SocialSec23: 9th International Symposium on Security and Privacy in Social Networks and Big Data

Data Reconstruction Attack Against Principal Component Analysis

Saloni Kwatra
Vicenç Torra
saloni.kwatra@umu.se

UMEÅ UNIVERSITY

August 12, 2023

Our Work

We formalize a data reconstruction attack theory against Principal Component Analysis (PCA) by extending a former work about Membership Inference Attack (MIA) against PCA.

Aim of the Membership Inference Attack (MIA)

Given a trained ML model and some data point, decide whether this point was part of the model's training sample or not.

Data Reconstruction attacks

The goal of the adversary in a reconstruction attack is to extract the data used in the training or inferences of a machine learning model.

MIA against Principal Component Analysis(PCA)

- MIA against PCA [2] was studied for the first time
- The attacker intercepts some of the principal components and infers whether a particular sample participated in the computation of principal components.
- The theory is that the samples belonging to the training set will incur lower reconstruction error in comparison with the samples not belonging to the training set.

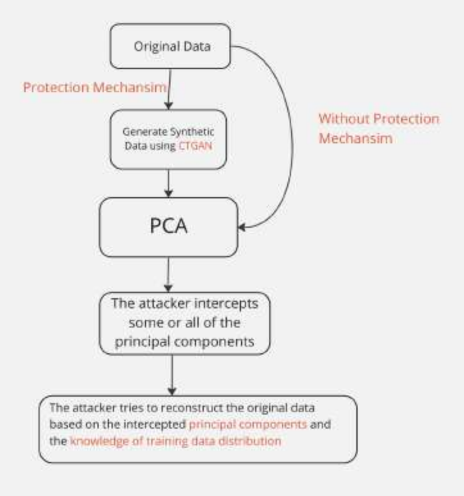
Data Reconstruction Attack

- Suppose X is an original data matrix of size $n \times p$ after subtracting the mean. Let V be the $p \times k$ matrix of some k eigenvectors to reduce the dimension.
- The matrix of PCA projection scores (Z) with the dimension $n \times k$ is $Z = XV$. To reconstruct all the original variables from a subset of principal components/eigenvectors, we can map it back to p dimensions with V^T.
- Reconstructed matrix, $\hat{X} = ZV^T$. Since we have a projection scores matrix, $Z = XV$, we obtain $\hat{X} = XVV^T$.
- We do not have access to the original data X; we assume that the attacker has knowledge about the distribution of X. Therefore, the attacker can synthesize the data X_{syn} with a similar distribution as X and reconstruct the original data using $\hat{X} = ZV = X_{syn}V^TV$.

Generation of Synthetic Data

- We use a Conditional Tabular Generative Adversarial Network (CTGAN) to generate the synthetic data
- To show experimental results, we generate the synthetic data using different percentages of records from the original data, including {10%, 30%, 50%, 70%, 100%}

Attack Methodology



Data reconstruction attack against Principal Component Analysis

Description of datasets

Dataset	**Number of Samples**	**Number of Attributes**
Heart-scale	270	13
Mushrooms	8124	112
a9a	32561	123

Compared Methodologies

- No Protection Mechanism: the data curator uses no protection mechanism at all
- Differentially Private Principal Component Analysis (DPPCA): the data curator applies DPPCA, which involves perturbing the covariance matrix

Reconstruction Accuracy (R.A.)

Definition

Suppose S is the synthetic data obtained after the alignment, and O is the original data. Let n be the total number of samples in the original and the synthetic data, O_j be the value of the sensitive attribute from the original data, which the attacker aims to infer, and S_j is the inferred attribute in the synthetic data corresponding to the sensitive attribute O_j. Let δ be the deviation between the original and the synthetic attribute that can be tolerated to measure the level of inference for a record. The lower the δ, the closer the values of S_j and O_j must be to each other. The Reconstruction Accuracy, $I.A.$, for the continuous attributes, is defined as follows:

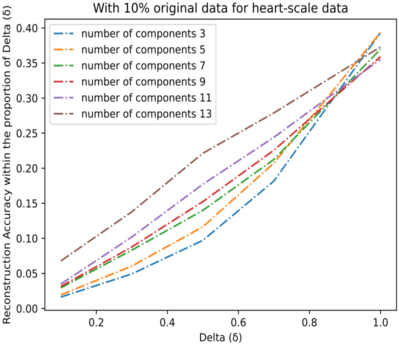
$$R.A. = \frac{\#\left\{\hat{S}_j : |\frac{O_j - S_j}{S_j}| \leq \delta, j = 1 \ldots n\right\}}{n} \quad (1)$$

where $\#$ means count. I.A. is the percentage of inferred entries for which the relative errors are within δ.
For the categorical data, the above formula is more strict (as we are counting only the **exact matches**) and changes to

$$R.A. = \frac{\#\left\{\hat{S}_j : O_j == S_j, j = 1 \ldots n\right\}}{n} \quad (2)$$

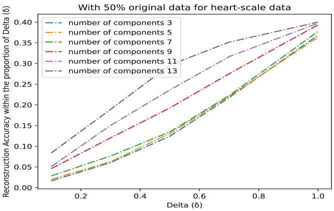
R.A. on the heart-scale data (CTGAN) I

- It is noted that there is not much difference in the R.A. when the CTGAN uses less percentage (e.g., 10%) of samples from the original data compared to using all the samples from the original data for generating the synthetic data.

