CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 1435-1450
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Geometry-Based Differentially Private Synthetic Tabular Data Generation via K-Means Clustering with Bounded and Discrete Feature Constraints

Robby*, Agus Sukmana, and Erwinna Chendra

Center for Mathematics and Society, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia

Abstract

Most clustering-based differentially private synthetic data generation methods assume unconstrained continuous feature spaces and offer no mechanism for hard feature bound enforcement or discrete-valued attribute handling, which limits their practical applicability to real-world tabular data where such constraints are common. This paper proposes a geometry-based mechanism that generates synthetic tabular data by application of Laplace noise jointly to K-means cluster centroids and within-cluster radial distances, calibrated via a data-dependent sensitivity approximation. Three components distinguish the approach from prior work: coordinate-wise centroid reflection to enforce hard feature bounds after perturbation, coordinate-wise clipping to enforce bounds on reconstructed synthetic points, and randomized rounding for discrete features as a post-processing step. A utility-driven calibration strategy selects the privacy budget ε to meet a user-specified target Adjusted Rand Index (ARI), which makes the privacy–utility trade-off directly interpretable. Baseline comparisons on a two-dimensional illustrative example show that the proposed mechanism achieves ARI $= 0.666$ at $\varepsilon \approx 1.60$, which substantially outperforms direct coordinate-wise noise addition at the same budget (ARI $= 0.199$), while it matches the non-private synthesis baseline (ARI $= 0.624$). Across 30 independent runs the mechanism achieves mean ARI $= 0.629 \pm 0.108$, which confirms that the calibration target is reliably met under stochastic variation.

Keywords: ARI-guided Calibration; Bounded Features; Differential Privacy; K-means Clustering; Synthetic Tabular Data.

1. Introduction

The increased availability of sensitive data in domains such as healthcare, recommendation systems, and distributed computing has raised significant concerns about privacy preservation during data sharing and analysis. Differential privacy (DP) has emerged as a rigorous and widely accepted framework for the limitation of information leakage while it enables statistical inference and machine learning on sensitive datasets.

Clustering-based methods, particularly K-means and its variants, are widely used in privacy-preserving data analysis because of their simplicity, scalability, and interpretability. A substantial body of work has investigated differentially private K-means clustering under various threat

*Corresponding author. E-mail: robby@unpar.ac.id

Submitted: March 02, 2026 Reviewed: May 21, 2026 Accepted: June 06, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.41403

models and computational settings, including centralized, local, and distributed environments [1–4], showing that meaningful clustering structure can be retained through mechanisms such as distance perturbation, centroid noise injection, and exponential mechanisms.

Beyond clustering itself, differentially private synthetic data generation has drawn increased attention as a way to share data while the disclosure risk is reduced. Several methods use clustering as an intermediate representation and build synthetic data from privatized cluster summaries [5–7]. Other approaches apply DP to generative models: Jordon, Yoon, and Schaar [8] proposed PATE-GAN for synthetic tabular data with formal privacy guarantees, while Park et al. [9] introduced a DP tabular synthesizer that preserves marginal distributions under a fixed privacy budget. Broader studies also highlight challenges of balance between privacy, utility, and interpretability in deep learning and federated settings [10–13].

Despite these advances, a gap remains in how existing clustering-based synthesis methods handle practical data constraints. MC-GEN [5] and the co-clustering approach of Benkhelif et al. [6] assume unconstrained continuous feature spaces and provide no mechanism for hard feature bound enforcement or discrete-valued attribute handling. PrivSyn [7] supports mixed data types through marginal-based synthesis, but does not use cluster geometry as the generative scaffold and does not offer a utility-driven budget selection strategy. When these constraints are ignored, synthetic records may violate feasibility requirements or contain invalid values, which reduces practical utility even when formal privacy guarantees hold. Additionally, existing methods require practitioners to set ε manually without a direct connection to a downstream quality target [14].

This paper proposes a geometry-based privacy-preserving synthetic data generation mechanism built upon K-means clustering that explicitly addresses these limitations. The main methodological contributions are: (i) a joint Laplace perturbation of cluster centroids and within-cluster radial distances combined with coordinate-wise centroid reflection and point clipping to enforce hard feature bounds during synthesis — a combination absent from prior clustering-based DP synthesis methods; (ii) a utility-driven calibration strategy that selects the privacy budget ε to meet a user-specified target Adjusted Rand Index, which makes the privacy–utility trade-off directly interpretable without manual ε tuning; and (iii) randomized rounding as a post-processing step for discrete features, which preserves integrality at no additional privacy cost under the post-processing property of differential privacy. A low-dimensional numerical example illustrates the method and analyzes clustering stability through the Adjusted Rand Index and cluster label changes. Baseline comparisons against non-private synthesis and direct coordinate-wise noise addition are included to contextualize the utility of the proposed mechanism.

2. Methodology

This section presents the methodological foundation of the proposed synthetic data generation mechanism. It begins with the basic concepts of differential privacy and the Laplace mechanism, followed by K-means clustering and the evaluation metrics used to assess the preservation of clustering structure.

2.1. Differential Privacy

Differential privacy (DP), first formalized by Dwork et al. [15], provides a formal guarantee that the output of a randomized algorithm is insensitive to the contribution of any single data record. Formally, a randomized mechanism $\mathcal{M}$ satisfies ε-differential privacy if, for any pair of neighboring datasets D and D' and for any measurable set $\mathcal{S}$,

$$\mathbb{P}[\mathcal{M}(D) \in \mathcal{S}] \leqslant e^{\varepsilon}\, \mathbb{P}[\mathcal{M}(D') \in \mathcal{S}].$$

The datasets D and D' are said to be *neighboring*, denoted by $D \sim D'$, if they differ in the presence or value of exactly one data record, i.e.,

$$|D \triangle D'| = 1,$$

where $\triangle$ denotes the symmetric difference operator. The privacy parameter $\varepsilon > 0$ controls the strength of the privacy guarantee, with smaller values corresponding to stronger protection.

In this work, the mechanism is designed within the central (trusted curator) model: all computations on the private dataset are performed internally and only the outputs of the perturbation procedure are released.

Theorem 1 (Sequential Composition). *Let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be randomized mechanisms such that $\mathcal{M}_i$ satisfies ε_i-differential privacy for each $i \in \{1, 2, \ldots, m\}$. Then the mechanism defined by the sequential release*

$$\mathcal{M}(D) = (\mathcal{M}_1(D), \ldots, \mathcal{M}_m(D))$$

satisfies $(\varepsilon_1 + \cdots + \varepsilon_m)$-differential privacy.

This composition result is central to the proposed synthetic data generation procedure. It allows the total privacy budget to be decomposed across multiple randomized queries on the private dataset.

These concepts — the DP definition, neighboring-dataset relation, Laplace mechanism, and sequential composition theorem — form the theoretical foundation of the proposed mechanism. The present work applies Laplace noise decomposed via sequential composition across centroid and radial-distance queries under the trusted curator model, but does not establish a formal end-to-end ε-DP proof: the sensitivity is a data-dependent approximation rather than a derived bound, and the adaptive budget-selection step has not been formally analyzed. The DP framework above should therefore be read as a motivating foundation rather than a guarantee that applies directly to the full mechanism.

