On Privacy Protection by Synthetic Data Generation

Submitted to
Graduate School of Information Science and Technology
The University of Osaka

July 2025

Takayuki MIURA

List of Publications

Journals (Peer-reviewed)

1. Takayuki Miura, Masanobu Kii, Toshiki Shibahara, Kazuki Iwahana, Tetsuya Okuda, Atsunori Ichikawa, and Naoto Yanai. Setsubun: Revisiting membership inference game for evaluating synthetic data generation. *Journal of Information Processing*, Vol. 32, pp. 757–766, 2024.

2. Takayuki Miura, Toshiki Shibahara, Masanobu Kii, Atsunori Ichikawa, Juko Yamamoto, and Koji Chida. On Rényi differential privacy in statistics-based synthetic data generation. *Journal of Information Processing*, Vol. 31, pp. 812–820, 2023.

International Conferences (Peer-reviewed)

1. Tomoya Matsumoto, Takayuki Miura, Toshiki Shibahara, Masanobu Kii, Kazuki Iwahana, Osamu Saisho, and Shingo Okamura. Differentially Private Sequential Data Synthesis with Structured State Space Models and Diffusion Models. In *NeurIPS Safe Generative AI Workshop*, 8 pages, 2024.

2. Takayuki Miura, Toshiki Shibahara, and Naoto Yanai. MEGEX: Data-free model extraction attack against gradient-based explainable AI. In *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems*, SecTL '24, p. 56–66, New York, NY, USA, 2024. Association for Computing Machinery.

3. Takayuki Miura, Eizen Kimura, Atsunori Ichikawa, Masanobu Kii, and Juko Yamamoto. Evaluating synthetic data generation techniques for medical dataset. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: HEALTHINF*, pp.

315–322. INSTICC, SciTePress, 2024.

4. Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pp. 77–83. IEEE, 2023.

Domestic Conference Paper (not Peer-reviewed)

1. 三浦尭之, 竹内弘史, 櫛部義幸, 紀伊真昇, 芝原俊樹, 山本充子, 市川敦謙, 石原一郎, 千田浩司. PrivBayes^{+}: Staircase メカニズムを用いた PrivBayes の性能向上. 第 202 回マルチメディア通信と分散処理・第 108 回コンピュータセキュリティ合同研究発表会 (CSEC108), 8 pages, mar 2025.

2. 武内祐哉, 岩花一輝, 三浦尭之, 芝原俊樹, 山下恭佑. インコンテキスト学習による Jailbreak に対するバックドアを利用した防御手法の一検討. 2025 年　暗号と情報セキュリティシンポジウム（SCIS2025), 8pages, jan 2025.

3. 岩花一輝, 伊東燦, 山田真徳, 芝原俊樹, 山下智也, 三浦尭之. モデル合成に対するプライバシーリスク評価. コンピュータセキュリティシンポジウム 2024 論文集, 8 pages, oct 2024.

4. 三浦尭之, 竹内弘史, 櫛部義幸, 紀伊真昇, 芝原俊樹, 山本充子, 市川敦謙, 石原一郎, 千田浩司. 高次元クエリに対する Staircase メカニズムの適用. コンピュータセキュリティシンポジウム 2024 論文集, 8 pages, oct 2024.

5. 松本知優, 三浦尭之, 芝原俊樹, 紀伊真昇, 岩花一輝, 税所修, 岡村真吾. 構造化状態空間モデルと拡散モデルを用いた差分プライベートな系列データ合成. コンピュータセキュリティシンポジウム 2024 論文集, 8 pages, oct 2024.

6. 松本知優, 杉浦一瑳, 手島宏貴, 三浦尭之, 矢内直人. 処理時間を考慮に入れた匿名化コンテストの提案. 情報処理学会第 86 回全国大会, 2pages, mar 2024.

7. 三浦尭之, 権英哲, 芝原俊樹, 矢内直人. 勾配系の説明がついた 3 層 ReLU ネットワークに対するモデル抽出攻撃. 2024 年　暗号と情報セキュリティシンポジウム（SCIS2024), 8pages, jan 2024.

8. 三浦尭之, 紀伊真昇, 市川敦謙, 岩花一輝, 芝原俊樹, 奥田哲矢, 矢内直人. 合成データに対するメンバーシップ推論攻撃評価フレームワークの拡張ム. マルチメディア、分散、協調とモバイル (DICOMO2023) シンポジウム, 8pages, jul 2023.

9. 松本知優, 三浦尭之, 矢内直人. 拡散モデルのメンバーシップ推論耐性の評価. マルチメディア、分散、協調とモバイル (DICOMO2023) シンポジウム, 8pages, jul 2023.

10. 三浦尭之, 紀伊真昇, 市川敦謙, 山本充子, 木村映善. 合成データ生成技術の医療データへの適用と課題. 第 27 回日本医療情報学会春季学術大会, 8 pages, jul 2023.

11. 三浦尭之, 紀伊真昇, 市川敦謙, 岩花一輝, 芝原俊樹, 奥田哲矢, 山本充子, 矢内直人. 合成データ生成の出力を評価するメンバーシップ推論攻撃フレームワーク. コンピュータセキュリティシンポジウム 2022 論文集, 8 pages, oct 2022.

12. 三浦尭之, 紀伊真昇, 芝原俊樹, 市川敦謙, 山本充子, 千田浩司. 合成データ生成のランダム性が持つ Renyi 差分プライバシー性の評価. コンピュータセキュリティシンポジウム 2022 論文集, 8 pages, oct 2202.

13. 三浦尭之, 芝原俊樹, 矢内直人. MEGEX：勾配系の説明可能な AI に対するデータフリーモデル抽出攻撃. コンピュータセキュリティシンポジウム 2022 論文集, 8 pages, oct 2024.

14. 三浦尭之, 芝原俊樹, 矢内直人. 勾配系の説明付きモデルに対するデータフリーモデル抽出攻撃. 2022 年　暗号と情報セキュリティシンポジウム（SCIS2022), 8pages, jan 2022.

15. 岩花一輝, 三浦尭之, 奥田哲矢, 矢内直人. 電子指紋は機械学習の二段階攻撃に使えるか？. 2022 年　暗号と情報セキュリティシンポジウム（SCIS2022), 8pages, jan 2022.

Abstract

Advances in deep learning and computational power have increased the value of data analysis, particularly when leveraging personal information in domains such as healthcare, finance, and recommendation systems. To enable data sharing while preserving privacy, anonymization techniques have been explored. However, conventional methods like k-anonymity suffer from severe utility loss in high-dimensional settings due to the curse of dimensionality, making it difficult to simultaneously ensure privacy and maintain data utility. As an alternative, synthetic data generation has gained traction. These methods extract generative parameters from real datasets to produce new, statistically similar data. To formally guarantee privacy, differentially private synthetic data generation has been proposed, where noise is added to the parameter extraction process. However, this often leads to reduced practical utility, and successful societal implementation remains limited due to the ongoing challenge of achieving a viable privacy-utility trade-off in high-dimensional data.

This thesis seeks to promote the societal implementation of synthetic data generation by addressing its core challenges. Synthetic data generation methods are categorized into two types: (i) non-differentially private (non-DP) approaches, and (ii) differentially private (DP) approaches. In this thesis, we aim to advance the societal implementation of synthetic data generation by analyzing challenges for each type respectively.

In Chapter 3, we propose a privacy evaluation framework to overcome limitations of existing evaluation frameworks; (1) they cannot evaluate the worst-case because a target sample is chosen randomly; and (2) the decision criterion of an adversary's inference is black box since the adversary conducts membership inference by using machine learning models. To cope with limitation (1), we introduce a statistical distance and propose the way to choose a vulnerable target sample with respect to the distance. To cope with limitation (2), we propose two interpretable and simple inference methods. One is a method with typical statistics scores, and the other

is a method with the number of samples close to the target sample with respect to Euclidean distance. We conduct extensive experiments on two datasets and five synthesis algorithms to confirm the effectiveness of our framework. The experiments show that our framework enables us to evaluate privacy in synthetic data generation techniques more tightly from the perspective of the statistical distance.

In Chapter 4, we address the challenge of evaluating the utility of differentially private synthetic data generation methods. We conduct experiments with a Diagnosis Procedure Combination (DPC) dataset to evaluate the quality of synthetic data generated by statistics-based, graphical model-based, and deep neural network-based methods. Further, we implement differential privacy for theoretical privacy protection and assess the resultant degradation of data quality. The findings indicate that a statistics-based method called Gaussian Copula and a graphical-model-based method called AIM yield high-quality synthetic data regarding statistical similarity and machine learning model performance. The chapter also summarizes issues pertinent to the practical application of synthetic data derived from the experimental results.

In Chapter 5, we aim to improve utility, and theoretically evaluate Rényi differential privacy, which is a kind of relaxations of differential privacy, of the randomness in data generation of a synthetic data generation method that uses the mean vector and the covariance matrix of an original dataset. Specifically, for a fixed privacy parameter $\alpha > 1$, we show the condition of the privacy parameter ε such that the synthetic data generation satisfies (α, ε)-Rényi differential privacy under a bounded neighboring condition and an unbounded neighboring condition, respectively. In particular, under the unbounded condition, when the size of the original dataset and synthetic dataset is 10 million, the mechanism satisfies $(4, 0.576)$-Rényi differential privacy. We also show that when we translate it into the traditional (ε, δ)-differential privacy, the mechanism satisfies $(4.46, 10^{-14})$-differential privacy.

Finally, Chapter 6 summarizes this thesis, gives several concluding remarks, and discusses our future work.

Contents

Chapter 1

Introduction

1.1 Background

Driven by breakthroughs in deep neural networks and continuous improvements in computational capabilities, the value derived from data analysis has been steadily increasing in recent years. Especially, personal information constitutes a highly valuable asset for data analysis and is used in various domains, including healthcare [9, 52, 64, 96, 97], finance [20, 38, 93], and marketing [61, 70]. Although unstructured data types such as images and text are actively utilized, the majority of personal data is often curated and employed in the form of structured tabular datasets [12].

The use of such data necessitates careful consideration of the privacy of individuals included in the dataset. Despite the application of ostensibly safe processing, such as removing names or identification numbers and performing random sampling, there have been reported cases in which individual privacy was still compromised through publicly available datasets. In 2008, Netflix published a dataset for a movie recommendation algorithm competition, in which certain user attributes were removed prior to its release. Narayanan et al. subsequently showed that, despite the removal of explicit identifiers, the released dataset still contained sufficient information to enable the re-identification of users, thereby highlighting the inherent risks of insufficient anonymization [91]. In addition, Nissim et al. have reported that a substantial amount of personal information can be reconstructed from the results of official statistics [21, 26].

To mitigate the aforementioned privacy risks while releasing data that preserves similar statistical properties to the original dataset, various techniques such as anonymization have been proposed. One of the distinctive challenges inherent to

these technologies lies in the setting where it is not possible to differentiate between data users and potential adversaries. In other words, it is necessary to release data that simultaneously ensures privacy protection and retains utility. Traditional anonymization techniques based on approaches such as k-anonymity [110] are known to suffer significant utility loss due to the curse of dimensionality when the information pertaining to an individual becomes high-dimensional [3, 122]. Due to the curse of dimensionality, it becomes challenging to achieve an appropriate trade-off between privacy protection and data utility.

To address such privacy-utility trade-offs, privacy protection through synthetic data has garnered increasing attention [56, 112]. A synthetic data generation technique extracts generative parameters, such as statistical values and machine learning model parameters, from the original dataset and generates new data of the same format based on these parameters. Furthermore, in order to provide formal privacy guarantees, differentially private synthetic data generation—where the data synthesis process is designed to satisfy differential privacy [31]—has also been actively studied. Differential privacy is a widely used privacy-preserving criterion that quantifies the extent to which input data can be inferred from the output of a randomized algorithm (See Section 2.3). In general, differentially private synthetic data generation is achieved by applying a differentially private mechanism to the function that extracts the generative parameters. However, the addition of noise to satisfy differential privacy often results in a trade-off that falls short of practical utility, and the extent to which utility can theoretically be achieved remains unclear.

1.2 Goal and Contributions of Thesis

In this thesis, we aim to advance the societal implementation of synthetic data generation in the real world. Successful social implementation necessitates numerous real-world trials. To promote real-world trials, the relationship between utility and privacy must be thoroughly analyzed from a theoretical perspective. This necessitates a two-step research approach. The first step (a) involves the privacy evaluation. The second step (b) consists of validating the utility of the system using real-world data under privacy guarantees. Given that public data may already incorporate certain privacy measures, evaluating the system on real-world data is essential for a reliable and accurate evaluation. Synthetic data can also be categorized into two types: (i) data that is not generated under differential privacy (non-DP), and (ii) data that is generated with differential privacy (DP). The current limita-

Table 1.1: The levels of maturity and remaining challenges associated with the two types of synthetic data generation methods. (a) Privacy represents the extent to which a method has been theoretically analyzed with respect to formal privacy guarantees, such as differential privacy. (b) Real Data Utility refers to the evaluation of data utility conducted using real-world datasets. The check mark ✓expresses that the problem is resolved, and the hyphen - expresses that the problem cannot be solved yet because the preceding stage has not been resolved.

Method	(a) Privacy	(b) Real Data Utility
(i) Data Synthesis	Challenge in Chapter 3	-
(ii) DP Data Synthesis	✓	Challenge in Chapter 4

tions of each method are indicated in the corresponding positions in Table 1.1, and are discussed in detail in the following sections.

1.2.1 Privacy Evaluation Framework for Synthetic Data Generation

In Chapter 3, we address the limitation of privacy evaluation of the approach (i) [88, 130, 131]. Synthetic data is often considered privacy-preserving even without the application of differential privacy, as it appears to be disconnected from the original data [12, 116]. However, in the absence of differential privacy protection, synthetic data has been shown to be vulnerable to privacy attacks, such as membership inference attacks [109, 63, 58]. A membership inference attack is a type of privacy attack in which an adversary attempts to infer whether a particular data point was included in the training dataset based on the output of a machine learning model or synthetic data generator [107]. The success of such an attack may result in several critical consequences, and a membership inference attack is also used as a standard auditing tool for evaluating privacy protection [68, 120].

The existing evaluation frameworks has limitations from two perspectives:

(1) it cannot evaluate the worst-case because a target sample is chosen randomly; and

(2) the decision criterion of an adversary's inference is black box since the adversary conducts membership inference by using machine learning models.

In this chapter, we propose a framework to overcome the above limitations in a simple and clear fashion. To cope with limitation (1), we introduce a statistical distance and propose the way to choose a vulnerable target sample with respect to the distance. To cope with limitation (2), we propose two interpretable and simple inference methods. One is a method with typical statistics scores, and the other is a method with the number of samples close to the target sample with respect to Euclidean distance. We conduct extensive experiments on two datasets and five synthesis algorithms to confirm the effectiveness of our framework. The experiments show that our framework enables us to evaluate privacy in synthetic data generation techniques more tightly.

As a result, it was found that, when using non-DP synthetic data, it is necessary, at a minimum, to remove outliers. Alternatively, given that the theoretical relationship between differential privacy and membership inference attacks is well established [121, 66], it is advisable to use DP synthetic data to ensure a certain level of privacy protection.

1.2.2 Utility Evaluation of Synthetic Data Generation with Real Medical Dataset

Based on the results presented in Chapter 3, it was concluded that differential privacy is required when outliers in the dataset cannot be identified or addressed in advance. In Chapter 4, we address the limitation of the approach (ii) [89, 133, 132]. Differentially private synthetic data generation, a variety of methods have been proposed [74, 82, 83, 123]. The privacy protection of these methods is theoretically guaranteed under the formal definition of differential privacy, which checks (ii)-(a). However, there remain challenges in evaluating the quality of the generated synthetic data. In most cases, widely proposed differentially private synthetic data generation methods are evaluated using publicly available datasets. Without comprehensive studies comparing these methods using real-world data, it is difficult to conduct a rigorous evaluation of their practical viability. Nonetheless, such empirical evaluations on actual datasets remain insufficient.

Given that privacy is of paramount importance and data analysis is highly active in this domain, the this study focuses on the medical field. Anticipation surrounds the use of real-world data for data analysis in medicine and healthcare, yet handling sensitive data demands ethical review and safety management, presenting bottlenecks in the swift progression of research. Consequently, numerous techniques have emerged for generating synthetic data, which preserves the features of the original

data. Nonetheless, the quality of such synthetic data, particularly in the context of real-world data, has yet to be sufficiently examined. In this chapter, we conduct experiments with a real Diagnosis Procedure Combination (DPC) dataset provided by Ehime University Hospital to evaluate the quality of synthetic data generated by statistics-based, graphical model-based, and deep neural network-based methods. Further, we implement differential privacy for theoretical privacy protection and evaluate the resultant degradation of data quality. The findings indicate that a statistics-based method called Gaussian Copula [74] and a graphical-model-based method called AIM [82] yield high-quality synthetic data regarding statistical similarity and machine learning model performance. The chapter also summarizes issues pertinent to the practical application of synthetic data derived from the experimental results.

1.2.3 Evaluating Differential Privacy of Synthetic Data Generation without Adding Intentional Noise

In Chapter 5, we discuss methods for ensuring differential privacy without the addition of noise, with the aim of improving utility [90, 134, 135]. Privacy protection with synthetic data generation often uses differentially private statistics and model parameters to quantitatively express theoretical security. However, these methods do not take into account privacy protection due to the inherent randomness of data generation. Such approaches have rarely been proposed, and the only existing prior work remains at a theoretical level, lacking the capability for concrete numerical evaluation [77]. In practical applications, it is essential that security metrics be expressed in concrete numerical terms. Therefore, it is necessary to quantitatively evaluate the differential privacy guarantees derived from the inherent randomness in the synthetic data generation process.

In this chapter, we theoretically evaluate Rényi differential privacy [87], which is a kind of relaxation of differential privacy, of the randomness in data generation part in a synthetic data generation method that uses the mean vector and the covariance matrix of an original dataset. Specifically, for a fixed $\alpha > 1$, we show the condition of ε such that the synthetic data generation satisfies (α, ε)-Rényi differential privacy under a bounded neighboring condition (Definition 2.3.1) and an unbounded neighboring condition, respectively. In particular, under the unbounded condition, when the size of the original dataset and synthetic dataset is 10 million, the mechanism satisfies $(4, 0.576)$-Rényi differential privacy. We also show that when we translate it into the traditional (ε, δ)-differential privacy, the mechanism satisfies

$(4.46, 10^{-14})$-differential privacy.

1.2.4 Organization of Thesis

This thesis aims to facilitate the societal implementation of synthetic data generation technologies by addressing the challenges outlined in the previous section. The structure of the remainder of this thesis is as follows. Chapter 2 is the preliminary of this thesis. We first introduce the notations, and formulate tabular datasets. We also introduce differential privacy and synthetic data generation. In Chapter 3, we point out the limitations of the existing privacy evaluation framework [109], and propose the way to improve the limitations. As a result, we observe that outliers with respect to the Mahalanobis distance poses a higher risk of privacy leakage from the perspective of membership inference. Next, in Chapter 4, to enable a realistic evaluation of synthetic data utility and privacy protection, we evaluated five different synthetic data generation techniques using actual patient data provided by Ehime University Hospital. In Chapter 5, in order to improve data utility, we propose a novel approach that achieves differential privacy without relying on explicit noise injection. Finally, in Chapter 6, we summarize the obtained results, conclude this thesis, and discuss directions for future work.

Chapter 2

Preliminary

In this chapter, we introduce the fundamental notations, definitions, and concepts that underpin the remainder of this thesis. These preliminaries provide the theoretical and technical groundwork necessary to understand our proposed methods and analyses.

We begin by clarifying the mathematical symbols and notations used throughout the thesis. Then, we formalize the structure of tabular datasets, which are the primary data modality studied in this work. Next, we present the definitions and properties of differential privacy, including its relaxations such as Rényi differential privacy and zero-concentrated differential privacy, along with the key mechanisms and theorems that support privacy-preserving data analysis. Finally, we define the concept of synthetic data generation and introduce the main algorithmic frameworks considered in this thesis, setting the stage for the methodological developments in the subsequent chapters.

2.1 Mathematical Notations

This thesis uses $\mathbb{R}$ to denote the set of real numbers, $\mathbb{Z}$ to denote the set of integers, and $\mathbb{Z}_{\geq 0}$ to denote the set of natural numbers. For a natural number $n \in \mathbb{Z}_{\geq 0}$, a set $[n]$ is defined as

$$[n] := \{x \in \mathbb{Z}_{\geq 0} \mid 1 \leq x \leq n\}.$$

We also denote the closed interval from $a \in \mathbb{R}$ to $b \in \mathbb{R}$ as $[a, b]$. Namely, we denote

$$[a, b] := \{x \in \mathbb{R} \mid a \leq x \leq b\}.$$

For a set S, we denote the number of elements as $|S| \in \mathbb{Z}_{\geq 0}$ and the power set of S as 2^S. The natural number $|S|$ is also called the cardinality of the set. Namely, we

denote

$$2^S := \{A \mid A \subset S\}.$$

The probabilistic simplex is denoted by

$$\Delta^d := \{x \in \mathbb{R}^d \mid \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}.$$

Although the simplex is $(d-1)$-dimensional as a manifold and should thus be denoted as Δ^{d-1}, we use the notation Δ^d in this thesis to emphasize the correspondence with the number of attribute values, which is more important in this thesis context.

2.2 Tabular Dataset

In this section, we explain tabular datasets in research of synthetic data generation. A tabular dataset refers to a structured form of data organized in a two-dimensional table consisting of rows and columns. Tabular datasets are essential because they provide a clear, structured way to organize data, making it easy to analyze, visualize, and share. They are widely supported by data tools and are the standard input for many machine learning and statistical methods, enabling efficient data processing and automation. As an example of such data, in the medical domain, electronic health records (EHRs) often consist of tables where each row corresponds to a patient and each column represents clinical attributes such as age, blood pressure, laboratory test results, diagnoses, and prescribed medications. This structured representation is widely used in predictive modeling tasks, such as disease diagnosis, risk stratification, and treatment outcome prediction.

To formalize the tabular dataset, let $\mathcal{D}$ denote the set of all possible datasets under consideration. Attributes can be categorized into two types: categorical attributes and numerical attributes. For categorical attributes, the attribute values are typically drawn from a finite set $\{a_1, \ldots, a_m\}$, where each a_i represents one of the possible discrete categories. For example, if the attribute represents gender, the set of possible attribute values is expressed by the finite set {male, female}. Numerical attributes take either integer (discrete) values or real (continuous) values. A typical example is age, which can be represented as an integer indicating the number of years.

Categorical and numerical attributes can be transformed into each other using appropriate methods tailored to their respective data types. Numerical attributes

can be regarded as categorical attributes by applying appropriate clustering techniques that partition the continuous values into discrete groups. Conversely, categorical attributes can be treated as numerical attributes by converting them into one-hot vectors. Depending on the method, synthetic data generation techniques assume different formats of tabular data: some are designed for categorical attributes only, others for numerical attributes only, and some are tailored to handle mixed-type data consisting of both categorical and numerical attributes. As noted above, such transformations between categorical and numerical attributes are feasible. Therefore, in this thesis, we assume that appropriate transformations have been applied when necessary.

Here, tabular datasets are assumed to consist of M columns corresponding to categorical attributes, with each row representing an individual record. The dataset is composed of N individuals, each described by a single row. Let $A_1, \dots, A_M$ denote the sets of possible values for each attribute. For each $i \in \{1, \dots, M\}$, we define $d_i := |A_i|$ as the cardinality of the set A_i. In this setting, the information corresponding to a single individual (i.e., a row in the table) can be represented as

$$x \in A_1 \times \cdots \times A_M =: A,$$

and the entire dataset consisting of N individuals can be expressed as

$$D \in A^N = \mathcal{D}.$$

Let $r = \{r_1, \dots, r_t\} \subset [M]$. Then, we set $d_r := d_{r_1} \times \cdots \times d_{r_t}$. For a dataset $D \in \mathcal{D}$, we define a function

$$p_r : \mathcal{D} \to [0,1]^{d_r}$$

that extracts the marginal joint distribution over the $(r_1, \dots, r_t)$-th attributes from a dataset D. Namely, for $a \in A_{r_1} \times \cdots \times A_{r_t}$, the a-th component of $p_r(D) \in [0,1]^{d_r}$ is

$$p_r(D)_a = \frac{1}{N} |\{x \in D \mid (x_{r_1}, \dots, x_{r_t}) = a\}| \in [0,1].$$

In this manner, the frequency of tabular data is represented as a probability distribution, which is then learned by synthetic data generation methods.

2.3 Differential Privacy

In this subsection, we introduce differential privacy.

2.3.1 Overview of differential privacy

Differential privacy, proposed by Dwork in 2006 [31], is a privacy-preserving metric that quantitatively expresses the degree of privacy protection when releasing statistical information derived from private data. The degree of privacy protection is represented by the privacy loss parameter ε, and for a given setting of ε, the randomness of the output is increased to the extent necessary to guarantee the corresponding level of privacy protection. Differential privacy has been employed in services provided by companies such as Apple [5] and Google [35, 10], and was also applied in the release of data from the 2020 United States Census to ensure privacy protection [113].

Furthermore, two key properties—namely the *Composition Theorem* and the *Post-processing Theorem*—are considered to be major factors contributing to the widespread adoption of differential privacy. The Composition Theorem states that guarantees the overall level of differential privacy when multiple differentially private mechanisms are combined. The Post-processing Theorem states, in essence, that once an output has been protected under differential privacy, any subsequent processing of that output—so long as it does not access the original data—preserves the same differential privacy guarantees. These two theorems enable the guarantee of privacy protection at the system level, even in complex data analyses involving the combination of multiple mechanisms or further transformations of already processed data. This flexibility, which allows differential privacy to be applied across a wide range of use cases, is considered to be a major factor contributing to its widespread adoption.

Moreover, the standard composition theorem may lead to a conservative estimation of security, potentially guaranteeing only a weaker level of security than what the mechanism intrinsically satisfies. In the context of the privacy-utility trade-off, such overly conservative estimations are highly undesirable and should be avoided whenever possible. To address this issue and achieve tighter composition bounds, several relaxations of differential privacy, such as Rényi differential privacy and zero-concentrated differential Privacy, have been proposed.

These properties are, in fact, used in the context of synthetic data generation and analysis, which forms the primary focus of this study. During the training of the generative parameters, it is necessary to add noise into multiple output results, which can be evaluated based on the composition theorem. Furthermore, once differentially private generative parameters have been obtained, any data subsequently sampled from them is guaranteed to satisfy differential privacy by virtue of the

post-processing theorem.

2.3.2 Definitions

In this section, we provide the definitions of differential privacy and its relaxations. First, we define neighboring datasets, which is an essential concept for differential privacy.

Definition 2.3.1 (Neighboring Datasets)**.** *Datasets $D, D' \in \mathcal{D}$ are* **neighboring datasets** *if D and D' are different only in one record. When datasets have a fixed size n, we call the neighboring condition a* **bounded condition** [69]*. In this case, neighboring means changing the value of exactly one record. When datasets have no such restriction, we call the neighboring condition an* **unbounded condition** [69]*. In this case, neighboring means either adding or removing one record.*[1]

(ε, δ)-differential privacy [31, 32] is defined as follows.

Definition 2.3.2 (Differential Privacy [31])**.** *A randomized function $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ satisfies* **(ε, δ)-differential privacy** *((ε, δ)-DP) if for any neighboring $D, D' \in \mathcal{D}$ and any output range $S \subset \mathcal{Y}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta.$$

In particular, $\mathcal{M}$ satisfies ε-DP if it satisfies $(\varepsilon, 0)$-differential privacy (ε-DP). We also call the values ε and δ the privacy loss budgets. We regard non-differentially private algorithm as $\varepsilon = \infty$.

The value δ represents the allowable failure probability, and it is generally recommended to set δ to a value smaller than $1/N$, where N denotes the number of records in the dataset.

Remark 2.3.3. *We can interpret the condition "for any neighboring $D, D' \in \mathbb{D}$" as considering, in a sense, the worst-case input. That is, the above inequality is guaranteed to hold even for a dataset $D \in \mathcal{D}$ in which a single individual's change has the greatest possible impact on the output*

[1] This difference is important for the sensitivity of queries. For example, the sensitivity of the mean value query under the bounded condition is twice as large as that under the unbounded condition.

The original definition of differential privacy is based on the ratio of probabilities. However, numerous subsequent definitions and relaxations have been proposed, many of which aim to bound the divergence of probability density functions.

The following Hockey stick divergence is helpful for describing the differential privacy [27].

Definition 2.3.4 (Hockey Stick Divergence)**.** *Let P, Q be probability distributions on $\mathbb{R}^d$. For $\alpha \geq 0$, the* **α-hockey-stick divergence** *is given as*

$$D_\alpha^h(P||Q) := \sup_{S \subset \mathbb{R}^d} \max\{P(S) - \alpha \cdot Q(S), 0\}.$$

Here, $P(S)$ and $Q(S)$ are $\int_S P(x)dx$ and $\int_S Q(x)dx$ respectively.

By using this divergence, we can interpret a mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{R}^d$ satisfying (ε, δ)-differential privacy. For any neighboring datasets $D, D' \in \mathcal{D}$, it holds that

$$D_{e^\varepsilon}^h(\mathcal{M}(D)||\mathcal{M}(D')) \leq \delta.$$

As a basic type of divergences, we introduce Rényi divergence, which is necessary to define Rényi differential privacy.

Definition 2.3.5 (Rényi Divergence)**.** *Let P, Q be probability distributions on $\mathbb{R}^d$. For $\alpha > 1$, the* **Rényi divergence** *of order α is*

$$D_\alpha(P||Q) := \frac{1}{\alpha - 1} \log \left(\int_{\mathbb{R}^d} P(x)^\alpha Q(x)^{1-\alpha} dx \right).$$

Definition 2.3.6 (Rényi Differential Privacy [87])**.** *For $\alpha > 1$ and $\varepsilon > 0$, a randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{R}^d$ satisfies (α, ε)-***Rényi differential privacy** *((α, ε)-RDP) if for neighboring datasets $D, D' \in \mathcal{D}$,*

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \varepsilon.$$

The smaller ε is, the stronger the protection, and the larger α is, the stronger the protection. To satisfy (α, ε)-RDP for any α is equivalent to ε-DP.

