HHS Public Access

Author manuscript

IEEE Trans Knowl Data Eng. Author manuscript; available in PMC 2019 March 01.

Published in final edited form as:

IEEE Trans Knowl Data Eng. 2018 March 1; 30(3): 573–584. doi:10.1109/TKDE.2017.2773545.

Selecting Optimal Subset to release under Differentially Private M-estimators from Hybrid Datasets

Meng Wang,
Department of Biomedical Informatics, University of California at San Diego, CA, 92093 U.S., and now is with the Department of Genetics, Stanford University, CA, 94305, U.S

Zhanglong Ji,
Department of Biomedical Informatics, University of California at San Diego, CA, 92093 U.S

Hyeon-Eui Kim,
Department of Biomedical Informatics, University of California at San Diego, CA, 92093 U.S

Shuang Wang,
Department of Biomedical Informatics, University of California at San Diego, CA, 92093 U.S

Li Xiong, and
Department of Computer Science, Emory University, GA, 30322 U.S

Xiaoqian Jiang
Department of Biomedical Informatics, University of California at San Diego, CA, 92093 U.S

Abstract

Privacy concern in data sharing especially for health data gains particularly increasing attention nowadays. Now some patients agree to open their information for research use, which gives rise to a new question of how to effectively use the public information to better understand the private dataset without breaching privacy. In this paper, we specialize this question as selecting an optimal subset of the public dataset for M-estimators in the framework of differential privacy (DP) in [1]. From a perspective of non-interactive learning, we first construct the weighted private density estimation from the hybrid datasets under DP. Along the same line as [2], we analyze the accuracy of the DP M-estimators based on the hybrid datasets. Our main contributions are (i) we find that the bias-variance tradeoff in the performance of our M-estimators can be characterized in the sample size of the released dataset; (2) based on this finding, we develop an algorithm to select the optimal subset of the public dataset to release under DP. Our simulation studies and application to the real datasets confirm our findings and set a guideline in the real application.

Index Terms

differential privacy; M-estimators; hybrid datasets

1 Introduction

A common challenge in clinical studies is accessing sufficient amount of data of the subjects with a condition of interest. This becomes even more challenging when the study is set in a limited setting of a single institution and targets a rare disease. Hence, data sharing is of

great interest to biomedical research for accelerating scientific discovery which can be translated into novel treatment methods. However, revealing sensitive information about patients or individuals is an important privacy concern in sharing healthcare data. This concern is gaining particularly increasing attention in healthcare as shown in [3–8]. Nowadays, to protect private information to some extent, accessing raw data from multiple institutions requires not only Institutional Review Boards (IRBs) approval but also conforming to the institutional Data Use Agreement (DUA) conditions, which becomes a time-consuming and laborious process. It would be ideal to have a mechanism to support healthcare data analysis in a privacy-preserving manner by maximizing the utility of available dataset while reducing the burden of undergoing the complex process around IRB and DUA.

In the literature of protecting private information from synthetic data, Rubin first proposed the idea of fully synthetic data to represent “no actual individual” [9], which attracted a lot of attentions. Later, Reiter provided guidance on specifying the number and size of Rubin’s synthetic data generation model to ensure valid statistical procedures [10]. But Abowd and Vilhuber discussed potential “disclosure risk” of such method through inference disclosure [11]. In the imputation model. [12] described partial synthesis of survey data collected by the Cancer Care OutcomesResearch and Surveillance (CanCORS) project and discussed the key steps of selecting variables for synthesis, specification of imputation models, measurements of data utility, and disclosure risk.

Due to the risk concerns that were previously identified, modern synthetic data generation methods mostly driven by differential privacy. Differential privacy (DP) [1] uses perturbation techniques giving a mathematically rigorous privacy guarantee even if an attacker has arbitrarily auxiliary information. There are many and an increasing number of research fields in applying the concept of DP and DP mechanisms to machine learning [13], statistics [14; 15] and medicine application [6; 16–18]. Under the framework of DP, Mohammed et al. developed decide-tree based synthetic data release using a top down model that dynamically specifies sibling nodes and perturbs counts satisfying differential privacy [19]. Machanavajjhala et al. introduced another synthetic data generation model for commuting patterns of the population of the United States [20], which are very sparse in nature. Hall et al. developed yet another method to release differentially private functional data in the reproducing kernel Hilbert space (RKHS) by introducing an appropriate Gaussian process to the function of interest in [21]. These methods only retain utility for special classes of functions, which are not generalizable to medical data at large.

In reality, some patients are willing to sign the open consents to make their information freely available, which gives rise to a new problem of integrating the public information to infer from the private information. Several works show that hybridizing public and private information can significantly improve the utility of the DP mechanisms in specific learning goals. In logistic regression, Ji et al. [22] developed DP algorithm based on importance sampling for the hybrid datasets. A recent work [23] modified the update step in Newton-Raphson method in logistic regression under DP based on public data and private data together. Another similar work [24] developed a hybrid DP support vector machine (SVM)

model that also takes advantage of aforementioned hybrid datasets. These models are still specific to certain machine learning models.

Different from existing task-specific models, we propose a new algorithm to publish data under DP using public and private datasets. It is not designed for a special task but the algorithm works for a general task such as estimation, regression and classification. To achieve this goal, we construct a weighted private density from the hybrid datasets under DP (in Section 3). We focus on protecting privacy in M-estimators, which are well studied in statistics [25] without privacy concerns. Under some regularities, M-estimators are robust to the outliers. This property makes it possible for M-estimators to maintain high utility under DP as discussed in [14].

Our workflow to infer the M-estimators under DP is summarized in Figure 1. Consider a situation that the researcher (the data user) would like to conduct exploratory data analysis to get M-estimators (for example, the coefficients of the covariates in the logistic regression) over a private dataset before making the formal IRB application. The researcher asks for a best model under DP from a trusted third party (TTP), which can access both public and private datasets. One way is to directly infer the M-estimators from a private dataset under DP. [2] provides a good framework for low dimensional data by perturbing the private histogram but this method could become useless for high dimensional data. Noticing the related public information, the TTP could take the result from the public dataset to represent that for the private dataset. However, we need to be extra careful here. If the distribution of the public dataset is the same as that of the private dataset, it will be perfect to directly infer from the public dataset without costing any privacy budget. But in most cases, the open-consent population itself could have bias, such as the young and high educated people are probably more willing to make their information open for research so that some covariate in the open-consent population has smaller variation than that in the private population. Hence if we directly learn from the data from consented population, the results could not fully represent the private data.

Due to the possible bias from public dataset, how to effectively use the public information to better understand the private information without releasing privacy is a challenging question. Our main contributions (in Section 4) are (i) we provide a theoretical support that there could exist an optimal subset of the public dataset, not always the whole public dataset, giving the highly accurate and privacy protected M-estimators and (ii) we further develop a DP selection procedure to obtain this optimal subset. Given the private data of interest, our method generates a DP hybrid cohort by referencing the public data to pick the most representative samples to facilitate the study, which ensures the valid statistical procedures and prevents against inference disclosure problem in Rubin's case [9]. The simulation studies and an application in real datasets in logistic regression (in Section 5) show that the optimal public subset from our selection procedure performs better than the perturbed histogram method in [2] and promisingly our procedure can be applied in high dimensional data. We discuss future directions and limitations and finally make a conclusion in Section 6.

2 Preliminary

2.1 Setting of the problem

Let $D = \{x_1, \ldots, x_{nD}\}$ be the private dataset and $E = \{y_1, \ldots, y_{nE}\}$ the public dataset where y's are distinct. Each data point $x_i \in D$ is a realization of a random variable X from a density function f_D and $y_i \in E$ a realization of Y from f_E. We generally call the density function for the probability mass function of a discrete variable. We assume $X_i \overset{\text{iid}}{\sim} f_D, Y_i \overset{\text{iid}}{\sim} f_E$ and X's and Y's are independent. Due to the bias from open-consent preselection, we consider generally $f_D \neq f_E$. Denote the sample space of X by $\mathcal{X}_D$ and Y by $\mathcal{X}_E$.

Notation—The symbol $|\cdot|$ of a set means the cardinality of the set. For $a, b \in \mathbb{R}$, denote $a \vee b = \max\{a, b\}$. For two sequences of reals (a_n) and (b_n): $a_n \sim b_n$ when $a_n/b_n \to 1$; $a_n = o(b_n)$ when $a_n/b_n \to 0$; $a_n = O(b_n)$ when a_n/b_n is bounded. For two real random sequences (a_n), (b_n), $a_n = O_p(b_n)$ if a_n/b_n is bounded in probability, i.e., $\sup_n \mathbb{P}(|a_n/b_n| > x) \to 0$ as $x \to \infty$. For a random variable X and density function f, $X \sim f$ means X has density F. $X \sim Laplace(\lambda)$ means X has density $\frac{1}{2\lambda} \exp(-|x|/\lambda)$.

2.2 Differential Privacy

Differential privacy proposed by [1] provides guarantees on the privacy of the dataset against any arbitrary external attacks. The concept of differential privacy (DP) is designed to protect the worst case that the attacker knows the information of all the patients except one. To protect this case, DP considers two neighbor datasets D and D' which differ in only one patient. The rigorous definition of ε-differential privacy is in Definition 1. The definition of DP is symmetric in D and D'. By the comparability property of DP in Definition 4, it is easily applied to the case where D and D' differ in more than one patients.

