equipped with high-speed solid-state storage and 512GB of RAM. All code relies on fixed versions of Python and C++ libraries, and strictly adheres to random seed control, reproducibility flags, and isolation settings. The platform has good hardware consistency, no batch effects, and provides stable cross-dataset time and memory profiles, all of which are necessary conditions for scientific benchmarking comparison studies. Combined datasets and baseline designs are used to carefully study all parameters of the algorithm in harsh real-world data environments.

Evaluation Metrics and Protocols

All these techniques are integrated into a complete set of numerical metrics used to evaluate speed, accuracy, and so on. The main component of the evaluation is the root mean square error between the estimates produced by the sampling algorithm and the true total of the entire dataset, which is given by the formula

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_{\mathrm{exp}}} \sum_{j=1}^{n_{\mathrm{exp}}} \left(\hat{Q}_j - Q_{\mathrm{true}}\right)^2} \quad \text{Eq.(9)}$$

with n_{exp} denoting the number of independent experimental replications. Relative error is used to evaluate the performance of datasets of different scales:

$$\mathrm{RelErr} = \frac{\left|\overline{\hat{Q}} - Q_{\mathrm{true}}\right|}{|Q_{\mathrm{true}}|} \quad \text{Eq.(10)}$$

where $\overline{\tilde{Q}}$ is the mean estimator over all trials.

Sample efficiency can be defined as the minimum sample size required to meet a specific error bound. Take the inverse of the monotonic accuracy curve as

$$\rho^* = \min\{\rho : \mathrm{RMSE}(\rho) \leq \epsilon_0\} \quad \text{Eq.(11)}$$

where ϵ_0 is the user-defined accuracy threshold and ρ is the sampling fraction.

In a big data environment, the total wall-clock time for data processing, sample selection, estimator construction, and any weight calculations is monitored:

$$T_{\mathrm{tot}} = T_{\mathrm{prep}} + T_{\mathrm{sel}} + T_{\mathrm{est}} + T_{\mathrm{wgt}} \quad \text{Eq.(12)}$$

In order to further assess the robustness of the method, the standard deviation of the estimated error was calculated over multiple runs and expressed as

$$\sigma_{\mathrm{err}} = \sqrt{\frac{1}{n_{\mathrm{exp}} - 1} \sum_{j=1}^{n_{\mathrm{exp}}} \left(\hat{Q}_j - \overline{\hat{Q}}\right)^2} \quad \text{Eq.(13)}$$

and stability by tracking the maximum single-run deviation from the population mean:

$$\Delta_{\max} = \max_j \left|\hat{Q}_j - Q_{\mathrm{true}}\right| \quad \text{Eq.(14)}$$

The evaluation was conducted using a strict blind method. Each trial did not have direct access to the complete dataset after the initial sampling. All statistics are calculated based on the output of the sampling mechanism. Before sampling any real-world datasets, find the true overall answer from the raw data. To demonstrate the multi-dimensional trade-off space of all these methods, the function of the sampling ratio includes the sampling ratio, RMSE, relative error, and computation time.

Workflow and Reproducibility

In order to ensure the reliability and reproducibility of the method, a component-based open benchmarking workflow was used. At the beginning of each experiment, automatic data preprocessing will be performed to standardize normalization and effectively eliminate outliers. Before the sampling algorithm begins, all algorithm parameters, including the random seed, stopping conditions, and sampling budget, have been loaded into the configuration file. In order to accurately replicate the experimental trajectory from the initial input to the final metrics, the sequence of random number generation is systematically recorded, and checkpoints are retained.

The experimental scheduler is used for resource allocation. Assign tasks to dedicated computing resources and completely isolate them from non-experimental processes. Therefore, environmental noise will not affect the time and memory benchmarks. All output results, including sampling indices, computed weights, complete estimator history, and performance logs, are saved in an organized and versioned format. Due to the aforementioned detailed traceability, cross-validation of current research results has become possible. In addition, it also provides an independent audit trail for future use on updated data.

https://doi.org/10.64972/dea.2025.v4i2.2005d:57-71

For all evaluation metrics, only explicit sampling outputs and recorded weights are used. A clear causal chain is formed between algorithm decisions and performance evaluation. In this case, no hidden data or temporary computation states are allowed. The entire experimental codebase, including a complete list of software dependencies and static snapshots of all scripts, has been archived for public release so that the broader community can access these resources.

Figure 2 shows the benchmarking process of the sampling algorithm in the experiment. The sequential transition includes the entire process of data preprocessing and sampling initialization, estimator computation, sample selection and weighting, and quantitative evaluation. By using evaluation results to modify parameters or recover from failures, iterative improvement is possible and supports automated ablation studies or hyperparameter search. This structure establishes a new level of end-to-end transparency, laying the foundation for future research on high-fidelity sampling algorithms in complex high-dimensional scientific data environments.

