통계연구(2023), 제28권 제4호, 1-24/ DOI https://doi.org/10.22886/jkos.2023.28.4.1

데이터 유용성이 개선된 Differential Privacy 방법들에 대한 사례 검증 연구

박성률[1] 오영주[2]

요약

Differential Privacy는 데이터에 잡음을 추가하는 비식별화 기술이다. 이 기술은 프라이버시 수준을 하이퍼-파라메타 ϵ의 수준으로 통제가 가능하며, ϵ와 전역 민감도를 이용하여 잡음을 추가한다. 잡음을 추가하는 양에 따라 데이터 프라이버시 수준이 결정되고, 많은 잡음은 데이터의 프라이버시를 강력하게 해주지만 그에 반해 데이터의 유용성은 낮아진다. 그 이유는 프라이버시와 유용성은 trade-off 관계에 있기 때문이다. 본 논문에서는 같은 프라이버시 수준에서 유용성을 높여주는 가우시안 Differential Privacy과 재현자료를 생성하는 베이지안 모형 기반의 Modips에 대한 데이터 유용성 성능을 평가하고자 한다. 기존 Differential Privacy와 가우시안 Differential Privacy, 그리고 베이지안 모형 기반 Differential Privacy 알고리즘을 소개하고, 실제 데이터를 이용해서 두 방법이 적용된 통계량을 계산한 후 유용성 평가 측도를 이용하여 데이터 유용성 성능을 분석한다.

주요용어 : 데이터 비식별화, 가우시안, 베이지안 모형, 데이터 유용성

1. 서론

경제, IT, 과학, 통계 분야 등에서 사용하고 있는 데이터 정의는 약간의 차이는 있지만 대체로 속성의 특성을 나타내는 의미 있는 정보들을 말하며, 민감한 정보들이 다수 포함하고 있다는 점에서 유사하다고 할 수 있다. 여기서 민감한 정보들은 자료 분석이나 통계 작성에 매우 중요한 변수들로 활용이 되고 있으나 현실적으로 민감 정보들은 개인(또는 기업)의 프라이버시 문제 때문에 통계 이용자들에게 원활하게 제공하지 못하고 있다.

이러한 문제점을 보완하기 위한 방법으로 민감한 정보의 프라이버시를 보장하면서 데이터 유용성을 높일 수 있는 비식별화 방법들이 제안되어 왔다. 전통적인 비식별화 방법에서 가장 많이 활용되고 있는 방법들은 민감한 정보를 삭제하거나 대체하는 방법으로 이 방법들은 자료를 분석할 때 데이터의 정보 손실이 발생하여 자료 분석 시

[1] 대전광역시 서구 한밭대로 713 통계센터 6층, 통계청 통계개발원 통계방법연구실. E-mail: srpark08@korea.kr
[2] 대전광역시 서구 청사로 189 정부대전청사 3동, 통계데이터허브국 통계데이터기획과. E-mail: oyj1928@korea.kr

정확성과 신뢰성을 장담할 수 없게 된다. 특히 민감한 정보는 데이터 분석에서 매우 중요한 정보를 가지고 있어 더욱 신중하게 처리해야함이 틀림없다. 전통적인 비식별화 방법의 단점을 보완하기 위해서 제안된 여러 방법 중 Differential Privacy(이하, DP)는 기존 문제점을 개선할 수 있을 것으로 여겨진다. 그 이유는 데이터 정보 손실을 방지할 수 있을 뿐만 아니라 데이터 특성을 고려한 잡음을 추가하기 때문에 원자료와 유사한 값들을 만들어 줄 수 있기 때문이다. 하지만 적정한 ϵ을 정해야하는 문제가 여전히 남아 있는 숙제다.

원자료와 동일한 분포를 따르도록 생성된 가상자료인 재현자료(Synthetic Data) 역시 데이터 정보 손실 문제를 해결하기 위해 이용할 수 있으나, 연결이나 추론 등 다양한 공격을 통해 정보노출이 발생할 수 있다는 한계를 지닌다. 이에 따라, 기존 재현자료보다 정보 노출을 더 강력하게 제어해야 한다는 필요에 의해 DP 기준을 충족시키는 재현자료 생성기법들이 제안되었다.

본 논문에서는 DP 방법들(ϵ-DP, ϵ,δ-DP 등) 중 동일한 ϵ 수준에서 데이터 유용성이 높은 방법이 무엇인지 알아보기 위해 공개 데이터에 대한 실증분석을 수행하였다. 또한, DP 기준을 충족시키는 재현자료는 기존 재현자료보다 데이터 유용성이 낮다는 한계가 있으므로, ϵ 수준에 기반 한 데이터 유용성을 살펴볼 수 있도록 비교 결과를 제시함으로써 타 연구들과 차별성을 두었다.

논문의 구성은 2장에서는 Global Privacy 분야에서 제안된 여러 DP 방법론에 대한 사례 연구들을 소개하고, 이 중에서 논문에서 적용할 알고리즘을 선정하였다. 3장에서는 2006년 Cynthia Dwork이 제안한 DP의 고전적인 정의(ϵ DP, ϵ,δ DP)에 대하여 소개한다. 4장에서는 DP를 통계적 귀무가설의 관점에서 알고리즘을 적용하는 가설검정 기반의 Gaussian DP를 소개하고, 5장에서는 베이지안 체계에서 모델에 기반한 DP가 적용된 재현자료 생성 방법인 Bayesian model DP를 소개하고자 한다. 6장에서는 전자공시시스템에서 공개하고 있는 기업체의 재무정보 데이터를 활용하여 세 방법들에 대해 DP가 적용된 값들을 구하고, 데이터 유용성을 비교하였다. 마지막 7장에서는 연구 내용을 정리하고 시사점을 제시하였다.

2. 관련연구

DP는 자료를 수집하는 환경(기관 신뢰, 기관 불신)에 따라 Global Privacy와 Local Privacy로 구분된다. Global Privacy는 기관에서 수집한 자료를 일반 이용자가 이용할 수 있게 공개할 때 잡음 추가하는 방식이고, Local Privacy는 개인이 자료를 수집할 기관에 정보를 제공할 때 DP 기준에 맞춰 제공하는 방식이다.(박민정, 2018)

DP 분야에서 발표되고 있는 논문들은 주로 Global Privacy에 기반한 연구들이다. 그 이유는 공학, 통계, 수학 등 여러 분야에서 민감한 원자료를 대체하는 데 있어 비

식별화 방식(DP 적용)을 적용했을 때 나타나는 데이터 안전성과 활용성에 대해 연구자들의 높은 관심 때문이다. 본 논문에서도 Global Privacy에 초점을 맞춰 DP 적용에 따라 노출 제어의 효과 대비 데이터 유용성이 낮아지는 문제점을 보완한 연구들을 살펴봤다. 아래의 선행연구들은 이 취지에 부합하여 살펴본 연구들이다. 동일한 ϵ 하에서 데이터 유용성을 높일 수 있는 다양한 이론들을 살펴보았다.

Li 등(2014)은 1차원 또는 저차원에서만 적용되는 기존 DP 알고리즘의 단점을 보완하는 방법으로 marginal 분포들과 Gaussian Copula 함수를 추정하고 이렇게 생성된 결합분포로부터 ϵ DP가 결합분포를 생성한 후 샘플링을 통해 재현자료를 생성하는 알고리즘을 제안했다.

Liu(2022)은 베이지안 모델에 기반하여 재현자료 내 원자료에 대응되는 개별 데이터가 없도록 DP 개념을 적용한 Modips 알고리즘을 제안하였다.

Zhang 등(2017)은 데이터셋의 전체 분포를 근사하기 위해 저차원 분포들을 결합하여 간단하게 이용하여 고차원 데이터를 효과적으로 처리하지 못하고, 입력 데이터에 많은 속성이 포함될 경우 많은 양의 잡음이 추가되는 Bayesian Network를 보완하기 위해 PrivBayes 알고리즘을 제안하였다. 이 방법은 데이터셋의 각 변수들과 변수들 간 방향성의 관계들을 이용하여 Bayesian Network를 형성하고, DP를 만족하는 재현자료 생성 방법이다.

Abadi 등(2016)은 기존 SGD 알고리즘에 기울기 클리핑 기법을 적용하여 평균 기울기의 민감도 한계를 제한하여 반복적인 학습을 통해 발행되는 프라이버시 예산을 효율적으로 계산하는 DPSGD 알고리즘을 제안하였다.

Abay 등(2018)은 프라이버시 예산의 효율적 분석과 심층 학습 기술을 활용하여 더 높은 데이터 유용성을 갖는 DPSYN 알고리즘을 제안하였다. 이 방식은 다중 특징들 사이의 관계를 포착할 수 있는 딥러닝 모델 학습을 통해 재현자료를 생성하는 방법이다.

Jordon 등(2019)은 개별 데이터 영향을 엄격하게 제한하여 엄격한 DP를 보장하는 PATE(Private Aggregation of Teacher Ensembles) 프레임워크를 수정하여 GAN에 적용하는 PATE-GAN 알고리즘을 제안하였다.

Dong 등(2022)은 통계 가설 관점에서 주어진 유의수준에서 최소의 제2종 오류를 trade-off function으로 나타냄으로써 DP를 적용하는 GDP(Gaussian Differential Privacy) 방법을 제안하였다.

본 논문에서는 위 선행연구에서 제시된 여러 방법들 중 통계 가설 검정 기반의 Gaussian DP와 모형 기반의 Modips를 통해 기존에 제시된 (ϵ, $\epsilon\delta$)DP 대비 데이터 유용성 성능을 검증하고자 한다. 검증을 위해서는 통계량(평균과 분산)을 계산하고, 계산된 통계량을 여러 측도(MAE, RMSE, MAPE)를 활용하여 검증하였다. 3장부터 5장까지는 기존 (ϵ, $\epsilon\delta$)DP 모델, Gaussian DP 모델, 그리고 Modips 모델의 알고리즘에 대해 살펴본다.

