Generating Sufficiency-based Non-Synthetic Perturbed Data

Rathindra Sarathy
Oklahoma State University

Krish Muralidhar
University of Kentucky

This paper was recently published in the Transactions on Data Privacy, Vol. 1, No. 1.

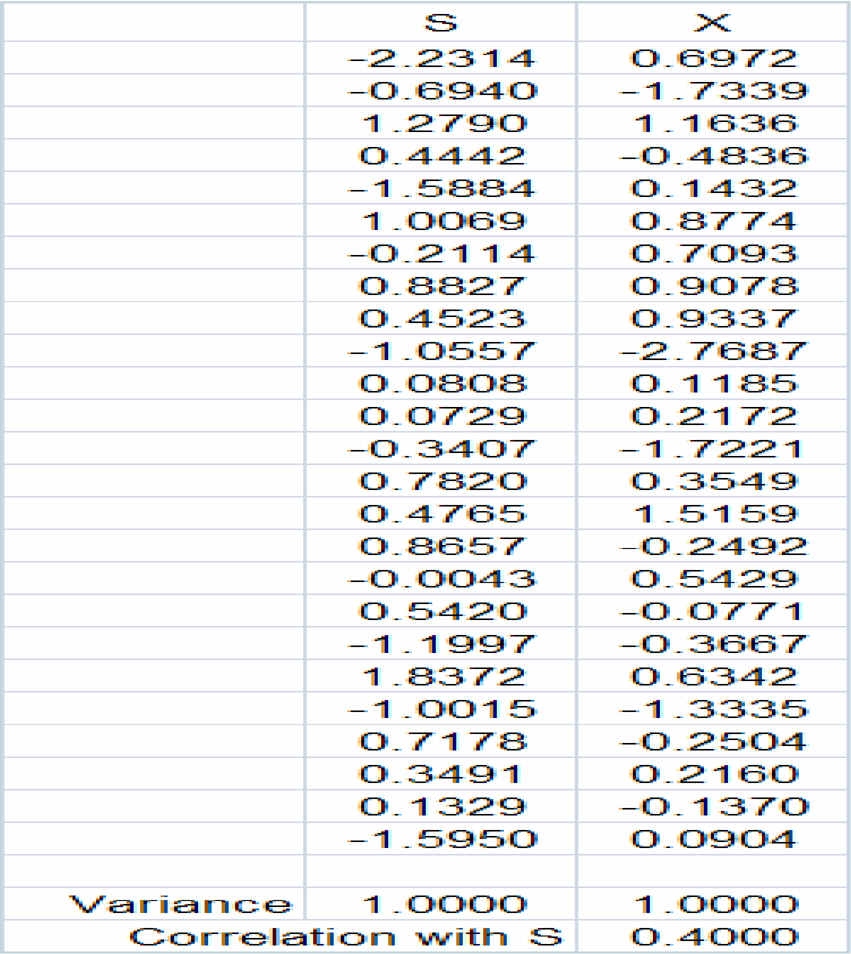
Additive Models for Data Perturbation

- Simple additive noise
- Kim's additive noise
- General Additive Data Perturbation (GADP)
- Information Preserving Statistical Obfuscation (IPSO)
- Sufficiency Based Additive Noise

Running Example

- We will use the following running example to illustrate the differences between the techniques.
- The example has 25 observations and two variables (one non-confidential and one confidential). Both variables have zero mean and unit variance. The data has a bivariate normal distribution.