Laplace mechanism [1] in Definition 3 is one of mechanisms that achieve ε-differential privacy by adding Laplace noise on private information, which is the mechanism we use in this paper. The amount of noise added is determined by the privacy parameter ε, and the sensitivity of the statistics in Definition 2 which characterizes the worst case among all neighbor datasets.

Definition 1. (Differential Privacy [1])—*A randomized mechanism $\mathcal{M}$ is ε-differentially private if, for all datasets D and D' which differ on at most one individual and for any measurable subset S of the range of $\mathcal{M}$,*

$$\left|\log \frac{\mathbb{P}(\mathcal{M}(D) \in S)}{\mathbb{P}(\mathcal{M}(D') \in S)}\right| \leq \varepsilon, \quad \forall S \subseteq range(\mathcal{M}).$$

In Definition 1, D and D' are called *neighbor datasets* with Hamming distance $H(D, D') = |D \backslash D'| = |D' \backslash D| = 1$. ε is the privacy parameter. Small ε tells the distribution of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ is not easily distinguishable so that in some sense this protects the information of the patients that D and D' are different in. Note here the randomness of $\mathcal{M}$ is in the perturbation added to protect privacy not in the data itself.

Definition 2. (Sensitivity for Laplace Mechanism [1])—*The sensitivity of a statistic T in the Laplace Mechanism is the smallest number S(T) such that*

$$\|T(D) - T(D')\|_1 \leq S(T),$$

for all D, D′ with H(D, D′) = 1.

Definition 3. (Laplace Mechanism [1])—*Releasing (T(D) + noise) where noise has Laplace(S(T)/ε) distribution is an ε-differentially private mechanism satisfying Definition 1, where S(T) is the sensitivity of T in Definition 2.*

Definition 4. (Sequential Composition [1])—*If we apply k independent statistics $T_1, \ldots, T_k$ with corresponding privacy parameters $\varepsilon_1, \ldots, \varepsilon_k$, then any function $g(T_1, \ldots, T_k)$ is $\sum_{i=1}^{k} \varepsilon_i$-differentially private.*

2.3 M-estimator

In our case, the parameter we are interest in is the one for the private distribution,

$$\theta^*(\mathcal{X}_D) = \arg\min_{\theta \in \Theta} M_{\mathcal{X}_D}(\theta), \quad (1)$$
$$\text{where } M_{\mathcal{X}_D}(\theta) = \int_{\mathcal{X}_D} m(x, \theta) f_D(x) \mathrm{d}x,$$

and Θ is the parameter space of θ and $m(x, \theta)$ is called *contrast function*. Given the private dataset $D = \{x_1, \ldots, x_{nD}\}$ and under the independence assumption of X's, the M-estimator for $\theta^*(\mathcal{X}_D)$ is the minima of the sum of contrast functions,

$$\hat{\theta}_{(D)} = \arg\min_{\theta \in \Theta} M_D(\theta), \quad (2)$$
$$\text{where } M_D(\theta) = \sum_{x_i \in D} \frac{1}{n_D} m(x_i, \theta).$$

For example, taking $m(x, \theta)$ as absolute loss function $|x - \theta|$, $\hat{\theta}$ is the sample median; taking $m(x, \theta)$ as squared loss function $(x - \theta)^2$, $\hat{\theta}$ is the sample mean; taking $m(x, \theta)$ as the negative log-likelihood function, $\hat{\theta}$ is maximum likelihood estimator. More examples and properties of M-estimators can be found in [25].

Since D is private, we can not release $\hat{\theta}(D)$ in (2) without any privacy protection. We take $\hat{\theta}(D)$ as a baseline to evaluate the performance of the privacy protected M-estimator. Our

goal is to develop a M-estimate under DP integrating public information to achieve high utility.

3 DP M-estimator from Hybrid Datasets

3.1 Our DP M-estimator

In a non-interactive fashion, a direct way to release a DP version of the M-estimator is to perturb the density function. Once we get a privacy protected density, based on the synthetic data points, we can not only ask for the mean, the median of the private dataset, but also learn the private dataset form linear regression, logistic regression, in a series of estimating M-estimators.

Only based on the private dataset D, Lei in [2] proposed a method to get DP M-estimator from perturbed histogram. The key is to perturb the private density function under DP. His strategy is as follows. We first partition the sample space (assuming bounded) into cubic cells with equal bandwidth in each dimension, next add Laplace noise to the counts of the points in D lying in each cell under DP, then generate the synthetic data by replicating the center point in each cell in the number of the DP perturbed count, finally obtain the DP M-estimator from the synthetic points. But this model does not scale well if the dimensionality is high since adding noise to all high dimensional cells may render the estimated density useless. In our setting, this problem can be alleviated if we carefully make use of the public dataset. We compare their performances in the simulation studies in Section 5.

At the presence of the related public dataset, we would like to take use of the public data points to represent the private data points in a privacy-preserving manner and construct a DP hybrid weighted density function. Our idea is to assign a weight to each distinct public point, where the weight in one public point is proportional to the count of the private points such that they are closer to this point rather than any other public points. In statistics, the hybrid weighted density is formulated in

$$\hat{f}_D^{hyb}(x) = \sum_{i=1}^{n_E} w_i \, \mathbb{1}_{\{x = y_i\}}, \quad (3)$$

where

$$w_i = \frac{|D_{y_i}|}{n_D}, \; y_i \in E,$$

$$D_{y_i} = \{x \in D : \|x - y_i\| \le \|x - y_j\|, \text{ for all } j \neq i\}.$$

By the definition of D_{y_i}, we can see if a point in E surrounded by more points in D, it gains large weight and otherwise it gains small weight. In this way, it relieves the bias of estimating from E to some extent.

In (3) to estimate the private density, we relate its empirical probability to that of the public empirical probability by the weight w_i. Without privacy concerns, there are several ways in density estimation. The book [26] summarizes several approaches, such as moment matching, probabilistic classification, and density-ratio fitting with kernel smoothing. In this paper, our contribution is not to find an optimal way to estimate private density under DP but instead under the constructed hybrid weighted density to establish a procedure to better use the public information to reduce unnecessary privacy budget allocation and therefore improve the utility.

Definition 5. (DP M-estimator from hybrid datasets.)—Under the notation in Section 2, the DP M-estimator from hybrid datasets is

$$\hat{\theta}(E, \tilde{w}) = \arg\min_{\theta \in \Theta} M_{(E, \tilde{w})}(\theta), \quad (4)$$

where

$$M_{(E, \tilde{w})}(\theta) = \sum_{i=1}^{n_E} \tilde{w}_i m(y_i, \theta),\ y_i \in E,$$

$$\tilde{w}_i = w_i + Z_i/n_D,\ w_i = \frac{|D_{y_i}|}{n_D},$$

$$D_{y_i} = \{x \in D: \|x - y_i\| \le \|x - y_j\| \text{ for all } j \ne i\}$$

$$Z_i \overset{\text{iid}}{\sim} Laplace(2/\varepsilon),\ i = 1, \ldots, n_E.$$

Proposition 1— $\{\tilde{w}_i\}_{i=1}^{n_E}$ *in Definition 5 satisfies ε-differential privacy.*

Proof: To better understand the concept of DP, we give the proof for Proposition 1. Suppose D' and D are neighbor datasets, and $x_{i_0} \in D \backslash D'$ and $x_{i'_0} \in D' \backslash D$. Let $T(D) = (|D_{y_1}|, \ldots, |D_{y_{nE}}|)$ and $T(D') = (|D'_{y_1}|, \ldots, |D'_{y_{n_E}}|)$. Since we add independent Laplace noise to T, it suffices to show that the sensitivity of T is 2. In the case that x_{i_0} and $x_{i'_0}$ have the same representative point in E, then $T(D) = T(D')$. In the other case that they have different representatives, denote $x_{i_0} \in D_{y_j}$ and $x_{i'_0} \in D'_{y'_j}$, where $y_j \ne y'_j$. Then $|D_{y_j}| = |D'_{y_j}| + 1$ and $|D_{y'_j}| = |D'_{y'_j}| - 1$ and $T(D)$ and $T(D')$ are the same in other elements. Hence, $\|T(D) - T(D')\|_1 \le 2$ for all neighbor datasets D and D', which gives $S(T) = 2$ by Definition 2.

As we can see in Definition 5, adding Laplace noise can lead some weights $\tilde{w}_i$'s negative. Many negative weights can cause non-convexity of $M_{(E, \tilde{w})}(\theta)$. Hence, it is necessary to define a truncated version of the weights. Parallel to the perturbed histogram with

nonnegative counts in [2], we define the truncated perturbed weights and obtain the DP M-estimator from hybrid datasets with nonnegative weights in Definition 6.