The composition theorem [32, 67] holds for Rényi differential privacy as well as (ε, δ)-DP. Furthermore, Rényi differential privacy can be translated into (ε, δ)-DP.

Proposition 2.3.7 (Translation from (α, ε)-RDP to (ε, δ)-DP [87])**.** *If $\mathcal{M}$ is an (α, ε)-RDP mechanism, it also satisfies $(\varepsilon + \frac{\log \frac{1}{\delta}}{\alpha - 1}, \delta)$-DP for any $0 < \delta < 1$.*

Next, we define zero concentrated differential privacy.

Definition 2.3.8 (zero-Concentrated Differential Privacy [13]). *Let $\rho > 0$. A randomized function $\mathcal{M} : \mathcal{D} \to \mathbb{R}^d$ satisfies ρ-***zero-concentrated differential privacy** *(ρ-zCDP) if for all $\alpha > 1$ and neighboring datasets $D, D' \in \mathcal{D}$,*

$$D_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \rho \cdot \alpha.$$

Proposition 2.3.9 (Translation from ρ-zCDP to (ε, δ)-DP [87]). *If $\mathcal{M}$ is a ρ-zCDP mechanism, it also satisfies (ε, δ)-DP for any $\varepsilon \geq 0$ and*

$$\delta = \inf_{\alpha > 1} \frac{e^{(\alpha-1)(\alpha\rho-\varepsilon)}}{\alpha - 1} \left(1 - \frac{1}{\alpha}\right)^\alpha.$$

2.3.3 Differentially Private Mechanisms

We introduce mechanisms satisfying differential privacy. The simplest way to achieve differential privacy is to add random noise to the result of a query or computation. To determine the scale of the noise, the following sensitivity is important.

Definition 2.3.10 (Sensitivity). *Let $q : \mathcal{D} \to \mathbb{R}^d$ be a query. We define the L_1* **sensitivity** *as*

$$\Delta_1 := \max_{D, D' \in \mathcal{D}, D \sim D'} ||q(D) - q(D')||_1$$

and the L_2 **sensitivity** *as*

$$\Delta_2 := \max_{D, D' \in \mathcal{D}, D \sim D'} ||q(D) - q(D')||_2,$$

where $D \sim D'$ expresses that D and D' are neighboring.

Definition 2.3.11 (Additive Noise Mechanism). *Let $q : \mathcal{D} \to \mathbb{R}^d$ be a query. Let P be a probabilistic distribution, and X be a random variable with P. A randomized function $\mathcal{M} : \mathcal{D} \to \mathbb{R}^d$ defined as*

$$\mathcal{M}(D) = q(D) + X$$

is called an **additive noise mechanism**.

Definition 2.3.12 (Laplace Mechanism). *Define a probabilistic distribution on $\mathbb{R}^d$ as*

$$f_{\varepsilon,\Delta}(x) := \left(\frac{\varepsilon}{2\Delta}\right)^d e^{-\frac{\varepsilon}{\Delta}\|x\|_1}.$$

An additive noise mechanism using a random variable X_{lap} with $f_{\varepsilon,\Delta}(x)$ is called **Laplace mechanism**. *This mechanism satisfies ε-differential privacy. This is equivalent to applying the Laplace mechanism with $d = 1$ independently to each component. Therefore, the implementation is straightforward.*

Definition 2.3.13 (Staircase Mechanism [42, 41]). *For $r > 0$, set*

$$B_d(r) := \{x \in \mathbb{R}^d \mid \|x\|_1 < r\},$$

which is an open L_1-ball with radius r in $\mathbb{R}^d$. Let $\gamma \in [0, 1]$. Define a probabilistic distribution on $\mathbb{R}^d$ as

$$f_{\varepsilon,\Delta,\gamma}(x) = C_\gamma e^{-k\varepsilon} \quad \text{if} \quad x \in B_d((k+\gamma)\Delta) \backslash B_d((k+\gamma-1)\Delta), \tag{2.1}$$

where $k \in \mathbb{Z}_{\geq 0}$. Here, C_γ is a normalization term and described as

$$C_\gamma = \frac{d!}{(1-e^{-\varepsilon})2^d \Delta^d S_d(\gamma, \varepsilon)},$$

and we set

$$S_d(\gamma, \varepsilon) := \sum_{k=0}^{\infty} e^{-k\varepsilon}(\gamma + k)^d. \tag{2.2}$$

An additive noise mechanism using a random variable X_{st} with $f_{\varepsilon,\Delta,\gamma}(x)$ is called **Staircase mechanism**. *For any $\gamma \in [0, 1]$, this mechanism satisfies ε-differential privacy.*

Definition 2.3.14 (Gaussian Mechanism). *Let $f(x)$ be a probabilistic distribution on $\mathbb{R}^d$ described as*

$$f_\sigma(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^d e^{-\frac{||x||_2}{\sigma^2}}.$$

An additive noise mechanism using a random variable X_{gau} with $f_\sigma(x)$ is called **Gaussian mechanism**. *If the inequality*

$$\sigma^2 > \frac{2\Delta \log \frac{1.25}{\delta}}{\varepsilon^2}$$

holds, the Gaussian mechanism satisfies (ε, δ)-differential privacy. If the inequality

$$\sigma > \frac{\Delta^2}{2\rho}$$

holds, the Gaussian mechanism satisfies ρ-zero concentrated differential privacy [13].

The Gaussian mechanism is widely employed in various applications, including Differentially Private Stochastic Gradient Descent (DP-SGD), which integrates differential privacy into stochastic gradient descent [1].

Beyond additive noise mechanisms, the Exponential Mechanism is the most widely recognized approach to achieving differential privacy without directly modifying outputs via noise addition.

Definition 2.3.15 (Exponential Mechanism [84]). *Set the candidates of outputs as* $\mathcal{C} = \{c_1, \ldots, c_t\}$. *Let* $s : \mathcal{D} \times \mathcal{C} \to \mathbb{R}$ *be a score function. Set*

$$\Delta := \sup_{D, D' \in \mathcal{D}, D \sim D', c \in \mathcal{C}} |s(D, c) - s(D', c)|.$$

Then, a randomized function $\mathcal{M} : \mathcal{D} \to \mathcal{C}$ *such that*

$$\Pr[\mathcal{M}(D) = c] = \frac{\exp(\frac{\varepsilon}{2\Delta} s(D, c))}{\sum_{c' \in \mathcal{C}} \exp(\frac{\varepsilon}{2\Delta} s(D, c'))}$$

is called **exponential mechanism**. *This mechanism satisfies* ε*-differential privacy. According to an existing work* [82]*, it also satisfies* $\frac{\varepsilon^2}{8}$*-zCDP.*

2.3.4 Properties

In this section, we introduce the important properties of differential privacy.

Composition Theorem

When multiple outputs are generated from mechanisms that each satisfy differential privacy, the overall privacy guarantee of the system can be expressed as the sum of the individual privacy loss parameters ε. This property enables formal guarantees of privacy even when combining various processing steps, making it possible to construct complex data analysis while preserving differential privacy. Such composability is one of the key features that contribute to the widespread adoption of differential privacy in practice.

Proposition 2.3.16 (Composition of Differential Privacy [67]). *Let* $\mathcal{M}_1 : \mathcal{D} \to \mathbb{R}^{d_1}$ *be* $(\varepsilon_1, \delta_1)$*-DP and* $\mathcal{M}_2 : \mathcal{D} \times \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ $(\varepsilon_2, \delta_2)$*-DP. Then the mechanism* $\mathcal{M} : \mathcal{D} \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ *defined as*

$$\mathcal{M}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D, \mathcal{M}_1(D)))$$

satisfies $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$*-DP.*

Proposition 2.3.17 (Advanced Composition of Differential Privacy [67]). *Let* $\mathcal{M}_i : \mathcal{D} \to \mathbb{R}^d$ *be* (ε, δ)*-DP for* $1 \leq i \leq k$. *Then the k-fold adaptive mechanism* $\mathcal{M} : \mathcal{D} \to \mathbb{R}^{d \times k}$ *defined as*

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

satisfies $(k\varepsilon(e^{\varepsilon} - 1) + \varepsilon\sqrt{2k \log(1/\delta')}, k\delta + \delta')$*-DP. Here,* $\mathcal{M}_i$ *can look at the dataset* D *and the previous outputs* $\mathcal{M}_1(D), \ldots, \mathcal{M}_{i-1}(D)$.

For Rényi differential privacy, the following composition theorem is known.

Proposition 2.3.18 (Composition of Rényi Differential Privacy [87]). *Let $\mathcal{M}_1 : \mathcal{D} \to \mathbb{R}^{d_1}$ be (α, ε_1)-RDP and $\mathcal{M}_2 : \mathcal{D} \times \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ (α, ε_2)-RDP. Then the mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ defined as*

$$\mathcal{M}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D, \mathcal{M}_1(D)))$$

satisfies $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP.

For zero concentrated differential privacy, the following composition theorem is known.

Proposition 2.3.19 (Composition of zero-Concentrated Differential Privacy). *Let $\mathcal{M}_1 : \mathcal{D} \to \mathbb{R}^{d_1}$ be ρ_1-zCDP and $\mathcal{M}_2 : \mathcal{D} \times \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ ρ_2-zCDP. Then the mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ defined as*

$$\mathcal{M}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D, \mathcal{M}_1(D)))$$

satisfies $(\rho_1 + \rho_2)$-zCDP.

As a refinement of the standard composition theorems, Gopi et al. [46] introduced the Numerical Composition method, which enables the computation of tighter cumulative privacy bounds by numerically evaluating the privacy loss rather than relying on conservative analytical estimates.

In general, methods for analyzing the privacy loss of differential privacy mechanisms aim to derive upper bounds, ensuring conservative guarantees from the perspective of safety. However, some studies have also investigated lower bounds, which indicate regions where privacy guarantees are empirically violated, based on the results of practical attacks [92, 106].

Post-Processing Theorem

Proposition 2.3.20 (Post-Processing Theorem). *Let $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ be an ε-DP mechanism and $f : \mathcal{Y} \to \mathcal{Z}$ be a deterministic (or probabilistic) function. Then, the composition of $f \circ \mathcal{M} : \mathcal{D} \to \mathcal{Z}$ also satisfies ε-DP.*

What this proposition implies is that, once an output satisfies differential privacy, any subsequent processing that does not depend on the underlying dataset does not affect the overall privacy guarantee. This property ensures that post-processing operations—so long as they are independent of the original data—do not compromise differential privacy.

2.4 Synthetic Data Generation

2.4.1 Overviews

In this paper, we define synthetic data generation as follows.

Definition 2.4.1 (Synthetic Data Generation)**.** *Let $\mathcal{F}_{\text{ext}} : \mathcal{D} \to \mathbb{R}^r$ be a deterministic function and $\mathcal{F}_{\text{gen}} : \mathbb{R}^r \to \mathcal{D}$ be a probabilistic function. We call a composition of these function $\mathcal{F} : \mathcal{D} \to \mathcal{D}$ a synthetic data generation. We also call $\mathcal{F}_{\text{ext}} : \mathcal{D} \to \mathbb{R}^r$ an extraction function and $\mathcal{F}_{\text{gen}} : \mathbb{R}^r \to \mathcal{D}$ a generation function.*

Definition 2.4.2 (Differentially Private Synthetic Data Generation)**.** *If the extraction $\mathcal{F} : \mathcal{D} \to \mathcal{D}$ satisfies differential privacy, the synthetic data generation is called differentially private synthetic data generation. In most cases, the extraction $\mathcal{F}_{\text{ext}} : \mathcal{D} \to \mathbb{R}^r$ satisfies differential privacy, and thus, the entire function $\mathcal{F} : \mathcal{D} \to \mathcal{D}$ also satisfies differential privacy by the postprocessing theorem.*

A wide range of synthetic data generation methods that can be formulated in this manner have been extensively studied by many researchers [119]. Broadly speaking, synthetic data generation methods can be categorized into three classes: synthesis based on statistical summaries [48], synthesis using graphical models, and synthesis using deep learning models [2]. In particular, recent years have seen a growing number of synthetic data generation approaches that leverage large language models (LLMs) as deep learning-based generators [114, 126, 47, 23]. There is also a way to synthesize plausible tokens in the in-context learning under differential privacy [118].

In addition, the study of evaluation methods for synthetic data quality has become an active area of research [36, 57, 62, 104, 117]. Although the specific evaluation methodologies are detailed in Chapter 4, it is common to evaluate synthetic data quality using two major approaches: (i) measuring the statistical similarity between the synthetic and real datasets, and (ii) evaluating the performance of machine learning models trained on the synthetic data. Regarding utility, there exist benchmarking studies such as [112] that systematically evaluate the performance of synthetic data generation methods. Some of these studies also consider additional aspects beyond utility, such as generation time [129].

In addition to tabular data, there has been active research on differentially private data synthesis for other data modalities [125]. Notably, recent studies have explored the generation of time-series data such as electrocardiograms (ECG) and electroencephalograms (EEG) with differential privacy [79].

Algorithm 1 PrivBayes

Require: $D \in A^N$: Dataset, $\varepsilon = (\varepsilon_g, \varepsilon_p)$: privacy loss budget, $N' \in \mathbb{N}$: the number of outputs, θ: utility threshold

Ensure: $\hat{D} \in A^{N'}$: synthetic dataset

1: $G \leftarrow \texttt{Str}(D, V, \varepsilon_g, \mathrm{PC}_\theta, I)$

2: $P \leftarrow \texttt{Param}(D, G, \varepsilon_p)$

3: Based on the learned graph structure G and the parameter set P, generate a synthetic dataset $\hat{D}$ consisting of N' records.

4: **return** $\hat{D}$

Furthermore, there has been research on active synthetic data generation, which aims to generate data for labeling purposes while preserving privacy [102].

2.4.2 Synthesis Algorithms

In this section, we explain several synthesis algorithms we focus on in this thesis. In particular, since the focus here is on methods targeting categorical attributes, it is assumed that any numerical attributes have been appropriately transformed into categorical ones.

PrivBayes

PrivBayes is a Bayesian network-based differentially private synthetic data generation method proposed by Zhang et al. [123]. The learning process of PrivBayes can be devided into two stages: structure learning `Str`() as shown in Algorithm 2 and parameter learning `Param`() as shown in Algorithm 3. The overall procedure is shown in Algorithm 1 and Figure 2.1.

In the structure learning step, the attributes $A_1, \ldots, A_M$ are regarded as nodes, and connected by edges if they have significant relationships. To facilitate efficient data generation, the structure is learned as a directed acyclic graph (DAG). In this graph, each directed edge points from a parent node to its child node. The strength of the relationship between attributes is measured using metrics such as mutual information.

Definition 2.4.3 (Mutual Information)**.** *Let X be a random variable with a probabilistic distribution on $[d_1]$, and Π a random variable with a probabilistic distributions*

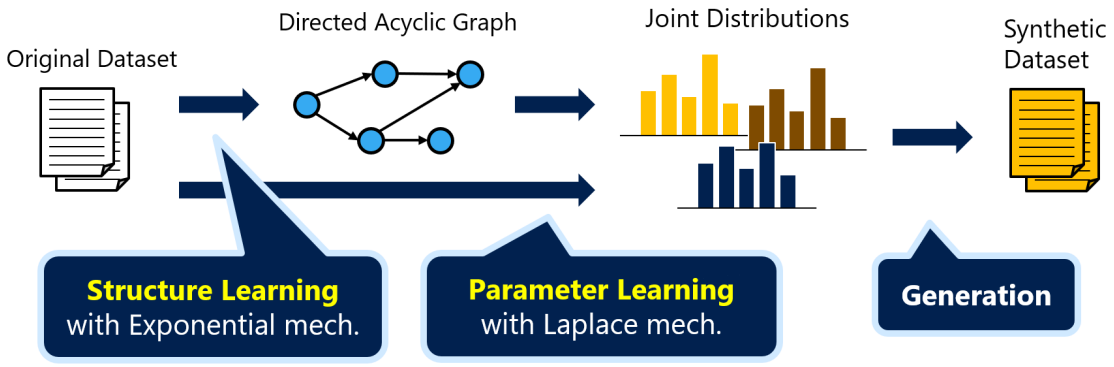

Figure 2.1: The overview of PrivBayes: First, a graph structure is learned in the structure learning phase, where each attribute is represented as a node. Next, in the parameter learning phase, the joint distributions are estimated based on frequency counts. Finally, new synthetic data is generated using the learned graph structure and the estimated joint distributions.

on $[d_2]$. *Then, a real value*

$$I(X,\Pi) := \sum_{\substack{x\in\mathrm{dom}(X),\\ \pi\in\mathrm{dom}(\Pi)}} \Pr[(X,\Pi)=(x,\pi)] \log \frac{\Pr[(X,\Pi)=(x,\pi)]}{\Pr[X=x]\Pr[\Pi=\pi]}$$

is called **mutual information** *between* X *and* Π.

For notational convenience, we denote the set of nodes by $V = [M]$. In PrivBayes, structure learning is performed by iteratively selecting parent sets for each node through the following process:

1. All feasible combinations of parent nodes are enumerated for a given node.
2. Among these candidates, the parent set is selected based on statistical criteria, such as maximizing mutual information.

Define a function $\mathrm{PC} : V \times 2^V \to 2^{(2^V)}$ whose inputs are a node and all possible parents, and output is a set of candidates of parents. Namely, for a node $X \in V$ and the candidate of parents $V' \subset V$, we obtain the candidates of parents $\mathrm{PC}(X, V') = \{P_1, \ldots, P_t\}$. Next, among the valid candidate parent sets, one is selected in a differentially private manner using the Exponential Mechanism, which assigns higher selection probabilities to those with greater mutual information. Directed edges are then added from each selected parent node to its corresponding child node,

Algorithm 2 Structure learning: $\texttt{Str}(D, V, \varepsilon_g, \mathrm{PC}, s)$

Require: D: Dataset, V: Nodes, ε_g: privacy loss budget for structure learning, PC: parents candidate function, s: score function

Ensure: E: Edge

$E = \emptyset, V_{\mathrm{done}} = \emptyset$
Choose the first node $X_1 \in V$ at random, and V_{done}.append(X_1).
for $i = 2, \ldots, d$ **do**
    $C_{\mathrm{tmp}} = \emptyset$
    for $Y \in V \backslash V_{\mathrm{done}}$ **do**
        for $T \in \mathrm{PC}(Y, V_{\mathrm{done}})$ **do**
            C_{tmp}.append((Y, T))
        end for
    end for
    $(X_i, P) \leftarrow$ Exp-Mech(s, D, C_{tmp})
    V_{done}.append(X_i)
    for $Y \in P$ **do**
        E.append((Y, X_i))
    end for
end for
return E

thereby constructing the directed graph. The overall procedure is summarized in Algorithm 2. Note that, in the algorithm, mutual information is abstracted as a scoring function $s : \mathcal{D} \times V \times 2^V \to \mathbb{R}$.

In practical implementations, including those of DataSynthesizer[2] [98] and Synthcity[3] [99], the candidate parent sets are determined by the function $\mathrm{PC}_k : V \times 2^V \to 2^{(2^V)}$, which defines a mapping from a node and a set of previously considered nodes to a family of candidate subsets. This function constrains the number of parent nodes to a fixed value $k \in \mathbb{Z}_{\geq 0}$ and can be formally expressed as follows.

$$\mathrm{PC}_k(X, V') = \{V'' \in 2^{V'} \mid |V''| = k\}.$$

In the parameter learning phase, the joint distribution is learned in a differentially private manner based on the graph structure $G = (V, E)$ obtained from the structure

[2] `https://github.com/DataResponsibly/DataSynthesizer`
[3] `https://github.com/vanderschaarlab/synthcity`

Algorithm 3 Parameter learning: $\texttt{Param}(D, G, \varepsilon_p)$

Require: D: Dataset, $G = (V, E)$: graph structure, ε_p: parameter privacy loss budget, $\mathcal{A}$: additive noise mechanism

Ensure: P: joint distribution with respect to nodes

$P = []$
for $i \in V$ **do**
  if i has no parents **then**
    $p \leftarrow p_i(D)$
  else
    Collect the candidates of parents of i, and make a set r.
    $p \leftarrow p_r(D)$
  end if
  $\hat{p} \leftarrow \mathcal{A}(p)$
  Set all negative entries in $\hat{p}$ to zero.
  $\hat{p} \leftarrow \hat{p} / \sum_i \hat{p}_i$.
  P.append($\hat{p}$)
end for
return P

learning step. The learned graph is assumed to be topologically sorted. Parameter learning begins by aggregating the frequencies from each record and normalizing them by the number of total records. After applying the Laplace mechanism to add noise into the estimated frequencies, negative values (if any) are rounded up to zero to ensure validity. The resulting values are then normalized to form a valid probability distribution, i.e., their sum is adjusted to be one. The detailed procedure for parameter learning is presented in Algorithm 3.

AIM

The adaptive and iterative mechanism (AIM) is a graphical model based synthetic data generation proposed by McKenna [82]. In this approach, the parameters of the graphical model are updated so as to minimize the workload error defined below, while carefully managing the privacy budget based on zero-Concentrated Differential Privacy (zCDP). This method is known for generating high-quality data [18].

Definition 2.4.4 (Workload)**.** *A workload $\mathcal{W}$ is a list of marginal queries $r_1, \ldots, r_k \subset$*

$[M]$. *We also define* $\mathcal{W}_+$ *as*

$$\mathcal{W}_+ := \{s \in 2^{[M]} \mid s \subset r, r \in \mathcal{W}\}.$$

Thus, it holds that

$$\mathcal{W} \subset \mathcal{W}_+.$$

Definition 2.4.5 (Workload Error)**.** *Let* $\mathcal{W}$ *be a workload which consists of a list of marginal queries* $r_1, \ldots, r_k$ *and* $c_1, \ldots, c_k \geq 0$ *be associated weights. The error of a synthetic dataset* $\hat{D}$ *is defined as:*

$$E(D, \hat{D}) := \frac{1}{k \cdot |D|} \sum_{i=1}^{k} c_i ||p_{r_i}(D) - p_{r_i}(\hat{D})||_1.$$

Here, the numbers of attribute values of $A_1, \ldots, A_M$ are $d_i := |A_i| < \infty$, respectively. Set $d := d_1 \times \cdots \times d_M$ and $\theta \in \Delta^d$ as a parameter, which is an initialized total joint distribution.

The overview of algorithm of AIM is as follows.

1. Initialize θ by Algorithm 5.
2. Choice a target query $r \in \mathcal{W}$ by Exponential mechanism.
3. Add Gaussian noise into the contingency table.
4. The parameter θ is optimized based on the Private-PGM framework. The detail of Private-PGM is described in a paper [83].
5. Steps 2, 3, and 4 are repeated within the limits of the privacy budget.
6. Generate synthetic dataset with Private-PGM.

Algorithm 4 AIM: An Adaptive and Iterative Mechanism

Require: D: Dataset, $\mathcal{W}$: workload, ρ: privacy loss budget

Ensure: $\hat{D}$: Synthetic Dataset

Hyper-Parameters: MAX-SIZE=80MB, $T = 16d$, $\alpha = 0.9$

$\sigma_0 = \sqrt{T/(2\alpha\rho)}$ ▷ The budget of Gaussian mechanism

$\epsilon_0 \leftarrow \sqrt{8(1-\alpha)\rho/T}$ ▷ The budget of Exponential mechanism

$\rho_{\text{used}} \leftarrow 0$

$t \leftarrow 0$

Initialize θ_t using Algorithm 5

for $r \in \mathcal{W}$ **do**

    $w_r = \sum_{s \in \mathcal{W}} c_s \mid r \cap s \mid$

end for

while $\rho_{\text{used}} < \rho$ **do**

    $t \leftarrow t + 1$

    $\rho_{\text{used}} \leftarrow \rho_{\text{used}} + \frac{1}{8}\epsilon_t^2 + \frac{1}{2\sigma_t^2}$

    $C_t \leftarrow \{r_t \in \mathcal{W}_+ \mid \text{JT-SIZE}(r_1, \ldots, r_t) \leq \frac{\rho_{\text{used}}}{\rho} \cdot \text{MAX-SIZE}\}$

    select $r_t \in C_t$ using the exponential mechanism with:

$$q_r(D) = w_r \left(\|p_r(D) - p_r(\theta_{t-1})\|_1 - \sqrt{2/\pi} \cdot \sigma_t \cdot d_r \right)$$

    measure marginal on r_t:

$$\tilde{y}_t = p_{r_t}(D) + \mathcal{N}(0, \sigma_t^2 \mathbb{I})$$

    estimate data distribution using Private-PGM:

$$\theta_t = \arg\min_{\theta \in S} \sum_{i=1}^{t} \frac{1}{\sigma_i} \|p_{r_i}(\theta) - \tilde{y}_i\|_2^2$$

    anneal ϵ_{t+1} and σ_{t+1} using Algorithm 6

end while

generate synthetic data $\hat{D}$ from $\hat{p}_t$ using Private-PGM

return $\hat{D}$

Algorithm 5 Initialize θ_t (subroutine of Algorithm 4)

for $r \in \{r \in \mathcal{W}_+ \mid |r| = 1\}$ **do**
  $t \leftarrow t + 1, \quad \sigma_t \leftarrow \sigma_0, \quad r_t \leftarrow r$
  $\tilde{y}_t = p_r(D) + \mathcal{N}(0, \sigma_t^2 I)$
  $\rho_{\text{used}} \leftarrow \rho_{\text{used}} + \frac{1}{2\sigma_t^2}$
end for
$\theta_t = \arg\min_{\theta \in S} \sum_{i=1}^{t} \frac{1}{\sigma_i} \|p_{r_i}(\theta) - \tilde{y}_i\|_2^2$ ▷ Private-PGM

Algorithm 6 Budget annealing (subroutine of Algorithm 4)

if $\|p_{r_t}(\theta_t) - p_{r_t}(\theta_{t-1})\|_1 \leq \sqrt{2/\pi} \cdot \sigma_t \cdot d_{r_t}$ **then**
  $\epsilon_{t+1} \leftarrow 2 \cdot \epsilon_t$
  $\sigma_{t+1} \leftarrow \sigma_t / 2$
else
  $\epsilon_{t+1} \leftarrow \epsilon_t$
  $\sigma_{t+1} \leftarrow \sigma_t$
end if
if $(\rho - \rho_{\text{used}}) \leq 2\left(\frac{1}{2\sigma_{t+1}^2} + \frac{1}{8}\epsilon_{t+1}^2\right)$ **then**
  $\epsilon_{t+1} = \sqrt{8 \cdot (1 - \alpha) \cdot (\rho - \rho_{\text{used}})}$
  $\sigma_{t+1} = \sqrt{1/(2 \cdot \alpha \cdot (\rho - \rho_{\text{used}}))}$
end if

Chapter 3

Privacy Evaluation Framework for Synthetic Data Generation

3.1 Introduction

The development of machine learning has led to a growing interest in the use of data containing information about individuals. In utilizing such personal data, reducing privacy leakage risk via anonymization techniques [110, 115] is crucial. Although anonymization for high-dimensional tabular datasets is difficult [3], synthetic data generation is known to generate high-quality and privacy-preserved datasets [56, 112].

There are two major approaches to achieving privacy in synthetic data generation. The first approach is to guarantee theoretical privacy as a *pre*-evaluation of datasets. In the pre-evaluation, intentional noise is often added to the output to satisfy differential privacy [31]. Differential privacy assumes the worst-case input dataset (See Remark 2.3.3), and this assumption causes low-quality output datasets [112] as a critical issue. On the other hand, the second approach is to check vulnerabilities after the data synthesis as *post*-evaluation of datasets. In the post-evaluation, synthetic data generation is evaluated regarding resistance to a specific attack to test vulnerabilities for a given dataset. Since differential privacy as the pre-evaluation often needs excessively strong noise to utilize data, there are several situations where post-evaluation is superior. Indeed, in the black-box setting, mechanisms can achieve better robustness to privacy leakage than theoretical results from differential privacy [65, 109]. To obtain a more accurate understanding of privacy risks, membership inference attacks should be introduced to evaluate the privacy of synthetic data generation [120].

One attempt to remedy the above described background is to design an evaluation framework for synthetic data against membership inference attacks through game-based definition [109]. However, this framework has two limitations. First, target samples are chosen in a random way, and the framework cannot choose vulnerable samples in advance. While this framework can only evaluate the average risk of synthetic data, it should be able to evaluate worst-case vulnerability risk as well by following the same spirit of DP. If a company utilizes synthetic data generated from its collected personal data, then it needs to know which individuals are most at risk. Second, since an adversary's inference method is only based on machine learning models, the decision criteria of membership inference is a black box. Namely, the inference results should be interpreted to understand the risk of synthetic data. This interpretation will give us insights into the main factors of privacy leakage and the adversary's advantages.

In this chapter, we propose a framework for evaluation called Setsubun[1] which gives simple and clear solutions to the limitations described above. To cope with the first limitation, we define a new membership inference game with a stronger adversary than the existing game-based definition [109], and then introduce Mahalanobis distance [16] to choose a target sample as a concrete approach. Compared to the conventional Euclidean distance, the Mahalanobis distance is more suitable for detecting outliers, as it takes into account the correlations among variables in the dataset. This enables us to choose high-risk target samples efficiently in a statistical fashion. To cope with the second limitation, we propose interpretable inference methods: a statistic-based inference method and a sample-distance-based inference method. The former proposed method infers with typical statistics scores, whereas the latter proposed method infers with samples whose distance is close to the target samples described above. We also utilize Inference Measure (IM) as an evaluation metric which enables us to evaluate not only binary classification but also the area-under-curve (AUC) [51] of membership inference attacks.

