A Differentially Private Bayesian Approach to Replication Analysis

by

Chengxin Yang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Surya T. Tokdar

_____
Amy H. Herring

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2022

ABSTRACT

A Differentially Private Bayesian Approach to Replication Analysis

by

Chengxin Yang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Supervisor

_____
Surya T. Tokdar

_____
Amy H. Herring

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2022

Abstract

Replication analysis is widely used in many fields of study. Once a research is published, other researchers will conduct analysis to assess the reliability of the published research. However, what if the data are confidential? In particular, if the data sets used for the studies are confidential, we cannot release the results of replication analyses to any entity without the permission to access the data sets, otherwise it may result in privacy leakage especially when the published study and replication studies are using similar or common data sets. In this paper, we present two methods for replication analysis. We illustrate the properties of our methods by a combination of theoretical analysis and simulation.

Contents

Abstract **iv**

List of Figures **vii**

Acknowledgements **viii**

1 Introduction **1**

2 Review of Replication Analysis and Differential Privacy **3**

2.1 Review of Replication Analysis 3

2.2 Review of Differential Privacy 5

3 The differentially private replication framework **8**

3.1 The framework for alternative data (AD) 9

3.2 The framework for alternative data (AM) 12

3.3 The two frameworks in algorithm-style 15

4 Theoretical Properties **17**

4.1 Properties of AD 17

4.2 Properties of AM 24

5 Work flow of the frameworks **26**

5.1 Work flow of AD 26

5.1.1 Constructing the tolerance region 26

5.1.2 Choosing the δ in AD 29

5.1.3 Choosing the M in AD 36

5.1.4 MCMC sampling and interpretation 40

5.2 Work flow of AM 41

5.2.1 Choosing the M in AM 41

5.2.2 Interpretation and reference contour plot 45

6 Empirical Illustrations on AD framework **52**

6.1 Consistency of r_0 53

6.2 Consistency of r 57

7 Application **59**

7.1 Application of AD 59

7.2 Application of AM 63

7.3 Comments on the applications 66

8 Conclusions **68**

Bibliography **69**

List of Figures

5.1 Examples of the simulated contours of $\bar{\nu}$. Every contour value, the $\bar{\nu}$, is the average of 500 randomly sampled ν. 48

6.1 Simulated posterior median of r without inflation. . Red points are empirical $\frac{S}{M}$, black points are the posterior median of r. 55

6.2 Simulated posterior median of r with inflation. Red points are empircal $\frac{S}{M}$, black points are the posterior median of r. 56

6.3 $residual = median(r|S^R) - \frac{S}{M}$. Red line is average $residual$. For each trial, black point is $residual$, purple point is posterior se of r. 57

7.1 Reference contour for r_{obs}. Contour value is the distance between 0.5 and the interval $(q_{0.1}, q_{0.9})$. (a) is for U_1, (b) is for U_2. 61

7.2 Posterior distribution of r. (a) is for U_1 and (b) is for U_2. Red lines indicate $\delta = 0.5$. 62

7.3 Reference contour plot for IPUMS. (a) and (c) are on $M = 25$, (b) and (d) are on $M = 50$. Each contour value is average on 500 ν. . . . 64

7.4 Distribution of $\hat{\theta}_n$ under repeated verifications. (a) is for U_1 and (b) is for U_2. Red lines show the 5% and 95% quantiles respectively. . . . 66

Acknowledgements

First and foremost, I want to express my great appreciation to my advisor, Dr. Jerome Reiter, for his invaluable and inspiring guidance during this research project, as well as the help with my career in statistics. His diligence, helpfulness and insight have always inspired me to step forward.

I want to thank Dr. Surya Tokdar for his wonderful lectures in statistical inference that inspired me to pursue the research in statistics. I would also like to send my appreciation to Dr. Olanrewaju Akande for his generous help with my statistical career, as well as the great support during my work at Duke. Further thanks give to Dr. Amy Herring, my advisor when I was admitted as a Master's student for her enthusiasm and genuine advice. I also want to thank all my friends at Duke, and my previous friends from Guangzhou, for lightening my life.

Finally, thanks to my parents and partner, for their unconditional love and generous support to my career.

Chapter 1

Introduction

In many fields of study, replication analysis is an important part of rigorous scientific discovery. Once a data-based result is published, it will be replicated and tested by other researchers to see if the result is repeatable and reliable. Apart repeating and checking, replication analysis can also be motivated by different reasons. A possible reason is some changes on the side of data (Hoeppner, 2019). Researchers may want to see whether the results in the research published hold in other datasets or when different subsets of data subjects are used for analysis. Another reason is the desire of using a different model. It's common to have a set of candidate models which perform similarly to each other in some model selection criteria and each of them is supported by some background knowledge. In this case, researchers may want to see whether the conclusion will change a lot if they adopt another candidate model.

Replication analysis can cause concerns for data privacy when the study is conducted on confidential data or related to sensitive information. Imagine a simple case where replication researchers want to check if the published research changes substantially when including or removing some outliers. If they report the sample mean with the outliers included or excluded, depending on what was done in the published research, it will cause loss in data privacy since people can predict the values of these outliers by comparing the two studies and the released sample means. Generally speaking, privacy loss can take place whenever a determined function of the confidential data set is released and it can be a serious issue if the replication studies are conducted many times by different researchers on a common confidential data set. To date, we know of no methods specifically for replication analysis that

satisfy any popular privacy protection standards. So, there is a practical need to develop methods for replication analysis while protecting the privacy of confidential data.

In this thesis, we're going to propose two differentially private frameworks for releasing results of replication analyses. One is called framework for alternative data (AD) which is designed for replication study motivated by the changes on the side of data while the model itself remains the same as the published research (data-motivated). The other is called the framework for alternative model (AM) which focuses on the replication study motivated by a different model specification (model-motivated). We restrict our scope to linear regression models and focus on the effect sizes of variables of interest, but the frameworks can be extended to other families of models and can be used to other questions of interest in replication analysis, e.g., the significance level of a conclusion. We investigate the properties of our frameworks by a combination of theoretical analysis and simulation study and demonstrate how they work in answering the questions in replication analysis while satisfying differential privacy (DP) constraints.

Chapter 2

Review of Replication Analysis and Differential Privacy

2.1 Review of Replication Analysis

Replication analysis has been a common practice in rigorous scientific discoveries because repeatability is considered a basic criteria in evaluating a scientific finding. In the setting of replication analysis, typically there will be an original finding coming from a previous study (in this thesis, we use the terms original, published and previous interchangeably). For some reasons, the researchers themselves, or some other researchers, may want to conduct a new search based on the same, or pretty similar methodologies as the previous study and see whether the conclusions change or not. As stated before, most of the reasons can be categorized into data-motivated or model-motivated. An example in data-motivated replication study is that, in social science, once a data-based study is published, other researchers may want to replicate a similar study in a different region to see if the published results apply in other geographic regions. For the model-motivated situation, there may be a set of candidate models which perform similarly to each other in some model selection criteria, but the researchers published their results based on one single final model. It also may be that the researchers made certain modeling assumptions in the published analysis, and other researchers may believe different assumptions would better describe the data. Under this circumstance, we may want to see if the published results are consistent among alternative models.

There are some existing statistical tools for conducting a replication analysis. For

example, if researchers are curious about the significance of effect sizes, they may conduct two hypothesis test, one for the original study and the other for replication study, which might be called as "vote counting" (Hoeppner, 2019). However, as pointed out (Valentine et al., 2011), there can be issues in the interpretation on the probabilistic sides, since the p-value in one study doesn't have the same meaning for the other one. Misleading results can even happen under underpowered replication studies (Simonsohn, 2015; Hoeppner, 2019). If researcher are curious about the magnitude of effect sizes, they may compare the two point estimations for the effect of interest, but this methodology doesn't consider the uncertainty of different estimations. As a substitute, using confidence intervals is becoming more preferable (Hoeppner, 2019). Overall, there is no single dominating approach in replication analysis and different interests can lead to different approaches. However, the methods merely using frequentist point estimations or hypothesis testing are under criticism recently, and more emphasis is being put on methods incorporating uncertainty quantifications. In addition, Bayesian approaches are getting more attention (Hoeppner, 2019) because of the technical restriction of frequentist approaches and the epistemic goal of replication analysis.

The replication analysis can be more complicated if the researchers are conducting studies on confidential data and they need to comply with some requirements on privacy protection. There have been many discussions on the importance of protecting the privacy of data, and there are also federal laws related to data privacy (Reiter 2018; Barrientos et al. 2019; Dwork et al. 2014). In the fields like social science and health data analysis where the data may contain much sensitive personal information, e.g., income, job, race, medical history and health conditions, privacy protection is often required.

A practical challenge under replication analysis is to decide the approach we

should take to preserve privacy. Generating synthetic data (Rubin 1993; Feinberg 1994; Raghunathan et al. 2003; Reiter et al. 2007) can retain the original workflow of replication analysis, but it will also suffer from the defects of this method (Barrientos 2019). If the synthetic data doesn't retain the correlation between variables of interest in our study, even if the replication study indicates significant difference from the published one, we will not be able to overturn the published one since we cannot differentiate whether the difference is caused by poor synthetic data or the difference between the two studies. For the method we use, the differential privacy, challenges will also arise, especially the difficulties in perturbing some certain quantities to obtain the privacy protection, since the common replication practice often refer to the raw estimations (Hoeppner, 2019), which can suffer from infinite sensitivity under differential privacy context.

2.2 Review of Differential Privacy

As a golden standard of privacy protection, differential privacy (DP) provides an attractive approach in combining the replication analysis with privacy protection. It adds noise to the statistics before release (Dwork et al., 2014). In the context of replication analysis, it has many advantages. First, it provides an exact probabilistic bound of the amount of information that can be leaked at most. Second, different from the synthetic data approach, it will not suffer from the quality of the synthetic data, which is very hard to evaluate. Because DP has much better probabilistic interpretation, we can have a better idea on the extent to which we're potentially perturbed from the raw replication result.

Suppose we have an algorithm $\mathcal{A}$ which takes a data set $\mathbf{D}$ as the input and outputs a numerical quantity o, namely, $\mathcal{A}(\mathbf{D}) = o$. Then, define the neighboring data set $\mathbf{D}'$ such that $\mathbf{D}$ and $\mathbf{D}'$ only differs in one record, i.e., there is only one

sample $d' \in \mathbf{D}'$ and $d \in \mathbf{D}$ such that $\mathbf{D} \setminus \{d\} = \mathbf{D}' \setminus \{d'\}$. Then, ϵ-differential privacy (ϵ-DP) is define in the following way.

Definition (ϵ-DP): An algorithm $\mathcal{A}$ satisfies ϵ-DP where $\epsilon > 0$ if for any pair $(\mathbf{D}, \mathbf{D}')$ and any output $o \in range(\mathcal{A})$, it holds that $Pr(\mathcal{A}(\mathbf{D}) = o) \leq e^{\epsilon} Pr(\mathcal{A}(\mathbf{D}') = o$.

ϵ is often named as privacy budget, and smaller ϵ implies stricter privacy guarantee. DP has three fundamental properties. Suppose $\mathcal{A}_1$ and $\mathcal{A}_2$ are ϵ_1-DP and ϵ_2-DP algorithms separately. First, for any data set $\mathbf{D}$, releasing both $\mathcal{A}_1(\mathbf{D})$ and $\mathcal{A}_2(\mathbf{D})$ satisfies $(\epsilon_1 + \epsilon_2)$-DP. Second, when $\mathbf{D}_1 \cap \mathbf{D}_2 = \emptyset$, releasing both $\mathcal{A}_1(\mathbf{D})$ and $\mathcal{A}_2(\mathbf{D})$ satisfies $max\{\epsilon_1, \epsilon_2\}$-DP. Third, for any algorithm $\mathcal{A}_3$, releasing $\mathcal{A}_3(\mathcal{A}_1(\mathbf{D}))$ satisfies ϵ_1-DP, which says that post-processing the output of ϵ-DP algorithms doesn't cause extra loss of privacy.

A common method to ensure ϵ-DP is the Laplace Mechanism (Dwork et al., 2006). Define the sensitivity of a function f to be $\Delta(f) := max_{(\mathbf{D},\mathbf{D}')}||f(\mathbf{D}) - f(\mathbf{D}')||$ where $(\mathbf{D}, \mathbf{D}')$ is any pair of neighboring data sets. Then, the released quantity after the Laplace mechanism is $LM(\mathbf{D}) = f(\mathbf{D}) + \eta$ where $\eta \sim Laplace(0, \frac{\Delta(f)}{\epsilon})$. Releasing $LM(\mathbf{D})$ satisfies ϵ-DP.

In practice, the Laplace noise can sometimes be too large such that the information contained in the quantity of interest will be obscured. A useful technique to reduce the Laplace noise is the sub-sample and aggregate method (Nissim et al., 2007). Roughly speaking, the idea is to reduce the sensitivity of f by partitioning the full data set $\mathbf{D}$ into M subsets $\{\mathbf{D}^{(l)}\}_{l=1}^{M}$, calculate $f(\mathbf{D}^{(l)})$ on each subset $\mathbf{D}^{(l)}$ and take the new $f(\mathbf{D})$, denoted as $f^{new}(\mathbf{D})$, to be the sample mean of $\{f(\mathbf{D}^{(l)})\}_{l=1}^{M}$, namely $f^{new}(\mathbf{D}) = \frac{1}{M}\sum_{l=1}^{M} f(\mathbf{D}^{(l)})$. In this way, for any neighboring data set $\mathbf{D}'$, it can only cause one of $\{f(\mathbf{D}^{(l)})\}_{l=1}^{M}$ to change its value while all the other remain the same. Thus, we have the $\Delta(f^{new}) = \frac{1}{M}\Delta(f)$. When adding Laplace noise to f^{new},

we sample the new noise $\eta^{new} \sim Laplace(0, \frac{\Delta(f)}{M\epsilon})$ such that $\mathbb{V}ar(\eta^{new}) = \frac{1}{M^2}\mathbb{V}ar(\eta)$. So, it's much less likely to have a noise large enough to overwhelm the quantity of interest.

Chapter 3

The differentially private replication framework

We begin by formalizing the replication analysis in the context of comparing the effect of a given variable, and we will specify the relevant objects and notations along the way of illustration. Suppose we have a confidential data set $\mathbf{D}$ with size n which was used for fitting the model in the original study. We have p variables $X_1, ..., X_p$ and response Y. The original researchers fitted a linear regression model named as $Model_0$ on $\mathbf{D}$, and published the effect of variable of interest $X \in \{X_1, ..., X_p\}$ by releasing the corresponding fitted coefficient $\hat{\gamma}_o$, which is an estimation of γ, the true coefficient of X in $Model_0$. Replication researchers also have access to $\mathbf{D}$, or another data set $\mathbf{D}^*$ that contains all the variables used by the original researchers. They cannot release the result of replication analysis directly and they want to release a DP measure of the replication analysis.

Generally speaking, by conducting a replication study, we aim to see whether the effect size changes substantially or not. However, the formalization of question is different for data-motivated and model-motivated situations. In the data-motivated situation, the replication researcher fit $Model_0$ on $\mathbf{D}^*$, which implicitly indicates that the model assumption remains the same as the original study. Namely, we will only have one true coefficient of X. Thus, we're actually comparing two estimations of γ when we comparing the results from original and replication study. In contrast, in model-motivated situation, we're going to specify a new model $Model_1$, namely, the model assumption changes. Thus, there will be a new true coefficient of X in $Model_1$, denoted as β. Under this setting, what we're ultimately interested in is the

difference between γ and β.

3.1 The framework for alternative data (AD)

Under this setting, we only have one model specification $Model_0$. Let's write it as $\mathbb{E}[y_i|\mathbf{x}_i] = \gamma_{intercept} + \boldsymbol{\gamma}^T\mathbf{x}_i$ where $\boldsymbol{\gamma}^T = (\gamma_1, ..., \gamma_p)$ and $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$. The illustration in this paper will be based on normal assumptions, namely $y_i|\mathbf{x}_i \sim \mathcal{N}(\gamma_0 + \boldsymbol{\gamma}^T\mathbf{x}_i, \sigma^2)$, but it can be extended to Generalized linear model as well as other families of model. For the variable of interest X, the corresponding true coefficient in $Model_0$ is denoted as γ. The original researchers fitted the model on confidential data set $\mathbf{D}$, and they published $\hat{\gamma}_o$ as the estimation for γ. Now, the replication researchers are going to fit this model on the new confidential data set $\mathbf{D}^*$.

Denote the estimation of γ in the replication study as $\hat{\gamma}_r$. There can be two ways for interpretation. The first is to take γ as a fixed parameter. In this sense, if $\hat{\gamma}_r$ is far from $\hat{\gamma}_o$, we may say that the original estimation is sensitive to the change of data. Another interpretation takes γ as a given but changeable parameter. It can change by nature, e.g., change with time, and we're making inference about this value. This situation can be common when the original research was published many years ago and the replication researchers want to see if the conclusion still holds or not. If $\hat{\gamma}_r$ differs a lot from $\hat{\gamma}_o$, we may say that the effect of the variable X changes. However, different interpretations will not lead to different methodologies in modeling, and we're always comparing $\hat{\gamma}_r$ and $\hat{\gamma}_o$.

Because we're curious about how far $\hat{\gamma}_r$ is from $\hat{\gamma}_o$, it's intuitive to set a tolerance region denoted as $U(\hat{\gamma}_o; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ are the parameters defining the region. If $\gamma \in U(\hat{\gamma}_o; \boldsymbol{\alpha})$, we tend to say the original result is valid. As it turns out, the specification of the tolerance region can significantly influence the work flow as well as the power of this method, and it cannot be arbitrarily specified. Generally speaking, there are

two kinds of tolerance region depending on whether $U(\hat{\gamma}_o; \boldsymbol{\alpha})$ is fixed or not. Being fixed namely means that the tolerance region doesn't depend on any varying quantity or parameters that we can adjust in this framework. An example of a fixed $U(\hat{\gamma}_o; \boldsymbol{\alpha})$ is $[0, +\infty)$, which can be used to test the sign of the coefficient. We will provide more discussions on the tolerance region in later sections.