2.2. Laplace Mechanism

A standard approach to the achievement of differential privacy is the Laplace mechanism. Given a function $f(D) \in \mathbb{R}^k$ with ℓ_1-sensitivity

$$\Delta_1 f = \max_{D \sim D'} \|f(D) - f(D')\|_1,$$

the Laplace mechanism releases $\mathcal{M}(D) = f(D) + \eta$, where each component of η is drawn independently from a Laplace distribution with scale parameter $\Delta_1 f / \varepsilon$.

2.3. K-Means Clustering

Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ denote the original dataset. K-means clustering partitions the data into K clusters by minimization of the within-cluster sum of squared distances,

$$\sum_{j=1}^{K} \sum_{x_i \in C_j} \|x_i - \mu_j\|_2^2,$$

where μ_j denotes the centroid of cluster C_j. The resulting centroids provide a compact geometric summary of the data.

2.4. Evaluation Metrics

To assess the extent to which clustering structure is preserved in the synthetic data, we use both label-based and distribution-based measures.

2.4.1. Adjusted Rand Index (ARI)

Let a dataset of n observations be clustered in two different ways, which yields partitions $\mathcal{P} = \{P_1, \ldots, P_r\}$ and $\mathcal{Q} = \{Q_1, \ldots, Q_s\}$. Let

$$n_{ij} = |P_i \cap Q_j|, \quad a_i = \sum_{j=1}^{s} n_{ij}, \quad \text{and} \quad b_j = \sum_{i=1}^{r} n_{ij}.$$

The adjusted Rand index (ARI) is defined as

$$\mathrm{ARI} = \frac{\displaystyle\sum_{i=1}^{r}\sum_{j=1}^{s}\binom{n_{ij}}{2} - \frac{\displaystyle\sum_{i=1}^{r}\binom{a_i}{2}\sum_{j=1}^{s}\binom{b_j}{2}}{\binom{n}{2}}}{\displaystyle\frac{1}{2}\left[\sum_{i=1}^{r}\binom{a_i}{2} + \sum_{j=1}^{s}\binom{b_j}{2}\right] - \frac{\displaystyle\sum_{i=1}^{r}\binom{a_i}{2}\sum_{j=1}^{s}\binom{b_j}{2}}{\binom{n}{2}}}.$$

The ARI takes values in $[-1, 1]$, where a value of 1 indicates identical clusterings, 0 corresponds to agreement expected by chance, and negative values indicate less agreement than expected under random assignment.

2.4.2. Confusion Matrix

To complement the adjusted Rand index, a confusion matrix provides a cluster-wise comparison between two labelings. Given original cluster labels $\mathcal{P} = \{P_1, \ldots, P_r\}$ and synthetic cluster labels $\mathcal{Q} = \{Q_1, \ldots, Q_s\}$, the confusion matrix $M \in \mathbb{Z}_{\geqslant 0}^{r \times s}$ is defined by

$$M_{ij} = |P_i \cap Q_j|,$$

that is, the number of data points assigned to cluster P_i in the original data and to cluster Q_j in the synthetic data. Each row of the confusion matrix describes how the points from a given original cluster are redistributed across synthetic clusters. Unlike ARI, which summarizes agreement into a single scalar value, the confusion matrix provides an interpretable view of cluster-level distortions induced by the privacy mechanism.

3. Synthetic Data Generation Procedure

Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ denote the original dataset. The following inputs are assumed to be specified:

- the number of clusters K,
- a target adjusted Rand index $\mathrm{ARI}_{\mathrm{target}} \in (0, 1)$,
- lower and upper bounds for each feature, and
- feature types (continuous or discrete).

The goal is to generate a synthetic dataset $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$ that preserves cluster structure up to the desired ARI level while Laplace perturbation is applied, calibrated through the sensitivity approximation described below.

Prior to clustering, the dataset X is transformed feature-wise by standardization. Let $\bar{x}_j$ and σ_j denote the empirical mean and standard deviation of feature j. Each data point is mapped to a standardized representation

$$x_i^{(s)} = \left(\frac{x_{i1} - \bar{x}_1}{\sigma_1}, \frac{x_{i2} - \bar{x}_2}{\sigma_2}, \ldots, \frac{x_{id} - \bar{x}_d}{\sigma_d}\right).$$

All subsequent clustering, distance computations, and noise injection are performed in the standardized space. After synthetic data generation, the inverse transformation is applied to recover data on the original scale.

3.1. Private Cluster Representation

Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ denote a private (standardized) dataset held by a trusted data curator. The proposed synthetic data generation mechanism is designed within the central (trusted curator) model: all computations on X are performed internally, and only the output of the perturbation procedure parameterized by a privacy budget ε is released.

Given a fixed privacy budget $\varepsilon > 0$, K-means clustering is first applied to X to obtain cluster assignments $\{C_1, \ldots, C_K\}$ and corresponding centroids $\{\mu_1, \ldots, \mu_K\}$. These quantities are treated as intermediate representations and are never released. The mechanism represents each cluster using two classes of statistics:

- cluster centroids, which encode the translational location of each cluster, and
- radial distances of individual data points to their assigned centroids, which encode within-cluster dispersion.

3.1.1. Sensitivity Calibration

To apply the Laplace mechanism, the ℓ_1-sensitivity of each released quantity must be specified. In this work, the sensitivity is set to the maximum pairwise Euclidean distance between cluster centroids in standardized space,

$$\Delta = \max_{j \neq k} \|\mu_j - \mu_k\|_2.$$

This choice serves as a data-dependent proxy sensitivity used to calibrate noise for both the centroid and radial distance perturbations. A formal sensitivity derivation that accounts for the full K-means update is left as future work; the use of Δ constitutes a practical calibration heuristic under the assumption that the cluster structure is stable across single-record changes.

3.1.2. Neighboring Datasets

Under the add/remove neighboring-dataset definition, $D \sim D'$ if they differ in exactly one record. The addition or removal of a single point can shift all K centroids, alter assignments, and change radial distances within the affected cluster. A tight closed-form worst-case sensitivity for K-means centroids is difficult to establish, as the centroid shift depends on cluster sizes, data geometry, and which point changes [2]. The approximation Δ is motivated by the observation that a single record can displace a centroid by at most the cluster radius, which is bounded by the maximum pairwise centroid distance; radial distances $r_i = \|x_i - \mu_k\|_2$ are similarly bounded by Δ under the same stability assumption. This approximation holds well when clusters are large and well separated, but may underestimate the true sensitivity when clusters are small or near-balanced. Accordingly, Δ serves as a practical uniform sensitivity proxy throughout this work, with the understanding that it is a heuristic rather than a rigorous global bound.

3.1.3. Noise Injection

To provide privacy-preserving perturbation motivated by differential privacy principles, Laplace noise is injected into both components. The total privacy budget ε is decomposed as

$$\varepsilon = \varepsilon_\mu + \varepsilon_r, \qquad \varepsilon_\mu = \varepsilon_r = \varepsilon/2.$$

For each cluster k, Laplace noise with scale Δ/ε_μ is added independently to each coordinate of the centroid μ_k, which yields a noisy centroid $\tilde{\mu}_k$. Similarly, for each data point $x_i \in C_k$, Laplace noise with scale Δ/ε_r is added to its radial distance

$$r_i = \|x_i - \mu_k\|_2,$$

which produces a noisy distance $\tilde{r}_i$.