Definition 6. (DP M-estimator from hybrid datasets with nonnegative weights.) —In the same notation in Definition 5, replacing $\tilde{w}_i$'s by

$$\tilde{w}_i^+ = \tilde{w}_i \vee 0,$$

the DP M-estimator from hybrid datasets with nonnegative weights is

$$\hat{\theta}(E, \tilde{w}^+) = \arg\min_{\theta \in \Theta} M_{(E, \tilde{w}^+)}(\theta),$$
$$\text{where } M_{(E, \tilde{w}^+)}(\theta) = \sum_{i=1}^{n_E} \tilde{w}_i^+ m(y_i, \theta), y_i \in E.$$

Proposition 2— $\{\tilde{w}_i^+\}_{i=1}^{n_E}$ *in Definition 6 satisfies ε-differential privacy.*

Proof: By Definition 6, $\tilde{w}_i^+$ is a measurable function of $\tilde{w}_i$. From Proposition 1, $\{\tilde{w}_i\}_{i=1}^{n_E}$ satisfies DP so does $\{\tilde{w}_i^+\}_{i=1}^{n_E}$ by Definition 1.

3.2 Algorithm to Get Our DP M-estimator

To obtain the DP M-estimator from hybrid datasets in practice, we summarize the steps in the *Algorithm 1*. We first rescale both the public dataset and the private dataset to $[0, 1]^p$ where p is the data dimension; next calculate the weights in Definition 5 from the rescales datasets to establish the hybrid density; then add Laplace noise to the weights under DP; finally we obtain the DP M-estimator for the user assigned contrast function. To obtain the DP M-estimator with nonnegative weights, the steps follow from *Algorithm 1* except replacing $\tilde{w}$ by $\tilde{w}^+$ in Definition 6.

Before constructing the hybrid density, we add a preprocessing step to rescale the datasets in Step 1. We claim that it is a necessary step from two aspects. In one aspect, rescaling the data in each dimension to [0, 1] makes the scales of all dimensions comparable. For a categorical variable, suppose it has k categories. We first transform the labels to k-dimensional dummy variables, (1, 0, …, 0) for label 1, (0, 1, 0, …, 0) for label 2, …, (0, …, 0, 1) for label k. Hence, the l_2 distance between two different labels is $\sqrt{2}$. We further divide the transformed labels by $\sqrt{2}$ so that the l_2 distance between any two categories inside [0, 1]. In another aspect, we need to guarantee privacy protection in each step involving the private dataset. Hence we rescale the private dataset according to the range of the public dataset in case that the private data point outside the range of public dataset could expose privacy.

To see the effect of rescaling, we first illustrate the original datasets and rescaled ones from a simulation in Figure 2. In the simulated data, we consider the private dataset D containing $n_D = 10^4$ points has a two-dimensional covariate X_D generated from $MVN((0, 0)^T, diag(1, 1))$ where $diag(\cdot)$ denotes a diagonal matrix and a 0–1 response variable generated from a logistic regression model based on X_D and coefficient $\beta_D = (0.2, 0.4)$, where MVN is the abbreviation for multivariate normal distribution. Similarly for the public dataset E, $n_E = 10^4$, X_E is generated from $MVN((0, 0)^T, diag(0.5, 1))$ and $\beta_E = (0.5, -0.1)$. We simulate the case that D and E have different distributions due to open-consent bias. The lower panels in Figure 2 show the rescaled private and public datasets. We can see since the covariate-1 has less variation in original E than that in original D, after rescaling D according to the range of E, several points in D in large deviation in covariate-1 are compressed to the boundaries in the rescaled $\tilde{D}$. Rescaling step changes the relative positions of two data points. However, what really matters is how much it changes the density weights because the change in the weights will more or less affect the accuracy of the M-estimator. To illustrate this idea, we compare the weights calculated from original D and E to those from rescaled $\tilde{D}$ and $\tilde{E}$ and highligt the points with different weights from two ways of calculation in rescaled $\tilde{E}$ in Figure 3. We can see the points with large different weights mainly lie in the boundaries of covariate-1. (Note that 0–1 response variable also contribute to the calculation of the weights.) From this simulation, we get an idea that the step of rescaling could add more variation on the weights especially on the points compressed to the boundaries. This is another cost of protecting privacy, besides adding Laplace noise on the weights. If the public dataset and the private are not quite different like in this simulation, the number of the points that have more than 2 counts different in weights is only 42 out of 10^4 points. If the model is not extremely sensitive to the outliers, we think a few points with large deviated weights could not make a large difference to the final result. On the other hand, this indicates that if the public dataset and the private are extremely different such that in some dimension they have quite different distributions, rescaling step may deteriorate the accuracy in the final estimation. In this case, we suggest finding a more comparable public dataset in order to apply the privacy-protection procedures. More discussion is in Section 6

Algorithm 1

To obtain DP M-estimator from hybrid datasets.

Input: private dataset D, public dataset E with distinct points, privacy parameter ϵ and objective function M.
Output: DP M-estimator from hybrid datasets $\hat{\theta}(E, w)$.

1 *To rescale D and E to* $[0, 1]^p$ *where p is the dimension of the data points.* Combine D and E by the variables in E. Let z_{ij} be i-th observation in j-th variable of the combined dataset. If j-variable is continuous, rescale z_{ij} to

$$\tilde{z}_{ij} = \frac{z_{ij}^{trunc} - a_j}{b_j - a_j}$$

where $a_j = \min_{i = 1, \ldots, n_D + n_E} \{z_{ij} \in E\}$,

$b_j = \max_{i = 1, \ldots, n_D + n_E} \{z_{ij} \in E\}$,

$$z_{ij}^{trunc} = a_j \mathbb{1}\{z_{ij} < a_j\} + b_j \mathbb{1}\{z_{ij} > b_j\} + z_{ij} \mathbb{1}\{a_j \le z_{ij} \le b_j\}$$

If j-th variable is categorical with k-labels, write z_{ij} in terms of k-dimensional dummy variable then rescale it by dividing $\sqrt{2}$.

2 *To find the weights w_i's in Definition 5 based on rescaled hybrid datasets from Step 1.* Define $\tilde{x}$'s and $\tilde{y}$'s are the points in the rescaled dataset $\tilde{D}$ and the rescaled $\tilde{E}$. The weight associated with $\tilde{y}_i \in \tilde{E}$ is $w_i = |\{\tilde{x} \in \tilde{D}: \|\tilde{x}-\tilde{y}_i\|_2 \le \|\tilde{x}-\tilde{y}_j\|_2, i \ne j\}|/n_D$.

3 *Adding Laplace noise to the weights.* To each w_i obtained in Step 2 adding independent Laplace noise, $\tilde{w}_i = w_i + Z_i/n_D$ where $Z_i \overset{\text{iid}}{\sim} Laplace(2/\varepsilon)$, $i = 1, \ldots, n_E$.

4 *To obtain DP M-estimator from hybrid datasets.* Minimizing the objective function $M_{(E, \tilde{w})}(\theta)$ in Definition 5 where $\tilde{w}$ is obtained in Step 3, get $\hat{\theta}(E, \tilde{w}) = \arg\min_\theta M_{(E, \tilde{w})}(\theta)$.

4 Optimal Subset in Releasing Public Dataset under DP

In the previous section, we establish the DP M-estimator from hybrid datasets. In this section, we first evaluate its performance from bias and variance decomposition Subsection 4.1. We find that this bias and variance tradeoff can be characterized in the number of the public data points to release. This makes it possible that there may exist an optimal subset of public dataset which could give more accurate M-estimators under DP. This is our main contribution. In Subsection 4.2, we further develop an algorithm to select the optimal subset under DP in practice.

4.1 Bias and Variance Tradeoff

At a first look, one may doubt since we have the public dataset, why only to use part of them. It is true with more points in the public dataset, they can be more representative of the private dataset. However, under the framework of DP, we need to add Laplace noise to the weights of the public data points. More points to release, more noise to add. Hence, there is a tradeoff between bias and variance, where the bias is from using public dataset to estimate the private dataset and inadequate sample size while the variance from adding noise. We find this tradeoff is in the sample size of the public subset. It is analogous to the case where the whole dataset is private where the tradeoff is in the bandwidth of the perturbed histogram in [2]. However there how to choose the cell bandwidth is a tricky question in practice but in our case, selecting the public subset can be optimized.

To analyze the accuracy of our M-estimator, we consider the contrast function is well-defined in the same conditions as in [2]. Assume the contrast function $m(x, \theta): \mathcal{X} \times \Theta \to \mathbb{R}$ where $\mathcal{X} = \mathcal{X}_D \cup \mathcal{X}_E$ is the union of the private and public sample spaces and Θ is the parameter space of θ, satisfies condition $(A1) - (A3)$ listed below.

(A1) $g(x, \theta) := \frac{\partial}{\partial\theta} m(x, \theta)$ exists and $\|g(x, \theta)\|_2 \le C_1$ on $\mathcal{X} \times \Theta$ where C_1 is a constant;

(A2) $g(x, \theta)$ is Lipschitz in x and θ: $\|g(x_1, \theta) - g(x_2, \theta)\|_2 \le C_2\|x_1 - x_2\|_2$ for all $\theta \in \Theta$ and $\|g(x, \theta_1) - g(x, \theta_2)\|_2 \le C_2\|\theta_1 - \theta_2\|_2$ for all $x \in \mathcal{X}$, where C_2 is a constant;

(A3) $m(x, \theta)$ is convex in θ for all $x \in \mathcal{X}$ and $M_{\mathcal{X}_D}(\theta)$ is twice continuously differentiable with $M''_{\mathcal{X}_D}(\theta^*(\mathcal{X}_D)) := \int_{\mathcal{X}_D} f_D(x) \frac{\partial}{\partial \theta} g(x, \theta^*(\mathcal{X}_D)) dx$ positive definite where $M_{\mathcal{X}_D}(\theta)$ and $\theta^*(\mathcal{X}_D)$ are defined in (1).