3. Differential Privacy

2006년 Cynthia Dwork이 발표한 DP는 새로운 개념의 정보 보호 방법론이다. DP는 프라이버시를 정량적으로 모델화하여 프라이버시 보호 정도를 측정할 수 있는 방법으로 데이터의 분포 특성을 고려하여 데이터의 유용성은 유지하면서 개인정보를 보호하기 위해 잡음을 추가하는 방법이라고 할 수 있다.

DP는 프라이버시 일부 노출을 감수하면서 원본 데이터의 유사한 특성을 갖도록 데이터를 익명화시키는 것이 중요하다. 그 이유는 프라이버시와 데이터의 유용성 간에는 Trade-off 관계가 존재하기 때문이다. DP는 특정 개인의 존재 유무가 다른 두 데이터베이스에 대하여 어떠한 질의의 출력 값에 랜덤성을 추가하여 두 버전에 대한 질의 응답값이 확률적으로 일정 값 이하의 차이를 갖도록 함으로써 차분 공격을 어렵게 하는 프라이버시 모델이라고 할 수 있다.

$$P[K(D_1)\in S] \le e^{\epsilon} P[K(D_2)\in S] \tag{3.1}$$

$$\frac{P[K(D_1)\in S]}{P[K(D_2)\in S]} \le e^{\epsilon} \tag{3.2}$$

식 (3.1)은 ϵ DP의 정의로, 1개의 데이터만 다른 임의의 두 데이터베이스 D_1과 D_2에 대하여, 그리고 임의의 집합 S에 대하여, 비식별화 알고리즘 K가 식 (3.1)을 만족하면 K가 ϵ DP를 만족하는 알고리즘이라고 정의한다. 이 정의는 D_1을 비식별 처리한 결과 $K(D_1)$과 D_2를 비식별 처리한 결과 $K(D_2)$가 동일한 집합 S에 포함될 확률의 비율이 주어진 ϵ에 대하여 e^{ϵ}보다 작거나 같다는 것을 의미한다. 이러한 의미에서 ϵ은 비식별화 알고리즘 K의 프라이버시 손실을 계량화하는 모수이며 프라이버시 예산(Privacy Budget)이라고 부른다. 참고로 비식별화 알고리즘 K의 대표적인 예는 평균, 총합, 도수 등 비식별화 처리가 필요한 통계량에 잡음을 임의로 더하는 알고리즘이다.

식 (3.1)과 식 (3.2)를 통해 알 수 있는 DP의 특징은 첫째, ϵ값이 작을수록 K로 비식별 처리한 두 결과의 확률분포가 유사해지므로 비식별 처리한 결과를 구분하는 것이 더 어려워지게 된다. 따라서 구분 불가능성 즉, 프라이버시를 강화하기 위하여 더 큰 잡음을 더한다는 것을 의미하므로 비식별 처리한 결과의 유용성은 낮아지게 된다. 둘째, ϵ값이 커지면 반대로 두 응답 값의 확률분포 차이가 커져서 프라이버시는 약화(프라이버시 손실 많아짐), 잡음이 적은 데이터, 데이터 유용성 높아지고, 데이터 안전성이 낮아지는 효과가 나타난다.

DP 개념 중 전역 민감도(Global Sensitivity)의 개념을 살펴봐야한다. 이 개념은 특정 개인을 추가 또는 제거했을 때 생기는 변화량의 최대값을 말한다. 즉, 데이터의 경계가 모호하거나 이상치가 큰 경우 민감도를 설정하기 어렵거나 지나치게 큰 값이 설정되는 문제도 함께 발생된다. 작은 전역 민감도는 잡음을 적게 만들고, 큰 전역 민감도는 잡음을 많게 만든다.

ϵ DP에서 비식별화 알고리즘 K의 대표적인 매커니즘은 라플라스 매커니즘이다. 이 방법은 라플라스 매커니즘은 $Laplace(\mu,b)$ 분포를 이용하여 임의의 잡음(X)을 추가하는 방법으로 도출 과정은 식 (3.3)과 같다.

$$X=\mu-b\ sgn(U)\ \ln(1-2|U|) \quad \textbf{(3.3)}$$

식 (3.3)에서 X는 잡음, μ=0(라플라스 함수 기준점), U는 $(-1/2 \le U \le 1/2)$ 사이 랜덤 값으로 이 값을 사용하여 잡음 생성, sgn은 입력 값이 양수이면 1, 음수이면 -1 반환, b는 $\Delta f/\epsilon$이다. 이 때 전역 민감도(Δf)는 통계량을 계산하는 함수 f에 대하여 $\Delta f=\max|f(D_1)-f(D_2)|$로 정의된다. 만약 $f(D)$가 주어진 데이터베이스 D에서 계산한 총합이라면 Δf는 주어진 데이터베이스에서 데이터의 최대값과 최소값의 차이가 된다.

예를 들면 원본 데이터가 1.5인 값에 잡음을 삽입하여 비식별 속성 값을 계산하고자 할 때, $\mu=0$, $U=0.3$, $\Delta f=1$, $\epsilon=0.2$ 값이 주어졌을 때 잡음(X)은 1.9897이다. 따라서 원본 데이터에 잡음을 더하여 만든 비식별 데이터 값은 3.4897이 됨을 알 수 있다.

ϵ DP는 두 확률의 차이를 상대적인 관계로 제어하기 때문에 매우 강한 정도의 정보 보호라고 할 수 있다. 하지만 이렇게 강한 DP는 현실적으로 구현이 불가능한 경우가 발생될 수 있기 때문에 기준을 조금 완화한 ϵ,δ DP를 제안하였다.

$$P[K(D_1\in S)] \le e^{\epsilon}P[K(D_2\in S)]+\delta \quad \textbf{(3.4)}$$

식 (3.4)는 δ로 두 확률의 차이를 절대적 차이 이내에 있게 하는 조건을 주어 기존 ϵ DP 기준보다 조금 완화된 기준을 제시한다. 또한 데이터 주체가 ϵ을 초과하는 프라이버시 손실을 겪을 확률이 δ로 제한되도록 보장한다. 실제 ϵ과 δ는 모두 사용자가 값을 선택하여 적용하고, 이 때 δ값의 범위는 에러 발생 가능성으로 확률 0 ~ 1 사이의 값을 가진다. 예를 들면, $\epsilon=0.2, \delta=0.1$ 값을 가지면 원본 데이터셋 D_1의 확률분포가 $\epsilon=0.2$를 적용한 비식별 데이터 D_2의 확률분포의 차이에 $\delta=0.1$ 만큼의 에러 발생 가능성을 가진다는 의미이다.

4. Gaussian Differential Privacy

2022년 Domg 등이 제시한 Gaussian Differential Privacy(이하, GDP)는 기존 DP 방법을 통계적 귀무가설의 관점에서 보는 것이다. GDP는 잡음을 삽입하는 방법 중 동일한 조건(ϵ,δ)하에서 분산이 가장 작게 나타나는 방법으로 자료의 유용성을 가장 높일 수 있는 것이 장점이다. GDP는 어떤 자료가 주어졌을 때, 그 자료를 만드는데 한 개인이 포함되었는지 여부가 알려지는 것을 개인 정보가 노출되었다고 할 수 있다. 통계 가설의 관점에서 귀무가설(주어진 자료에 개인이 포함 되었다)과 대립가설(주어진 자료에 개인이 포함되지 않았다)을 설정하면 주어진 유의수준하에서 제2종 오류(type II error)가 클수록 개인이 포함되었는지 알기 어렵다는 것을 의미한다.

GDP는 주어진 유의수준에서 최소의 제2종 오류를 trade-off function으로 나타낸다. 여기서 기준이 되는 trade-off function은 귀무가설이 표준정규분포, 대립가설이 평균이 μ이고 분산이 1인 정규분포이다. 주어진 trade-off function보다 더 큰 값을 가지는 DP 알고리즘에 대해서 μ GDP를 만족한다고 한다. μ가 작을수록 두 분포 사이의 거리가 가까워지고 제2종 오류가 커지기 때문에 개인 정보 보호 수준이 강력해진다.

GDP 적용 절차로 1단계에서는 DP를 적용하고자 하는 대상($f(x)$)을 결정하고, 민감도(sensitivity)를 계산한다. 여기서 민감도는 기본 데이터 셋의 변경이 쿼리 결과에 미칠 수 있는 영향을 의미하며, 민감도가 클수록 잡음 분포의 분산이 커지게 된다.

Δf를 $f(x)$의 민감도라고 하고, 한 값의 변화로 인해 바뀌는 $f(x)$의 최대 변화량은 식(4.1)과 같이 계산할 수 있다.