To verify the effectiveness of our proposed framework, we conduct an experiment with two datasets and five data synthesis algorithms: Gaussian Copula [108], Bayesian Networks [123], MWEM-PGM [83], AIM [82], and Conditional Tabular GAN [116]. As a result, we show that the AUC scores of membership inference attacks increase up to 0.4 points by using Mahalanobis distance to choose a target sample, which has a significant impact on the evaluation of membership inference

[1]The groundhog day in the title of the paper [109] is often February 2, whereas Setsubun in Japan is often February 3. We chose the word Setsubun as the next day of the groundhog day.

attacks. We also demonstrate that the proposed interpretable inference methods achieve higher AUC scores than or equal to those of the existing black-box inference method [109]. In other words, the proposed interpretable inference method is able to evaluate privacy more tightly.

3.2 Preliminaries and Related Work

We introduce synthetic data generation techniques and membership inference attacks as related work. Basic notations are as follows. This study focuses on tabular format datasets. In a tabular dataset, a row corresponds to a person, and a column corresponds to an attribute. Let $A_1, \ldots, A_d$ be attributes. We can express a record as an element $x \in A := A_1 \times \cdots \times A_d$. If a dataset D contains N records, we can regard D as an element of A^N and set a universe of datasets as $\mathcal{D} = A^N$.

3.2.1 Synthetic Data Generation

A synthesis algorithm $\mathcal{F} : \mathcal{D} \to \mathcal{D}$ is decomposed into two steps. The first step is an extraction of generative parameters $\mathcal{F}_{ext} : \mathcal{D} \to \mathbb{R}^r$. For example, the extraction is utilized to compute statistics or train machine learning models. The second step is generation, where we generate synthetic data from extracted generative parameters $\mathcal{F}_{gen} : \mathbb{R}^r \to \mathcal{D}$.

In this chapter, we classify data synthesis algorithms for tabular datasets into three types. The first method is a basic statistics or copula-based method [75, 6]. The second method is a graphical-model-based method [123, 124, 82, 83]. The third method is a deep-neural-network-based method [116, 37, 127, 128, 19, 73, 72, 76].

3.2.2 Membership Inference Attacks against Synthetic Data Generation

A membership inference attack is an attack against a trained machine learning model, where an adversary infers whether a target sample is contained in the training dataset of the model by using its query results [107]. A membership inference attack has been utilized for an evaluation metric of privacy in recent years [120]. While research on membership inference attacks originally focuses on a discriminative model [107, 14, 120], several works discuss attacks against generative models. Specifically, many attacks have been found against generative adversarial networks (GANs) [17, 54, 59] and diffusion models [15, 60, 30, 80]. However, these works

mainly discuss attacks based on model parameters instead of outputs of models. It is important to evaluate outputs of generative models, including synthetic data generation, as well as model parameters. A framework to evaluate membership inference attacks against generative models based on outputs is discussed by Stadler et al. [109].

To the best of our knowledge, only the work by Stadler et al. discusses a game-based evaluation framework for synthetic data generation. However, as mentioned in Section 4.1, it contains two limitations, and hence we aim to give their solutions. As described in Section 3.3 in detail, an adversary in our framework is stronger than that in the work by Stadler et al.: for instance, an adversary can arbitrarily choose a high-risk target sample, including the worst-case, as well as knowing the entire dataset. We also note that our framework allows us to deeply understand evaluation results through interpretable inference methods.

3.3 Proposed Framework

In this section, we propose a privacy evaluation framework for synthetic data generation. The membership inference game by Stadler et al. [109] is elegant for evaluating the privacy in synthetic datasets for a fixed dataset. However, it has two limitations as described here:

1. The method of target choice is random. Although the risk should be evaluated in the worst-case, random choice may cause evaluation far from such a worst-case risk. For example, when a company utilizes synthetic data generated from its collected personal data, risk management needs to know which individuals are most at risk. It is important to be able to quantitatively select outliers.

2. The method of inference is only by machine learning models. Black-boxing inferences may make the risk analysis unnecessarily difficult. Indeed, these are often unable to provide the reason why data are at risk. The inference method also indicates that how to generate samples to reduce the risk of membership inference is unclear despite the fact that a target sample is at risk.

Our proposed framework overcomes these limitations and gives them simple and clear solutions:

1. We modify the framework to allow an adversary to choose the most vulnerable sample from the original dataset. We introduce Mahalanobis distance [16] to choose empirically vulnerable samples.

2. We propose simple and interpretable inference methods. One is a statistics-based inference that uses the likelihood ratio of datasets. If the inference in this method is successful, it means that membership for a target sample has a significant impact on typical statistics scores. Another method is a sample-distance-based inference, which uses the distance from the target sample. If the inference in this method is successful, it means that samples whose distance is close to a target sample are generated.

These solutions enable us to evaluate the risk of membership inference against synthetic datasets more tightly.

3.3.1 Definition of Membership Inference Game

The proposed membership inference game is defined as follows. The game consists of three steps between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$ as shown in Figure 4.1. The adversary tries to infer membership of a target sample, and the challenger interacts with the adversary during the game and checks if the adversary succeeds in the inference.

(1) Target Choice. In this step, $\mathcal{C}$ sends a dataset D to $\mathcal{A}$. The adversary $\mathcal{A}$ chooses an arbitrary target sample t from D, and sends it to $\mathcal{C}$. The challenger $\mathcal{C}$ makes a positive dataset $D_{pos} = D$ and a negative dataset $D_{neg} = D \backslash \{t\}$.

(2) Synthesis. The challenger $\mathcal{C}$ flips a coin $b \leftarrow \{0, 1\}$. For $b = 0$, $\mathcal{C}$ sets $D_{orig} = D_{pos}$. Otherwise, for $b = 1$, $\mathcal{C}$ sets $D_{orig} = D_{neg}$. Then, $\mathcal{C}$ generates synthetic dataset $D_{target} \leftarrow S(D_{orig})$, and sends it to $\mathcal{A}$.

(3) Inference. The adversary $\mathcal{A}$ computes an inference measure (IM) from D_{target}, D_{pos}, D_{neg}, t. The adversary returns $b' = 0$ if the IM is large, and returns $b' = 1$ otherwise.

We say that an adversary $\mathcal{A}$ wins the game if $b = b'$ holds. Otherwise, we say that $\mathcal{A}$ loses the game.

Note: The main difference between our framework and Stadler et al. [109] is in the Target Choice phase. In this phase, while the adversary $\mathcal{A}$ in our framework chooses a target sample t *after* receiving a dataset D in the Target Choice phase, an adversary in the framework by Stadler et al. chooses it *before* receiving a dataset

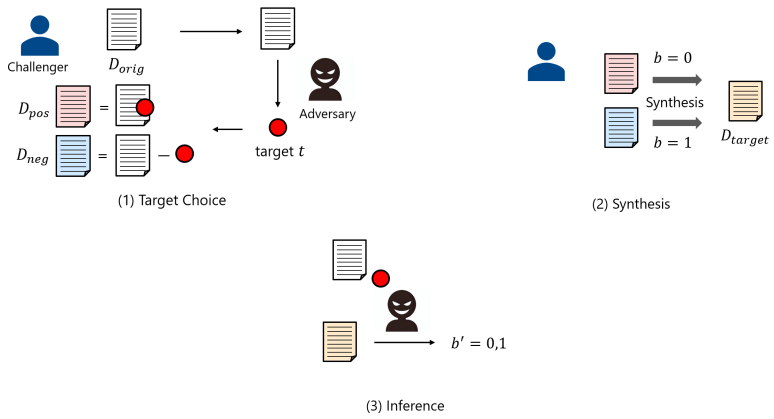

Figure 3.1: (1) Target Choice. (2) Synthesis. (3) Inference.

D in the Syntheis phase[2]. It implies that the adversary $\mathcal{A}$ in our framework can arbitrarily choose a target sample t to maximize its advantages for winning the game because of knowing D itself. Namely, if we obtain a synthesis algorithm such that $\mathcal{C}$ wins the game described above, it will also win the game by Stadler et al. implicitly. We describe how to choose t with the Mahalnobis distance [16] on the Target Choice phase and inference methods on the Inference phase in the remaining parts of this section.

3.3.2 Target Choice

In the target choice phase, the adversary seeks the most vulnerable sample. It is widely known that outlier records are vulnerable to membership inference attacks [14]. Based on the above insight, it is considered plausible to select outliers in a quantitative fashion. We then consider the high-risk target sample as a statistical outlier and, hence, utilize the Mahalanobis distance, which is simple and requires few computational costs.

Definition 3.3.1 (Mahalanobis distance [16]). *For a dataset* $D = \{x_i \in \mathbb{R}^d\}_{i=1,\ldots,n}$, *we set the mean vector*

$$\mu_D = \frac{1}{n}\sum_{i=1}^{n} x_i$$

[2]Although the original description of the framework in [109] does not contain the Target Choice and Syntheis phases explicitly, they correspond to the phases until an adversary returns the guess on the likability privacy game in [109].

and the covariance matrix

$$\Sigma_D = \frac{1}{n}\sum_{i=1}^{n} {x_i}^t x_i - \mu^t \mu,$$

where ${}^t x_i$ *and* ${}^t \mu$ *are transposed vectors of* x_i *and* μ*. For a point* $x \in \mathbb{R}^d$*,*

$$M(D, x) := \sqrt{{}^t(x - \mu_D)\Sigma_D^{-1}(x - \mu_D)}$$

is called the **Mahalanobis distance** *of a point* x*. The larger* $M(D, x)$ *is, the more we can assume that* x *is an outlier in the dataset* D*.*

This distance is the same as the conventional Euclidean distance from the mean μ when $\Sigma_D = I_d$. Compared to the Euclidean distance, the Mahalanobis distance is more suitable for detecting outliers, as it takes into account the correlations among variables in the dataset.

In this study, we assume the adversary chooses the sample with the largest Mahalanobis distance as the target sample.

3.3.3 Inference Methods

The goal of an adversary in the inference step is to infer whether $D_{orig} = D_{pos}$ or $D_{orig} = D_{neg}$ for a given synthetic dataset D_{target}. In this study, we define inference measure (IM), which becomes large when $D_{orig} = D_{pos}$ and small when $D_{orig} = D_{neg}$. We then propose a statistics-based method and a sample-distance-based method as well as utilizing the existing machine-learning-based method [94].

We note that, although machine-learning-based methods are utilized for privacy evaluation in many works [94, 121, 103, 14, 80], their inference results are often uninterpretable. By contrast, the statistics-based method infers results based on typical statistics scores, e.g., mean and covariance matrix. Likewise, the sample-distance-based method infers results based on samples whose distance is close to a target sample. The two methods described above are more interpretable than the machine-learning-based methods for the above mentioned reasons.

Machine-learning-based Inference

In the conventional machine-learning-based inference methods, an adversary trains a machine learning model that infers whether $D_{orig} = D_{pos}$ or $D_{orig} = D_{neg}$ from D_{target}. The model aims to return 1 if D_{target} is obtained from $D_{orig} = D_{pos}$ and 0 if D_{target} is obtained from $D_{orig} = D_{neg}$. Since the synthetic dataset D_{target} itself is

too large to be an input, we investigate two types of inputs as well as Oprisanu et al. [94]:

(1) a chunk of several records;

(2) the histogram of the synthetic dataset.

For (1), we regard a chunk of c_1 records as an input for an inference model $f_{dataset} : \mathbb{R}^{c_1} \to [0, 1]$. For (2), we concatenate histograms of all attributes. We also use a function

$$\text{logit}(x) := \log\left(\frac{x}{1-x}\right)$$

to make it easier to see the distributions of the model's output [14]. To sum up, we define the inference measure as

$$\text{IM}_{\text{dataset}} := \text{logit}(f_{dataset}(D_{target})), \tag{3.1}$$

$$\text{IM}_{\text{hist}} := \text{logit}(f_{hist}(D_{target})), \tag{3.2}$$

where f_*'s are machine learning models.

Statistics-based Inference

In the proposed statistics-based inference method, an adversary assumes that records in $\mathcal{F}(D_{pos})$ and $\mathcal{F}(D_{neg})$ follow some multivariate Gaussian distributions respectively, and attempts to distinguish them by the likelihood ratio at D_{target}. First, an adversary generates c_2 synthetic datasets from D_{pos} and D_{neg} respectively: $D^1_{pos}, \ldots, D^{c_2}_{pos} \leftarrow \mathcal{F}(D_{pos})$, $D^1_{neg}, \ldots, D^{c_2}_{neg} \leftarrow \mathcal{F}(D_{neg})$. Next, the adversary computes

$$\mu_{pos/neg} := \frac{1}{c_2}\sum_{i=1}^{c_2} m(D^i_{pos/neg}),$$

$$\Sigma_{pos/neg} := \frac{1}{c_2}\sum_{i=1}^{c_2} cov(D^i_{pos/neg}),$$

where $m : \mathcal{D} \to \mathbb{R}^d$ is the mean vector function and $cov : \mathcal{D} \to \mathbb{R}^{d\times d}$ is the covariance matrix function. Let f_{pos} be the probabilistic density function of the multivariate Gaussian distribution $\mathcal{N}(\mu_{pos}, \Sigma_{pos})$, and f_{neg} be that of $\mathcal{N}(\mu_{neg}, \Sigma_{neg})$. Then, the adversary computes

$$\text{IM}_{\text{stat}} := \log f_{pos}(m(D_{target})) - \log f_{neg}(m(D_{target})). \tag{3.3}$$

If the IM_{stat} is larger, then the likelihood of D_{target} in f_{pos} is larger. If smaller, that in f_{neg} is smaller.

Sample-distance-based Inference

In the proposed sample-distance-based inference method, an adversary works on the hypothesis that there are more records close to the target sample if $D_{orig} = D_{pos}$. We define a function $dist : \mathcal{D} \times \mathbb{R}^d \times \mathbb{N} \to \mathbb{R}$ as

$$dist(D, t, c_3) := \sum_{x \in D_{t,c_3}} ||x - t||_2,$$

where D_{t,c_3} is a set of samples up to the c_3-th nearest ones with respect to the Euclidean distance, which is the most basic distance for two points. Note that we can use another distance function in this method, but the Mahalanobis distance may be considered inappropriate in this context, as it measures the distance from the mean of the data while taking the covariance structure into account, rather than quantifying the distance between two arbitrary points in the space. We define the inference measure as

$$\mathrm{IM}_{\mathrm{dist}} := -dist(D_{target}, t, c_3). \tag{3.4}$$

If the $\mathrm{IM}_{\mathrm{dist}}$ is larger, more records close to t exist in D_{target}, and we see that the probability of $D_{orig} = D_{pos}$ is higher.

3.4 Experiments

In this section, we conduct an experiment to evaluate our proposed framework. The goal of the experiment is to confirm the effectiveness of the proposed framework.

3.4.1 Experimental Settings

The detailed experimental settings are described below. The detailed experiment algorithm is shown in Algorithm 7.

Datasets

We make use of two datasets: (1) Adult Dataset [29], which consists of nine categorical attributes and six numerical attributes, and (2) California Housing Dataset [95], which consists of nine numerical attributes. For Adult Dataset, we removed records with some missing values, and the number of records was reduced to 30,162. California Housing Dataset consists of 20,640 rows. One row corresponds to one block, but we regard one row as one individual.

Algorithm 7 Experiment algorithm

Require: D_{orig}: original dataset, $\mathcal{F}$: synthetic data generation algorithm, N_{exp}: the number of experiments, N_{gen}: the number of generations

Ensure: pos_list, neg_list : values of IMs

```
pos_list = []
neg_list = []
for i = 1, ..., N_exp do
    t ← D_orig
    D_pos ← D_orig
    D_neg ← D_orig \ {t}
    M ←Prepare inference models(D, t).
    θ_pos ← F_ext(D_pos)
    θ_neg ← F_ext(D_neg)
    for j = 1, ..., N_gen do
        D_pos,target ← F_gen(D_pos)
        D_neg,target ← F_gen(D_neg)
        pos_list.append(M(D_pos,target))
        neg_list.append(M(D_neg,target))
    end for
end for
```

Synthesis Algorithms

We implement five synthesis algorithms: Gaussian Copula (`gcopula`) [108], Bayesian Networks (`bayes`) [123], MWEM-PGM (`mwem-pgm`) [83], AIM (`aim`) [82], and Conditional Tabular GAN (`ctgan`) [116]. We use `gcopula` as a statistics-based synthesis, `bayes`, `mwem-pgm`, `aim` as graphical-model-based synthesis, and `ctgan` as a deep-neural-network-based synthesis. Although `mwem-pgm` and `aim` are proposed as differentially private mechanisms, we change these mechanisms to non-differentially private ones in this study.

Target Choices

We compare the following two target choice methods for a given dataset D. In the **random** case, we randomly choose a target record t from D. In the **mah-max** case, we choose the sample $x_i \in D$ with the largest Mahalanobis distance $M(D, x_i)$ as the target record t. The former case is identical to the previous work [109] while the latter case is identical to our proposed framework. Through comparison

Algorithm 8 Prepare inference models

Require: D_{orig}: original dataset, t: target sample

Ensure: inference model M

1: $D_{pos} \leftarrow D_{orig}$
2: $D_{neg} \leftarrow D_{orig} \backslash \{t\}$
3: $\theta_{pos} \leftarrow \mathcal{F}_{ext}(D_{pos})$
4: $\theta_{neg} \leftarrow \mathcal{F}_{ext}(D_{neg})$
5: Train an inference model M

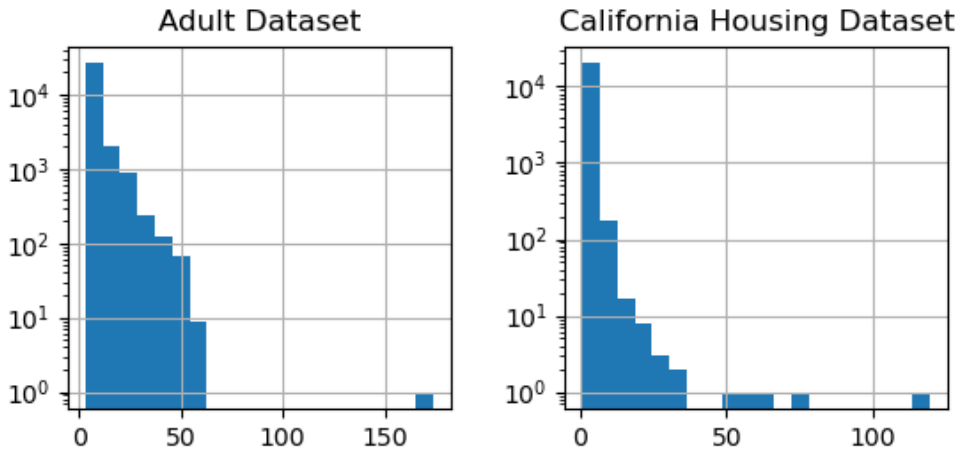

Figure 3.2: Distributions of the Mahalanobis distance. The vertical axis has a logarithmic scale for readability. We found one outlier whose Mahalanobis distance is the largest in each dataset.

between these two target choice methods, we confirm whether a target sample in the worst-case has a significant impact on the experimental results.

We measured the Mahalanobis distance for any sample $x_i \in D$. The largest Mahalanobis distance in Adult Dataset is 173.67, and that in California Housing Dataset is 119.52. The distributions of the Mahalanobis distance for each record are shown in Figure 3.2. Although one might think that the number of outliers is limited, outliers with respect to the Mahalanobis distance exist in both datasets. Based on these outliers in the datasets, we discuss the AUC scores for the worst-case in this thesis. Investigating a relationship between the Mahalanobis distance of the target sample and the AUC scores still remains an open problem.

Table 3.1: Experimental Settings.

Dataset	Adult Dataset, California Housing Dataset
Target Choice	random, mah-max
Synthesis	`gcopula`, `bayes`, `mwem-pgm`, `aim`, `ctgan`
Inference	RF_dataset, RF_hist, LR_dataset, LR_hist, MLP_dataset, MLP_hist, XGB_dataset, XGB_hist, statistics, sample

Inference Methods

We implement ten inference methods that can be categorized into three types: eight machine learning model-based methods, one statistics-based method, and one sample-distance-based method. For machine learning model-based method, we use Random Forest (RF_dataset, RF_hist), Logistic Regression (LR_dataset, LR_hist), Multi Layer Perceptron (MLP_dataset, MLP_hist), and Graident Boosting (XGB_dataset, XGB_hist), which are the same models used in the previous works [109, 94]. In the *_dataset methods, we regard concatenated ten records as one input; that is, we set $c_1 = 10$. The reason for $c_1 = 10$ is to provide the same setting in the previous work [109]. We also set $c_2 = 10$ for statistics-based inference and set $c_3 = 10$ for distance-based inference similarly to c_1. When we conducted a preliminary experiment to evaluate the impact of c_2 and c_3 on experimental results, we did not find a significant difference between the results with respect to c_2 and c_3. Since the computational cost for experiments is also heavy, we $c_2 = c_3 = 10$ to reduce the computational cost. We leave to find other suitable parameters as an open problem.

Number of Trials

For datasets (2 patterns), target choices (2 patterns), synthesis algorithms (5 patterns), and inference methods (10 patterns), the combination of them has $2 \times 2 \times 5 \times 10 = 200$ patterns as shown in Table 3.1. For each pattern, we compute generative parameters three times, $\theta^1_{pos}, \theta^2_{pos}, \theta^3_{pos} \leftarrow \mathcal{F}_{ext}(D_{pos})$ and $\theta^1_{neg}, \theta^2_{neg}, \theta^3_{neg} \leftarrow \mathcal{F}_{ext}(D_{neg})$. For each θ, we generate synthetic datasets that are the same size as the original dataset and compute IMs 100 times. The detailed experiment algorithm is described in Algorithm 7.

Evaluation of Membership Inference Attacks

We introduce the inference measures (IMs). IMs are output scores of the computation. Using IMs, we can compute the AUC, which is a popular measure for evaluating membership inference [55, 86, 103, 80]. We also plot frequency distributions of IMs with $D_{orig} = D_{pos}$ and those with $D_{orig} = D_{neg}$.

Stadler et al. proposed a privacy measure, Privacy Gain [109], but two problems have been pointed out [44]. First, Privacy Gain represents the difference between the advantage of an adversary receiving the original data and an adversary receiving only its synthetic data. Privacy Gain is unavailable in the proposed framework: specifically, our adversary is stronger than that in Ref. [109] because the adversary always knows the original dataset in the target choice phase. Second, Privacy Gain often becomes unstable: for instance, it causes a step-like variation that differs from the intuition of privacy evaluation.

Using AUC enables us to avoid the problems described above. AUC is a well-established measure and it is available even to our adversary. AUC is also interpretable in the context of a typical statistical method and essentially represents the success rate of specific attacks. Consequently, we utilize AUC to evaluate the privacy of synthetic data through IMs.

3.4.2 Experimental Results and Discussion

AUC Scores for Each Condition

The summary of the main results is shown in Figure 3.3 and all AUC scores are shown in Figures 3.4 and 3.5. Since the AUC scores vary across datasets, we separate graphs by datasets. We calculate the mean AUC scores of all synthesis algorithms for each dataset. For Adult Dataset, most of AUC scores are around 0.5, which means that the adversary cannot distinguish the membership at all. When the target choice is random, all inference methods are around 0.5, but when it is mah-max, the results by RF_hist, XGB_hist and sample are relatively larger AUC scores. In particular, the AUC score of sample-distance-based inference, which is a part of our proposed framework, is around 0.8. For California Housing Dataset, when the target choice is random, all results are around 0.5. The results of all inference methods and mah-max increase and are more than 0.6.

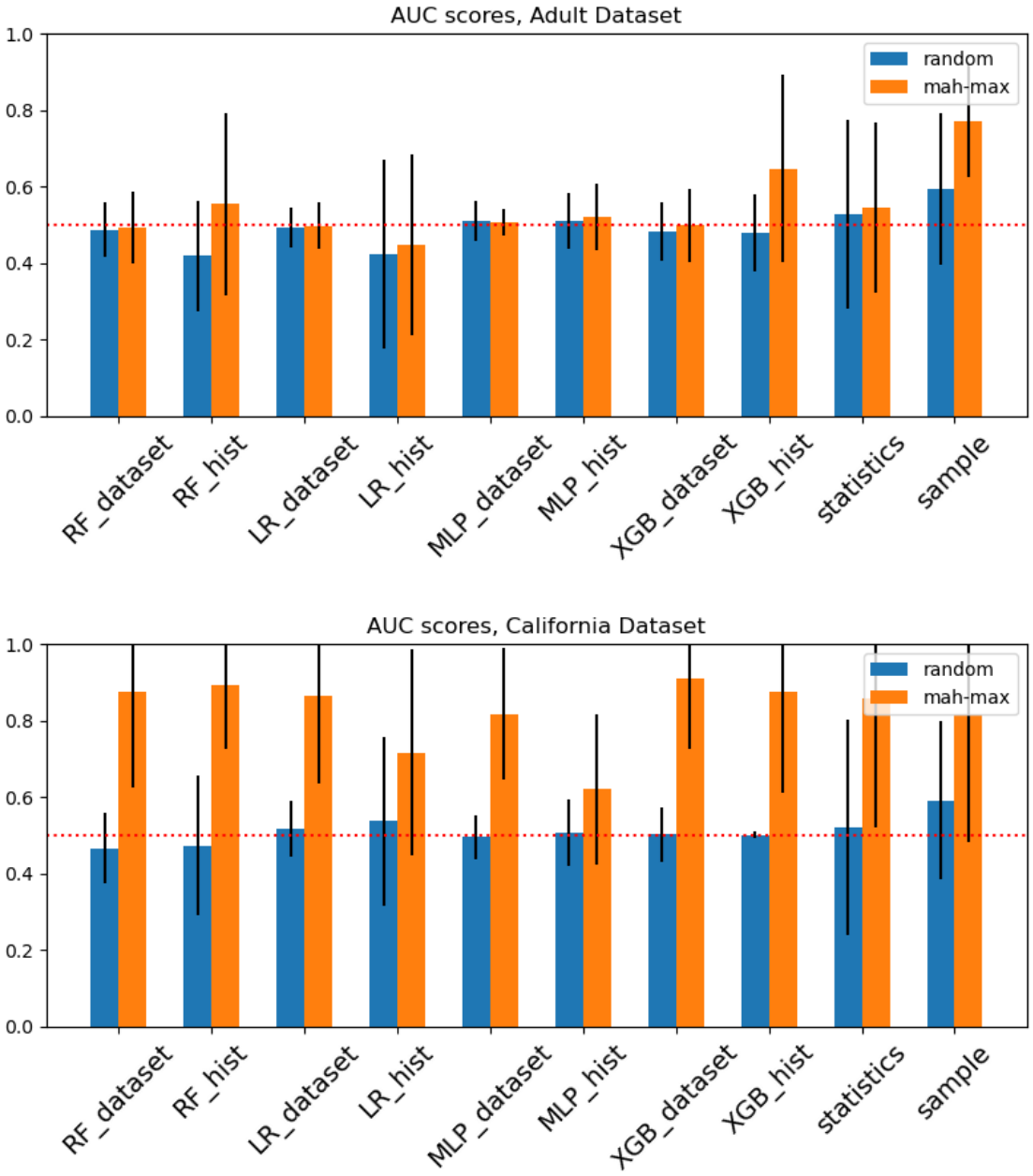

Figure 3.3: The mean AUC scores of all synthesis methods for each inference method. Blue bars are the results when the target choice is random, and orange bars are the result when the target choice is mah-max.

Difference in Target Choices

The difference in target choices is significant. From Figure 3.3, we see that the AUC scores when the target choice is mah-max highly outperform those when it

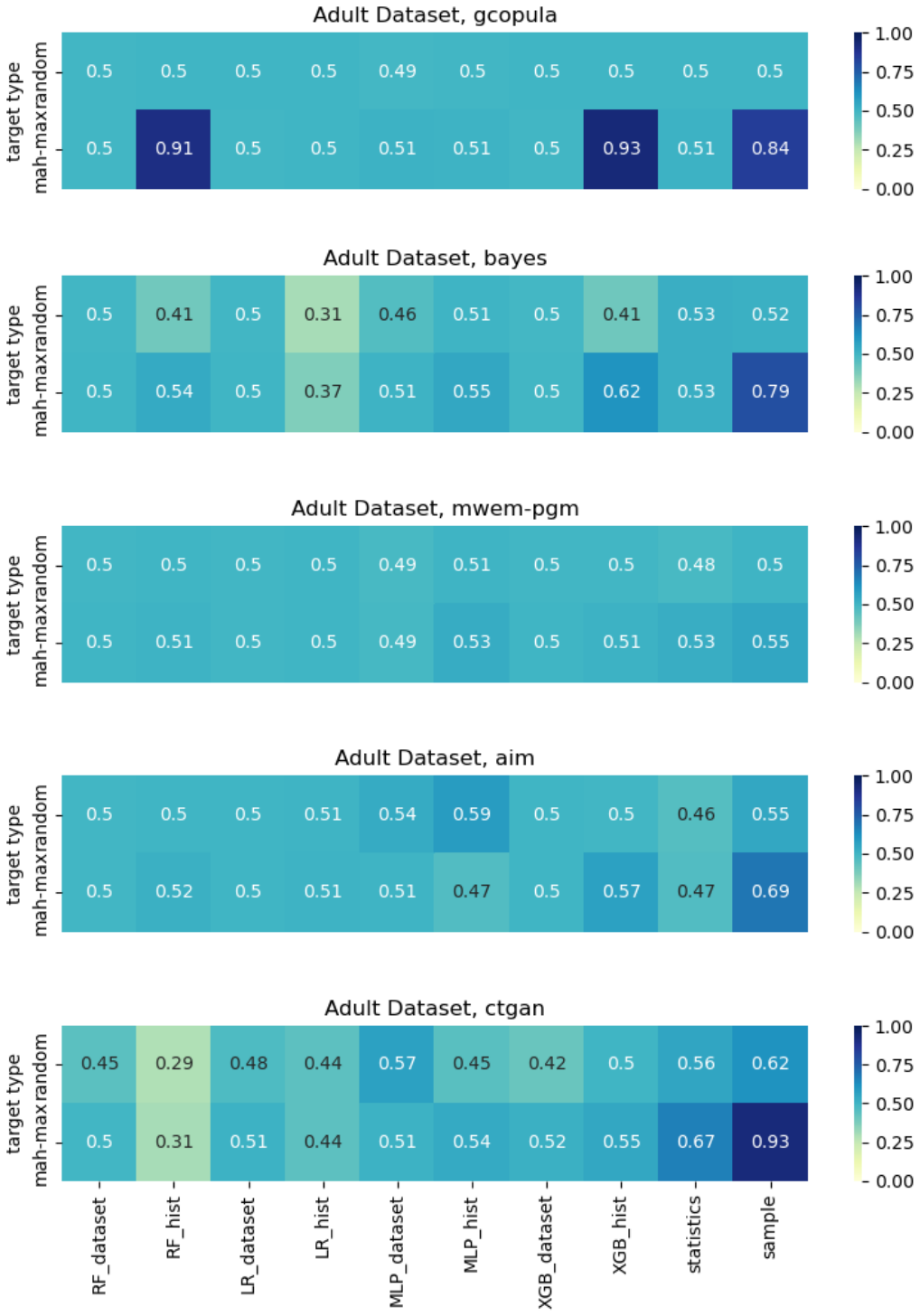

Figure 3.4: AUC scores for Adult Dataset.