To quantify the bias from using weighed density from hybrid datasets to estimate the private density, we take the measurement of the distance between two datasets D and $E' \subseteq E$.

$$d(D, E') = \frac{1}{n_D} \sum_{x \in D} \min_{y \in E'} \|x - y\|_2 . \quad (5)$$

In (5), $d(D, E')$ takes the average of the distance between a point x in D and E' (i.e., $\min_{y \in E'} \|x-y\|_2$) over all the points in D. Without considering DP, obviously $d(D, E') \geq d(D, E)$ for $E' \subseteq E$. Hence, increasing the sample size of the released dataset decreases the distance to the private dataset D.

Lemma 1—*Let* $g(x, \theta) = \frac{\partial}{\partial \theta} m(x, \theta)$ *Under the notation in Section 2 and condition (A2), we have*

$$\left\| \sum_{i=1}^{n_E} w_i g(y_i, \theta) - \sum_{j=1}^{n_D} \frac{1}{n_D} g(x_j, \theta) \right\|_2 = O(d(D, E)), \quad (6)$$

where $d(D, E)$ *is defined in* (5).

Proof: To prove Lemma 1, recall that $w_i = |D_{y_i}|/n_D$ in Definition 5 where $D_{y_i} = \{x \in D: \|x - y_i\|_2 \leq \|x-y_j\|_2, i \neq j\}$. Since the points in E are distinct, for each $x_j \in D$, there exists a unique point $y_{k_j} \in E$ such that $y_{k_j} = \arg\min_{y \in E} \|x_j - y\|_2$. Based on the definition of D_{y_i}, $|D_{y_i}| g(y^j, \theta) = \sum_{\{j:x_j \in D_{y_i}\}} g(y_{k_j}, \theta)$. Since $g(x, \theta)$ is Lipschitz in x for all θ by condition ($A2$), we have

$$\begin{aligned} \text{(LHS) in (6)} &= \left\| \sum_{j=1}^{n_D} \frac{1}{n_D} g(y_{k_j}, \theta) - \sum_{j=1}^{n_D} \frac{1}{n_D} g(x_j, \theta) \right\| \\ &\leq C_2 \cdot \frac{1}{n_D} \sum_{j=1}^{n_D} \|y_{k_j} - x_j\|_2 = C_2 \cdot d(D, E), \end{aligned}$$

which completes the proof.

Proposition 3—*Under the notation in Section 2 and the contrast function m satisfing conditions (A1)–(A3), there exists a local minimizer* $\hat{\theta}(E, \vec{w})$ *for* $M_{(E, \vec{w})}(\theta)$ *such that*

$$\|\theta(E,\tilde{w}) - \theta^*(\mathscr{X}_D)\|_2 = O_p\left(\frac{n_E \log(n_E)}{n_D} \vee d(D,E) \vee \frac{1}{\sqrt{n_D}}\right), \text{ as } n_E, n_D \to \infty,$$

where $M_{(E,\tilde{w})}(\theta)$ *is in Definition 5, the true parameter* $\theta^*(\mathscr{X}_D)$ *for* $M_{\mathscr{X}_D}(\theta)$ *in* (1) *and* $d(D, E)$ *is defined in* (5)*. The same result holds for* $\hat{\theta}(E, \tilde{w}^+)$ *in Definition 6.*

The proof for Proposition 3 is a straightforward extension of Theorem 3 in [2], hence we only sketch the key ideas here. To get the convergence rate of $\hat{\theta}(E, \tilde{w})$, applying Lemma 9 in [2], it suffices to show that $\sup_{\theta^*(\mathscr{X}_D) \in \Theta_0} \|M'_{(E,\tilde{w})}(\theta) - M'_{\mathscr{X}_D}(\theta)\|_2$ where Θ_0 is a compact neighbor set around $\theta^*(\mathscr{X}_D)$, is bounded in the same order as $\|\hat{\theta}(E, \tilde{w}) - \theta^*(\mathscr{X}_D)\|_2$. Recall $\tilde{w}_i = w_i + Z_i/n_D$ where Z_i is the Laplace noise and $w_i = |D_{y_i}|/n_D$ in Definition 6. In our case, we get the bias-variance decomposition in

$$\begin{aligned} M'_{(E,\tilde{w})}(\theta) - M'_{\mathscr{X}_D}(\theta) &= \sum_{i=1}^{n_E} \tilde{w}_i g(y_i,\theta) - \int_{\mathscr{X}_D} g(x,\theta) f_D(x)\mathrm{d}x \\ &= \sum_{i=1}^{n_E} \frac{Z_i}{n_D} g(y_i,\theta) + \left(\sum_{i=1}^{n_E} \frac{|D_{y_i}|}{n_D} g(y_i,\theta) - \sum_{i=1}^{n_D} \frac{g(x_i,\theta)}{n_D}\right) + \left(\sum_{i=1}^{n_D} \frac{g(x_i,\theta)}{n_D} - \mathbb{E}_{f_D} g(X,\theta)\right). \end{aligned} \tag{7}$$

The first term in (7) comes from the variation error by adding Laplace noise; the second term in the parenthesis comes from the bias by using the weighted density from hybrid datasets; the third term in the parenthesis comes from the sampling error of inadequate sample size. The convergence rates in the first and third terms are a direct extension from Theorem 3 in [2] and the error bound for the second term is from Lemma 1. The convergence rate of $\hat{\theta}(E, \tilde{w}^+)$, by Definition 6 follows in the same fashion as Theorem 5 in [2].

From Proposition 3, we can see that there is a clear bias-variance tradeoff in the estimation accuracy and it can be characterized by the sample size of the release dataset. The term $(n_E \log(n_E)/n_D)$ in the convergence rate of $\hat{\theta}(E, \tilde{w})$ is increasing in n_E while $d(D, E)$ is decreasing in n_E. This motivates us to find an optimal subset of public dataset to better estimate the private dataset especially in practical application.

4.2 Selection Algorithm

To implement the procedure of selecting optimal public subset for M-estimators under DP, we summarize the steps in Algorithm 2. We add Laplace noise in all the steps that involve private dataset D to make sure the whole process guarantees privacy. Algorithm 2 can be implemented for the procedure of selecting the optimal subset from the hybrid density with nonnegative weights by replacing $\tilde{w}$ by $\tilde{w}^+$ in Definition 6.

To find the optimal public subset under DP, Algorithm 2 contains three parts. Part I: to get a sequence of subset candidates $\{\tilde{E}_{(i)}\}_{i=1}^{k}$. After rescaling the datasets to $\tilde{D}$ and $\tilde{E}$ in step 1, we get the weights for all the points in $\tilde{E}$ then add Laplace noise to those weights in step 2. We order all the points in $\tilde{E}$ by their DP weights in step 3. We consider if a data point in $\tilde{E}$ gains more weight, this point is more representative to the private points and thus should have high priority to be selected. Assigning a sequence of the sample sizes $\{n_i\}_{i=1}^{k}$, $\{\tilde{E}_{(i)}\}_{i=1}^{k}$ is a sequence of increasing sets of the ordered points, i.e., $\tilde{E}_{(1)}$ contains the points in the first n_1 largest DP weights and $\tilde{E}_{(2)}$ contains the points in the first n_2 largest DP weights and so on so forth. Regarding to choose $\{n_i\}_{i=1}^{k}$, we take into the consideration of the trend of $d(\tilde{D}, \tilde{E})$. From Proposition 3, the upper bound of the estimation error depends on $d(\tilde{D}, \tilde{E})$, and $d(\tilde{D}, \tilde{E}_{(i)})$ is increasing in the size of $\tilde{E}_{(i)}$. Figure 4 shows that $d(\tilde{D}, \tilde{E})$ decreases first then stays flat in both a low dimension case $p = 2$ and high dimension case $p = 100$ in the simulations of multivariate normal (MVN) variables. From our analysis, the estimation error from adding Laplace noise is increasing in n_i. Hence, we would expect that the turning point for the estimation error can not happen where $d(\tilde{D}, \tilde{E}_{(i)})$ is flat. Therefore, the trend of $d(\tilde{D}, \tilde{E}_{(i)})$ gives us a sense on the general range of the size of the optimal public subset.

Part II: to get DP M-estimators from candidate subsets $\{\tilde{E}_{(i)}\}_{i=1}^{k}$. In step 4, we apply Algorithm 1 to get M-estimators under DP based on each data pair $(\tilde{E}_{(i)}, \tilde{D})$. Applying the sequential composition rule in Section 2.2 (since we add Laplace noise independently), we equally arrange the privacy parameter ε_2 to the weights of the weighted density under each data pair.