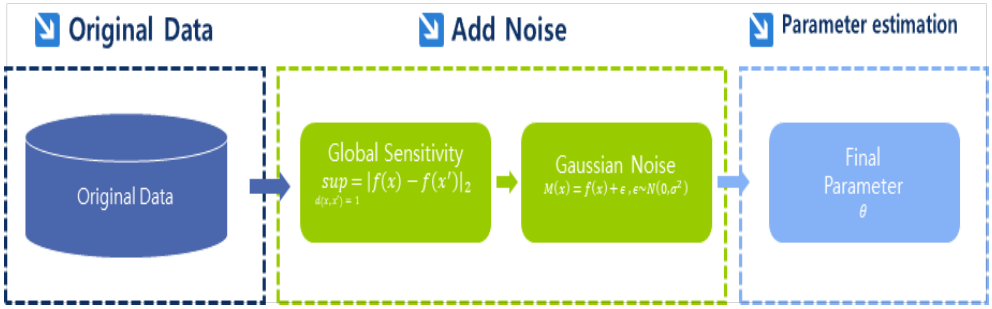
$$\sup_{d(x,x')=1} = |f(x)-f(x')|_2 \tag{4.1}$$

예를 들어, 식(4.2)에서 $f(x)$를 5개의 범주를 가진 히스토그램, $f_i(x)$를 i번째 범주의 관측값이라고 할 때 x와 x'의 차이를 한 자료가 추가되거나 제외된 경우라고 하면 위 식은 1이 되고, x와 x'의 값 하나만 서로 다른 값을 가진다고 하면 식(4.2)은 $\sqrt{2}$가 된다.

$$|f(x)-f(x')|_2 = \sqrt{\sum_{i=1}^{5}|f_i(x)-f_i(x')|} \tag{4.2}$$

2단계에서는 1단계의 정보를 바탕으로 DP가 적용된 $M(x)$를 계산하는 것이다. DP가 적용된 $f(x)$는 $M(x)=f(x)+\varepsilon$과 같이 정의되고, $\varepsilon \sim N(0,\sigma^2)$에서 $\sigma^2=\frac{\Delta f^2}{\mu^2}$이다. $M(x)$을 활용해서 통계량을 계산하거나 자료를 생성하는데 활용할 수 있다. 또한, μ GDP에서

개인 정보 보호 수준의 비교를 위해서 ϵ,δ DP의 ϵ,δ로 변환도 가능하다. μ GDP를 만족하는 $M(x)$는 $\delta=\Phi(-\frac{\epsilon}{\mu}+\frac{\mu}{2})-e^{\epsilon}\Phi(-\frac{\epsilon}{\mu}-\frac{\mu}{2})$인 ϵ,δ DP를 만족한다. 즉 주어진 δ 수준에서 ϵ을 비교할 수 있다.



<그림 4.1> GDP 알고리즘

5. Bayesian model Differential Privacy Data Synthesis

Liu(2022)는 베이지안 체제에서 베이지안 충분 통계량과 근사 분포 등을 이용하여 DP를 충족시키는 재현자료를 생성하는 기법인 Model-based Differentially Private Synthesis(이하, Modips)를 제안하였다. 본 연구에서는 Modips를 이용하여 재현자료를 생성하기 위해 라플라스 매커니즘을 활용하며, $Laplace(0,\Delta\epsilon^{-1})$로부터 독립적인 확률표본을 생성함으로써 ϵ-DP를 만족하는 재현자료를 생성한다.

베이지안 통계학은 고전 통계학과 달리 선험적 정보를 반영하여 분석 및 예측하는 통계학의 한 분야이다. 고전 통계학과 베이지안 통계학의 큰 차이점은 모수를 바라보는 관점이 다르다는 것이다. 고전 통계학은 모수를 고정된 값으로 고려하는 반면, 베이지안 통계학은 모수를 분포를 가지는 확률변수로써 고려한다.

$$f(\theta \mid \boldsymbol{x})=\frac{f(\theta,\boldsymbol{x})}{f(\boldsymbol{x})}=\frac{f(\boldsymbol{x} \mid \theta)f(\theta)}{f(\boldsymbol{x})} \quad \textbf{(5.1)}$$

여기서, $f(\theta \mid \boldsymbol{x})$는 사후분포, $f(\theta)$는 사전분포, $f(\boldsymbol{x} \mid \theta)$는 우도함수를 의미한다.

위의 사후분포를 활용하여, 관측되지 않은 새로운 자료 x^*에 대한 확률분포는 다음과 같이 추정될 수 있다. 이는 사후예측분포라 불리며, 본 논문에서 재현자료 생성을 위해 이용된다.

$$f(x^* \mid \boldsymbol{x})=\int_{\theta} f(x^* \mid \theta)f(\theta \mid \boldsymbol{x})d\theta \quad \textbf{(5.2)}$$

DP를 적용한 재현자료 생성 절차는 베이지안 모델 생성, 익명화, 재현자료 생성하는 3단계로 이루어져 있다. 1단계는 주어진 데이터 $\boldsymbol{x}$에 대한 베이지안 모델을 생성하는 단계로 원자료(D)와 여러 개(m개)의 사전 확률 분포($f(\theta)$)를 이용하여 m개의 사후 확률 분포($f(\theta \mid \boldsymbol{x})$)를 계산한다.

2단계는 익명화(sanitization) 단계로 Direct sanitization, SBS, sanitization of Approximate Distribution 3가지 익명화 방법이 있다.

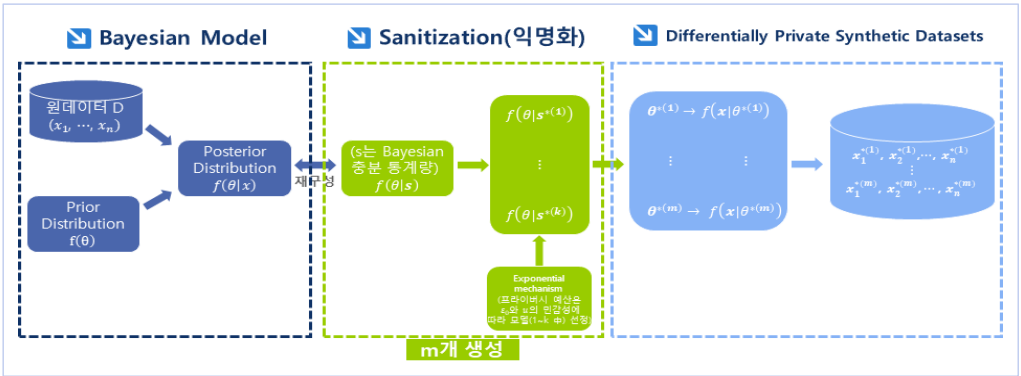
첫째, Direct sanitization 기법은 DP 메커니즘을 통해 직결로 익명화하는 방법으로 베이지안 알고리즘 $f(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta) f(\theta)$에 대하여 익명화 결과 값은 $f^*(\theta \mid \boldsymbol{x}) = f(\theta \mid \boldsymbol{x}) + e$이고, 여기서 $e \sim Laplace(0, \Delta\epsilon^{-1})$로 라플라스 잡음을 추가하여 익명화를 진행한다.

둘째, 베이지안 충분성을 통한 익명화(SBS) 기법은 베이지안 충분통계량 s를 이용하여 사후 확률 분포($f(\theta \mid \boldsymbol{x})$)를 $f(\theta|s)$로 재구성하여 s를 익명화한다.

셋째, 근사 분포의 익명화는 사후 확률 분포($f(\theta \mid \boldsymbol{x})$)를 계산하는 게 불가능에 가까울 정도로 어려운 경우 다루기 쉬운 $g(\theta \mid \boldsymbol{x})$로 근사하여 익명화하는 기법이다.

이 논문에서는 베이지안 충분통계량 $\boldsymbol{s}$를 이용하여 사후 확률 분포($f(\theta \mid \boldsymbol{x})$)를 $f(\theta \mid \boldsymbol{s})$로 재구성하여 $\boldsymbol{s}$를 익명화하는 베이지안 충분성을 통한 익명화(Sanitization through Bayesian Sufficiency, SBS) 기법을 이용한다. 이는 $f(\theta \mid \boldsymbol{x})$가 비록 닫힌 형태이더라도, Direct sanitization 기법을 통해 $f(\theta \mid \boldsymbol{x})$를 익명화하는 것은 $f^*(\theta \mid \boldsymbol{x})$가 proper density function이 되거나 1로 적분되지 않을 수 있기 때문이다. 또한, 본 논문에서는 가우시안 분포를 이용하여 사후확률분포를 간단하게 계산할 수 있으며, 이를 통해 베이지안 모델에 사용자가 지정한 ϵ보다 작은 ϵ_0을 지닌 m개의 DP 베이지안 모델을 생성할 수 있다.

3단계는 DP 메커니즘을 통해 선정된 베이지안 모델에 대응하여 사후 표본 $\theta^{*(i)}$를 만들고, 이 값을 이용하여 $f(\boldsymbol{x} \mid \theta^{*(i)})$에서 $\tilde{\boldsymbol{x}}^{*(i)}$를 생성한다. 최종적으로 m개의 재현자료 데이터 셋이 생성된다.



<그림 5.1> Modips 알고리즘

6. 데이터 유용성 성능 평가

6.1 raw data 소개

본 논문에서는 기존 ϵ,δ DP 대비 GDP와 Modips의 데이터 유용성 성능 평가를 위해 금융감독원 전자공시시스템(Data Analysis Retrieval and Transfer System, DART)의 기업 재무정보를 이용하였다. 이 데이터는 상장법인 등이 제출한 공시서류를 누구나 인터넷을 통해 조회할 수 있고, 기업별 다양한 재무정보도 다운이 가능한 공개 자료이다.

실증분석을 위해 사용된 데이터(2,233개)는 2021년 사업보고서로 재무상태표, 손익계산서, 현금흐름표 항목 중 분석 결과에 영향을 미치는 분야별(산업대분류, KSIC) 데이터 크기, 결측값, 극단값, 정규성 가정 등을 고려하여 2개 변수(당기순이익(Profit Loss), 재무활동으로 인한 현금 흐름표(Cash Flows From Used In Financing Activities))에 대해 1,476개를 확정하였다.