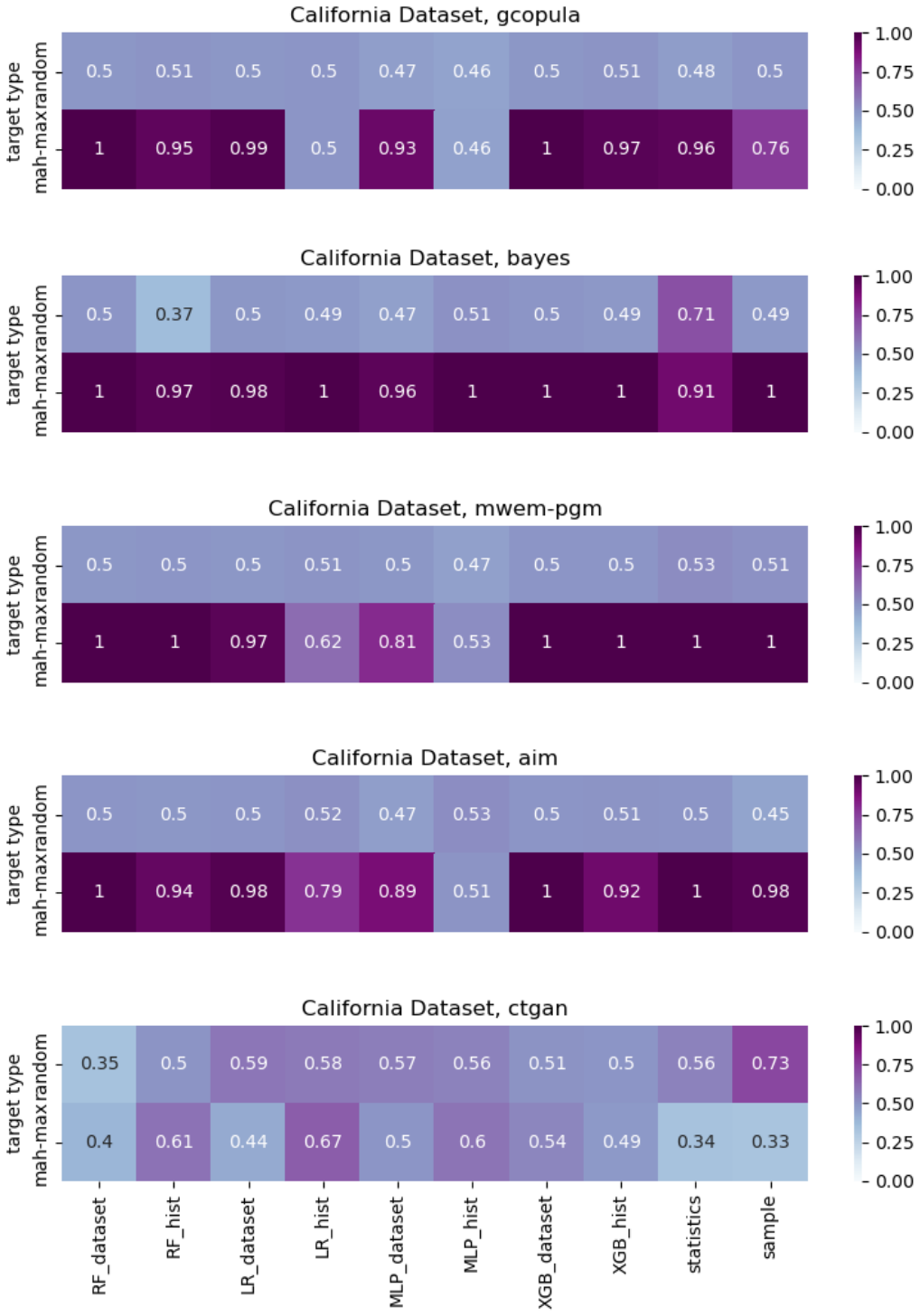

Figure 3.5: AUC scores for California.

is random. AUC scores increase up to 0.4 points with California Housing Dataset. By choosing an outlier sample as a target sample, the risk of membership inference increased significantly. Thus, our framework is effective in evaluating the risk of synthetic data generation.

Difference in Inference Methods

Among inference methods, the sample-distance-based method provides high performance in both datasets. When the dataset is Adult Dataset, it provides the best AUC score. When the dataset is California Housing Dataset, the AUC scores of most of inference methods, including our proposals, are large as shown in Figure 3.3. Despite the simplicity and interpretability of our proposed methods, they are as accurate or better than other inference methods.

For machine learning model-based methods, we found that RF and XGB generally yielded better results. For LR and MLP, we found that retaining data as a dataset (*_dataset) resulted in slightly better performance. In contrast, RF performed slightly better when using histogram-based data retention (*_hist). For XGB, we observed that the histogram approach (*_hist) yielded better performance on the Adult dataset, whereas the dataset-based approach (*_dataset) performed marginally better on the California dataset.

Difference of Synthesis Algorithms

We also show AUC scores for each synthesis algorithm in Figure 3.6. All AUC scores are shown in Figures 3.4 and 3.5. When the dataset is Adult Dataset, the scores are around 0.5 for each synthesis algorithm. When the dataset is California Housing Dataset, the AUC scores except for `ctgan` are large. In particular, the result of `bayes` is the largest.

AUC Scores and Utility

We also check the utility of each synthetic data. We generate a synthetic dataset D_{syn} with the same size as the original dataset D_{orig} and evaluate the utility by the mean of the L_1 distance for each attribute, which is a function widely used in the evaluation of synthetic data [112]. For a categorical attribute with K categorical values, the L_1 distance is computed as

$$L_1(p^{orig}, p^{syn}) = \sum_{i=1}^{K} |p_i^{orig} - p_i^{syn}|,$$

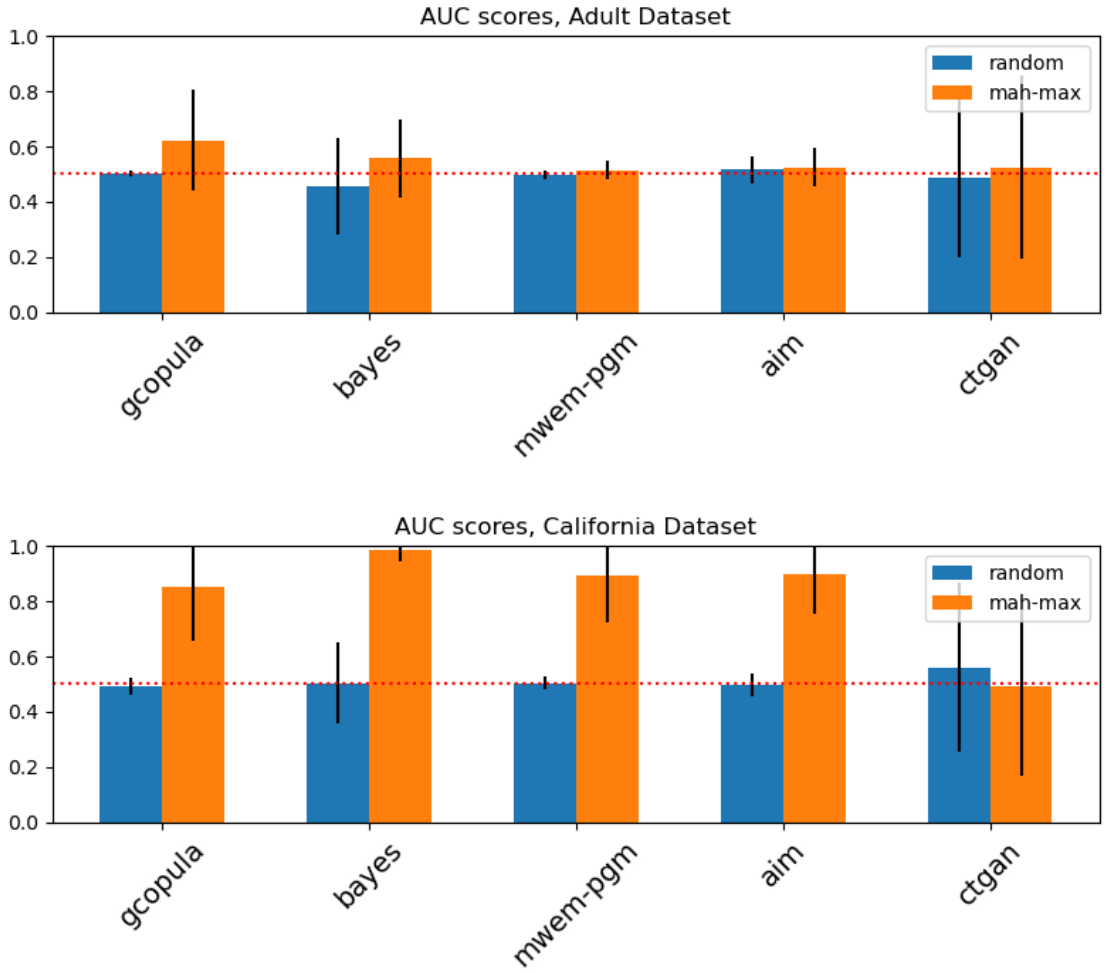

Figure 3.6: The mean AUC scores of all inference methods for each synthesis method. Blue bars are results when the target choice is random, and orange bars are when the target choice is mah-max.

where p^{orig} is a normalized frequency distribution of the original dataset and p^{syn} is that of the synthetic dataset. Thus, the range of L_1 distance is from 0 to 2. For a numerical attribute, we separated them into 20 groups with equal-width ranges.

The graphs shown in Figure 3.7 are the means of the L_1 distance for all attributes. The result showed that the quality of synthetic data generated by the other synthesis than `ctgan` is high. We found several insights from Figure 3.6 and Figure 3.7. The high-quality synthesis methods, such as `gcopula`, `bayes`, `mwem-pgm` and `aim` deliver high AUC scores for California Housing Dataset and around 0.5 for Adult Dataset. Although the average AUC scores of `ctgan` are low, in some cases, e.g.,

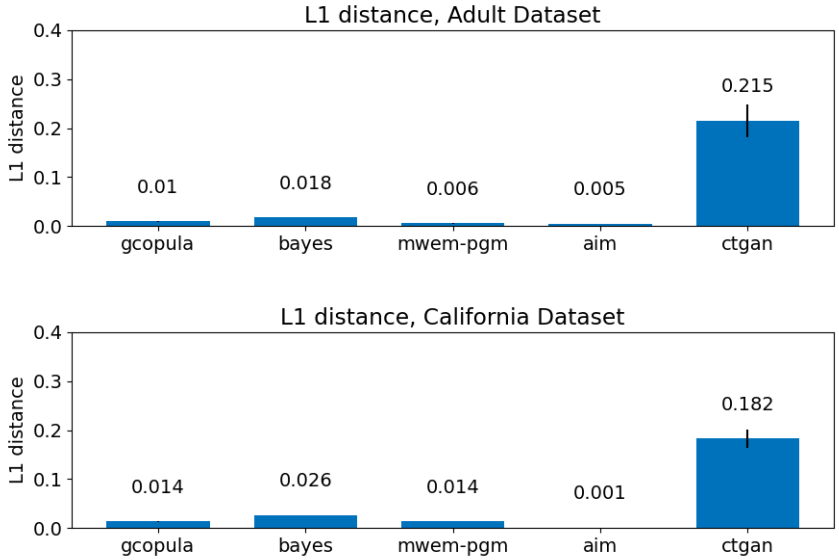

Figure 3.7: The above figures represent the means of the L_1 distance for all attributes, where the measurement is executed for each synthesis algorithm.

[Adult, `ctgan`, mah-max, sample], those are high. In summary, we found cases with high-utility-and-low-AUC-score and also cases with low-utility-and-high-AUC-score. We believe that the above results are evidence for the hypothesis that the results of membership inference attacks are independent of utility. Namely, evaluating membership inference attacks is crucial regardless of utility.

Frequency Distributions

We discuss whether IMs can provide more implications about experimental results than the AUC scores. We plot IMs and draw frequency distributions in Figure 3.8 and Figure 3.9. IMs of $D_{orig} = D_{pos}$ are red, and those of $D_{orig} = D_{neg}$ are blue. These are classified into four kinds of results. Here, each result is denoted by a tuple of [Dataset, Target, Synthesis, Inference] below.

- The case where two distributions are completely indistinguishable from each other: for example, [Adult, `gcopula`, random, RF_dataset] and [California, `mwem-pgm`, random, MLP_dataset]. The AUC score is then 0.5.

- The case where two distributions are somewhat indistinguishable from each other: for example, [Adult, `gcopula`, mah-max, RF_dataset] and [California, `aim`, mah-max, XGB_hist]. The AUC scores are then 0.91 and 0.92.

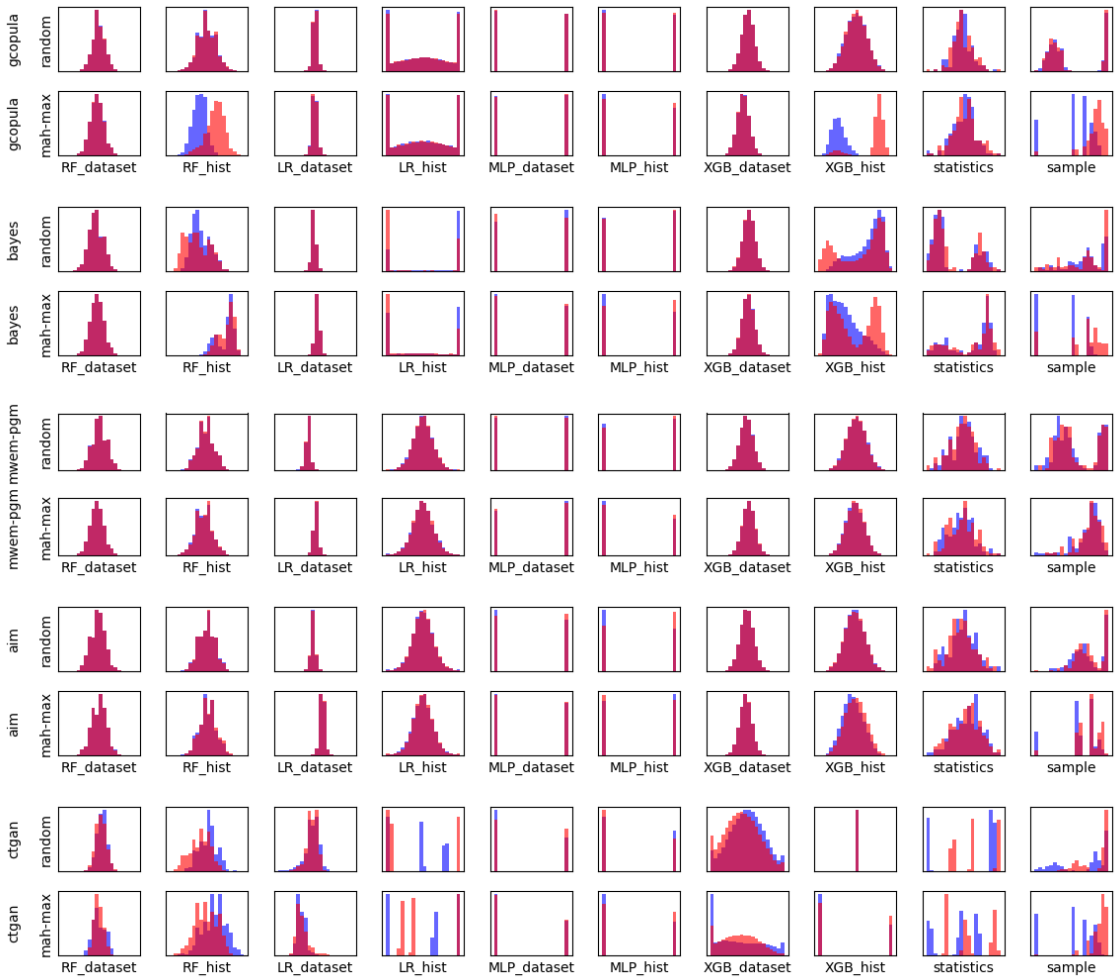

Figure 3.8: Frequency distributions of IM (Adult Dataset). IMs of $D_{orig} = D_{pos}$ are red, and those of $D_{orig} = D_{neg}$ are blue. Horizontal-axis indicates that IM is greater as it goes to the right. The more red is on the right side, the larger the AUC.

- The case where two distributions are completely distinguishable from each other: for example, [California, `bayes`, mah-max, sample] and [California, `mwem-pgm`, mah-max, RF_hist]. Although the AUC scores of both cases are equal to 1, we can find that the distributions of both cases are different.

- The case where distributions are unstable: for example, [Adult, `ctgan`, mah-max, statistics]. We can see that the low quality of synthetic data by `ctgan` causes this result.

The above cases have two implications. The first implication is that the use of IMs provides a chance to visualize the results, such as Figures 3.8 and 3.9. The second

implication is that the visualization of IMs enables us to more deeply understand why membership inference attacks are successful than AUC scores.

3.5 Conclusion

In this chapter, we propose a privacy framework to evaluate resilience against membership inference attacks for synthetic data generation techniques. We introduced Mahalanobis distance to choose a target sample. By way of the experiment, we showed that the performance of membership inference attacks increased when we used the introduced method. We also propose two interpretable inference methods. By way of the experiment, we showed that they are stronger than or equal to the existing black-box inference method. From these results, we can conclude the proposed framework enables us to evaluate the privacy of synthetic data generations more tightly.

As a result, we found that, when using non-DP synthetic data, we must at least remove outliers to ensure privacy. Alternatively, since the theoretical relationship between differential privacy and membership inference attacks is well established [121, 66], we recommend using DP synthetic data to achieve a certain level of privacy protection.

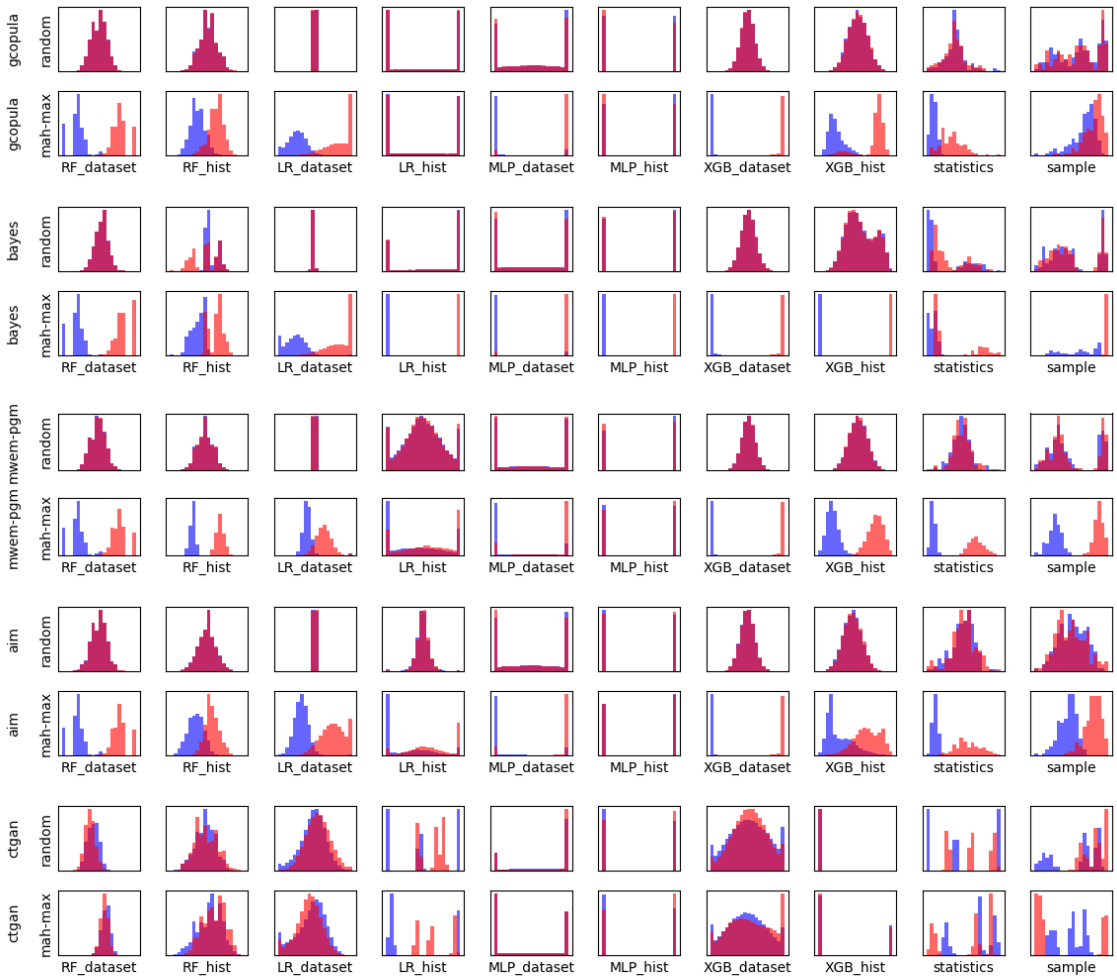

Figure 3.9: Frequency distributions of IM (California Housing Dataset). IMs of $D_{orig} = D_{pos}$ are red, and those of $D_{orig} = D_{neg}$ are blue. Horizontal-axis indicates that IM is greater as it goes to the right. The more red is on the right side, the larger the AUC.

Chapter 4

Utility Evaluation of Synthetic Data Generation with Real Medical Dataset

4.1 Introduction

Given that privacy is of paramount importance and data analysis is highly active in this domain, the this study focuses on the medical and healthcare field. Real-world data collected from healthcare settings has attracted attention for propelling new clinical research due to its non-invasive nature for patients and its potential to constitute big data, thereby reducing bias. Including personal information in the data necessitates a substantial investment of person-hours for ethical review procedures and data protection, thereby impeding the prompt progression of medical research. Anonymization techniques, which reduce the risk of identifying individuals, are crucial in providing data to third parties without patient consent and streamlining the research approval process. Unlike secure computation [22, 105], which facilitates data analysis in encrypted form, these techniques afford analysts the advantages of viewing anonymized data that possess similar properties to the original in a format equivalent to actual data and conducting analyses in an exploratory manner. However, conventional anonymization methods, such as k-anonymity [110], encounter an issue where the quality of the anonymized data significantly diminishes as the data becomes high-dimensional [3].

The technology of synthetic data generation has been recognized for its ability to produce new data while preserving the original statistical properties of high-dimensional data [56, 112, 108, 123, 82, 83, 116]. Specifically, this technology enables

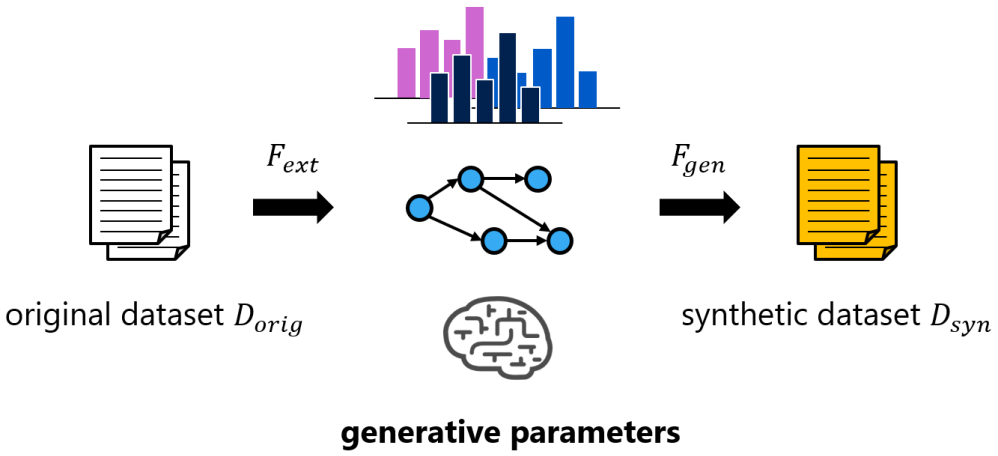

Figure 4.1: Overview of synthetic data generation. The first step is to extract generative parameters $\mathcal{F}_{ext} : \mathcal{D} \to \mathbb{R}^p$. The second step is to generate synthetic data from the extracted generative parameters $\mathcal{F}_{gen} : \mathbb{R}^p \to \mathcal{D}$.

the expedited analysis of synthetic data in a relatively unrestricted environment, potentially abbreviating the approval process. Upon securing useful results, researchers can directly apply them to the original data, deriving final results and potentially mitigating research costs [33]. Nevertheless, to the best of our knowledge, few studies have concurrently deployed various synthetic data generation techniques to authentic medical data [8, 24]. Moreover, few studies have applied various synthetic data generation techniques to real medical data, and insufficient knowledge has been accumulated on the differences among the techniques and the quality of the generated synthetic data.

In the previous chapter, we demonstrated that non-DP synthetic data poses privacy risks for outliers. However, these risks can be theoretically evaluated by applying differential privacy, which provides a formal framework for privacy risk evaluation [121, 66]. In this chapter, we generate synthetic data by using statistics-based, graphical-model-based, and deep-neural-network-based approaches and evaluate the quality of the resultant synthetic data. Due to the heightened significance of privacy-preserving data analysis in healthcare, we focus on medical data in this work. Utilizing the Diagnosis Procedure Combination (DPC) dataset from Ehime University Hospital as the original dataset, we evaluate generated synthetic data from three critical perspectives:

- distribution distances, as a metric for univariable,
- differences in correlation matrices, as a metric for bivariable,
- machine learning model performances, as a metric for multivariable.

Furthermore, we incorporate differential privacy (DP) [31] into each synthetic data generation method, serving as a theoretical privacy framework.

Consequent to the experimental results, we obtained the following conclusions:

- The incorporation of DP enhances privacy protection while concurrently diminishing the quality of synthetic data
- The magnitude of quality degradation is contingent upon the synthesis method employed. Gaussian Copula [75] and AIM [82] sustained comparatively superior quality even after applying DP.

4.2 Related Work

4.2.1 Synthetic Data Generation

Numerous methods have been proposed for generating synthetic data, especially concerning tabular formatted data, while ensuring DP. Synthetic data generation approaches for tabular datasets can be categorized into three types. The first type is founded on basic statistics [75, 6]. The second type leverages graphical models [123, 124, 82, 83]. Tabular formatted data can be regarded as features extracted by humans. Since the graphical models learn relationships among attributes, they produce high-quality synthetic data [112]. The third is the deep-neural-network-based method [116, 37, 128, 19, 73, 72, 76]. In this research, we evaluate one statistics-based method, three graphical-model-based methods, and one deep-neural-network-based method, utilizing a real medical dataset for the assessment.

4.2.2 Synthetic Data Generation for Medical Data

Researchers have directed substantial interest toward using synthetic data generation in the medical field, mainly focusing on image data [49, 111]. In these applications, practitioners employ synthetic data for data augmentation and privacy protection. However, the predominant methods, which are image-specific, present

difficulties when applied to tabular data and do not account for DP. Although Hernandez et al. investigated a tabular healthcare dataset [56], their research concentrates exclusively on deep neural network-based synthetic data generation without considering DP. Our research evaluates several synthetic data generation techniques in conjunction with DP.

4.3 Methodology

Our experiment comprises three components: datasets, synthetic data generation algorithms, and evaluation methods. The experiment aims to evaluate the differences among synthesis algorithms and analyze DP's influence. An overview of the experiment is as follows:

- Apply a synthesis algorithm $\mathcal{F} : \mathcal{D} \to \mathcal{D}$ to the original dataset D_{orig}. The generated synthetic dataset $\mathcal{F}(D_{orig}) = D_{syn}$ is the same size as the original dataset D_{orig}.
- By using an evaluation method $E : \mathcal{D} \times \mathcal{D} \to R$, compare D_{syn} with D_{orig}.

4.3.1 Dataset

This research uses a DPC dataset from Ehime University Hospital. This dataset has been extracted from the data warehouse, which encompasses DPC data from 2010 to 2013, to analyze the impact of 15 attributes on length of hospital stay: gender, type of admission, emergency admission, length of stay, height, weight, smoking, pregnancy, independent eating, independence in activities of daily living, independent mobility, major diagnostic category, surgery, subclassification, and secondary disease. Table 4.1 delineates the information for each category. All categorical data are encoded into one-hot vectors. Records containing missing values were excluded from the dataset, and the number of records became 9,666.