Part III: to release the estimation error from each candidate subsets $\{\tilde{E}_{(i)}\}_{i=1}^{k}$ under DP and pick the optimal subset. In step 4, we evaluate the performance of $\hat{\theta}(\tilde{E}_{(i)}, \tilde{w})$ by comparing to the baseline $\hat{\theta}(D)$. One can take criterion as $\|\hat{\theta}(\tilde{E}_{(i)}, \tilde{w}) - \hat{\theta}(D)\|_2$. But in order to release the criterions, we need to perturb them under DP in step 5. Considering the case that the sensitivity of the criterion is large, then selecting the optimal subset based on the noisy criterions could be useless. From the prospective of privacy, we consider a transformation function h on θ with small sensitivity to guarantee a certain utility. We require the function h is Lipschitz then $\|h(\hat{\theta}(\tilde{E}_{(i)}, \tilde{w})) - h(\hat{\theta}(D))\|_2$ shares the same convergence rate as $\|\hat{\theta}(\tilde{E}_{(i)}, \tilde{w}) - \hat{\theta}(D)\|_2$ and thus the bias-variation tradeoff could still exist. In step 6, the subset with the smallest criterion is the optimal set to release.

Algorithm 2

To obtain optimal subset of public dataset for M-estimators under DP.

Input: private dataset D, public dataset E with distinct points, privacy parameters ε_1, ε_2, ε_3, objective function M and transformation function h.
Output: Optimal subset of E with associated weights under DP.

1 To rescale D and E to $\tilde{D}$ and $\tilde{E}$ according to Step 1 in Algorithm 1.

2 To get noisy weights $\tilde{w}_i = \tilde{w}_i(\tilde{E}, \tilde{D})$ for the points in $\tilde{E}$ according to Step 2–3 in Algorithm 1 under privacy parameter ε_1.

3 To sort the points in $\tilde{E}$ in the order of $\tilde{w}(\tilde{E}, \tilde{D})$ obtained in Step 2 in decreasing order.

4 Consider a sequence of increasing sets $\{\tilde{E}_{(i)}\}_{i=1}^{k}$ of ordered points in $\tilde{E}$. To apply Algorithm 1 to each data pair $(\tilde{E}_{(i)}, \tilde{D})$ under privacy parameter ε_2/k to get its M-estimator $\hat{\theta}(\tilde{E}_{(i)}, \tilde{w})$ where $\tilde{w} = \tilde{w}(\tilde{E}_{(i)}, \tilde{D})$. To evaluate the performance of $\hat{\theta}(\tilde{E}_{(i)}, \tilde{w})$ by the criterion $t_i = \|h(\hat{\theta}(\tilde{E}_{(i)}, \tilde{w})) - h(\hat{\theta}(D))\|_2$, $i = 1, \ldots, k$.

5 To add independent Laplace noise to the performance criterion t_i's obtained in Step 4. Releasing $\tilde{t}_i = t_i + Z_i$, $i = 1, \ldots, k$, where $Z_i \overset{\text{iid}}{\sim} Laplace(S(t_i)/\varepsilon_3)$ and $S(t_i)$ is the sensitivity of t_i.

6 To select the optimal subset $\tilde{E}_{(i^*)}$ where $i^* = \arg\min_{i=1,\ldots,k} \tilde{t}_i$ and release $\tilde{E}_{(i^*)}$ and its noisy weights.

5 Simulation and Application

In this section, we take logistic regression as an example to implement Algorithm 1–2. Both simulation studies and application on the real datasets are coherent to our finding that bias-variance tradeoff phenomenon can be characterized in the sample size of the released dataset. Our example sets a guideline to select optimal subset under DP in practice.

5.1 Simulation Studies

We simulate the private dataset D and public dataset E from logistic regression model. We set $n_D = 10^4$, $n_E = 10^4$ and consider the covariates in a low dimensional space with $p = 2$ and a high dimensional space with $p = 100$. In the low dimensional case, to get D, we first generate independent covariates (X_{i1}^D, X_{i2}^D) from $\mathcal{N}((0, 0)^\top, diag(1, 1))$, $i = 1, \ldots, n_D$. Given (X_{i1}^D, X_{i2}^D), we generate $Y_i^D \mid (X_{i1}^D, X_{i2}^D) \sim Bernoulli(p_i^D)$ and model $logit(p_i^D) = \beta_1^D X_{i1}^D + \beta_2^D X_{i2}^D$ setting $\beta_1^D, \beta_2^D \overset{\text{iid}}{\sim} \text{Unif}(-1, 1)$ where $logit(p) = \log\left(\frac{p}{1-p}\right)$. Then D is a set of n_D $(Y_i^D, X_{i1}^D, X_{i2}^D)$'s. Similarly, to get E, we first generate independent covariates (X_{i1}^E, X_{i2}^E) from $\mathcal{N}((0, 0)^\top, diag(0.5, 1))$, $i = 1, \ldots, n_E$ then generate $Y_i^E \mid (X_{i1}^E, X_{i2}^E) \sim Bernoulli(p_i^E)$. Model $logit(p_i^E) = \beta_1^E X_{i1}^E + \beta_2^E X_{i2}^E$ setting $\beta_1^E, \beta_2^E \overset{\text{iid}}{\sim} \text{Unif}(0, 1/2)$. Then E is a set of n_E $(Y_i^E, X_{i1}^E, X_{i2}^E)$'s. In the high dimensional case, for D, we consider $(X_{i1}^D, \ldots, X_{i100}^D)$ from $\mathcal{N}((0, \ldots, 0)_{100}^\top, I_{100 \times 100})$ and $\beta_i^D \overset{\text{iid}}{\sim} \text{Unif}(-1, 1)$, $i = 1, \ldots, n_D$. For E, we consider $(X_{i1}^E, \ldots, X_{i100}^E)$ from $\mathcal{N}((0, \ldots, 0)_{100}^\top, 0.5 I_{100 \times 100})$ and $\beta_i^E \overset{\text{iid}}{\sim} \text{Unif}(0, 1/2)$, $i = 1, \ldots, n_E$.

Based on the simulated datasets, we investigate how the sample size of released public dataset effects the performance of logistic regression under DP. In the simulation, we set $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.8\varepsilon, 0.2\varepsilon, \varepsilon)$ where $\varepsilon = 0.5, 1, 3$, so the total privacy parameter is 2ε by the sequential composition of DP. Here we put more weight to ε_1 than ε_2 from our practical experience. How to optimize the allocation of the total privacy budget can be further investigated. To Implement Algorithm 2, for $p = 10$, we start our search of optimal subset candidates $\{\tilde{E}_{(i)}\}_{i=1}^{k}$ from the size 10 up to 3000 with increment 50, and for $p = 100$, start from the size 100 up to 10000 with increment 200. In the simulations we choose a large

range to get the optimal subset but with the information of the decreasing range of $d(\tilde{D}, \tilde{E})$ in Figure 4, we can further narrow down our search to save more privacy budget. For each data pair $(\tilde{E}_{(i)}, \tilde{D})$, we get DP M-estimators from hybrid datasets with nonnegative weights $\hat{\beta}(\tilde{E}_{(i)}, \vec{w}^{+})$.

To evaluate the performance on the sequence of subset candidates, we consider the total prediction error defined by $t_i = \|(\delta h)\|_2$ for candidate $\tilde{E}_{(i)}$, where $(\delta h) = h_D(\hat{\beta}(\tilde{E}_{(i)}, \vec{w}^{+})) - h_D(\hat{\beta}(D))$. Instead of directly measuring difference in $\hat{\beta}$ from different datasets, we take the transformation function $h_D(\hat{\beta}) = \frac{e^{X^D \hat{\beta}}}{1 + e^{X^D \hat{\beta}}}$ where X^D is the covariate matrix of D and thus $h_D(\hat{\beta})$ is a vector of n_D elements. The prediction error of a candidate public subset measures the sum of total differences between the predicted probabilities with $\hat{\beta}$ estimated from the candidate under DP on the private covariates for dataset D and those with $\hat{\beta}$ estimated from D and its private covariates across all the observations. As we take $\hat{\beta}(D)$ as the baseline, $\hat{\beta}(D)$ itself contains private information. To compare the difference to $\hat{\beta}(D)$, we need to add additional perturbation when report this difference. Taking a transformation on measuring error $\|\hat{\beta}(\tilde{E}_{(i)}, \vec{w}^{+}) - \hat{\beta}(D)\|$., we need the transformation function maintains its tradeoff property and has a small sensitivity. It is easy to check that h is Lipschitz in this case that X^D is bounded (note X^D is not random when adding Laplace noise). Hence (δh) captures the tradeoff. Define (δh) and $(\delta h)'$ as neighbor vectors only differing in j_0-th component and denote their common components as $(\delta h)_{-j0}$. Noticing that $(\delta h)_{j0}, (\delta h)'_{j_0} \in [0, 1]$, we have

$$|t_i((\delta h)) - t_i((\delta h)')| = \left| \sqrt{\|(\delta h)_{-j_0}\|_2^2 + (\delta h)_{j_0}^2} - \sqrt{\|(\delta h)_{-j_0}\|_2^2 + (\delta h)_{j_0}'^2} \right| \leq |(\delta h)_{j_0} - (\delta h)'_{j_0}| \leq 1,$$

where for the first inequality, the equality achieves when $\|(\delta h)_{-j0}\|_2 = 0$. Hence the sensitivity of t_i is 1 and we perturb t_i by adding noise $Laplace(1/\varepsilon_3)$. From the privacy-protected releasing curve of t_i's, we take the optimal subset as the one with the smallest prediction error.