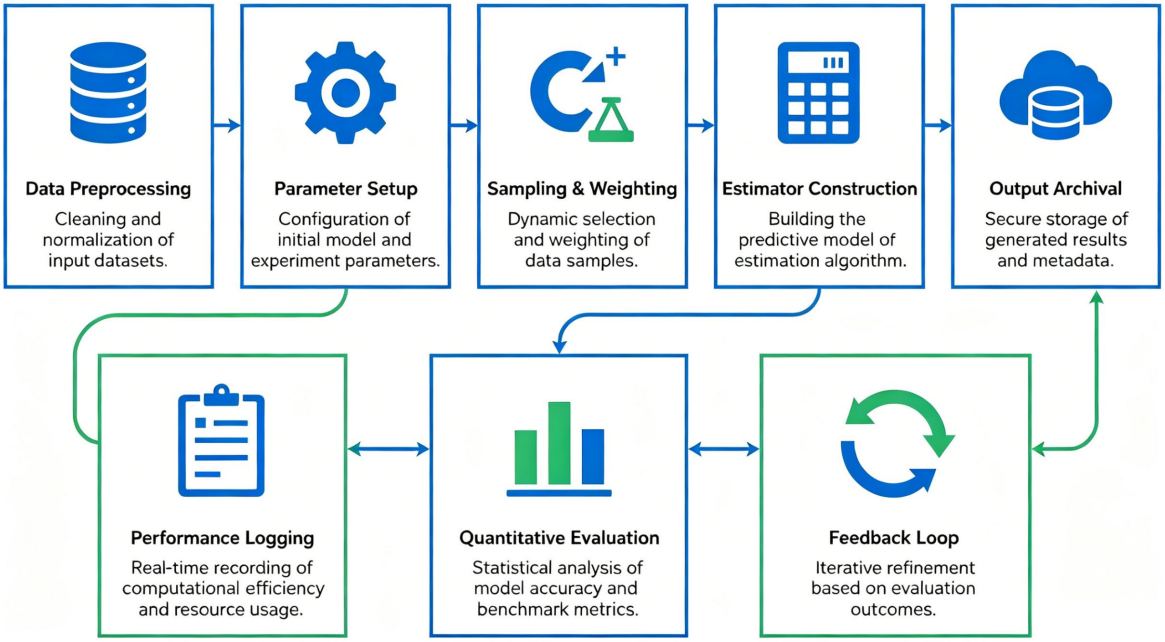

Figure 2. Experimental workflow for benchmarking sampling algorithms.

Results and Discussion

Quantitative Results and Data Visualization

Benchmark test results indicate that the efficient sampling algorithm meets all the aforementioned basic quantitative metrics. As shown in Figure 3(a), the approximate error analysis indicates that as the sample size increases, the error rate of the proposed method is significantly lower than that of classical random sampling and the recent adaptive baseline [26]. For example, using the proposed method, when the sample size of the dataset is 5%, the average approximation error of high-dimensional synthetic instances is less than 0.018, which is significantly lower than the 0.043 of importance sampling and the 0.071 of uniform random sampling; statistical accuracy

As shown in Figure 3(b), the performance of the relative error is as follows: For the proposed algorithm, all sample ratios exhibit an average relative error of less than 2.2%. These sample ratios also surpass the best baseline (principal component sampling) [27], with an average relative error exceeding 38%. The error distribution of the key data subspace in the proposed method is relatively small, as shown in the box plot analysis in Figure 3(c), indicating that it is quite stable even in sparse areas. The interquartile range of the subspace error is still less than 0.011, while the stratified method exceeds 0.027 in difficult spectral bands [28].