Following Barrientos et al. (2018), define $\theta_0 = \mathbb{I}(\gamma \in U(\hat{\gamma}_o; \boldsymbol{\alpha}))$ where $\mathbb{I}(A)$ is the binary indicator function which takes value of 1 when A is true, otherwise 0. Of course, θ_0 is the ultimate quantity of interest which directly answers our question, but we cannot calculate it directly since we can never know the true value of γ. To approximate it, let's denote $\hat{\gamma}_{r,n}$ to be the estimation of γ based on n samples from $\mathbf{D}^*$, and define the pseudo parameter $\theta_n = \mathbb{I}(Pr[\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha})] \geq \delta)$, where $\delta \in (0, 1)$ is called as the degree of certainty needed. We need $Pr[\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha})]$ to be at least δ to make the conclusion that $\theta_0 = 1$. As it turns out, $\delta = \frac{1}{2}$ will be a robust choice for fixed tolerance region. No matter the δ, if $\hat{\gamma}_{r,n}$ is a consistent estimator of γ and $U(\hat{\gamma}_o; \boldsymbol{\alpha})$ itself doesn't depend on n, it's guaranteed that $\lim_{n\to\infty} \theta_n = \theta_0$.

However, we cannot report $\hat{\gamma}_{r,n}$ directly, nor any other deterministic functions of the confidential data set since it will violate the DP. We have to add noise to a quantity itself or its predecessor before we release it. Unfortunately, we cannot add noise directly to θ_n or the estimation of it either, since it will cause problems in the interpretation because it's a binary variable and the scale of the noise can be big enough to obscure the true value of θ_n. Instead, we release a "noisy version" of θ_n, namely denoted as $r_0 = Pr[\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha})]$ derived from the sub-sample and aggregate method (Nissim et al., 2007; Barrientos et al., 2019).

Denote the sample size of $\mathbf{D}^*$ as N. We randomly and evenly (ignoring the slight error when rounding N/M to an integer) partition $\mathbf{D}^*$ into M disjoint subsets $\mathbf{D}_1^*, ..., \mathbf{D}_M^*$, where M is an integer that can be selected by the users. In each $\mathbf{D}_l^*$,

we calculate the corresponding consistent estimator for γ denoted as $\hat{\gamma}_r^{(l)}$, which will often be the MLE in practice. Each $\hat{\gamma}_r^{(l)}$ can be viewed as an independent draw from the distribution of $\hat{\gamma}_{r,n}$ where $n = [\frac{N}{M}]$. Define $W_l = \mathbb{I}(\hat{\gamma}_r^{(l)} \in U(\hat{\gamma}_o; \boldsymbol{\alpha}))$ for each $\mathbf{D}_l^*$. Because it's a binary random variable, each W_l can be viewed as an independent draw from $Bernoulli(r_0)$, where r_0 stands for the Bernoulli parameter. Under this setting, we can make inference on r_0 through $S = \sum_{l=1}^{M} W_l$. To satisfy the DP constraint, we add Laplace noise to S. Denote $S^R = S + \eta$ where $\eta \sim Laplace(0, \frac{1}{\epsilon})$. The global sensitivity of S is exactly equaling to 1, because if we change one observation in the data set, S itself can change up to 1 since at most one of the partitions can switch from 1 to 0 or vice versa. By partitioning and aggregating, we can significantly reduce the effective Laplace noise that can obscure our inference result.

Unfortunately, interpreting S^R itself can be tricky. It is not bounded in $(0, M)$ nor guaranteed to be an integer. In addition, a single S^R itself cannot give uncertainty quantification of r_0. For better interpretation, we define a random quantity r in our Bayesian model as the substitution of r_0, and adopt MCMC for posterior sampling which has no bearing on the DP properties of S^R to improve the interpretation and describe the uncertainty. Our Bayesian model is proposed as

$$S^R|S \sim Laplace(S, \frac{1}{\epsilon}),\ S|r \sim Binomial(M, r),\ r \sim \phi_0,$$

where ϕ_0 is the prior distribution for r.

In this model specification, S^R is our observation and S is lying in the middle layer of the model on which we're averaging. Since r itself is the parameter in $W_l \overset{i.i.d.}{\sim} Bernoulli(r)$ which represents the probability, it can be conventional to take $\phi_0 = Beta(1, 1)$.

The posterior r conditioning on the observed S^R already have a fine interpretation by Barrientos el al. (2018). It can be interpreted as the approximate posterior

probability that the estimated coefficient in each subset is inside the tolerance region, and the replication researchers can make their conclusion based on the posterior distribution of r as they like. Furthermore, after δ is chosen, they can approximate θ_n by $\hat{\theta}_n$, which is defined as $Pr(r \geq \delta | S^R)$. As what we will illustrate in later sections, if $\delta = 0.5$, then $\hat{\theta}_n$ will also be a good reference for guessing θ_0.

3.2 The framework for alternative data (AM)

Under this setting, we have two model specifications, one is the published $Model_0$ and $Model_1$ for replication study. We're interested in whether the conclusions made by these two models are considerably different. Let's write $Model_0$ the same way as before, and $Model_1$ as $\mathbb{E}[y_i|\mathbf{x}_i] = \beta_0 + \boldsymbol{\beta}^T H(\mathbf{x}_i)$, where $\boldsymbol{\beta}^T = (\beta_1, ..., \beta_p, ..., \beta_{p+q})$ and $H(\mathbf{x}_i) = (h_1(x_{i1}), ..., h_p(x_{ip}), ..., h_{p+q}(x_{i,p+q}))$. $h_k(.)$ can be any function applied to the independent variable, e.g., $h_k(.) \equiv 0$ if we want to delete the k-th variable. Suppose the variable of interest is X, and denote the coefficient of X in $Model_0$ as γ, and the coefficient of X in $Model_1$ as β. Assume no transformation is conducted on X. The question of interest in replication analysis is comparing γ and β, and the data set used by the replication researchers to fit the two models is denoted as $\mathbf{D}^*$.

A point need to be clarified is that we don't require the replication researchers to fit $Model_1$ on $\mathbf{D}$. However, it will lead to slightly different interpretation when $\mathbf{D} = \mathbf{D}^*$ and $\mathbf{D} \neq \mathbf{D}^*$. For the former situation, we're not only evaluating the difference between the two models but also evaluating the reliability of $\hat{\gamma}_o$ directly. If $\mathbf{D} \neq \mathbf{D}^*$, we're evaluating the difference between the two models but not necessarily making conclusion on the reliability of $\hat{\gamma}_o$, since answering whether the true coefficients vary across different data sets requires additional knowledge. The exact interpretation will be subject to the background information as well as the two confidential data sets. For example, suppose $\mathbf{D}$ in the previous study comes from region A, while

the replication researchers only have access to the confidential data set $\mathbf{D}^*$ of region B. If the replication researchers make the conclusion that $Model_0$ and $Model_1$ give significantly different results based on $\mathbf{D}^*$, they might doubt the original research and request the original researchers to re-conduct the study by using $Model_1$ based on $\mathbf{D}$, but they should be conservative in announcing that the original research is not reliable since some truths across different geographical regions can be very different. However, if additional background knowledge indicates homogeneity in the truth across geographical regions, the replication researchers will be able to make stronger statements on the validity of $\hat{\gamma}_o$ itself.

As stated before, we can never compare γ and β directly since they're blind to us. However, comparing $\hat{\beta}$ directly with $\hat{\gamma}_o$ may not sufficient even if $\mathbf{D}^* = \mathbf{D}$, since it doesn't incorporate the uncertainty in estimating γ and brings in the inherent difference $\hat{\gamma}_o - \gamma$ to our analysis. The approach that we're going to take here is adopting the sub-sampling and aggregate approach on $\mathbf{D}^*$, fitting $Model_0$ and $Model_1$ on each subset and comparing the fitted coefficients. In this way, we can incorporate the randomness in estimating γ and β as well as evaluate the difference between two models on average so as to evaluate the reliability of the conclusions made by them. The design above doesn't block us from testing the reliability of the published results. If we want to test the validity of $\hat{\gamma}_o$, it's suggested to adopt $\mathbf{D}^* = \mathbf{D}$. As a remedy for not being able to do so, the reference for interpretation in section 5 will be based on the published results, which can also partially reveal the validity of $\hat{\gamma}_o$.

Since we're no longer comparing a new estimator to a constant but comparing two parameters through comparing two estimators under repeated experiments, it's intuitive to compare their confidence intervals (CI), which can also be preferred by replication analysis community (Hoeppner, 2019). One approach to compare CIs is to use the overlap between two CIs (Maghsoodloo and Huang 2009; Karr et al.,

2006; Mirjam el al., 2011; Matthew el al., 2005). There are many advantages by using CI, one of which is that it gives better probabilistic interpretation than using an arbitrarily set tolerance region, since CI provides the probability of covering the true parameter under repeated experiments. Another advantage is that CI takes the sample size of each sub-data set into consideration automatically because the length of CI should be adjusted by the sample size, at least for many conventional parametric CIs.

Now, we build our framework for comparing two models formally. Denote the sample size of $\mathbf{D}^*$ as N. Let $Model_0$ and $Model_1$ be specified before. Let M be the number of partitions given by the users, and we randomly and evenly partition $\mathbf{D}^*$ into disjoint $\mathbf{D}_1^*, ..., \mathbf{D}_M^*$ with each subset containing $n = [N/M]$ samples (ignoring the rounding error). For each data set $\mathbf{D}_l^*$, fit $Model_0$ and $Model_1$ separately and get the corresponding α-confidence intervals of interest $U_\gamma(\mathbf{D}_l^*; \alpha) = [L_\gamma(\mathbf{D}_l^*; \alpha), U_\gamma(\mathbf{D}_l^*; \alpha)]$ and $U_\beta(\mathbf{D}_l^*; \alpha) = [L_\beta(\mathbf{D}_l^*; \alpha), U_\beta(\mathbf{D}_l^*; \alpha)]$. In this paper, we choose $\alpha = 95\%$, and the confidence intervals are equal-tailed (have same probability in left and right tail). If $U_\gamma(\mathbf{D}_l^*; \alpha) \cap U_\beta(\mathbf{D}_l^*; \alpha) \neq \emptyset$, define $[L_l, U_l]$ as $U_\gamma(\mathbf{D}_l^*; \alpha) \cap U_\beta(\mathbf{D}_l^*; \alpha)$, otherwise define $[L_l, U_l]$ as $[0, 0]$. Then, we calculate the the overlapping between the two intervals

$$\nu^{(l)} = \frac{1}{2}[\frac{U_l - L_l}{U_\gamma(\mathbf{D}_l^*; \alpha) - L_\gamma(\mathbf{D}_l^*; \alpha)} + \frac{U_l - L_l}{U_\beta(\mathbf{D}_l^*; \alpha) - L_\beta(\mathbf{D}_l^*; \alpha)}]$$

according to the overlapping measure in length (Karr et al., 2006). Denote ν^* as the overlapping measure calculated based on $U_\gamma(\mathbf{D}^*; \alpha)$ and $U_\beta(\mathbf{D}^*; \alpha)$, and $\bar{\nu} = \frac{1}{M}\sum_{l=1}^{M} \nu^{(l)}$. It's not hard to see the overlapping measure is bounded in $[0, 1]$. If $\nu = 1$, the two intervals are exactly the same as each other, while $\nu = 0$ indicates that the joint part is no more than a single point.

We cannot release any of the values above directly due to the DP constraints. However, we should not add Laplace noise to ν directly either, because it's a quantity

that is bounded in $[0, 1]$, which means it can be easily out of boundary even when we adopt a common ϵ, e.g., $\epsilon = 1$ (Jaewoo et al, 2011; Reiter, 2018). Using the subsample and aggregate approach, we take the average $\bar{\nu}$, by which the global sensitivity is multiplied by the factor $1/M$. Thus, the noise added will no longer be drawn from $Laplace(0, \frac{1}{\epsilon})$ but $Laplace(0, \frac{1}{\epsilon M})$, which significantly reduces the probability of being out of boundary after adding the noise. As it turns out, another reason for using $\bar{\nu}$ is that ν can be highly variable which is an undesired property.

Then, our Bayesian model is proposed as

$$\nu^L | \bar{\nu} \sim Laplace(\bar{\nu}, \frac{1}{\epsilon M}),\ \bar{\nu} \sim \psi_0,$$

where ψ_0 is the prior distribution for $\bar{\nu}$. Because $\bar{\nu}$ is bounded from 0 to 1, it's natural to take ψ_0 as a Beta distribution, and $\psi_0 = Beta(1, 1)$ for the weak prior. Finally, the researchers will release the posterior samples of $\bar{\nu}$ and make inference based on the these samples. Our proposed way is to make a posterior credible intervals for the average overlap measure.

3.3 The two frameworks in algorithm-style

AD:

- Input: confidential data set $\mathbf{D}^* = \{(x_i, y_i)\}_{i=1}^N$. The model ($Model_0$). The number of subsets M. The differential privacy parameter ϵ. The variable of interest X and the original estimated coefficient $\hat{\gamma}_o$. The interval $U(\hat{\gamma}_o, \boldsymbol{\alpha})$. The required degree of uncertainty δ.
- Take a random partition of $\mathbf{D}^*$: $\mathcal{P} = \{\mathbf{D}_1, ..., \mathbf{D}_M\}$.
- For each subset, fit the model and get $\hat{\gamma}_r^{(l)}$ and $W_l = \mathbb{I}(\hat{\gamma}_r^{(l)} \in U(\hat{\gamma}_o, \boldsymbol{\alpha}))$. Then calculate $S = \sum_{l=1}^M W_l$.

- Generate $S^R \sim Laplace(S, \frac{1}{\epsilon})$. Then conduct MCMC sampling for the Bayesian model

$$S^R|S \sim Laplace(S, \frac{1}{\epsilon}),\ S|r \sim Binomial(M, r),\ r \sim \phi_0$$

- Output the posterior distribution of r through MCMC. Then, make inference based on the posterior distribution of r. One approach is to calculate $Pr(r \geq \delta|S^R)$.

AM:

- Input: confidential data set $\mathbf{D}^* = \{(x_i, y_i)\}_{i=1}^N$. The number of subsets M. The differential privacy parameter ϵ. The variable of interest X. The $Model_0$ (original model) and $Model_1$ (alternative model).

- Take a random partition of $\mathbf{D}^*$: $\mathcal{P} = \{\mathbf{D}_1, ..., \mathbf{D}_M\}$.

- For each subset, fit the two models and calculate the overlap measure $\nu^{(l)}$. Then, calculate $\bar{\nu} = \frac{1}{M}\sum_{l=1}^M \nu^{(l)}$.

- Generate $\bar{\nu}^L \sim Laplace(\bar{\nu}, \frac{1}{\epsilon M})$. Then sampling posterior samples of $\bar{\nu}$ based on the Bayesian model

$$\bar{\nu}^L \sim Laplace(\bar{\nu}, \frac{1}{\epsilon M}),\ \bar{\nu} \sim \psi_0$$

- Calculate the posterior credible intervals of $\bar{\nu}$ based on the posterior samples.

Chapter 4

Theoretical Properties

In this section, we're going to discuss some theoretical properties of the core quantities in both frameworks. The properties will be on convergence to show the consistency of our method. Other theoretical properties will be given in later sections.

4.1 Properties of AD

In this situation, we're focusing on the reliability of the published $\hat{\gamma}_o$, and $\theta_0 = \mathbb{I}(\gamma \in U(\hat{\gamma}_o; \boldsymbol{\alpha}))$ is the ultimate quantity of interest. Before the sub-sample and aggregate procedure, we defined $r_0 = Pr(\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha}))$ and $\theta_n = \mathbb{I}(Pr(\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha})) \geq \delta)$, where n indicates the sample size of each partition; after combining with sub-sample and aggregate procedure, we have S, S^R and r and their posterior samples. As it turns out, these quantities form a chain of convergence which provide the guarantee that the posterior r conditioning on S^R will nearly converge to θ_0 in probability.

Property a1: *If $\hat{\gamma}_{r,n}$ is a consistent estimator of γ, $U(\hat{\gamma}_o; \boldsymbol{\alpha})$ is fixed, and γ is not on the boundary of $U(\hat{\gamma}_o; \boldsymbol{\alpha})$, then $\lim_{n\to\infty} \theta_n = \theta_0$ almost everywhere.*

Proof:

θ_0 is a binary indicator, which takes a fixed value zero or one. If $\hat{\gamma}_{r,n}$ is a consistent estimator of γ, we will have $\lim_{n\to\infty} Pr(|\hat{\gamma}_{r,n} - \gamma| \geq \epsilon) = 0$ for any $\epsilon > 0$. Thus, by simply taking $\epsilon = inf_{x \in U(\hat{\gamma}_o; \boldsymbol{\alpha})} |\gamma - x|$, we have

$$\lim_{n\to\infty} r_0 = \lim_{n\to\infty} Pr(\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha})) = \theta_0,$$

from which we can get exactly $\lim_{n\to\infty} \theta_n = \theta_0$ for any $\delta \in (0, 1)$. $\square$

Then, denote $r|S$ as the posterior r conditioning on S, and similarly for $r|S^R$. They are different from the the sampling distribution of r. Take $r|S$ for example. It can be viewed as a random variable that is a function of S. After fixing S, the distribution of $r|S$ will be determined. The following properties illustrate the convergence of posterior r conditioning on S or S^R. The convergence is in the sense that we take S (or S^R) as a random variable, and under each realization of S (or S^R), the posterior r will tend to be closer to S/M (or S^R/M).

Property a2: *If we take ϕ_0 as a Beta distribution, then, for any fixed sub-group sample size $n > p$, where p indicates the number of coefficients in the model, we have the posterior conditional probability*

$$\lim_{M\to\infty} Pr(|r - \frac{S}{M}| \geq \epsilon|S) = 0$$

which can be denoted as $r|S \xrightarrow{P} \frac{S}{M}$ when $M \to \infty$.