4.3.2 Synthesis Algorithm

In this research, we implement five synthesis algorithms, as listed in Table 4.2. Generally, a synthesis algorithm $\mathcal{F} : \mathcal{D} \to \mathcal{D}$ is decomposed into two steps, as shown in Fig.4.1. The first step is to extract generative parameters $\mathcal{F}_{ext} : \mathcal{D} \to \mathbb{R}^p$. Generative parameters are compressed information needed for the generation, such

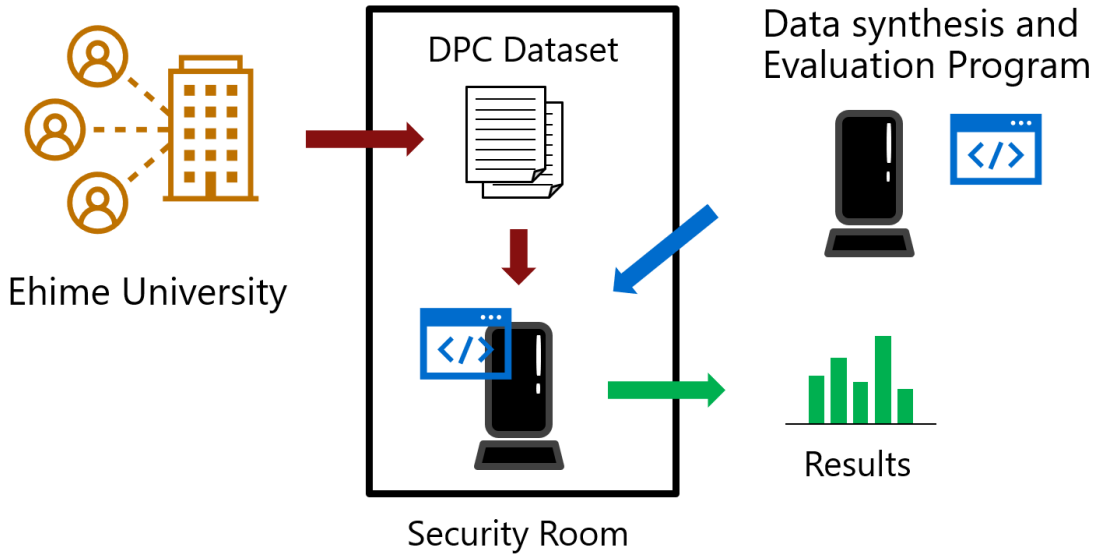

Figure 4.2: Overview of the experiment. In the experiments, we brought the DPC dataset provided by Ehime University, along with the programs for data synthesis and evaluation, into a secure room. Only the results of the experiments were taken out of the secure environment.

as basic statistics or trained machine learning model parameters. The second step is to generate synthetic data from the extracted generative parameters $\mathcal{F}_{gen} : \mathbb{R}^p \to \mathcal{D}$.

Moreover, we use DP, which is known as the gold standard of the privacy protection framework [31, 32] (See Definition 2.3.2). We add intentional noise to the generative parameter $\theta = \mathcal{F}_{ext}(D)$ to satisfy DP. This research investigates the case $\varepsilon = \infty, 8, 4, 2, 1$ and $\delta = 10^{-5}$.

Statistics-based Methods

We evaluate the Gaussian Copula-based synthetic data generation as a statistics-based method [108, 75]. The Gaussian Copula's generative parameters are the original dataset's mean vector μ, the correlation matrix S, and the marginal distribution $H_1, \ldots, H_d$. For the DP version, we use the implementation by Li et al. [75]. We denote this method by `GCopula`.

Table 4.1: Names and types of attributes of DPC dataset. (n) means that the number of the attribute values is n.

	Name	Type
1	Gender	categorical (2)
2	Type of admission	categorical (7)
3	Emergency admission	categorical (2)
4	Length of Stay	numerical
5	Height	numerical
6	Weight	numerical
7	Smoking	categorical (2)
8	Pregnancy	categorical (2)
9	Independent eating	categorical (4)
10	Independence in Activities of Daily Living	categorical (4)
11	Independent Mobility	categorical (5)
12	Major diagnostic category	categorical (18)
13	Surgery	categorical (9)
14	Subclassification	categorical (10)
15	Secondary disease	categorical (3)

Graphical-Model-based Methods

We evaluate PrivBayes [123], MWEM-PGM [83], and AIM [82] as graphical-model-based methods. PrivBayes trains important relations between attributes and expresses the relation as a directed acyclic graph. When generating data, attribute values are sampled in accordance with the graph. AIM and MWEM-PGM are similar methods that learn conditional probability tables to satisfy DP and sample data from them. These methods are denoted by `Bayes`, `MWEM`, and `AIM`.

Deep-Neural-Network-based Methods

We evaluate Conditional Tabular Gan, CTGAN [116], as a deep-neural-network-based method. The differentially private version of CTGAN is implemented by smart-noise[1]. In this method, we train deep neural networks with DP-SGD [1]. This method is denoted by `CTGAN`.

[1] `https://docs.smartnoise.org/synth/index.html`

Table 4.2: Synthesis algorithms in our experiment.

Synthesis algorithm	Description	Generative parameter
Gaussian Copula [75]	`GCopula`	Statistics
PrivBayes [123]	`Bayes`	Directed acyclic graph, conditional probability
MWEM-PGM [83]	`MWEM`	Total joint distribution
AIM [82]	`AIM`	Total joint distribution
CTGAN [116]	`CTGAN`	Model parameter of deep neural network

Algorithm 9 Experiment algorithm

Require: D_{orig}: original dataset, F: synthesis algorithm, E: evaluation function, N_{exp}: number of experiments

Ensure: v_{mean}, v_{std}

```
ans = []
for i = 1, ..., N_exp do
    D_syn ← F(D_orig)
    v ← E(D_orig, D_syn)
    ans.append(v)
end for
v_mean ← mean of ans
v_std ← standard deviation of ans
return v_mean, v_std
```

4.3.3 Evaluation Methods (Quality of Synthetic Data)

In this research, we evaluate the quality of the synthetic dataset D_{syn}, which is the same size as the original dataset D_{orig}, from three perspectives: distribution distances, machine learning model performances, and differences in correlations. Distribution distance is a broad measure, and machine learning model performance is a narrow measure [28, 25]. We also evaluate the absolute difference in correlations to compare relations explicitly. Let $E : \mathcal{D} \times \mathcal{D} \to R$ be an evaluation function.

Evaluation by Distribution Distances

The first evaluation is by statistical distribution distances $E_{dist} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ between D_{orig} and D_{syn}. For each attribute, we evaluate the statistical distance of 1-way marginals. For the statistical distances, we use L1 distance, L2 distance, Hellinger distance, and Wasserstein distance. The definitions are as follows.

Definition 4.3.1 (L_p norm)**.** *For $x, y \in \Delta^d$, the L_p* **norm** *is defined as*

$$||x - y||_p := \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

We use the case when $p = 1$ or $p = 2$.

In a previous work, Hellinger distance was regarded as the best utility metric to rank synthetic data generation algorithms [34].

Definition 4.3.2 (Hellinger distance)**.** *For $x, y \in \Delta^d$, the* **Hellinger distance** *is defined as*

$$\mathrm{Hel}(x, y) := \left(\sum_{i=1}^{d} \sqrt{x_i} - \sqrt{y_i} \right)^2.$$

Definition 4.3.3 (Wasserstein distance)**.** *For $x, y \in \Delta^d$, the* **Wasserstein distance** *or the* **Earth-Mover distance** *is defined as*

$$\mathrm{Was}(x, y) := \inf_{\gamma \sim \Gamma(x,y)} \mathbb{E}_{(a,b)\sim\gamma}[|a - b|],$$

where $\Gamma(x, y)$ is the set of all couplings of x and y. A coupling γ is a joint probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are x and y on the first and second factors, respectively.

Evaluation by the Difference of Machine Learning Model Performances

The second evaluation is the differences in machine learning model performances. Since DPC datasets are often used to predict the length of hospital stays, we train a regression model to predict length of stay (fourth attribute in Table 4.1) with LightGBM, which is a simple but high-performing machine learning model. We compare machine learning models trained by D_{syn} with D_{orig}.

The accuracy of models is evaluated by using the root-mean-square error (RMSE). For a trained model f, the error is defined by

$$RMSE(f, D) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2},$$

where $D = \{(x_i, y_i)\}_{i=1,\ldots,n}$. We evaluate RMSE of a trained model with a synthetic dataset D_{syn}. Thus, the evaluation function $E_{ml} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is defined as $E_{ml}(D_{orig}, D_{syn}) = RMSE(f_{syn}, D_{orig})$, where f_{syn} is a trained model with D_{syn}.

Evaluation by the Difference of Correlation Matrices

The third evaluation is the difference in correlation matrices. The correlation matrix is defined as follows:

Definition 4.3.4 (Correlation matrix)**.** *For data samples $x^1, \ldots, x^m \in \mathbb{R}^d$, set its mean vector as $\mu \in \mathbb{R}^d$. Then, a matrix $R \in \mathbb{R}^{d \times d}$ whose (i, j)-th component is*

$$R_{ij} = \frac{\sum_{k=1}^{d}(x_i^k - \mu_i)(x_j^k - \mu_j)}{\sqrt{\sum_{k=1}^{d}(x_i^k - \mu_i)^2}\sqrt{\sum_{k=1}^{d}(x_j^k - \mu_j)^2}}$$

is called the **correlation matrix**.

We calculate the correlation matrices of D_{orig} and D_{syn}. We evaluate only numerical attributes and compute the absolute error of each component. Thus, the evaluation function $E_{cor} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{n \times n}$ is defined as

$$(E_{cor}(D_{orig}, D_{syn}))_{i,j} = |R_{ij}^{orig} - R_{ij}^{syn}|,$$

where n is the number of numerical attributes.

4.4 Experimental Results

We generated synthetic data five times under the same conditions and calculated the average of the evaluation values. In this section, we report the results.

4.4.1 Distribution Distance Results

Figures 4.3 and 4.4 display the evaluation results by distribution distances, separating the graphs of categorical and numerical attributes due to differing scales. The results of all attributes are shown in Figure 4.5, 4.6, 4.7, and 4.8. Values represent the means of all categorical or numerical attributes, respectively. Notably, the distance is regarded as a loss.

First, the losses for $\varepsilon = \infty$, representing a non-differentially private case, are small. Also, the losses significantly increase as the values of ε decrease, enhancing the robustness of the protection by DP.

`CTGAN` and differentially private `Bayes` exhibit more substantial losses when synthesizing algorithms are compared, while `GCopula`, `MWEM`, and `AIM` demonstrate lesser losses.

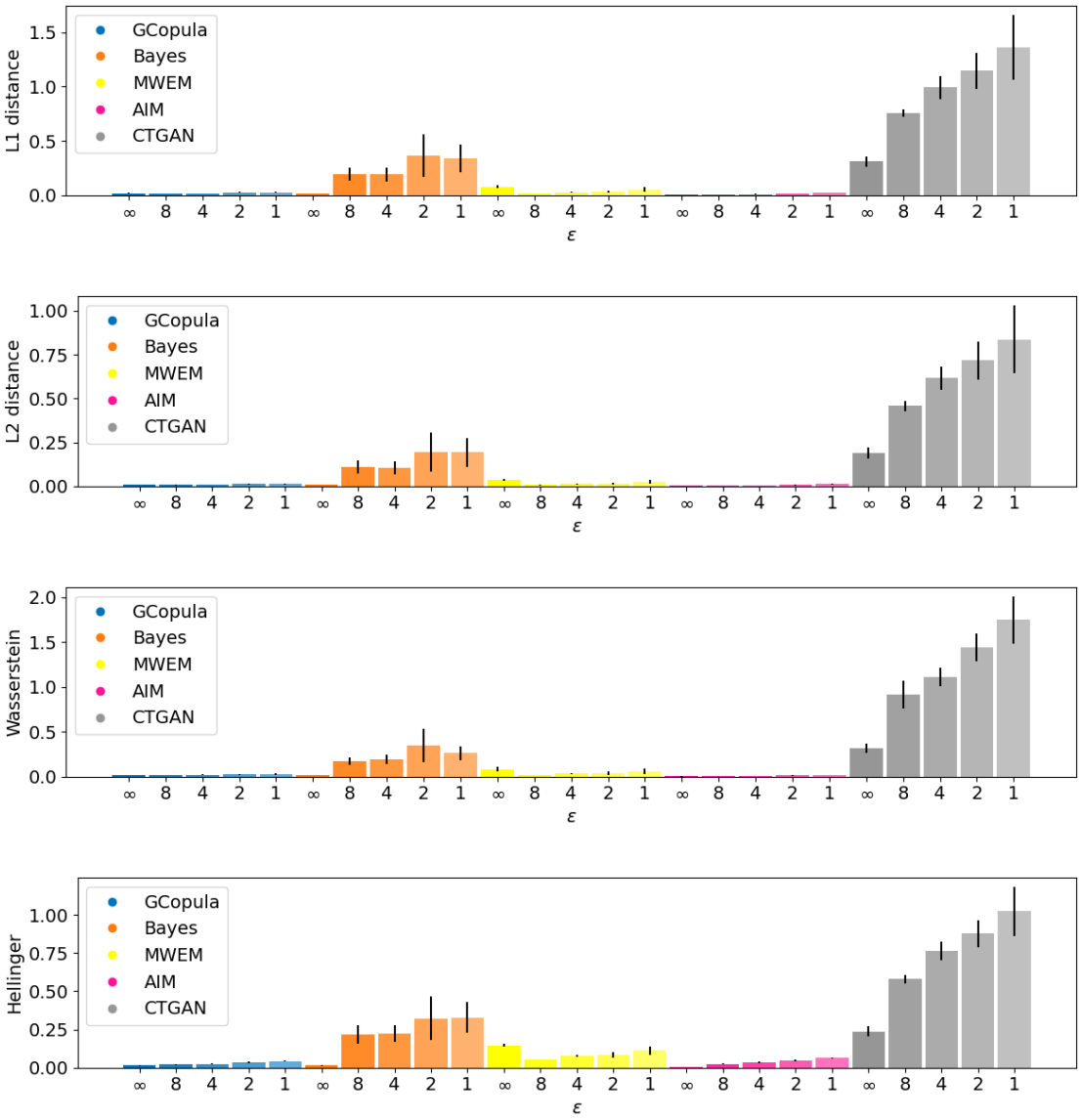

Figure 4.3: Result of categorical attributes distance. L1 distance, L2 distance, Hellinger distance, and Wasserstein distance from the top.

4.4.2 Machine Learning Model Performance Results

Fig. 4.9 illustrates the results of machine learning model performances, with the red line expressing RMSE for the original dataset. Non-differentially private results for each synthesis algorithm ($\varepsilon = \infty$) align closely with the original. The quality of the synthetic data discernibly declines as ε increases. Specifically, the results from differentially private `Bayes` and `CTGAN` are inferior, while those of `GCopula` and `AIM`

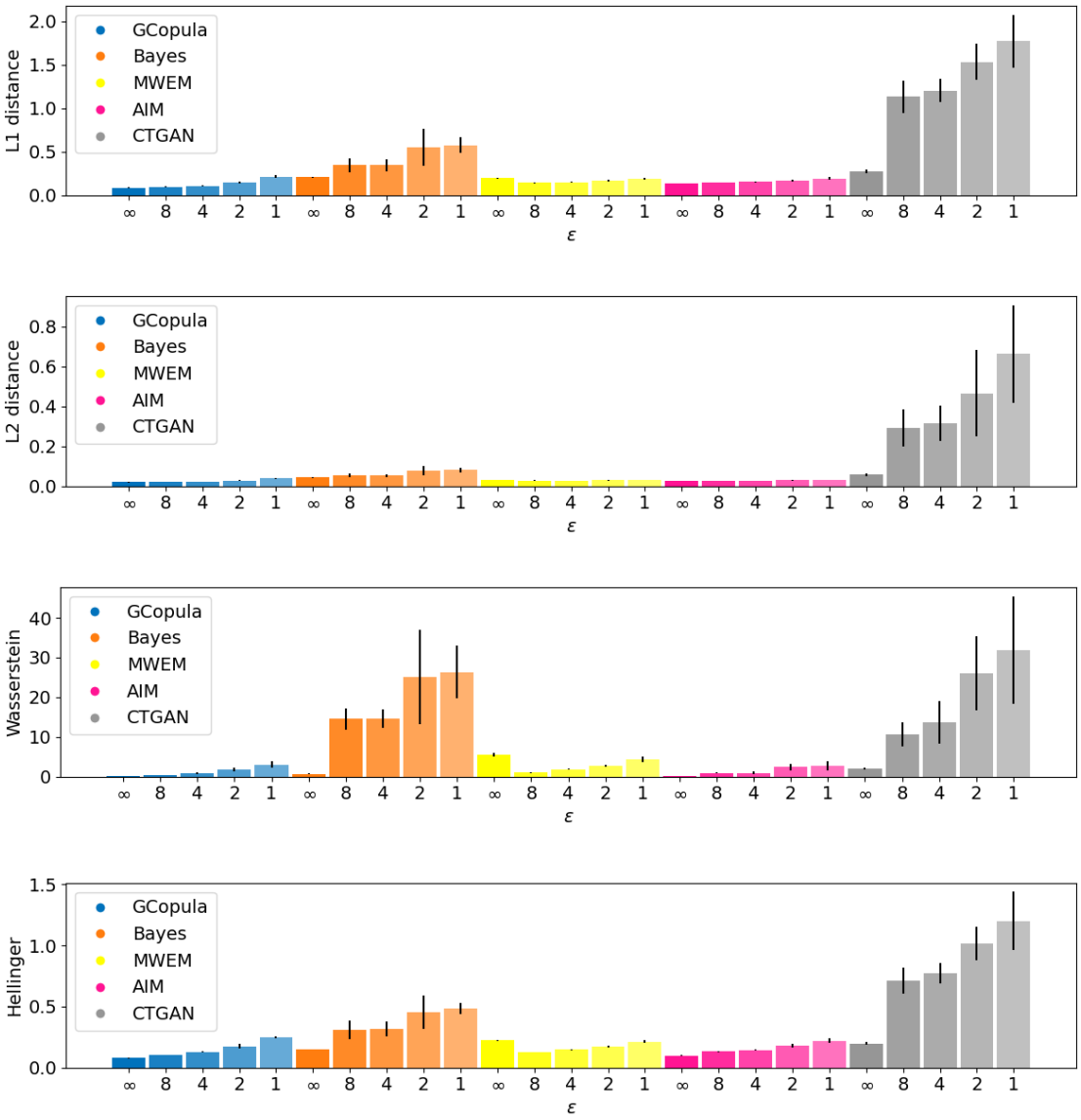

Figure 4.4: Result of numerical attributes distance. L1 distance, L2 distance, Hellinger distance, and Wasserstein distance from the top.

remain proximate to the original results, even when differentially private.

The results of distribution distances for each attribute are shown in Fig. 4.5, 4.6, 4.7 and 4.8.

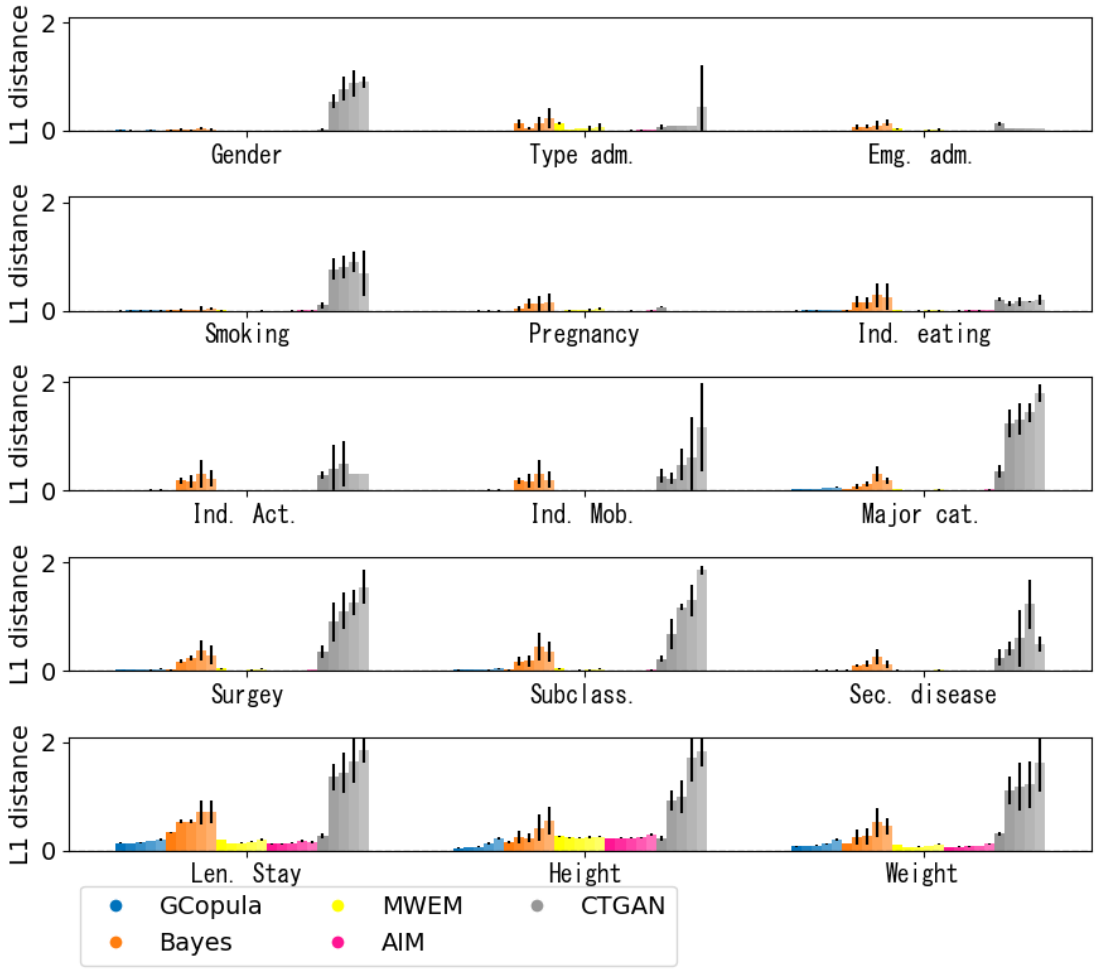

Figure 4.5: The values of L1 distance of each attribute.

4.4.3 Difference in Correlations Results

Fig. 4.10 presents the results in cases where $\varepsilon = \infty$, the absolute losses of `GCopula` and `Bayes` are small. Additionally, losses become more significant as ε increases, resulting in differentially private `CTGAN` being the worst.

4.5 Discussion

4.5.1 Quality of Synthetic Data

The three evaluation methods reveal that the losses associated with non-differentially private synthesis remain sufficiently small, while DP diminishes the quality of syn-

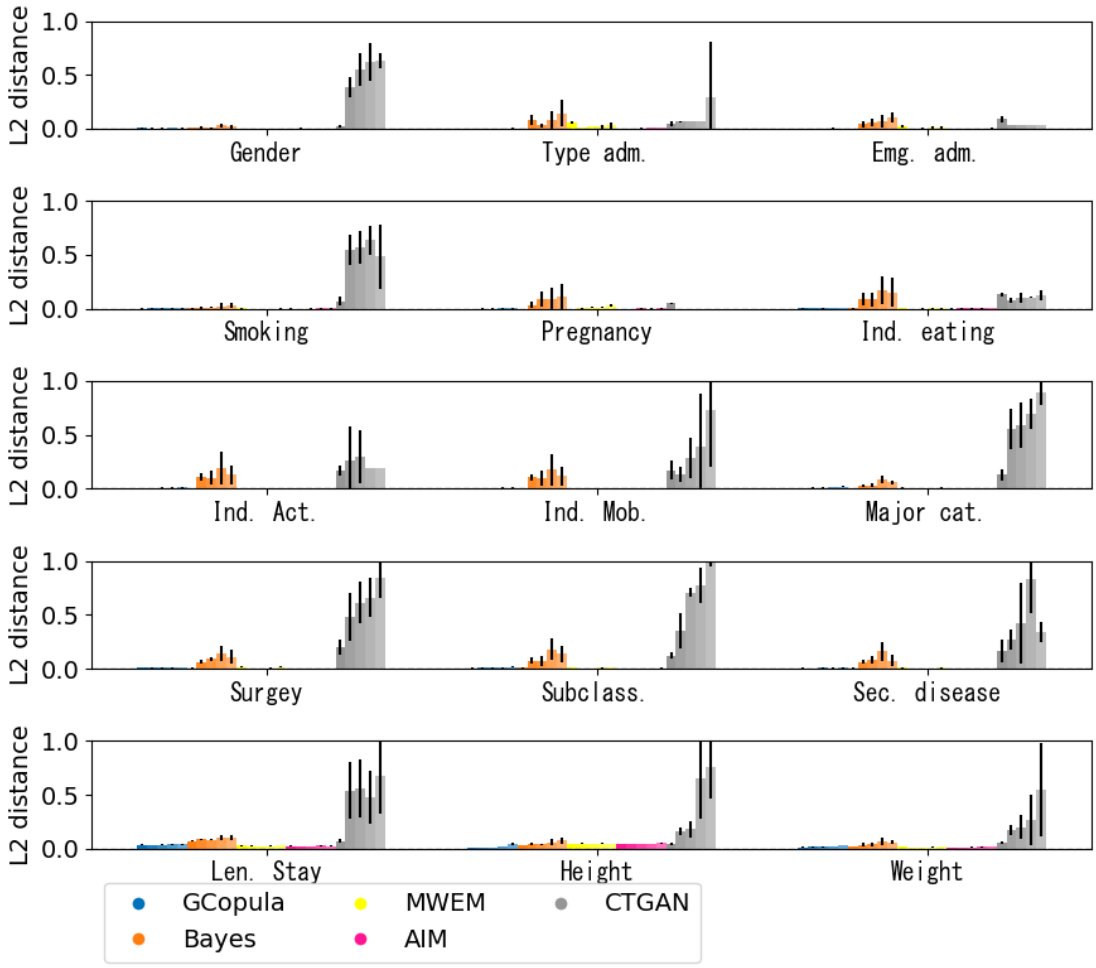

Figure 4.6: The values of L2 distance of each attribute.

thetic data. In differentially private cases, the magnitude of the losses varies among synthesis methods. This indicates the potential for enhancing the quality of synthetic data by strategically devising DP. Notably, the recently proposed `AIM` achieves noteworthy experimental results consistently. `AIM` manifests negligible deterioration in the quality of the synthetic data when implementing DP.

4.5.2 Evaluation Methods

This study employs L1 distance, L2 distance, Hellinger distance, and Wasserstein distance as evaluative metrics, which are widely utilized in studies measuring the quality of synthetic data and prove highly useful when assessing the "relative" qual-

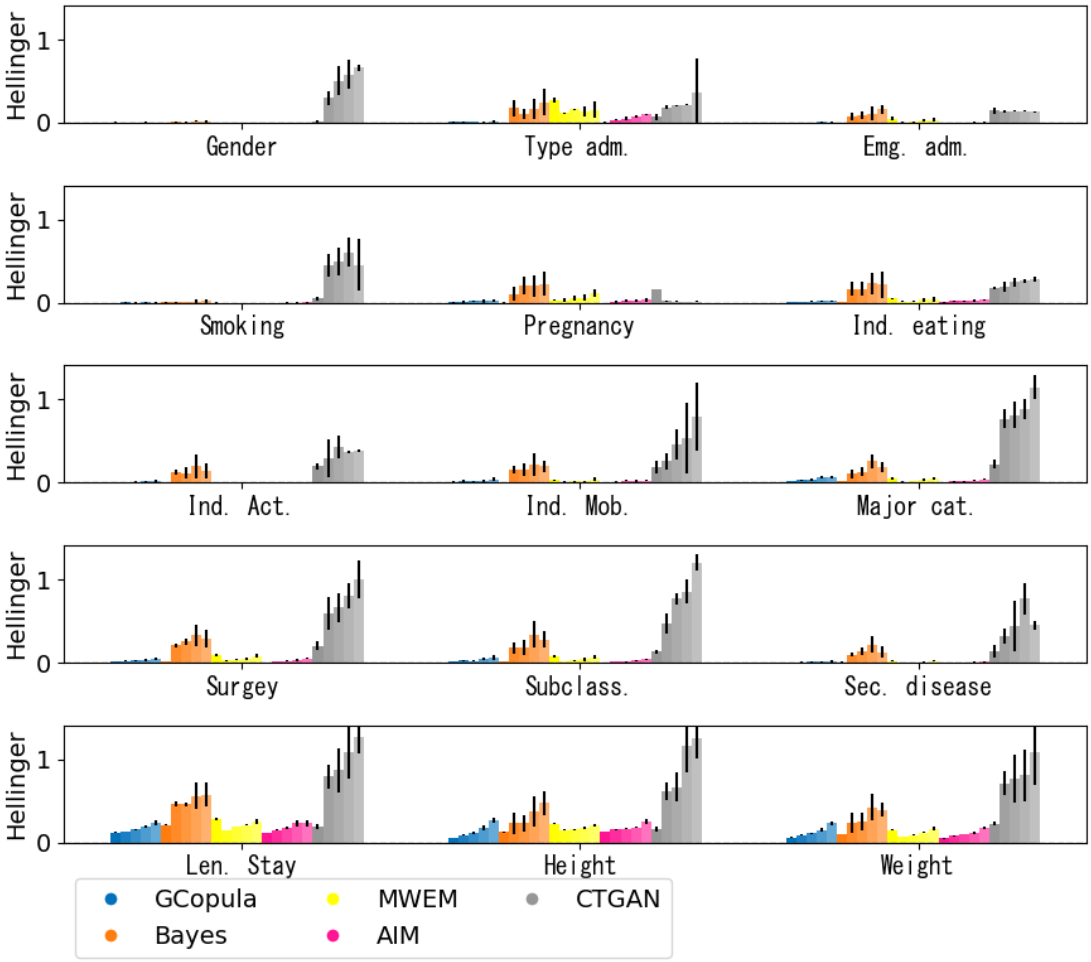

Figure 4.7: The values of Hellinger distance of each attribute.

ity thereof. These metrics indicate that `AIM` exhibits notably superior results to other methods.

Conversely, to facilitate absolute evaluations with qualitative significance, it is necessary to assume realistic use cases for evaluations by machine learning performance and ascribe meaning to the magnitude of errors.

