Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Differentially Private Confidence Interval for Extrema of Parameters

Xiaowen Fu[1] | Yang Xiang[1,2] | Xinzhou Guo*[1]

[1]Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[2]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China

Correspondence
*Xinzhou Guo, Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Email: xinzhoug@ust.hk

Summary

This paper aims to construct a valid and efficient confidence interval for the extrema of parameters under privacy protection. The usual statistical inference on the extrema of parameters often suffers from the selection bias issue, and the problem becomes more acute, as in many application scenarios of extrema parameters, we often need to protect the privacy of the data. In this paper, we focus on the exponential family of distributions and propose a privatized parametric bootstrap method to address selection bias in the extrema of parameters problem under the scheme of differential privacy. While the usual privatized parametric bootstrap does not address selection bias appropriately, we prove that with a privatized bias correction term, the proposed parametric bootstrap method can lead to a valid and efficient confidence interval for the extrema of parameters. We illustrate the proposed method with the Gaussian case and regression case and demonstrate the merits of the proposed method by revisiting the national supported work program.

KEYWORDS:
Bootstrap; Privacy; Selection bias; Statistical inference

1 | INTRODUCTION

Confidence interval (CI) refers to a range of plausible values for estimates of an unknown parameter of the population. Compared with point estimate, CI not only measures the magnitude of the parameter but also quantifies uncertainty of estimation, and serves as one foundation for statistical inference. CI is widely used in different scientific disciplines. For example, Sandercock (2015) shows that reporting CI has become a standard in medical journals since the late 1980s.

Extrema parameter refers to the maximum or minimum of some parameters of population and bears practical implications in many real-world problems. It is widely recognized that constructing a CI for the extrema parameter is challenging due to selection bias, see Thomas and Bornkamp (2017) and Magnusson and Turnbull (2013), and simply using sample analogue of the extrema parameter would not lead to valid statistical inference. How to address selection bias and construct a valid and efficient CI for the extrema parameter is an important problem bearing both methodological and practical importance. Some attempts have been made. Fuentes, Casella, and Wells (2018) and Hall and Miller (2010) propose valid CI based on simultaneous controls so the resulting CIs tend to be conservative and are not efficient. Rosenkranz (2016) and Stallard, Todd, and Whitehead (2008) have considered some ad-hoc methods to adjust selection bias, but those methods lack theoretical justifications. Bornkamp, Ohlssen, Magnusson, and Schmidli (2017) and Woody, Padilla, and Scott (2022) consider Bayesian approach, which is clearly model-dependent and often lacks frequentist interpretation. Recently, several bootstrap-based CI for extrema parameter have been proposed in Guo and He (2021) and Guo et al. (2022) among others.

One important application scenario of extrema parameter is subgroup analysis which aims to uncover and confirm treatment effect heterogeneity within a population. In clinical trials, it is often the case that a treatment is more effective for some patients than others; see for example Mologen (2018). When this happens, the extrema parameter can be used to represent the treatment effect of the best subgroup and a valid and efficient CI for the extrema parameter can help researchers better understand the treatment and know where and to what extent the treatment is most

useful. Despite the potential benefits, analyzing the best subgroup by directly accessing clinical trial data might face several disincentive issues and in particular, the growing concern of privacy leakage given the sensitive nature of health data as discussed in Xiang and Cai (2021).

To protect privacy, several schemes have been proposed. For example, Sweeney (2002) considers forward K-anonymity,Machanavajjhala, Kifer, Gehrke, and Venkitasubramaniam (2007) proposes l-diversity and Li, Li, and Venkatasubramanian (2006) considers t-closeness. One scheme drawing great attentions recently is differential privacy (DP) proposed in Dwork, McSherry, Nissim, and Smith (2006) which aims to protect privacy by making sure the impact of an arbitrary single substitution in the database is small enough so that the adversarial may not be able to speculate the real data set. To achieve DP, as proposed by Dwork et al. (2006), we often need to add some amounts of noise to the estimate or data. This will clearly induce additional randomness in statistical inference. Therefore, the usual CI or bias correction method, assuming the data is public, would not lead to satisfactory results with private data as discussed in Dwork, Roth, et al. (2014). In this paper, we aim to address the uncertainty quantification problem induced by the noise and consider statistical inference on extrema parameter under DP.

Statistical inference under DP has been considered for several scenarios. For example, Dimitrakakis, Nelson, Mitrokotsa, and Rubinstein (2014), Dimitrakakis, Nelson, Zhang, Mitrokotsa, and Rubinstein (2017), Zhang, Rubinstein, and Dimitrakakis (2016) study DP of Bayesian inference, Rogers, Roth, Smith, and Thakkar (2016), Balle, Barthe, Gaboardi, Hsu, and Sato (2020), Gaboardi, Lim, Rogers, and Vadhan (2016) do hypothesis testing under DP, and Rinott, O'Keefe, Shlomo, and Skinner (2018) study DP in frequency tables. There are some works on constructing CI with DP. For example, Karwa and Vadhan (2017) proposes a private algorithm to estimate a range for the data and derives a private CI. Du, Foot, Moniot, Bray, and Groce (2020) proposes private simulation and quantile methods for estimating mean and variance. Wang, Kifer, and Lee (2018) proposes the method to construct CI under differential privacy for empirical risk estimation which can be applied to logistic regression and support vector machines (SVM). However, the existing statistical inference approaches under DP are not directly applicable to extrema parameter due to the issue of selection bias.

To address selection bias, we often adopt calibrated bootstrap approaches as introduced in Guo and He (2021) and Hall and Miller (2010) while the usual bootstrap is invalid for extrema parameter. Under DP, some usual bootstrap methods have been proposed. Covington, He, Honaker, and Kamath (2021) develops bag of little bootstraps (BLB) to privately estimate sampling distribution of parameters, Chadha, Duchi, and Kuditipudi (2021) proposes a private confidence sets with bootstrap, Brawner and Honaker (2018) uses the bootstrap with DP to estimate standard errors "for free", Dunsche, Kutta, and Dette (2022) presents a test for multivariate mean comparisons under pure-DP with bootstrap, and Ferrando, Wang, and Sheldon (2022) proposes a method to construct confidence intervals with differentially private parametric bootstrap. However, similar to their analogues without DP, the usual bootstrap procedures under DP fail to address selection bias issue appropriately and can not deliver valid CI for the extrema parameter. To sum up, de-biased CI for the extrema parameter under DP is still lacking, and we aim to bridge this gap in this paper.

In this paper, we propose a valid and efficient CI for the extrema parameter under the scheme of differential privacy. We focus on the exponential family of distributions and develop a privatized parametric bootstrap approach to address selection bias in the extrema parameter under the scheme of differential privacy. The proposed method is easy-to-implement, efficient, and can be adapted to different practical scenarios. The main contribution of our work can be summarized as follows, (1) propose a privatized parametric bootstrap procedure to address selection bias in the extrema parameter under differential privacy; (2) account for randomness induced by noise term in constructing CI with private data; (3) implement our proposed method with Gaussian case and regression case which have broad applications in practice; and (4) propose strategies to avoid wasting privacy budget on nuisance parameter.