Proof:

Suppose $\phi_0 = Beta(a_0, b_0)$, then we have the likelihood (full joint distribution) function

$$\begin{aligned} f(S^R, S, r) &= f_{S^R}(s^R|S, \frac{1}{\epsilon}) \cdot f_S(s|M, r) \cdot f_r(r) \\ &\propto e^{-\epsilon|s^R - s|} \cdot C_M^s r^s (1-r)^{M-s} \cdot r^{a-1}(1-r)^{b-1}\mathbb{I}(0 \leq r \leq 1) \\ &\propto e^{-\epsilon|s^R - s|} \cdot C_M^s r^{s+a-1}(1-r)^{M-s+b-1} \cdot \mathbb{I}(0 \leq r \leq 1) \end{aligned}$$

Then, we have the following full conditional in MCMC

$$f(r|S = s, S^R = s^R) \propto r^{s+a-1}(1-r)^{M-s+b-1} \cdot \mathbb{I}(0 \leq r \leq 1),$$

which is a $Beta(s + a, M - s + b)$ distribution. Thus,

$$\mathbb{E}(r|S, S^R) = \frac{S + a}{M + a + b} := \mu_{r|S,S^R}, \quad \mathbb{V}ar(r|S, S^R) = \frac{\mu_{r|S,S^R}(1 - \mu_{r|S,S^R})}{M + a + b + 1}$$

They are two functions of the random variable S. It's not hard to see $f(r|S, S^R)$ doesn't depend on S^R, thus $f(r|S, S^R) = f(r|S)$. When $M \to \infty$, by the law of large number, we have $\mu_{r|S} = \frac{S+a}{M+a+b} \to \frac{S}{M}$ and $\mathbb{V}ar(r|S) = \frac{\mu_{r|S}(1-\mu_{r|S})}{M+a+b+1} \to 0$. The decrease of variance indicates the density is getting concentrated around the mean, thus we have $\lim_{M\to\infty} Pr(|r - \frac{S}{M}| \geq \epsilon|S) = 0$. In addition, this proof is invariant to the MCMC because after conditioning on S, the posterior r will be independent with S^R, and the full conditional of r will be the exact distribution of $r|S$ conditioning on the S. $\square$

Property a3: *Suppose the conditions in Property a2 hold and the Laplace noise* $\eta \sim$ *Laplace*$(0, \frac{1}{\epsilon})$. *Then, for any given* $\delta > 0$, *The conditional probability*

$$\lim_{M\to\infty} Pr(|r - \frac{S^R}{M}| \geq \delta|S) = 0.$$

Proof:

Because $|r - \frac{S^R}{M}| = |r - \frac{S}{M} + \frac{S}{M} - \frac{S^R}{M}| \leq |r - \frac{S}{M}| + |\frac{S}{M} - \frac{S^R}{M}|$, we have

$$Pr(|r - \frac{S^R}{M}| \leq \delta|S) \geq Pr(|r - \frac{S}{M}| \leq \frac{\delta}{2}, |\frac{S}{M} - \frac{S^R}{M}| \leq \frac{\delta}{2}|S),$$

because $|r - \frac{S}{M}| \leq \frac{\delta}{2}$ and $|\frac{S}{M} - \frac{S^R}{M}| \leq \frac{\delta}{2}$ will lead to $|r - \frac{S^R}{M}| \leq \delta$, but the inverse is not necessarily correct. Since r and S^R are conditionally independent for given S, we have

$$Pr(|r - \frac{S^R}{M}| \leq \delta|S) \geq Pr(|r - \frac{S}{M}| \leq \frac{\delta}{2}|S) \cdot Pr(|\frac{S}{M} - \frac{S^R}{M}| \leq \frac{\delta}{2}|S).$$

From Property 2, we know that $Pr(|r - \frac{S}{M}| \leq \frac{\delta}{2}|S) \to 1$. In addition, $Pr(|\frac{S}{M} - \frac{S^R}{M}| \leq \frac{\delta}{2}|S) = 1 - e^{-\frac{M\delta\epsilon}{2}} \to 1$. Thus, $Pr(|r - \frac{S^R}{M}| \leq \delta|S) \to 1$. Equivalently,

$$\lim_{M\to\infty} Pr(|r - \frac{S^R}{M}| \geq \delta|S) = 0.$$

We can similarly denote it as $r|S \xrightarrow{P} \frac{S^R}{M}$ when $M \to \infty$. $\square$

Property a4: *The posterior r conditioning on S converges to θ_0 in probability when $n, M \to \infty$. Equivalently,*

$$\lim_{n,M\to\infty} Pr(|r - \theta_0| \geq \epsilon|S) = 0$$

for any $\epsilon > 0$.

Proof:

According to the law of large numbers, we have $\frac{S}{M} \xrightarrow{P} r_0$ when $M \to \infty$ for given n. In addition, by Property a2, $r|S \xrightarrow{P} \frac{S}{M}$ when $M \to \infty$. Since the combination of convergence in probability preserve the convergence, we have $r|S \xrightarrow{P} r_0$. From Property 1, we have $r_0 \to \theta_0$ when $n \to \infty$. Thus, $r|S \xrightarrow{P} \theta_0$ when $n, M \to \infty$. The property is still in the sense that S is considered as a random variable, and for each realization of S, the probability that the posterior r given S is far from the true θ_0 will converge to zero. $\square$

Interpretation of these convergence properties

Since we hide the true S in our framework because of differential privacy, the posterior turns out to be $f(r, S|S^R)$ where r and S are treated as parameters and r is the core quantity used to make conclusions. The closed form for $f(r|S^R)$ is hard to obtain, so we evaluate the convergence property by $f(r|S)$. Though we only have $r|S^R$ in practice, the scale of Laplace noise will be smaller when M gets larger. Thus, the influence of the Laplace part $e^{-\epsilon|s-s^R|}$ in the full likelihood will decay. In addition, the properties above can give us an idea that to which extent and in which sense the subgroup sample size n and the number of partitions M will influence the performance of out methods.

The four properties above give parts of the guarantees to what we expect. First, the Property a2 guarantees for given S, the posterior r will converge to $\frac{S}{M}$ when

$M \to \infty$ fox fixed n. Thus, r will be a better track of $\frac{S}{M}$ when M grows big. Second, the effects of Laplace noise on the posterior distribution of r decays as M grows big, which is supported by Property a3. In addition, Property a4 guarantees the conditional consistency of r, i.e., the posterior of r will converge to θ_0 in probability conditioning on S.

We need to mention that Property a2 to Property a4 above are being conditioned on S. They partially ignore the randomness of S in the model. Though the properties above may not be held strictly when we consider $r|S^R$ rather than $r|S$, they will not be too far away from what we expect. As we will see in the Theorem 1, under a certain approximation, the posterior mean of r will not be far from r_0. Then, we can better approximate the posterior of S by Mr_0 when M gets larger, which in return helps the posterior r conditioning on S^R to be closer to r_0. As a result, we can show that $r|S^R \xrightarrow{P} \theta_0$. We still take the randomness coming from the whole verification process, i.e., S^R is also taken as random in this convergence.

Theorem 1: *Given the conditions that $\hat{\gamma}_{r,n}$ is a consistent estimator of γ, $U(\hat{\gamma}_o; \boldsymbol{\alpha})$ is fixed, and γ is not on the boundary of $U(\hat{\gamma}_o; \boldsymbol{\alpha})$. Under the approximation $S|S^R \sim$ Laplace$(S^R, \frac{1}{\epsilon})$, we have the posterior $r|S^R$ converges to r_0 in probability when $M \to \infty$. In addition, $r|S^R$ converges to θ_0 in probability when $n, M \to \infty$.*

Proof:

First, derive the true posterior distribution of $r|S, S^R$. Refer to the proof of Property a2, we have

$$f(S^R, S) \propto e^{-\epsilon|s^R - s|} \cdot C_M^s \cdot \int_0^1 r^{s+a-1}(1-r)^{M-s+b-1} dr$$

Notice that $(1-r)^{M-s+b-1} = \sum_{k=0}^{M-s+b-1} C_{M-s+b-1}^k r^k (-1)^k$, thus, we have the k-th term of $r^{s+a-1}(1-r)^{M-s+b-1}$ to be $C_{M-s+b-1}^k r^{s+a-1+k}(-1)^k$. In addition, the integral

of the k-th term is

$$
\begin{aligned}
C^k_{M-s+b-1}(-1)^k \int_0^1 r^{s+a-1+k} &= \frac{C^k_{M-s+b-1}(-1)^k}{s+a+k}(-1)^k \cdot (r^{s+a+k}|_0^1) \\
&= \frac{C^k_{M-s+b-1}(-1)^k}{s+a+k}(-1)^k \\
&:= m_k.
\end{aligned}
$$

Thus we have

$$
f(S^R, S) \propto e^{-\epsilon|s^R - s|} \cdot C^s_M \cdot \sum_{k=1}^{M-s+b-1} m_k.
$$

Combine with the full joint distribution, we have

$$
f(r|S^R, S) \propto \frac{f(r, S^R, S)}{f(S^R, S)} \propto r^{s+a-1}(1-r)^{M-s+b-1}\mathbb{I}(0 \le r \le 1),
$$

from which we can get $r|S^R, S \sim Beta(S+a, M-S+b)$.

Then, let's derive $\mathbb{E}(r|S^R)$. Notice that

$$
\mathbb{E}(r|S^R) = \int r \int f(r|s, s^R) f(s|s^R) ds dr = \iint r f(r|s, s^R) f(s|s^R) dr ds,
$$

where $\int r f(r|s, s^R) dr = \mathbb{E}(r|S, S^R) = \frac{S+a}{M+a+b}$. Then, we have

$$
\mathbb{E}(r|S^R) = \int \frac{s+a}{M+a+b} f(s|s^R) ds = \frac{S^R + a}{M+a+b}, \quad f(s|s^R) \propto e^{-\epsilon|s - s^R|}.
$$

For $\mathbb{V}ar(r|S^R)$, we have

$$
\begin{aligned}
\mathbb{V}ar(r|S^R) &= \int (r - \mathbb{E}(r|s^R))^2 \int f(r|s, s^R) f(s|s^R) ds dr \\
&= \iint (r - \mathbb{E}(r|s^R))^2 f(r|s, s^R) f(s|s^R) dr ds.
\end{aligned}
$$

Notice that $\mathbb{E}(r|S^R) \xrightarrow{P} \mathbb{E}(r|S, S^R)$ when $M \to \infty$ in each random trail (partitioning and aggregating process), thus we have

$$\mathbb{V}ar(r|S^R) \xrightarrow{P} \int \mathbb{V}ar(r|S, S^R) f(s|s^R) ds.$$

Recall that $\frac{\mathbb{E}(r|S,S^R)[1-\mathbb{E}(r|S,S^R)]}{M+a+b+1} = \frac{1}{M+a+b+1} \frac{S+a}{M+a+b} \frac{M+a-S}{M+a+b}$. Thus we have

$$\begin{aligned}
\int \mathbb{V}ar(r|S, S^R) f(s|s^R) ds &= \int \frac{Ms + a(M+a) + s^R(s^R + 2s) - (s - s^R)^2}{(M+a+b+1)(M+a+b)^2} f(s|s^R) ds \\
&= \frac{Ms^R + a(M+a) + 3(s^R)^2 - \frac{2}{\epsilon^2}}{(M+a+b+1)(M+a+b)^2} \\
&\xrightarrow{P} 0,
\end{aligned}$$

by which we have $\mathbb{V}ar(r|S^R) \xrightarrow{P} 0$ for each random trail when $M \to \infty$. ϵ is the Laplace parameter. Notice that we use the fact $\frac{(s^R)^2}{M^3} \xrightarrow{P} 0$.

Then, by Chebyshev's inequality, we have

$$Pr(|r - \mathbb{E}(r|S^R)| \geq k|S^R) \leq \frac{1}{k^2} \mathbb{V}ar(r|S^R) \xrightarrow{P} 0, \ k > 0.$$

In addition, we have $\mathbb{E}(r|S^R) \xrightarrow{P} \mathbb{E}(r|S, S^R)$. And thus, $r|S^R \xrightarrow{P} \frac{S+a}{M+a+b}$.

Since we have $\frac{S+a}{M+a+b} \xrightarrow{P} \frac{S}{M}$ by Property a2, and $\frac{S}{M} \xrightarrow{P} r_0$ by law of large number, we have $r|S^R \xrightarrow{P} r_0$. Combine with Property a1, we have $r|S^R \xrightarrow{P} \theta_0$ when $n, M \to \infty$. $\square$

Theorem 1 exactly shows the convergence of our posterior r to the quantity of interest θ_0. It guarantees that, for M that is large enough, the posterior distribution we have for r will have a high probability to be concentrated around r_0, where the randomness comes from each trail of the whole verification procedure. In addition, the essential approximation we make here contains two parts. The first part is that

S and S^R are coming from the same distribution, since we have

$$f(S^R, S) \propto f(S^R|S)f(S) \propto e^{-\epsilon|s^R - s|} \cdot C_M^s \cdot \sum_{k=1}^{M-s+b-1} m_k,$$

from which we know $S|S^R \sim \mathit{Laplace}(S, \frac{1}{\epsilon})$ is actually assuming $f(S^R) \propto C_M^s \cdot \sum_{k=1}^{M-s+b-1} m_k \propto f(S)$. The second part is that we continulize the discretely distributed S, which is a mixed binomial distribution after marginalizing out the uncertainty in r, when we have the posterior $S|S^R$. These two assumptions are more valid when M is large, and we actually have the approximated $S^R \xrightarrow{D} S$ when $M \to \infty$ if we ignore the discretization error, where D indicates convergence in distribution. Thus, up to the approximation error illustrated above, namely, up to a vanishing Laplace noise and continulization error, the posterior r conditioning on S^R converges to θ_0 in probability when $n, M \to \infty$.

4.2 Properties of AM

ν itself is not like the r in the AD framework that has desirable limiting properties. As we will see, it should almost always convergences to 0 in practice. However, it turns out that such undesirable convergence is dependent on n and M by which we can control it. In addition, the speed of convergence to zero can be fairly slow, which makes it easy to come around.

Property b1: *For $\gamma \neq \beta$, if the two equal-tailed confidence intervals are constructed on consistent estimators $\hat{\gamma}$ and $\hat{\beta}$, then $\nu \xrightarrow{P} 0$ when the subset sample size $n \to \infty$.*

Proof:

For consistent estimator $\hat{\gamma}$, we have $\lim_{n\to\infty} Pr(|\hat{\gamma} - \gamma| \leq \epsilon) = 1$ and similar result for $\hat{\beta}$. Denote $\hat{\gamma} - \gamma \sim F_1$ and $\hat{\beta} - \beta \sim F_2$ and thus the two confidence intervals with

equal tail probability are $[\hat{\gamma} - F_1(1 - \frac{\alpha}{2}), \hat{\gamma} - F_1(\frac{\alpha}{2})]$ and $[\hat{\beta} - F_2(1 - \frac{\alpha}{2}), \hat{\beta} - F_2(\frac{\alpha}{2})]$ where $F_{\cdot}(\alpha)$ indicates the α-quantile. Since $F_{\cdot}(1 - \frac{\alpha}{2}) - F_{\cdot}(\frac{\alpha}{2}) \to 0$ when $N \to \infty$, we can just take $\epsilon = \frac{|\gamma - \beta|}{4}$ by which the probability of the right bound of the interval with smaller center is smaller than the left bound of the interval with larger center will goes to 1. Thus, $\nu \xrightarrow{P} 0$. $\square$

This property is not desirable sometimes but it's consistent with our intuition. For larger sample size, we can make more confident statements on the true coefficients. Thus, even though the two coefficients are close to each other, if we can make very confident statements about them based on very large data set, the two confidence intervals will not overlap in general. The most important effect of this property is that n will affect the mean of ν, which leads to a trade-off in choosing appropriate M. In addition, this property can potentially make merely reporting ν to the users not very helpful since it doesn't incorporate the information about the accuracy (standard deviation) of our estimations of the true coefficients, which is one of the reasons for providing additional interpretation for ν in the next section.

Chapter 5

Work flow of the frameworks

We've proposed the frameworks for the same model specification and different model specifications and seen some theoretical properties, but there are still some details need to be addressed. To make it clearer, we'll go through the work flow of the two frameworks in a detailed way in this section.

5.1 Work flow of AD

In this framework, we assume that the model specification in the original study and replication study are the same, and we're interested in the reliability of the original estimation. The nature of the question of interested requires us to conduct the study based on a new confidential data set $\mathbf{D}^* = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. In general, the differential privacy parameter ϵ is pre-specified by exterior factors. The degree of uncertainty δ and the number of partitions M are the parameters that we need to adjust. In addition, we need to propose a meaningful tolerance region U. The choices of these quantities are discussed in this section.

5.1.1 Constructing the tolerance region

Roughly speaking, the tolerance region U should be proposed based on practical need. Depending on whether U is a function of the other parameters, i.e., n, M and other sub-group specific quantities, the region can be classified into fixed region and varying region.

For a fixed region U, it will remain the same no matter the value of other parameters. A very common one could be $[0, +\infty)$ (and its counterpart $(-\infty, 0]$). Con-

ceptually, this region tests the sign of the effect. For example, if we want to check whether the positive effect of a factor is still significantly positive within a different population, we might adopt this one-sided tolerance region. Similarly, $[\hat{\gamma}_o/2, +\infty)$ tests whether the effect size preserved at least a half or not.

Another intuitive way to construct the tolerance region might be based on the original estimation of standard error, i.e., $U(\hat{\gamma}_o; \alpha) = [\hat{\gamma}_o - \alpha \cdot \hat{\sigma}(\hat{\gamma}_o), \hat{\gamma}_o + \alpha \cdot \hat{\sigma}(\hat{\gamma}_o)]$, where $\hat{\sigma}(\hat{\gamma}_o)$ is the estimated standard error of $\hat{\gamma}_o$ in the original study. By choosing an appropriate α, this U can be the estimated confidence interval with any confidence level in the original study. However, a fundamental issue in the partitioning-and-aggregating procedure is the difference in uncertainty between the estimation within each subset and that based on the whole data set in the original study. Especially, in practice, we should have $n_0 >> n$ for most of the time, where n_0 is the sample size of $\mathbf{D}$. Thus, the variance of $\hat{\gamma}_r^{(l)}$ will be much larger than that of $\hat{\gamma}_o$. The difference in the scale of uncertainty can drastically pull each W_l to 0 if U is not wide enough. If U is proposed in a probabilistically meaningful way, e.g., U is the original 95% confidence interval, it can lose the probabilistic interpretation when we come to each subgroup because of the change in uncertainty. It motivates us to adjust U for the difference in uncertainty especially when we propose U based on a probabilistic interpretation, which directly results in the second type of tolerance region. The varying tolerance region is a function of n and n_0.

First, we give two assumptions here. Use $\mathbb{X}$ to denote the design matrix of any full data set and $\mathbb{X}_l$ to denote the design matrix of l-th subset where we partition the full set into M subset evenly. A reasonable assumption is that $M(\mathbb{X}_l'\mathbb{X}_l) \approx \mathbb{X}'\mathbb{X}$ for all $l = 1, ..., M$ under the condition that the size of $\mathbb{X}_l$ is enough to make it representative of the background sampling population.

Assumption 1 (Barrientos et al., 2019): *Use $\mathbb{X}$ to denote the design matrix of*

any full data set and $\mathbb{X}_l$ to denote the design matrix of l-th subset where we partition the full set into M subsets evenly. If we believe each subset has enough size to be representative of the whole population, it approximately holds that $M(\mathbb{X}_l^T\mathbb{X}_l) \approx \mathbb{X}^T\mathbb{X}$ for any $l = 1, ...M$.