4.5.3 Towards Practical Use

Discussion has yet to emerge regarding whether using synthetic data for personal data is subject to the agenda of Ethics Review Committees. Conversely, Guo et al. have reported that they did not require an ethical review because the synthetic

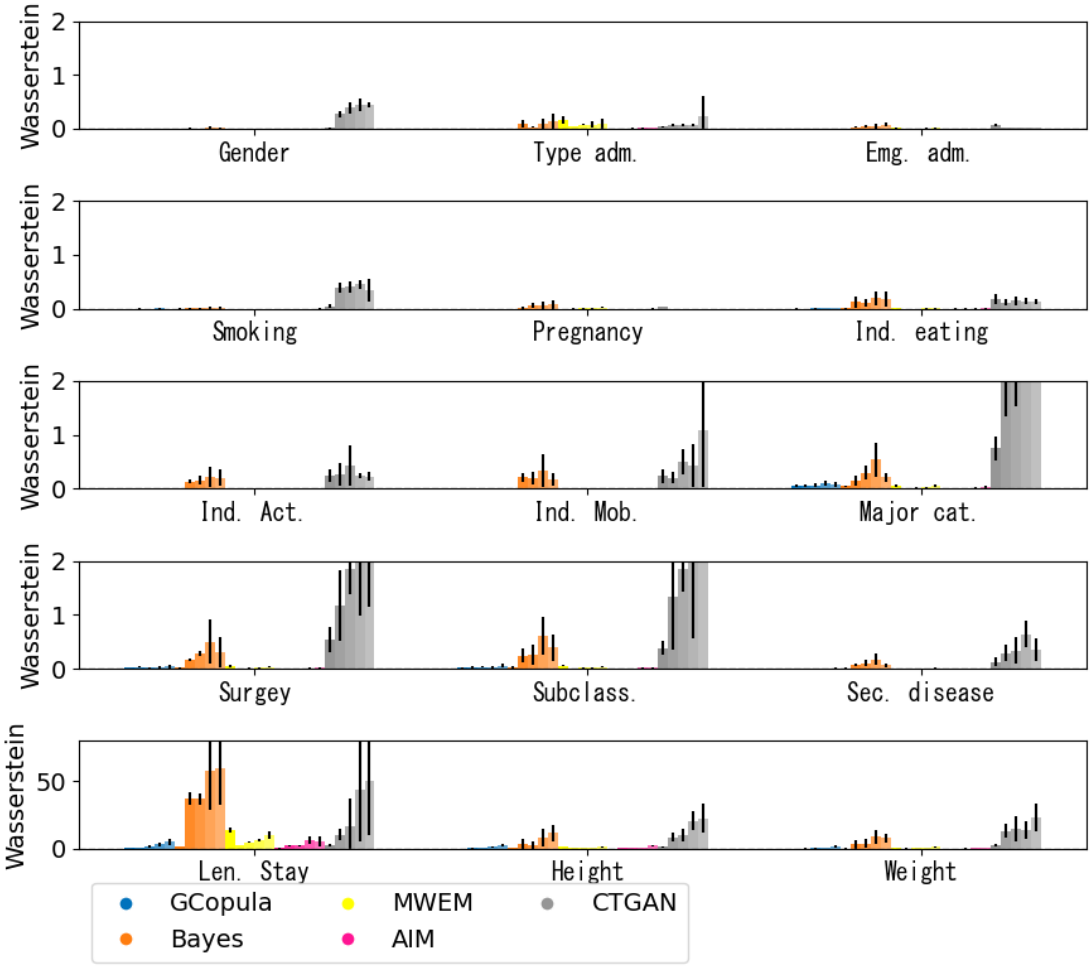

Figure 4.8: The values of Wasserstein distance of each attribute.

data contained no information that could lead to the identification of individual patients [50]. It has been posited that, should synthetic data gain recognition as a viable option for privacy considerations, obtaining approval from ethics committees may become unnecessary [7]. In a case wherein an organization inadvertently disclosed the personal information of numerous individuals online while testing a cloud solution, the Norwegian Data Protection Authority (Datatilsynet) highlighted that testing could have been conducted by processing synthetic data or using less personal data [2]. This ruling also implies that synthetic data may be recognized as having the potential to exclude information that leads to personal identification.

Furthermore, DP can potentially enhance the security of such synthetic data.

[2] https://www.dataguidance.com/news/norway-datatilsynet-fines-nif-nok-12m-disclosing

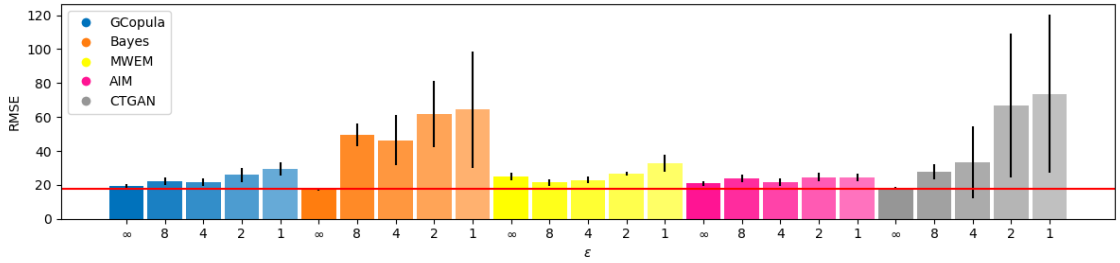

Figure 4.9: Results of machine learning model performances: RMSEs of a trained LightGBM regression model. The red line expresses RMSE for the original dataset

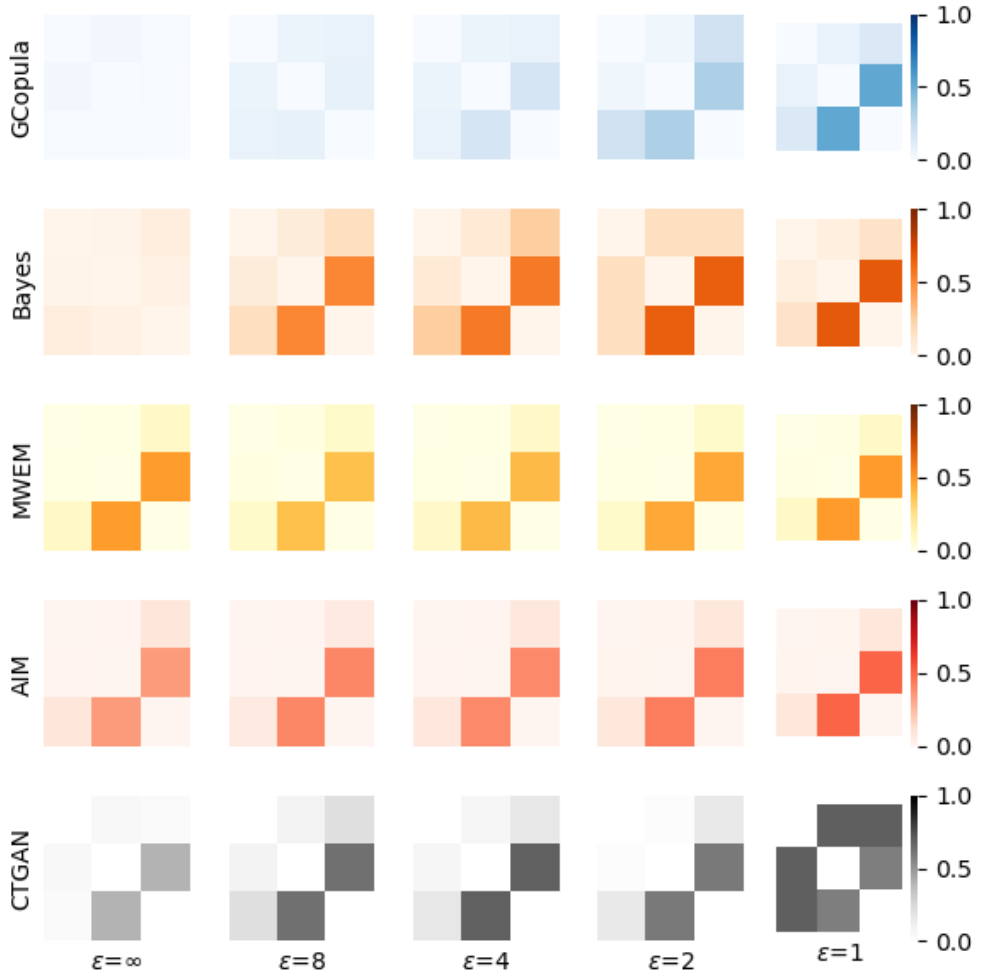

Figure 4.10: Results of differences in correlations.

Therefore, DP is anticipated to minimize discussions concerning anonymous processing and expedite the progression of research. Nonetheless, studies have examined attacks that deduce the original data from synthetic data [109], necessitating further research to ensure its security.

4.6 Conclusion

In this research, employing the a Diagnosis Procedure Combination (DPC) dataset, we experimentally evaluated synthetic data generation techniques' effectiveness using statistic-based, machine-learning model-based, and deep neural network-based methods. The investigation clarified the differences in performance among the methods, attributing them to variations in the amount of source data and the degree of accuracy degradation when implementing differential privacy. In particular, we found that the methods using Gaussian copula and AIM produced high-quality results even under differential privacy settings. However, we also observed a slight degradation in data quality due to the application of differential privacy. Further, we discussed issues that must be addressed to apply synthetic data generation techniques more effectively.

Ethical Considerations

The Ethics Review Committee of Ehime University Hospital approved this study ("Quality evaluation of synthetic data generation methods preserving statistical characteristics," Permission number 2012001), and we conducted it in accordance with the committee's guidelines.

Chapter 5

Evaluating Differential Privacy of Synthetic Data Generation without Adding Intentional Noise

5.1 Introduction

Personal data is expected to be utilized in various fields such as finance, healthcare, and medicine, but sharing personal data collected by one organization with another organization requires attention to individual privacy. Traditional anonymization techniques such as k-anonymization [110] and randomized response [115] have struggled to find a good trade-off between utility and privacy for high-dimensional data [3]. In contrast, a synthetic data generation technique has emerged as a privacy protection method that preserves data utility even for high-dimensional data such as images and tabular data with multi-attributes [11]. In synthetic data generation, generative parameters are extracted from the original raw dataset, and then synthetic data are generated randomly as shown in Figure 5.1(a). The synthetic data are in the same format as the original data and should be statistically similar to them. Typical generative parameters are statistics of original data or trained parameters of deep neural networks [108, 75, 6, 40, 123, 124, 82, 45, 116, 71, 100]. After synthetic data are generated, they are shared with other organizations, but the generative parameters are typically discarded without being disclosed.

To guarantee privacy protection theoretically, differential privacy [31] is used as a standard framework. By adding randomness in generative parameter calculation, the generative parameters become differentially private [1, 81, 123]. The post-processing property of differential privacy guarantees that synthetic data generated with dif-

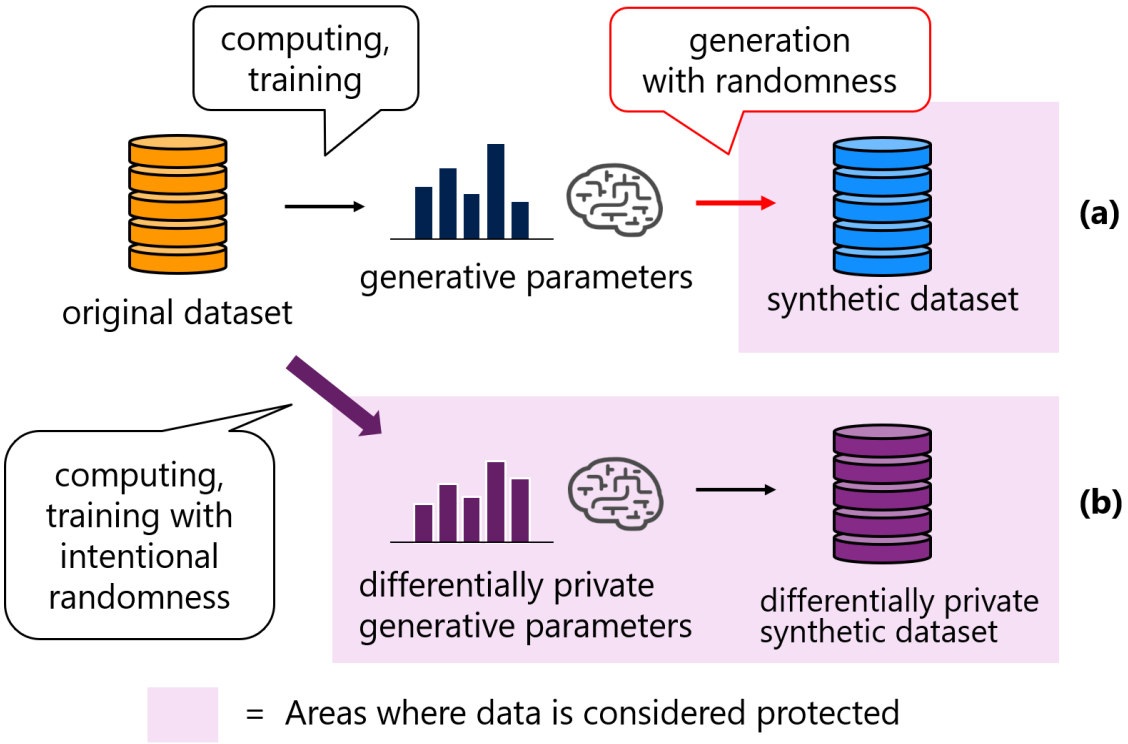

Figure 5.1: (a) Output = Only synthetic data: The generative parameters are discarded after data are generated. We evaluate privacy protection by the randomness in generation.
(b) Output = Generative parameters: By computing or training generative parameters with intentional randomness, we obtain differentially private generative parameters that also generate differentially private synthetic data.

ferentially private generative parameters also satisfy differential privacy as shown in Figure 5.1(b). Although the synthetic data generated with non-differentially private generative parameters have high utility, those with differentially private parameters are known to have lower utility [112].

We address this problem by evaluating differential privacy of randomness in data generation when using non-differentially private generative parameters. As mentioned above, in the context of anonymization, the generative parameters are discarded without disclosing them to the public. When the output does not include generative parameters but only consists of synthetic data, it can be regarded that the synthetic data has already been protected due to the inherent randomness, even if the generative parameters are not protected with differential privacy, as shown in Figure 5.1(a). By evaluating privacy protection in data generation quantitatively,

theoretically guaranteed synthetic data can be obtained without degrading the utility. Moreover, by incorporating this evaluation into traditional methods, we can decrease the amount of random noise while ensuring the same level of privacy.

In this chapter, we regard a record as a d-dimensional vector and focus on a synthetic data generation mechanism with the mean vector and the covariance matrix of the original dataset shown in Figure 5.2. We theoretically evaluate Rényi differential privacy [87], which is a relaxed concept of differential privacy, by randomness in generation for the method. We explicitly derive the condition of ε such that the synthetic data generation mechanism satisfies (α, ε)-Rényi differential privacy for a fixed $\alpha > 1$ under the unbounded neighboring condition (Theorem 5.3.1) and the bounded neighboring condition (Corollary 5.3.2). Furthermore, we conduct a numerical evaluation with reference to the Adult dataset [29] and compute ε concretely. We demonstrate that when the size of the original dataset is 10 million and the mechanism outputs data the same size as the input dataset, it satisfies $(4, 0.576)$-Rényi differential privacy under the unbounded condition and $(4, 2.307)$-Rényi differential privacy under the bounded condition (Table 5.1). If they are translated into the traditional (ε, δ)-differential privacy, the mechanism satisfies $(4.46, 10^{-14})$ and $(9.21, 10^{-14})$ differential privacy under the unbounded and bounded condition, respectively (Figure **??**). These values are mostly similar to ones used by US Census [113].

5.2 Preliminaries

In this section, I introduce basic notations and concepts for later discussion.

5.2.1 Notations

In this chapter, we denote the determinant of a square matrix $A \in \mathbb{R}^{d \times d}$ by $|A| := \det A$. The transposes of a vector $x \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ are denoted by ${}^t x \in \mathbb{R}^{1 \times d}$ and ${}^t A \in \mathbb{R}^{d_2 \times d_1}$. We assume that datasets are tabular but all discussions can be applied to other datasets such as images since we consider records as vectors. In a tabular dataset, a record is expressed as a combination of several attribution values. Each attribution value is a numerical value and normalized into a range $[-1, 1]$. Thus, a record is regarded as a vector $x \in [-1, 1]^d$, and a dataset with n records is regarded as $D = \{x_i\}_{i=1,\dots,n} \in [-1, 1]^{d \times n} =: \mathcal{D}$.

5.2.2 Synthetic Data Generation with Mean Vector and Covariance Matrix

In this chapter, we focus on a simple synthetic data generation with the mean vector and the covariance matrix of the original dataset $\mathcal{M}_G : \mathcal{D} \to [-1,1]^d$ as shown in Figure 5.2. This method is identical to the Gaussian copula [108] with the assumption that the marginal distributions are all normal distributions.

The mechanism $\mathcal{M}_G$ generates synthetic data as follows. First, for dataset $D = \{x_i\}_{i=1,\dots,n} \in \mathcal{D}$, the mean vector $\mu \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ are computed:

$$\mu := \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\Sigma := \frac{1}{n}\sum_{i=1}^{n} x^t x - \mu^t \mu.$$

Next, a sample is drawn from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, and its values are cut into the range $[-1,1]^d$.

We denote by $\mathcal{M}_G^n : \mathcal{D} \to [-1,1]^{d\times n}$ the mechanism that simultaneously outputs n records by $\mathcal{M}_G$. By Proposition 2.3.18, we see that if $\mathcal{M}_G$ satisfies (α, ε)-RDP, then $\mathcal{M}_G^n$ also satisfies $(\alpha, n\varepsilon)$-RDP.

5.2.3 Properties of Symmetric Matrices

We explain the properties of symmetric matrices for the proof of the main theorem.

Definition 5.2.1 (symmetric matrix)**.** *A square matrix A is called* **symmetric** *if $A = {}^tA$ holds.*

Definition 5.2.2 (eigenvalue, eigenvector)**.** *Let $A \in \mathbb{R}^{d\times d}$ be a matrix. A complex number $\lambda \in \mathbb{C}$ is an eigenvalue of A, if there exists a non-zero vector $x \in \mathbb{C}^d$ such that $Ax = \lambda x$ holds. The vector x is also called an eigenvector of A.*

Definition 5.2.3 (positive-definite, semi-positive definite)**.** *For a d-dimensional symmetric matrix A, the following two conditions are equivalent:*
(1) *For all $x \in \mathbb{R}^d \backslash \{0\}$, it holds ${}^txAx > 0$ (≥ 0);*
(2) *All eigenvalues of A are positive real numbers (non-negative).*
If A satisfies these conditions, then A is called **positive-definite (positive semi-definite)**.

The following two lemmas are well-known facts [53].

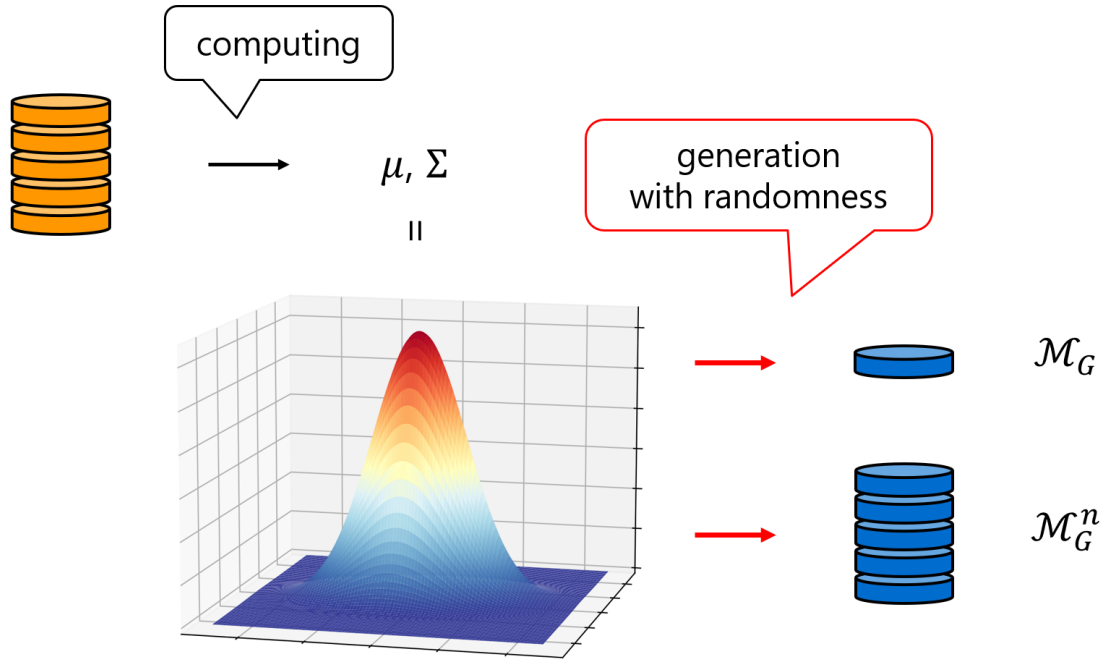

Figure 5.2: Synthetic data generation algorithms $\mathcal{M}_G$ and $\mathcal{M}_G^n$

Lemma 5.2.4. *Let A, B be positive-definite symmetric matrices. If AB is symmetric, then AB is also positive-definite.*

Proof. Since AB is symmetric, we have

$$AB = {}^t(AB) = {}^tB{}^tA = BA.$$

Thus, there exists an orthogonal matrix $P \in \mathbb{R}^{d\times d}$ such that ${}^tPAP = D_A$ and ${}^tPBP = D_B$, where D_A and D_B are diagonal and positive-definite. Let $x \in \mathbb{R}^d\backslash\{0\}$. Then we have

$${}^txABx = {}^t({}^tPx){}^tPA{}^tPPBP({}^tPx) = {}^t({}^tPx)D_AD_B({}^tPx) > 0.$$

Thus, we see that AB is also positive-definite. □

Lemma 5.2.5. *Let A be a positive-definite symmetric matrix. For an invertible matrix S that is the same size as A, ${}^tSAS$ is also positive-definite.*

Proof. Let $x \in \mathbb{R}^d\backslash\{0\}$. Since S is invertible, we see that $Sx \neq 0$. Since A is positive-definite, we have

$${}^tx({}^tSAS)x = {}^t(Sx)A(Sx) > 0.$$

Thus, ${}^tSAS$ is also positive-definite. □

Proposition 5.2.6. *Let A, B, C be positive-definite symmetric real matrices. If ABC is symmetric, then ABC is also positive-definite.*

Proof. Set $D := ABC = CBA$. Since C is positive-definite, we can obtain the spectral decomposition

$$C := \sum_{i=1}^{d} \lambda_i {\theta_i}^t \theta_i,$$

where $\lambda_i > 0$ for all $i = 1, \ldots, d$. Then we set $S := \sum_{i=1}^{d} \sqrt{\lambda_i} {\theta_i}^t \theta_i$. We see that S is symmetric and $C = S^2$ holds. We have

$$S^{-1} D S^{-1} = S^{-1} A S^{-1} S B S = S B S S^{-1} A S^{-1}.$$

By applying $S^{-1}AS^{-1}$ and SBS to Lemma 5.2.4 and Lemma 5.2.5, we see that $S^{-1}DS^{-1}$ is positive-definite. Thus, D is also positive-definite. □

5.3 Main Theorem

In this chapter, we prove the upper bound of ε such that the mechanism $\mathcal{M}_G$ satisfies (α, ε)-Rényi differential privacy for a fixed α. We assume that all datasets have a limitation for the minimum eigenvalue of their covariance matrices. Specifically, for a fixed $\sigma > 0$, we define the set of datasets as

$$\mathcal{D}_\sigma := \{D \in [-1, 1]^{n \times d} \mid z \in S^{d-1}, {}^t z \Sigma_D z \geq \sigma\}.$$

We also set $\tau := \frac{4d}{\sigma}$.

First, the result under the unbounded condition is the following theorem. We assume that the number of records in an original dataset is n and that in its neighboring dataset is $n + 1$.

Theorem 5.3.1. *Under the unbounded condition, let $\alpha > 1$. We assume that*

$$\begin{aligned} &\frac{n}{n+1} < \tau, \\ &\alpha < \min\{n+1, \frac{n^2}{\tau(n+1) - n}\}. \end{aligned} \tag{5.1}$$

Then, the synthetic data generation mechanism $\mathcal{M}_G$ satisfies $(\alpha, \varepsilon_{UB}(\alpha, n))$-RDP

for $\varepsilon_{UB}(\alpha, n) := \max\{\varepsilon_{\alpha 1}, \varepsilon_{\alpha 2}\}$. *Here,*

$$\begin{aligned}\varepsilon_{\alpha 1} &= \frac{\alpha}{2}\cdot\frac{\tau}{(n+1)(n+1-\alpha)} \\ &\quad + \frac{\alpha d}{2(\alpha-1)}\log\frac{n}{n+1} - \frac{d}{2(\alpha-1)}\log(1-\frac{\alpha}{n+1}) \\ &\quad - \frac{1}{2(\alpha-1)}\log\min\{1, \frac{1+\alpha\frac{n\tau}{(n+1)(n+1-\alpha)}}{(1+\frac{\tau}{n+1})^{\alpha}}\}\end{aligned}$$

and

$$\begin{aligned}\varepsilon_{\alpha 2} &= \frac{\alpha}{2}\cdot\frac{\tau}{n(n+\alpha)-\alpha(n+1)\tau} \\ &\quad + \frac{\alpha d}{2(\alpha-1)}\log\frac{n+1}{n} - \frac{d}{2(\alpha-1)}\log(1+\frac{\alpha}{n}) \\ &\quad - \frac{1}{2(\alpha-1)}\log\min\{1, \frac{1-\frac{\alpha(n+1)\tau}{(n+\alpha)n}}{(1-\frac{\tau}{n})^{\alpha}}\}.\end{aligned}$$

Next, under the bounded condition, we obtain the following statement as a corollary of Theorem 5.3.1.

Corollary 5.3.2. *Under the bounded condition, let* $\alpha > 1$. *We set*

$$c := \min\{n+1, \frac{n^2}{\tau(n+1)-n}\}$$

and assume that

$$\alpha < \frac{c^2}{2c-1}. \tag{5.2}$$

Then, the synthetic data generation mechanism $\mathcal{M}_G$ *satisfies* $(\alpha, \varepsilon_B(\alpha, n))$*-RDP for the following*

$$\varepsilon_B(\alpha, n) = \inf_{\frac{c-1}{c-\alpha}<p<\frac{c}{\alpha}} \frac{\alpha-\frac{1}{p}}{\alpha-1}\varepsilon_{UB}(p\alpha, n) + \varepsilon_{UB}(\frac{p\alpha-1}{p-1}, n+1). \tag{5.3}$$

Proof. For any neighboring datasets D_1, D_2 under the bounded condition, there exists a dataset D_3 such that D_1 and D_3 are neighboring and D_2 and D_3 are neighboring under the unbounded condition. Then, to obtain Equation (5.3), we use the following Lemma 5.3.3. Here, the weak triangle inequality holds for all $p > 1$, and the following condition is necessary:

$$\max\{p\alpha, \frac{p\alpha-1}{p-1}\} < c.$$

This is equivalent to

$$\frac{c-1}{c-\alpha} < p < \frac{c}{\alpha}.$$

The existence of p is equivalent to Equation (5.2). □

By the following lemma, the result with the unbounded condition can be reduced to the bounded condition.

Lemma 5.3.3 (Weak triangle inequality [87]). *Let P, Q, R be probability distributions on $\mathbb{R}^d$. Let $\alpha > 1$. If it holds*

$$\frac{1}{p} + \frac{1}{q} = 1,$$

then we have

$$D_\alpha(P||Q) \leq \frac{\alpha - \frac{1}{p}}{\alpha - 1} D_{p\alpha}(P||R) + D_{q(\alpha - \frac{1}{p})}(R||Q).$$

5.4 Proof of Theorem 5.3.1

In this section, we prove Theorem 5.3.1. The following proposition is essential.

Proposition 5.4.1 (Gil et al. [43]). *Let $\alpha > 1$ and $\mathcal{N}(\mu_1, \Sigma_1)$, $\mathcal{N}(\mu_2, \Sigma_2)$ be multivariate normal distributions. If a matrix*

$$T_\alpha := \alpha\Sigma_1^{-1} + (1 - \alpha)\Sigma_2^{-1}$$

is positive-definite, then it holds

$$\begin{aligned} &D_\alpha(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \\ &= \frac{\alpha}{2} {}^t(\mu_1 - \mu_2)\Sigma_\alpha^{-1}(\mu_1 - \mu_2) - \frac{1}{2(\alpha - 1)} \log \frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}, \end{aligned}$$

where $\Sigma_\alpha := (1 - \alpha)\Sigma_1 + \alpha\Sigma_2$.

For neighboring datasets $D_1, D_2 \in \mathcal{D}_\sigma$, we set the mean vectors as μ_1, μ_2 and the covariance matrices as Σ_1, Σ_2. If $D_\alpha(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) \leq \varepsilon$, the mechanism $\mathcal{M}_G$ satisfies (α, ε)-RDP. Here we set

$$L_1 := {}^t(\mu_1 - \mu_2)\Sigma_\alpha^{-1}(\mu_1 - \mu_2),$$

$$L_2 := \frac{|\Sigma_\alpha|}{|\Sigma_1|^{1-\alpha}|\Sigma_2|^\alpha}.$$

Then we see

$$D_\alpha(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) = \frac{\alpha}{2} L_1 - \frac{1}{2(\alpha - 1)} \log L_2. \tag{5.4}$$

Thus, an upper bound ε is described by the maximum of L_1 and the minimum of L_2. The outline of proof is as follows. First, by using the different record, we represent

the difference between mean vectors and the difference between covariance matrices (Lemma 5.4.2). Next, we determine the positive-definiteness of T_α (Lemma 5.4.3). Finally, we compute the upper bound of L_1 (Lemma 5.4.4) and the lower bound of L_2 (Lemma 5.4.5).

