Differential Privacy

Xiaotong Shen
xshen@umn.edu
School of Statistics
University of Minnesota

Joint with Xuan Bi

SC Big Data Health Science Center Conference, 2023

Main references

Shen, X., Bi, X., and Shen, R. (2022). Data Flush. *Harvard Data Science Review*, 4(2). https://hdsr.mitpress.mit.edu/pub/x1ozqj10.

Bi, X. and Shen, X. (2022). Distribution-invariant differential privacy. *Journal of Econometrics*.

Data privacy

- Data privacy has become important, especially in health care.
- Benefits of privacy protection:
  - Promoting data sharing;
  - Addressing legal, ethical, and fairness-related concerns;
  - Configuring data-security policies.
- Some popular methods:
  (1) Differential privacy;
  (2) Secure multi-party computation–Federated learning;
  (3) Homomorphic encryption;
- Benefits of Differential privacy:
  - Executable without the involvement of other parties;
  - Effective for broad applications;
  - One technical solution to data privacy at Low cost.

Differential privacy

Differential privacy [Dwork, 2008] quantifies amount of privacy protection for downstream data analysis:

- Recognizes that privacy can be undermined even after data de-identification; e.g., "tallest person in room" is an identifier)
- Privatization mechanism m satisfies ε-*differential privacy*:

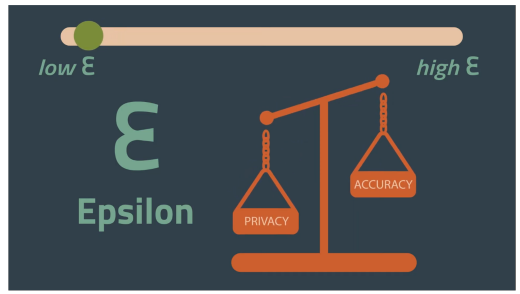
$$\frac{P\big(m(\boldsymbol{Z}) \in B | \boldsymbol{Z} = \boldsymbol{z}\big)}{P\big(m(\boldsymbol{Z}) \in B | \boldsymbol{Z} = \boldsymbol{z}'\big)} \leq e^{\varepsilon},$$

for event B & adjacent $\boldsymbol{z}$, $\boldsymbol{z}'$ (substitute a single observation)

  - ε: budget of protection. Small $\varepsilon \to$ strict privacy protection but may reduce statistical accuracy of downstream analysis.
  - Lemma (Privacy leakage): Any hypothesis test to identify Z_{i0}'s value by testing $H_0 : Z_{i_0} = \mu_0$ vs $H_a : Z_{i_0} = \mu_1 \neq \mu_0$, based on iid copy of $\boldsymbol{Z}$ $\tilde{\boldsymbol{Z}}^{(1)}, \ldots, \tilde{\boldsymbol{Z}}^{(M)}$ has power $\leq \alpha e^{M\varepsilon}$ for significance level $\alpha > 0$.

Trade-off between protection and accuracy

- Consensus view
  - Trade-off: Large noise (small budget ε) $\rightarrow$ better privacy protection but less accuracy (less usefulness) for data sharing.



- Differentially private mechanisms:
  - Laplace mechanism: $M(\boldsymbol{Z}) = f(\boldsymbol{Z}) + (e_1, \cdots, e_p)$, e_i iid $\sim Laplace(1/\varepsilon)$ [Dwork et al., 2006, Dwork and Roth, 2014]
  - Exponential mechanism: Sampling from a certain exponential distribution. [McSherry and Talwar, 2007]
  - Minimax optimal procedure: through conditional independence construction [Duchi et al., 2018]
  - Synthetic (imputation) data methods [Wasserman and Zhou, 2010, Snoke and Slavković, 2018, Gong and Meng, 2020, Liu et al., 2021]
  - ...

Distribution invariant privatization (DIP)

- Question: Can we design a differentially private mechanism preserving $\boldsymbol{Z}$'s distribution?
- DIP : Univariate $Z_i \sim$ unknown cont CDF F.

$$\text{Base distr} : U_i \sim U(0,1)$$
$$F_{\boldsymbol{V}} : \text{CDF of } \boldsymbol{V} = (V_i)_{i=1}^n (V_i = U_i + e_i)$$
$$\widetilde{Z}_i := \underbrace{\tilde{F}_{\boldsymbol{Z}}^{-1} \circ F_{\boldsymbol{V}}}_{\boldsymbol{G}}(U_{r(i)} + e_i) \sim \tilde{F}_{\boldsymbol{Z}}, \quad e_i \sim Laplace(0, 1/\epsilon).$$

- Generalization
  - Discrete: continualization; Multivariate: normalizing flows, chain rule.
  - Holdout sample: constructing $\tilde{F}_{\boldsymbol{Z}}$ or estimating $\boldsymbol{G}$ via NF.
    1. public data $\sim$ same distribution: e.g., American Community Survey (public) & Census (private) are from same population.
    2. subset of raw sample: never allowed to be altered, queried, or released; rest sample is privatized & released.