Assumption 2: *Use $\mathbb{P}$ to denote the design matrix of the original data set with sample size n_0, and $\mathbb{V}$ to denote the design matrix of the replication data set with sample size N. We believe they're both large enough to be representative of the population and there is no huge heterogeneity of these two data set, the approximation $\frac{n_0}{N}(\mathbb{V}^T\mathbb{V}) \approx \mathbb{P}^T\mathbb{P}$ holds.*

Generally speaking, larger the sample sizes are, the more likely Assumption 1 and Assumption 2 are reasonable. Based on these two assumptions, we have $\sigma(\hat{\gamma}_r) \approx \sqrt{\frac{n_0}{n}}\sigma(\hat{\gamma}_o)$, where $n = [N/M]$ is the subgroup sample size. Thus, a varying tolerance region by inflating $U(\hat{\gamma}_o; \alpha) = [\hat{\gamma}_o - \alpha \cdot \hat{\sigma}(\hat{\gamma}_o), \hat{\gamma}_o + \alpha \cdot \hat{\sigma}(\hat{\gamma}_o)]$ can be proposed as $U^*(\hat{\gamma}_o; \alpha^*) = [\hat{\gamma}_o - \alpha\sqrt{\frac{n_0}{n}} \cdot \hat{\sigma}(\hat{\gamma}_o), \hat{\gamma}_o + \alpha\sqrt{\frac{n_0}{n}} \cdot \hat{\sigma}(\hat{\gamma}_o)]$, where $\alpha^* = \alpha\sqrt{\frac{n_0}{n}}$.

For the choice of the tolerance region U, we suggest to construct it according to the question of interest at first, and be careful with small regions. For example, if we want to test if the replication effect size is within 10% from the published one, we may use $U = [\hat{\gamma}_o \pm \frac{|\hat{\gamma}_o|}{10}]$. However, as illustrated above, the result may suffer from the larger uncertainty of the estimation in each replication subset unless the size of uncertainty is small compares to the effect size. If one wants to test the reproducibility of the model by conducting this verification process on the published 95% confidence interval, she can use the varying tolerance region by inflating the standard error as introduced above. One cautionary note on inflating the standard error is that, if $n > n_0$, namely, we have a super large replication data set, we shouldn't deflate the standard error because we may enlarge the bias between $\hat{\gamma}_o$ and γ given that U is designed to be centered around $\hat{\gamma}_o$. In that case, though it's not likely to happen

often in practice, we'd better choose the M that can make $n \approx n_0$ to make the scale uncertainty the same for previous and replication study.

If there is no preference on different tolerance regions, we suggest to use a one-sided tolerance region, i.e., to verify whether the effect size is larger or smaller than a threshold, especially $[0, +\infty)$ or $(-\infty, 0]$ since they're assessing the sign of a covariate. In addition, as it turns out later, one-sided fixed U has clearer interpretation on the control of the pseudo type-II error, and it will generally suffer less from the increase in the uncertainty of estimation in the replication subset since it's immune from one side of the fluctuation of the estimation. In addition, we suggest using fixed rather than varying tolerance region unless we have particular reasons for that due to the potential change of θ_0 and ambiguity in probabilistic meaning, which will be discussed later.

5.1.2 Choosing the δ in AD

So far, we define $\delta \in (0, 1)$ as the degree of uncertainty that we need to address the conclusion $\theta_N = 1$. Essentially, δ itself doesn't play any role in the validity of the convergence properties in Section 4. However, it does affect the robustness and stability of this method, namely, it affects the power. In addition, the choice of δ is different for fixed and varying tolerance regions.

Varying tolerance region

The δ for varying U is given in an ad hoc way comparing to that of the fixed one. By definition, varying U itself is, at least, a function of n, indicating that for different number of partitions M, we may actually change our ultimate value of interest θ_0. For example, the true γ may fall outside of the original 95% confidence interval (CI), but after inflating this level by a large enough factor $\sqrt{n_0/n}$, the inflated region will cover the γ, under which the θ_0 is actually a function of N and M. This change in

θ_0 is not desirable actually.

As mentioned above, the intuition of constructing a varying region is adjusting for the change in the scale of uncertainty. This adjustment is based on an adjustable fixed region, namely, the region that will not lose the interpretation after being adjusted. For example, the region $[0, +\infty)$ cannot be adjusted. Arguably, being adjustable itself should be based on an interpretation that is meaningful in probability. For instance, an original 95% confidence interval might be given as $U(\hat{\gamma}_o; \alpha) = [\hat{\gamma}_o - 1.96 \cdot \hat{\sigma}(\hat{\gamma}_o), \hat{\gamma}_o + 1.96 \cdot \hat{\sigma}(\hat{\gamma}_o)]$. When we decide to use this U into the replication analysis, unless we care about this single realization of confidence interval (under which we may just use this U for the following analysis without any adjustment), what we are more likely to be interested in is whether the 95% confidence interval will have high coverage of the newly estimated ones given that the scales of uncertainty are the same for the previous estimation and the new ones. This probabilistic interpretation will be broken if we don't inflate the original confidence interval to match the new scale of uncertainty.

The discussion above suggest we select the δ based on the probabilistic interpretation, i.e., it should roughly correspond to the theoretical one that it should have. For example, for the 95% CI given above, theoretically, the mechanism that generates this CI should cover γ under repeated experiment 95% of the times, indicating that we can guess the probability that this single realization covering the truth is 0.95 (though it's not a random event actually). Suppose $U(\hat{\gamma}_o; \alpha)$ covers γ, given that $\hat{\gamma}|\mathbf{X} \sim \mathcal{N}(\gamma, \sigma^2)$ where σ^2 is the variance with the scale corresponding to the size of $\mathbf{X}$, the inflated region $U^* = [\hat{\gamma}_o - 1.96 \cdot \sqrt{n_0/n} \cdot \hat{\sigma}(\hat{\gamma}_o), \hat{\gamma}_o + 1.96 \cdot \sqrt{n_0/n} \cdot \hat{\sigma}(\hat{\gamma}_o)]$ should cover the newly estimated $\hat{\gamma}_r^{(l)}$ with high probability. Even if the γ is on the edge of $U(\hat{\gamma}_o; \alpha)$, after inflation, the bulk of the distribution of $\hat{\gamma}_r^{(l)}$ will still be covered a lot by U^*. After choosing the appropriate δ, i.e., if $\gamma \in U(\hat{\gamma}_o; \alpha)$, then we almost surely

will not make the conclusion that the original 95% CI is unreliable, the pseudo type-I error (the 95% CI is reliable, but we falsely say that it's not reliable) will be roughly controlled at the level of 5%, which is just the probability that the CI generating mechanism doesn't generate a CI that covers the truth.

The ad hoc method to provide this δ is based on the discussion above. Suppose γ is on the edge of $U(\hat{\gamma}_o; \alpha)$ (the left and right edge will give same result), then, calculate the aggregated probability that $\mathcal{N}(\gamma, \frac{n_0}{n} \cdot \hat{\sigma}^2(\hat{\gamma}_o))$ in U^*, and denote it as the reference δ^*. Considering we will have Laplace random noise, as well as the partitioning can also lead to some uncertainty though we have Assumption 2, we suggest choosing $\delta \leq \delta^*$. For example, we can be 10% conservative, and choose $\delta = 0.9 \times \delta^*$.

In addition, the choice of δ will only be necessary when we calculate $Pr(r \geq \delta | S^R)$. In practice, we will have the whole posterior distribution of r, and the analyst can directly look at this distribution to make inference. Calculating $Pr(r \geq \delta | S^R)$ is only one way for interpreting the result, and directly report the posterior of r can also be a way to describe the performance of the interval rule under the new confidential data set.

Fixed tolerance region

Different from the varying tolerance region, the fixed tolerance region itself often doesn't require a probabilistic interpretation, e.g., we don't have a pre-specified expectation on the probabilistic property of $[0, +\infty)$, but instead, we just want to test whether the effect remains positive or not. Technically, once the fixed tolerance region U is given, θ_0 itself will not change no matter how the other parameters change, for which we can more easily analyze the power of this method. As it turns out, for the one-sided fixed U, a robust choice of δ is 0.5.

As evident in Section 4, the validity of the convergences does not depend on δ, but the speed of the convergence depends on it. In the AD framework, we conduct the

partitioning, aggregate the indicator in each partition, and finally, get the observed empirical fraction $r_{obs} = \frac{S^R}{M}$. We conduct the posterior sampling based on this observation and calculate $\hat{\theta}_n = Pr(r \geq \delta | r_{obs}) = Pr(r \geq \delta | S^R)$. Since the partitioning is random, r_{obs} is actually a random variable. If δ is not chosen carefully, e.g., δ is chosen in the bulk of the distribution of r_{obs}, the verification result can vary significantly if we conduct it repeatedly, which is definitely an undesirable situation that can make the single verification conclusion not reliable. Thus, we should choose the δ that is far away from the distribution of r_{obs} to make the verification result stable under repeated verifications. However, for given data set and M, moving δ can actually shift the verification result, but we don't want the conclusion to be subject to subjective choices of δ. To illustrate the choice of δ for fixed tolerance region U, we first need the following assumption.

Assumption 3 (Consistency Assumption): *If one wants to make the conclusion that $\theta_0 = 1$, then she should choose $\delta \geq \frac{1}{2}$.*

Assumption 3 is an intuitive but fundamental assumption. Our goal is to make an excellent guess of the true value of θ_0, such that we should choose the δ that can help us to make the correct conclusion the most easily and definitely. If $\theta_0 = 1$, obviously, the best choice of δ is $\delta = 0$, and $\delta = 1$ for $\theta_0 = 1$. However, it's sneaky and impractical. The Assumption 3 basically requires people to make conclusion based on the majority, i.e., if one branch of a mutually exclusive dichotomous event is correct, then it should happen at least half of the times under repeated experiments.

Property a5: *Under Assumption 3, $\delta = 0.5$ will give the closest $\hat{\theta}_n$ to θ_0 if we make a correct conclusion on the value of θ_0. Namely, we make the conclusion that $\theta_0 = 1$ when $\theta_0 = 1$, and the conclusion that $\theta_0 = 0$ when $\theta_0 = 0$.*

Proof:

Suppose $\theta_0 = 1$. For any given S^R, namely, r_{obs} as defined above, we have

$Pr(r \geq 0.5|S^R) \geq Pr(r \geq \delta^*|S^R)$ where $\delta^* > 0.5$ because of the function $\hat{\theta}_n(x) = Pr(r \geq x|S^R)$ is monotonously decreasing with larger x. In addition, $Pr(r \geq x|S^R)$ is bounded from 0 to 1, and thus,

$$\underset{x \geq 0.5}{\operatorname{argmin}}\{1 - Pr(r \geq x|S^R)\} \iff x = 0.5.$$

It indicates that to minimize the distance between $\hat{\theta}_n(\delta)$ to $\theta_0 = 1$, we should choose $\delta = 0.5$.

Similarly, for $\theta_0 = 0$, we have

$$\underset{x \leq 0.5}{\operatorname{argmin}}\{Pr(r \geq x|S^R)\} \iff x = 0.5.$$

Thus, if we want to make a right guess, $\delta = 0.5$ can make $\hat{\theta}_n$ closest to δ_0. $\square$

Property a5 actually tells us, under any realization of the verification process, namely we observe the r_{obs}, $\delta = 0.5$ will give us the $\hat{\theta}_n$ that obtains the minimal error if we want to make a right guess on θ_0. This guarantee is given in the sense of each realization of the verification process. It also requires us to make a conclusion about the true value of θ_0 such that we have Assumption 3 to avoid the sneaky guess. In contrast, the following property gives us another justification of using $\delta = 0.5$ under the case when we don't necessarily decide to make a conclusion about the value of θ_0, but would rather report the value of $\hat{\theta}_n$ without making a further guess.

Property a6: *$\delta = \delta^o$ where $\mathbb{E}[Pr(r \geq \delta^o|S^R)] = \frac{1}{2}$ is the δ that minimizes $max\{\mathbb{E}[|1-\hat{\theta}_n|], \mathbb{E}[|\hat{\theta}_n - 0|]\}$. The expectation is taken w.r.t. S and the Laplace noise η. In addition, denote random quantity $\xi = \frac{S}{M}$, then any consistent estimator for $\mathbb{E}[\xi]$ is a consistent estimator for δ^o in the sense of $M \to \infty$ for given fixed n, with an approximation error that is introduced by a Laplace noise and the continulization of a mixed Binomial random variable.*

Proof:

First, define the function $L(\hat{\theta}_n) = \mathbb{I}(\theta_0 = 1)\mathbb{E}[|1 - \hat{\theta}_n|] + \mathbb{I}(\theta_0 = 0)\mathbb{E}[|\hat{\theta}_n|]$. We can see L is the average loss function of our guess from the true θ_0. And thus, $max\{\mathbb{E}[|1 - \hat{\theta}_n|], \mathbb{E}[|\hat{\theta}_n - 0|]\}$ actually define the worst case of the average loss over all choices of δ. To minimize it over $\mathbb{E}[|\hat{\theta}_n|]$, it's obvious that $\mathbb{E}[\hat{\theta}_n] = \frac{1}{2}$ is the optimal solution. Then, we need to find the δ such that $\mathbb{E}[\hat{\theta}_n] = \frac{1}{2}$.

According to the proof from Theorem 1, we have $r|S^R, S \sim Beta(S+a, M-S+b)$. Thus,

$$\mathbb{E}_r[r|S, S^R] = \frac{S+a}{M+a+b}, \quad median_r(r|S, S^R) \approx \frac{S+a-\frac{1}{3}}{M+a+b-\frac{2}{3}}.$$

For any n, the two quantities above converge to $\xi = \frac{S}{M}$ in probability when $M \to \infty$.

Let's take $\hat{\delta^o}$ to be any consistent estimator for $\mathbb{E}[\xi]$. Then, we need to show $\mathbb{E}_{S,\eta}[Pr(r \geq \hat{\delta^o}|S^R, S)] \xrightarrow{P} \frac{1}{2}$ when $M \to \infty$. Notice the fact that $median_r(r|S, S^R) \xrightarrow{P} \xi$, and, $\xi \xrightarrow{P} \mathbb{E}\xi$ when $M \to \infty$, from which we get $median_r(r|S, S^R) \xrightarrow{P} \mathbb{E}\xi$. Thus we have

$$\begin{aligned}
\lim_{M\to\infty} \mathbb{E}_{S,\eta}[Pr(r \geq \hat{\delta^o}|S^R, S)] &= \lim_{M\to\infty} \int_S \int_\eta Pr(r \geq \hat{\delta^o}|S^R = s^R, S = s) f(s,\eta) d\eta ds \\
&= \int_S \int_\eta \lim_{M\to\infty} Pr(r \geq \hat{\delta^o}|S^R = s^R, S = s) f(s,\eta) d\eta ds \\
&\xrightarrow{P} \int_S \int_\eta \lim_{M\to\infty} Pr(r \geq \mathbb{E}\xi|S^R = s^R, S = s) f(s,\eta) d\eta ds \\
&\xrightarrow{P} \int_S \int_\eta \frac{1}{2} f(s,\eta) d\eta ds = \frac{1}{2}
\end{aligned}$$

In addition, the posterior distribution of $r|S, S^R$ is exactly the same as $r|S$. The difference between $r|S$ and $r|S^R$ is a Laplace noise and the continulization error as

we introduced in Theorem 1. If we ignore these errors, we have $\lim_{M\to\infty} \mathbb{E}_{S,\eta}[Pr(r \geq \hat{\delta^o}|S^R)] \xrightarrow{P} \frac{1}{2}$, which essentially gives us $\hat{\delta^o} \xrightarrow{P} \delta^o$. $\square$

Property a6 itself doesn't necessarily justify $\delta = 0.5$ since the δ^o is a function of the random mechanism and other parameters. However, it justifies $\delta = 0.5$ in the following sense. Suppose we have a one-sided tolerance region, without losing generality, denote is as $U = [a, +\infty]$. Given that the estimator of the coefficient is symmetrically distributed around the true value, $median(\xi)$ and $\mathbb{E}[\xi]$ will be exactly 0.5 if true $\gamma = a$, i.e., lying on the boundary. If the true γ is not far away from a, we should expect that $\mathbb{E}[\xi]$ will not be far from 0.5. In practice, we never know the true value of γ. If $\gamma << a$ or $\gamma >> a$, then 0.5 shouldn't locate at the bulk of the distribution of ξ, namely, we should get a relatively big or small r_{obs} such that we're not likely to make mistake when making inference about θ_0. The most dangerous situation is that γ is near the boundary of the region such that the center of the distribution of ξ is close to 0.5, where we can easily make a mistake of the inference, or at least have an unstable result. For example, when γ is slightly bigger than a, even the truth is $\theta_0 = 1$, the r_{obs} still has large enough probability to take a small value in a single trial since the bulk of the distribution of ξ can cover a big area in $[0, 0.5]$. There is high probability that we observe a small r_{obs} and make the wrong conclusion that $\theta_0 = 0$. In such case, we would rather result in a moderate $\hat{\theta}_n$ and report this value, rather than get an extreme $\hat{\theta}_n$ but draw a wrong answer.

Thus, Property a6 shows that, as a pre-specified threshold before the verification process, $\delta = 0.5$ is the best one to control the average error around the region where γ is near the boundary. When γ in on the boundary, $\delta = 0.5$ will give the best protection because $\mathbb{E}[\xi] = 0.5$. When γ is getting away from the boundary, $\mathbb{E}[\xi]$ starts to deviate from 0.5, but the practical need of controlling the error is also decreasing. In this sense, $\delta = 0.5$ automatically gives the amount of protection based

on the level of risk in making a serious mistake.

As a conclusion, we classify Property a5 and a6 into two pseudo types of error for the fixed region. Pseudo type-I error means that we fail to make a correct conclusion even if we could have made it because $median(\xi)$ is not close to 0.5 (namely, γ is far from boundary of one-sided U). Pseudo type-II error means that we make a strong false conclusion when the truth is hard to identify (namely, γ is close to boundary of one-sided U). $\delta = 0.5$ is the choice that almost obtains the minimal value in these two types of errors simultaneously. Not strictly speaking, $\delta = 0.5$ gets the largest power in the sense of hypothesis testing, and also obtains the least average error in the sense of making estimation.