The remainder of the paper is organized as follows. In Section 2, we introduce some preliminaries on the extrema parameter, selection bias and differential privacy. In Section 3, we illustrate our proposed method under exponential family distribution, where we construct a CI for the extrema parameter with differential privacy and selection bias correction. In Section 4, we show the bootstrap theory about the consistency of our method and the protection of privacy. In Section 5, we apply our proposed frame to multivariate Gaussian and linear regression scenarios. In Section 6, we introduce a strategy to save the privacy budget. In Sections 7 and 8, we give the results of the simulation and the real data application respectively, and Section 9 gives concluding remarks.

2 | PRELIMINARIES

2.1 | Extrema Parameter and Selection Bias

Let $\beta = (\beta_1, \ldots, \beta_k) \in \mathbb{R}^k$ denote some parameters of interest in a population and $\hat{\beta}_j$ denote an estimate for β_j for $j = 1, \cdots, k$ from data sets $\mathcal{X}$. W.L.O.G., our goal is to construct a lower confidence limit for the maximum $\beta_{\max} = \max_{j=1,\cdots,k} \beta_j$. Due to selection bias, the sample analogue of $\beta_{\max}$, $\hat{\beta}_{\max} = \max_{j=1,\cdots,k} \hat{\beta}_j$, is biased even when $\hat{\beta}_j$ is consistent to β_j, and CI simply relying on $\hat{\beta}_{\max}$ would not be valid; see Nadarajah and Kotz (2008) for theoretical derivations.

2.2 | Differential Privacy

Differential privacy is a scheme for privacy protection that if changing an individual data in the data set does not cause much change in the outcome, the adversarial may not be able to speculate the real data set as defined in Definition 1.

Definition 1. (Differential privacy, Dwork et al. (2006))
A mechanism $\mathcal{A}$ is said to satisfy ε - differential privacy (ε-DP) if for all pairs $x, x^{'} \in \mathcal{X}$ which differ in only one entry, and for any outcome $O \subseteq \text{range}(\mathcal{A})$, we have

$$|\ln(\frac{Pr(\mathcal{A}(x) \in O)}{Pr(\mathcal{A}(x^{'}) \in O)})| \leq \varepsilon \tag{1}$$

Under DP, ε is a parameter to control privacy leakage and is called privacy budget. A smaller ε indicates better privacy protection at the potential cost of accuracy. Differential privacy has the composition properties as shown in Zhao (2017), which we use later for privacy budget allocation for parameters and privacy guarantee in cross-validation.

Lemma 1. (Sequential composition theorem, Zhao (2017))
Let M_i each provide ε_i-DP, then the sequence of $M_i(X)$ provides ($\sum_i \varepsilon_i$)-DP.

Lemma 2. (Parallel Composition Theorem, Zhao (2017))
Let M_i each provide ε-DP. Let D_i be arbitrary disjoint subsets of the input domain D. The sequence $M_i(X \cap D_i)$ provides ε-DP.

To achieve differential privacy, we often need to add some amount of noise to the data and the amount is often determined by the sensitivity defined in Dwork et al. (2006).

Definition 2. (Sensitivity, Dwork et al. (2006))
The sensitivity of a function Γ is the smallest number $S(\Gamma)$ such that for all $x, x^{'} \in \mathcal{X}$ which differ in a single entry,

$$||\Gamma(x) - \Gamma(x^{'})||_1 \leq S(\Gamma) \tag{2}$$

For a random algorithm Γ, to achieve ε-DP, we often consider the Laplace mechanism. The Laplace mechanism introduces additional randomness to protect privacy, which usually brings damage to accuracy and efficiency in inference and we need to appropriately account for it.

Definition 3. (Laplace mechanism, Dwork et al. (2006))
For all function Γ that maps data sets to $\mathbb{R}^d$, $\Gamma(\mathbf{x}) + w$ is ε-DP, where $w = \{w_k\}_{k=1}^d$ is the added Laplacian noise with entry $w_k \sim \text{Lap}(S(\Gamma)/\varepsilon)$, and Lap denotes a zero-mean Laplacian distribution with scale $S(\Gamma)/\varepsilon$.

3 | DIFFERENTIALLY PRIVATE CONFIDENCE INTERVAL FOR EXTREMA PARAMETER

In this section, we introduce the framework of deferentially private CI for extrema parameter. The framework built on exponential family could be naturally extended to other parametric families of distribution.

3.1 | Exponential Family Distribution

For a sample x_i from the data set $\mathcal{X} = \{x_i\}_{i=1}^n$, a family of distribution is called the exponential family if the density function is

$$p(x_i; \alpha) = h(x_i) e^{\alpha^T T(x_i) - A(\alpha)} \tag{3}$$

where h is a base function, α is the natural parameter, T is the sufficient statistics function, and $A(\alpha)$ is the log-partition function. Exponential family distribution includes many common distributions, such as guassian distribution and binomial distribution.

In practical applications, the parameter of our interest $\beta \in \mathbb{R}^{k_1}$ can be viewed as a function of α; i.e. $\beta = f(\alpha)$. With appropriate nuisance parameter $\gamma \in \mathbb{R}^{k_2}$, there exists a partition of $\alpha = (\alpha_1, \alpha_2)$ and an a 1-1 mapping $\mathbf{f}$: $(\alpha_1, \alpha_2) = \mathbf{f}(\beta, \gamma)$ where α_2 only depends on the nuisance parameter γ and

$$\begin{cases} \alpha_1 = f_1(\beta, \gamma) \\ \alpha_2 = f_2(\gamma). \end{cases} \tag{4}$$

α_1 can be generated by gathering all the terms including β, and α_2 is then easy to determine. With this reparameterization. the exponential family distribution can be rewritten as

$$p(x_i; \alpha_1, \alpha_2) = h(x_i) e^{\alpha_1^T T_1(x_i) + \alpha_2^T T_2(x_i) - A(\alpha_1, \alpha_2)} \tag{5}$$

where $(T_1(x_i), T_2(x_i)) = T(x_i)$. Our goal is to construct private CI for $\beta_{\max} = \max_{j=1,\cdots,k_1} \beta_j$.

Considering data sets $\mathcal{X} = \{x_i\}_{i=1}^n$ and plugging Eq. 4 into Eq. 5, the log-likelihood is

$$\ln p(\mathcal{X}; \beta, \gamma) = \sum \ln h(x_i) + f_1(\beta, \gamma) \sum T_1(x_i) + f_2(\gamma) \sum T_2(x_i) - nA(\beta, \gamma) \tag{6}$$

where $A(\beta, \gamma) = A(f_1^{-1}(\beta, \gamma), f_2^{-1}(\gamma))$ and $\sum$ denotes the simplified symbol for $\sum_{i=1}^n$ throughout the paper. Then, the MLE estimate is a solution for Eq. 7.

$$\begin{aligned} \frac{\partial \ln p(X; \beta, \gamma)}{\partial \beta} &= \frac{\partial f_1(\beta, \gamma)}{\partial \beta} \sum T_1(x_i) - n \frac{\partial A}{\partial \beta} = 0 \\ \frac{\partial \ln p(X; \beta, \gamma)}{\partial \gamma} &= \frac{\partial f_1(\beta, \gamma)}{\partial \gamma} \sum T_1(x_i) + \frac{\partial f_2(\gamma)}{\partial \gamma} \sum T_2(x_i) - n \frac{\partial A}{\partial \gamma} = 0 \end{aligned} \tag{7}$$

We write out the solution in preparation for the partially privatized case discussed in Section 5. The solution can be written as

$$\begin{cases} \hat{\beta} = g_1(\sum T_1(x_i), \sum T_2(x_i)) \\ \hat{\gamma} = g_2(\sum T_1(x_i), \sum T_2(x_i)) \end{cases} \tag{8}$$

where g_1, g_2 are functions of sufficient statistics to stand for the MLE of β, γ.