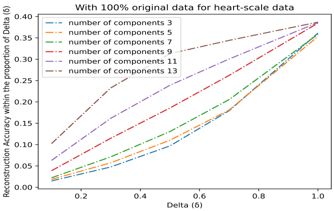


Used 10% of the original data

R.A. on the heart-scale data (CTGAN) II

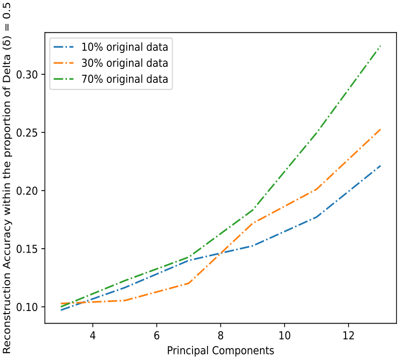


50% of the original data



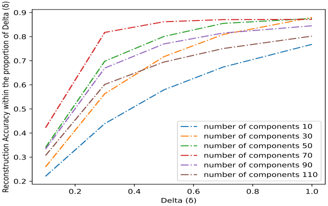
Used 100% of the original data

R.A. on the heart-scale data (CTGAN) III

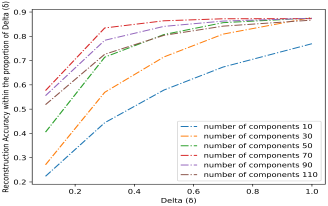


R.A. vs no. of principal components with δ= 0.5

R.A. on the a9a data (CTGAN) I

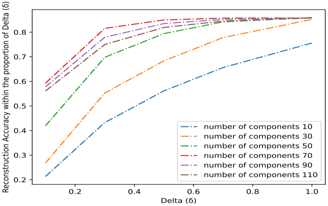


10% original data

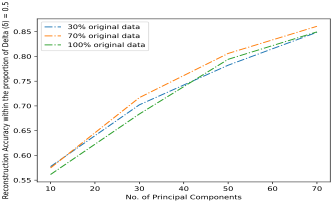


50% original data

R.A. on the a9a data (CTGAN) II

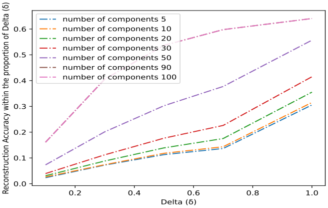


100% original data

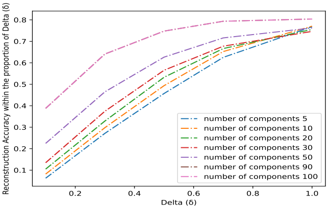


R.A. vs no. of principal components with δ= 0.5

R.A. on the mushroom data(CTGAN) I

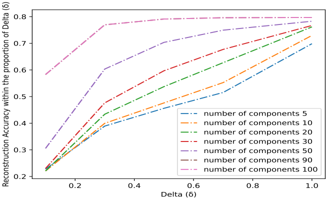


10% original data

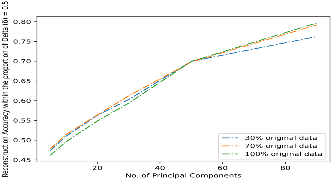


50% original data

R.A. on the mushroom data(CTGAN) II



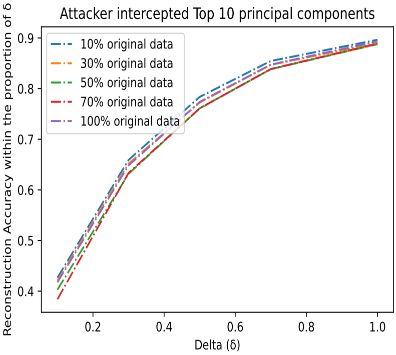
100% original data



R.A. vs no. of principal components with δ= 0.5

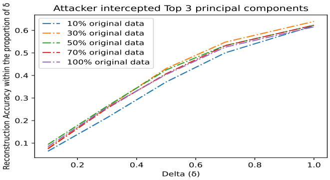
No Protection mechanism I

- When no protection mechanism is used, we show that the R.A. increases in comparison with the case when DPPCA is used, and when the principal components are computed on the synthetic dataset.