<표 6.1> 산업분류별 분석 데이터

한국표준산업분류(KSIC)	raw data	Modified data
전체	1,891	1,476
C. 제조업	1,454	1,135
G. 도매 및 소매업	161	123
J. 정보통신업	276	218

* KSIC: Korean Standard Industrial Classification

(자료 출처 : 금융감독원 전자공시시스템)

본 논문에서 데이터 유용성 검증을 위해 선정한 방법은 기존 ϵ,δ DP와 GDP, Modips이다. ϵ,δ DP와 GDP는 원자료 추정량에 잡음(가우시안)을 입력해서 DP가 적용된 추정량을 생성하는 반면, Modips는 원자료 분포를 이용하여 추정량을 만들고, 그 추정량에 잡음(라플라스)을 입력한 후 새로운 추정량을 이용하여 재현자료를 생성하는 방법이다. 현재 발표되고 있는 DP 알고리즘은 어떤 방법이 좋은지 판단할 수 없기 때문에 GDP와 Modips 등 알고리즘을 바로 비교하기에 무리가 있다. 따라서 분석의 방향은 기존 ϵ,δ DP 방법과의 비교를 통해 데이터 유용성에 효과가 있는지 검토를 하고자 한다.

<표6.2>는 모의실험에 활용할 2개의 분석 유형에 대한 설계 결과이다. 유형1은 ϵ,δDP와 GDP 모두 모수에 잡음을 입력하는 방식으로 데이터 유용성 측도로 MAE, RMSE, MAPE를 활용한다. 유형2는 원자료와 Modips에 대한 비교로 재현자료를 생성한 결과에 대해 통계량, 95%신뢰구간 중첩성을 데이터 유용성 측도로 활용한다. 원자료와 비교하는 이유는 ϵ,δ DP로 생성한 재현자료는 Global Sensitivity 너무 커지기

때문에 재현자료 생성 결과에 대한 정확도에 문제가 발생되기 때문이다. 따라서 Modips는 ϵ의 변화에 따른 추정량의 변화를 살펴본다.

<표 6.2> 유형별 모의실험 분석 설계

비교 대상			잡음 입력 방식(추정량)	데이터 유용성 측도
유형	기존 방법	제안 방법		
1	ϵ,δ DP	μ GDP	ϵ,δ DP : $\hat{\theta}$ + 잡음 → 평균, 분산	MAE, RMSE, MAPE
			μ GDP : $\hat{\theta}$ + 잡음 → 평균, 분산	
2	원자료	Modips	Modips : $\hat{\theta}$ + 잡음 → 재현자료 → 통계량	통계량, 95%신뢰구간 중첩성

6.2 GDP 모의실험 결과

GDP의 데이터 유용성에 대한 검증을 위해서 Original data 통계량 대비 ϵ,δ DP의 적용 결과와 GDP의 적용 결과를 비교하고자 한다. 모의실험의 목적은 ϵ의 변화에 따른 데이터 유용성 성능에 대한 평가로 각 방법별 데이터 유용성 변화, ϵ,δ DP와 GDP의 성능 비교, 데이터 크기(산업분류별) 차이에 따른 성능을 비교할 것이다.

ϵ은 $\{0.01, 0.05, 0.1, 0.5, 1, 3\}$ 6개 범주에 따라 변화량 측정 및 성능을 비교하고, $\delta = 10^{-5}$는 고정(GDP는 μ값 지정, μ는 ϵ과 δ값으로 계산 가능), 반복횟수는 10,000번으로 Hyper-parameter를 설정한다. 또한 두 값들의 비교를 위한 데이터 유용성 성능 측도로 MAE(Mean Absolute Error), RMSE(Root Mean Squared Error), MAPE(Mean Absolute Percentage Error)를 사용하였다.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x|,\ \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x)^2}.\quad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x}{x}\right| \times 100\% \qquad \textbf{(6.1)}$$

<표 6.3>는 당기순이익, <표 6.4>은 재무활동으로 인한 현금 흐름표 전체 데이터에 ϵ,δ DP와 μ GDP을 적용한 결과로 두 방법 모두 ϵ 값이 작을수록 Original data와 차이가 크게 나타나 정보 보호 효과는 높지만 데이터 유용성 떨어지는 효과를 보인 반면 ϵ 값이 커질수록 Original data와 유사해져 데이터 유용성은 높지만 정보 보호 효과는 떨어짐을 확인 할 수 있다.

ϵ,δ DP와 μ GDP의 성능은 ϵ 크기와 상관없이 μ GDP가 ϵ,δ DP보다 MAE, RMSE, MAPE 평가 측도가 작음을 알 수 있다. 이는 μ GDP 방법으로 정보 보호 기술을 적용했을 때 데이터 유용성 측면에서 더 유용한 결과임을 보이는 결과이다. 특히 ϵ 값이 0.1이하일 경우 그 차이는 더 커지고 있으며, ϵ 값이 0.5보다 커지면서 차이는 점점 줄어들어 Original data와 비슷한 값을 보임을 알 수 있다.

<표 6.3> 당기순이익_ϵ,δ DP, μ GDP

(단위 : 백만원, %)

			ϵ=0.01	ϵ=0.05	ϵ=0.1	ϵ=0.5	ϵ=1	ϵ=3
Original	Mean	4,356.9						
ϵ,δ DP	Mean		4,460.5	4,307.8	4,356.8	4,356.1	4,359.6	4,356.6
	MAE		20,387.6	4,178.6	2,041.4	417.0	208.1	69.2
	RMSE		25,553.6	5,253.8	2,562.7	521.7	261.1	86.6
	MAPE(%)		467.9	95.9	46.9	9.6	4.8	1.6
μ GDP	Mean		4,437.0	4,355.2	4,376.8	4,352.7	4,354.4	4,355.7
	MAE		10,654.8	2,484.8	1,314.9	306.2	160.4	60.0
	RMSE		13,348.5	3,114.3	1,647.9	384.3	201.4	75.5
	MAPE(%)		244.6	57.0	30.2	7.0	3.7	1.4

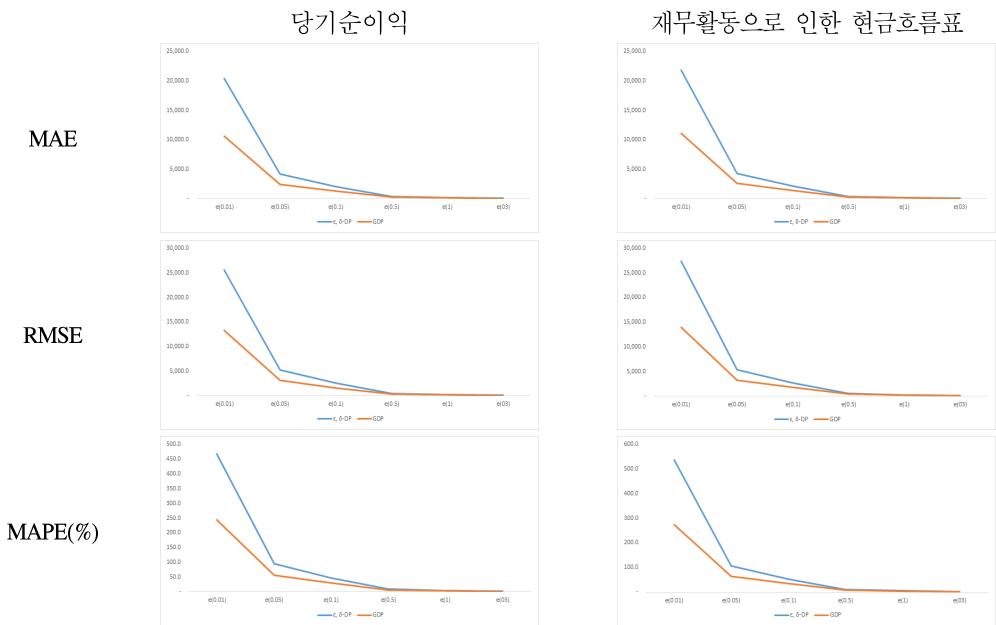
※ ϵ=0.01(μ=0.00404732), ϵ=0.05(μ=0.0170583), ϵ=0.1(μ=0.03212011), ϵ=0.5(μ=0.139521), ϵ=1(μ=0.2653662), ϵ=3(μ=0.7056499)

<표 6.4> 재무활동으로 인한 현금 흐름표_ϵ,δ DP, μ GDP

(단위 : 백만원, %)

			ϵ=0.01	ϵ=0.05	ϵ=0.1	ϵ=0.5	ϵ=1	ϵ=3
Original	Mean	4,051.7						
ϵ,δ DP	Mean		4,135.2	4,044.3	4,031.6	4,046.4	4,051.1	4,051.0
	MAE		21,760.1	4,262.3	2,161.4	435.2	216.2	72.4
	RMSE		27,389.1	5,360.0	2,699.4	545.1	271.0	90.6
	MAPE(%)		537.1	105.2	53.3	10.7	5.3	1.8
μ GDP	Mean		4,094.3	4,055.1	4,042.5	4,045.0	4,048.0	4,050.8
	MAE		11,111.8	2,622.4	1,390.6	323.5	167.0	63.1
	RMSE		13,915.5	3,291.3	1,744.1	405.7	209.9	79.4
	MAPE(%)		274.2	64.7	34.3	8.0	4.1	1.6

※ ϵ=0.01(μ=0.00404732), ϵ=0.05(μ=0.0170583), ϵ=0.1(μ=0.03212011), ϵ=0.5(μ=0.139521), ϵ=1(μ=0.2653662), ϵ=3(μ=0.7056499)



<그림 6.1> ϵ,δ DP, μ GDP의 데이터 유용성 성능 비교

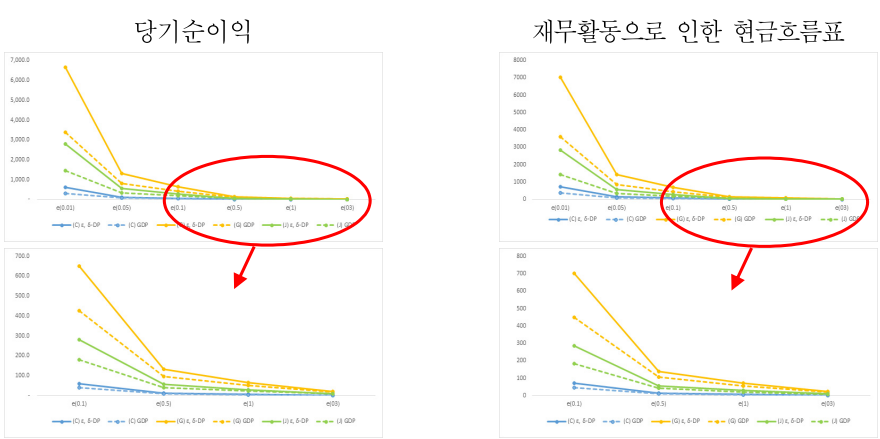
데이터 크기에 따른 두 방법 간의 데이터 유용성 성능을 비교하기 위해서 산업분류별 DP를 적용하였다. <표 6.5>에서 산업분류별 데이터 크기는 C(1,135개) > J(218개) > G(123개) 순으로 Original data 대비 ϵ,δ DP 적용 결과와 μ GDP 적용 결과의 MAPE를 비교하였다. 그 결과 DP 적용 방법에 관계없이 데이터가 클수록 데이터 유용성이 높음을 알 수 있다. 또한 ϵ,δ DP와 μ GDP를 비교하면, μ GDP의 데이터 유용성이 더 좋음을 알 수 있다. 이 결과는 전체 데이터 분석 결과와 같이 ϵ,δ DP와 μ GDP 사이의 차이는 ϵ이 작을 때는 큰 차이를 보이다 ϵ이 커지면 그 차이는 점차 줄어들고 있다.