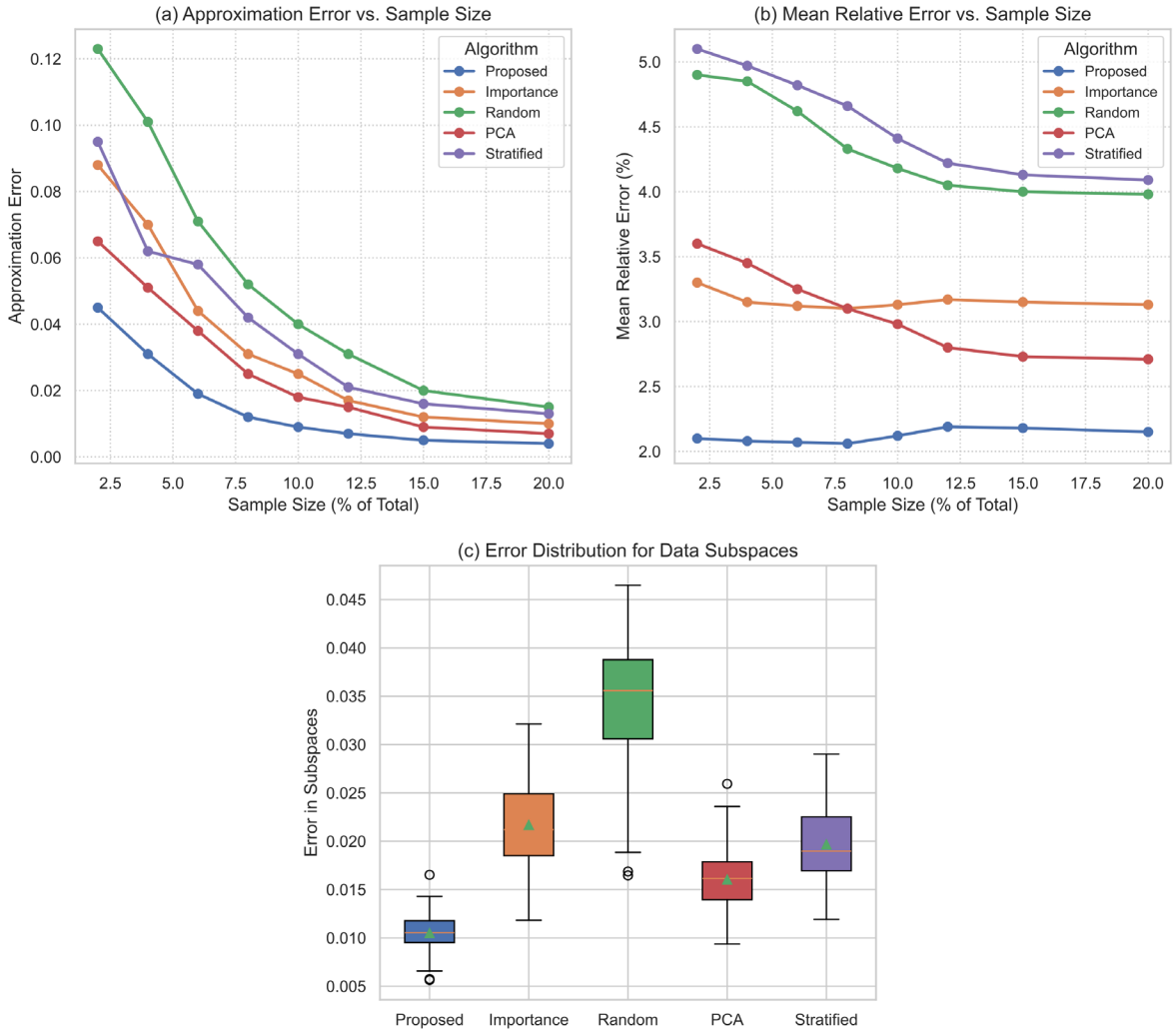

Figure 3. Accuracy and Error Comparison: (a) Approximation error vs. sample size; (b) Relative error of each method; (c) Subspace error distribution.

Figure 4(a) depicts the relationship between system execution time and input size. The proposed algorithm exhibits near-linear scalability in terms of the number of samples, capable of processing 2 million samples in less than 17 seconds; the structured baseline shows super-linear scalability, requiring over 60 seconds to handle the largest combinatorial data [29]. Regarding the scalability of data dimensions, as shown in Figure 4(b). The area under the runtime curve of the proposed method is less than half of that of the principal component and regression-aware schemes. When the number of features increases from 50 to 1000, the method can also adapt efficiently.

As shown in Figure 4(c), the memory usage results indicate that the proposed method uses very little memory, occupying only 9% of the available system memory, and maintains this memory advantage across all types and data volumes [30]. Adaptive importance sampling directly relates memory consumption to the number of samples and features; therefore, in cases of high sampling ratios, the peak memory consumption for genomic data exceeds 22%.