3.2 | Privatized Parametric Bootstrap CI

To construct differentially private CI for $\beta_{\max}$, we propose a privatized parametric bootstrap algorithm to address selection bias in the extrema parameter. Take the lower confidence limit as an example, the proposed algorithm is summarized in Algorithm 1. There are three key elements in the algorithm: (1) privatized point estimate; (2) privatized parametric bootstrap and (3) privatized bias-correction term.

To achieve differential privacy in point estimation, following Eq. (8), we add Laplacian noises to sufficient statistics in Eq. 8 in Step 2 as follows

$$\begin{cases} \hat{\beta}^{priv} = g_1(\sum T_1(x_i) + w_1, \sum T_2(x_i) + w_2) \\ \hat{\gamma}^{priv} = g_2(\sum T_1(x_i) + w_1, \sum T_2(x_i) + w_2) \end{cases} \tag{9}$$

where $w_j \sim Lap(\triangle_i/\varepsilon_j)(j = 1, 2)$ is the Laplacian noise. $\triangle_j$ is the sensitivity of sufficient statistics $\sum T_j(x)$, and ε_j is the privacy budget allocated to it.

To avoid accessing data repeatedly, parametric bootstrap is adopted here. In specific, following the idea in Ferrando et al. (2022), we generate bootstrap estimate from the estimated model based on $\hat{\beta}^{priv}$ and $\hat{\gamma}^{priv}$ as shown in Step 5. To account for randomness induced in Laplace scheme, we add a Laplace noise and calculate bootstrap estimate $\hat{\beta}_j^{*,priv}$ in Step 6.

It is well known that the usual bootstrap can not address selection bias; see Bornkamp et al. (2017). Following the idea of Guo and He (2021), we consider a modified bootstrap estimate under DP, $\hat{\beta}_{j,modified}^{*,priv}$, to correct selection bias,

$$\hat{\beta}_{j,modified}^{*,priv} = \hat{\beta}_j^{*,priv} + d_j, j = 1, \cdots, k_1, \tag{10}$$

where d_j is the distance of privatized estimate $\hat{\beta}_j^{priv}$ and its extrema based on the original data sets

$$d_j = (1 - n^{r-0.5})(\hat{\beta}_{\max}^{priv} - \hat{\beta}_j^{priv}), j = 1, \cdots, k_1. \tag{11}$$

where n is the size of the total population ,and $r \in (0, 0.5)$ is a tuning parameter, which determines the adjustments to the estimate. With a smaller r, the adjustment gets stronger, and the coverage probability gets larger at the sacrifice of the efficiency of confident limit. With the proposed calibrated private bootstrap estimate, we can address selection bias under DP.

We adopt cross-validation to choose the tuning parameter r, and the detailed algorithm and related theory is included in the Appendix A.

Algorithm 1 Privatized parametric bootstrap inference on extrema problems with bias-correction

Data: $x_1, \cdots, x_n$

Result: Privatized lower confidence limit of $\beta_{\max}$

1: Add Laplacian noise to sufficient statistics and calculate the privatized MLE $\hat{\beta}^{priv}, \hat{\gamma}^{priv}$ based on Eq. 9
2: Estimate privatized extrema $\hat{\beta}^{priv}_{\max} = \max_{j=1,\cdots,k_1}\{\hat{\beta}^{priv}_j\}$
3: Calculate bias-correction term $d_j = (1 - n^{r-0.5})(\hat{\beta}^{priv}_{\max} - \hat{\beta}^{priv}_j), j = 1, \cdots, k_1$
4: **for** $b = 1 : B$ **do**
5: Generate bootstrap sample based on exponential family (5) and parameter transformation (4) $x^*_1, \cdots, x^*_n \sim p(f_1(\hat{\beta}^{priv}, \hat{\gamma}), f_2(\hat{\gamma}^{priv}))$
6: Add Laplacian noise to sufficient statistics and calculate the privatized MLE $\hat{\beta}^{*,priv}$ based on (9)
7: Calculate $T^{*,priv} = \sqrt{n}(\max\{\hat{\beta}^{*,priv}_j + d_j - \hat{\beta}^{priv}_{\max}\})$
8: **end for**
9: Let $c_\alpha = quantile(T^{*,priv}, 1-\alpha)$. The level $1-\alpha$ lower confidence limit is $L^{priv} = \hat{\beta}^{priv}_{\max} - c_\alpha/\sqrt{n}$.

4 | THEORETICAL INVESTIGATIONS

By composition properties, we can show that based on parametric bootstrap, the proposed method can achieve privacy protection under DP as summarized in Theorems 1.

Theorem 1. (Differential privacy for our proposed method (Algorithm 1))
By Lemma 1 and Lemma 2, the level $1-\alpha$ lower confidence limit $L^{priv} = \hat{\beta}^{priv}_{\max} - c_\alpha/\sqrt{n}$ remains $(\epsilon_1 + \epsilon_2)$-DP since we can at most infer the estimated model (i.e. $\hat{\beta}^{priv}$ and $\hat{\gamma}^{priv}$) from the bootstrap sample.

While the usual bootstrap estimate $\hat{\beta}^{*,priv}_{\max}$ fails to address selection bias under DP, Theorem 2 states that with a correction term and if $0 < r < 0.5$, our proposed calibrated bootstrap in Algorithm 1 can deliver valid lower confidence limit for $\beta_{\max}$ and the lower confidence limit is efficient as it achieves the nominal level as n goes to infinite.

Theorem 2. (Consistency of privately modified extrema parameter with bootstrap)
For any tuning parameter $0 < r < 0.5$, we have that $\hat{\beta}^{*,priv}_{\max,modified}$ is consistent:

$$\sup_{x\in R} |P^*(\sqrt{n}(\beta^*_{\max,modified} - \hat{\beta}^{priv}_{\max}) \le x) \quad - P(\sqrt{n}(\hat{\beta}^{priv}_{\max} - \beta_{\max}) \le x)| \to 0 \tag{12}$$

as $n \to \infty$, in probability w.r.t P.

Note that Theorem 2 is for the private estimate where additional randomness is introduced due to the noise term for privacy protection. Therefore, Theorem 2 indeed states our proposed procedure can address selection and account for additional randomness under DP, which is different from and beyond the classical theoretical argument for calibrated bootstrap based on public data; see Guo and He (2021).

