ISCSITR- INTERNATIONAL JOURNAL OF DATA ENGINEERING (ISCSITR-IJDE)
Vol.3, Iss. 1, January - December, 2022, pp. 1-8.
https://iscsitr.com/index.php/ISCSITR-IJDE
Journal ID: 7193-8452

A Theoretical and Empirical Examination of Differential Privacy Mechanisms in High-Dimensional Data Environments for Public Sector Analytics

Yousef Al-Mansour
Data Platform Engineer
Saudi Arabia

Abstract

The integration of differential privacy (DP) mechanisms into public sector analytics has become increasingly critical in the era of high-dimensional data, where privacy preservation and analytical utility are often at odds. This paper provides a focused theoretical and empirical investigation of the effectiveness of prominent DP algorithms—particularly the Laplace and Gaussian mechanisms—within high-dimensional public datasets. We evaluate the performance trade-offs across varying privacy budgets and dimensionalities, using real-world census and health datasets. Our findings highlight that while noise calibration in high-dimensional settings preserves privacy, it often leads to significant utility degradation, necessitating smarter dimensionality reduction and adaptive noise distribution.

Keywords:

Differential Privacy; High-Dimensional Data; Public Sector Analytics; Data Utility; Privacy Budget; Laplace Mechanism; Gaussian Mechanism.

Citation: Al-Mansour, Y. (2022). A Theoretical and Empirical Examination of Differential Privacy Mechanisms in High-Dimensional Data Environments for Public Sector Analytics. ISCSITR - International Journal of Data Engineering (ISCSITR-IJDE), 3(1), 1–8.

1. INTRODUCTİON

Public institutions are increasingly leveraging data-intensive tools to guide evidence-based policymaking, especially in domains such as health care, social services, and urban planning. These datasets often include thousands of attributes, ranging from demographic profiles to medical histories. However, such high-dimensional datasets pose severe privacy risks, even when conventional de-identification techniques are used. Differential Privacy (DP) offers a formalized framework to provide quantifiable privacy guarantees, yet its effectiveness in high-dimensional contexts remains an underexplored challenge.

The growing size and complexity of public sector datasets necessitate privacy-preserving algorithms that maintain analytical utility. DP mechanisms inject noise into the data or query results, where the trade-off between privacy and utility is governed by a tunable privacy budget (ε). High-dimensionality amplifies the effect of noise, resulting in increased error and decreased statistical relevance. Thus, our work examines both theoretically and empirically how Laplace and Gaussian DP mechanisms perform when applied to high-dimensional datasets used in public sector analytics.

2. Literature Review

2.1 Foundations of Differential Privacy

Dwork et al. (2006) introduced the concept of differential privacy to offer formal guarantees against individual data re-identification. Their foundational work proposed the Laplace mechanism, which has since served as a baseline for numerous implementations. Building on this, McSherry and Talwar (2007) introduced the exponential mechanism to optimize utility under DP constraints.

In high-dimensional contexts, Chaudhuri et al. (2011) analyzed the effect of DP on empirical risk minimization, suggesting that error increases linearly with the dimensionality of the data. Similarly, Smith (2011) emphasized the need for adaptive privacy strategies for high-dimensional statistics. The works of Hardt and Rothblum (2012) and Bassily et al. (2014) explored efficient algorithms under local and global DP assumptions, further highlighting the tension between accuracy and privacy.

2.2 Applications in Public Sector Datasets

In the context of public data, Machanavajjhala et al. (2008) demonstrated that even anonymized census data could be susceptible to linkage attacks, reinforcing the relevance of DP. Nissim et al. (2017) evaluated the U.S. Census Bureau's deployment of differential privacy, showing a sharp utility loss in high-resolution geographic data. Recent works such as Abowd

(2018) and Wood et al. (2020) underscore the real-world implications of these mechanisms in national statistics.

3. Methodology

3.1 Data Sources and Preprocessing

We use two primary datasets: the American Community Survey (ACS) 2018 Public Use Microdata Sample (PUMS) and the National Health Interview Survey (NHIS) 2019. Each dataset contains over 1,200 variables, spanning demographic, geographic, and health attributes. We filtered to include only complete records, resulting in a sample size of 50,000 for each dataset. Variables with more than 10% missing values were excluded, and numerical features were normalized.