For two-sided U, γ being close to boundary of U doesn't necessarily indicate that $median(\xi) \approx 0.5$, such that the region that $\delta = 0.5$ protects may not be the boundary. It will still obtain the largest pseudo power, i.e., make a conclusion with correct direction the most likely to be drawn, but the control on the potential error $max\{\mathbb{E}[|1-\hat{\theta}_n|], \mathbb{E}[|\hat{\theta}_n - 0|]\}$ is weaker. In fact, the region for γ where $median(\xi) \approx 0.5$ is inside the two-sided U, so there is a larger tendency to result in smaller $\hat{\theta}_n$ and make the conclusion that $\theta_0 = 0$. In practice, we suggest be more conservative about small $\hat{\theta}_n$ when we use the two-sided U, and be careful with drawing the conclusion $\theta_0 = 0$.

5.1.3 Choosing the M in AD

For fixed tolerance region

We have shown the guarantees that $r_0 = Pr(\hat{\gamma}_{r,n} \in U(\hat{\gamma}_o; \boldsymbol{\alpha}))$ and θ_n converge to θ_0 with the subset sample size $n \to \infty$ and $r|S \xrightarrow{P} r_0$ when the number of partitions $M \to \infty$. After conducting the verification procedure, we take $\hat{\theta}_n = Pr(r \geq \delta | S^R)$ and use it to make inference about θ_0. The convergences themselves guarantee the

consistency of our estimation for θ_0, but part of the reason that we adopt a threshold δ and take $Pr(r \geq \delta | S^R)$ as the estimation instead of referring to the posterior r directly is the convergences can slow down when n or M are already relatively big. The threshold δ is actually a tolerance that can help with reducing the error between the estimation and the true θ_0 with restriction on the amount of data.

There is a trade-off in M and n. The practical restriction we are faced with is that the sample size N of $\mathbf{D}^*$ is limited, for which we have to strike the balance between M and $n = [N/M]$. n controls the closeness of r_0 and θ_n to θ_0, while M controls the closeness of the posterior r to r_0. From another point of view, n controls the variance of each $\hat{\gamma}_r^{(l)}$ while M controls the variance of S/M and the influence of Laplace noise η. Generally speaking, if M is too small, any single result in each partition matters, as well as we may suffer from large Laplace noise that can obscure the truth. If M is too large, then the variance of $\hat{\gamma}_r^{(l)}$ can explode. Actually we can prove that $\xi = S/M \xrightarrow{P} 0, \frac{1}{2}$ when the variance of $\hat{\gamma}_r^{(l)}$ goes to infinity for two-sided and one-sided tolerance region respectively, which is corresponding to the growth of M. If M is too large, the uncertainty of each $\hat{\gamma}_r^{(l)}$ will also shade the true θ_0. Thus, M should be chosen carefully and moderately.

In addition, the choice of M is based on the choice of δ. For fixed tolerance region, we have shown that $\delta = 0.5$ is a reasonable choice from different perspectives, so we will choose M based on $\delta = 0.5$. Without knowing the true θ_0, we hope the M can provide a robust $\hat{\theta}_n$, namely, based on the chosen δ and M, the verification should remain stable under repeated experiments, though in practice it will only be conducted once. The robustness implicitly requires that a large part of the sampling distribution of r_{obs} is lying on one side of 0.5. For each combination of M and n, the location of the distribution of r_{obs} is intrinsically decided by γ. For one-sided region U, the more far away the γ is from the boundary of U, the more possible that r_{obs} is

far from 0.5. For two-sided region U, this trend is not monotonous around the two boundaries of U, but we still have a similar trend in the large picture. A reasonable way to choose M is that, when γ is leaving from the boundary of U, we choose the M such that a large proportion of the distribution of r_{obs} is lying on one side of 0.5 the earliest. Namely, we make the least requirement on the distance between γ and the boundary of U to make a robust inference. To determine M, we use the following simulation while avoid the use of differential privacy budget.

- Step 1: Suppose the originally published research provide the estimated standard error $\hat{\sigma}(\hat{\gamma}_o)$. Propose a tolerance region U. Ideally, U is one-sided as we suggest.

- Step 2: Give a grid of values for γ and M based on which we do the simulation. For every γ and M in this grid (which corresponds to $n = [N/M]$), simulate $\hat{\gamma}_r^{(l)} \sim \mathcal{N}(\gamma, \frac{n_0}{n}\hat{\sigma}^2(\hat{\gamma}_o))$ for $l = 1, ..., M$. Based on these M many simulated estimations, calculate the S. Then, simulate a Laplace noise $\eta \sim Laplace(0, 1/\epsilon)$ and calculate the empirical $r_{obs} = \frac{S+\eta}{M}$. Repeat this process for K many times. Then, we can get the empirical distribution of r_{obs} for the given γ and M.

- Step 3: Change the choice of γ and M, and repeat Step 2, by which we can make different empirical distributions of r_{obs} under different M and γ. Go through this process for all the combinations of γ and M.

- Step 4: Calculate a summary statistic T for each empirical distribution of r_{obs}, and we can create a contour plot against M and γ. Here, the summary statistics T we suggest is defined in this way. Denote the empirical 10% and 90% quantiles of r_{obs} as $q_{0.1}$ and $q_{0.9}$. If $0.5 \in [q_{0.1}, q_{0.9}]$, define $T = 0$, otherwise $T = distance(0.5, [q_{0.1}, q_{0.9}])$ where distance means the absolute distance. Namely,

if $T > 0$, then the probability that r_{obs} falls in one of the two sides of 0.5 is less than 10%, which we consider as a threshold for robustness.

Examples of this simulated reference for choosing M and the example of the whole procedure will be given in the application in section 7. By referring to the simulated reference, we can see that under the choice $\delta = 0.5$, for which γ we can make a clear and strong conclusion while for which we cannot. Typically, for the one-sided U, for every given M, the contour will reach the mininum when γ is close to the boundary of U, and the contour value will roughly monotonously increase when the γ gets further from the boundary. After getting this contour, we should choose a moderate M that can make the contour value increase and becomes larger than 0 relatively fast.

For varying tolerance region

Unlike the fixed one, the choice of M here tends to be more subjective and ad hoc comparing to that of the fixed region. For varying region, the speed of the convergence doesn't matter that much, since we inflate the tolerance region to make it match the scale of uncertainty under sample size n. If the Assumption 2 holds, S/M shouldn't vary a lot with different M. Thus, the choice of M is much more heuristic.

First, we need to consider the space of S/M. If $M = 10$, then $\frac{S}{M}$ can only take values in $\{10\%, 20\%, ..., 100\%\}$, which can be too crude. We suggest to pick the M that at least have the space as delicate as 0.05, namely, $M \geq 20$. Another problem that we should consider is the Laplace noise. If we take the same criterion as before, i.e., to keep the influence of Laplace noise on r_{obs} less than 5% under $\epsilon = 1$, we need to pick at least $M = 40$.

In addition, we need to decide under which M we still feel comfortable with the Assumption 1 and Assumption 2. The performance of these two approximation doesn't merely depend on M, but it depends more on size of each subset. Namely, it

depends on whether the subset can still be representative of the full set. Given the prerequisite that we still feel comfortable with these assumptions, we suggest to use a M that is not too small, e.g., $M = 50$, to get a moderately delicate space of S/M as well as reduce the influence of Laplace noise.

5.1.4 MCMC sampling and interpretation

After choosing the M, the next step is to conduct the MCMC sampling by which we can get a set of posterior samples $\{(r_{(t)}, S_{(t)})\}_{t=1}^{T}$. To make inference based on the posterior samples, we can compute the empirical posterior probability $\hat{Pr}(r \geq \delta | S^R) = \sum_{t=1}^{T} \mathbb{I}(r \geq \delta)/T$ as the estimation of θ_n by which we make inference on θ_0.

As discussed above, for fixed region, we pick $\delta = 0.5$, while for varying region, we provide an ad hoc method to choose δ. Since the mechanism of proposing these two δ's are different, the interpretation tend to be a little different.

For fixed region, we no longer interpret δ as the degree of uncertainty that we need to make the conclusion. Instead, we're making conclusion based on the majority. If $\hat{\theta}_n$ takes an extreme value, i.e., 0.9, then we can feel comfortable to make the conclusion that $\theta_0 = 1$, and vice versa for small $\hat{\theta}_n$. If it takes a moderate value, especially around 0.4 to 0.6, we don't suggest to make any guess about the true value of θ_0. We can be conservative and claim that the verification is not perfectly passed, but there is still no strong evidence that $\theta_0 = 0$, and vice versa.

For varying region, δ should be interpreted as an expected approximated confidence level. We suggest to report the δ^*, δ and $\hat{\theta}_n$ together. If δ takes the influence of Laplace noise as well as the background δ^* into consideration, namely, δ is at least smaller than δ^* by an extreme absolute Laplace noise, we should expect a large $\hat{\theta}_n$, e.g., 0.8, after tolerating the randomness in the posterior sampling. The more useful aspect of this method for varying region should be claiming that the previous model

is not valid under the new data, namely we 're rejecting the null. If we set the δ that is tolerant enough, but result in $\hat{\theta}_n < 0.5$, we should reject the null.

5.2 Work flow of AM

In this framework, there are two different model specifications for the published study and replication study. We're interested in whether the results made by these two models are different from each other in a considerable scale. The replication study will be conducted on a confidential data set $\mathbf{D}^* = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, which can be a new one, or the same one as the published study. Again, the differential privacy parameter ϵ should be pre-specified by exterior requirements. For the confidence interval, we fix it to be the 95%-confidence interval with equal tail probabilities, since lower confidence level can potentially make the probabilistic statement made by the ν, the overlap of two intervals, very weak, because it will only describe a small amount of the uncertainty about the estimations of γ and β. After choosing the number of partition M, we will be able to step into the posterior sampling and make inference based on the posterior samples of $\bar{\nu}$.

Two questions remain to be answered. The first is the choice of M, and the second is the interpretation of the posterior distribution of $\bar{\nu}$. In this section, they are discussed in detail.

5.2.1 Choosing the M in AM

Generally speaking, we confront the similar issues as those in the AD framework. As we choose larger M, the average overlap $\bar{\nu}$ will be computed based on more partitions, which will potentially decrease the uncertainty of it, and reduce the influence of Laplace error. However, the point estimations made based on each partition will be less stable. In addition, the length of confidence intervals will increase for larger M

because of the larger standard deviation under smaller partitions. Considering all the factors above, for larger M, we tend to have more stable $\bar{\nu}^L$, but we will lose the sensitivity in detecting $|\beta - \gamma|$, i.e., for a fixed change in $|\beta - \gamma|$, the expected change in the overlap will be smaller. The M that we suggest to use is one that can control the influence of the Laplace noise and the sampling variance of $\bar{\nu}$ conditioning on the data we have to a reasonable scale, as well as preserve enough sensitivity in detecting the change of $|\beta - \gamma|$.

With the chosen M and the prior ψ_0, which can naturally be a Beta distribution, we can directly sample from the posterior distribution

$$p(\bar{\nu}|\bar{\nu}^L) \propto e^{-M\epsilon|\bar{\nu}-\bar{\nu}^L|} \cdot \bar{\nu}^{a-1}(1-\bar{\nu})^{b-1} \cdot \mathbb{I}(0 \leq \bar{\nu} \leq 1).$$

The part $\bar{\nu}^{a-1}(1-\bar{\nu})^{b-1} \cdot \mathbb{I}(0 \leq \bar{\nu} \leq 1)$ is the prior Beta kernel, while $e^{-M\epsilon|\bar{\nu}-\bar{\nu}^L|}$ is a Laplace kernel centered around $\bar{\nu}^L$ and decays exponentially. Thus, this posterior distribution can be viewed as a weighted Beta distribution with the weight centered at $\bar{\nu}^L$ and decays exponentially when the value departs from this center.

A decomposition on the sampling ν

As shown before, if we generate ν from the underlying generating mechanism, we will have $\nu \xrightarrow{P} 0$ when $n \to \infty$ if $\beta \neq \gamma$, which is a condition that almost definitely holds in practice. Thus, n is exactly influencing $\mathbb{E}[\nu]$, so we can denote the expectation as $\mathbb{E}[\nu|n]$ to indicate it's a function of n. By changing the number of partitions, we're changing the distribution of ν. Since we're conducting the replication study on $\mathbf{D}^*$ with sample size N, the expecation of the overlap under the full data is $\mathbb{E}[\nu|N]$. However, in order to reduce the influence of Laplace noise as well as the uncertainty in the observed ν, we adopt the partitioning-and-aggregating procedure to obtain a sample mean of ν. As a directly result of this procedure, we're working with ν under sample size $n = [N/M]$ and the new expectation of interest $\mathbb{E}[\nu|n]$. If we view $\mathbb{E}[\nu|n]$

as a substitute for $\mathbb{E}[\nu|N]$, and $\bar{\nu}$ is calculated under the subset sample size n, then we have

$$\bar{\nu}^L - \mathbb{E}[\nu|N] = (\bar{\nu}^L - \mathbb{E}[\nu|n]) + (\mathbb{E}[\nu|n] - \mathbb{E}[\nu|N]) := error + bias$$

The decomposition above is similar to the bias-variance trade-off, which shows the way that $\bar{\nu}$ departs from $\mathbb{E}[\nu|N]$. As for the bias term $\mathbb{E}[\nu|n] - \mathbb{E}[\nu|N]$, generally speaking, it increases with larger M. For the error term $\bar{\nu}^L - \mathbb{E}[\nu|n]$, if n is fixed, then it converges to zero in probability when $M \to \infty$, but here n decreases when M gets larger, which shades this convergence.

Different from the traditional bias-variance trade-off, we're not necessarily minimizing the sum of error and bias. In practice, we have $\bar{\nu}^L$ under the sample size n as an approximation for $\mathbb{E}[\nu|n]$. Then, we make inference based on posterior $\bar{\nu}$, which can be viewed as the compromise of our prior belief on the overlap and the estimated $\mathbb{E}[\nu|n]$ after taking the uncertainty introduced by Laplace error into account. We can interpret the result in the sense of overlap, or build a further step to relate the difference between γ and β with the posterior $\bar{\nu}$. The interpretation isn't necessarily based on sample size N, and we can conduct this step based on n instead. In this sense, we're not actually require that $\bar{\nu}^L$ is a good approximation to $\mathbb{E}[\nu|N]$.

This decomposition provides us an insight in selecting the M. First, we hope that the error term will not be very large, such that $\bar{\nu}^L$ can be representative of $\mathbb{E}[\nu|n]$, and the verification process will be more robust. Secondly, the bias term represents how sharply we're shifting the distribution of ν. The more we shift the distribution of ν, the less sensitive ν will be in detecting a certain change in $|\beta - \gamma|$ in general.

Selecting the M

Given the decomposition above and the potential loss in sensitivity if we have large bias, we cannot choose the M as large as possible. Instead, we should choose a smaller M that is enough for getting a stable $\bar{\nu}$ as well as reduce the effect of

Laplace noise. Fortunately, since $\nu \in [0,1]$, we have $max\ \mathbb{V}ar[\nu] = \frac{1}{4}$, so that we can bound $\mathbb{V}ar[\bar{\nu}]$ by $\mathbb{V}ar[\bar{\nu}] \leq \frac{1}{M}\mathbb{V}ar[\nu] \leq \frac{1}{4M}$ no matter the n we have. In addition, as $\bar{\nu}^L = \bar{\nu} + \eta$ where $\eta \sim Laplace(0, \frac{1}{M\epsilon})$ and η is independent with ν, we have $\mathbb{V}ar[\bar{\nu}] \leq \frac{1}{4M} + \frac{2}{M^2\epsilon^2}$, Based on this upper bound and Chebyshev's inequality, we have

$$Pr(|\bar{\nu}^L - \mathbb{E}[\nu|n]| \geq \omega) \leq \frac{1}{\omega^2}(\frac{1}{4M} + \frac{2}{M^2\epsilon^2}).$$

In most of the practices, this upper bound for the error $|\bar{\nu}^L - \mathbb{E}[\nu|n]|$ should be very loose, since it's not likely to have the density of ν being concentrated at the edges of $[0,1]$. If the standard error of $\hat{\beta}$ and $\hat{\gamma}$ are similar, they should provide the confidence interval with similar lengths, such that the distribution of ν will be likely to be unimodal. In addition, the term $\frac{1}{M^2+\epsilon^2}$ goes to zero very quickly when M grows, which can even be ignored if M is large enough. Overall, this upper bound is given in the sense of controlling the worst case, but we can expect much better in practice.

Our goal in choosing M is to select the M that can control $|\bar{\nu}^L - \mathbb{E}[\nu|n]|$ under acceptable probability, while trying to preserve more sensitivity of ν. Naturally, the selection of M will be related to the next section where we build the connection between $\mathbb{E}[\nu|n]$ and the difference between γ and β. Here, we suggest setting a reasonable M at first, and then, checking whether it's sensitive enough for practical application. For example, if we're comfortable with assuming ν is evenly distributed, the choice $M = 25$ can control $|\bar{\nu}^L - \mathbb{E}[\nu|n]|$ under 0.2 with probability larger than 0.9. If ν is unimodal, we can expect smaller $|\bar{\nu}^L - \mathbb{E}[\nu|n]|$. The influence of Laplace noise can be directly checked by the variance of the noise, and the requirement of controlling it can be subject to the users. Then, we can go forward to see if $M = 25$ will preserve reasonable sensitivity. If we want better control of $|\bar{\nu}^L - \mathbb{E}[\nu|n]|$ and the Laplace noise, we can choose $M = 50$, and check the whether enough sensitivity is preserved. The choice of M can be done in a comparative way, i.e., for a given set

of M that can satisfy our need in controlling the error, we can check their sensitivity and pick the best one.

5.2.2 Interpretation and reference contour plot

There is still one step left for us to build, which is the connection between $|\gamma - \beta|$ and ν, or the characteristics of the distribution of ν. Theoretically, it's not easy to interpret ν directly without any additional assumptions. A single ν itself can be viewed as a quantity evaluating the distance between two interval rules, but it's hard to interpret it directly in the sense of probability (Karr et al, 2006), and misuse of this overlap is often happening (Knol et al, 2011). Even if we take a pseudo approximated interpretation, i.e., interpret an estimated 95%-CI as having probability of covering the true coefficient to be 95%, it's still hard to interpret ν even when $\nu = 1$. In addition, this pseudo probabilistic guarantee can weaken fast while ν gets smaller. The consequence of the vague meaning of ν is that we can hardly have any prior-to-experiment expectation on it: which values ν will likely take for a certain difference between the two true parameters and the uncertainty in estimations.

For some situations, it's acceptable to leave the results as the posterior ν. For example, some previous studies may provide prior knowledge on the overlap of confidence intervals between different models. It can also be the case where we don't need to translate ν into the difference between coefficients, e.g., we're comparing the similarities between several models, and the overlap can be used as the measure for the similarity between models. For other cases where we need to build the connection between ν and $|\beta - \gamma|$, we propose two ways to make this translation. The first one is the reference contour plot, which we suggest to choose M as well as to take it as a rough reference rather than making an estimation for $|\beta - \gamma|$. The second is the inverted credible interval given additional assumptions on the standard errors, by

which we can exactly get the posterior credible intervals for $|\beta - \gamma|$.