5 | APPLICATION

In this section, we apply our proposed framework to two important scenarios: (1) multivariate Gaussian and (2) linear regression. For simplicity, we only discuss the implementation of the privatized point estimation and privatized parametric bootstrap as the detailed algorithm naturally follows from the framework in Algorithm 1.

5.1 | Multivariate Gaussian Case

Consider a k-dimensional multivariate Gaussian where $\mathbf{x}_i \sim N(\mu, \Sigma)$ for $i = 1, \ldots, n$, $\beta = \mu \in \mathbb{R}^k$ is the parameter of interest and $\gamma = \Sigma \in \mathbb{R}^{k\times k}$ is an nuisance parameter. Then, $\beta_{\max}$ represents the largest population mean and often bears practical implications, such as the best subgroup effect in clinical studies. For $\mathbf{x}_i$, the density function is

$$p(\mathbf{x}_i; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{(\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)}{2}} \tag{13}$$

with two sufficient statistics: $T_1(\mathbf{x}_i) = \mathbf{x}_i, T_2(\mathbf{x}_i) = \mathbf{x}_i\mathbf{x}_i^T$. Given data sets $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, the log-likelihood is

$$\ln p(\mathcal{X}; \mu, \Sigma) = -\frac{nk\ln(2\pi)}{2} - \frac{n\ln(|\Sigma|)}{2} - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)^T\Sigma^{-1}(\mathbf{x}_i - \mu), \tag{14}$$

and the privatized point estimate by MLE is

$$\begin{cases} \hat{\mu}^{priv} = \frac{1}{n}(\sum_{i=1}^{n}\mathbf{x}_i + w_1) \\ \hat{\Sigma}^{priv} = \frac{1}{n-1}(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T + w_2) - \frac{1}{n(n-1)}(\sum_{i=1}^{n}\mathbf{x}_i + w_1)(\sum_{i=1}^{n}\mathbf{x}_i + w_1)^T \end{cases} \tag{15}$$

where $w_i \sim Lap(\triangle_i/\varepsilon_i)(i = 1, 2)$ is the Laplacian noise. $\triangle_i$ is the sensitivity of sufficient statistics $\sum T_i(x)$, and ε_i is the privacy budget allocated to it.

For the privatized bootstrap, we generate $\mathbf{x}_1^*, \cdots, \mathbf{x}_n^* \sim N(\hat{\mu}^{priv}, \hat{\Sigma}^{priv})$ and the estimate

$$\hat{\mu}^{*,priv} = \frac{1}{n}(\sum_{i=1}^{n}\mathbf{x}_i^* + w_1^*) \tag{16}$$

where w_1^* is the Laplacian noise generated by the same distribution of w_1 to account for the additional randomness in privacy protection. With $\hat{\mu}^{priv}$ and $\hat{\mu}^{*,priv}$, we can proceed following Algorithm 1.

5.2 | Linear Regression

We consider the linear regression case. While inspired by exponential family, some modifications are adopted to better fit the protection of privacy in parametric bootstrap detailed later. Consider a linear model $y_i = \mathbf{x}_i^T\beta + e_i$, where $e_i \sim N(0, \sigma^2)$, $\beta, \mathbf{x}_i \in \mathbb{R}^k, y_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Then, $\beta_{\max}$ represents the largest regression coefficient and often bears practical implications, such as the strongest signal in genetic association studies. The density function is

$$p(\mathbf{x}_i, y_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i - \mathbf{x}_i^T\beta)^2}{2\sigma^2}}. \tag{17}$$

Let $X \in \mathbb{R}^{n\times k}$ denote the matrix with the i^{th} row equal to $\mathbf{x}_i^T$, and $\mathbf{y}, \mathbf{e} \in \mathbb{R}^n$ be the vectors with the i^{th} entries y_i and e_i, respectively. Then the linear regression problem becomes $\mathbf{y} = X\beta + \mathbf{e}$, and the log-likelihood is

$$\ln P(X, \mathbf{y}; \beta, \sigma^2) = -\frac{\mathbf{y}^T\mathbf{y} - 2\beta^T X^T\mathbf{y} + \beta^T X^T X\beta}{2\sigma^2} - \frac{n}{2}\ln(\sigma^2), \tag{18}$$

and the privatized point estimate for β by MLE is

$$\hat{\beta}^{priv} = (X^T X + w_1)^{-1}(X^T\mathbf{y} + w_2). \tag{19}$$

where $w_i \sim \text{Lap}(\triangle_i/\varepsilon_i)(i = 1, 2)$ is the Laplacian noise, $\triangle_i$ is the sensitivity of sufficient statistics $\sum T_1 = X^T X$, and $\sum T_2 = X^T\mathbf{y}$. Following Ferrando et al. (2022), we adopt a bias-corrected estimate for σ^2 with additional Laplacian noise

$$\hat{\sigma}^{2,priv} = \frac{1}{n-k}\left[\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\beta}^{priv})^2\right] + w_3, \tag{20}$$

where $w_3 \sim \text{Lap}(\triangle_3/\epsilon_3)$ is additional the Laplacian noise, and $\triangle_3$ is the sensitivity of the variance. This scheme is also beneficial to the partially privatized settings to be discussed in the next section.

If we use the estimated model $\mathbf{y} = X\hat{\beta}^{priv} + \mathbf{e}$ to generate bootstrap sample, we need to access the original data X many times and sacrifice privacy budget to protect X. Ferrando, Wang, and Sheldon (2020) suggests we rewrite the privatized MLE $\hat{\beta}^{priv}$ in Eq. 19 and consider the following bootstrap estimate for β

$$\sqrt{n}\hat{\beta}^{*,priv} = \sqrt{n}(S^{priv} + \frac{1}{n}w_1^*)^{-1}S^{priv}\hat{\beta}^{priv} + (S^{priv} + \frac{1}{n}w_1^*)^{-1}(C^{*,priv} + \frac{1}{\sqrt{n}}w_2^*) \tag{21}$$

where $S^{priv} = \frac{1}{n}X^T X + \frac{1}{n}w_1$, and we generate $C^{*,priv} \sim N(0, \hat{\sigma}^{2,priv}S^{priv})$. w_1, w_2 are corresponding Laplacian noises, and w_1^*, w_2^* are the Laplacian noises generated by the same distribution of w_1, w_2 to account for the addition randomness in privacy protection. With $\hat{\beta}^{priv}$ and $\hat{\beta}^{*,priv}$, we can proceed following Algorithm 1.

6 | PARTIALLY PRIVATE METHOD

In this section, we introduce a strategy to save privacy budget when the calculation and bootstrapping of parameter of interest β only depends on part of sufficient statistics T. We discuss the implementation of the strategy in two applications for multivariate Gaussian and linear regression.

The consistent property for partially privatized cases still holds by replacing $\hat{\alpha}^{priv}$ with $\hat{\alpha} = (\hat{\beta}^{priv}, \gamma)$ in the proof, and the details are contained in supporting information.

6.1 | General Privacy Budget Reduction

In some practical applications, the release of the estimate of parameter of interest β might only depends on part of sufficient statistics as stated in Theorem 3.