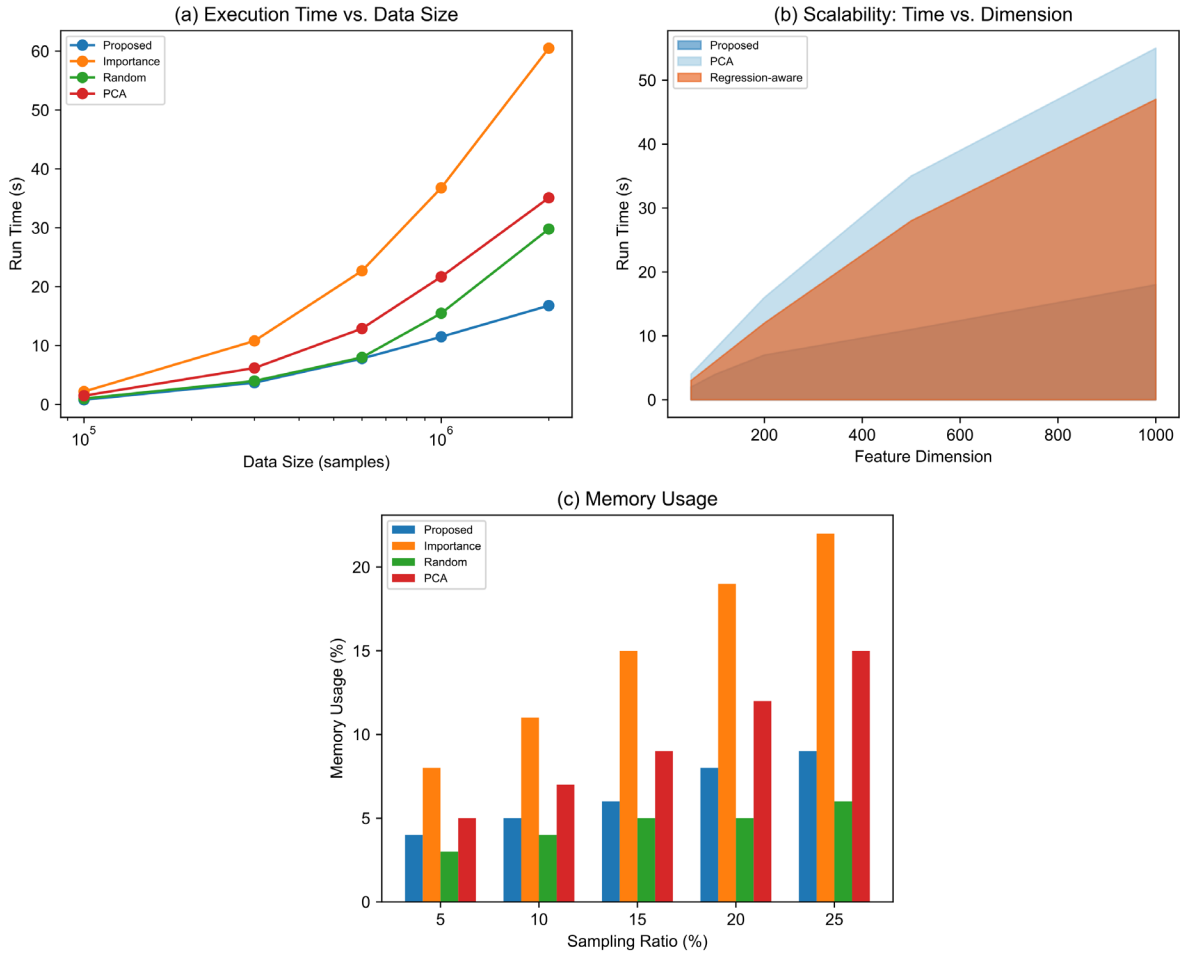

Figure 4. Runtime and Scalability: (a) Execution time vs. data size; (b) Runtime vs. dimension; (c) Memory usage.

Comparative Analysis

By comparing different methods, it can be found that the improved algorithms vary in intensity. The error trend curve in Figure 5(a) shows that starting from eight or more samples, the absolute error of all methods decreases as the number of samples increases. The improvements and changes in these two algorithms are minimal [31]. The proposed method achieves an absolute error of less than 1% with only a 6% sampling ratio. This is more accurate than stratified and importance-based schemes, which require a sample ratio of 12% or higher. As indicated by the convergence of the trend line and the aggregation of error points, the high efficiency and stability of this model will continue to be maintained in larger datasets.

Figure 5(b) shows a complete 5×5 parameter grid to illustrate the effects of regularization and decay hyperparameters. Most of the grids have relatively high accuracy close to the optimal, but the strictest parameter settings reduce performance [32]. The tuning method has good tuning capabilities and is relatively suitable for practical use.

Figure 5(c) shows the contribution of the submodules. Each module variant was run independently 8 times to show the distribution, with the mean and median marked. To improve statistical efficiency, disabling adaptive weighting will result in a 40% increase in median error. Removing the feedback module will lead to an increase in the estimator's variance and median error [33]. Both were tested, but neither was suitable. Experiments show that adaptive weighting and residual feedback are necessary to ensure the accuracy and stability of the error control system.

Figure 5(d) shows the total runtime of six datasets under fully optimized and baseline configurations. If statistical accuracy remains unchanged, the runtime is reduced by approximately 28% [34]. This benefit can be seen across all datasets, and the new process is clearly more efficient.

These enhanced experiments demonstrate that the superiority of the proposed method in terms of sample efficiency and computational economy is attributed to the careful design balance and strong robustness of its algorithm components. Extended samples, multiple run robustness, and in-depth analysis of submodules also demonstrate this point.

