When Does Randomization Fail to Protect Privacy?

Wenliang (Kevin) Du

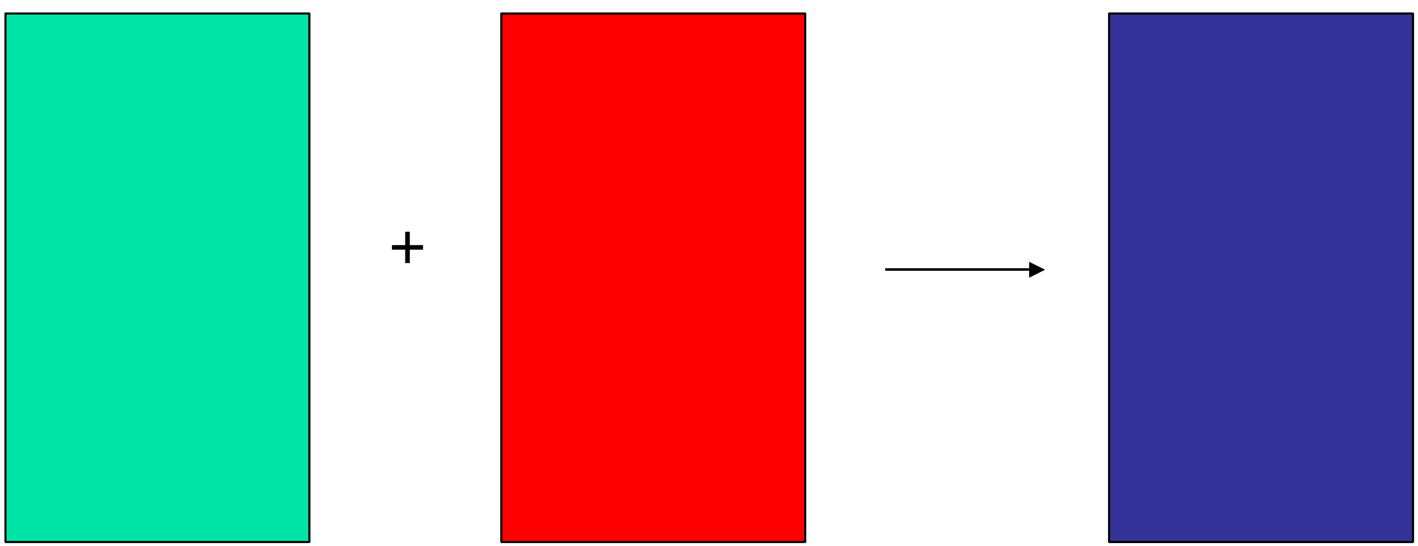
Department of EECS, Syracuse University

Random Perturbation

Agrawal and Srikant's SIGMOD paper.

Y = X + R

Original Data X   Random Noise R   Disguised Data Y



Random Perturbation

- Most of the security analysis methods based on randomization treat each attribute separately.
- Is that enough?
  - Does the relationship among data affect privacy?

As we all know …

We can't perturb the same number for several times.

If we do that, we can estimate the original data:

- Let t be the original data,
- Disguised data: $t + R_1, t + R_2, \ldots, t + R_m$
- Let $Z = [(t+R_1) + \ldots + (t+R_m)] / m$
- Mean: $E(Z) = t$
- Variance: $Var(Z) = Var(R) / m$

This looks familiar …

This is the data set (x, x, x, x, x, x, x, x)

Random Perturbation:

$(x+r_1, x+r_2, \ldots\ldots, x+r_m)$

We know this is NOT safe.

Observation: the data set is highly correlated.

Let's Generalize!

Data set: $(x_1, x_2, x_3, \ldots\ldots, x_m)$

If the correlation among data attributes are high, can we use that to improve our estimation (from the disguised data)?

Introduction

- A heuristic approach toward privacy analysis
- Principal Component Analysis (PCA)
- PCA-based data reconstruction
- Experiment results
- Conclusion and future work

Privacy Quantification: A Heuristic Approach

Our goal:

to find a best-effort algorithm that reconstructs the original data, based on the available information.