DIP

Laplace noise $e \sim Laplace(0, 1/\epsilon)$ for ε-differentially private:

Theorem 1 (DIP, [Bi and Shen, 2022])

1. *DIP is ε-differentially private; privatized sample follows $\tilde{F}_{\mathbf{Z}} \approx F$.*
2. *Computational complexity of DIP is $O(dn \log n)$.*

- No trade-off: privacy protection & statistical accuracy.
- Trade-off: released size & approx accuracy of F if sample splitting.

Numerical examples

- **Laplace mechanism (LPM)** [Dwork et al., 2006, Dwork and Roth, 2014]:
  Good for many types of bounded data & easy to implement;
- **Exponential mechanism (EXM)** [McSherry and Talwar, 2007]:
  Accommodates high sensitivity & works well for discrete/categorical data;
- **Minimax optimal procedure mechanism (OPM)** [Duchi et al., 2018]:
  Work for many canonical families & with exhibit lower and upper bounds on minimax risk;
- Oracle: non-private (NP) $\leftarrow$ based on raw sample.

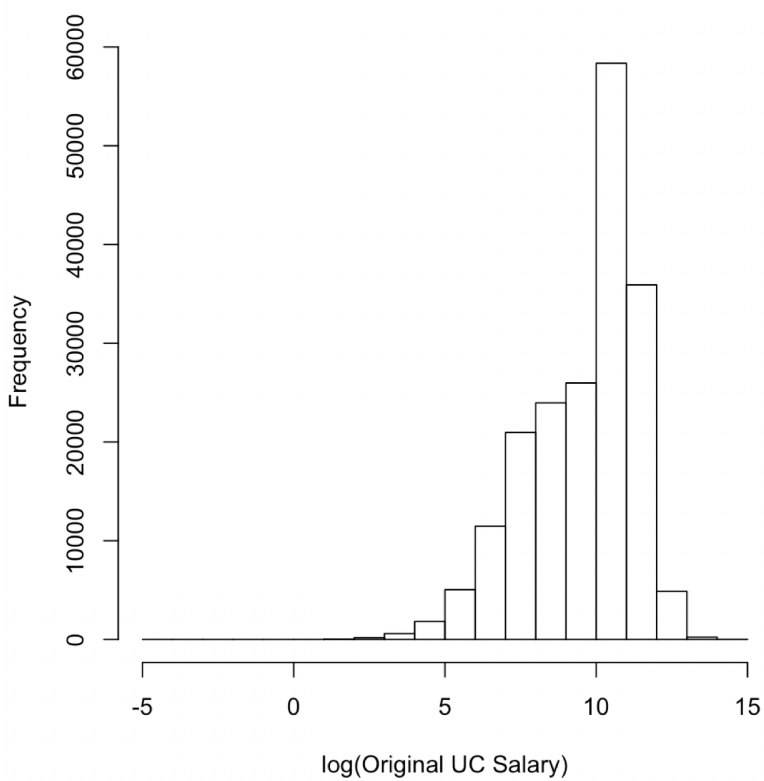
- Note: **DIP** uses 25% and 75% for training and holdout samples, while other methods use 100% for training.