<표 6.5> 데이터 크기(산업분류)별 ϵ,δ DP, μ GDP의 MAPE 비교

(단위 : 백만원)

	당기순이익						재무활동으로 인한 현금흐름표					
	C		G		J		C		G		J	
	ϵ,δDP	μGDP	ϵ,δDP	μGDP	ϵ,δDP	μGDP	ϵ,δDP	μGDP	ϵ,δDP	μGDP	ϵ,δDP	μGDP
ϵ=0.01	604.5	306.5	6,647.0	3,394.4	2,805.2	1,440.3	713.2	372.1	7,014.5	3,592.6	2,818.0	1,440.7
ϵ=0.05	121.6	71.9	1,307.8	800.1	565.0	342.6	142	87.3	1,414.2	855.2	563.3	340.6
ϵ=0.1	60.4	39.1	649.7	426.4	280.4	179.9	72.5	46.0	702.9	449.8	285.8	182.1
ϵ=0.5	12.2	9.0	131.5	96.6	57.3	41.2	14.4	10.7	139.3	106.6	56.9	42.1
ϵ=1	6.0	4.6	65.7	50.9	28.2	22.1	7.1	5.5	70.4	54.7	28.4	21.7
ϵ=3	2.0	1.8	21.9	18.9	9.5	8.2	2.4	2.1	23.4	20.8	9.5	8.2

※ϵ=0.01(μ=0.00404732), ϵ=0.05(μ=0.0170583), ϵ=0.1(μ=0.03212011), ϵ=0.5(μ=0.139521), ϵ=1(μ=0.2653662), ϵ=3(μ=0.7056499)



<그림 6.2> 데이터 크기(산업분류별)별 ϵ,δ DP, μ GDP의 MAPE 비교

DP 관련 여러 논문에서 데이터 안전성과 유용성을 고려하여 최적의 ϵ 값을 선정하는 기준에 대해 발표된 논문은 아직 없다. 이 사실에 대한 증명은 여전히 진행 중에 있으며, 반드시 해결해야 할 숙제이다. 본 논문에서 모의실험 결과를 통해 알 수 있는 사실은 동일한 ϵ 수준에서도 데이터 유용성이 높은 방법론이 존재한다는 것이다. μ GDP 방법론은 기존 ϵ,δ DP 보다 데이터 유용성 측도인 MAE, RMSE, MAPR 값이

상대적으로 낮게 나타남을 알 수 있었다. 특히 ϵ 값이 작을수록 두 값의 차이는 더 커짐을 알 수 있다. 이 결과를 통해 DP 분야에서 최적의 ϵ 수준에 대한 기준 마련은 어렵지만 동일한 ϵ 에서 기존 방법보다 데이터 유용성을 높여주는 방법을 활용할 수 있다는 결론을 내릴 수 있다.

6.3 Bayesian model Differential Privacy 모의실험 결과

1) 베이지안 모델 생성

베이지안 모델을 활용하여 DP를 충족시키는 재현자료를 생성하고자 한다. <표 6-6>의 각 변수에 대해 정규성 검정을 시행한 결과, 관측값들은 정규분포를 따른다고 가정된다. 평균 μ와 정밀도 λ를 지닌 정규분포의 확률표본들을 $x_1, \cdots, x_n$라 할 때, 우도함수(likelihood function)는 다음과 같이 나타낼 수 있다.

$$p(x_1, \cdots, x_n \mid \mu, \lambda) \sim \prod_{i=1}^{n} N(x_i \mid \mu, \lambda^{-1}) = \prod_{i=1}^{n} \sqrt{\frac{\lambda}{2\pi}} \exp\left\{-\frac{\lambda(x_i-\mu)^2}{2}\right\} \quad \textbf{(6.2)}$$

여기서 정밀도는 분산의 역수이며, 항상 양의 값을 지닌다. 본 논문에서는 알려지지 않은 모수 (μ, λ)에 대해, 정규-감마 사전분포(Normal-Gamma Prior distribution)를 이용한다.

$$\mu \sim N(\mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \quad \textbf{(6.3)}$$

$$\lambda \sim Gam(\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \lambda^{\alpha_0 - 1} \exp\{-\beta_0 \lambda\} \quad \textbf{(6.4)}$$

여기서 $(\mu_0, \sigma_0^2, \alpha_0, \beta_0)$는 (μ, λ)에 대한 사전분포의 초모수(hyper-parameter)이다.

결합 사후분포(joint posterior distribution)는 우도함수에 사전분포를 곱하여 얻을 수 있으며, 다음과 같이 구해진다.

$$
\begin{aligned}
p(\mu,\lambda \mid \boldsymbol{x}) &\propto \left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^{n}\lambda(x_i-\mu)^2}{2}\right\} \\
&\quad * \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \\
&\quad * \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\lambda^{\alpha_0-1}\exp\{-\beta_0\lambda\}
\end{aligned}
\tag{6.5}
$$

$$
\begin{aligned}
&\propto \lambda^{\frac{n}{2}+\alpha_0-1}\exp\left\{-\frac{1}{2}\lambda\left(\sum_{i=1}^{n}x_i^2-\mu\sum_{i=1}^{n}2x_i+n\mu^2+2\beta_0\right)-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \\
&= \lambda^{\frac{n}{2}+\alpha_0-1}\exp\left\{-\frac{1}{2}\lambda\left(\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2+2\beta_0\right)-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \\
&= \lambda^{\frac{n}{2}+\alpha_0-1}\exp\left\{-\frac{1}{2}\lambda\left(ns^2+n(\overline{x}-\mu)^2+2\beta_0\right)-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}
\end{aligned}
\tag{6.6}
$$

이때 위의 함수는 $(\overline{x},\, s^2)=\left(\frac{1}{n}\sum_{i=1}^{n}x_i,\ \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)$에 대한 함수로 나타낼 수 있으며, 따라서 $\boldsymbol{s}=(\overline{x},s^2)$는 베이지안 충분통계량이 된다.

2) 익명화(Sanitization)

ϵ-DP를 만족하는 재현자료를 생성하기 위해, 우리는 라플라스 매커니즘을 활용하여 $Laplace(0,\Delta_1\epsilon^{-1})$로부터 독립적인 확률표본을 생성함으로써 프라이버시 예산 $(\epsilon-\epsilon_0)/m$을 가지고 충분통계량 $\boldsymbol{s}$를 익명화할 수 있다.

ϵ-DP인 라플라스 매커니즘을 통해 익명화된 $\boldsymbol{s}^*$는 $\boldsymbol{s}^*=\boldsymbol{s}+\boldsymbol{e}$이며, 여기서 $\boldsymbol{s}=(s_1,\cdots,s_r)$는 r차 통계량 $\boldsymbol{e}$는 $Laplace(0,\Delta_1\epsilon^{-1})$로부터의 r개의 독립적인 확률표본들, $\Delta_1=\max_{\boldsymbol{x},\boldsymbol{x}',d(\boldsymbol{x},\boldsymbol{x}')=1}\parallel \boldsymbol{s}(\boldsymbol{x})-\boldsymbol{s}(\boldsymbol{x}')\parallel$은 $\boldsymbol{s}$의 l_1 전역 민감성(l_1 global sensitivity, l_1 GS)을 나타낸다.

두 개의 이웃하는 데이터 셋들이 모든 가능한 방식과 모든 가능한 데이터 셋들에 대해 1개의 개체가 차이 난다고 정의될 때, 민감성은 “global”이 된다. $\boldsymbol{s}$에 대해 GS가 더 크다는 것은 기존의 $\boldsymbol{s}$가 드러날 노출위험이 더 크다는 것을 의미하며, 더 큰 민감성을 상쇄하기 위해 $\boldsymbol{s}$에 대해 더 큰 교란이 필요하다. 이에 따라, Δ_1이 커지거나, ϵ이 작아지면 $\boldsymbol{s}^*$의 분포는 넓게 퍼지게 된다.

본 논문에서는 모수(μ,λ)의 분포에 대해 아무런 지식을 가지고 있지 않으므로, 사전분포의 분산(μ의 분산: σ_0^2, λ의 분산: α/β^2)을 $1e+10$으로 매우 크게 설정함으로써 무정보 사전분포(non-informative prior distribution)로 두고 분석을 수행하였다.