Theorem 3. (Estimates for parameters under special case for partial privacy)
In the framework of exponential family shown in section 3.1, if $\frac{\partial A}{\partial f_1} = l(\beta)$, l is some function, then the MLE has the form,

$$\begin{cases} \hat{\beta} = g_1(\sum T_1(x_i)) \\ \hat{\gamma} = g_2(\sum T_1(x_i), \sum T_2(x_i)). \end{cases} \tag{22}$$

From Eq. 22, we see that $T_2(x)$ has nothing to do with the release of $\hat{\beta}$, which implies that for estimation, we may save the privacy budget without adding Laplacian noise to $T_2(x)$ or in specific.

Theorem 4. (Differential privacy for partially privatized case)
Let

$$\begin{cases} \hat{\beta}^{priv} = g_1(\sum T_1(x_i) + w_1) \\ \hat{\gamma} = g_2(\sum T_1(x_i) + w_1, \sum T_2(x_i)). \end{cases} \tag{23}$$

We obtain an ε_1-DP estimation instead of the $(\varepsilon_1 + \varepsilon_2)$-DP with Eq. 23.

Furthermore, if the release of $\hat{\beta}^*$ has nothing to do with $\hat{\gamma}$ or only depends on part of $\hat{\gamma}$, we may only need to add noise to the relevant part in estimating γ and save the privacy budget in Algorithm 1. We state the partially privatized scheme of Gaussian case and regression case, and more details are included in supporting information.

6.2 | Partially Privatized Multivariate Gaussian Case

In applications, we might be only interested in some of the population means. Suppose

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_i^1 \\ \mathbf{x}_i^2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right) \tag{24}$$

where $\mathbf{x}_i^1 \in \mathbb{R}^{n_1}$, $\mathbf{x}_i^2 \in \mathbb{R}^{n_2}$. Then $\beta = \mu_1$ is the parameter of interest, and $\gamma = (\mu_2, \Sigma)$ is the nuisance parameter, where $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. We can check that in this multivariate Gaussian case, $\frac{\partial A}{\partial f_1} = \frac{1}{\mu_1}$, which satisfies the condition in Theorem 3. Note that in parametric bootstrap, we only need to estimate μ and Σ_{11} instead of all, we can modify Eq. 15 and derive the partially privatized MLE estimate by only adding Laplacian noise w_1, w_2 to sufficient statistics $\sum_{i=1}^{n_1} \mathbf{x}_i^1$, $\sum_{i=1}^{n_1} \mathbf{x}_i^1 (\mathbf{x}_i^1)^T$.

We have the bootstrap estimate

$$\hat{\mu}_1^{*,priv} = \frac{1}{n_1}(\sum_{i=1}^{n_1} \mathbf{x}_i^* + w_1^*) \tag{25}$$

where w_1^* is the Laplacian noise generated by the same distribution of w_1. Then we calculate the $T^{*,priv}$ and the privatized lower confident limit as in Algorithm 1. As for the bias-selection correction parts in step 3 and step 7 in Algorithm 1, we plug in partial sample size n_1. With this modification, we can save the privacy budget for $\mu_2, \Sigma_{12}, \Sigma_{21}$ and Σ_{22}.

6.3 | Linear Regression with Nuisance Parameters

In many applications, we often consider a linear regression model $y_i = \mathbf{z}_i^T \beta + \mathbf{x}_i^T \gamma + e_i$, $i = 1, \cdots, n$. Take subgroup analysis as an example, $\mathbf{z}_i \in \mathbb{R}^{k_1}$ can be the interaction terms between subgroup indicators and treatment indicator, $\mathbf{x}_i \in \mathbb{R}^{k_2}$ can be pre-treatment covariates, and y_i is the response vector, $e_i \sim N(0, \sigma^2), i.i.d.$ is the white noise; see Imai, Ratkovic, et al. (2013). Then, $\beta \in \mathbb{R}^{k_1}$ is the parameter of interest and $\beta_{\max}$ represents the best subgroup effect, and $\gamma \in \mathbb{R}^{k_2}$ is the nuisance parameter.

Let $Z \in \mathbb{R}^{n \times k_1}$ denote the matrix with the i^{th} row equal to $\mathbf{z}_i^T$, $X \in \mathbb{R}^{n \times k_2}$ denote the matrix with the i^{th} row equal to $\mathbf{x}_i^T$ and $\mathbf{y}, \mathbf{e} \in \mathbb{R}^n$ denote the vectors with the i^{th} entries y_i and e_i, respectively. In some real problems such as randomized trails, we have $Z^T X = 0$. We can check

that $A(\beta, \gamma, \sigma^2) = \frac{1}{2}\ln\sigma^2$. Thus $\frac{\partial A}{\partial \beta} = 0$. The condition of Theorem 3 is satisfied. We then follow the scheme of Eq. 19 to construct privatized MLE for β by only adding Laplacian noises w_1, w_2 to 2 sufficient statistics $\sum T_1 = Z^T Z$ and $\sum T_3 = Z^T \mathbf{y}$ related to β.

We adopt a bias-corrected estimate $\hat{\sigma}^{2,priv}$ for σ^2 with additional Laplacian noise following the form of Eq. 20 by plugging $\hat{\beta}^{priv}, \hat{\gamma}$. Similarly, we form a bootstrap estimate $\hat{\beta}^{*,priv}$ that follows the idea of Eq. 21:

$$\sqrt{n}\hat{\beta}^{*,priv} = \sqrt{n}(S^{priv} + \frac{1}{n}w_1^*)^{-1}S^{priv}\hat{\beta}^{priv} + (S^{priv} + \frac{1}{n}w_1^*)^{-1}(C^{*,priv} + \frac{1}{\sqrt{n}}w_2^*), \tag{26}$$

where $S^{priv} = \frac{1}{n}Z^T Z + \frac{1}{n}w_1$, and we generate $C^{*,priv} \sim N(0, \hat{\sigma}^{2,priv}S^{priv})$. w_1, w_2 are corresponding Laplacian noises, and w_1^*, w_2^* are the Laplacian noises generated by the same distribution of w_1, w_2 to account for the additional randomness in privacy protection. With $\hat{\beta}^{priv}$ and $\hat{\beta}^{*,priv}$, we can proceed Algorithm 1.

7 | SIMULATION

In this section, we take multivariate Gaussian case as an example and conduct Monte Carlo simulation to demonstrate the benefits of the proposed method. Results of other scenarios can be found in the supporting information.

We consider a simple setting with data $\mathbf{x_1}, \cdots, \mathbf{x_n} \sim N(\mu, \Sigma)$ where Σ is identity matrix. We consider two cases for μ following the settings of existing literature, the vector $(0, \cdots, 0, 0)$ with $k's$ zeros and the vector $(0, \cdots, 0, 1)$ with $(k-1)'s$ zeros and 1 one. The best subpopulation effect in the two cases are 0 and 1 respectively, and we denote BS=0 and BS=1 for the two cases.The parameter of interest β is constructed as the first k_1 items of μ: $\beta = (\mu_1, \cdots, \mu_{k_1})$. We take $k_1 = k$ for the wholly privatized case, and $k_1 = k/2$ for the partially privatized case as an example in the simulation. We generate random samples of size $n = 400$ for each subpopulation and use 1000 Monte Carlo samples in evaluating the empirical coverage and average distance from the true maximum value and the estimated 95% lower confidence limit. We denote such distance as 'length' for simplification. We consider tuning parameter $r = 1/30, 1/15, 1/10, 1/5$, and the tuning parameter chosen by cross-validation. We adopt two kinds of noise with privacy budget $\varepsilon = 1.5$ and $\varepsilon = 5$, to simulate different levels of privacy protection.