Under the sequential composition theorem, if Δ were a valid global ℓ_1-sensitivity for both query types, the joint release of noisy centroids and noisy distances would satisfy ε-differential privacy. As noted in Section 3.1.1, however, Δ is a data-dependent heuristic approximation; a formal privacy guarantee conditional on this approximation is assumed throughout but not proved. These noisy quantities fully determine the subsequent synthetic data reconstruction procedure.

3.2. ARI-Guided Calibration of the Privacy Budget

The privacy parameter ε determines the magnitude of noise injected into the noisy cluster representation described in Section 3.1 and therefore controls the trade-off between privacy protection and clustering fidelity. Rather than a fixed ε a priori, we adopt a utility-driven calibration strategy in which the desired level of clustering similarity is specified through a target Adjusted Rand Index (ARI).

Let $c = (c_1, \ldots, c_n)$ denote the cluster labels obtained by application of K-means to the original dataset X, and let $\tilde{c}^{(\varepsilon)}$ denote the cluster labels obtained by reclustering of the synthetic dataset generated by the mechanism parameterized by ε. For a fixed value of ε, the induced clustering similarity is quantified by

$$\text{Utility}(\varepsilon) = \text{ARI}(c, \tilde{c}^{(\varepsilon)}).$$

As the scale of the injected noise decreases with increased ε, larger values of ε generally lead to synthetic data that more closely preserve the original clustering structure. Accordingly, the function $\text{Utility}(\varepsilon)$ tends to be non-decreasing in ε in practice; however, due to the stochastic nature of both K-means initialization and Laplace noise injection, this monotonicity is an empirical tendency rather than a deterministic guarantee.

Given a user-specified target $\text{ARI}_{\text{target}} \in (0, 1)$, the calibration selects

$$\varepsilon^\star = \inf\{\varepsilon > 0 : \text{Utility}(\varepsilon) \geqslant \text{ARI}_{\text{target}}\},$$

approximated in practice by evaluating $\text{Utility}(\varepsilon)$ over a finite grid and selecting the value that minimizes $|\text{Utility}(\varepsilon) - \text{ARI}_{\text{target}}|$. Because the grid has finite resolution and each evaluation uses a single stochastic realization, the achieved ARI may differ from $\text{ARI}_{\text{target}}$ in either direction. Only the single synthetic dataset generated using $\varepsilon^\star$ is released; no intermediate results are disclosed.

The privacy implications of this adaptive selection deserve comment. Under standard DP theory, selecting $\varepsilon^\star$ from the private data is not equivalent to post-processing: each grid evaluation applies the synthesis mechanism to the private dataset, creating a data-adaptive query whose privacy cost is not fully accounted for by $\varepsilon^\star$ alone. Certification via the propose-test-release framework [2] would require additional budget allocation and is left for future work; the mechanism is best understood as a practically motivated privacy-preserving procedure rather than one with a formally certified end-to-end ε-DP guarantee.

3.3. Geometric Reconstruction of Synthetic Points

Given the noisy cluster representation obtained in Section 3.1, synthetic data points are reconstructed using a geometric procedure that separates cluster translation from within-cluster dispersion. For each cluster k, let $\tilde{\mu}_k$ denote the noisy centroid. For each original data point $x_i \in C_k$, its distance to the (non-released) true centroid is computed internally as

$$r_i = \|x_i - \mu_k\|_2.$$

As described in Section 3.1, Laplace noise with scale Δ/ε_r is added to obtain a noisy radius $\tilde{r}_i$. Since radial distances are non-negative by definition, any negative value produced by the noise is

clamped to zero:

$$\tilde{r}_i = \max\{0,\ r_i + \eta_i\}, \qquad \eta_i \sim \mathrm{Lap}(0, \Delta/\varepsilon_r).$$

This clamping operation is a deterministic post-processing step applied to the already-perturbed quantity and does not introduce additional information leakage beyond what is already present in $\tilde{r}_i$.

To prevent disclosure of directional information, the original direction, $(x_i - \mu_k)/\|x_i - \mu_k\|_2$, is not used in the reconstruction. Instead, a new direction vector $v_i \sim U(S^{d-1})$ is sampled independently and uniformly from the unit sphere in $\mathbb{R}^d$. Synthetic points are then reconstructed by translation of the noisy radial component around the noisy centroid according to

$$\tilde{x}_i = \tilde{\mu}_k + s_i\, \tilde{r}_i\, v_i,$$

where the scalar $s_i \in (0, 1)$ is sampled such that the resulting points are uniformly distributed within a d-dimensional ball of radius $\tilde{r}_i$. Specifically, s_i is drawn as

$$s_i = U_i^{1/d}, \qquad U_i \sim U(0, 1).$$

3.4. Handling Feature Bounds

Let $\mathcal{J}_b \subseteq \{1, \ldots, d\}$ denote the index set of features for which lower and upper bounds are specified. For each bounded feature $j \in \mathcal{J}_b$, the original data satisfy $L_j \leqslant x_{ij} \leqslant U_j$. Features not in $\mathcal{J}_b$ are treated as unbounded. All bound enforcement is performed in standardized space. Let $\bar{x}_j$ and σ_j denote the empirical mean and standard deviation used to standardize feature j. The corresponding bounds in standardized space are

$$L_j^{(s)} = \frac{L_j - \bar{x}_j}{\sigma_j}, \qquad U_j^{(s)} = \frac{U_j - \bar{x}_j}{\sigma_j}, \qquad j \in \mathcal{J}_b.$$

After Laplace noise is added to each cluster centroid, the resulting noisy centroid $\tilde{\mu}_k$ may violate bounds in one or more bounded coordinates. Feasibility is enforced by a coordinate-wise reflection procedure applied independently to each bounded feature directly in standardized space. The reflection rule for each bounded feature $j \in \mathcal{J}_b$ is

$$\tilde{\mu}_{kj} \leftarrow \begin{cases} 2L_j^{(s)} - \tilde{\mu}_{kj}, & \text{if } \tilde{\mu}_{kj} < L_j^{(s)}; \\ 2U_j^{(s)} - \tilde{\mu}_{kj}, & \text{if } \tilde{\mu}_{kj} > U_j^{(s)}; \\ \tilde{\mu}_{kj}, & \text{otherwise.} \end{cases}$$

For unbounded features $j \notin \mathcal{J}_b$, no correction is applied. This coordinate-wise reflection ensures that all noisy centroids lie within the specified bounds. As the transformation operates solely on the already-perturbed centroid values, it does not introduce additional information about the original dataset beyond what is already in $\tilde{\mu}_k$.

At this stage, synthetic points generated around $\tilde{\mu}_k$ are not yet required to satisfy feature bounds. During geometric reconstruction, a synthetic point $\tilde{x}_i$ generated as described in Section 3.3 may violate bounds in one or more bounded coordinates. Feasibility is enforced by coordinate-wise clipping to the admissible interval: for each bounded feature $j \in \mathcal{J}_b$, the synthetic coordinate is set to

$$\tilde{x}_{ij}^{(s)} \leftarrow \min\{U_j^{(s)}, \max\{L_j^{(s)}, \tilde{x}_{ij}^{(s)}\}\}.$$

This clipping step operates on already-perturbed outputs and does not introduce additional information leakage. While clipping may concentrate synthetic points near the boundary when the noisy centroid lies close to a bound, this is an acceptable trade-off for computational efficiency compared to rejection sampling.