In the simulation study for logistic regression, we compare the performance of the estimators for β: (1) $\hat{\beta}(E)$ naïvely from public dataset, (2) $\hat{\beta}(\tilde{E}_{(i)}, w)$ from weighted density of data pair $(\tilde{E}_{(i)}, \tilde{D})$ without adding noise, and (3) $\hat{\beta}(\tilde{E}_{(i)}, \vec{w}^{+})$ from weighted density of data pair $(\tilde{E}_{(i)}, \tilde{D})$ under DP, and (4) under the low dimensional case, $\hat{\beta}$(perturbed D) from [2]. To make it fair to the perturbed histogram method, we estimate β under privacy parameter $\varepsilon_1 + \varepsilon_2$. We report their average performance in terms of prediction error t from 10 repeats on the whole procedure. We can see from Figure 5 and Figure 6 that there is a clear tradeoff in the performance of $\hat{\beta}(\tilde{E}_{(i)}, \vec{w}^{+})$ in the red curve under DP, i.e. the prediction error decreases first then increases with the released sample size, when the privacy parameter ε is not too small. Increasing ε, the performance of $\hat{\beta}(\tilde{E}_{(i)}, \vec{w}^{+})$ approaches to that of $\hat{\beta}(\tilde{E}_{(i)}, w)$ as expected. The blue line is the performance of $\hat{\beta}(E)$. Since it does not change with the subset

of E, it is a straight line. We can see due to the bias of public dataset, directly estimating from public dataset to represent private dataset has large prediction error compared to our selecting subset procedure. Figure 5 and Figure 6 indicate that under a small privacy parameter, releasing about 10% public points could give a better performance under DP. In the case of $p = 2$, Figure 5 shows the prediction error from the optimal subset of the public dataset is smaller than that from the method in [2] only perturbing the private dataset (in pink line). Under the high dimensional case, our DP procedure of selecting the optimal subset to release shows its priority. When $p = 100$, adding noise to say 10^p cubes where to partition 10 bins in each dimension would make the perturbed histogram useless but the tradeoff curve could still exist which makes our selection doable.

5.2 Applying to Real Datasets

We have two clinical datasets from different institutes in the diagnosis of acute myocardial infarction from [27]. One is from Edinburgh institute with 1253 patients and the other is from Sheffield institute with 500 patients. We are interested in selecting an optimal public dataset in a logistic regression model. The response in logistic regression is 0–1 disease variable and the covariates in this study are Pain in left arm, Pain in right arm, Nausea, Hypo perfusion, ST elevation, New Q waves, ST depression, T wave inversion and Sweeting, which are all 0–1 categorical variables and all measured in both institutes.

Since the data is collected from different sites, there is an intrinsic bias if using one dataset to approximate the other. In our study, we mimic the case that one dataset is public while the other is private. We first take Sheffield dataset as public and Edinburgh as private and then converse their roles. Note that since the datasets are of discrete variables, it is necessary to remove the replicates in the public dataset when getting the weights for the weighted density to avoid ambiguity of assigning the weights. The distinct data points in Sheffield institute are 153 and for Edinburgh are 181. We compare the performance of three M-estimators (from the public dataset, from the hybrid dataset without adding noise to the weights and from the hybrid datasets under DP) in logistic regression for the real datasets, in the same manner as in the previous section of simulation studies.

In Figure 7, it demonstrates the case that Sheffield dataset is public while Edinburgh is private, and in Figure 8, it demonstrates the case that Edinburgh dataset is public while Sheffield is private. In both cases, applying Algorithm 1–2, we consider $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$ and set the total of privacy parameter is 3ε where $\varepsilon = (1, 3, 10)$. We add 10 points each time to the sequence sorted subsets of the public dataset. To mimic the real application, we report the prediction error defined in Algorithm 2) from one repeat.

From both Figure 7 and Figure 8, we can see the tradeoff phenomenon in the performance of the DP M-estimator from hybrid datasets (the red curve) is still clear, as we discussed in the previous section. To select the optimal subset under DP, the performance of the M-estimator from hybrid datasets without adding noise (the black curve) can not be released since it has private information. However, the turning point in the red curve suggests an optimal sample size to release. In reality to protect the privacy, we can only see the red curve and the blue line for the performance of M-estimator from the public dataset. Comparing those performance gives us a guideline to select the optimal subset under DP.

6 Discussion and Conclusion

In this paper, we discuss a privacy problem in a situation where both the private and public datasets are present. How to effectively take use of the public information to better understand the private information without releasing privacy is a prime challenge and especially important in the healthcare data analysis. In our work, we formulate this big question to how to effectively select public data points under the framework of differential privacy in the analysis of M-estimator. We first proposed our DP weighted density estimation based on the hybrid datasets which make it possible to do non interactive learning. Along the same line as [2], we gave the convergence rate of our DP M-estimator from hybrid datasets. Based on that, we found that the bias-variance tradeoff can be characterized in the sample size of the released dataset. This inspired us to explore an algorithm to implement the selection procedure. The framework of our Algorithm 1–2 gives a guideline for answering more general questions. As we previously pointed, this paper focused on a small question but can be generalized to more variants in both statistical interests and privacy concerns.

To integrate public and private information, in Section 3, we used the public points to estimate the private density weighing them by how they are representative to the private points in (3). In Section 4, we defined a distance $d(D, E)$ in (5) to measure similarity of these two datasets and we showed the accuracy of the DP M-estimator highly depends on $d(D, E)$. However, in reality, how the public dataset at hand is representative to the private dataset is a data-dependent question. As we saw in our investigation in the effect of rescaling in Subsection 3.2, when these two sets are not extremely different, our proposed procedure still could provide a DP M-estimation in high utility. If we consider a case that only males are in the public dataset while both males and females in the private dataset, then we can never infer the coefficient for the female in the private dataset from the public dataset. In this case under the concern of privacy, either we do not consider the gender effect or try to find another more related public dataset. Our work intends to give a general framework to select the optimal public subset to provide another option besides directly learning from the public dataset and only from the private dataset itself, as we explained in the workflow in Figure 1.

From a statistical point of view, to estimate the private density from public data points, we took the weighted empirical probabilities in (3). As we mentioned in the Section 3, one can further develop other density estimation methods in the context of DP to get a better convergence rate.

From the privacy concerns, more work can be done in effectively distribute the privacy parameters ($\varepsilon_1, \varepsilon_2, \varepsilon_3$) in Algorithm 2. Since the privacy parameter controls of the utility of the randomized mechanism, the question in how to optimize the distribution of the privacy parameter should be further investigated. For example, instead of releasing the entire performance curve (the red curve in our simulation studies) to pick the optimal sample size, one can apply other DP mechanism such as exponential mechanism to improve the utility.

In conclusion, we proposed Algorithm 1–2 as a strategy to select an optimal public subset to get DP M-estimator in high utility, which serves as a guideline to real world applications.

Acknowledgments

We would like to thank the associate editor and the reviewers. MW would like to thank Ery Arias-Castro and Anthony Gamst for helpful discussion and the training from UCSD. XJ was supported in part by the Patient-Centered Outcomes Research Institute (PCORI) under contract ME-1310-07058, the National Institute of Health (NIH) under award number R01GM118574 and U01EB023685.

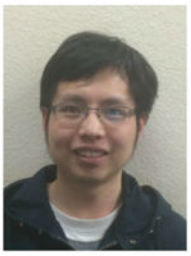
References

1. Dwork C, , McSherry F, , Nissim K, , Smith A. Theory of cryptography Springer; 2006 Calibrating noise to sensitivity in private data analysis; 265284
2. Lei J. Differentially private m-estimators. Advances in Neural Information Processing Systems. 2011:361–369.
3. Cox LH. Protecting confidentiality in small population health and environmental statistics. Statistics in medicine. 1996; 15(17):1895–1905. [PubMed: 8888482]
4. Fienberg SE. Statistical perspectives on confidentiality and data access in public health. Statistics in medicine. 2001; 20(9–10):1347–1356. [PubMed: 11343356]
5. Paiva T, Chakraborty A, Reiter J, Gelfand A. Imputation of confidential data sets with spatial locations using disease mapping models. Statistics in medicine. 2014; 33(11):1928–1945. [PubMed: 24395116]
6. O'Keefe CM, Rubin DB. Individual privacy versus public good: protecting confidentiality in health research. Statistics in medicine. 2015; 34(23):3081–3103. [PubMed: 26045214]
7. Overhage JM, Overhage LM. Sensible use of observational clinical data. Statistical methods in medical research. 2013; 22(1):7–13. [PubMed: 21828172]
8. Nelson JC, Cook AJ, Yu O, Zhao S, Jackson LA, Psaty BM. Methods for observational post-licensure medical product safety surveillance. Statistical methods in medical research. 2015; 24(2): 177–193. [PubMed: 22138688]
9. Rubin DB. Statistical disclosure limitation. Journal of official Statistics. 1993; 9(2):461–468.
10. Reiter JP. Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics. 2002; 18(4):531.
11. Abowd JM, , Vilhuber L. International Conference on Privacy in Statistical Databases Springer; 2008 How protective are synthetic data?; 239246
12. Loong B, Zaslavsky AM, He Y, Harrington DP. Disclosure control using partially synthetic data for large-scale health surveys, with applications to cancors. Statistics in medicine. 2013; 32(24):4139–4161. [PubMed: 23670983]
13. Friedman A, , Schuster A. Data mining with differential privacy. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining; ACM; 2010 493502
14. Dwork C, , Lei J. Differential privacy and robust statistics. Proceedings of the forty-first annual ACM symposium on Theory of computing; ACM; 2009 371380
15. Dwork C, , Feldman V, , Hardt M, , Pitassi T, , Reingold O, , Roth AL. Preserving statistical validity in adaptive data analysis. Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of; ACM; 2015 117126
16. Mohammed N, Jiang X, Chen R, Fung BC, Ohno-Machado L. Privacy-preserving heterogeneous health data sharing. Journal of the American Medical Informatics Association. 2013; 20(3):462–469. [PubMed: 23242630]
17. Zhao Y, Wang X, Jiang X, Ohno-Machado L, Tang H. Choosing blindly but wisely: differentially private solicitation of dna datasets for disease marker discovery. Journal of the American Medical Informatics Association. 2014 pp. amiajnl–2014.
18. Dankar FK, , El Emam K. The application of differential privacy to health data. Proceedings of the 2012 Joint EDBT/ICDT Workshops; ACM; 2012 158166
19. Mohammed N, , Chen R, , Fung B, , Yu PS. Differentially private data release for data mining. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining; ACM; 2011 493501