Reference Contour Plot

The reference contour plot is used to build the bridge between $\mathbb{E}[\nu|n]$ and $|\gamma - \beta|$. In this simulation, we will have a grid of values for different $|\gamma - \beta|/|\gamma|$ and $\sigma(\hat{\gamma})/\sigma(\hat{\beta})$. Under each combination of these two quantities, we simulate $\hat{\gamma}$, $\hat{\beta}$ and the two confidence intervals directly from two normal distributions repeatedly, and then calculate the average overlap measure. The mathematical description of the simulation is provided as follow.

For given data set, we assume the corresponding estimations

$$\hat{\boldsymbol{\gamma}} \sim \mathcal{N}(\boldsymbol{\gamma}, \Sigma_0), \quad \hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \Sigma_1)$$

For the variable of interest X, we can specify

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix} \sim \mathcal{N}(\begin{pmatrix} \gamma \\ \beta \end{pmatrix}, \begin{pmatrix} \mathbb{V}ar(\hat{\gamma}) & \mathbb{C}ov(\hat{\gamma}, \hat{\beta}) \\ \mathbb{C}ov(\hat{\beta}, \hat{\gamma}) & \mathbb{V}ar(\hat{\beta}) \end{pmatrix}).$$

The value of γ is specified as $\hat{\gamma}_o$, and $\sigma(\hat{\gamma})$ is specified as $\hat{\sigma}(\hat{\gamma}_o)\sqrt{n_0 \cdot n/M}$, i.e, the reference is based on and in comparison with the published result. We will take $\frac{|\gamma-\beta|}{|\gamma|}$ and $\frac{\sigma(\hat{\beta})}{\sigma(\hat{\gamma})}$ as the two axes of the contour plot. The grid of axes can be specified by the researchers as they like to address their interests, e.g., the grid containing $|\gamma - \beta|/|\gamma| = 1$ can be used to test the sign of the effect.

Then, for every given pair $(\frac{|\gamma-\beta|}{|\gamma|}, \frac{\sigma(\hat{\beta}_k)}{\sigma(\hat{\gamma}_k)})$, we generate $(\hat{\gamma}, \hat{\beta})$ from the bivariate normal distribution above and the corresponding confidence intervals based on normal distribution, calculate ν, and then repeat this procedure for K times and return $\bar{\nu}$ and show its value in the corresponding grid in the contour plot. Figure 5.1 provides several examples of the reference plots with different parameters.

For the correlation $corr(\hat{\gamma}, \hat{\beta})$, it should be specified by the replication researchers such that there will not be privacy loss. The global sensitivity of the estimation for

the correlation has an upper bound of 2. The estimated correlation can actually change from 1 to a negative value very close to -1 due to the high sensitivity to outliers. Thus, if we want to estimate the correlation as well as protect the privacy, we have to add a large Laplace noise which can likely make the estimation senseless. It's also a reason that blocks us from using the reference contour plot for making point estimation on $|\gamma - \beta|/|\gamma|$, since the correlation cannot be estimated accurately without privacy loss. For many practical applications, $corr(\hat{\beta}, \hat{\gamma})$ should be very high, since $\hat{\beta}$ and $\hat{\gamma}$ obtained under the same subsample should be correlated in some ways. According to our current simulations, we often get the estimated correlations to be larger than 0.9. Though it's not theoretically guaranteed, we can still expect that we should have high correlation in practice. Typically, as shown in Figure 5.1, high correlation will make the contours more discriminating, i.e., contour value will change more sharply for a given change in $|\gamma - \beta|$. In contrast, the least discriminating plot is obtained under no correlation, i.e., $corr(\hat{\beta}, \hat{\gamma}) = 0$.

In addition, for given ratio in the standard deviation, the average overlap will be roughly decreasing when the relative difference between the two coefficients increases. By referring to the contour plot, the researchers can also have an overview of the extend to which the difference in standard deviation will influence the overlap.

The simulated reference is based on directly sampling the fitted coefficients from a bivariate normal distribution. The two confidence intervals are also based on normal distributions. The normal assumption is based on conditioning on the given data, so it is reasonable as long as we are comfortable with Assumption 1 and Assumption 2. In practice, the standard error is unknown, which will result in confident intervals based on Student's t-distribution. However, as long as Assumption 1 holds, the normal-based confidence interval will not be far away from that one based on Student's t-distribution since $\hat{\sigma} \xrightarrow{P} \sigma$ when $n \to \infty$, where σ can be $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$. Assumption

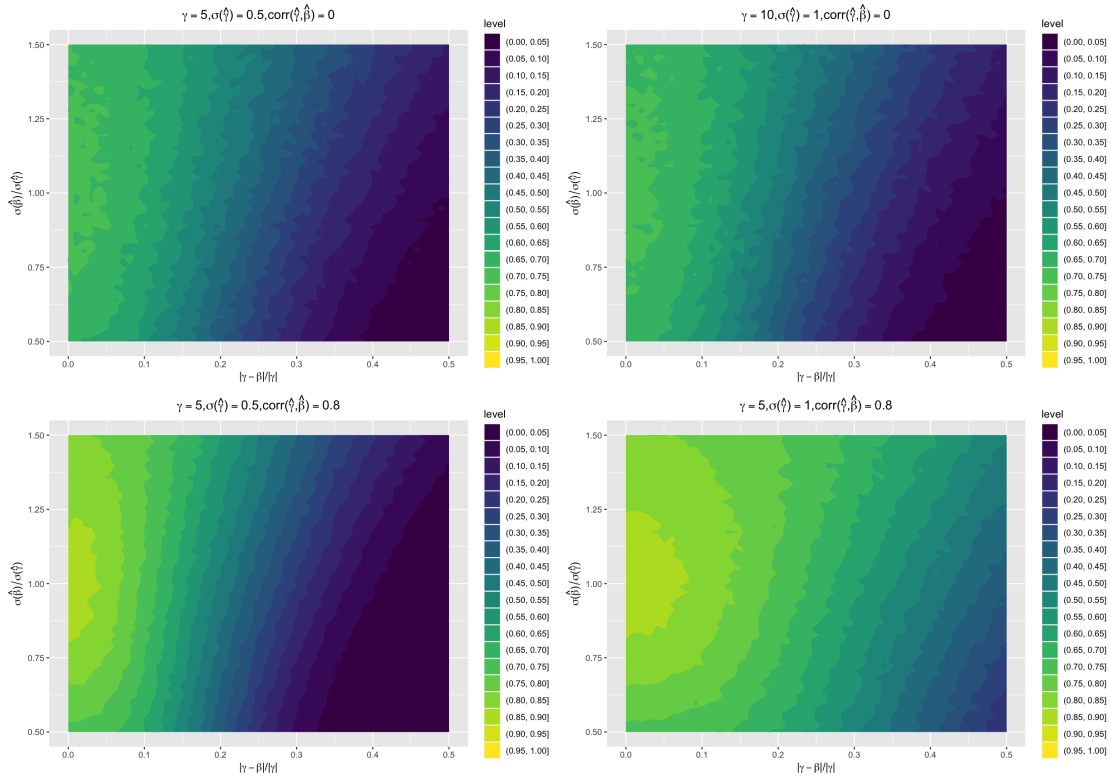

Figure 5.1: Examples of the simulated contours of $\bar{\nu}$. Every contour value, the $\bar{\nu}$, is the average of 500 randomly sampled ν.

1 implicitly guarantees the stability of the estimator $\hat{\sigma}$ when estimating the true standard deviation.

For the choice of the axes, we take one of the axis to be the relative difference between γ and β for better control of the plot and clearer visualization. We can clearly see how the overlap will be theoretically if β is different from γ to a certain extent. This choice of axis can also make the reference plot nearly invariant to the scale of the actual coefficients. As shown in Figure 5.1, the pair $(5, 0.5)$ and $(10, 1)$ for γ and $\sigma(\hat{\gamma})$ almost lead to the same reference. In addition, for $|\gamma - \beta| = \gamma - \beta$ and $|\gamma - \beta| = -\gamma + \beta$, the sampling distribution of ν will be exactly the same, so we take the absolute difference $|\gamma - \beta|$ rather than the raw $\gamma - \beta$.

Inverted function for the overlap

The method in this section gives a simple further step to translate ν directly into $|\beta - \gamma|$. Under additional assumptions on $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$ such that we have the length of the two confidence intervals, we can easily invert ν into the difference $|\beta - \gamma|$. Suppose that we've assumed the exact values of $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$. Then, denote the length of the two intervals to be l_1 and l_2. Denote the length of the overlap part is c. Without losing generality, suppose $l_1 \geq l_2$.

By the inverting definition of overlap measure ν in section 3.2, we have

$$c = \frac{2\nu}{1/l_1 + 1/l_2}.$$

Then, the overlap length c can be expressed as

$$c = \begin{cases} 0, \ |\beta - \gamma| \geq \frac{1}{2}(l_1 + l_2) \\ \frac{1}{2}(l_1 + l_2) - |\beta - \gamma|, \ |\beta - \gamma| \geq \frac{1}{2}(l_1 - l_2) \\ l_2, \ 0 \leq |\beta - \gamma| \leq \frac{1}{2}(l_1 - l_2) \end{cases} .$$

Combine the equations above, the inverted function will be given as

$$f := \nu \mapsto |\beta - \gamma| : \begin{cases} |\beta - \gamma| \geq \frac{1}{2}(l_1 + l_2), \ \nu = 0 \\ |\beta - \gamma| = \frac{1}{2}(l_1 + l_2) - \frac{2\nu}{1/l_1 + 1/l_2}, \ 0 < \nu < \frac{1}{2}(1 + \frac{l_2}{l_1}) \\ |\beta - \gamma| \leq \frac{1}{2}(l_1 - l_2), \ \nu = \frac{1}{2}(1 + \frac{l_2}{l_1}) \end{cases} .$$

If we further define that, $f(\nu) = \frac{1}{2}(l_1 + l_2)$ when $\nu = 0$ and $f(\nu) = \frac{1}{2}(l_1 - l_2)$ when $\nu = \frac{1}{2}(1 + \frac{l_2}{l_1})$, it will be a function mapping each ν to a single real value. If we map the posterior credible interval of $\bar{\nu}$ by this mapping rule, we will get an interval that can be interpreted as the expected credible interval of $|\beta - \gamma|$ averaging on independent trials of the verification process under the assumptions on l_1 and l_2. By referring to this expected credible interval, we can provide an exact bound for the difference of the two coefficients with fine probabilistic interpretation.

Theoretically, people can make any assumption on $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$. However, we may feel uncomfortable with making an arbitrary assumption on these two quantities.

A useful assumption is $\sigma(\hat{\beta}) = \sigma(\hat{\gamma}) = \hat{\sigma}(\hat{\gamma}_o)$ under the sample size n_0, and we call is as the null assumption. This assumption is roughly saying that the two models perform similarly in practice, and the published result is validated. We can tailor this null assumption into several perspectives.

Firstly, it's a direct requirement of similarity in practical performance of the two candidate models. Even if $\beta = \gamma$, to get similar result from these two models in practice, we still need them to have similar efficiency, i.e., the alternative model will not suffer from significantly larger variance in the estimation, otherwise we can still likely get two different results from the two models when we apply them simultaneously. Secondly, for the exact value of the standard deviation, the previously estimated $\hat{\sigma}(\hat{\gamma}_o)$ is likely the best reference we can have without using the confidential data. By taking $\sigma(\hat{\beta}) = \sigma(\hat{\gamma}) = \hat{\sigma}(\hat{\gamma}_o)$ under the sample size n_0, we're somewhat assuming the homogeneity of the effect of interest across $\mathbf{D}$ and $\mathbf{D}^*$, by which we can indirectly test the published result. If the change in the model specification is somewhat unrelated to the effect of interest, we should expect that the uncertainty in estimating the targeted effect will not change a lot when we take the alternative model as long as no computational issue arises.

If the verification result indicates significant difference of the two models, we will have all possible interpretations implying the unreliability or sensitivity of the original study to the model specification in some ways. For example, if the posterior 90% credible interval is $[500, 800]$, and the previous estimated γ is 200, we can conclude that the two models may give significantly different estimated effects. Otherwise, to overturn this conclusion, at least one of the assumptions, i.e., similar efficiency of the two models or reliability of the previously estimated uncertainty quantification, is invalid.

On the other hand, if the posterior credible interval for $|\beta - \gamma|$ under the null

assumption indicates small difference comparing to $\hat{\gamma}_o$, we're not likely to underestimate the difference of these two model specifications, as long as we're not seriously underestimating $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$. It's easy to see the inverted mapping $\nu \mapsto |\beta - \gamma|$ is linear. To seriously underestimate $|\beta - \gamma|$, at least we need to underestimate l_1, i.e., one of $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$. It's equivalent to saying that the uncertainty quantification provided by $\mathbf{D}^*$ is much higher than $\mathbf{D}$. Thus, if the two models kind of pass the verification process after we check the credible interval for $|\beta - \gamma|$ under the null assumption, we can make the interpretation that one of the following statement is true: there is no significant difference in β and γ, or the uncertainty quantification provided by $\mathbf{D}^*$ is much higher than $\mathbf{D}$.

If $\mathbf{D} = \mathbf{D}^*$, we can even get finer interpretation. According to the illustration above, part of the difference between γ and β can be attributed to the potentially invalid assumption on $\sigma(\hat{\beta})$ and $\sigma(\hat{\gamma})$, and further be attributed to the heterogeneity across $\mathbf{D}$ and $\mathbf{D}^*$. Under the case where $\mathbf{D} = \mathbf{D}^*$, there is no heterogeneity across data sets, such that the assumption on $\sigma(\hat{\gamma})$ has to be correct as long as Assumption 1 holds. Then, all the difference between the two coefficients can be attributed to the two models themselves, which can significantly improve the strength of our statements on $|\beta - \gamma|$.

Chapter 6

Empirical Illustrations on AD framework

In this section, we use simulation studies to show some convergence properties in section 4 as well as demonstrate the overall performance of AD method. Larger M is desirable in the sense of controlling the Laplace noise, as well as lowering the sampling variance of r if other factors are fixed. But with modest N, large M will also suffer from small sample size in each partition, which results in larger variance of the estimated coefficients. Here, we illustrate how the core quantities in our framework can change with ϵ, M and n. Specifically, we aim to demonstrate that the influence of Laplace noise decreases with larger M. In addition, we will provide an example to show how the posterior r will be effected by M and n. For better control of the simulation study, this part will be based on where we mimic conducting the replication study on a new data set.

In this simulation, we assume the original research is conducted on $\mathbf{D}$ and the replication study is on $\mathbf{D}^*$. Both of them contain three predictors X_1, X_2 and X_3. For both $\mathbf{D}$ and $\mathbf{D}^*$, we sample these variables from the following distributions: $X_1 \sim Uniform[0, 10]$, $X_2 \sim \mathcal{N}(5, 1)$ and $X_3 \sim Bernoulli(0.5)$. For $\mathbf{D}$, we assume $Y|X_1, X_2, X_3 \sim \mathcal{N}(2X_1 + X_2 + 3X_3, 3^2)$. For $\mathbf{D}^*$, we assume $Y|X_1, X_2, X_3 \sim \mathcal{N}(2X_1 + 0.9X_2 + 3X_3, 3^2)$. The simulated data sets are designed in this way mainly for convenience and illustration purpose. $\mathbf{D}$ contains $n_0 = 10,000$ samples while we don't limit the size of $\mathbf{D}^*$ right now for the purpose of convenience. Suppose we're interested in the effect of X_2, and we want to see if the estimation on the effect size of X_2 in the original study is generalized, i.e., the effect size doesn't vary across different data sets obviously. Based on the simulation of $\mathbf{D}$, we have $\hat{\gamma}_o = 0.971$, and

the estimated standard error $\hat{\sigma}(\hat{\gamma}_o) = 0.031$.

The parameters we can adjust are ϵ, M, n and the tolerance region U. We take $U = [\hat{\gamma}_o - \alpha \cdot \hat{\sigma}(\hat{\gamma}_o), \hat{\gamma}_o + \alpha \cdot \hat{\sigma}(\hat{\gamma}_o)]$. There are two versions of α: $\alpha = \alpha_0$ and the inflated $\alpha = \alpha_0\sqrt{n_0/n}$. We can adjust the results of simulation simply by tuning α. Here, we pick $M \in \{10, 50, 90\}$ and $n \in \{10, 20, ..., 1000\}$, meaning that for every simulation result with specified M and n, we need a data set exactly with $M \times n$ samples. For each setting of parameters, we conduct the sub-sample and aggregate procedure to generate S^R followed by MCMC to obtain the posterior samples, and repeat this process for 20 times. For each MCMC sampling, the burn-in sample size is 200, after which we have 1000 posterior samples. For each set of posterior samples of r, we calculate its median, meaning that we will have 20 medians for each (M, n), and then we plot the posterior medians of r against subset sample size n for each M. In order to be more informative and reliable, we use the medians of posterior samples from repeated trails of simulation instead of the posterior samples from a single trail.

6.1 Consistency of r_0

Property a1 states that $r_0 \to \theta_0$ if $\hat{\gamma}_{r,n}$ is a consistent estimator, and such convergence happens when the sample size n goes to infinity. r_0 itself is interpreted as the probability parameter of the Bernoulli variable W, and it is estimated by the empirical $\frac{S}{M}$. Under the DP context and the hierarchical structure, r is used as a noisy version of r_0 which is shown to converge to $\frac{S}{M}$ when M grows large. Not strictly specking, the random quantity r should be centered around r_0, and the center (median) of the distribution of r demonstrates how well r_0 is converging to θ_0. In some senses, it determines the confidence and ability we have in differentiating whether $\theta_0 = 1$ or not.

In this simulation, we know the true value of θ_0 such that we can evaluate how

well we're approximating the answer. Since we want to show the changes in the distribution of r, we take $\epsilon = 1$ to avoid being overwhelmed by the Laplace noise. For the tolerance parameter α, we deliberately choose it as $\alpha = 1, 3, 6$ to show different situations. For $\alpha = 1$, the true γ, which is 0.9, is outside of the tolerance region. For $\alpha = 3$, γ is inside the tolerance region, but very close to the boundary. For $\alpha = 6$, γ is well inside the tolerance region.