이에 따라, 사전분포 1개가 재현자료 생성에 이용되어 $m=1$이 된다. 위에서 설명한 바와 같이, 프라이버시 예산 $(\epsilon-\epsilon_0)/m$을 가지고 충분통계량 $\boldsymbol{s}$를 익명화하기 때문에

$m=1$인 재현자료는 $m>1$인 경우보다 프라이버시 예산이 크고, 잡음의 양이 적게 설정된다.

가우시안 우도함수와 사전분포들에 기반한 사후분포에서의 베이지안 충분통계량은 표본 평균과 분산 $(\overline{x}, s^2)$이며, 이때 l_1 전역 민감성은 각각 $(c_1-c_0)/n$, $(c_1-c_0)^2/n$이 된다. 여기서 $[c_0, c_1]$은 가우시안 데이터의 경계로 설정된다.

<표 6.6> $\overline{x}$에 적용되는 산업 대분류 별 l_1 전역 민감성

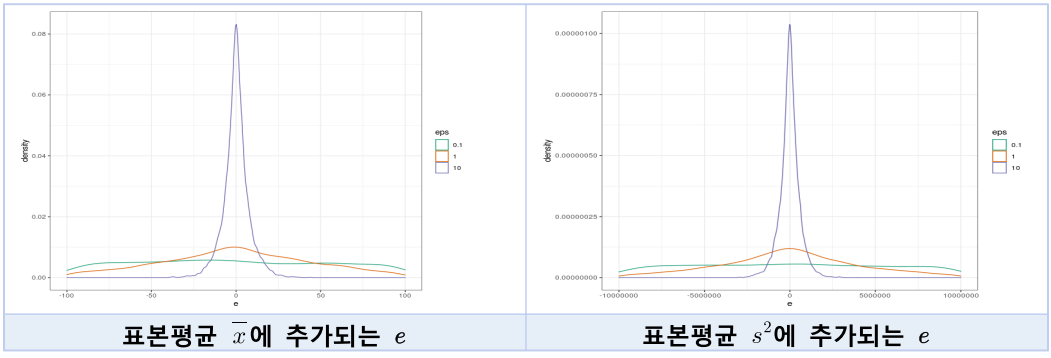
(단위 : 백만원)

항목명 \ 산업대분류	C 제조업	G 도매 및 소매업	J 정보통신업	전체(C+G+J)
당기순이익(손실)	53.39	51.63	47.02	53.39
재무활동으로 인한 현금흐름	56.11	49.29	55.02	56.27

<표 6.7> s^2에 적용되는 산업 대분류 별 l_1 전역 민감성

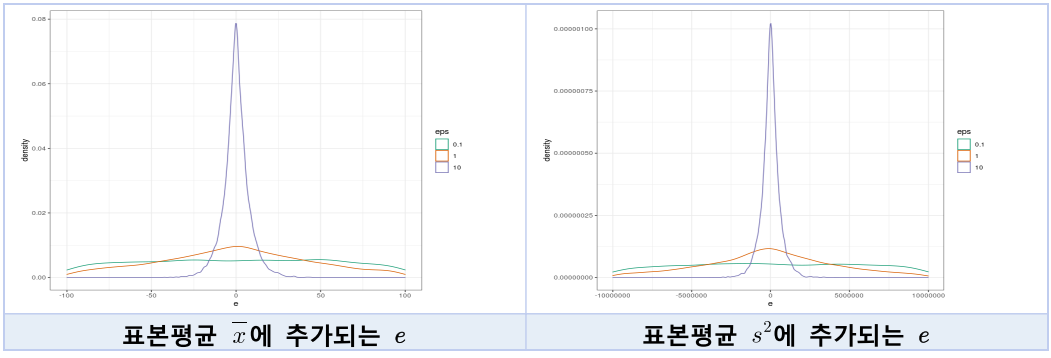
(단위 : 백만원)

항목명 \ 산업대분류	C 제조업	G 도매 및 소매업	J 정보통신업	전체(C+G+J)
당기순이익(손실)	4,207,979.50	3,934,741.15	3,263,038.32	4,207,979.50
재무활동으로 인한 현금흐름	4,646,129.32	3,585,772.36	4,468,007.02	4,673,708.93



표본평균 $\overline{x}$에 추가되는 e **표본평균 s^2에 추가되는 e**

<그림 6.3> 당기순이익(손실) 내 $\boldsymbol{s}=(\overline{x}, s^2)$에 추가되는 $Laplace\left(0, \Delta_1\epsilon^{-1}\right)$ 밀도



표본평균 $\overline{x}$에 추가되는 e **표본평균 s^2에 추가되는 e**

<그림 6.4> 재무활동으로 인한 현금흐름 내 $s=(\bar{x}, s^2)$에 추가되는 $Laplace\left(0, \Delta_1\epsilon^{-1}\right)$ 밀도

전체 산업(C+G+J)을 대상으로 '당기순이익(손실)'과 '재무활동으로 인한 현금흐름' 변수 내 충분통계량 $s=(\bar{x}, s^2)$에 추가되는 $\mathcal{L}\left(0, \Delta_1\epsilon^{-1}\right)$ 밀도를 살펴본 결과, 표본평균보다 표본 분산의 l_1 전역 민감성이 더 크게 나타나므로(기존의 s가 드러날 노출 위험이 더 크게 나타나므로) s에 대해 더 큰 교란이 필요하며, e의 분포가 넓게 퍼지는 것을 확인할 수 있다.

3) 재현자료 생성

깁스샘플링에 사용되는 μ와 λ의 완전 조건부 분포(Full Conditional Distribution, FCD)는 다음과 같다.

$$\begin{aligned} p(\mu \mid \lambda, \boldsymbol{x}) &\propto \exp\left\{-\frac{1}{2}\mu^2\left(\lambda n+\sigma_0^{-2}\right)+\mu\left(\lambda\sum_{i=1}^{n}x_i+\mu_0\sigma_0^{-2}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left\{\mu^2\left(\lambda n+\sigma_0^{-2}\right)-2\mu\left(\lambda\sum_{i=1}^{n}x_i+\mu_0\sigma_0^{-2}\right)\right\}\right\} \\ &\sim N\left(\frac{\lambda n\bar{x}+\mu_0\sigma_0^{-2}}{\lambda n+\sigma_0^{-2}}, \left(\lambda n+\sigma_0^{-2}\right)^{-1}\right) \end{aligned} \tag{6.7}$$

$$\begin{aligned} p(\lambda \mid \mu, \boldsymbol{x}) &\propto \lambda^{\frac{n}{2}+\alpha_0-1}\exp\left\{-\frac{1}{2}\lambda\left(ns^2+n(\bar{x}-\mu)^2+2\beta_0\right)\right\} \\ &\sim Gam\left(\frac{n}{2}+\alpha_0,\ \frac{ns^2+n(\bar{x}-\mu)^2}{2}+\beta_0\right) \end{aligned} \tag{6.8}$$

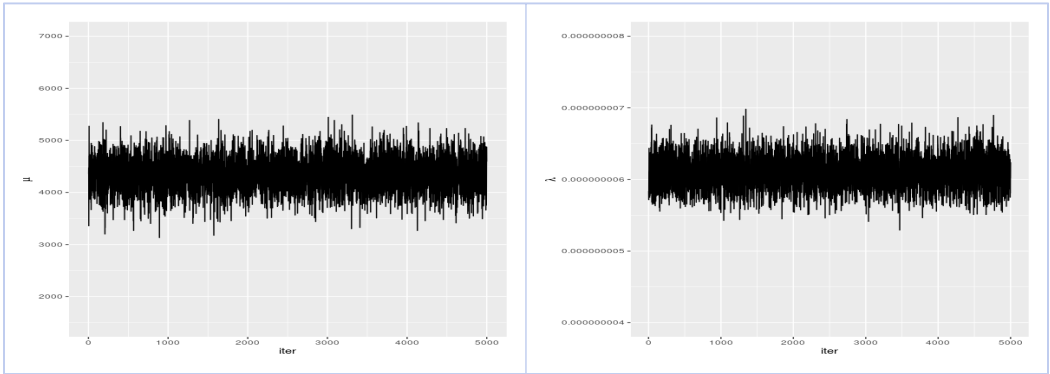
깁스 샘플링 알고리즘은 다음과 같이 작동된다.

[1단계] 초기값 μ_0, λ_0을 설정
[2단계] $m=1$로 설정
[3단계] λ_{m-1}을 사용하여 $p(\mu \mid \lambda, \boldsymbol{x})$로부터 μ_m를 샘플링
[4단계] μ_{m-1}를 사용하여 $p(\lambda \mid \mu, \boldsymbol{x})$로부터 λ_m를 샘플링
[5단계] $m=m+1$로 설정
[6단계] 3단계부터 5단계까지 M번 반복

위 절차에 따라, μ, λ의 사후 기댓값은 다음과 같이 얻을 수 있다.

$$\begin{aligned} \hat{\mu} &= \hat{E}(\mu \mid \underline{x}) = \frac{1}{M-b}\sum_{m=b+1}^{M}\mu_m, \\ \hat{\lambda} &= \hat{E}(\lambda \mid \underline{x}) = \frac{1}{M-b}\sum_{m=b+1}^{M}\lambda_m \end{aligned} \tag{6.9}$$

여기서 b은 burn-in 구간이고, $M=5000$, $b=4000$으로 설정하였다.



<그림 6.5> μ와 λ의 trace plot

<그림 6.5>의 trace plot 결과를 보면, 깁스 샘플링 수행 결과 μ와 λ가 정상적으로 수렴함을 확인할 수 있다.