Definition

$$PM_F = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} L(D_{i,j}^{*}, D_{i,j})$$

How to use the correlation?

High Correlation   Data Redundancy

Data Redundancy   Compression

Our goal: Lossy compression:

We do want to lose information, but

We don’t want to lose too much data,

We do want to lose the added noise.

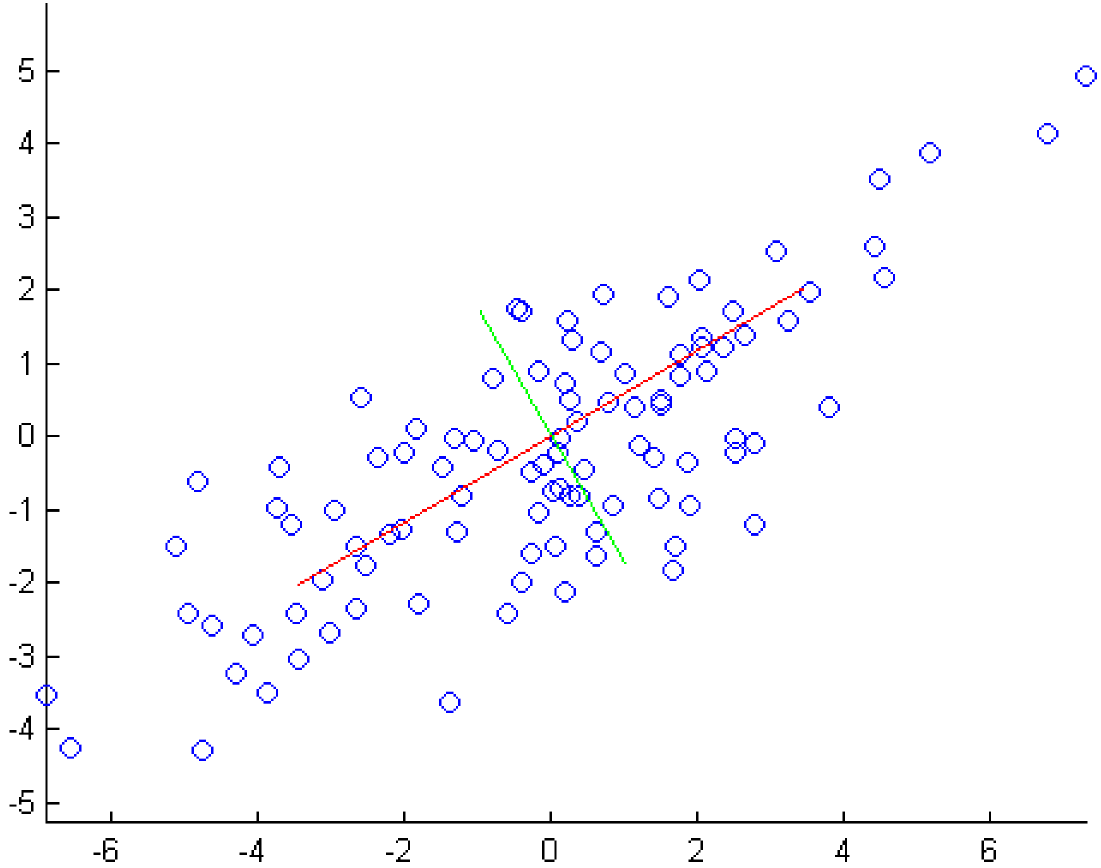
PCA Introduction

The main use of PCA: reduce the dimensionality while retaining as much information as possible.