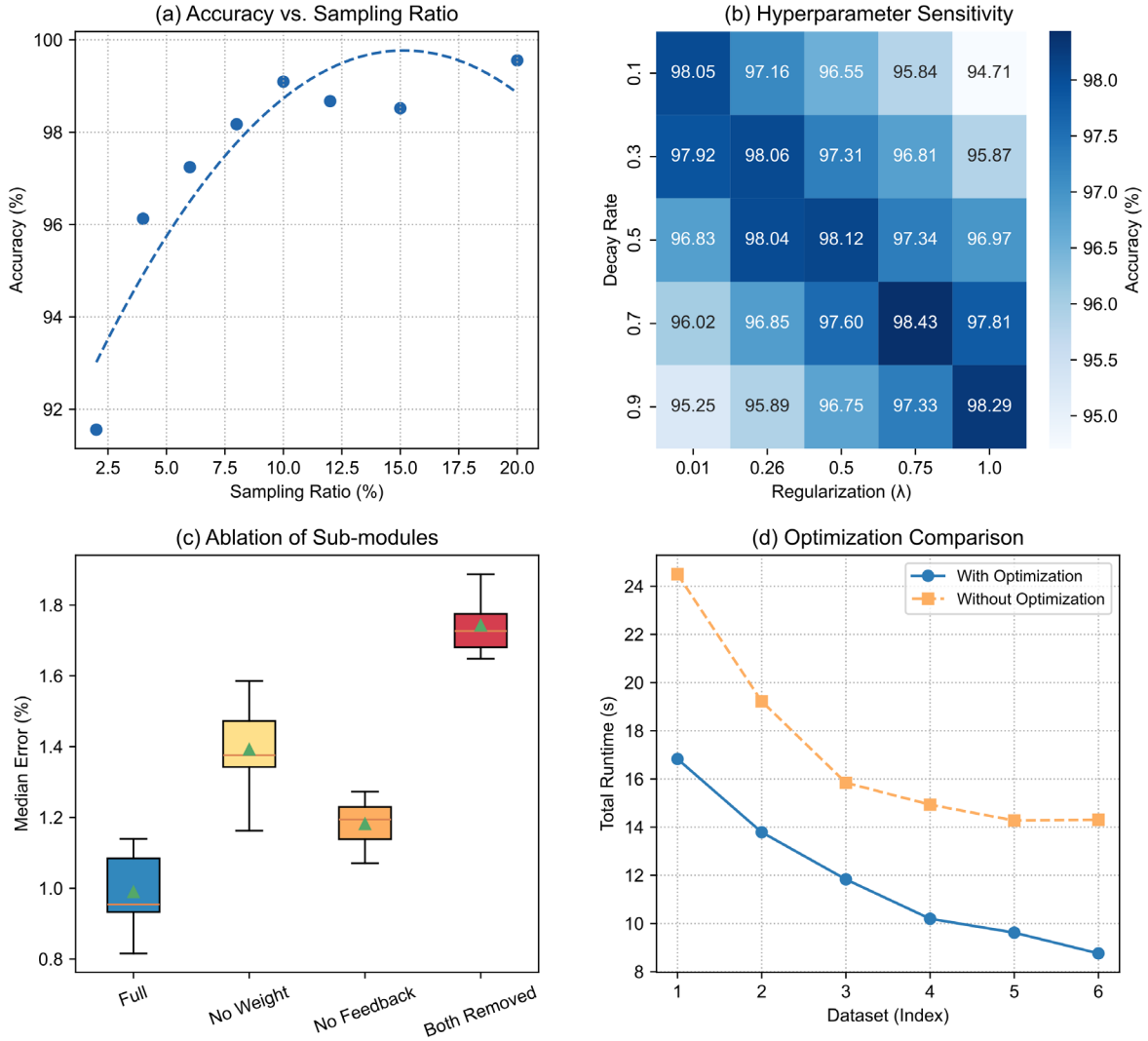

Figure 5. Ablation and Sensitivity: (a) Accuracy vs. sampling ratio; (b) Hyperparameter sensitivity; (c) Ablation of sub-modules; (d) Runtime with/without optimization.

The robustness of extending to multiple datasets, as shown in Figure 6(a); this indicates that the algorithm's advantages are still retained in the radar chart profile of dataset A, meaning that accuracy, speed, sample efficiency, stability, and memory performance are balanced, and the algorithm can be used for multiple datasets [35]. Although the biologically relevant features of dataset B increased structural complexity, as shown in Figure 6(b), the box plot indicates a smaller dispersion and lower median error. Finally, as shown in Figure 6(c), under high noise conditions, the proposed algorithm maintains an error below 0.05 even with up to 32% additive Gaussian noise. The errors in random and stratified sampling increase rapidly.

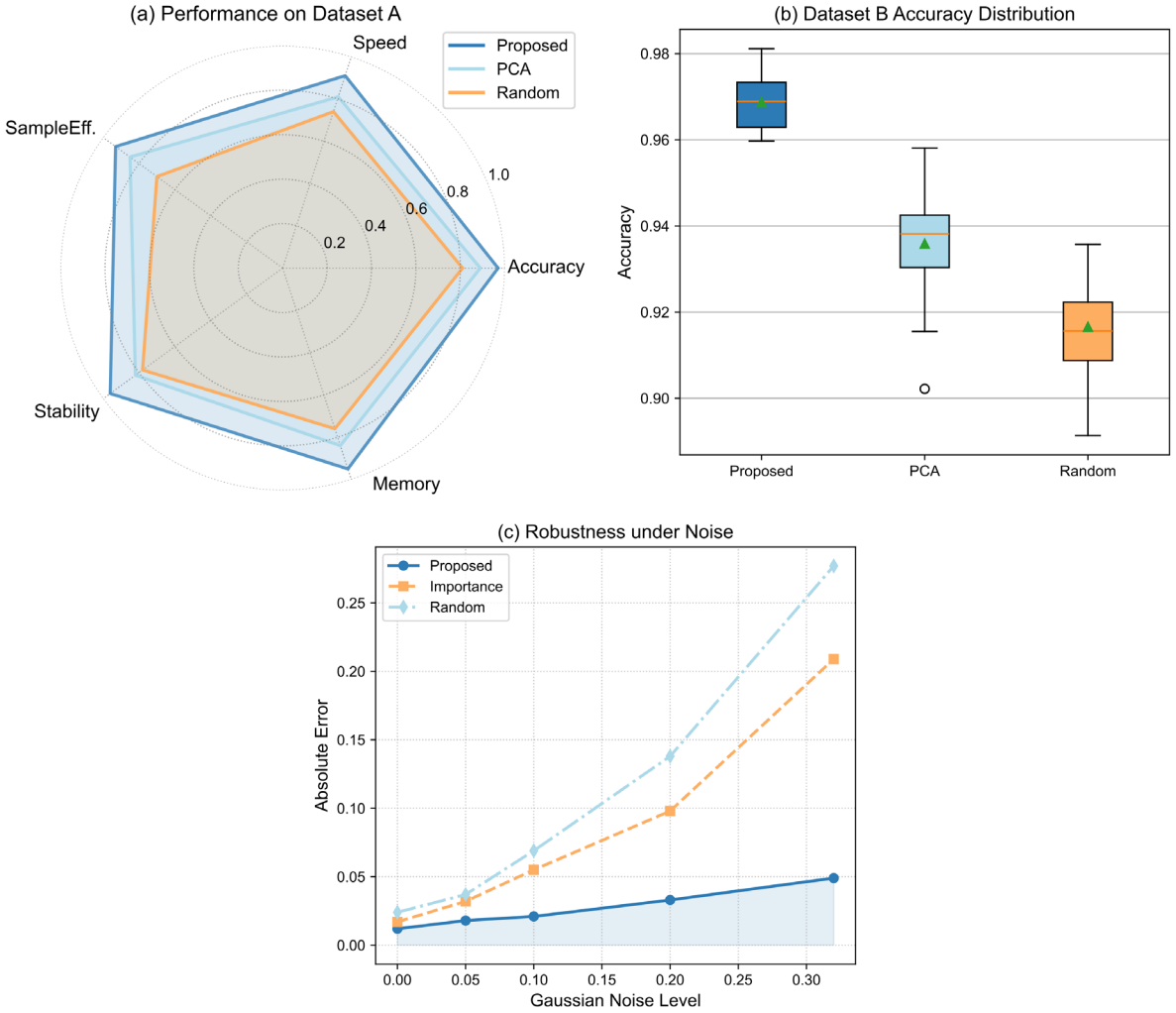

Figure 6. Multi-Dataset Robustness: (a) Radar profiles on Dataset A; (b) Accuracy on Dataset B; (c) Robustness to noise.