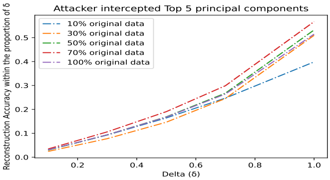


a9a data

No Protection mechanism II



Heart-scale data

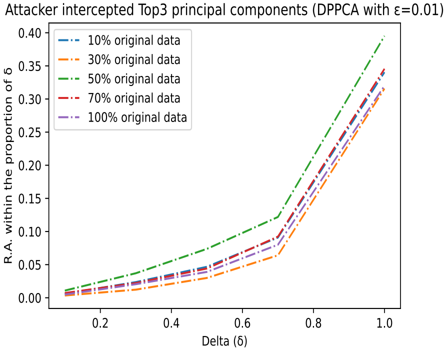


Mushrooms data

R.A. without protection mechanism prior to the computation of principal components

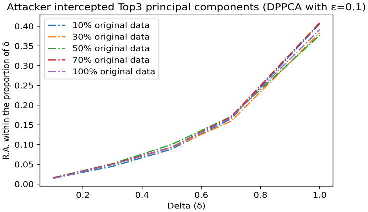
Results with DPPCA on the heart-scale data I

- Lesser the value of ϵ (higher privacy), the shallower the graph for reconstruction accuracy (less reconstruction).

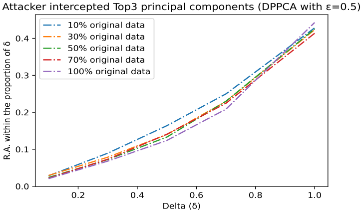


$\epsilon = 0.01$ for DPPCA

Results with DPPCA on the heart-scale data II

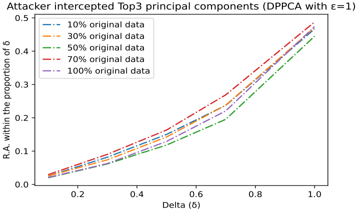


$\epsilon = 0.1$ for DPPCA

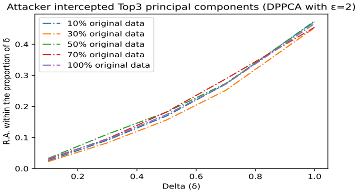


$\epsilon = 0.5$ for DPPCA

Results with DPPCA on the heart-scale data III

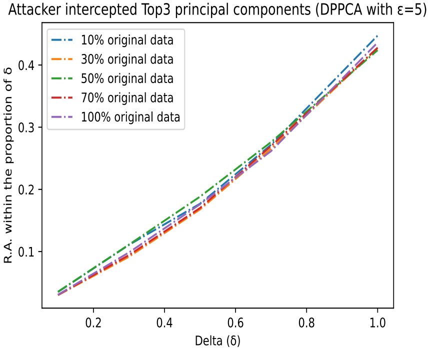


$\epsilon = 1$ for DPPCA



$\epsilon = 2$ for DPPCA

Results with DPPCA on the heart-scale data IV



$\epsilon = 5$ for DPPCA

Summary

- We demonstrated a data reconstruction attack theory against Principal Component Analysis.
- We compared two defense strategies, including DPPCA, and synthetic data against the proposed attack.

Thank You Very Much

References I

[1] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.

[2] Zari, O., Parra-Arnau, J., Ünsal, A., Strufe, T., & Önen, M. (2022, September). Membership inference attack against principal component analysis. In Privacy in Statistical Databases: International Conference, PSD 2022, Paris, France, September 21–23, 2022, Proceedings (pp. 269-282). Cham: Springer International Publishing.

[3] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR), 54(11s), 1-37.

[4] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems, 32.

[5] Imtiaz, H., & Sarwate, A. D. (2016, March). Symmetric matrix perturbation for differentially-private principal component analysis. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2339-2343). IEEE.

References II

[6] Bentley, J. W., Gibney, D., Hoppenworth, G., & Jha, S. K. (2020). Quantifying membership inference vulnerability via generalization gap and other model metrics. arXiv preprint arXiv:2009.05669.

[7] Farokhi, F., & Kaafar, M. A. (2020). Modelling and quantifying membership information leakage in machine learning. arXiv preprint arXiv:2001.10648.

[8] Jha, S. K., Jha, S., Ewetz, R., Raj, S., Velasquez, A., Pullum, L. L., & Swami, A. (2020). An Extension of Fano's Inequality for Characterizing Model Susceptibility to Membership Inference Attacks. arXiv preprint arXiv:2009.08097.

[9] Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018, July). Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st computer security foundations symposium (CSF) (pp. 268-282). IEEE.