3.5. Post-processing and Discrete Feature Handling

All synthetic data generated in Sections 3.1–3.4 are represented in standardized space. As a final post-processing step, the inverse standardization is applied to recover synthetic data on the original feature scale. Let $\bar{x}_j$ and σ_j denote the empirical mean and standard deviation used during standardization. Each synthetic data point $\tilde{x}_i^{(s)}$ is mapped back to the original scale via

$$\tilde{x}_{ij} = \sigma_j \tilde{x}_{ij}^{(s)} + \bar{x}_j, \qquad j \in \{1, 2, \ldots, d\}.$$

Let $\mathcal{J}_d \subseteq \{1, \ldots, d\}$ denote the index set of features that are discrete-valued. For continuous features $j \notin \mathcal{J}_d$, the inverse-transformed values are retained without modification. For each discrete feature $j \in \mathcal{J}_d$, the inverse-transformed value $\tilde{x}_{ij}$ is converted to a valid discrete value by randomized rounding. Specifically, let $\lfloor \tilde{x}_{ij} \rfloor$ and $\lceil \tilde{x}_{ij} \rceil$ denote the floor and ceiling of $\tilde{x}_{ij}$. If these two values coincide, the value is already an integer and is retained. Otherwise, the final discrete synthetic value is defined as

$$\tilde{x}_{ij} = \begin{cases} \lceil \tilde{x}_{ij} \rceil, & \text{with probability } 1/2; \\ \lfloor \tilde{x}_{ij} \rfloor, & \text{with probability } 1/2. \end{cases}$$

This equal-probability rounding rule is simple and avoids dependence on the exact fractional part of the inverse-transformed value. If discrete features are subject to predefined bounds, the rounded values are subsequently clipped to the admissible integer set. Under the post-processing property of differential privacy, application of any deterministic or randomized function to the output of a differentially private mechanism does not increase the privacy loss [2]. Randomized rounding thus adds no privacy cost beyond that already incurred by the Laplace perturbation.

Randomized rounding is applied independently to each coordinate and observation, so the joint dependence structure between discrete and continuous features may be partially disrupted. In low-dimensional, well-separated settings this effect is typically small, but it may grow under strong feature correlation or when inverse-transformed values fall far from integers. Dependence-aware rounding strategies are a direction for future work.

3.6. Algorithm Summary

The complete synthetic data generation mechanism described in Sections 3.1–3.5 is summarized in Algorithm 1.

Algorithm 1 Differentially Private Synthetic Data Generation via Cluster Geometry

Require: Private dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$; number of clusters K; target clustering similarity $\mathrm{ARI}_{\mathrm{target}} \in (0, 1)$; feature bounds $\{(L_j, U_j)\}_{j \in \mathcal{J}_b}$; feature type sets $\mathcal{J}_b$ (bounded), $\mathcal{J}_d$ (discrete)

Ensure: Synthetic dataset $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$

1: **Standardization:** standardize X feature-wise to obtain $X^{(s)}$; transform bounds to standardized space as $L_j^{(s)} = (L_j - \bar{x}_j)/\sigma_j$, $U_j^{(s)} = (U_j - \bar{x}_j)/\sigma_j$

2: **Clustering:** apply K-means to $X^{(s)}$ to obtain labels c and clusters $\{C_1, \ldots, C_K\}$ with centroids $\{\mu_k\}_{k=1}^K$

3: **Sensitivity:** compute $\Delta = \max_{j \neq k} \|\mu_j - \mu_k\|_2$

4: **Privacy calibration:** select $\varepsilon^\star$ internally such that $\mathrm{ARI}(c, \tilde{c}^{(\varepsilon^\star)}) \approx \mathrm{ARI}_{\mathrm{target}}$; release only the final synthetic dataset

5: **Privacy budget split:** set $\varepsilon_\mu = \varepsilon_r = \varepsilon^\star/2$

6: **Centroid privatization:** for each cluster k, add Laplace noise with scale Δ/ε_μ independently to each coordinate to obtain $\tilde{\mu}_k = \mu_k + \delta_k$, where $\delta_k \sim \mathrm{Lap}(0, \Delta/\varepsilon_\mu)^d$

7: **Centroid reflection (feasibility correction):** for each bounded feature $j \in \mathcal{J}_b$, apply coordinate-wise reflection in standardized space: if $\tilde{\mu}_{kj} < L_j^{(s)}$, set $\tilde{\mu}_{kj} \leftarrow 2L_j^{(s)} - \tilde{\mu}_{kj}$; if $\tilde{\mu}_{kj} > U_j^{(s)}$, set $\tilde{\mu}_{kj} \leftarrow 2U_j^{(s)} - \tilde{\mu}_{kj}$

8: **Distance privatization:** for each $x_i \in C_k$, compute $r_i = \|x_i - \mu_k\|_2$ and add Laplace noise with scale Δ/ε_r; clamp to obtain $\tilde{r}_i = \max\{0, r_i + \eta_i\}$

9: **Geometric reconstruction:** for each $x_i \in C_k$, sample a direction $v_i \sim U(S^{d-1})$ and a scaling $s_i = U_i^{1/d}$ with $U_i \sim U(0,1)$; compute $\tilde{x}_i^{(s)} = \tilde{\mu}_k + s_i \tilde{r}_i v_i$

10: **Bound enforcement (clipping):** for each bounded feature $j \in \mathcal{J}_b$, set $\tilde{x}_{ij}^{(s)} \leftarrow \min\{U_j^{(s)}, \max\{L_j^{(s)}, \tilde{x}_{ij}^{(s)}\}\}$

11: **Inverse transformation:** map $\tilde{X}^{(s)}$ back to the original feature scale via $\tilde{x}_{ij} = \sigma_j \tilde{x}_{ij}^{(s)} + \bar{x}_j$

12: **Discrete post-processing:** apply randomized rounding (equal probability) to features $j \in \mathcal{J}_d$; clip to integer bounds if applicable