20. Machanavajjhala A, , Kifer D, , Abowd J, , Gehrke J, , Vilhuber L. Privacy: Theory meets practice on the map. Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on; IEEE; 2008 277286
21. Hall R, Rinaldo A, Wasserman L. Differential privacy for functions and functional data. Journal of Machine Learning Research. Feb.2013 14:703–727.
22. Ji Z, Elkan C. Differential privacy based on importance weighting. Machine learning. 2013; 93(1): 163–183. [PubMed: 24482559]
23. Ji Z, Jiang X, Wang S, Xiong L, Ohno-Machado L. Differentially private distributed logistic regression using private and public data. BMC medical genomics. 2014; 7(Suppl 1):S14. [PubMed: 25079786]
24. Li H, Xiong L, Ohno-Machado L, Jiang X. Privacy preserving rbf kernel support vector machine. BioMed research international. 2014; 2014
25. Van der Vaart AW. Asymptotic statistics Vol. 3. Cambridge university press; 2000
26. Sugiyama M, , Suzuki T, , Kanamori T. Density ratio estimation in machine learning Cambridge University Press; 2012
27. Kennedy R, Fraser H, McStay L, Harrison R. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. European heart journal. 1996; 17(8):1181–1191. [PubMed: 8869859]

Biographies

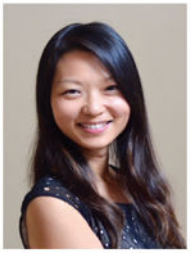
MengWang is a postdoctoral fellow in the department of genetics at Stanford University from February, 2016 to now. She received the B.S. degree in mathematical statistics from Nankai University, China in 2010, and the Ph.D. degree in mathematics specialized in statistics from University of California at San Diego (UCSD) in 2014. She was a postdoctoral fellow in the department of biomedical informatics at UCSD, 2014–2015. Her research interests include statistics, high-dimensional data, statistical application in healthcare data privacy and biology.



Zhanglong Ji is a PhD student in Department of Computer Science at the University of California, San Diego. Before coming to UCSD, he majored in Statistics in Peking University. His research focuses on differential privacy and machine learning.

Hyeoneui Kim received the B.S. degree in nursing (BSN) and the Masters degree in Public Health (MPH) from Seoul National University, South Korea. She earned PhD in Health Informatics from the University of Minnesota at Twin Cities. She received post-doctoral training from Brigham and Women's Hospital in Boston. Dr. Kim is currently an associate professor of Division of Biomedical Informatics at UC San Diego. Her research focuses on standardized data representation, secondary use of healthcare data, clinical informatics and consumer health informatics.

Shuang Wang (S'08–M'12) received the B.S. degree in applied physics and the M.S. degree in biomedical engineering from the Dalian University of Technology, China, and the Ph.D. degree in electrical and computer engineering from the University of Oklahoma, OK, USA, in 2012. He was worked as a postdoc researcher with the Department of Biomedical Informatics (DBMI), University of California, San Diego (UCSD), CA, USA, 2012 – 2015. Currently, he is an assistant professor at the DBMI, UCSD. His research interests include machine learning, and healthcare data privacy/security. He has published more than 60 journal/conference papers and 2 book chapters. He was awarded a NGHRI K99/R00 career grant. Dr. Wang is a senior member of IEEE.



Li Xiong Li Xiong is Professor of Computer Science (and Biomedical Informatics) and holds a Winship Distinguished Research Professorship at Emory University. She has a PhD from Georgia Institute of Technology, an MS from Johns Hopkins University, and a BS from University of Science and Technology of China, all in Computer Science. She and her research group, Assured Information Management and Sharing (AIMS), conduct research

that addresses both fundamental and applied questions at the interface of data privacy and security, spatiotemporal data management, and health informatics. She is a recipient of a Google Research Award, IBM Faculty Innovation Award, Cisco Research Award, and Woodrow Wilson Fellowship. Her research is supported by NSF (National Science Foundation), NIH (National Institute of Health), AFOSR (Air Force Office of Scientific Research), and PCORI (Patient-Centered Outcomes Research Institute).

Xiaoqian Jiang is an assistant professor in the Department of Biomedical Informatics at the University of California San Diego. He received his PhD in Computer Science from Carnegie Mellon University. He is an associate editor of BMC Medical Informatics and Decision Making and serves as an editorial board member of Journal of American Medical Informatics Association. He works primarily in health data privacy and predictive models in biomedicine. Dr. Jiang is a recipient of NIH K99/R00 award and he won the distinguished paper award from American Medical Informatics Association Clinical Research Informatics (CRI) Summit in 2012 and 2013.

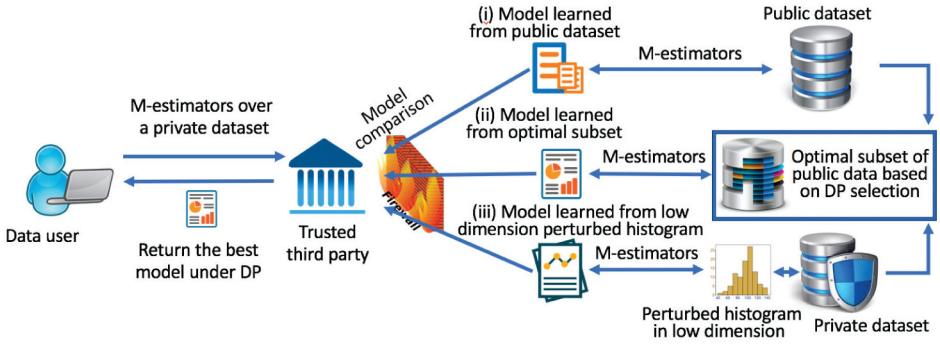


Fig. 1.

Workflow of the proposed framework. (1) A trusted third party (TTP) can access both public and private datasets. (2) A data user would like to infer an M-estimator over a private dataset. (3)The TTP compares the results from (i) directly learning from the public dataset (ii) applying our DP selection procedure to learn from an optimal subset of the public dataset (iii) in low dimension, applying the perturbed histogram method only learning from the private dataset. (4) The TTP reports the user the best model under DP to learn the private information.

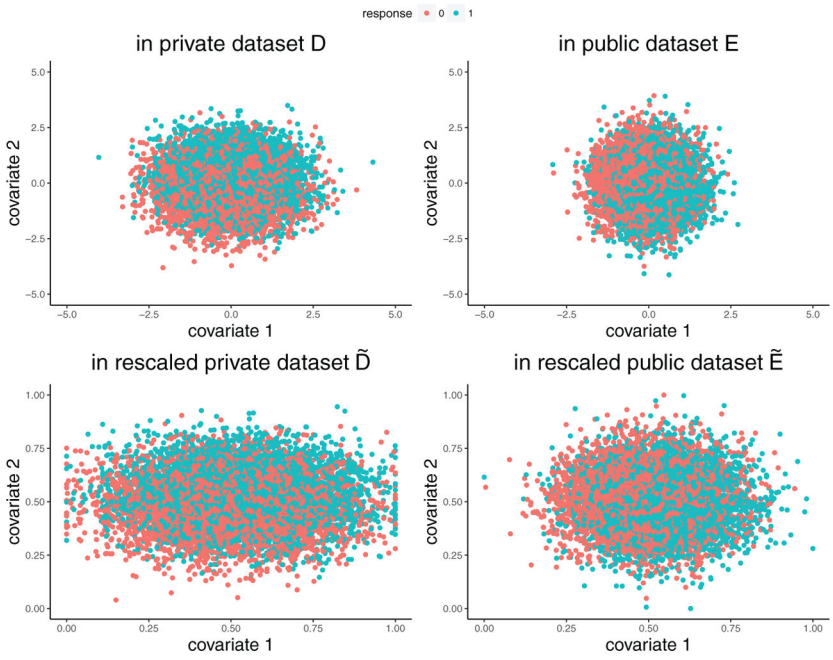


Fig. 2.