For comparison, we adopt the naive privatized method, where we construct the CI by normal approximation with the estimated privatized extrema and its standard error considered in Guo and He (2021). The non-private naive method has the same structure except for Laplacian noise. We also compare a semi-naive bootstrap method by setting tuning parameter $r = 0.5$, which implies that we do not add a bias correction term in the bootstrap. For simplification, we use the following abbreviations: (1) **WPB**: wholly privatized bootstrap; (2) **PPB**: partially privatized bootstrap; (3) **NPB**: non-privatized parameter bootstrap; (4) **WPN**: wholly privatized naive method; (5) **PPN**: partially privatized naive method; (6) **NPN**: non-privatized naive method; (7) **rWPB**: wholly privatized bootstrap without account for randomness induced by Laplace noise. We consider different scenarios to demonstrate the benefits of our proposed method as follows. The simulation results are summarized in Table B2.

We start with a the 2-dimensional case $k = 2$ and set the privacy budget with privacy budget $\varepsilon = 1.5$ to see the effect of bias correction. From Figure 1, we see that bias-correction plays an important role in achieving nominal coverage in both private settings and non-private settings, and the proposed methods work well with cross-validation. To demonstrate the merits of accounting for randomness in bootstrap, we skip Step 6 in Algorithm 1; i.e. implement **rWPB**. The results are shown in Table B1 in AppendixB. We can observe that the coverage is unsatisfying. Thus it is essential to account for randomness induced by noise terms in constructing CI with private data.

Figures 2 and 3 show the results with tuning parameters under cross-validation for bootstrap methods testing with 2-dimensional data and 8-dimensional data. We can see that bootstrap methods work well on bias correction compared to naive methods. We can also observe a trade-off between length and privacy protection. With a smaller ε, the privacy protection is stronger with the sacrifice in length and efficiency. With a larger ε, the length decreases due to less randomness induced by Laplacian noise, but the privacy protection becomes weaker. We can also learn that modifying partially privatized methods works well and leads to more efficient CI for extrema parameters. The results also show that when the dimension gets higher, the advantage of our method becomes more significant. Detailed results are listed in Table B2 in AppendixB.

Figure 1 Coverage and length with different tuning parameters with 2-dimensional data

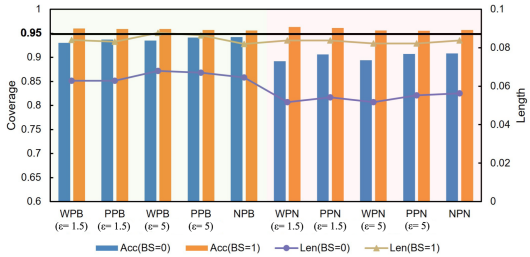

Figure 2 Simulation on 2-dimensional data

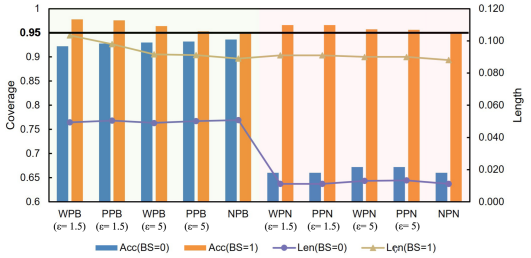

Figure 3 Simulation on 8-dimensional data

8 | APPLICATION

In this section, we apply the proposed method to the national supported work (NSW) program, where we aim to identify and assess the subpopulation of workers for whom the job training program is most beneficial. In particular, we focus on the subset of NSW considered in Dehejia and Wahba (1999) and LaLonde (1986), where the treatment and control groups consist of randomly chosen 297 and 425 such workers respectively.

Following the setup of subgroup analysis of NSW in Imai et al. (2013), we consider the regression model $Y \sim DZ + X$ for the empirical application here where we consider compared to the earnings before the program (measured in 1975), the salary increase after the job training program (measured in 1978) as the response vector Y and we consider four subpopulations considered in the setup of Imai et al. (2013) defined by the binary variables of years of education (less than the median or not), and earnings in 1975 (higher than the median or not) and the interaction matrix DZ is formed by interaction terms between subpopulation indicators Z and treatment indicators D, and the matrix X represents the baseline covariates which consist of 4 subpopulation indicators and seven pre-treatment covariates including age, years of education, and the log of one plus 1975 earnings, as well as binary indicators for race, marriage status, college degree, and whether the individual was unemployed in 1975. To implement the proposed method, we consider tuning parameter $r = -\infty, 1/30, 1/15, 1/10, 1/5, 0.5$, where $r = -\infty$ corresponds to a simultaneous method, and $r = 0.5$ implies that we do not add bias correction term in the bootstrap. For partially privatized scheme, we add noise to nuisance parameters and to sufficient statistics except $X^T X$. Here we assume X is public information, which corresponds to the scheme in Sec. 5.3. We test the privacy budget with $\varepsilon = 500$ and $\varepsilon = 2000$ for both the wholly-privatized schemes and partially privatized schemes.

The results are summarized in Figure 4, and Table B3 in AppendixB. All methods imply that subpopulation 3 is the best subpopulation. Under non-privatized setting, with the naive methods, the lower confidence bound is positive, but the results of parametric bootstrap methods with bias correction show negative effects, which indicates that without bias correction, we might draw overly optimistic conclusions. The proposed method, whatever it is partially private or fully private and the budget is, leads to the same conclusion as that from the parametric bootstrap method that the best subgroup is not statistically significant in the 95% level. In addition, partially privatized method can save privacy budget as indicated by the distance between the lower confidence bound and the confidence bound drawn from non-privatized parametric bootstrap, especially when the magnitude of sufficient statistics $X^T X$ is large. Finally, our results are consistent with that of Imai et al. (2013), which implies subpopulation 3, consisting of people with low education and high earning in 1975, is the best subpopulation, but previous work fails to provide trustworthy inference results.

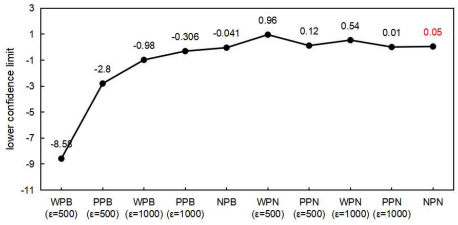

Figure 4 CI of real data example

9 | CONCLUSION

We propose a method to construct a CI for the extrema parameter under privacy scheme, which is efficient and easy to be implemented. We validate it by both analysis and experiments. Via a carefully designed privatized bootstrap procedure, selection bias in extrema parameter is appropriately adjusted under differential privacy and the randomness induced by Laplace noise is well accounted for. We also propose a partially privatized strategies which can help avoid wasting privacy budget for some application scenarios.