13: **return** $\tilde{X}$

3.7. Computational Complexity

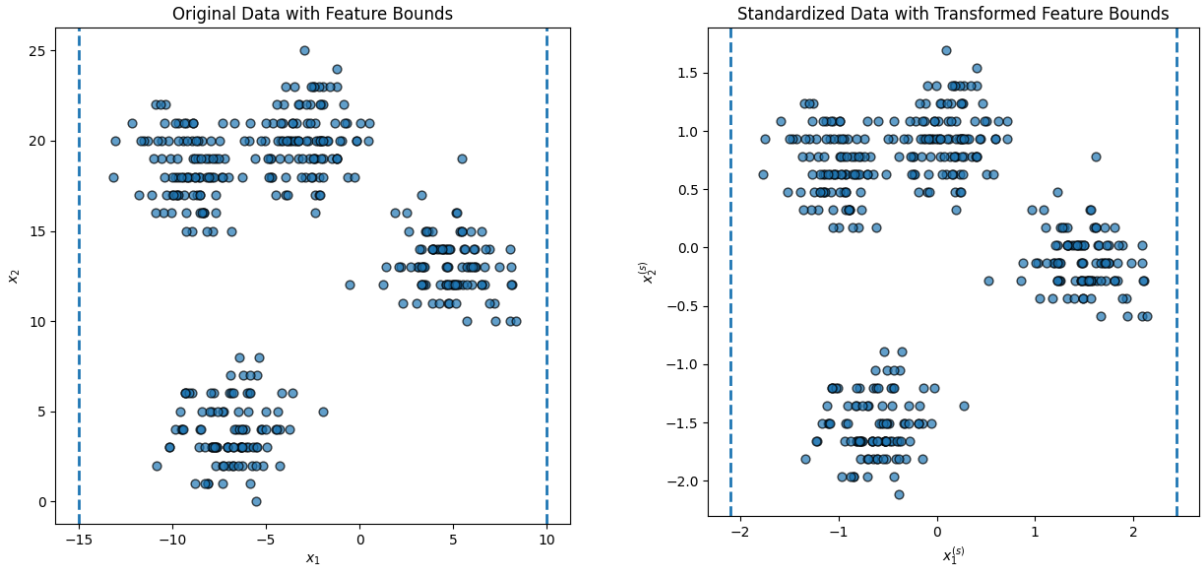
Let n denote the number of observations, d the number of features, K the number of clusters, and I the number of K-means iterations. The dominant cost of each stage of Algorithm 1 is as follows.

- Feature-wise means and standard deviations require $O(nd)$ time and are computed once.
- Each K-means iteration computes pairwise distances between all n points and K centroids at $O(nKd)$ per iteration, for a total of $O(nKdI)$ over I iterations.
- The computation of $\Delta = \max_{j \neq k} \|\mu_j - \mu_k\|_2$ requires $\binom{K}{2}$ pairwise centroid distances each of cost $O(d)$, which gives $O(K^2 d)$. For typical values of K this is negligible.
- Laplace noise is added independently to each of the Kd centroid coordinates at cost $O(Kd)$.
- Coordinate-wise reflection across K centroids and $|\mathcal{J}_b|$ bounded features costs $O(K|\mathcal{J}_b|) \subseteq O(Kd)$.
- One radial distance per observation is computed and perturbed at total cost $O(nd)$.
- Direction sampling from $U(S^{d-1})$ and synthetic point construction each cost $O(d)$ per point; the clipping step costs $O(|\mathcal{J}_b|)$ per point. The total over all n points is $O(nd)$.
- Rounding and clipping n discrete feature values costs $O(n|\mathcal{J}_d|) \subseteq O(nd)$.
- Evaluation of Utility(ε) over a grid of G candidate values requires G independent runs of the synthesis procedure, each of cost $O(nKdI + nd)$, for a total of $O(G \cdot nKdI)$.

The overall computational complexity is $O(G \cdot nKdI)$, driven entirely by the ARI calibration grid search. For fixed K, d, I, and G, the cost scales linearly in n, which means the mechanism is applicable to datasets of moderate size. High-dimensional settings increase the per-point cost of distance computation and direction sampling, but do not change the linear scaling in n. The main practical limitation is the calibration grid: large G is needed when the ARI–ε relationship is flat or non-monotonic, which may occur in low-signal or high-noise regimes.

4. Numerical Example

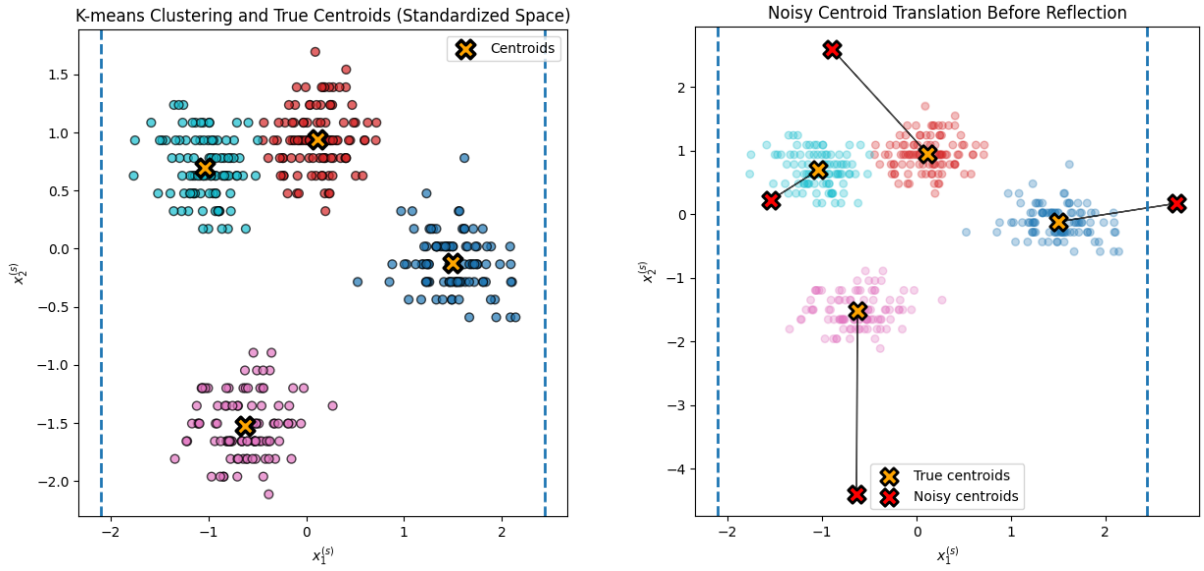
The proposed mechanism is illustrated on a two-dimensional dataset. All experiments were conducted using Google Colaboratory with Python 3.10.12 and the most recent stable versions of the required libraries. The dataset consists of $n = 400$ observations with two features (x_1, x_2): x_1 is continuous with known bounds $x_1 \in [-15, 10]$, and x_2 is discrete and unbounded. The data form four well-separated clusters of approximately equal size. The dataset and its standardized counterpart are shown in Fig. 1, with the admissible bounds for x_1 indicated by dashed vertical lines.



(a) **(b)**

Fig. 1: Scatter plots (a) original data with bounds and (b) standardized data with standardized bounds.

The calibration procedure takes $\mathrm{ARI}_{\text{target}} = 0.10$ as input and yields $\varepsilon \approx 1.6006$. K-means with $K = 4$ is applied to the standardized data; the four centroids are located at approximately $(-1.0,\ 0.7)$, $(0.2,\ 0.9)$, $(1.6,\ 0.0)$, and $(-0.7,\ -1.5)$, with $\Delta \approx 2.84$ standardized units.



(a) **(b)**

Fig. 2: (a) K-means clustering of the standardized dataset with $K = 4$. (b) Effect of Laplace noise injection on cluster centroids in standardized space.

Laplace noise is added to each centroid coordinate at scale $\Delta/\varepsilon_\mu \approx 3.55$ standardized units. The four observed displacements are approximately 0.23, 3.64, 0.80, and 2.86 standardized units; the largest exceeds Δ and pushes one centroid outside the admissible region for $x_1^{(s)}$, necessitating reflection (Fig. 2(b)). The reflected centroids and the data translated around them are shown in Fig. 3.