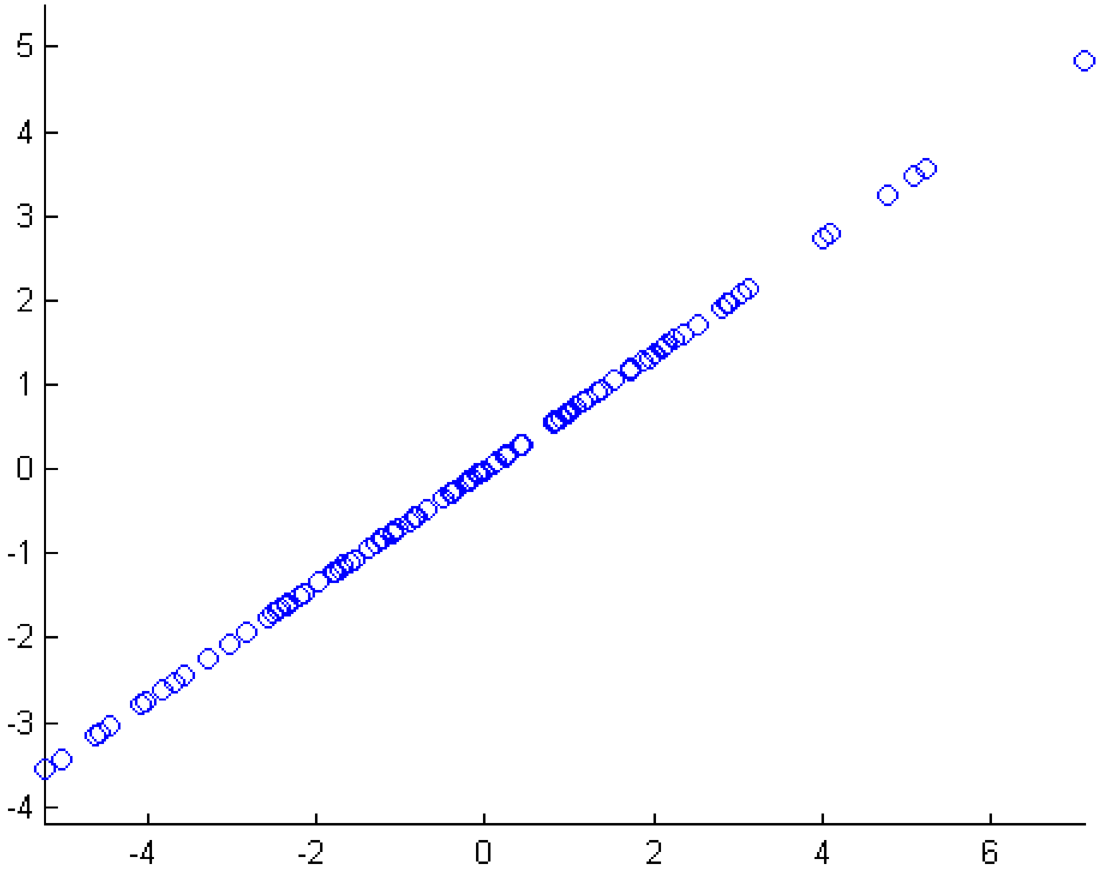
1st PC: containing the greatest amount of variation.

2nd PC: containing the next largest amount of variation.

Original Data



After Dimension Reduction



For the Original Data

They are correlated.

If we remove 50% of the dimensions, the actual information loss might be less than 10%.

For the Random Noises

They are not correlated.

Their variance is evenly distributed to any direction.

If we remove 50% of the dimensions, the actual noise loss should be 50%.

Data Reconstruction

Applying PCA

- Find Principle Components: $C = Q\,\Lambda Q^T$
- Set $\tilde{Q}$ to be the first p columns of Q.
- Reconstruct the data:

$$\tilde{X} = Y\tilde{Q}\tilde{Q}^T$$

$$= (X + R)\tilde{Q}\tilde{Q}^T = X\tilde{Q}\tilde{Q}^T + R\tilde{Q}\tilde{Q}^T$$

Random Noise R

How does RQQ^T affect accuracy?