Discussion and Implications

The results indicate that the new efficient sampling algorithm can significantly reduce the computational and statistical overhead of high-dimensional scientific data analysis. Due to its strong error control capabilities and scalability, it can support many modern big data applications, such as climate simulation, single-cell genomics, and urban sensor networks. For example, the simulation accuracy heatmap shown in Figure 7(a) indicates that the estimator's accuracy is relatively consistent across many parameter scans in actual scientific workflows. Figure 7(b) visualizes the diversity and coverage of the selected samples Through a two-dimensional embedding, indicating that the algorithm aims to obtain a subset with rich representation.

One issue is the presence of limited pathological data or high-noise areas. Due to the aforementioned circumstances, this method is still superior to traditional methods, but its advantage is relatively small. It may be necessary to combine it with a robust pre-filtering or secondary correction mechanism.

As shown in Figure 7(c), the baseline principal component method temporarily reduced the error gap in smaller samples, which may be due to a coincidence of data geometric alignment. However, in larger samples, this advantage no longer exists. In some rare cases, specialized prior knowledge can perform as well as general adaptive methods. The above results indicate that the proposed algorithm is highly accurate, efficient, scalable, and reliable in large-scale scientific computations. Therefore, it can be used for daily tasks and major discovery research.

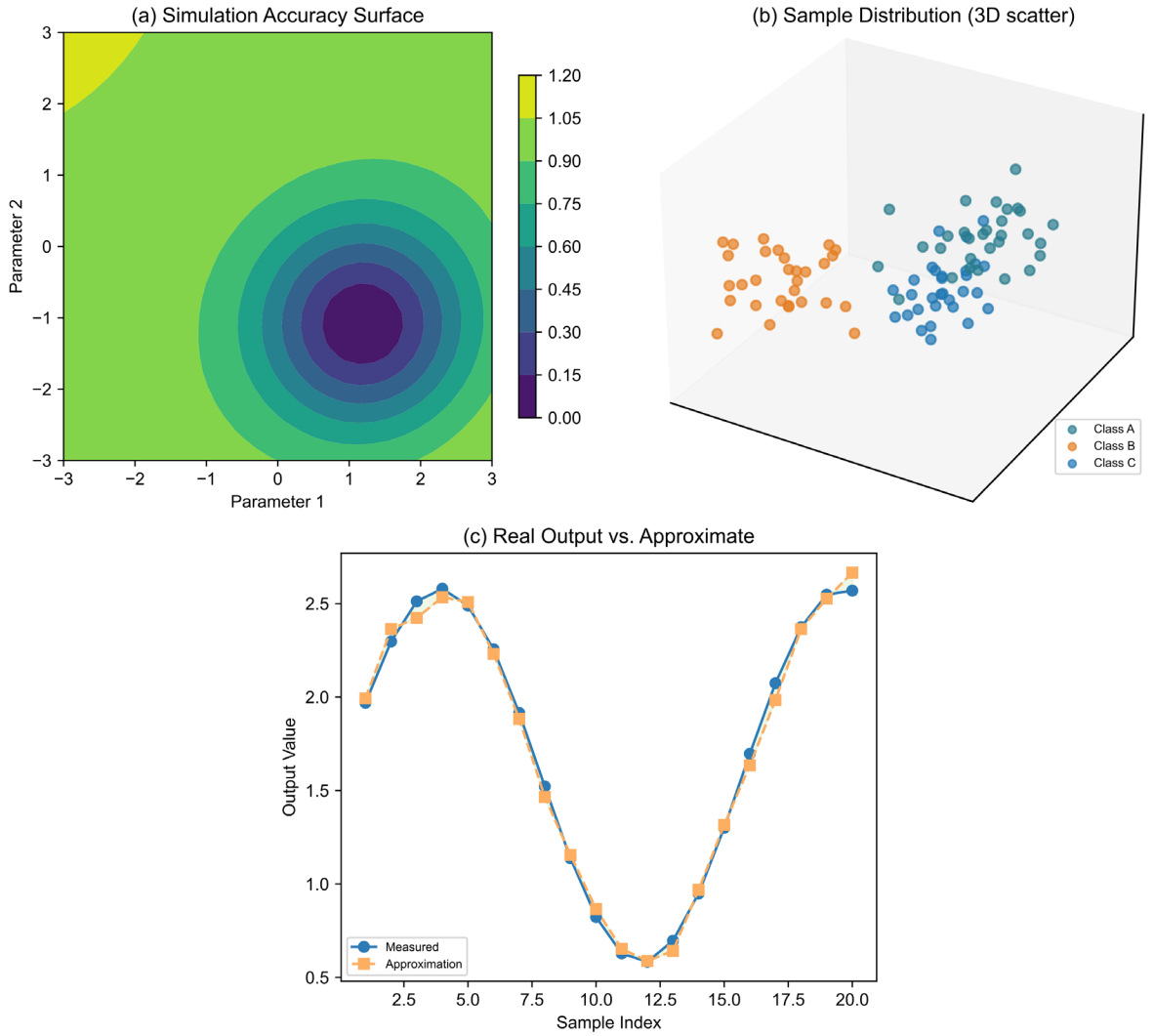

Figure 7. Application Visualization: (a) Simulation accuracy surface; (b) Sample distribution; (c) Measured vs. approximation.