Set $\#D_1 = n$ and $\#D_2 = n+s$, where $s = 1$ when we "add" a record and $s = -1$ when we "remove" a record. The common records are denoted by $x_1, \ldots, x_n \in [-1, 1]^d$ and the different record by $x \in [-1, 1]^d$. We set each mean vector as μ_1, μ_2 and covariance matrix as Σ_1, Σ_2. We also denote by σ_{min} the minimum eigenvalue of Σ_1. Note that $\sigma_{min} \geq \sigma$ by the assumption.

Lemma 5.4.2 (Representations of difference)**.** *The following equations hold:*

$$\mu_d := \mu_2 - \mu_1 = \frac{s}{n+s}x - \frac{s}{n(n+s)}\sum_{i=1}^{n} x_i,$$

$$X := \Sigma_2 - \frac{n}{n+s}\Sigma_1 = \frac{ns}{(n+s)^2}(x-\mu_1)^t(x-\mu_1).$$

Proof. By calculation, we see that

$$\begin{aligned}
\mu_d &:= \mu_2 - \mu_1 \\
&= \frac{1}{n+s}\sum_{i=1}^{n+s} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n+s}sx + \frac{1}{n+s}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{s}{n+s}x - \frac{s}{n(n+s)}\sum_{i=1}^{n} x_i.
\end{aligned}$$

□

The rank of X is one. X is semi-positive definite when $s = 1$ and semi-negative definite when $s = -1$.

By the following lemma, we can confirm the positive-definiteness of T_α.

Lemma 5.4.3 (Positive-definiteness of T_α)**.** *If the following two inequalities hold, T_α is positive-definite:*

$$\begin{aligned}
&\frac{n-1}{n} < \tau, \\
&\alpha < \min\{n+1,\ \frac{(n-1)^2}{\tau n - (n-1)}\}.
\end{aligned} \tag{5.5}$$

Proof. Since $T_\alpha = \Sigma_1^{-1}\Sigma_\alpha\Sigma_2^{-1} = \Sigma_2^{-1}\Sigma_\alpha\Sigma_1^{-1}$, by Lemma 5.2.6, the positive-definiteness of T_α is reduced to the positive-definiteness of Σ_α. By Lemma 5.4.2, we have

$$\Sigma_\alpha = (1-\alpha)\Sigma_1 + \alpha\left(\frac{n}{n+s}\Sigma_1 + X\right) = \left(1 - \frac{s\alpha}{n+s}\right)\Sigma_1 + \alpha X.$$

When $s = 1$, since Σ_1 is positive-definite and X is semi-positive definite, it is enough to be $\alpha < n+1$. We consider the case when $s = -1$. For an arbitrary vector $z \in \mathbb{R}^d$ whose norm is one, we seek a condition where the minimum of ${}^t z\Sigma_\alpha z$ is positive. Here we can consider that the vector $x - \mu_1$ is contained in a ball with a radius $2\sqrt{d}$. Thus, we obtain the minimum when the following two conditions hold:

- z is parallel to the eigenvector of the minimum eigenvalue σ_{min} of Σ_1;
- $x - \mu_1$ is parallel to z.

Hence we see that Σ_α is positive-definite if

$$\begin{aligned}
{}^t z\Sigma_\alpha z &= \left(1 + \frac{\alpha}{n-1}\right)\sigma_{min} - \alpha\frac{n}{(n-1)^2}4d \\
&= \sigma_{min} - \alpha\cdot\frac{4dn - (n-1)\sigma_{min}}{(n-1)^2} \\
&\geq \sigma - \alpha\cdot\frac{4dn - (n-1)\sigma}{(n-1)^2} \\
&> 0.
\end{aligned}$$

When the inequalities in Equation (5.5) hold, this inequality also holds. □

To compute the upper bound of Equation (5.4), we prove Lemma 5.4.4 and Lemma 5.4.5.

Lemma 5.4.4 (Upper bound of L_1)**.** *If $s = 1$, then we have*

$$L_1 \leq \frac{\tau}{(n+1)(n+1-\alpha)},$$

and if $s = -1$, then we have

$$L_1 \leq \frac{\tau}{(n-1)(n-1+\alpha) - \alpha n\tau}.$$

Proof. Now μ_d is contained in a ball with a radius $\frac{2\sqrt{d}}{n+s}$ by Lemma 5.4.2 and Σ_α is positive-definite by Lemma 5.4.3. By multiplying the reciprocal of the minimum of

${}^t z\Sigma_\alpha z$ for a unit vector $z \in \mathbb{R}^d$ by $\frac{4d}{(n+s)^2}$, we can obtain the maximum of ${}^t\mu_d\Sigma_\alpha^{-1}\mu_d$. Here, we see

$$ {}^t z\Sigma_\alpha z = {}^t z(1 - \frac{s\alpha}{n+s})\Sigma_1 z + \frac{s\alpha n}{(n+s)^2}({}^t z(x-\mu_1))^2.$$

Hence when $s = 1$, the minimum is

$$\left(1 - \frac{\alpha}{n+1}\right)\sigma_{min}.$$

When $s = -1$, since $x - \mu_1$ is contained in a ball with a radius $2\sqrt{d}$, the minimum is

$$(1 + \frac{\alpha}{n-1})\sigma_{min} - \frac{\alpha n}{(n-1)^2} \cdot 4d.$$

Thus, we obtain the inequality. □

Lemma 5.4.5 (Lower bound of L_2)**.** *It holds*

$$L_2 \geq \frac{(1 - \frac{s\alpha}{n+s})^d}{(\frac{n}{n+s})^{\alpha d}} \cdot \min\{1, \frac{1 + \frac{\alpha n s\tau}{(n+s-s\alpha)(n+s)}}{(1 + \frac{s\tau}{n+s})^\alpha}\}.$$

Proof. We see that

$$\begin{aligned} L_2 &:= \frac{|(1 - \frac{s\alpha}{n+s})\Sigma_1 + \alpha X|}{|\Sigma_1|^{1-\alpha}|\frac{n}{n+s}\Sigma_1 + X|^\alpha} \\ &= \frac{(1 - \frac{s\alpha}{n+s})^d |I + \frac{n+s}{n+s-s\alpha}\alpha\Sigma_1^{-1}X|}{(\frac{n}{n+s})^{\alpha d}|I + \frac{n+s}{n}\Sigma_1^{-1}X|^\alpha}. \end{aligned}$$

Since the rank of X is one and Σ_1^{-1} is invertible, the rank of $\Sigma_1^{-1}X$ is also one. Thus, there is only one non-zero eigenvalue, and it is set as λ. We also set

$$A := (1 - \frac{s\alpha}{n+s})^d / (\frac{n}{n+s})^{\alpha d}.$$

Since the other eigenvalues are all zero, we see

$$L_2 = \frac{1 + \frac{n+s}{n+s-s\alpha}\alpha\lambda}{(1 + \frac{n+s}{n}\lambda)^\alpha} \cdot A.$$

By differentiating this equation with respect to λ, we obtain

$$\frac{\partial L_2}{\partial \lambda} = \alpha(\alpha - 1) \cdot \frac{n+s}{n(n+s-s\alpha)} \cdot \frac{s - (n+s)\lambda}{(1 + \frac{n+s}{n}\lambda)^{\alpha+1}} \cdot A.$$

We see that $\frac{\partial L_2}{\partial \lambda} > 0$ when $\frac{s}{n+s} < \lambda$ and $\frac{\partial L_2}{\partial \lambda} < 0$ when $\frac{s}{n+s} > \lambda$. Hence the minimum of L_2 is obtained at the edges of the range of λ.

Next, we will find the range of λ, which is the only one non-zero eigenvalue of $\Sigma_1^{-1}X$. Since Σ_1 is positive-definite, we can obtain the spectral decomposition of Σ_1:

$$\Sigma_1 = \sum_{i=1}^{d} \sigma_i {p_i}^t p_i,$$

where $\sigma_1, \ldots, \sigma_d$ are the eigenvalues of Σ_1 and $p_1, \ldots, p_d$ are their eigenvectors whose norms are one. Since $p_1, \ldots, p_d$ is a basis of $\mathbb{R}^d$, there exist $r_1, \ldots, r_d \in \mathbb{R}$ such that

$$x - \mu_1 = \sum_{i=1}^{d} r_i p_i.$$

Squaring both sides, we obtain a condition

$$4d \geq \sum_{i=1}^{d} r_i^2 > 0.$$

Set

$$e_1 := \sum_{i=1}^{d} \frac{r_i}{\sigma_i} p_i.$$

Then we have

$$\begin{aligned} \Sigma_1^{-1} X e_1 &= \Sigma_1^{-1} \frac{ns}{(n+s)^2} \sum_{i=1}^{d} r_i p_i ((x-\mu_1)\cdot e_1) \\ &= \frac{ns}{(n+s)^2} ((x-\mu_1)\cdot e_1) e_1 \\ &= \frac{ns}{(n+s)^2} \left(\sum_{i=1}^{d} \frac{r_i^2}{\sigma_i} \right) e_1. \end{aligned}$$

Thus, we have

$$\lambda = \frac{ns}{(n+s)^2} \sum_{i=1}^{d} \frac{r_i^2}{\sigma_i}.$$

Therefore, we have

$$0 < \lambda \leq \frac{4dn}{(n+1)^2 \sigma_{min}} \leq \frac{4dn}{(n+1)^2 \sigma}$$

when $s = 1$, and

$$-\frac{4dn}{(n-1)^2 \sigma} \leq -\frac{4dn}{(n-1)^2 \sigma_{min}} \leq \lambda < 0$$

when $s = -1$. □

5.5 Numerical Evaluations

In Theorem 5.3.1 and Corollary 5.3.2, we obtain the concrete upper bounds. Thus, in this section, we compute the value ε concretely and observe the results. For the sake of clarity, the number of records in the original dataset is n_{in} (i.e., n in Section 5.3–5.4) and the number of records in the output dataset is n_{out}. Moreover, when $n_{in} = n_{out}$ holds, we denote them by $n_{in/out}$. Note that we used 1 for n_{out} in Section 5.3–5.4. In this section, based on the composition theorem, we compute ε by multiplying ones in Theorem 3.1 and Corollary 3.2 by n_{out}.

5.5.1 Setting of Numerical Parameters

We set $d = 6$, $\sigma = 0.01$ since the number of numerical attributions in Adult Dataset [29] is six and the minimum eigenvalue for the data normalized into $[-1, 1]$ is $\sigma_{min} = 0.01$. We also consider the case of $d = 9$ with reference to California Housing dataset [95]. Note that we compute ε without creating a concrete dataset and use only the number of records in dataset n_{in}, the number of attributions d and the minimal eigenvalue σ_{min}.

5.5.2 Relation between α and ε

The relations between α and ε are shown in Figure 5.3 (α-ε curves). For all curves, ε is monotonically increasing with respect to α. We also see that as n_{in} increases exponentially, ε becomes smaller at equal intervals on a logarithmic scale. In particular, if $n_{in} = 10^4$ and $d = 6$, the condition in Equation (5.1) is

$$\alpha < c := \min\{n_{in} + 1, \frac{n_{in}^2}{\tau(n_{in} + 1) - n_{in}}\} \approx 4.1679$$

and the condition in Equation (5.2) is

$$\alpha < \frac{c^2}{2c - 1} \approx 2.3680.$$

Thus, the curves stop at these values.

5.5.3 The Impact of n_{out} and d

In this subsection, we evaluate the impact of the number of outputs n_{out} and the number of attribution d. Throughout this subsection, we fix $\alpha = 4$.

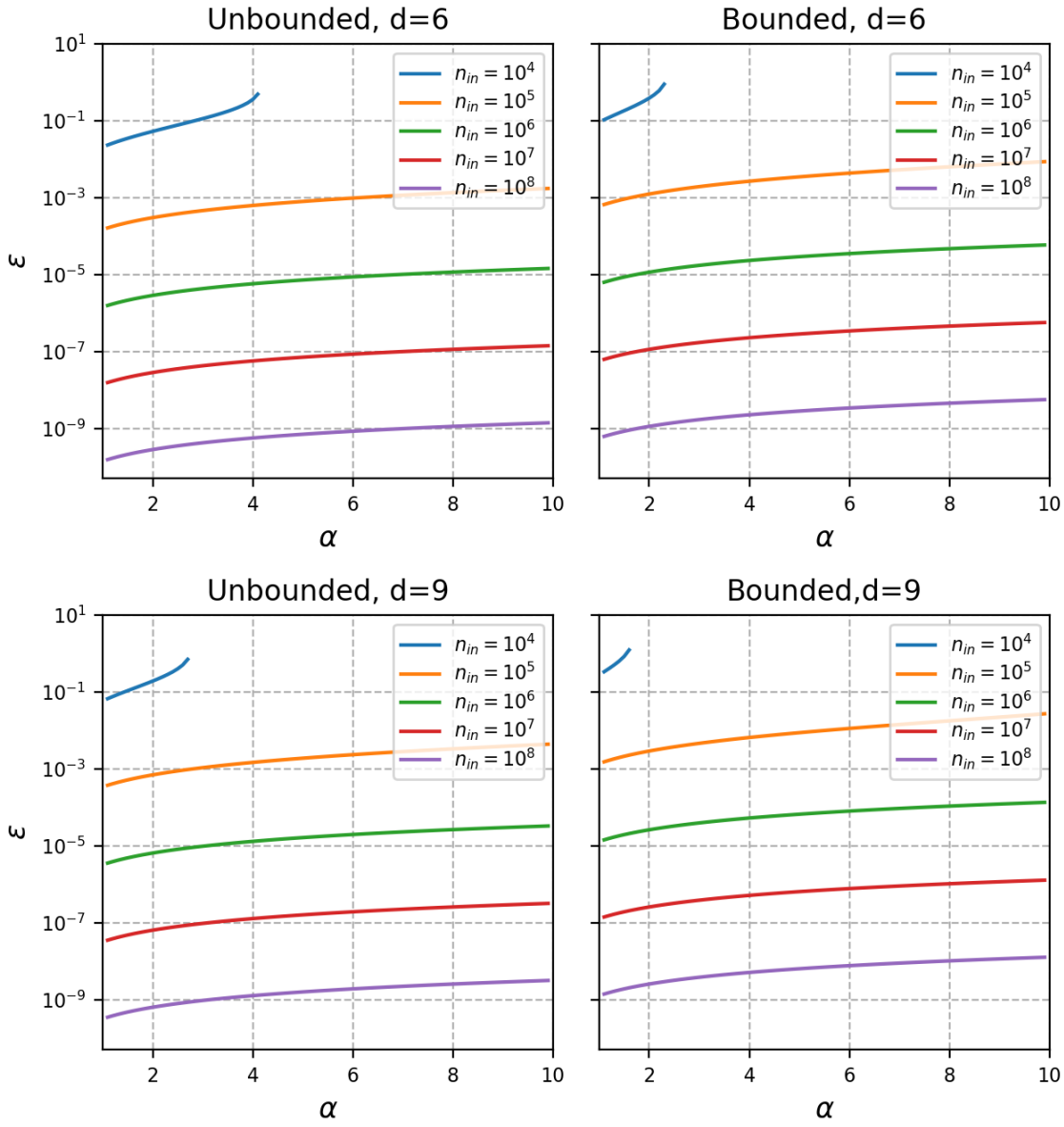

Figure 5.3: α-ε curve ($d = 6, 9$, $\sigma = 0.01$) : Vertical axis is logarithmic scale. The curves are drawn for each of the four sample sizes n_{in}

.

As the most basic case, we consider the case in which $n_{in} = n_{out} =: n_{in/out}$. The values of ε for which the mechanism $\mathcal{M}_G^{n_{in/out}}$ satisfies (α, ε)-RDP are shown in Table 5.1. By the composition theorem in Proposition 2.3.18, the values of ε are ones in Theorem 5.3.1 and Corollary 5.3.2 multiplied by $n_{in/out}$. We can show that values of ε are within a practical range when $n_{in/out} \geq 10^6$ under both conditions.

Table 5.1: Values of ε in the case that input and output are the same size $n_{in/out}$. $(\alpha = 4, d = 6, \sigma = 0.01)$

$n_{in/out}$	10^4	10^5	10^6	10^7	10^8
UB ε	3535.17	62.5859	**5.8064**	**0.5764**	**0.058**
B ε	-	266.7349	23.3577	**2.3071**	**0.23**

Table 5.2: The number of outputs n_{out} in the case that $\varepsilon = 1$. $(\alpha = 4, d = 6, \sigma = 0.01)$

n_{in}	10^4	10^5	10^6	10^7	10^8
UB n_{out}	3	1598	1.72×10^5	1.73×10^7	1.74×10^9
B n_{out}	0	375	4.28×10^4	4.33×10^6	4.34×10^8

In particular, under the unbounded condition, $\varepsilon = 0.5764$ when $n_{in/out} = 10^7$, which is very small. We also see that ε's under the unbounded condition are four times larger than those under the bounded condition.

Next, we compute the number of outputs n_{out} where $\mathcal{M}_G^{n_{out}}$ satisfies $(\alpha = 4, \varepsilon = 1)$-RDP. The result is shown in Table 5.2. When $n_{in} = 10^4$, we can output only three records under the unbounded condition. For $n_{in} \geq 10^7$, we can output records more than input records.

Lastly, we show the relation between ε and d in Figure 5.4. We can confirm that the values of ε increase as d increases. The values of ε under the bounded condition are about four times as large as those under the unbounded condition.

5.5.4 Translation into (ε, δ)-DP

By Proposition 2.3.9, we see that (α, ε)-RDP can be translated into (ε, δ)-DP. For a fixed δ, we seek α which gives the minimum of ε and plot them in Figures 5.5, 5.6, and 5.7.

According to [32], the value of δ should be less than $\frac{1}{n_{in}}$. The values translated into (ε, δ)-DP under the unbounded condition are shown in Figures 5.5, 5.6, and 5.7 as blue bars. When $\delta = 10^{-10}$, we see that $\varepsilon = 13.03$ for $n_{in/out} = 10^6$, $\varepsilon = 3.79$ for $n_{in/out} = 10^7$ and $\varepsilon = 1.23$ for $n_{in/out} = 10^8$. When $\delta = \frac{1}{n_{in}^2}$, we also see that $\varepsilon = 14.14$ for $n_{in/out} = 10^6$, $\varepsilon = 4.46$ for $n_{in/out} = 10^7$ and $\varepsilon = 1.71$ for $n_{in/out} = 10^8$. These values are reasonable [113].

The results under the bounded condition are shown in Figures 5.5, 5.6, and 5.7 as orange bars. When $\delta = 10^{-10}$, we see that $\varepsilon = 29.03$ for $n_{in/out} = 10^6$, $\varepsilon = 7.87$

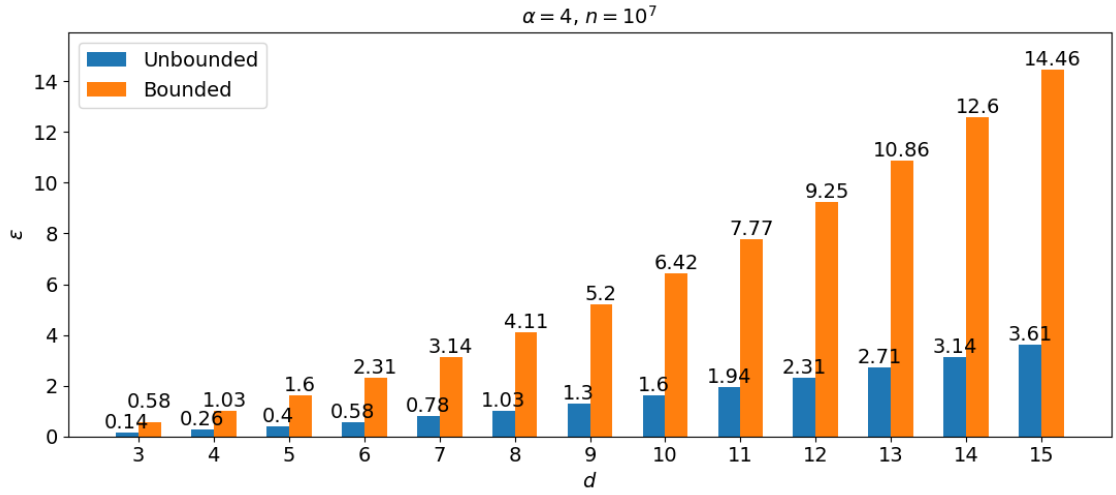

Figure 5.4: The values of ε for $(\alpha = 4, \varepsilon)$-RDP for each $3 \leq d \leq 15$.

for $n_{in/out} = 10^7$ and $\varepsilon = 2.36$ for $n_{in/out} = 10^8$. When $\delta = \frac{1}{n_{in}^2}$, we also see that $\varepsilon = 31.2$ for $n_{in/out} = 10^6$, $\varepsilon = 9.21$ for $n_{in/out} = 10^7$ and $\varepsilon = 2.97$ for $n_{in/out} = 10^8$.

The values of ε under the bounded condition are about twice as large as those under the unbounded condition.

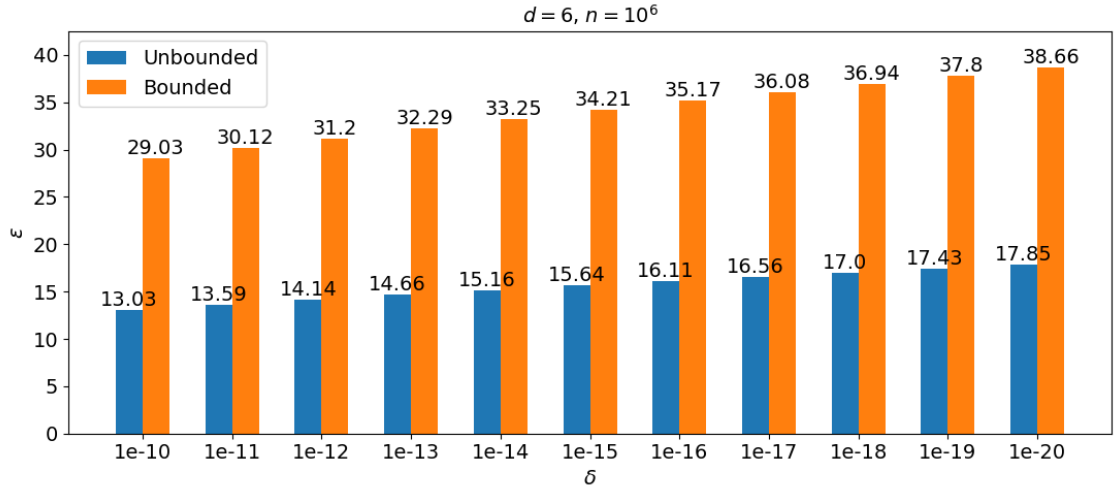

Figure 5.5: Values of ε in (ε, δ)-DP. $d = 6, \sigma = 0.01, n_{in/out} = 10^6$. Blue bars are unbounded cases and orange bars are bounded cases.

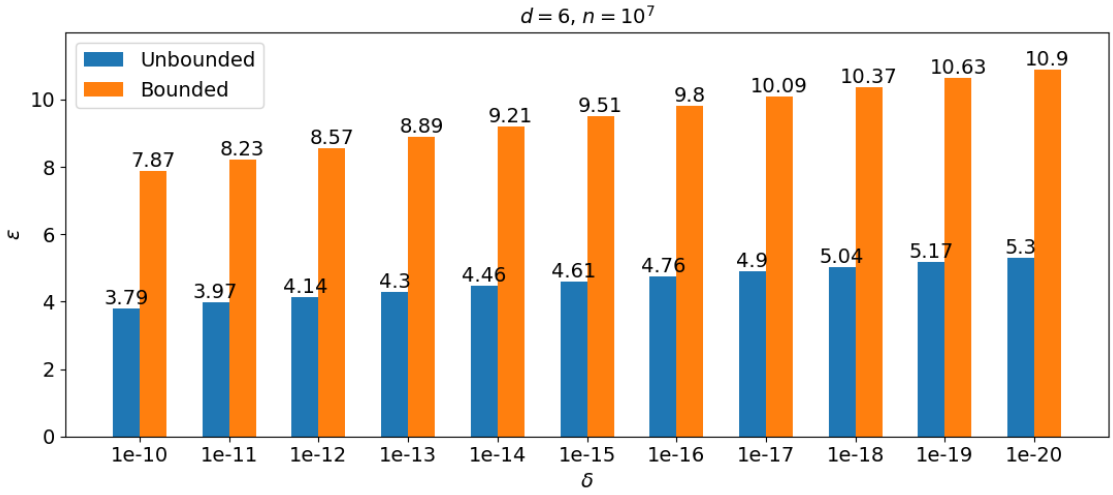

Figure 5.6: Values of ε in (ε, δ)-DP. $d = 6, \sigma = 0.01, n_{in/out} = 10^7$. Blue bars are unbounded cases and orange bars are bounded cases.

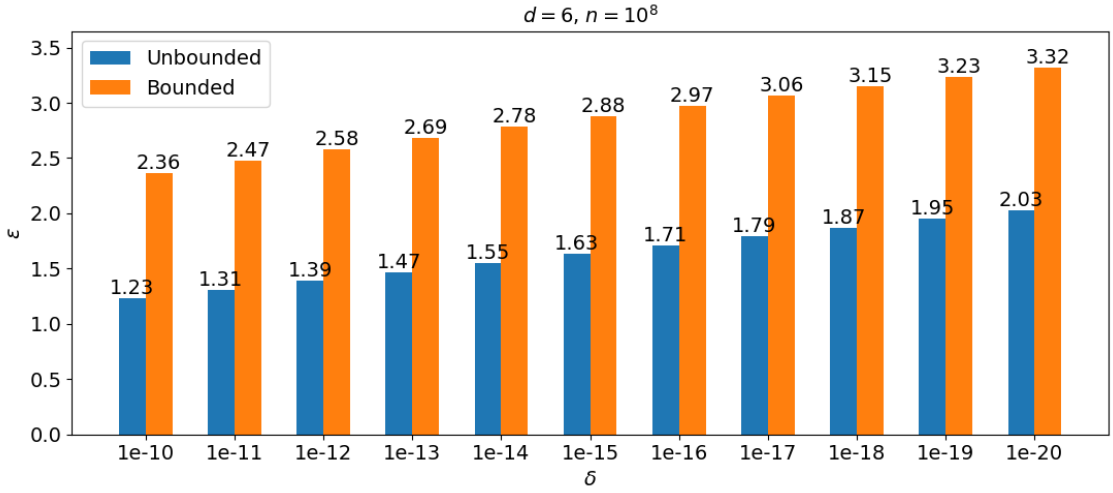

Figure 5.7: Values of ε in (ε, δ)-DP. $d = 6, \sigma = 0.01, n_{in/out} = 10^8$. Blue bars are unbounded cases and orange bars are bounded cases.

5.5.5 Impact of σ

Figure 5.8 shows the value of ε for each value of σ, which determines the minimum eigenvalue of the data range. As a default setting, σ was set to 0.01 with reference to the numerical attributes of the Adult Dataset. It was found that σ values down to approximately 0.005 yield practically acceptable ε values; however, when σ is

reduced to 0.001, ε becomes significantly larger.

This result is closely related to the observation in Chapter 3 that instances with large Mahalanobis distances tend to pose higher risks. When the minimum eigenvalue decreases, the risk associated with data distributed along the corresponding eigenvector direction increases.

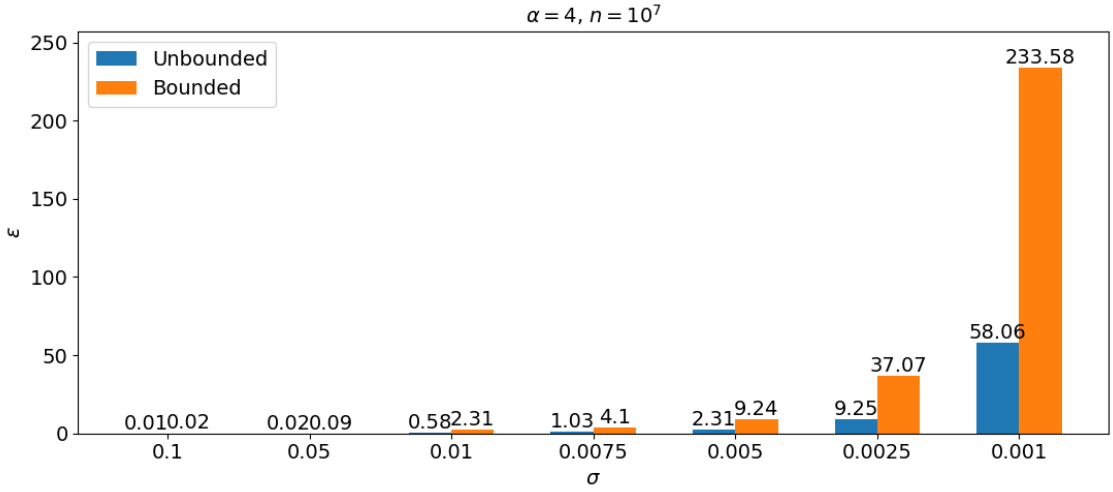

Figure 5.8: The values of ε for $(\alpha = 4, \varepsilon)$-RDP for each σ, where $d = 6$ and $n = 10^7$.

5.5.6 Summary of Results

To sum up the results of the numerical evaluations, we see the following:

- We see that ε is monotonically increasing with respect to α. This result is intuitive.

- If n_{in} increases exponentially, the curve becomes smaller at equal intervals on a logarithmic scale.

- When $n_{in} = 10^4$, a range where α satisfies the assumption of being very narrow. When $n_{in} = 10^7, 10^8$, the value of ε is practical.

- When σ falls below 0.001, the resulting ε becomes excessively large and falls outside the range considered practical.

5.6 Related Work

In this section, we describe the related work and mention the difference from our result.

5.6.1 Differentially Private Synthetic Data Generation for Tabular Data

In synthetic data generation, the post-processing property of differential privacy guarantees that synthetic data generated from differentially private generative parameters also satisfy differential privacy as shown in Figure 5.1(b). Methods to generate differentially private synthetic data for tabular data are classified into two types.

The first type is also called a "select-measure-generate" scheme [81]. Statistics and (conditional) probability distributions are used as the generative parameters. Typical statistics are mean vectors and covariance matrices of original datasets. In particular, synthetic data generation with copulas has been researched actively [108, 75, 6, 40]. To learn conditional distributions, graphical models such as Bayesian networks have been applied to synthetic data generation [123, 124, 82, 83].