It is of great interest to extend the work to the case of high-dimensional linear regression where appropriate studies of how to preserve sparsity under privacy protection is needed, and we may design private algorithms with the output perturbation and objective perturbation for the extrema parameter problem.

ACKNOWLEDGMENTS

This work was supported by HKUST IEG19SC04 and the Project of Hetao Shenzhen-HKUST Innovation Cooperation Zone HZQB-KCZYB-2020083.

APPENDIX

A CROSS-VALIDATION

To choose r, we suggest a data-adaptive cross-validated method. To start with, we consider a bias-reduced estimate $\hat{\beta}^{priv}_{\max,reduced}$ as follows:

$$\hat{\beta}^{priv}_{\max,reduced} = \hat{\beta}^{priv}_{\max} - E^*[\beta^{priv,*}_{\max,modified} - \hat{\beta}^{priv}_{\max}], \tag{A1}$$

where E^* denotes expectation under bootstrap distribution.

The idea of cross-validation is to choose r to minimize the mean square error between $\hat{\beta}^{priv}_{\max,reduced}(r)$ and $\beta^{priv}_{\max}$. Let $A = \{r_1, \cdots, r_m\}$ denote a set of possible tuning parameters in the range of $(0, 0.5)$ with $r_1 < \cdots < r_m$ and m is a finite integer. The following algorithm can be used to choose $r \in A$ under differential privacy based on our framework. For the j-fold in cross validation, we denote $\hat{\beta}^{priv}_j = \{\hat{\beta}^{priv}_{j,i}\}^{k_1}_{i=1}$ with $\hat{\beta}^{priv}_{j,i}$ as the i-th item of $\hat{\beta}^{priv}_j$.

Algorithm 2 Cross-validated choice of tuning parameter r

Data: $x_1, \cdots, x_n$

Result: Optimal choice of tuning parameter r

```
Randomly partition the data into v (approximately) equalsized subsamples
for l = 1, ..., m do
    for j = 1, ..., v do
        Use the jth subsample as the reference data and the rest as the training data
        Use the training data to obtain the bias-reduced estimate with DP via (A1) β̂^priv_{max,reduced,j}(r_l), with r_l as the tuning prarmeter
        Use the reference data to estimate β̂^priv_j
        for i = 1, ..., k_1 do
            Calculate the standard error σ̂^priv_{j,i} for β̂^priv_{j,i}
            Calculate accuracy of each choice h^priv_{j,i}(r_l) = (β̂^priv_{max,reduced,j}(r_l) − β̂^priv_{j,i})^2 − (σ̂^priv_{j,i})^2
        end for
    end for
end for
The tuning parameter is chosen to be argmin_{r_l}{ min_{i∈[k_1]} [Σ_{j=1}^{j=v} h^priv_{j,i}(r_l)]/v }.
```

Theorem 5. (Differential privacy for cross-validation (Algorithm 2))
According to Lemma 1, Lemma 2 and Theorem 1, the framework of cross-validation remains $(\epsilon_1 + \epsilon_2)$-DP.

B TABLES OF RESULTS IN SIMULATION AND APPLICATION

Table B1 Parametric bootstrap with no randomness of Gaussian case with $k = 2$

		BS=0		BS=1	
		coverage	**length**	**coverage**	**length**
rWPB	r = -∞	0.890	0.063	0.942	0.098
	r = 1/30	0.882	0.062	0.916	0.085
	r = 1/15	0.881	0.061	0.913	0.084
	r = 1/10	0.880	0.061	0.911	0.083
	r = 1/5	0.873	0.059	0.910	0.082
	r = 0.5	0.844	0.051	0.909	0.082
	cv	0.878	0.061	0.912	0.083

Table B2 Simulation results of Gaussian case with $k = 2$ and $k = 8$

		dim=2				dim=8			
		BS = 0		BS=1		BS=0		BS=1	
		coverage	length	coverage	length	coverage	length	coverage	length
WPB ($\varepsilon = 1.5$)	r =-∞	0.950	0.066	0.975	0.098	0.940	0.052	0.996	0.123
	r = 1/30	0.948	0.065	0.963	0.085	0.926	0.050	0.978	0.105
	r = 1/15	0.948	0.065	0.960	0.084	0.923	0.050	0.972	0.100
	r = 1/10	0.942	0.061	0.958	0.083	0.918	0.049	0.964	0.095
	r = 1/5	0.920	0.058	0.957	0.080	0.909	0.046	0.940	0.082
	r = 0.5	0.880	0.052	0.957	0.080	0.724	0.020	0.938	0.081
	cv	0.930	0.063	0.960	0.084	0.922	0.049	0.978	0.103
PPB ($\varepsilon = 1.5$)	r = -∞	0.954	0.068	0.975	0.098	0.939	0.052	0.996	0.123
	r = 1/30	0.948	0.065	0.965	0.086	0.935	0.050	0.978	0.105
	r = 1/15	0.944	0.064	0.962	0.085	0.923	0.050	0.972	0.100
	r = 1/10	0.944	0.064	0.959	0.083	0.919	0.049	0.964	0.095
	r = 1/5	0.936	0.062	0.959	0.082	0.909	0.046	0.942	0.082
	r = 0.5	0.884	0.054	0.959	0.082	0.723	0.020	0.940	0.081
	cv	0.937	0.063	0.959	0.083	0.928	0.050	0.976	0.098
WPB ($\varepsilon = 5$)	r = -∞	0.950	0.071	0.977	0.101	0.950	0.053	0.992	0.123
	r = 1/30	0.942	0.069	0.959	0.088	0.936	0.051	0.978	0.104
	r = 1/15	0.942	0.069	0.957	0.086	0.936	0.051	0.976	0.100
	r = 1/10	0.942	0.069	0.955	0.086	0.932	0.050	0.970	0.094
	r = 1/5	0.934	0.067	0.954	0.085	0.922	0.047	0.948	0.081
	r = 0.5	0.908	0.059	0.954	0.085	0.756	0.021	0.946	0.080
	cv	0.935	0.068	0.959	0.088	0.930	0.049	0.964	0.092
PPB ($\varepsilon = 5$)	r = -∞	0.950	0.071	0.977	0.101	0.955	0.055	0.993	0.125
	r = 1/30	0.942	0.069	0.958	0.088	0.938	0.054	0.982	0.105
	r = 1/15	0.942	0.069	0.957	0.086	0.934	0.051	0.980	0.100
	r = 1/10	0.942	0.069	0.956	0.086	0.934	0.051	0.972	0.093
	r = 1/5	0.934	0.067	0.954	0.085	0.929	0.050	0.948	0.081
	r = 0.5	0.906	0.059	0.954	0.085	0.796	0.029	0.946	0.080
	cv	0.941	0.067	0.957	0.086	0.932	0.050	0.953	0.091
NPB	r = -∞	0.950	0.071	0.975	0.098	0.939	0.052	0.996	0.123
	r = 1/30	0.942	0.070	0.965	0.085	0.937	0.051	0.978	0.105
	r = 1/15	0.942	0.069	0.962	0.084	0.923	0.050	0.972	0.100
	r = 1/10	0.940	0.069	0.959	0.083	0.919	0.049	0.964	0.095
	r = 1/5	0.934	0.067	0.956	0.082	0.908	0.046	0.942	0.082
	r = 0.5	0.908	0.059	0.956	0.082	0.725	0.020	0.940	0.081
	cv	0.942	0.065	0.956	0.082	0.936	0.051	0.952	0.089
WPN($\varepsilon = 1.5$)		0.892	0.052	0.963	0.084	0.660	0.011	0.966	0.091
PPN($\varepsilon = 1.5$)		0.906	0.054	0.961	0.084	0.660	0.011	0.966	0.091
WPN($\varepsilon = 5$)		0.894	0.052	0.956	0.082	0.672	0.013	0.957	0.090
PPN($\varepsilon = 5$)		0.907	0.055	0.955	0.082	0.672	0.013	0.956	0.090
NPN		0.908	0.056	0.957	0.084	0.660	0.011	0.952	0.088