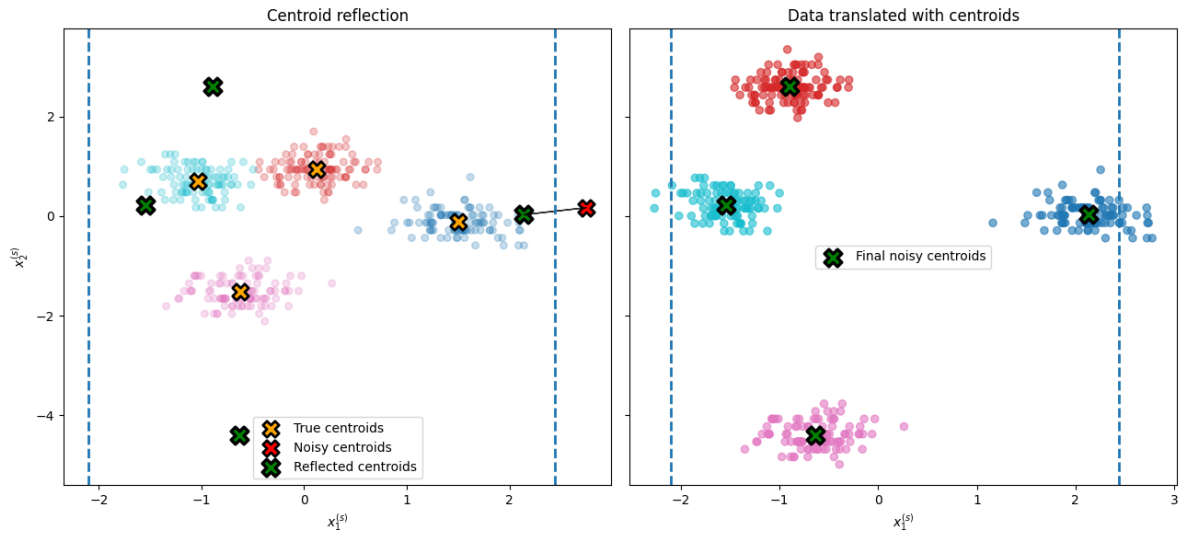
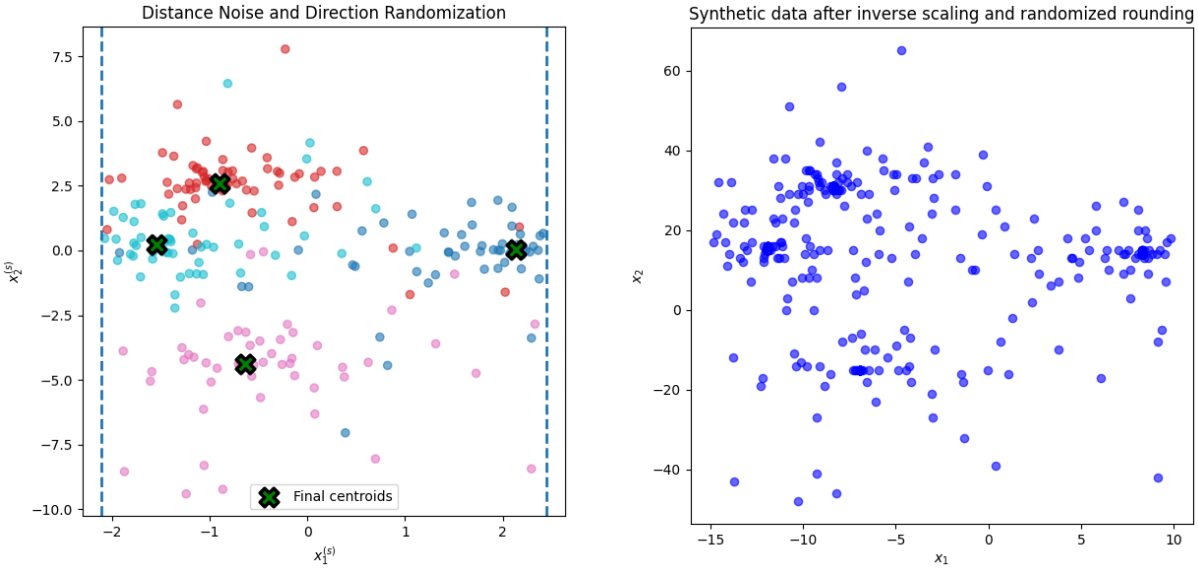


Fig. 3: Feasible noisy centroids obtained by reflection of noisy centroids back into the admissible bounds.

Radial distances are then perturbed at the same Laplace scale. The mean within-cluster radius is approximately 0.37 standardized units — roughly ten times smaller than the noise scale — so many noisy radii are substantially inflated and a non-trivial fraction are clamped to zero. The resulting dispersion and the final synthetic data after inverse standardization and randomized rounding of x_2 are shown in Fig. 4.



(a) **(b)**

Fig. 4: (a) Perturbation of within-cluster distances via Laplace noise prior to geometric reconstruction. (b) Synthetic data after inverse scaling and randomized rounding.

Fig. 5 compares K-means clustering on the original and synthetic data. The synthetic clusters are wider and more diffuse but their relative positions are preserved, because the four clusters are well separated and the data are low-dimensional. When clusters overlap or the signal-to-noise ratio is low, large radial and centroid noise may merge adjacent clusters, producing low ARI even at moderate ε.

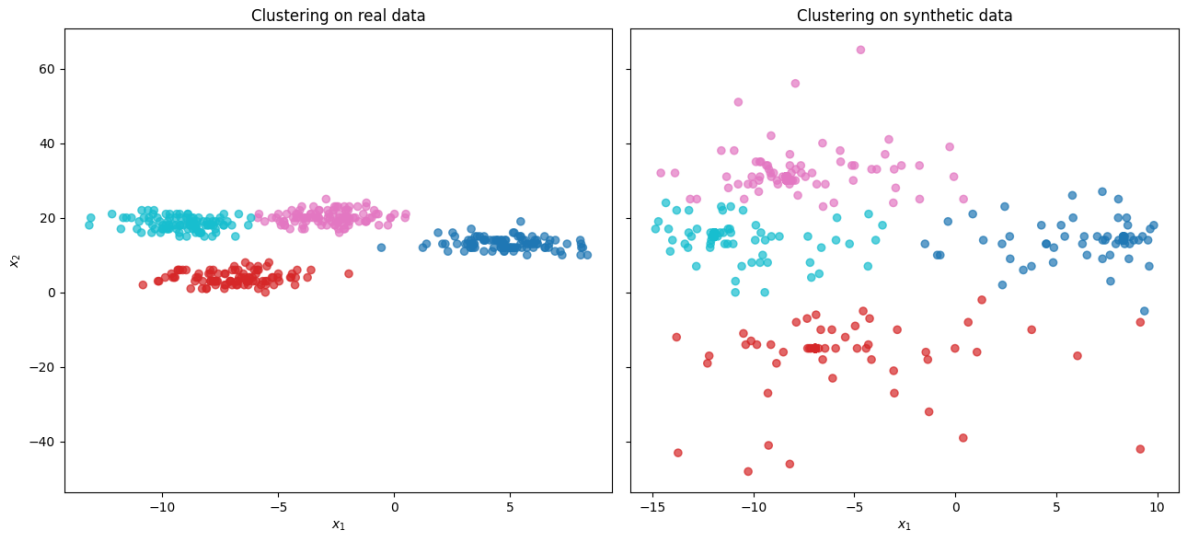


Fig. 5: Comparison of K-means clustering results on the original data (left) and the synthetic data (right).