Scatter plots of two covariates in the original private dataset D (upper left), original public dataset E (upper right) and rescaled private dates $\tilde{D}$ (lower left), rescaled public dates $\tilde{E}$ (lower right). The dots in blue indicate response 0 and in red indicate response 1.

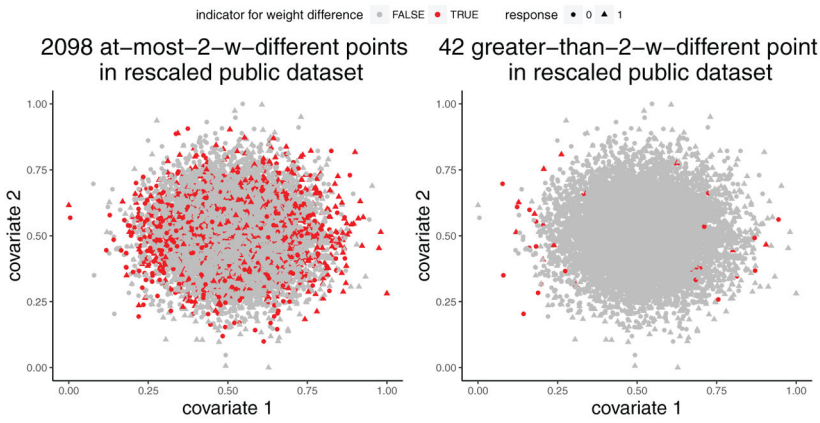


Fig. 3.

Scatter plots of two covariates in the rescaled public dataset $\tilde{E}$. The red dots indicates the points whose weights calculated from ($\tilde{D}$, $\tilde{E}$) have at most 2 counts difference (in the left panel), or greater than 2 counts difference (in the right panel), compared to the corresponding weights calculated from (D, E).

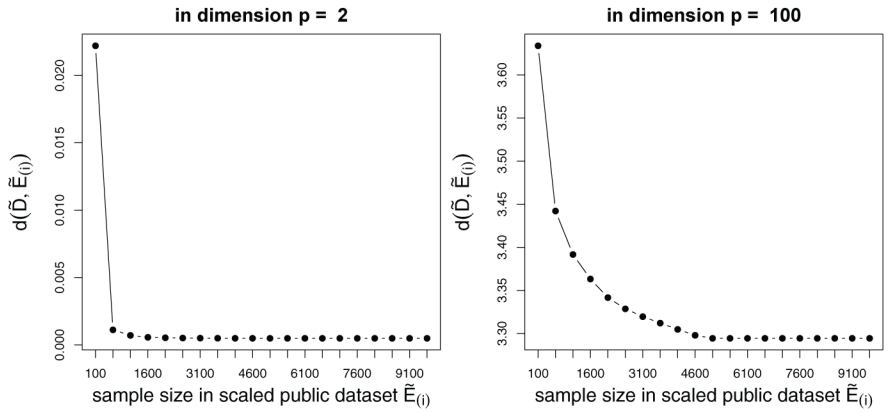


Fig. 4.

$d(\tilde{D}, \tilde{E}_{(i)})$ decreases with the sample size of $\tilde{E}_{(i)}$ under dimension $p = 2, 100$, where $\tilde{E}_{(i)}$ is the scaled subset of $\tilde{E}$ with sorted points defined in Algorithm 2 and the sample sizes of $\{\tilde{E}_{(i)}\}$ start from 100 to n_E with increment 500. In this simulation, D is a set of 10^4 points independently generated from $MVN((0, \ldots, 0)^\top, I_{p\times p})$ and E is a set of 10^4 points independently generated from $MVN((0, \ldots, 0)^\top, 0.5I_{p\times p})$, where $I_{p\times p}$ is the identity matrix.

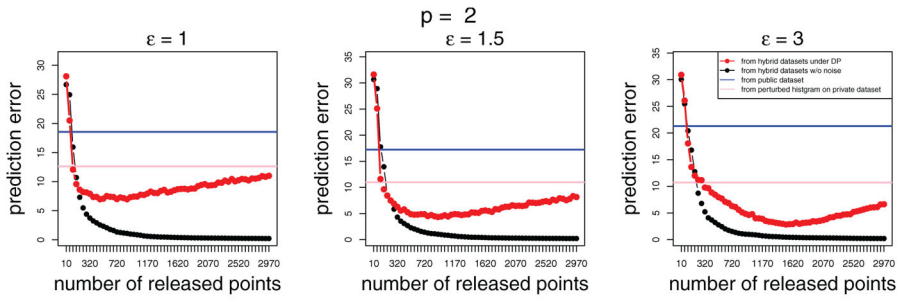


Fig. 5.

Prediction error varies with the released points from the public dataset under privacy parameter $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.8\varepsilon, 0.2\varepsilon, \varepsilon)$ where $\varepsilon = 0.5, 1, 3$ from replicating 10 times on the whole procedure. The black dotted curve is the prediction error from the M-estimator from the hybrid datasets without adding noise. The red dotted curve is from the M-estimator from the hybrid datasets under DP. The blue line is from the public dataset. The pink line is from the perturbed private dataset.

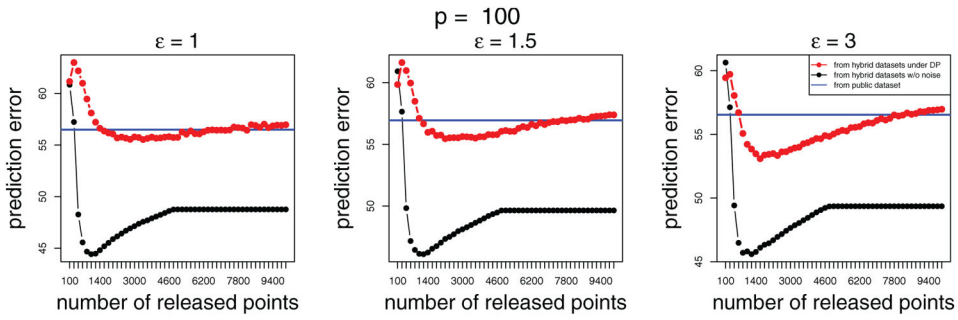


Fig. 6.

Prediction error varies with the released points from the public dataset under privacy parameter $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.8\varepsilon, 0.2\varepsilon, \varepsilon)$ where $\varepsilon = 0.5, 1, 3$ from replicating 10 times on the whole procedure.

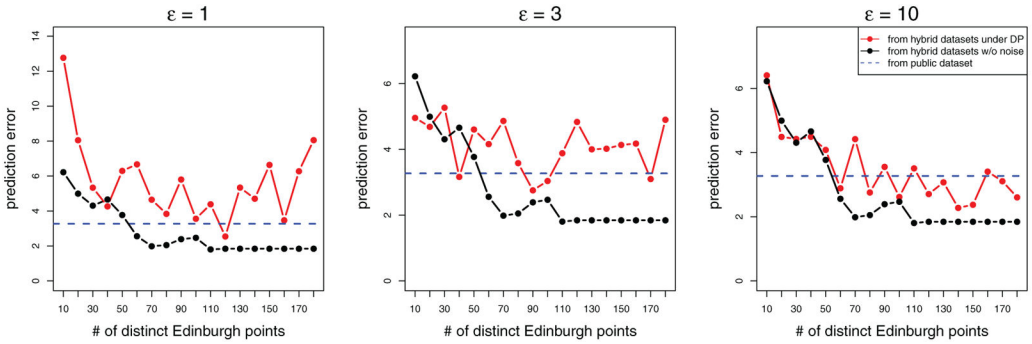


Fig. 7.

Sheffield dataset is public while Edinburgh is private. The sum privacy parameter is 3ε, where $\varepsilon = (1, 3, 10)$. The black dotted curve is the prediction error from the M-estimator based on hybrid datasets without adding noise to the weights. The red dotted curve is from M-estimator based on the hybrid datasets under DP. The blue dashed line is from the one based on public dataset. The study is only under one repeat.

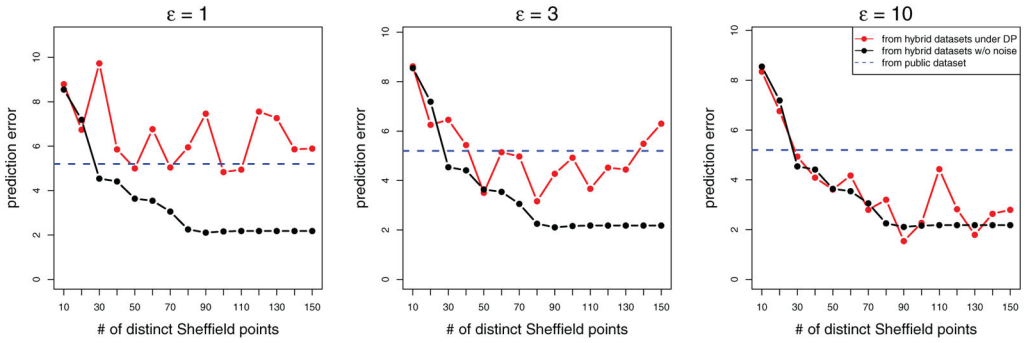


Fig. 8.

Edinburgh dataset is public while Sheffield is private. The sum privacy parameter is 3ε, where $\varepsilon = (1, 3, 10)$.