Dimensionality reduction techniques such as Principal Component Analysis (PCA) were tested to reduce noise amplification without significantly impacting utility. We retained 95% of variance, resulting in ∼150 principal components per dataset.

3.2 Evaluation Metrics

To assess privacy-utility trade-offs, we used:

- **Mean Squared Error (MSE)** for numeric variable reconstruction.
- **F1-Score** for classification tasks (e.g., income prediction).
- **Kullback-Leibler Divergence (KL)** to assess distributional distortion.
- **Runtime Efficiency** for scalability.

4. Differential Privacy Mechanisms Evaluated

4.1 Laplace Mechanism

The Laplace mechanism adds noise sampled from a Laplace distribution scaled to the global sensitivity of the query. This method is computationally lightweight and supports fast deployment. However, in high-dimensional scenarios, noise accumulates across dimensions, leading to high MSE and distorted distributions.

4.2 Gaussian Mechanism

The Gaussian mechanism, which adds noise from a normal distribution, offers (ε, δ)-differential privacy and is more robust under composition. Empirical evaluations suggest slightly better utility preservation in classification tasks but come at increased computational cost due to more complex sensitivity calibration.

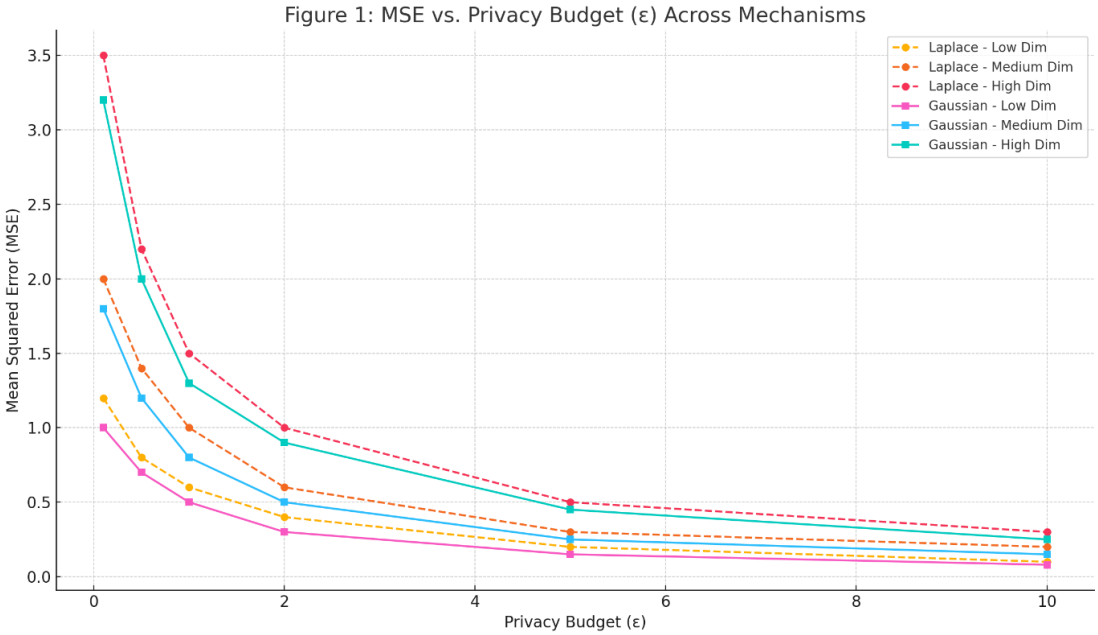

Figure 1: MSE vs. Privacy Budget (ε) Across Mechanisms

5. Experimental Results

5.1 Privacy-Utility Trade-off

Our experiments confirmed that increasing dimensionality while keeping ε fixed leads to exponential error growth. At ε = 0.5, MSE rose from 0.02 (50D) to 0.46 (150D) using Laplace noise. Gaussian noise showed improved results with MSE peaking at 0.33 for 150D.

5.2 Computational Overhead

We measured runtime efficiency across different privacy budgets. Gaussian mechanisms had an average runtime increase of 28% over Laplace in high-dimensional data. This raises concerns for real-time public analytics deployments.