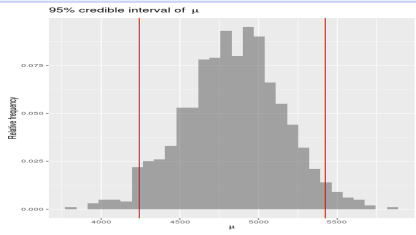
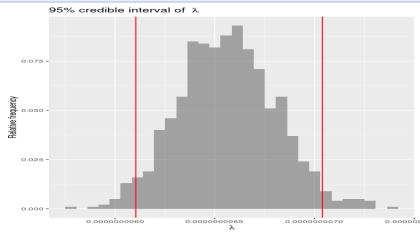
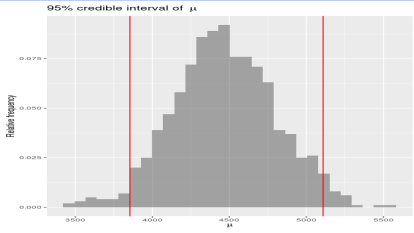
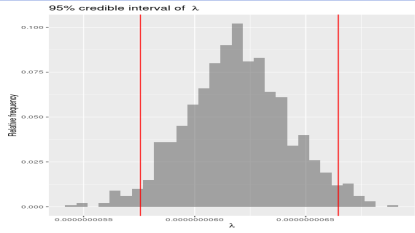
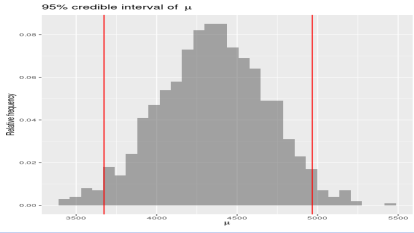
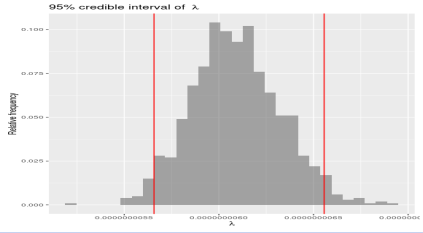
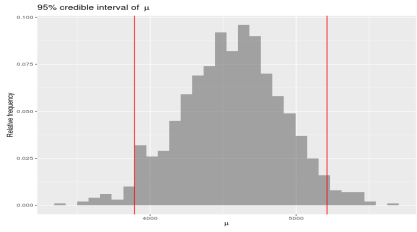
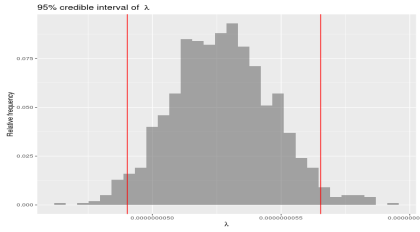
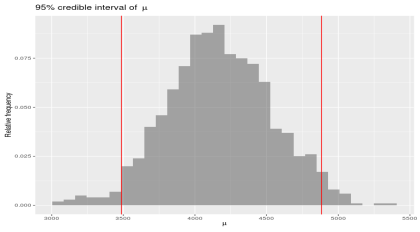
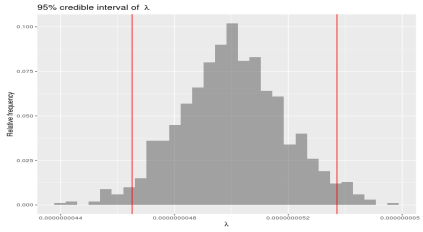
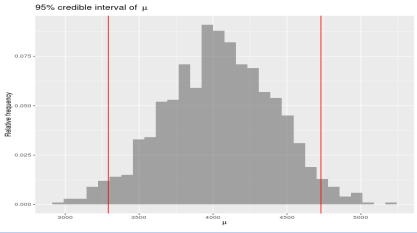
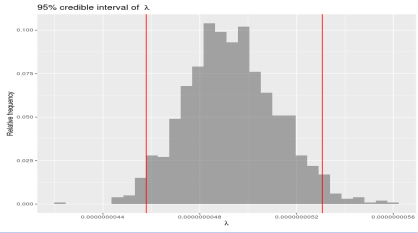
DP를 충족시키는 재현자료는 베이지안 충분통계량 $\boldsymbol{s}=(\overline{x}, s^2)$을 라플라스 매커니즘을 통해 익명화된 $\boldsymbol{s}^*=\boldsymbol{s}+\boldsymbol{e}$로 대체함으로써 간단하게 생성할 수 있다. 우리는 ϵ-DP인 라플라스 매커니즘을 통해 익명화된 $\boldsymbol{s}^*$에 기반하여, 결합 사후분포 $p(\mu,\lambda \mid \boldsymbol{s}^*)$로부터 사후표본 $\mu_m,\lambda_m,\ m=1,\cdots,M$ 를 생성한다.

<표 6.8> DP 수준별 베이지안 추정량(사후 기댓값)

(단위 : 백만원)

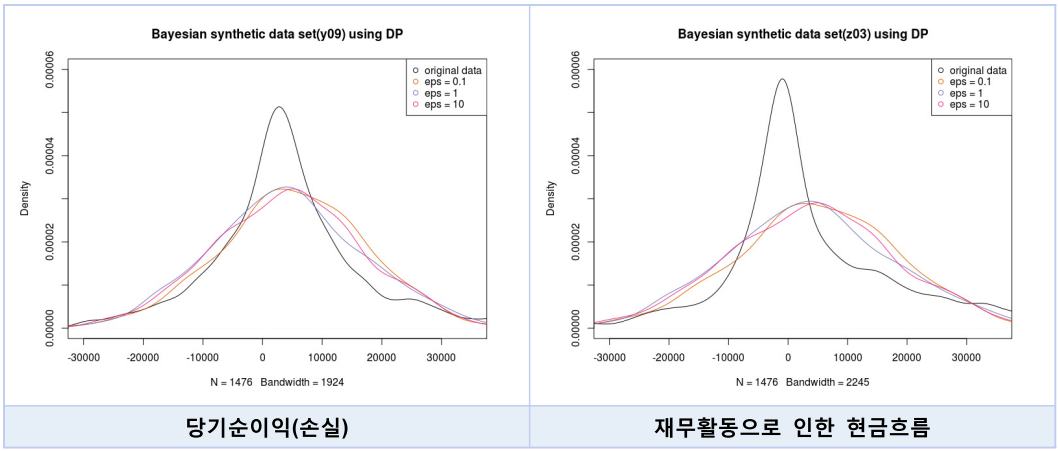
구분 / 항목명	모수	원자료 (표본평균/분산)	베이지안 추정량(사후 기댓값)		
			$\epsilon=0.1$	$\epsilon=1$	$\epsilon=10$
당기순이익(손실)	$\hat{\mu}$	4,356.87	4,838.92	4,468.82	4,348.23
	$1/\hat{\lambda}$	164,224,210.03	152,340,333.14	161,545,480.38	164,158,464.03
재무활동으로 인한 현금흐름	$\hat{\mu}$	4,051.75	4,558.05	4,170.51	4,042.24
	$1/\hat{\lambda}$	202,924,845.66	189,690,723.97	199,904,953.59	202,824,765.91

<표 6.8>에서 베이지안 추정량은 베이지안 충분통계량에 DP(ϵ=0.1,1,10)를 적용한 후 깁스 샘플링을 통해 추정한 사후 기댓값을 의미한다. ϵ값이 커질수록 베이지안 추정량이 표본평균 및 분산과 가까워지는 것을 확인할 수 있으며, 이에 따라 ϵ값이 커질수록 원자료의 분포와 유사해지는 것을 알 수 있다.

항 목	ϵ	μ	λ
당기 순이익 (손실)	0.1		
	1		
	10		
재무 활동 으로 인한 현금 흐름	0.1		
	1		
	10		

<그림 6.6> 항목별 DP를 적용한 μ, λ의 95% 신뢰구간

위 추정량들을 통해 구해진 재현자료의 분포는 다음과 같다.



<그림 6.7> 항목별 DP를 충족시키는 재현자료의 분포

4) 유용성 평가

본 논문에서는 위에서 생성한 재현자료의 유용성을 평가하고 그 결과를 비교하기 위해, 평균, 사분위수 등의 통계량과 함께 95% 신뢰구간 중첩성(Interval overlap) 측도 값을 비교하고자 한다.

신뢰구간(confidence interval, CI)은 표본집단에서 얻은 통계치를 이용하여 모집단의 모수가 어느 범위 안에 있는지 확률적으로 보여주는 구간이며, 자료들의 신뢰구간은 다음의 공식을 통해 구할 수 있다.

$$\left(\overline{X} - tS_{\overline{X}},\ \overline{X} + tS_{\overline{X}}\right) \qquad \textbf{(6.10)}$$

여기서 $\overline{X}$는 표본평균, t는 유의수준 $\alpha = 0.05$, 자유도 $df = n-1$일 때의 t값, $S_{\overline{X}} = \frac{S}{\sqrt{n-1}}$는 표준편차 S를 이용하여 계산한 수정표준오차이다.

신뢰구간 중첩성 측도 값은 원자료와 재현자료의 개별 추정량 각각에 대해 신뢰구간 공식을 이용하여 95% 신뢰구간을 추정하고, 두 신뢰구간의 교집합을 계산함으로써 얻어진다.

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)} \qquad \textbf{(6.11)}$$

여기서 (L_o, U_o)는 원자료의 95% 신뢰구간, (L_s, U_s)는 재현자료의 95% 신뢰구간을 의미한다. (L_i, U_i)는 원자료와 재현자료를 통해 구해진 두 신뢰구간의 교집합이다. 위

식에서 원자료와 재현자료의 신뢰구간이 중첩될수록 I 측도 값은 1에 가까워지고, 두 신뢰구간이 중첩되지 않으면 I 측도 값은 0이 된다. ϵ-DP를 만족하는 재현자료를 생성하고 DP 수준별 통계량을 비교한 결과는 다음과 같다.