Results of this simulation are shown in Figure 6.1. As we can see, the posterior medians of r become less sparsely distributed when M grows big, which is consistent to the fact that larger M can provide more refined description of r_0. However, the improvement made by larger M becomes much less obvious when $M \geq 50$, since the space we take in $[0, 1]$ when $M = 50$ is already reasonably fine at 0.02. In addition, for every M and α, the posterior median of r starts from a small value. It's corresponding to the fact that the standard error of estimation can be extremely large when the sample size is small. As n grows bigger, the standard error of estimation well get smaller, such that the estimation will be more stable and concentrated around the true coefficient. Thus, at the early stage when n grows bigger, even in the case where $\theta_0 = 0$, we can see the overall increase in the posterior median of r. However, when n continues growing, the medians will be closer to 1 if $\theta_0 = 1$ and closer to 0 if $\theta_0 = 0$ as guaranteed by the Property a1.

We can see such trend happens according to the simulation, but the speed is subject to the distance between γ and the nearest boundary of the tolerance region and the standard error of $\hat{\gamma}$. Under ordinary linear regression assumptions, the standard error will decrease with the rate $1/\sqrt{n}$, so the marginal improvement of larger n will decrease when n gets larger. For the case where the true γ is lying near the boundary of tolerance region, i.e., the case where $\alpha = 3$ and $\theta_0 = 1$, the posterior median of r seems to stick around a moderate uninformative value because we need

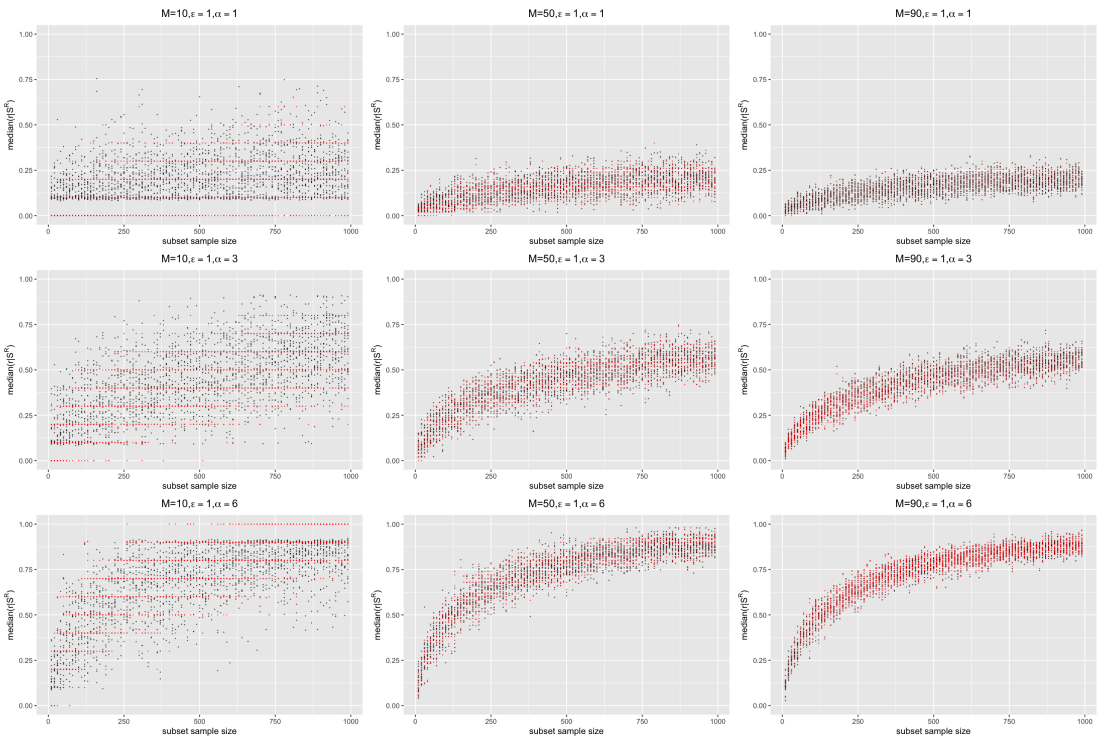

Figure 6.1: Simulated posterior median of r without inflation. . Red points are empirical $\frac{S}{M}$, black points are the posterior median of r.

a large amount of data to identify whether γ is inside the tolerance region and the convergence takes place slowly. For a typical research in practice under which we will have at least thousand samples, we can expect that by choosing a reasonable M and $\delta = 0.5$, we will be able to make the conclusion that we should make, while keeping conservative in the ambiguous case.

Figure 6.2 shows the situation where we inflate the tolerance region to the scale of uncertainty under subgroup sample size n, i.e., replace α by $\alpha \times \sqrt{n_0/n}$. We incorporate it here for the purpose of better illustration and comparison with the previous one. Under this situation, the true θ_0 is actually changing when n is getting larger, so we're not focusing on consistency, but focusing on the actual value of r and the influence of M.

As illustrated in section 5, inflating the tolerance region may be done when we

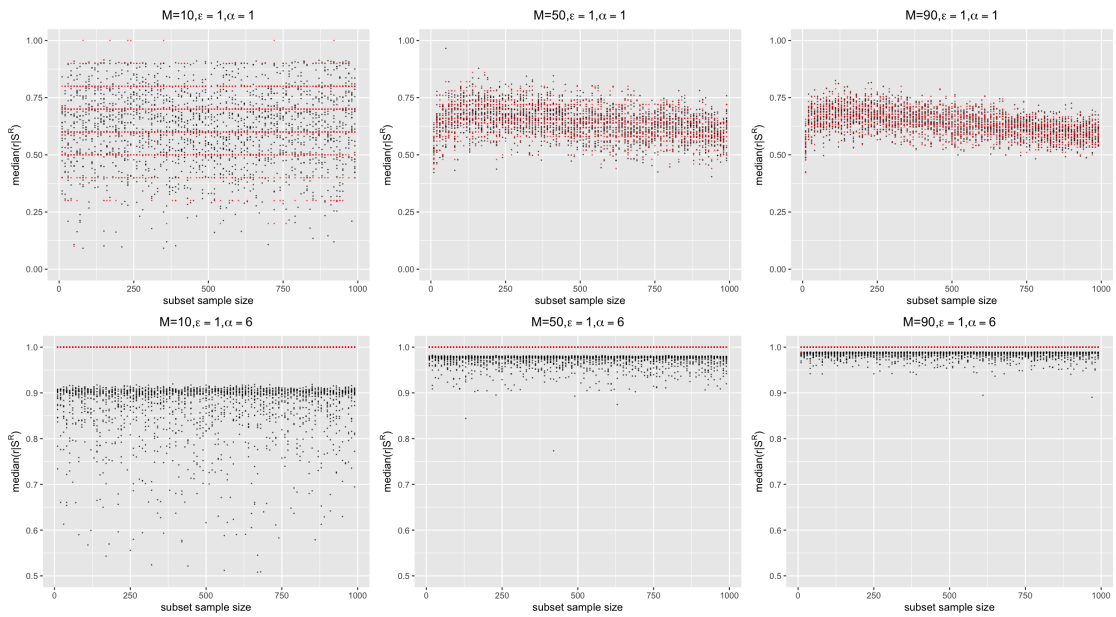

Figure 6.2: Simulated posterior median of r with inflation. Red points are empircal $\frac{S}{M}$, black points are the posterior median of r.

want to preserve the probability interpretation of a given interval rule. For $\alpha = 1$, the interval rule roughly corresponding to the 70% confidence interval, while for $\alpha = 6$, it roughly corresponds to the 99.9% confidence interval. Figure 6.2 reveals two important patterns. Firstly, as long as the subgroup sample size is enough for Assumption 1 to be reasonable, we can expect the interval rule will preserve the probabilistic guarantee significantly, as shown by the case with $\alpha = 1$. Secondly, due to the Laplace noise and the randomness in partitioning, we should tolerate the practical confidence level to be lower than the theoretical one slightly, i.e., δ should be specified smaller than it should be theoretically. It's revealed in the case where $\alpha = 6$ that the posterior medians are actually smaller than 1. In addition, similar phenomenon happens as before when M gets larger. The sampling variance of the posterior median of r reduces with larger M, which corresponds to finer grid and smaller effective Laplace error. Moreover, the marginal improvement of increasing M is decreasing with larger M.

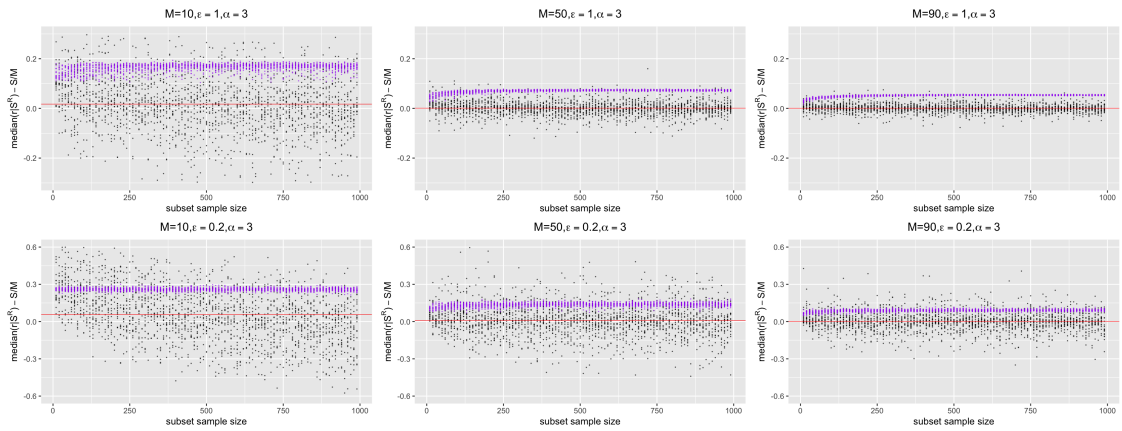

Figure 6.3: $residual = median(r|S^R) - \frac{S}{M}$. Red line is average $residual$. For each trial, black point is $residual$, purple point is posterior se of r.

6.2 Consistency of r

In this sub-section, we show that the effective error caused by Laplace noise decreases with larger M. Property a2 and Property a3 guarantee $r|S$ converges to $\frac{S}{M}$ and $\frac{S^R}{M}$, which indicates the decay of effective error. For $r|S^R$, it differs from $r|S$ by the approximation specified in Theorem 1, and such approximation will be finer in a global sense when M goes big. In this simulation, we continue with the setting in the previous section.

For the choice of parameters, we take $\epsilon \in \{0.2, 1\}$, $\alpha = 3$ and $M \in \{10, 50, 90\}$ simply for illustration. $\epsilon = 0.2$ indicates a strict DP requirement with large Laplace noise, while $\epsilon = 1$ is a more moderate one. Results are shown in Figure 6.3. With M grows bigger, the residual becomes much more concentrated around 0, and such trend holds for all values of n, which is corresponding to Property a2 and a3 that the convergence of posterior r to $\frac{S}{M}$ is controlled by M for any given legal n. In addition, the posterior standard error of r decreases with larger M, which indicates that not only the median (center), but also the whole posterior distribution of r is more tightly distributed around $\frac{S}{M}$ when M grows larger. For the speed of $r|S^R \to \frac{S}{M}$, as we can

expect, the marginal improvement is decreasing with larger M. As a direct result, for many practical applications, we can expect that the optimal M being chosen will not be large.

For smaller ϵ, the Laplace noise is much larger, resulting in the big divergence of the residuals from 0. Especially when $M \geq 50$, the residual and the posterior r are much more sparsely distributed when $\epsilon = 0.2$ than $\epsilon = 1$. However, the trend that posterior r is becoming more concentrated around $\frac{S}{M}$ with larger M doesn't change. The marginal improvement made by larger M is still decreasing, but less sharply when ϵ is smaller, i.e., the optimal M will be larger in order to control the effective Laplace noise.

Chapter 7

Application

In this section, we provide an example of implementing our frameworks. The data source is the IPUMS (https://usa.ipums.org/usa-action/variables/group). We extract the a subset from the IPUMS database of the year 2019 with the features: total income, family size, sex (2 levels), marital status (6 levels), race (9 levels), health insurance status (2 levels), education length (in year), school type (8 levels), employment status (4 levels). The extracted data set comprises 3,239,553 individuals. Then, we select a subset based on these criterion: income that is larger than 1, school type (private or public) is clearly identified, larger races (with sample size larger than 40,000), major marital status (married, married but now divorced, single), age from 18 to 65, no missing values in any of the features. We conduct this procedure mainly for simplicity in illustration rather than take it as a best practice for analyzing this data set. After that, we result in 160,364 samples. Suppose our goal is to study the relationship between school type (private or public) and the total personal income. The school type is binary which takes value 1 if being private, otherwise 0.

7.1 Application of AD

Suppose the original researchers first fitted a linear regression model without interaction terms and other transform of the variables on this full data set with 160,364 samples. Then, they deleted all the influential points with Cook's distances larger than $4/160364 \approx 2.5 \times 10^{-5}$, and resulted in a data set denoted as $\mathbf{D}$ with the sample size $n_0 = 154,442$. Namely, they deleted roughly 6,000 samples. Then, they refitted the model on $\mathbf{D}$, and published the result that $\hat{\gamma}_o = -26$ and $\hat{\sigma}(\hat{\gamma}_o) = 97$ after being

rounded to the nearest integer. Based on these estimations, the original researchers claimed that there is no significant relationship between the type of school and the total personal income.

Since deleting such a large proportion of samples is not common in practice and it might be controversial, we suppose the replication researchers want to replicate this study after putting back all the outliers deleted. Namely, the replication data set $\mathbf{D}^*$ is the full data set with $N = 160,364$. For the purpose of illustration, we propose two tolerance regions. The first is $U_1 = [500, +\infty)$, and the second question is $U_2 = [1000, +\infty)$. Both of them aim to address the question that whether graduating from private school has significantly positive influence on the total income, and the only difference between U_1 and U_2 is the scale of the difference. In addition, ϵ is chosen to be 1.

Choosing M based on simulated reference

As stated in Section 5.1, to select the M that can provide us with a robust result, we can refer to the simulated reference contour plot. Based on the published $\hat{\gamma}_o = -26$ and $\hat{\sigma}(\hat{\gamma}_o) = 97$, we create several simulated reference as shown in Figure 7.1 based on the procedure illustrated in Section 5.1.3.

To get the robustness, we need the sampling distribution of r_{obs} to be deviated from 0.5 for a larger range of true values of γ. Figure 7.1 directly shows the relationship between 0.5 and the distribution of r_{obs}. Figure 7.1 (a) and (b) show similar patterns, since they share the same random mechanism, i.e., normal distribution with the same standard error and the one-sided type of tolerance region, except for the different center. In addition, the plots are almost symmetric around the boundary of the tolerance region, which is a direct result of the fact that the sampling distribution for simulating the $\hat{\gamma}$ is symmetric.

From Figure 7.1, we can see that when M is too small, e.g., $M \leq 10$, the sampling

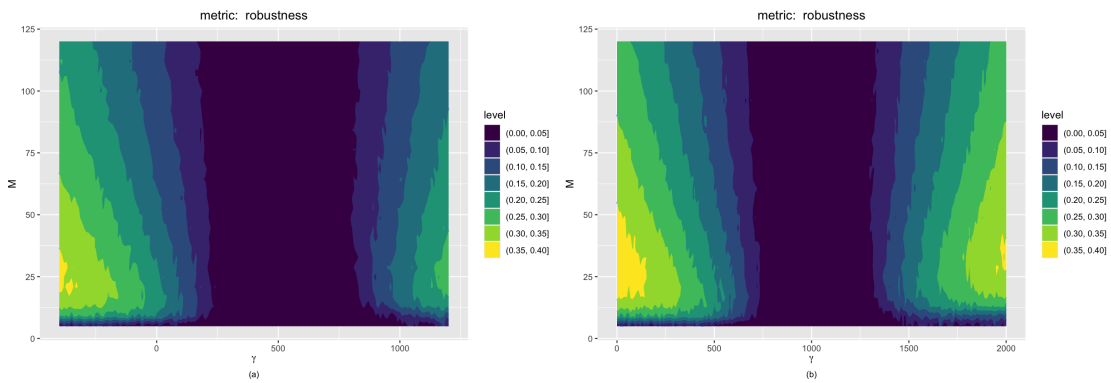

Figure 7.1: Reference contour for r_{obs}. Contour value is the distance between 0.5 and the interval $(q_{0.1}, q_{0.9})$. (a) is for U_1, (b) is for U_2.

distribution of r_{obs} can suffer from the Laplace noise a lot, indicating that the posterior inference of r will also suffer from the noise with a similar scale. In addition, with small M, the variance of S/M is large. When M gets as big as 20, the distribution of r_{obs} roughly leaves from 0.5 the most quickly with γ getting farther away from the boundary of U_1 or U_2. When M continues to get larger, the robustness is decreasing due to the larger uncertainty of each $\hat{\gamma}_r^{(l)}$, which makes the distribution of S closer to $Binomial(M, 0.5)$. As a result, r_{obs} is more potentially close to 0.5. Thus, to get the most robustness, we select the M that help our method to be the most robust, i.e., for larger range of the true values of γ, the distribution of r_{obs} is deviating from 0.5. For this example, the optimal M is roughly 20 for both U_1 and U_2.

If the region is two-sided, the contours will still be symmetric, but the trend above will be less monotonic, and we can imagine that we put one more center of symmetry in the plots we have now and distort the contours with the new center. Typically, the axis of symmetry of a two-sided region is a vertical line across the center of the region, and the two boundaries are the centers.

Conduct posterior sampling and make inference

After choosing $M = 20$, we take the partitioning procedure, get the aggregated indicator S, add Laplace noise to S and get the observation S^R. Then, we conduct

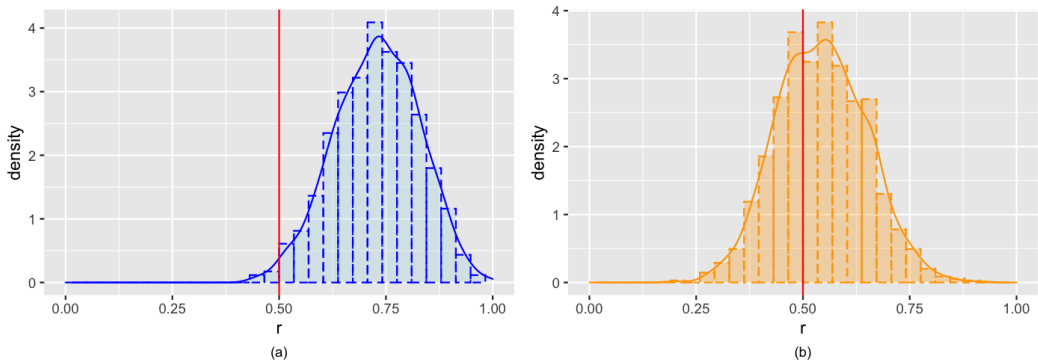

Figure 7.2: Posterior distribution of r. (a) is for U_1 and (b) is for U_2. Red lines indicate $\delta = 0.5$.

posterior sampling based on MCMC, get the posterior samples of r, and output this posterior distribution. Then, we can further calculate $\hat{\theta}_n = Pr(r \geq \frac{1}{2}|S^R)$, report this value, make inference about θ_n and furthermore, make inference about θ_0. In this example, we sample 1000 effect samples from the MCMC process as the posterior samples of r after burning the first 500 samples. In addition, the random seed is set to be 2021, and the platform we use for this example is R.