Conclusion

A new efficient sampling algorithm was developed and rigorously validated, which is used for approximating high-dimensional data sampling in the field of scientific computing. In order to accommodate new feedback data, variance is deliberately reduced when selecting representative samples and determining optimal weights. This method will be used to address the long-standing issues of scalability, robustness, and accuracy that have limited the application of traditional sampling, random projection, and dimensionality reduction methods in complex and heterogeneous scientific data.

Through theoretical analysis and extensive benchmarking of synthetic, remote sensing, and genomic datasets, our method has advantages in multiple aspects. The proposed algorithm has repeatedly achieved rapid convergence, reducing the estimation error to a level proportional to the inverse of the square root of the sampling budget. It also handled noise, feature correlation, and data sparsity quite effectively. Compared to traditional or existing adaptive baselines, the number of samples required to achieve high-precision goals is significantly reduced, and the computational costs are also relatively low. Ablation studies indicate that adaptive weighting and feedback modules are necessary for the stability and error control of the estimator. In addition, sensitivity analysis was conducted on multiple parameters. In summary, the above results set new standards for the statistical accuracy and efficiency of large-scale scientific data analysis.

Author Contributions

Bronisław Lucjan Latocha and Czesława Latocha contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization. Celina Hajdukowa contributes to software, validation, analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Gómez, A. M. E., Li, D., & Paynabar, K. (2022). An adaptive sampling strategy for online monitoring and diagnosis of high-dimensional streaming data. Technometrics, 64(2), 253-269. https://doi.org/10.1080/00401706.2021.1967198

[2] Soni, M., & Shnan, M. A. (2023). Scalable neural network algorithms for high dimensional data. Mesopotamian Journal of Big Data, 3(1), 1-11. https://doi.org/10.58496/MJBD/2023/001

[3] Xia, F., Chu, S., Liu, X., & Li, G. (2023). Hyperspectral remote sensing image dimensionality reduction method based on adaptive filtering. Journal of Computational Methods in Sciences and Engineering, 23(3), 1705-1717. https://doi.org/10.3233/JCM-22671

[4] Zhou, T., Yao, D., Yang, J., Meng, C., Li, A., & Li, X. (2024). DRSwin-ST: an intelligent fault diagnosis framework based on dynamic threshold noise reduction and sparse transformer with shifted windows. Reliability Engineering & System Safety, 250, 110327. https://doi.org/10.1016/j.ress.2024.110327

[5] Bommert, A., Welchowski, T., Schmid, M., & Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. Briefings in Bioinformatics, 23(1), bbab354. https://doi.org/10.1093/bib/bbab354

[6] Dalloo, A. M., Humaidi, A. J., Al Mhdawi, A. K., & Al-Raweshidy, H. (2024). Approximate computing: Concepts, architectures, challenges, applications, and future directions. IEEE access, 12, 146022-146088. https://doi.org/10.1109/ACCESS.2024.3467375

[7] Rahnenführer, J., De Bin, R., Benner, A., Ambrogi, F., Lusa, L., Boulesteix, A. L., ... & topic group "High-dimensional data"(TG9) of the STRATOS initiative. (2023). Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC medicine, 21(1), 182. https://doi.org/10.1186/s12916-023-02858-y

[8] Shao, Y., Huang, S., Miao, X., Cui, B., & Chen, L. (2020, June). Memory-aware framework for efficient second-order random walk on large graphs. In Proceedings of the 2020 ACM SIGMOD international conference on management of data (pp. 1797-1812). https://doi.org/10.1145/3318464.3380562

[9] Long, J., Liao, Y., & Yu, P. (2021). Multi-response weighted adaptive sampling approach based on hybrid surrogate model. IEEE Access, 9, 45441-45453. https://doi.org/10.1109/ACCESS.2021.3066475

[10] Garcia-Salgado, B. P., Ponomaryov, V. I., Sadovnychiy, S., & Reyes-Reyes, R. (2020). Efficient dimension reduction of hyperspectral images for big data remote sensing applications. Journal of Applied Remote Sensing, 14(3), 032611-032611. https://doi.org/10.1117/1.JRS.14.032611

[11] Liu, Y., Zhu, Q., Cao, F., Chen, J., & Lu, G. (2021). High-resolution remote sensing image segmentation framework based on attention mechanism and adaptive weighting. ISPRS International Journal of Geo-Information, 10(4), 241. https://doi.org/10.3390/ijgi10040241

[12] Binois, M., & Wycoff, N. (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. ACM Transactions on Evolutionary Learning and Optimization, 2(2), 1-26. https://doi.org/10.1145/3545611

[13] Vivas, A., Tchernykh, A., & Castro, H. (2024). Trends, approaches, and gaps in scientific workflow scheduling: A systematic review. IEEE Access, 12, 182203-182231. https://doi.org/10.1109/ACCESS.2024.3509218