In the second type, generative models with deep neural networks are used to generate synthetic data. The model parameters trained with the original data are regarded as the generative parameters. By training deep neural networks with differentially private stochastic gradient descent (DP-SGD) [1], we obtain differentially private model parameters. Methods based on generative adversarial networks (GAN) such as CTGAN [116], DPCTGAN [37], CtabGAN [127], and CtabGAN+ [128], are widely used. Methods other than GAN include variational autoencoders [19], flow-based generative models [73] and generative models with characteristic functions [76]. A method based on diffusion model such as TabDDPM [72] has also attracted attention recently.

In both types of approaches, generative parameters are computed by various differentially private mechanisms [1, 84] (Figure 5.1(b)). In contrast, we evaluate the differential privacy of randomness in data generation when using non-differentially private generative parameters.

5.6.2 Privacy Attacks against Synthetic Data Generation

Many methods empirically evaluate the privacy protection of synthetic data generations from attack success rates of membership inference attacks [107] and attribute inference attacks [39]. Most of them assume that an adversary has access to the target trained model such as GAN [17, 54, 59] and diffusion models [15, 60, 30, 80].

On the other hand, there are several methods where an adversary only has access to output synthetic data. Stadler et al. [109] discussed membership inference attacks and attribute inference attacks for tabular data in such a setting, and Oprisanu et al. [94] applied such attacks to genomic data. Annamalai et al. [4] conducted attribute inference with linear reconstruction in this setting.

Although these studies and ours share a common perspective in that they focus on the privacy protection of generated synthetic data alone, these studies differ from ours in that they experimentally evaluate synthetic data generation from an attack perspective. In contrast, our perspective is to prove Rényi differential privacy theoretically.

5.6.3 Differential Privacy of Randomness in Synthetic Data Generation

To the best of our knowledge, only Lin et al. [77] have evaluated the privacy protection by the randomness in outputs of synthetic data generations. They theoretically evaluated probabilistic differential privacy [85] of GAN-sampled data. However, the concretely evaluated bound is hard to compute since it needs a GAN's generalization error. In addition, they assume that training datasets are far larger than the number of model parameters. Thus, their main contribution is to give the theoretical bound, but we cannot compute the bound as a concrete numerical value.

In contrast, although we focus on only a simple synthetic data generation, we give the concretely computable bound.

5.7 Conclusion

In this chapter, we evaluated privacy protection due to the randomness of synthetic data generation without adding intentional randomness. We proved Rényi differential privacy of a synthetic data generation with a mean vector and covariance matrix (Theorem 5.3.1, Corollary 5.3.2). We also conducted numerical evaluations using the Adult dataset as a model case. Concretely, we demonstrated that the mechanism

$\mathcal{M}_G^n$ satisfies $(4, 0.576)$-RDP under the unbounded condition and $(4, 2.307)$-RDP under the bounded condition (Table 5.1). If they are translated into (ε, δ)-DP, $\mathcal{M}_G^n$ satisfies (ε, δ)-DP for a practical ε (Figures 5.5, 5.6, and 5.7). In future work, we will apply our evaluation method to more advanced synthetic data generation algorithms.

Chapter 6

Conclusion

This chapter provides a concise summary of the entire thesis, highlighting the main contributions and findings, and also outlines possible directions for future research.

6.1 Summary

This thesis has explored key challenges and solutions in the domain of synthetic data generation, with the ultimate goal of facilitating its safe and the societal implementation. Through a comprehensive investigation, this work has contributed to three main areas: the evaluation of privacy risks, the assessment of data utility, and the enhancement of privacy-preserving mechanisms in synthetic data generation.

First, we proposed a novel privacy evaluation framework to address limitations in prior membership inference attacks. By incorporating statistical distance-based sample selection and interpretable inference methods, we enabled clearer and more rigorous evaluation of worst-case privacy risks. This framework offers deeper insights into privacy vulnerabilities, particularly for outliers, and supports more transparent analysis without relying on black-box machine learning models.

Second, we evaluated the utility of synthetic data using a real-world medical dataset from Ehime University Hospital. Our experiments compared multiple generation methods—including statistical, graphical, and deep learning-based approaches —and demonstrated that Gaussian Copula and AIM generate synthetic data with high statistical fidelity and predictive performance. This practical evaluation underlines the viability and challenges of using synthetic data in sensitive domains such as healthcare.

Finally, we introduced a theoretical framework for analyzing differential privacy guarantees in synthetic data generation without relying on artificial noise. By mod-

eling privacy preservation through inherent randomness, we showed that Rényi differential privacy can be achieved under both bounded and unbounded neighboring conditions. These results offer a new perspective on achieving differential privacy while maintaining high data utility.

Together, these contributions advance the understanding of synthetic data generation techniques, and pave the way toward secure, ethical, and useful applications in data-driven fields.

6.2 Concluding Remarks

The key conclusions obtained through this thesis are summarized as follows.

Based on the framework proposed in Chapter 3, we identified the risks associated with synthetic data that are not protected by differential privacy. Our first key finding is that outliers in terms of Mahalanobis distance exhibit a high risk of membership inference attacks. As a result, we found that, when using non-DP synthetic data, we must at least remove outliers to ensure privacy. Alternatively, since the theoretical relationship between differential privacy and membership inference attacks is well established [121, 66], we recommend using DP synthetic data to achieve a certain level of privacy protection.

Thus, in Chapter 4, we addressed the challenges associated with Method (ii) in Table 1.1, that is the challenge of data utility evaluation of differentially private synthetic data generation with real-world data. Using real-world medical datasets, our findings are as follows:

- Adding noise for differential privacy leads to quality degradation even for real-world data, which exhibited a trend similar to that observed in public data.
- Among the evaluated methods, Gaussian Copula and AIM demonstrated superior performance.

Furthermore, this study also contributes the insight that, given the high data quality achieved by methods such as Gaussian Copula and AIM, identifying application scenarios where such quality is sufficient could serve as a practical pathway toward the societal implementation of synthetic data generation.

While Chapter 4 confirmed that the addition of noise contributes to quality degradation, Chapter 5 proposed an approach to guarantee differential privacy without intentional noise addition, relying solely on inherent randomness. As a result, we

showed that for simple methods, differential privacy can be ensured through inherent randomness alone. Moreover, we verified that the achieved privacy guarantees hold under epsilon values that are considered practical in real-world applications.

Building upon these insights, exploring real-world use cases is expected to facilitate the societal implementation of synthetic data generation.

6.3 Future works

This paper has pursued the goal of enabling the societal implementation of synthetic data generation by addressing the limitations inherent in prior approaches. Toward the realization of this objective, we conclude with a discussion of remaining challenges and highlight directions for future research. In particular, Section 6.3.2 is considered a key issue that should be addressed next.

6.3.1 Implementation of Comprehensive Evaluation Framework

In Chapter 3, we proposed a privacy evaluation framework that identifies high-risk records by selecting outliers based on the Mahalanobis distance. However, this framework does not incorporate alternative perspectives for outlier detection. For example, it is necessary to investigate whether similar characteristics emerge when using other outlier detection techniques, such as Isolation Forest (iForest)[78] or Support Vector Data Description (SVDD) [101], in place of Mahalanobis distance. Exploring the relationship between AUC scores and Mahalanobis distance is left for future work.

In addition, although Chapter 4 employed several utility metrics, it is desirable to evaluate utility alongside privacy risk. A meaningful direction for further research is to develop an integrated framework that simultaneously assesses the safety and utility of synthetic data generation methods, incorporating additional practical factors such as runtime performance.

6.3.2 Consideration of the real-world trials of Synthetic Data

Based on the results presented in Chapter 4, it was found that both Gaussian Copula and AIM are capable of generating synthetic data that closely resembles the original data, even under differential privacy constraints, when applied to real-world

datasets. For example, even under a strict privacy setting such as $\varepsilon = 1$, the differences in distributional error and machine learning model performance were minimal. It is, therefore, important to develop real-world application scenarios where such a level of privacy-utility trade-off is acceptable. Tabular formatted data requiring privacy protection are widely used in domains such as healthcare, finance, marketing, and product recommendation. Furthermore, such application scenarios are expected to help identify more concrete challenges toward societal implementation.

6.3.3 Differential Privacy Evaluation of More Practical Models without Adding Noise

In Chapter 5, we evaluated the inherent randomness of the data generation process and assessed its ability to satisfy Rényi differential privacy without the addition of explicit noise. However, the synthetic data generation model $\mathcal{M}_G : \mathcal{D} \to \mathcal{D}$ considered in the evaluation was a simple one, based solely on the mean vector and covariance matrix of the original dataset. If similar evaluations could be applied not only to such simple models, but also to graphical model–based methods such as AIM and deep learning neural network–based models, it would be expected that more useful synthetic data could be obtained. In particular, leveraging the properties of models such as diffusion models to simultaneously ensure both privacy guarantees and data utility remains an important direction for future research.

Acknowledgement

In the course of writing this thesis, I received invaluable support from many individuals. I would like to express my deepest gratitude to all of them.

First and foremost, I would like to extend my heartfelt thanks to Prof. Takanori Isobe, who kindly and promptly took over as my supervisor in the midst of an unexpected situation. His sincere advice and support played a crucial role in guiding me through the final stages of my doctoral studies. I am also deeply grateful to Prof. Makoto Onizuka and Prof. Atsuo Inomata for their valuable advice during the preparation for the preliminary review and for supporting me with regard to degree requirements. In particular, I am sincerely thankful to Prof. Onizuka for reviewing my thesis draft and sharing insightful perspectives on both academic writing and research presentations. I will make every effort to apply the perspectives I received from him on research presentations and paper writing in my future academic work. My sincere appreciation also goes to Prof. Tamami Nakano for her constructive comments on how to deepen the discussion in my research. Her advice greatly contributed to improving the final version of the manuscript.

I would like to express my profound gratitude to Prof. Toru Fujiwara, who has supported me since the beginning of my time at the laboratory. In addition to his direct guidance, observing how he mentored undergraduate and master's students was an invaluable learning experience for me. I am especially thankful to Assistant Prof. Kyosuke Yamashita, with whom I had the most frequent and insightful discussions. I learned a great deal from him, not only about research, but also about how to engage with students and what it means to foster a meaningful research environment.

My heartfelt thanks also go to Dr. Naoto Yanai, who inspired me to pursue my graduate studies at the University of Osaka and supported me throughout most of my time there. He taught me not only how to conduct research, but also how to manage a laboratory, write academic papers, and present myself at international conferences. I am especially thankful for his generous support during the revision

of Chapter 3, which was instrumental in improving the quality of the work. His guidance has been truly invaluable.

The results presented in Chapter 4 would not have been possible without the dedicated support of Prof. Eizen Kimura at Ehime University. I am deeply grateful for his efforts in organizing and providing the necessary data, setting up the experimental environment, assisting in debugging incomplete code, and offering valuable insights into the research direction. I sincerely hope to continue working together toward the practical application of synthetic data in clinical settings.

The research presented in Chapter 5 greatly benefited from the support of Associate Prof. Koji Chida at Gunma University. I am sincerely grateful for his assistance, not only in the technical aspects of the research, but also in discussions regarding the domestic deployment of related technologies and his active involvement in committee activities. I am especially thankful for his support as a supervisor at NTT when I decided to pursue my doctoral studies. I look forward to continuing our collaboration to further promote the practical use of synthetic data in Japan.

I would like to express my sincere appreciation to my colleagues at NTT Social Informatics Laboratories for their valuable support and insightful discussions throughout this work. In particular, I am deeply indebted to Dr. Toshiki Shibahara for his invaluable guidance on every aspect of my research, from the technical content to the overall approach. I am also sincerely grateful for his thoughtful support within the company, which greatly helped me in completing this thesis. I am also truly grateful to Kazuki Iwahana for his continuous support and invaluable advice on my research since his student days. The discussions we had in our laboratory at the University of Osaka remain irreplaceable memories for me. I would like to extend my sincere appreciation to my colleagues, including Masanobu Kii, Tetsuya Okuda, Dr. Atsunori Ichikawa, Juko Yamamoto, Osamu Saisho, Yusuke Yamasaki, and Satoshi Hasegawa, for their valuable support, especially in the preparation and writing of this thesis. Their support was indispensable to its completion. I am also grateful to my managers at NTT, including Toshiyuki Miyazawa, Tomoaki Washio, Sadahiro Ishizaki, and Hiromasa Tsugawa, for their thoughtful support and for fostering a work environment that enabled me to focus on my research.

My heartfelt thanks go to Hiroyuki Itakura, Ichiro Ishihara, and Susumu Kakuta from NTT TechnoCross Corporation, as well as Hiroshi Takeuchi and Yoshiyuki Kushibe from ARK Information Systems, INC., for their extensive support in both experimental work and discussions on research direction. Their contributions were essential to the progress of this study.

I would also like to sincerely thank Tomoya Matsumoto, who completed his master's degree at the University of Osaka, for the many valuable discussions we had, as well as for the opportunity to share experiences at international conferences in San Francisco and Vancouver, which significantly expanded my perspective. I look forward to seeing his continued success and hope to engage in further discussions with him as a fellow researcher in privacy-preserving technologies.

Last but not least, I would like to express my heartfelt thanks to my wife Ayaka, whom I married during the course of writing this thesis, for her unwavering support in our daily life. I am especially grateful for the good pressure she gave me at the Royal Host in Wakabayashi, which undoubtedly accelerated the completion of this thesis.

Reference

[1] 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 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526. Springer, 2018.

[3] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, volume 5, pages 901–909, 2005.

[4] Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. A linear reconstruction approach for attribute inference attacks against synthetic data. *arXiv preprint arXiv:2301.10053*, 2023.

[5] Apple. Apple differential privacy technical overview. `https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf`. Accessed: 2025-05-12.

[6] Hassan Jameel Asghar, Ming Ding, Thierry Rakotoarivelo, Sirine Mrabet, and Dali Kaafar. Differentially private release of datasets using gaussian copula. *Journal of Privacy and Confidentiality*, 10(2), 2020.

[7] Zahra Azizi, Chaoyi Zheng, Lucy Mosquera, Louise Pilote, and Khaled El Emam. Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, 11(4):e043497, 2021.

[8] Daniel Barth-Jones. The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. *Then and Now (July 2012)*, 2012.

[9] Abderrahim Oussama Batouche, Eugen Czeizler, Miika Koskinen, Tuomas Mirtti, and Antti Sakari Rannikko. Synergizing data imputation and electronic health records for advancing prostate cancer research: Challenges, and practical applications. *arXiv preprint arXiv:2311.02086*, 2023.

[10] Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnes, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th symposium on operating systems principles*, pages 441–459, 2017.

[11] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[12] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*, 2022.

[13] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[14] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *Proc. of IEEE S&P 2022*, pages 1897–1914. IEEE, 2022.

[15] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

[16] Mahalanobis Prasanta Chandra et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936.

[17] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 343–362, 2020.

[18] Kai Chen, Xiaochen Li, Chen Gong, Ryan McKenna, and Tianhao Wang. Benchmarking differentially private tabular data synthesis. *arXiv preprint arXiv:2504.14061*, 2025.

[19] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaarfar, and Haojin Zhu. Differentially private data generative models. *arXiv preprint arXiv:1812.02274*, 2018.

[20] Sung Hoon Cho, Jung Sook Jin, and Joo Seok Park. A study on the intention to use personal financial product recommendation mydata service. *The Journal of Bigdata*, 7(2):173–193, 2022.

[21] Edith Cohen, Haim Kaplan, Yishay Mansour, Shay Moran, Kobbi Nissim, Uri Stemmer, and Eliad Tsfadia. Data reconstruction: When you see it and when you don't. *arXiv preprint arXiv:2405.15753*, 2024.

[22] Ronald Cramer, Ivan Bjerre Damgård, and Jesper Buus Nielsen. *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, 2015.

[23] Sonia Cromp, Satya Sai Srinath Namburi GNVV, Mohammed Alkhudhayri, Catherine Cao, Samuel Guo, Nicholas Roberts, and Frederic Sala. Tabby: Tabular data synthesis with language models. *arXiv preprint arXiv:2503.02152*, 2025.

[24] Chris Culnane, Benjamin IP Rubinstein, and Vanessa Teague. Health data in an open world. *arXiv preprint arXiv:1712.05627*, 2017.

[25] Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022.

[26] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[27] Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions. *arXiv preprint arXiv:2207.04380*, 2022.

[28] Jorg Drechsler and JP Reiter. Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. *Journal of Official Statistics*, 25(4):589–603, 2009.

[29] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[30] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

[31] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

[32] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[33] Khaled El Emam. Seven ways to evaluate the utility of synthetic data. *IEEE Security & Privacy*, 18(4):56–59, 2020.

[34] Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR medical informatics*, 10(4):e35734, 2022.

[35] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[36] Erica Espinosa and Alvaro Figueira. On the quality of synthetic generated tabular data. *Mathematics*, 11(15):3278, 2023.

[37] Mei Ling Fang, Devendra Singh Dhami, and Kristian Kersting. Dp-ctgan: Differentially private medical data generation using ctgans. In *Artificial Intelligence in Medicine: 20th International Conference on Artificial Intelligence in Medicine, AIME 2022, Halifax, NS, Canada, June 14–17, 2022, Proceedings*, pages 178–188. Springer, 2022.

[38] Claudio Feijóo, José Luis Gómez-Barroso, and Peter Voigt. Exploring the economic value of personal information from firms ' financial statements. *International Journal of Information Management*, 34(2):248–256, 2014.

[39] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 17–32, 2014.

[40] Sébastien Gambs, Frédéric Ladouceur, Antoine Laurent, and Alexandre Roy-Gaumond. Growing synthetic data through differentially-private vine copulas. *Proceedings on Privacy Enhancing Technologies*, 2021(3):122–141, 2021.

[41] Quan Geng, Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

[42] Quan Geng and Pramod Viswanath. The optimal mechanism in differential privacy. pages 2371–2375, 2014.

[43] Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

[44] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *arXiv preprint arXiv:2211.10459*, 2022.

[45] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[46] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

[47] Mandeep Goyal and Qusay H Mahmoud. An llm-based framework for synthetic data generation. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00340–00346. IEEE, 2025.

[48] Elisabeth Griesbauer, Claudia Czado, Arnoldo Frigessi, and Ingrid Hobæk Haff. Tvinesynth: A truncated c-vine copula generator of synthetic tabular data to balance privacy and utility. *arXiv preprint arXiv:2503.15972*, 2025.

[49] John T Guibas, Tejpal S Virdi, and Peter S Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017.

[50] Aixia Guo, Randi E Foraker, Robert M MacGregor, Faraz M Masood, Brian P Cupps, and Michael K Pasque. The use of synthetic electronic health record

data and deep learning to improve timing of high-risk heart failure surgical intervention by predicting proximity to catastrophic decompensation. *Frontiers in digital health*, 2:576945, 2020.

[51] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[52] Gaspard Harerimana, Beakcheol Jang, Jong Wook Kim, and Hung Kook Park. Health big data analytics: a technology survey. *Ieee Access*, 6:65661–65678, 2018.

[53] David A Harville. Matrix algebra from a statistician's perspective, 1998.

[54] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: membership inference attacks against generative models. *Proceedings of Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

[55] Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Stealing links from graph neural networks. In *Proc. of USENIX Security 2021*, pages 2669–2686. USENIX Association, 2021.

[56] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

[57] Dayananda Herurkar, Ahmad Ali, and Andreas Dengel. Evaluating generative models for tabular data: Novel metrics and benchmarking. *arXiv preprint arXiv:2504.20900*, 2025.

[58] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data. *arXiv preprint arXiv:2211.06550*, 2022.

[59] Aoting Hu, Renjie Xie, Zhigang Lu, Aiqun Hu, and Minhui Xue. Tablegan-mca: Evaluating membership collisions of gan-synthesized tabular data releasing. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2096–2112, 2021.

[60] Hailong Hu and Jun Pang. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

[61] Weiming Huang, Baisong Liu, and Hao Tang. Privacy protection for recommendation system: a survey. In *Journal of Physics: Conference Series*, volume 1325, page 012087. IOP Publishing, 2019.

[62] Valter Hudovernik, Martin Jurkovič, and Erik Štrumbelj. Benchmarking the fidelity and utility of synthetic relational data. *arXiv preprint arXiv:2410.03411*, 2024.

[63] Jihyeon Hyeong, Jayoung Kim, Noseong Park, and Sushil Jajodia. An empirical study on the membership inference attack against tabular data synthesis models. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 4064–4068, 2022.

[64] Md Aminul Islam, Pretam Chandra, Bhupesh Kumar Mishra, SM Firoz, Ahmed Fahim, and Mozammel Hoque. Healthcare cost patterns and prediction: Investigating personal datasets using data analytics. *Authorea Preprints*, 2024.

[65] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proc. of USENIX Security 2019*, pages 1895–1912. USENIX Association, 2019.

[66] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *arXiv preprint arXiv:2005.10881*, 2020.

[67] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[68] Mishaal Kazmi, Hadrien Lautraite, Alireza Akbari, Qiaoyue Tang, Mauricio Soroco, Tao Wang, Sébastien Gambs, and Mathias Lécuyer. Panoramia: Privacy auditing of machine learning models without retraining. *Advances in Neural Information Processing Systems*, 37:57262–57300, 2024.

[69] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011.

[70] Min Sung Kim and Seongcheol Kim. Factors influencing willingness to provide personal information for personalized recommendations. *Computers in Human Behavior*, 88:143–152, 2018.

[71] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[72] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. *arXiv preprint arXiv:2209.15421*, 2022.

[73] Jaewoo Lee, Minjung Kim, Yonghyun Jeong, and Youngmin Ro. Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7345–7353, 2022.

[74] Haoran Li, Li Xiong, and Xiaoqian Jiang. Differentially private synthesization of multi-dimensional data using copula functions. In *Advances in database technology: proceedings. International conference on extending database technology*, volume 2014, page 475. NIH Public Access, 2014.

[75] Haoran Li, Li Xiong, Lifan Zhang, and Xiaoqian Jiang. Dpsynthesizer: Differentially private data synthesizer for privacy preserving data sharing. In *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, volume 7, page 1677. NIH Public Access, 2014.

[76] Seng Pei Liew, Tsubasa Takahashi, and Michihiko Ueno. PEARL: Data synthesis via private embeddings and adversarial reconstruction learning. In *International Conference on Learning Representations*, 2022.

[77] Zinan Lin, Vyas Sekar, and Giulia Fanti. On the privacy properties of gan-generated samples. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530. PMLR, 2021.

[78] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

[79] Tomoya Matsumoto, Takayuki Miura, Toshiki Shibahara, Masanobu Kii, Kazuki Iwahana, Osamu Saisho, and Shingo Okamura. Differentially private

sequential data synthesis with structured state space models and diffusion models. In *Neurips Safe Generative AI Workshop 2024*.

[80] Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 77–83. IEEE, 2023.

[81] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978*, 2021.

[82] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.*, 15(11):2599–2612, July 2022.

[83] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*, pages 4435–4444. PMLR, 2019.

[84] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[85] Sebastian Meiser. Approximate and probabilistic differential privacy definitions. *Cryptology ePrint Archive*, 2018.

[86] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *Proc. of IEEE S&P 2019*, pages 691–706. IEEE, 2019.

[87] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[88] Takayuki Miura, Masanobu Kii, Toshiki Shibahara, Kazuki Iwahana, Tetsuya Okuda, Atsunori Ichikawa, and Naoto Yanai. Setsubun: Revisiting membership inference game for evaluating synthetic data generation. *Journal of Information Processing*, 32:757–766, 2024.

[89] Takayuki Miura, Eizen Kimura, Atsunori Ichikawa, Masanobu Kii, and Juko Yamamoto. Evaluating synthetic data generation techniques for medical

dataset. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 2: HEALTHINF*, pages 315–322. INSTICC, SciTePress, 2024.

[90] Takayuki Miura, Toshiki Shibahara, Masanobu Kii, Atsunori Ichikawa, Juko Yamamoto, and Koji Chida. On rényi differential privacy in statistics-based synthetic data generation. *Journal of Information Processing*, 31:812–820, 2023.

[91] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[92] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pages 866–882. IEEE, 2021.

[93] Samuel Oladiipo Olabanji, Oluseun Babatunde Oladoyinbo, Christopher Uzoma Asonze, Tunboson Oyewale Oladoyinbo, Samson Abidemi Ajayi, and Oluwaseun Oladeji Olaniyi. Effect of adopting ai to explore big data on personally identifiable information (pii) for financial and economic data transformation. *Available at SSRN 4739227*, 2024.

[94] Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. On utility and privacy in synthetic genomic data. In *Proceedings 2022 Network and Distributed System Security Symposium. NDSS*, volume 22, pages 1–17, 2022.

[95] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[96] Caibe A Pereira, Rômulo CR Peixoto, Manuella P Kaster, Mateus Grellert, and Jônata Tyska Carvalho. Using data mining techniques to understand patterns of suicide and reattempt rates in southern brazil. In *BIOSTEC (2)*, pages 385–392, 2024.

[97] Vasileios C Pezoulas, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal*, 2024.

[98] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–5, 2017.

[99] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023.

[100] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[101] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[102] Osamu Saisho, Takayuki Miura, Kazuki Iwahana, Masanobu Kii, and Rina Okada. Active learning for human annotation of privacy-preserved synthetic data. In *Proc. of PSD2024*, pages 1–15, 2024.

[103] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Proc. of NDSS 2019*. The Internet Society, 2019.

[104] Gabriele Santangelo, Giovanna Nicora, Riccardo Bellazzi, Arianna Dagliati, et al. Synthcheck: A dashboard for synthetic data quality assessment. In *BIOSTEC (2)*, pages 246–256, 2024.

[105] Zihao Shan, Kui Ren, Marina Blanton, and Cong Wang. Practical secure computation outsourcing: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–40, 2018.

[106] Toshiki Shibahara, Takayuki Miura, Masanobu Kii, and Atsunori Ichikawa. Efficiently calculating stronger lower bound for differentially private sgd in black-box setting. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 970–975. IEEE, 2024.

[107] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[108] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[109] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[110] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[111] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020.

[112] Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238*, 2021.

[113] United States Census Bureau. Census bureau sets key parameters to protect privacy in 2020 census results. `https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html`. Accessed: 2025-05-12.

[114] Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[115] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[116] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

[117] Juko Yamamoto, Takayuki Miura, Rina Okada, Masanobu Kii, and Atsunori Ichikawa. Explaining and visualizing synthetic data quality using statistical distances. In *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2025.

[118] Yusuke Yamasaki, Kenta Niwa, Daiki Chijiwa, Takumi Fukami, and Takayuki Miura. Plausible token amplification for improving differentially private in-context learning based on implicit bayesian inference. In *International Conference on Machine Learning*. PMLR, 2025.

[119] Mengmeng Yang, Chi-Hung Chi, Kwok-Yan Lam, Jie Feng, Taolin Guo, and Wei Ni. Tabular data synthesis with differential privacy: A survey. *arXiv preprint arXiv:2411.03351*, 2024.

[120] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3093–3106, 2022.

[121] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proc. of CSF 2018*, pages 268–282. IEEE, 2018.

[122] Hessam Zakerzadeh, Charu C Aggarwal, and Ken Barker. Towards breaking the curse of dimensionality for high-dimensional privacy. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 731–739. SIAM, 2014.

[123] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 42(4), October 2017.

[124] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. {PrivSyn}: Differentially private data synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 929–946, 2021.

[125] Ying Zhao and Jinjun Chen. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28, 2022.

[126] Zilong Zhao, Robert Birke, and Lydia Chen. Tabula: Harnessing language models for tabular data synthesis. *arXiv preprint arXiv:2310.12746*, 2023.

[127] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2021.

[128] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan+: Enhancing tabular data synthesis. *arXiv preprint arXiv:2204.00401*, 2022.

[129] Yongrui Zhong, Yunqing Ge, Jianbin Qin, Shuyuan Zheng, Bo Tang, Yu-Xuan Qiu, Rui Mao, Ye Yuan, Makoto Onizuka, and Chuan Xiao. Privacy-enhanced database synthesis for benchmark publishing. *arXiv preprint arXiv:2405.01312*, 2024.

[130] 三浦尭之, 紀伊真昇, 市川敦謙, 岩花一輝, 芝原俊樹, 奥田哲矢, 山本充子, and 矢内直人. 合成データ生成の出力を評価するメンバーシップ推論攻撃フレームワーク. In **コンピュータセキュリティシンポジウム** *2022* **論文集**, pages 448–455, oct 2022.

[131] 三浦尭之, 紀伊真昇, 市川敦謙, 岩花一輝, 芝原俊樹, 奥田哲矢, and 矢内直人. 合成データに対するメンバーシップ推論攻撃評価フレームワークの拡張. In **マルチメディア、分散、協調とモバイルシンポジウム** *DICOMO2023* **論文集**, jul 2023.

[132] 三浦尭之, 紀伊真昇, 市川敦謙, 山本充子, and 木村映善. 合成データ生成技術の医療データへの適用と課題. In **第** *27* **回日本医療情報学会春季学術大会**, pages 68–69, jun 2023.

[133] 三浦尭之, 紀伊真昇, 市川敦謙, 千田浩司, and 木村映善. 医療データへの合成データ生成技術適用に向けた一検討. In **第** *97* **回コンピュータセキュリティ・第** *57* **回インターネットと運用技術合同研究発表会**, pages 1–8, may 2022.

[134] 三浦尭之, 紀伊真昇, 芝原俊樹, 市川敦謙, 山本充子, and 千田浩司. 合成データ生成のランダム性が持つ rényi 差分プライバシー性の評価. In **コンピュータセキュリティシンポジウム** *2022* **論文集**, pages 440–447, oct 2022.

[135] 三浦尭之, 紀伊真昇, 芝原俊樹, 市川敦謙, 山本充子, and 千田浩司. ベイジアンネットワークによる合成データ生成時のランダム性が持つ差分プライバシー性の評価. In **コンピュータセキュリティシンポジウム** *2023* **論文集**, pages 1389–1396, oct 2023.