Reclustering the synthetic data against the original labels yields ARI = 0.6656, well above the target of 0.10. Table 1 gives the corresponding confusion matrix, where each row is dominated by a single column. The achieved ARI greatly exceeds the target because $\varepsilon^\star$ is the smallest grid value satisfying $\text{Utility}(\varepsilon) \geqslant \text{ARI}_{\text{target}}$, so the result can overshoot considerably when the ARI–ε curve is steep, and the strong cluster separation makes the ARI robust to large perturbations.

Table 1: Confusion matrix between real cluster labels and synthetic cluster labels obtained by K-means.

Real cluster	Synthetic cluster			
	0	1	2	3
0	1	2	94	3
1	3	83	3	11
2	90	0	1	5
3	3	14	11	76

4.1. Baseline Comparison

To contextualize the utility of the proposed mechanism, two baseline methods are evaluated on the same illustrative dataset at the same privacy budget $\varepsilon \approx 1.60$.

- K-means clustering is applied and synthetic points are generated using the identical geometric reconstruction procedure: centroid-anchored ball sampling, feature clipping, and randomized rounding, but with no Laplace noise added to centroids or radii ($\varepsilon \to \infty$). This represents the best achievable clustering fidelity under the proposed generative structure in the absence of any privacy constraint.
- Laplace noise with scale Δ/ε is added independently to each coordinate of every data point directly in standardized space, without any use of cluster structure. Feature clipping and randomized rounding are then applied identically to the proposed method. This represents the simplest direct-perturbation alternative that provides the same nominal privacy budget.

Table 2 reports the ARI achieved by each method.

Table 2: ARI comparison of the proposed mechanism against two baseline methods at $\varepsilon \approx 1.60$ (single fixed-seed run).

Method	Privacy	ε	ARI
Non-private synthesis	None	∞	0.624
Proposed mechanism	Yes	≈ 1.60	0.666
Coordinate-wise noise	Yes	≈ 1.60	0.199

Several observations follow. The coordinate-wise noise baseline achieves ARI = 0.199, far below the proposed mechanism's ARI = 0.666 at the same ε. This demonstrates that the utility advantage of the proposed approach comes from the cluster-level perturbation structure: by the organization of noise around centroid geometry rather than independent application to each data point, the mechanism preserves inter-cluster separation even under substantial noise. The proposed mechanism also exceeds the non-private baseline (ARI = 0.624) in this single run. This counterintuitive result is a stochastic artifact of this particular seed: the noisy centroid displacement happens to produce synthetic clusters that are more cleanly separable than those produced by exact reconstruction from the true centroids. This outcome is not expected to hold in general, particularly on datasets with weaker cluster separation; the variability analysis in Section 4.2 shows that the ARI ranges from 0.381 to 0.790 across runs, which confirms that this single result should not be taken as representative.

4.2. Variability Across Random Seeds

Because both K-means initialization and Laplace noise injection are stochastic, the achieved ARI varies across independent runs of the mechanism. To characterize this variability, the proposed mechanism was run 30 times at $\varepsilon \approx 1.60$ (the value selected for $\text{ARI}_{\text{target}} = 0.10$) with different random seeds and all other settings fixed. Table 3 reports the resulting ARI distribution.

Table 3: Distribution of achieved ARI across 30 independent runs at $\varepsilon \approx 1.60$, $\text{ARI}_{\text{target}} = 0.10$.

Mean	Std. dev.	Min	Q1	Median	Max
0.629	0.108	0.381	0.576	0.647	0.790

The mean ARI of 0.629 across 30 runs comfortably exceeds the specified target of 0.10, which confirms that the calibration reliably produces utility above the threshold despite stochastic variation. The standard deviation of 0.108 indicates meaningful run-to-run variability: the ARI ranges from 0.381 to 0.790, and the interquartile range spans [0.576, 0.647]. This variability has two sources: K-means initialization, which can yield different cluster assignments and centroids across runs, and the Laplace noise realization, which may displace centroids by very different amounts. Practitioners who require a guaranteed minimum ARI in any given run should account for this variability by a wider target threshold or a run-selection strategy — accepting, however, that any data-dependent selection of which output to release has additional privacy implications beyond those of a single run, as discussed in Section 3.2.

4.3. Sensitivity to the Target ARI Parameter

To examine how the target ARI parameter influences the selected privacy budget and the achieved utility, the calibration procedure was applied to the same illustrative dataset configuration under three different target values: $\text{ARI}_{\text{target}} \in \{0.10, 0.30, 0.50\}$. A fixed random seed was used for all three runs to isolate the effect of the target value from stochastic variation; the variability analysis in Section 4.2 provides the broader distributional picture. Table 4 reports the selected $\varepsilon^\star$ and the achieved ARI for each target.

Table 4: Effect of the target ARI on the selected privacy budget and achieved clustering similarity. Results are from a single run with a fixed random seed on the illustrative dataset of Section 4.

Target ARI	Selected $\varepsilon^\star$	Achieved ARI
0.10	1.70	0.18
0.30	1.91	0.29
0.50	2.48	0.48

The results show that a higher target ARI requires a larger privacy budget: $\varepsilon^\star$ increases from 1.70 to 2.48 as the target rises from 0.10 to 0.50. This is consistent with the expected behavior of the Utility(ε) function: a stricter utility requirement can only be met by noise reduction, which corresponds to a larger ε. The achieved ARI tracks the target reasonably closely in each case, with small discrepancies attributable to the finite resolution of the candidate grid and the stochastic nature of both K-means and the Laplace perturbation. Specifically, the achieved ARI exceeds the target for the lowest target value (because the ARI–ε curve is steep there and the selected $\varepsilon^\star$ overshoots) and falls marginally below the target for the higher values (a consequence of grid granularity combined with a single stochastic evaluation per candidate).

The apparent monotonicity of $\varepsilon^\star$ in $\mathrm{ARI}_{\mathrm{target}}$ is intuitive: smaller noise (larger ε) preserves more cluster structure, so a stricter utility target requires a larger budget. As noted in Section 3.2, however, this monotonicity is an empirical tendency rather than a guarantee; in a single stochastic run the estimated utility function may be non-monotone, and the grid search could select a smaller ε for a larger target if a favorable noise realization occurred at that point. The results in Table 4 should therefore not be interpreted as a universal relationship: they reflect one realization on one dataset, and both $\varepsilon^\star$ values and the direction of the discrepancies will vary across datasets, seeds, and grid configurations.

4.4. Limitations

The results should be interpreted in light of several limitations. The evaluation is based on a single two-dimensional synthetic dataset with four well-separated clusters of similar size, a setting that is favorable to the proposed approach. Strong cluster separation makes the ARI relatively robust to centroid perturbation, while the low dimensionality limits distortion from clipping and rounding. Consequently, the present results cannot establish performance on higher-dimensional data, overlapping or imbalanced clusters, or datasets with more complex discrete-variable structures. In addition, the baseline comparison is limited to a non-private version and a coordinate-wise noise mechanism. Although the proposed method substantially outperforms direct coordinate-wise perturbation at the same privacy budget, comparisons with dedicated clustering-based differentially private synthesis methods, such as MC-GEN [5] and the co-clustering approach of [6], remain necessary.