Table B3 Real data example

	WPB ($\varepsilon = 500$)	**PPB** ($\varepsilon = 500$)	**WPB** ($\varepsilon = 1000$)	**PPB** ($\varepsilon = 1000$)	**NPB**	**WPN** ($O(0.1)$)	**PPN** ($O(0.1)$)	**WPN** ($O(0.01)$)	**PPN** ($O(0.01)$)	**NPN**
lower confidence limit	-8.58	-2.8	-0.98	-0.306	-0.041	0.96	0.12	0.54	0.01	0.05

References

Balle, B., Barthe, G., Gaboardi, M., Hsu, J., & Sato, T. (2020). Hypothesis testing interpretations and renyi differential privacy. In *International conference on artificial intelligence and statistics* (pp. 2496–2506).

Bornkamp, B., Ohlssen, D., Magnusson, B. P., & Schmidli, H. (2017). Model averaging for treatment effect estimation in subgroups. *Pharmaceutical statistics*, *16*(2), 133–142.

Brawner, T., & Honaker, J. (2018). Bootstrap inference and differential privacy: Standard errors for free. *Unpublished Manuscript*.

Chadha, K., Duchi, J., & Kuditipudi, R. (2021). Private confidence sets. In *Neurips 2021 workshop privacy in machine learning*.

Covington, C., He, X., Honaker, J., & Kamath, G. (2021). Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), 1053–1062.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., & Rubinstein, B. I. (2014). Robust and private bayesian inference. In *International conference on algorithmic learning theory* (pp. 291–305).

Dimitrakakis, C., Nelson, B., Zhang, Z., Mitrokotsa, A., & Rubinstein, B. I. (2017). Differential privacy for bayesian inference through posterior sampling. *Journal of machine learning research*, *18*(11), 1–39.

Du, W., Foot, C., Moniot, M., Bray, A., & Groce, A. (2020). Differentially private confidence intervals. *arXiv preprint arXiv:2001.02285*.

Dunsche, M., Kutta, T., & Dette, H. (2022). Multivariate mean comparison under differential privacy. In *International conference on privacy in statistical databases* (pp. 31–45).

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265–284).

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Ferrando, C., Wang, S., & Sheldon, D. (2020). General-purpose differentially-private confidence intervals. *arXiv preprint arXiv:2006.07749*.

Ferrando, C., Wang, S., & Sheldon, D. (2022). Parametric bootstrap for differentially private confidence intervals. In *International conference on artificial intelligence and statistics* (pp. 1598–1618).

Fuentes, C., Casella, G., & Wells, M. T. (2018). Confidence intervals for the means of the selected populations. *Electronic Journal of Statistics*, *12*(1), 58–79.

Gaboardi, M., Lim, H., Rogers, R., & Vadhan, S. (2016). Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on machine learning* (pp. 2111–2120).

Guo, X., & He, X. (2021). Inference on selected subgroups in clinical trials. *Journal of the American Statistical Association*, *116*(535), 1498–1506.

Guo, X., Wei, W., Liu, M., Cai, T., Wu, C., & Wang, J. (2022). Assessing heterogeneous risk of type ii diabetes associated with statin usage: Evidence from electronic health record data. *arXiv preprint arXiv:2205.06960*.

Hall, P., & Miller, H. (2010). Bootstrap confidence intervals and hypothesis tests for extrema of parameters. *Biometrika*, *97*(4), 881–892.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443–470.

Karwa, V., & Vadhan, S. (2017). Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604–620.

Li, N., Li, T., & Venkatasubramanian, S. (2006). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 ieee 23rd international conference on data engineering* (pp. 106–115).

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3–es.

Magnusson, B. P., & Turnbull, B. W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in medicine*, *32*(16), 2695–2714.

Mologen. (2018). Final analysis of impulse study confirms topline data with positive subgroup results. *MOLOGEN Press Releases*.

Nadarajah, S., & Kotz, S. (2008). Exact distribution of the max/min of two gaussian random variables. *IEEE Transactions on very large scale integration (VLSI) systems*, *16*(2), 210–212.

Rinott, Y., O'Keefe, C. M., Shlomo, N., & Skinner, C. (2018). Confidentiality and differential privacy in the dissemination of frequency tables. *Statistical Science*, *33*(3), 358–385.

Rogers, R., Roth, A., Smith, A., & Thakkar, O. (2016). Max-information, differential privacy, and post-selection hypothesis testing. In *2016 ieee 57th annual symposium on foundations of computer science (focs)* (pp. 487–494).

Rosenkranz, G. K. (2016). Exploratory subgroup analysis in clinical trials by model selection. *Biometrical Journal*, *58*(5), 1217–1228.

Sandercock, P. A. (2015). Short history of confidence intervals: Or, don't ask "does the treatment work?" but "how sure are you that it works?". *Stroke*, *46*(8), e184–e187.

Stallard, N., Todd, S., & Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference*, *138*(6), 1629–1638.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, *10*(05), 557–570.

Thomas, M., & Bornkamp, B. (2017). Comparing approaches to treatment effect estimation for subgroups in clinical trials. *Statistics in Biopharmaceutical Research*, *9*(2), 160–171.

Wang, Y., Kifer, D., & Lee, J. (2018). Differentially private confidence intervals for empirical risk minimization. *arXiv preprint arXiv:1804.03794*.

Woody, S., Padilla, O. H. M., & Scott, J. G. (2022). Optimal post-selection inference for sparse signals: a nonparametric empirical bayes approach. *Biometrika*, *109*(1), 1–16.

Xiang, D., & Cai, W. (2021). Privacy protection and secondary use of health data: Strategies and methods. *BioMed Research International*, *2021*.

Zhang, Z., Rubinstein, B. I., & Dimitrakakis, C. (2016). On the differential privacy of bayesian inference. In *Thirtieth aaai conference on artificial intelligence*.

Zhao, J. (2017). Composition properties of bayesian differential privacy. In *2017 ieee 28th annual international symposium on personal, indoor, and mobile radio communications (pimrc)* (pp. 1–5).

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.