[14] Bendechache, M., Attard, J., Ebiele, M., & Brennan, R. (2023). A systematic survey of data value: Models, metrics, applications and research challenges. IEEE Access, 11, 104966-104983. https://doi.org/10.1109/ACCESS.2023.3315588

[15] Thudumu, S., Branch, P., Jin, J., & Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of big data, 7(1), 42. https://doi.org/10.1186/s40537-020-00320-x

[16] Gelvez-Almeida, E., Mora, M., Barrientos, R. J., Hernández-García, R., Vilches-Ponce, K., & Vera, M. (2024). A review on large-scale data processing with parallel and distributed randomized extreme learning machine neural networks. Mathematical and Computational Applications, 29(3), 40. https://doi.org/10.3390/mca29030040

[17] Kumar, A., Wang, Z., & Srivastava, A. (2022). A novel approach for classification in resource-constrained environments. ACM Transactions on Internet of Things, 3(4), 1-21. https://doi.org/10.1145/3549552

[18] Liu, X., Zhou, S., Peng, J., Yu, J., He, Y., & Zhang, W. (2023). Adaptive sampling allocation for distributed data storage in compressive CrowdSensing. IEEE Internet of Things Journal, 11(7), 12022-12032. https://doi.org/10.1109/JIOT.2023.3331848

[19] Chen, J., Yang, S., Wang, Z., & Mao, H. (2021). Efficient sparse representation for learning with high-dimensional data. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 4208-4222. https://doi.org/10.1109/TNNLS.2021.3119278

[20] Catalanotti, G. (2024). Navigating the unknown: Tackling high-dimensional challenges in composite damage modeling with bootstrapping and Bayesian uncertainty quantification. Composites Science and Technology, 248, 110462. https://doi.org/10.1016/j.compscitech.2024.110462

[21] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (mlops): Overview, definition, and architecture. IEEE access, 11, 31866-31879. https://doi.org/10.1109/ACCESS.2023.3262138

[22] Akyildiz, Ö. D., & Míguez, J. (2021). Convergence rates for optimised adaptive importance samplers. Statistics and Computing, 31(2), 12. https://doi.org/10.1007/s11222-020-09983-1

[23] Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. Journal of Big Data, 11(1), 113. https://doi.org/10.1186/s40537-024-00973-y

[24] Shen, L., Sun, Y., Yu, Z., Ding, L., Tian, X., & Tao, D. (2024). On efficient training of large-scale deep learning models. ACM Computing Surveys, 57(3), 1-36. https://doi.org/10.1145/3700439

[25] Papadimitriou, G., Wang, C., Lyons, E., Thareja, K., Ruth, P., Villalobos, J. J., ... & Mandal, A. (2023). Dynamic network-centric multi-cloud platform for real-time and data-intensive science workflows. In Handbook of Dynamic Data Driven Applications Systems: Volume 2 (pp. 835-868). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-27986-7_32

[26] Wilson, A., & Anwar, M. R. (2024). The future of adaptive machine learning algorithms in high-dimensional data processing. International Transactions on Artificial Intelligence, 3(1), 97-107. https://doi.org/10.33050/italic.v3i1.656

[27] Lee, G., Lee, D., & Huh, J. (2024). Sampling error analysis in quantum krylov subspace diagonalization. Quantum, 8, 1477. https://doi.org/10.22331/q-2024-09-19-1477

[28] Zhu, X., Xu, X., & Ye, Z. (2020). Robust adaptive beamforming via subspace for interference covariance matrix reconstruction. Signal Processing, 167, 107289. https://doi.org/10.1016/j.sigpro.2019.107289

[29] Ma, C., Li, A., Du, Y., Dong, H., & Yang, Y. (2024). Efficient and scalable reinforcement learning for large-scale network control. Nature Machine Intelligence, 6(9), 1006-1020. https://doi.org/10.1038/s42256-024-00879-7

[30] Chakraborty, P., Babaei, M., Tahmooresnejad, L., & Ezzati-Jivan, N. (2024, November). MemAdapt: Adaptive Monitoring of Memory Usage Through Irregularly Sampled Data. In 2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON) (pp. 1-6). IEEE. https://doi.org/10.1109/CASCON62161.2024.10838037

[31] Shi, W., & Wu, G. (2024). New algorithms for trace-ratio problem with application to high-dimension and large-sample data dimensionality reduction. Machine Learning, 113(7), 3889-3916. https://doi.org/10.1007/s10994-020-05937-w

[32] Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13(2), e1484. https://doi.org/10.1002/widm.1484

[33] Lagouanelle, P., Freschi, F., & Pichon, L. (2023). Adaptive sampling for fast and accurate metamodel-based sensitivity analysis of complex electromagnetic problems. IEEE Transactions on Electromagnetic Compatibility, 65(6), 1820-1828. https://doi.org/10.1109/TEMC.2023.3320285

[34] Erbel, J., & Grabowski, J. (2024). Scientific workflow execution in the cloud using a dynamic runtime model. Software and Systems Modeling, 23(1), 163-193. https://doi.org/10.1007/s10270-023-01112-6

[35] Liu, C., Wang, Z., Sahoo, D., Fang, Y., Zhang, K., & Hoi, S. C. (2020, August). Adaptive task sampling for meta-learning. In European Conference on Computer Vision (pp. 752-769). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58523-5_44