Calculated from the posterior distributions of r as shown in Figure 7.2, we have $\hat{\theta}_n = 0.99$ for U_1, while $\hat{\theta}_n = 0.641$ for U_2. We can also see from the two plots in Figure 7.2 that the posterior distribution of r for U_1 is almost on the right side of 0.5, while it is roughly centered at 0.5 for U_2.

Because $\hat{\theta}_n$ gives a pretty strong estimation, we can make the conclusion that $\theta_0 = 1$ for U_1. Namely, for the replication data set, there is a significant positive relationship between graduating from private school and the total income, and the estimated effect size is very likely to be larger than 500. In contrast, $\hat{\theta}_n$ gives a much more mutual answer for U_2, indicating the high uncertainty in this verification process, thus we only report this value and don't make any conclusion for the truth of θ_0 for U_2. Namely, we can hardly draw a conclusion of whether the effect size is larger than 1000 or not.

Overall, the conclusion is that, the published study might be controversial, and it's not reproducible on the replication data set $\mathbf{D}$. On $\mathbf{D}$, there is a strong positive relationship between private school the income, and the fitted coefficient of graduating from private school is very likely to be larger than 500.

7.2 Application of AM

Suppose the original study is conducted on the full data set with 160,364 individuals, and the original researcher fitted a linear model without interaction terms, which is $Model_0$, after which they published the results that $\hat{\gamma}_o = 1010$ and $\hat{\sigma}(\hat{\gamma}_o) = 177$. To be consistent in the notations as we have used before, continue to denote this full data set as $\mathbf{D}^*$. Suppose the alternative model $Model_1$ for replication is an ordinary linear regression with interaction terms of age (continuous) with marital status (categorical) and race (categorical) respectively, which is proposed after exploratory analysis by which we find the effect of age on personal income may vary by different marital status and races. Suppose we are the replication researchers who want to see how different the conclusions given by the two model specifications will be. More specifically, we may want to see whether the effect from $Model_1$ may change the sign or change by a half in the size relative to the previously estimated one.

Choose the M and make reference contour

Since the difference between $Model_1$ and $Model_0$ is that $Model_1$ incorporates some interaction terms, and the increase in the number of parameters to be estimated is not large compare to the sample size of each subset even if we conduct a dense partitioning such as $M = 50$, we can expect that the change in the standard error of the coefficient of school type will not be large. Namely, we prefer to believe $\sigma(\hat{\gamma}) \approx \sigma(\hat{\beta})$, indicating the sampling variance of $\bar{\nu}^L$ will not be large. Thus, we conduct the partitioning with $M = 25, 50$, and see which M can meet our need. The reference plots are shown in

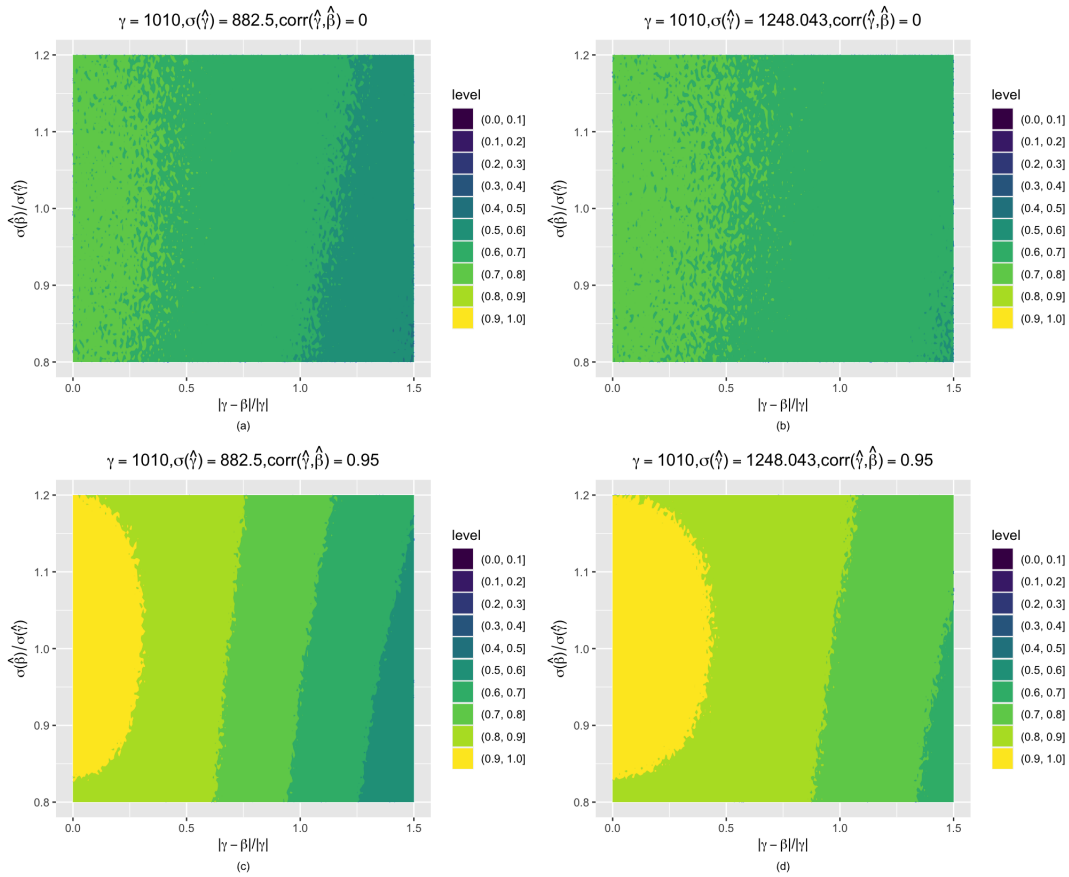

Figure 7.3: Reference contour plot for IPUMS. (a) and (c) are on $M = 25$, (b) and (d) are on $M = 50$. Each contour value is average on 500 ν.

Figure 7.3. The contour value is the estimation of $\mathbb{E}[\nu|n]$. For each M, we create two plots. One with no correlation between $\hat{\gamma}$ and $\hat{\beta}$, and the other one assumes the correlation is 0.95.

The reference without correlation, i.e., Figure 7.3 (a) and (b), provide the worst situation in the sense that the verification method will be least sensitive in detecting $|\beta - \gamma|$. In contrast, under high correlation, as shown in Figure 7.3 (c) and (d), the method will be more sensitive to the change in $|\beta - \gamma|$ since the influence of the randomness in the sampling of $\hat{\beta}$ and $\hat{\gamma}$ will reduce. Suppose we're caring about the relative difference from 0 to 1.5. Under the worst case, $M = 25$ can roughly provide ν with 3 different color bars, i.e., ν can change with a range with length

0.3, while $M = 50$ will only provide 2 color bars. In a practical situation, we can expect much higher correlation, and $M = 25$ will provide 5 effective color bars, while $M = 50$ will roughly provide 3 effective color bars. As illustrated before, the choice of M is depending on the control of Laplace error and the sampling sparsity of the average overlap, as well as considering the sensitivity of the method. $M = 25$ has been considered as eligible in controlling these errors, and the simulated contour also indicates that it can preserve enough sensitivity. In addition, larger M can lose sensitivity to an undesirable extent, which discourages us from choosing larger M. Thus, we adopt $M = 25$ as the final choice.

Make inference based on the posterior samples of $\bar{\nu}$

We conduct the verification process of AM framework once given random seed 2022 and number of partition $M = 25$. We have our observation of the noisy version average overlapping measure $\bar{\nu}^L \approx 1.03$, based on which we sample 1000 posterior samples of $\bar{\nu}$ with the default prior $\psi_0 = Beta(1, 1)$. The posterior 90% credible interval (from 5% quantile to 95% quantile) is $[0.878, 0.998]$ according to our posterior samples. Here, we decide to invert it into the credible interval for $|\beta - \gamma|$ under null assumption. The corresponding 90% credible interval for $|\beta - \gamma|$ is $[28, 473]$.

Then, we can make the inference statement on $|\beta - \gamma|$. With the given prior and Laplace noise, the difference between the two effect sizes will be within the range $[6, 424]$ with probability 90% under the null assumption after rounding the boundaries of the interval to nearest integers. With the originally estimated effect size being 1010, the change in the model specification will not change the sign of the effect, and with the probability larger than 0.9, the effect size will not reduce or increase by a half of the original one. Considering the data sets are the same for replication analysis and the original study, the statements above on the difference between β and γ will only be invalid if the two models differ significantly in the efficiency of estimating the

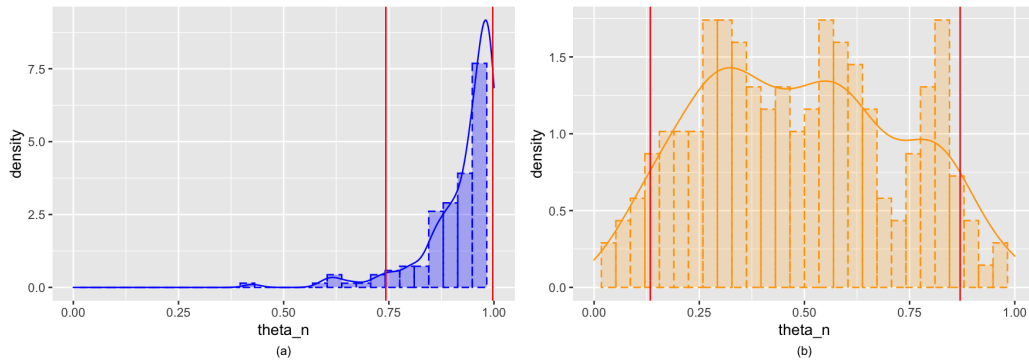

Figure 7.4: Distribution of $\hat{\theta}_n$ under repeated verifications. (a) is for U_1 and (b) is for U_2. Red lines show the 5% and 95% quantiles respectively.

targeted effect.

7.3 Comments on the applications

In this section, we give a brief evaluation on the performance of our methods in this application. It is purely for the purpose of illustration, and in practice, the following studies will now show up and be conducted by the replication researchers.

To evaluate the performance of AD framework, we repeated the verification process for 200 times and calculate $\hat{\theta}_n$ for each trial of the verification. The results for U_1 and U_2 are shown in Figure 7.4. Recall that on $\mathbf{D}$, we have $\hat{\gamma}_o = -26$ with $\hat{\sigma}(\hat{\gamma}_o) = 97$, and we have the fitted coefficient and standard error to be 1010 and 177 on $\mathbf{D}^*$. Because $U_1 = [500, +\infty)$ and it has the boundary that is far from 1010, we should expect that the $\hat{\theta}_n$ we get is much larger than 0.5. It is confirmed in Figure 7.4 (a) that most of the density of $\hat{\theta}_n$ under repeated verification is centered closely to 1. In contrast, the boundary of U_2, 1000, is close to 1010, so we can expect that it is hard to draw a conclusion on whether the coefficient of private school is larger than 1000. Correspondingly, Figure 7.4 (b) shows this fact that the $\hat{\theta}_n$ will be very likely to take a moderate value.

For the evaluation of AM framework, we fit both of the $Model_0$ and $Model_1$ directly on $\mathbf{D}^*$. The fitted coefficient and standard error of $Model_0$ are 1010 and 177, while these two quantities are 1086 and 176 for $Model_1$. Based on the fitted quantities on $\mathbf{D}^*$, we can expect that the difference between γ and β is not large. In section 7.2, the corresponding 90% credible interval for $|\beta - \gamma|$ is $[28, 473]$. Consider the estimated standard errors for $\hat{\beta}$ and $\hat{\gamma}$ on $\mathbf{D}^*$, this interval successfully resists the inflation of uncertainty caused by the partitioning. We result in an interval that has comparable uncertainty to the estimated quantities based on the full data set, namely, we are not losing much of the accuracy and meaningfulness during the partitioning procedure.

Chapter 8

Conclusions

In this thesis, we proposed two frameworks for replication analysis. AD framework models the probability of the fitted coefficient of a model falls within a certain tolerance region, while AM framework compares two fitted coefficients from two models. To protect the data privacy, we adopts the Laplace mechanism as well as the partitioning-and-aggregating procedure to reduce the effective Laplace noise that can obscure our quantities of interest. Not strictly speaking, one of the most significant ability and advantage of our method is that it preserves the power of the traditional hypothesis testing very well. In another word, we don't suffer a lot from the inflation of uncertainty caused by partitioning process. There should be three main reasons. First, as long as the subsets are still representative of the whole population, each subset should still be able to make meaningful inference, and the point estimation of the coefficient shouldn't vary drastically. Second, the intermediate quantities we use, r and ν, are both bounded from 0 to 1, such that the inherent variances of these two quantities will not be large, and the aggregating process to get r_{obs} and $\bar{\nu}^L$ can make the variances be more likely to be small. Last, for AD framework, the method for interpretation, namely, the $\hat{\theta}_n$ and the choice of δ, also helps with getting a robust result.

In addition, as a concept that we talk through this thesis, many of the properties of our methods are based on the prerequisite that each subset is still representative of the whole population. Thus, the users should check the subset sample size before implementing the two frameworks into a replication analysis.

Bibliography

[BBB+18] Andrés F Barrientos, Alexander Bolton, Tom Balmat, Jerome P Reiter, John M de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. Providing access to confidential research data through synthesis and verification: An application to data on employees of the us federal government. *The Annals of Applied Statistics*, 12(2):1124–1156, 2018.

[BHS16] Richard A Bettis, Constance E Helfat, and J Myles Shaver. The necessity, logic, and forms of replication. *Strategic Management Journal*, 37(11):2193–2203, 2016.

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE, 2015.

[BRMC19] Andrés F Barrientos, Jerome P Reiter, Ashwin Machanavajjhala, and Yan Chen. Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2):440–453, 2019.

[CBMR18] Yan Chen, Andrés F Barrientos, Ashwin Machanavajjhala, and Jerome P Reiter. Is my model any good: differentially private regression diagnostics. *Knowledge and Information Systems*, 54(1):33–64, 2018.

[CBRG18] Zachary Campbell, Andrew Bray, Anna Ritz, and Adam Groce. Differentially private anova testing. In *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pages 281–285. IEEE, 2018.

[CSS19] William Lee Croft, Jörg-Rüdiger Sack, and Wei Shi. Differential privacy via a truncated and normalized laplace mechanism. *arXiv preprint arXiv:1911.00602*, 2019.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[FKF10] Hermann Frank, Alexander Kessler, and Matthias Fink. Entrepreneurial orientation and business performance—a replication study. *Schmalenbach business review*, 62(2):175–198, 2010.

[GLRV16] Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on machine learning*, pages 2111–2120. PMLR, 2016.

[HABMA19] Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Pól Mac Aonghusa. The bounded laplace mechanism in differential privacy. *Journal of Privacy and Confidentiality*, 10(1), Dec. 2019.

[Hoe19] Sven Hoeppner. A note on replication analysis. *International Review of Law and Economics*, 59:98–102, 2019.

[KHP15] Fragkiskos Koufogiannis, Shuo Han, and George J Pappas. Optimality of the laplace mechanism in differential privacy. *arXiv preprint arXiv:1504.00065*, 2015.

[KKO+06] Alan F Karr, Christine N Kohnen, Anna Oganian, Jerome P Reiter, and Ashish P Sanil. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60(3):224–232, 2006.

[KKS15] Vishesh Karwa, Dan Kifer, and Aleksandra B Slavković. Private posterior distributions from variational approximations. *arXiv preprint arXiv:1511.07896*, 2015.

[KPG11] Mirjam J Knol, Wiebe R Pestman, and Diederick E Grobbee. The (mis) use of overlap of confidence intervals to assess effect modification. *European journal of epidemiology*, 26(4):253–254, 2011.

[KV17] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

[LC11] Jaewoo Lee and Chris Clifton. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[MH10] Saeed Maghsoodloo and Ching-Ying Huang. Comparing the overlapping of two independent confidence intervals with a single confidence interval for two normal population parameters. *Journal of Statistical Planning and Inference*, 140(11):3295–3305, 2010.

[MR12] David R McClure and Jerome P Reiter. Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality*, 4(1), 2012.

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[PA13] Sampo V Paunonen and Michael C Ashton. On the prediction of academic performance with personality traits: A replication study. *Journal of research in personality*, 47(6):778–781, 2013.

[Rei19] Jerome P Reiter. Differential privacy and federal data releases. *Annual review of statistics and its application*, 6:85–101, 2019.

[ROK09] Jerome P Reiter, Anna Oganian, and Alan F Karr. Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, 53(4):1475–1482, 2009.

[She17] Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.

[SM11] Rathindra Sarathy and Krishnamurty Muralidhar. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Priv.*, 4(1):1–17, 2011.

[Smi11] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822, 2011.

[SSJM21] Sarah Steven, Ronald Sophia, Megan Jose, and Matthew. Ipums usa: Version 11.0, 2021.

[WFW+15] Xi Wu, Matthew Fredrikson, Wentao Wu, Somesh Jha, and Jeffrey F Naughton. Revisiting differentially private regression: Lessons from learning theory and their consequences. *arXiv preprint arXiv:1512.06388*, 2015.

[WLK15] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[WM10] Oliver Williams and Frank McSherry. Probabilistic inference and differential privacy. *Advances in Neural Information Processing Systems*, 23, 2010.

[WPB06] Matthew W Wheeler, Robert M Park, and A John Bailer. Comparing median lethal concentration values using confidence interval overlap or ratio tests. *Environmental Toxicology and Chemistry: An International Journal*, 25(5):1441–1444, 2006.

[WSW15] Yue Wang, Cheng Si, and Xintao Wu. Regression model fitting under differential privacy and model inversion attack. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[ZRD16] Zuhe Zhang, Benjamin IP Rubinstein, and Christos Dimitrakakis. On the differential privacy of bayesian inference. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.