<표 6.9> DP 수준별 통계량 비교

(단위 : 백만원)

구분 / 항목명	통계량	원자료	재현자료		
			$\epsilon = 0.1$	$\epsilon = 1$	$\epsilon = 10$
당기순이익 (손실)	평균	4,356.87	5,170.54	4,114.67	4,326.07
	분산	164,224,210.03	157,881,065.27	160,043,386.76	162,677,248.71
	최솟값	-31,806.73	-37,530.18	-30,685.08	-44,206.40
	제1사분위수	-2,348.38	-2,591.03	-4,548.15	-4,587.47
	중앙값	3,314.23	5,128.45	4,013.85	4,484.55
	제3사분위수	9,986.03	13,581.50	12,254.87	12,854.90
	최댓값	47,003.02	44,778.63	38,944.49	49,409.25
재무활동으로 인한 현금흐름	평균	4,051.75	4,928.10	3,776.56	4,017.61
	분산	202,924,845.66	196,589,917.79	198,046,183.30	200,994,661.01
	최솟값	-36,258.99	-42,720.57	-34,934.98	-49,928.64
	제1사분위수	-3,202.39	-3,732.84	-5,860.04	-5,890.23
	중앙값	-16.67	4,881.13	3,664.41	4,193.77
	제3사분위수	11,195.36	14,313.68	12,831.79	13,497.82
	최댓값	46,797.58	49,125.78	42,521.55	54,129.79

ϵ값이 커질수록 재현자료의 평균 및 분산이 원자료와 유사해지고, 이에 따라 재현자료의 유용성이 높아지는 것을 확인할 수 있다. 이는 ϵ의 값이 커질수록 $\boldsymbol{s}$에 대한 교란이 적어지기 때문으로 파악된다. 95% 신뢰구간 중첩성을 확인해 본 결과는 다음과 같다.

<표 6.10> DP 수준별 신뢰구간 중첩성 비교

(단위 : 백만원)

구분 / 항목명	통계량	원자료	재현자료		
			$\epsilon = 0.1$	$\epsilon = 1$	$\epsilon = 10$
당기순이익 (손실)	신뢰구간(하한)	3,702.6	4,390.8	3,329.6	3,534.6
	신뢰구간(상한)	5,011.2	5,950.3	4,899.7	5,117.6
	중첩성(I)	1	0.44	0.84	0.91
재무활동으로 인한 현금흐름	신뢰구간(하한)	3,324.4	4,058.0	2,903.3	3,137.8
	신뢰구간(상한)	4,779.1	5,798.2	4,649.8	4,897.4
	중첩성(I)	1	0.46	0.84	0.91

ϵ의 값이 커질수록 신뢰구간의 중첩성이 1에 가까워지고, 이에 따라 재현자료의 유용성이 높아지는 것을 확인할 수 있다.

7. 결론

데이터는 우리 생활 속에서 쉽게 만들어지고 이러한 데이터의 다 출처 연계를 통해 많은 경제·사회 현상들을 분석하는 기초 자료로 사용되고 있다. 물론 이 과정에서 데이터의 (개인 혹은 기업) 정보 보호는 필수적이고 정보 보호를 위한 많은 방법들이 활용되고 있다. 과거에는 데이터를 공개하기 위해서 정보 보호가 어렵다고 생각되는 자료는 삭제하거나 감춘 후 제공하여 원자료의 특성을 제대로 반영할 수 없었다.

DP 알고리즘은 원자료의 특성에 잡음을 입력하는 방법으로 기존 비식별화 방법을 보완하기 위해서 제안되었다. 이는 데이터 정보 손실을 방지할 수 있으며, 데이터 특성을 고려한 잡음을 추가하기 때문에 원자료와 유사한 값들을 만들어 낼 수 있다. 하지만 아직 이 분야에서는 데이터 안전성과 유용성을 위한 적정 수준의 ϵ, 최적의 생성 및 평가 방법이 정해져 있지 않다는 한계점이 있다.

본 논문에서는 데이터 유용성 측면에서 효과가 있는 GDP와 Modips의 알고리즘과 평가 방법, 그리고 공개 데이터를 이용한 모의실험을 통해 실제 데이터에서의 두 알고리즘의 성능을 평가하였다.

μ GDP를 이용하는 방법은 기존 ϵ,δ DP보다 데이터 유용성이 좋음을 실증분석을 통해 확인하였다. 데이터 크기와 ϵ의 수준에 관계없이 μ GDP의 성능은 탁월하다. 특히 ϵ이 작은 구간에서 그 차이는 더 크게 나타난다. 그 이유는 잡음을 입력할 때 이용하는 알고리즘의 분산이 기존 ϵ,δ DP보다 작게 설정되어 있기 때문이다. 현재 DP 방법에서 최적의 ϵ 기준을 고려하기 어려운 상태에서 μ GDP는 프라이버시와 유용성을 같이 고려할 수 있는 좋은 방법으로 DP 적용 시 더 나은 결과를 얻을 수 있을 것이다.

Modips는 DP가 적용된 재현자료를 생성하는 방법으로 데이터를 활용하는 측면에서 가치가 높다(데이터 유용성이 높다)라고 할 수 있다. Modips는 데이터를 활용하는 사람들이 원하는 분석을 수행할 수 있도록 재현 데이터 셋을 제공하며, DP 매커니즘을 통해 익명화 과정을 수행함으로써 프라이버시를 더 강력하게 제어할 수 있다.

본 논문에서 익명화 과정은 베이지안 충분통계량에 DP 매커니즘을 적용함으로써 이루어지며, 이를 통해 DP를 충족시키는 재현자료를 생성할 수 있었다. DP 수준 별 재현자료 생성 결과, ϵ값이 커질수록 재현자료의 평균 및 분산이 원자료와 유사해지고, 신뢰구간의 중첩성이 1에 가까워지는 등 재현자료의 유용성이 높아지는 것을 확인할 수 있었다.

DP 기준을 충족시키는 재현자료는 기존 재현자료보다 데이터 유용성이 낮다는 한계가 있다. 그러나, ϵ 수준에 기반 한 데이터 유용성을 살펴보고, 자료의 안전성과 유용성의 적절 수준을 정한다면 DP가 적용되지 않은 기존의 재현자료보다 정보 노출을 더 강력하게 제어할 수 있을 것이다.

우리가 데이터를 활용하여 분석할 때 분석과 관련된 부문에서 제공하는 추정량을

사용하거나 혹은 실제 데이터를 이용하는 경우도 있다. 즉, 분석 방향에 따라 사용되는 자료가 다르듯이 DP를 적용할 때 통계 작성 목적에 따른 전략적인 대응이 필요하다. 앞 서 논문에서 제시했던 두 방법 중 추정량이 필요한 경우 μ GDP, 데이터 셋이 필요한 경우 Modips를 적용한다면 효과적인 DP 적용이 가능할 것으로 보인다.

(2023년 7월 18일 접수, 2023년 8월 25일 수정, 2023년 11월 28일 채택)

참고문헌

박민정, 이용희, 권성훈 (2018). 차등정보보호에 관한 연구. 통계개발원 2018 연구보고서.

정강수, 박석 (2018). 차분 프라이버시 기반 비식별화 기술에 대한 연구, 정보보호학회, 제28권 제2호.

C. Dwork (2006). Differential privacy, International Colloquium on Automata, Languages, and Programming, pp. 1-12.

C. Dwork (2008). Differential Privacy: A Survey of Results, International Conference on Theory and Applications of Models of Computation, pp. 1-19.

Fang Liu (2022). Model-based Differentially Private Data Synthesis and Statistical Inference in Multiply Synthetic Differentially Private Data, Transaction on Data Privacy 15(3). 141-175.

Haoran Li, Li Xiong, Xiaoqian Jiang (2014). Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions, International conference on extending database technology, vol 2014.

James Jordon, Jinsung Yoon, Mihaela van der Schaar (2019). PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees, ICLR 2019 Conference.

Jinshuo Dong, Aaron Roth, Weijie F. Su (2022). Gaussian differential privacy, Journal of the Royal Statistical Society Series B: Statistical Methodology, volume 84, Issue 1, February 2022, 3-37.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, Xiaokui Xiao (2017). PrivBayes: Private Data Release via Bayesian Networks, ACM Transactions on Database Systems, Issue 4, Article No:25, 1-41.

Martin Abadi, Andy Chu, Ian Goodfellow (2016). Deep Learning with Differential Privacy, ACM, 308-318.

Nazmiye ceren Abay, Yan Zhou, Bhavani M.Thuraisingham (2018). Privacy Preserving Synthetic Data Release Using Deep Learning, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 510-526.

A Case Study on Differential Privacy Methods with Improved Data Usability

Seong Ryul Park[3] Yeong Ju Oh[4]

Abstract

Abstract Differential Privacy is a de-identification technology that adds noise to data. The technology can control the level of privacy to the level of hyper-parameter ϵ, and adds noise using ϵ and global sensitivity. The amount of noise added determines the level of data privacy, and a lot of noise strengthens the privacy of the data, but the usefulness of the data decreases. The reason is that privacy and usefulness are in a trade-off relationship. In this paper, we would like to evaluate the data usability performance of Gaussian Differential Privacy, which increases usefulness at the same privacy level, and Modips, based on Bayesian models that generate reproduction data. We introduce the existing Differential Privacy, Gaussian Differential Privacy, and Bayesian model-based Differential Privacy algorithms, calculating the statistics applied by the two methods using actual data, the data usefulness performance is analyzed using a usefulness evaluation measure.

Key words : data de-identification, Gaussian, Baysian model, data usefulness

[3] Statistical Methodology Division, Statistical Research Institute, Statistics Korea, 6F, Statistical Center, 713 Hanbatdaero, Seo-gu, Daejeon 35220, E-mail : srpark08@korea.kr

[4] Statistics Korea, Statistical Data Planning Division, Government Complex-Daejeon, 189 Cheongsa-ro, Seogu, Deajeon, 35208, E-mail : oyj1928@korea.kr