Theorem:

$$Var(RQQ^T) = Var(R)\frac{p}{m},$$

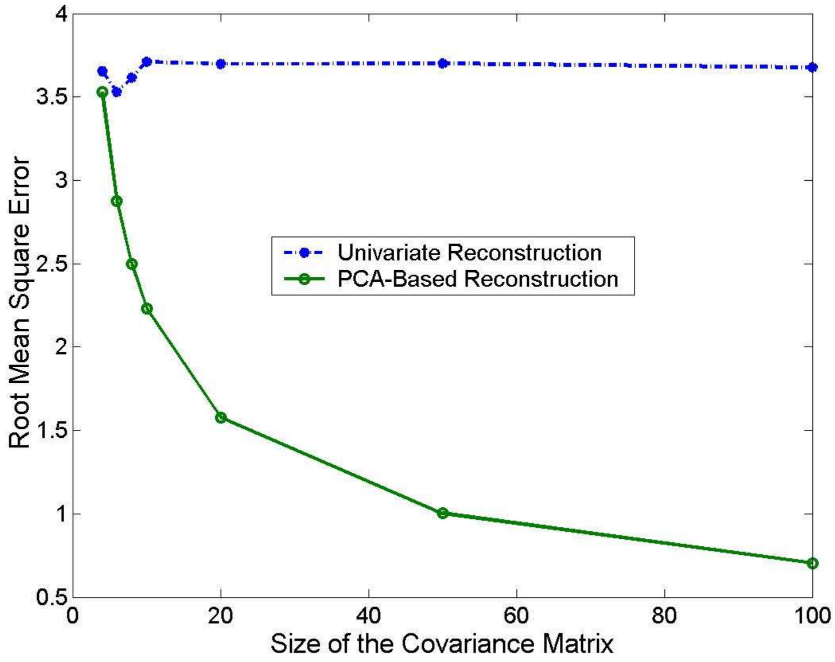
How to Conduct PCA on Disguised Data?

Estimating Covariance Matrix

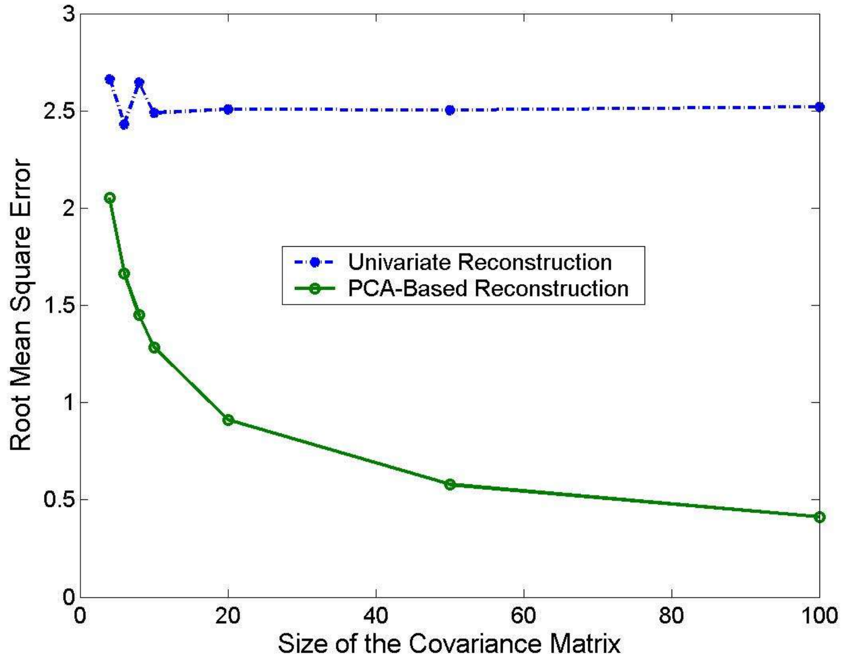
$$Cov(Y_i, Y_j) = Cov(X_i + R_i, X_j + R_j)$$

$$= \begin{cases} Cov(X_i, X_j) + \sigma^2, & \text{for } i = j \\ Cov(X_i, X_j), & \text{for } i \neq j \end{cases}$$