The reported utility metrics are also subject to stochastic variability and methodological assumptions. Across 30 independent runs at $\varepsilon \approx 1.60$, the ARI exhibited noticeable variation, indicating that a single realization should not be regarded as fully representative. Furthermore, the sensitivity Δ was defined as the maximum pairwise distance between K-means centroids, a data-dependent quantity rather than a worst-case bound derived from the clustering procedure itself. A rigorous analysis of centroid and cluster-assignment stability under single-record perturbations remains an open problem. Finally, the effects of clipping and independent rounding on high-dimensional or strongly correlated data were not investigated and may lead to greater distortion than observed in the present study.

5. Conclusion

This paper proposed a geometry-based privacy-preserving synthetic data generation mechanism that addresses two practical gaps in prior clustering-based DP synthesis: the absence of hard

feature bound enforcement and the lack of a utility-driven budget selection strategy. The method applies Laplace noise jointly to K-means cluster centroids and within-cluster radial distances, enforces bounds via coordinate-wise centroid reflection and point clipping in standardized space, and handles discrete features via randomized rounding as a cost-free post-processing step under the post-processing property of differential privacy. On the illustrative two-dimensional example the mechanism achieves mean ARI 0.629 ± 0.108 across 30 independent runs at $\varepsilon \approx 1.60$, well above the target of 0.10, and substantially outperforms direct coordinate-wise noise addition at the same budget (ARI 0.666 vs. 0.199). Several limitations bound the scope of these findings: the sensitivity is a heuristic data-dependent approximation, the adaptive ε-selection has not been formally analyzed under standard DP accounting, and the evaluation is confined to a single low-dimensional synthetic dataset without benchmarking against dedicated DP synthesis methods such as MC-GEN or PrivSyn.

Several directions remain open. Evaluation on multiple public datasets with varying dimensionality and cluster geometry, and benchmarking against dedicated DP synthesis methods, are the most pressing empirical needs. On the theoretical side, derivation of a rigorous worst-case K-means sensitivity bound, formal analysis of the adaptive budget-selection step via the propose-test-release framework, and study of clipping and rounding behavior in high-dimensional or strongly correlated settings are natural follow-ups.

CRediT Authorship Contribution Statement

Robby: Conceptualization, Software, Formal Analysis, Visualization, Writing – Original Draft Preparation, Project Administration, and Funding Acquisition. **Agus Sukmana**: Conceptualization, Methodology, Investigation, Writing – Review & Editing, and Supervision. **Erwinna Chendra**: Conceptualization, Methodology, Formal Analysis, Validation, Writing – Review & Editing, and Supervision.

Declaration of Generative AI and AI-assisted technologies

The authors used AI-assisted writing tools during the preparation of this manuscript. Specifically, ChatGPT (developed by OpenAI) and Claude (developed by Anthropic) were used only for language editing and grammatical refinement. No AI-generated content contributed to the conceptual development, theoretical formulation, or substantive research contributions of the paper.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research was made possible through the research funding provided by LPPM Parahyangan Catholic University in 2026. We sincerely thank them for their support.

Data and Code Availability

The data used in this study are entirely synthetic. The implementation code may be made available upon reasonable request to the corresponding author via email.

References

[1] Mengmeng Yang, Longxia Huang, and Cheng Pei Tang. "K-Means Clustering with Local Distance Privacy". In: *Big Data Mining and Analytics* (2023). DOI: 10.26599/BDMA.2022.9020050.

[2] Dong Su, Jianneng Cao, Ninghui Li, Elisa Bertino, Min Lyu, and Hongxia Jin. "Differentially Private K-Means Clustering and a Hybrid Approach to Private Optimization". In: *ACM Transactions on Knowledge Discovery from Data* (2017). DOI: 10.1145/3133201.

[3] Boyu Zhu, Yuan Zhang, Tingting Chen, and Sheng Zhong. "Differentially Private K-Means Publishing with Distributed Dimensions". In: *Proceedings of the IEEE Conference on Computer Supported Cooperative Work in Design*. 2024. DOI: 10.1109/CSWD61410.2024.10580021.

[4] Shuhui Fang, Xuejun Wan, Jun Wang, Lin Chai, Wenlin Pan, and Wu Wang. "HiDS Data Clustering Algorithm Based on Differential Privacy". In: *Proceedings of IEEE NaNA* (2024). DOI: 10.1109/NANA63151.2024.00029.

[5] Messaoud Saoudi. "MC-GEN: Multi-Level Clustering for Private Synthetic Data Generation". In: *Knowledge-Based Systems* (2023). DOI: 10.1016/j.knosys.2022.110239.

[6] Tarek Benkhelif, Françoise Fessant, Fabrice Clérot, and Guillaume Raschia. "Co-Clustering for Differentially Private Synthetic Data Generation". In: *Advances in Knowledge Discovery and Data Mining*. Springer, 2017. DOI: 10.1007/978-3-319-71970-2_5.

[7] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. "PrivSyn: Differentially Private Data Synthesis". In: *Proceedings of the VLDB Endowment* (2021). https://www.usenix.org/conference/usenixsecurity21/presentation/zhang-zhikun.

[8] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees". In: *International Conference on Learning Representations*. 2019. https://openreview.net/forum?id=S1zk9iRqF7.

[9] Noseong Park, Mihail Popescu, Youngja Park, Sungchul Kim, Ryan A. Rossi, and Franck Dernoncourt. "Differentially Private Tabular Data Synthesis Using Large Language Models". In: *arXiv preprint* (2023). DOI: 10.48550/arXiv.2304.10701.

[10] Jingwen Zhao, Yunfang Chen, and Wei Zhang. "Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions". In: *IEEE Access* (2019). DOI: 10.1109/ACCESS.2019.2909559.

[11] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep Learning with Differential Privacy". In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 308–318. DOI: 10.1145/2976749.2978318.

[12] Mónica Ribero, Jette Henderson, Sinead A. Williamson, and Haris Vikalo. "Federating Recommendations Using Differentially Private Prototypes". In: *Pattern Recognition* (2022). DOI: 10.1016/j.patcog.2022.108746.

[13] Jian-Zhi Zhao, Wenji Wang, Jiabao Wang, Songyang Zhang, Zhelin Fan, and Stan Matwin. "Privacy-Preserved Federated Clustering with Non-IID Data via GANs". In: *The Journal of Supercomputing* (2025). DOI: 10.1007/s11227-025-07006-2.

[14] Qaiser Razi, Souptik Datta, Vikas Hassija, G. S. S. Chalapathi, and Biplab Sikdar. "Privacy Utility Tradeoff Between PETs: Differential Privacy and Synthetic Data". In: *IEEE Transactions on Computational Social Systems* (2024). DOI: 10.1109/TCSS.2024.3479317.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography Conference*. Vol. 3876. Lecture Notes in Computer Science. Springer, 2006, pp. 265–284. DOI: 10.1007/11681878_14.