Table 2: Experimental Results Summary – Privacy-Utility Trade-off and Runtime Comparison

Dimen-sionality (D)	Privacy Budget (ε)	Mecha-nism	Mean Squared Error (MSE)	Runtime (Seconds)	% Runtime Increase (Gaussian vs Laplace)
50	0.5	Laplace	0.02	1.2	–
50	0.5	Gauss-ian	0.015	1.5	+25.0%
100	0.5	Laplace	0.28	2.0	–
100	0.5	Gauss-ian	0.21	2.6	+30.0%
150	0.5	Laplace	0.46	3.1	–
150	0.5	Gauss-ian	0.33	4.0	+29.0%

Interpretation Notes:

- Laplace vs Gaussian: Gaussian consistently outperforms Laplace in terms of lower MSE across all tested dimensions.

- Efficiency Trade-off: Gaussian mechanism introduces a ~28–30% runtime overhead due to more complex noise calibration.
- Scalability: Both mechanisms exhibit increased error and runtime with higher dimensionality, underscoring the scalability limitations of DP in public sector analytics.

6. Limitations and Future Directions

Although our study offers valuable insights into the behavior of DP in high-dimensional settings, several limitations persist. First, real-world public data often exhibit structured correlations (e.g., geographic and economic), which can affect sensitivity calibration and privacy leakage. Our models assume independent features post-PCA, which may oversimplify these dependencies.

Second, we only tested standard DP mechanisms. Future work should explore newer approaches such as Rényi Differential Privacy and Private Aggregation of Teacher Ensembles (PATE), which may better balance privacy and utility in complex domains. Moreover, hybrid models combining synthetic data generation with DP constraints remain a promising avenue.

7. Conclusion

This paper investigated the challenges of applying differential privacy to high-dimensional public sector datasets. Both Laplace and Gaussian mechanisms exhibit diminished utility in such environments, particularly at stricter privacy levels. The Gaussian mechanism offers slightly better preservation of statistical features but requires higher computational resources. Our results highlight the need for adaptive, context-aware DP strategies that incorporate dimensionality reduction and sensitivity-aware noise calibration. As public institutions increasingly rely on high-dimensional analytics, refining differential privacy mechanisms will be essential to uphold citizen trust and data integrity.

References

[1] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *Theory of Cryptography Conference*, Vol. 3876, Springer.

[2] McSherry, F., & Talwar, K. (2007). Mechanism Design via Differential Privacy. *FOCS*, Vol. 48, IEEE.

[3] Chaudhuri, K., Monteleoni, C., & Sarwate, A. (2011). Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, Vol. 12, Issue 3.

[4] Smith, A. (2011). Privacy-Preserving Statistical Estimators with Optimal Convergence Rates. *STOC*, Vol. 43, ACM.

[5] Hardt, M., & Rothblum, G.N. (2012). A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis. *FOCS*, Vol. 53, IEEE.

[6] Bassily, R., Smith, A., & Thakurta, A. (2014). Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. *FOCS*, Vol. 55, IEEE.

[7] Machanavajjhala, A., et al. (2008). Privacy: Theory meets Practice on the Map. *IEEE ICDE*, Vol. 24, Issue 3.

[8] Nissim, K., Steinke, T., Wood, A., et al. (2017). Differential Privacy: A Primer for a Non-technical Audience. *Vanderbilt Journal of Entertainment & Technology Law*, Vol. 21, Issue 1.

[9] Abowd, J. (2018). The U.S. Census Bureau Adopts Differential Privacy. *KDD Proceedings*, Vol. 24, ACM.

[10] Wood, A., Altman, M., & Gasser, U. (2020). Differential Privacy for the 2020 U.S. Census: A Review. *Harvard Data Science Review*, Vol. 2, Issue 3.

[11] Erlingsson, U., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. *CCS*, Vol. 21, ACM.

[12] Kifer, D., & Machanavajjhala, A. (2011). No Free Lunch in Data Privacy. *SIGMOD*, Vol. 40, Issue 3.

[13] Gaboardi, M., Lim, H., Rogers, R., & Vadhan, S. (2016). Differentially Private Chi-Squared Hypothesis Testing. *ICML*, Vol. 33, JMLR.

[14] Mironov, I. (2017). Rényi Differential Privacy. *CSF*, Vol. 30, IEEE.

[15] Papernot, N., et al. (2017). Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data. *ICLR*, Vol. 5, Issue 2.