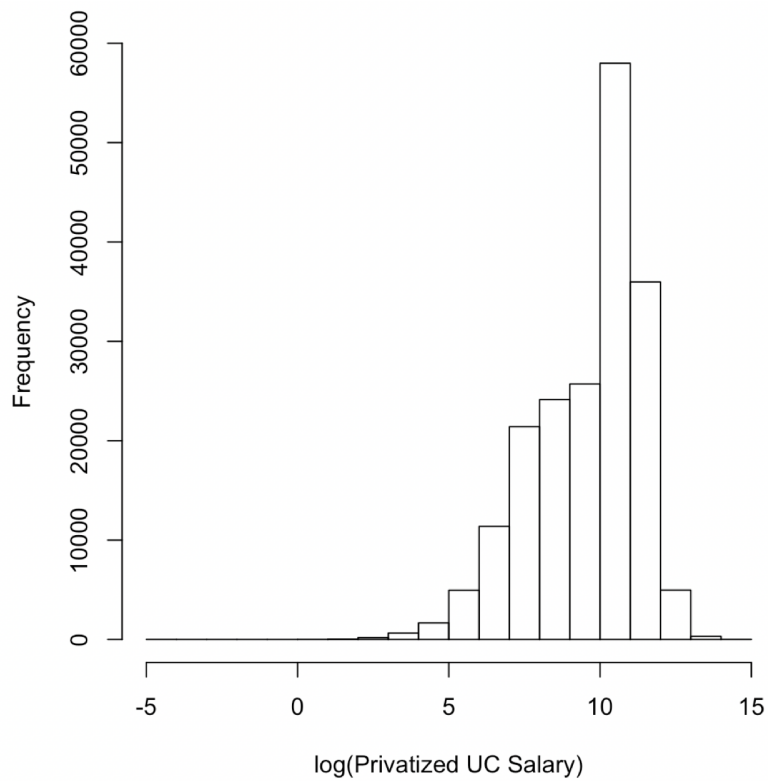
Benchmark: UC salary data

- University of California system salary data
  - collected in 2010
- Annual salaries of $n = 252,540$ employees
- The average salary is \$39,531.49 with a standard deviation of \$53,253.93
- Goal: mean estimation
  - Compare non-private & differentially private mean salaries
- Evaluation metric: Difference between private mean & non-private mean (baseline)

UC salary data

- Histograms of UC salary data before and after DIP.
- Empirical distribution has almost no change while individual data are privatized.





Benchmark: Bank data

- Portuguese bank marketing campaign data
  - Marketing campaign data collected from a Portuguese retail bank from 2008 to 2013 [Moro et al., 2014]
- $N = 30,488$ respondents
- The response variable (binary):
  - interest in a term deposit
- Covariates (continuous, binary, and categorical):
  - Age, employment status, marital status, education, loan status, default status, device type, and past contact histories
- Goal: Compare accuracy between private & non-private logistic regression
- Evaluation metric: Kullback-Leibler divergence

Benchmark: Recommender systems

- MovieLens data
- 25,000,095 movie ratings, collected from 162,541 users over 59,047 movies [Harper and Konstan, 2015]
- Movie ratings with values in $\{0.5, 1, 1.5, \ldots, 5\}$
- No covariates
- Goal: movie recommendation
  - Split the data randomly into a 75% training and a 25% test set
  - Privatize the training data
  - Train a collaborative filtering recommender system
  - Compare prediction accuracy on non-private test set
- Evaluation metric: Root mean square error

Real-world benchmark analysis results

	Dataset		
	UC Salary	Bank Marketing	MovieLens
Size	252,540	30,488	25,000,095
Type	Continuous	Multivariate	Discrete
Dim	1	10	1
Task	Mean Est.	Logistic Reg.	Collaborative Filtering
DIP	0.40 (0.31)	0.047 (0.003)	1.03 (4.96×10^{-4})
LRM	13.08 (9.95)	0.311 (0.004)	1.87 (1.03×10^{-3})
OPM	4.69 (3.44)	Infinity	2.61 (8.23×10^{-4})
EXM	N/A	N/A	1.11 (5.52×10^{-4})

Table 1: Privacy factor $\varepsilon = 1$. DIP holds out 25% data. DIP shows the best performance in differentially private mean estimation, logistic regression, and personalized recommendations.

Note: Essentially no loss of statistical accuracy by DIP privatization while satisfying differential privacy.

Conclusion

- Data Privacy (inference for privatized data)
  - Holdout sample + Laplace noise $\rightarrow$ differential privacy.
- DIP-privatization satisfies the differential privacy standard while retaining statistical accuracy of any downstream analysis.
- More work to expand to unstructured inference (Electrical Medical Records),.., Monte Carlo inference.

Thank you!

References I


Bi, X. and Shen, X. (2022). Distribution-invariant differential privacy. *Journal of Econometrics*.


Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201.


Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.


Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284.


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


Gong, R. and Meng, X.-L. (2020). Congenial differential privacy under mandated disclosure. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 59–70.


Harper, F. M. and Konstan, J. A. (2015). The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19.

References II


Liu, T., Vietri, G., and Wu, S. Z. (2021). Iterative methods for private synthetic data: Unifying framework and new methods. *Advances in Neural Information Processing Systems*, 34:690–702.


McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science*, pages 94–103.


Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31.


Snoke, J. and Slavković, A. (2018). pmse mechanism: differentially private synthetic data with maximal distributional similarity. In *International conference on privacy in statistical databases*, pages 138–159. Springer.


Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.