Experiment 1: Increasing the Number of Attributes

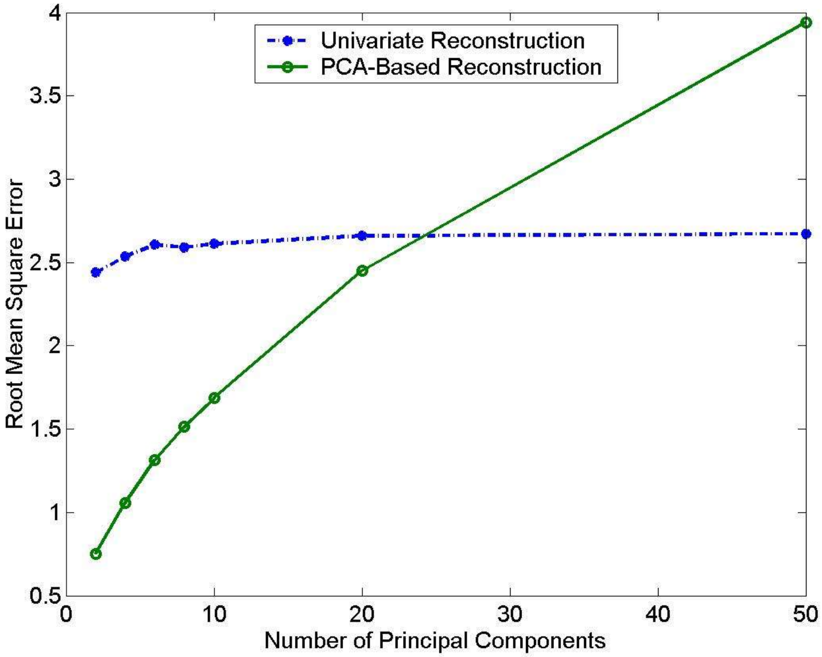


Normal Distribution

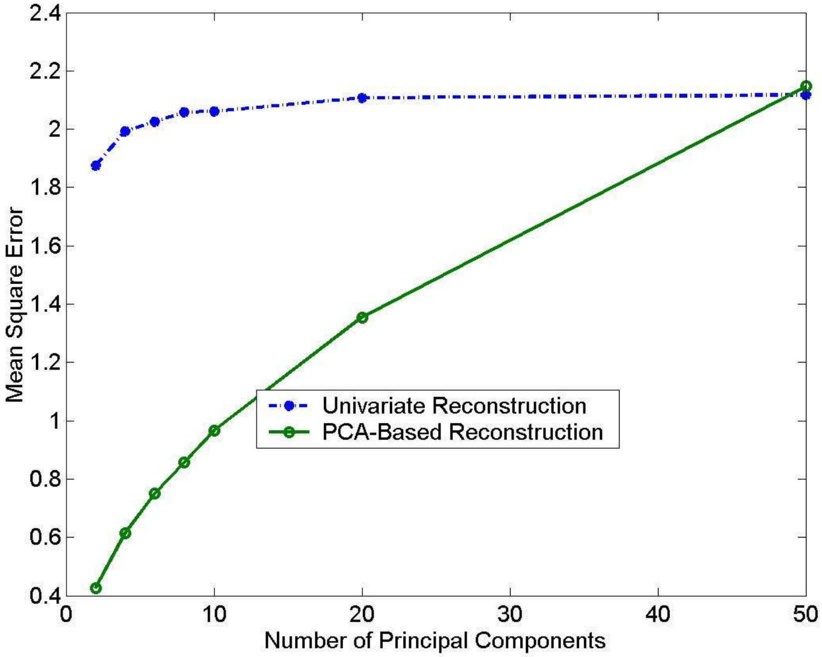


Uniform Distribution

Experiment 2: Increasing the number of Principal Components

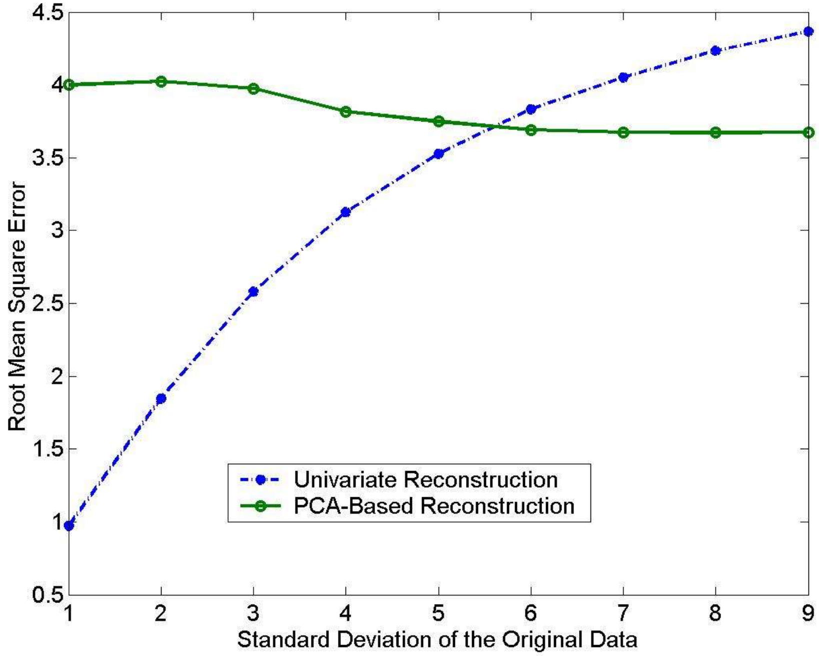


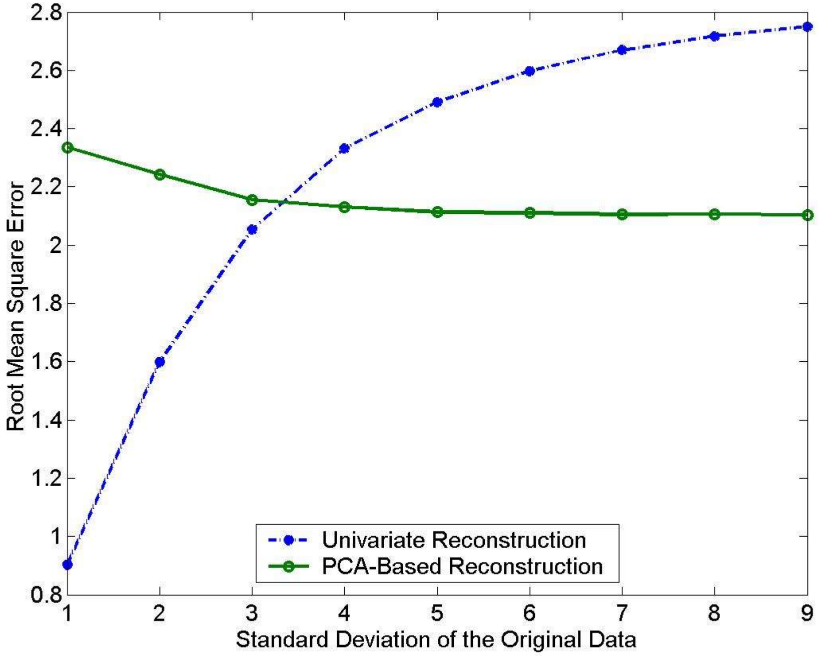
Normal Distribution



Uniform Distribution

Experiment 3: Increasing Standard Deviation of Noises





Normal Distribution

Uniform Distribution

Conclusions

- Privacy analysis based on individual attributes is not sufficient. Correlation can disclose information.
- PCA can filter out some randomness from a highly correlated data set.
- When does randomization fail:
  - Answer: when the data correlation is high.
  - Can it be cured?

Future Work

- How to improve the randomization to reduce the information disclosure?
  - Making random noises correlated?
- How to combine the PCA with the univariate data reconstruction?