Example Data Set



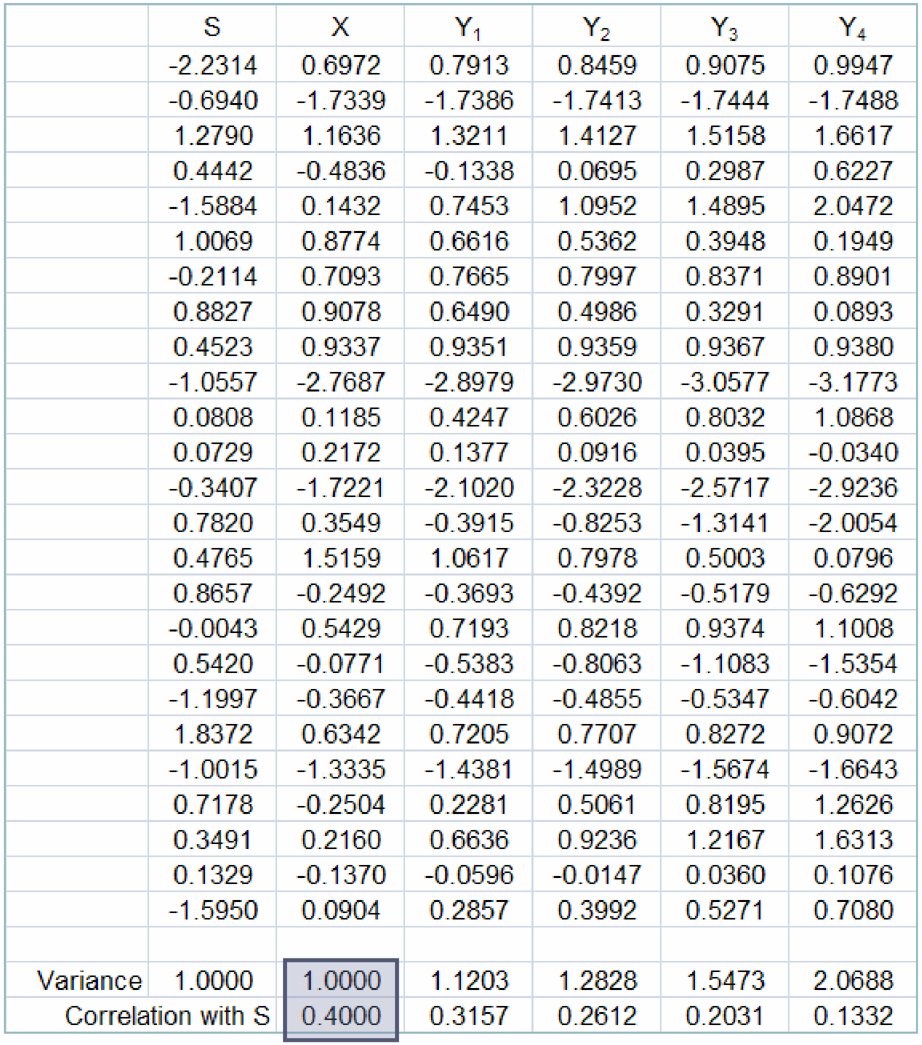
	S	X
	-2.2314	0.6972
	-0.6940	-1.7339
	1.2790	1.1636
	0.4442	-0.4836
	-1.5884	0.1432
	1.0069	0.8774
	-0.2114	0.7093
	0.8827	0.9078
	0.4523	0.9337
	-1.0557	-2.7687
	0.0808	0.1185
	0.0729	0.2172
	-0.3407	-1.7221
	0.7820	0.3549
	0.4765	1.5159
	0.8657	-0.2492
	-0.0043	0.5429
	0.5420	-0.0771
	-1.1997	-0.3667
	1.8372	0.6342
	-1.0015	-1.3335
	0.7178	-0.2504
	0.3491	0.2160
	0.1329	-0.1370
	-1.5950	0.0904
Variance	1.0000	1.0000
Correlation with S		0.4000

Simple Additive Noise

- Independent noise with mean 0 and a specified variance is added to the original confidential variable
  - $Y = X + e$
  - The variance of e dictates the level of perturbation
  - The procedure is then repeated for every confidential variable.
  - The variables are perturbed independent of one another

Additive Perturbation

(Variance of e = 0.10, 0.25, 0.50, 1.00)



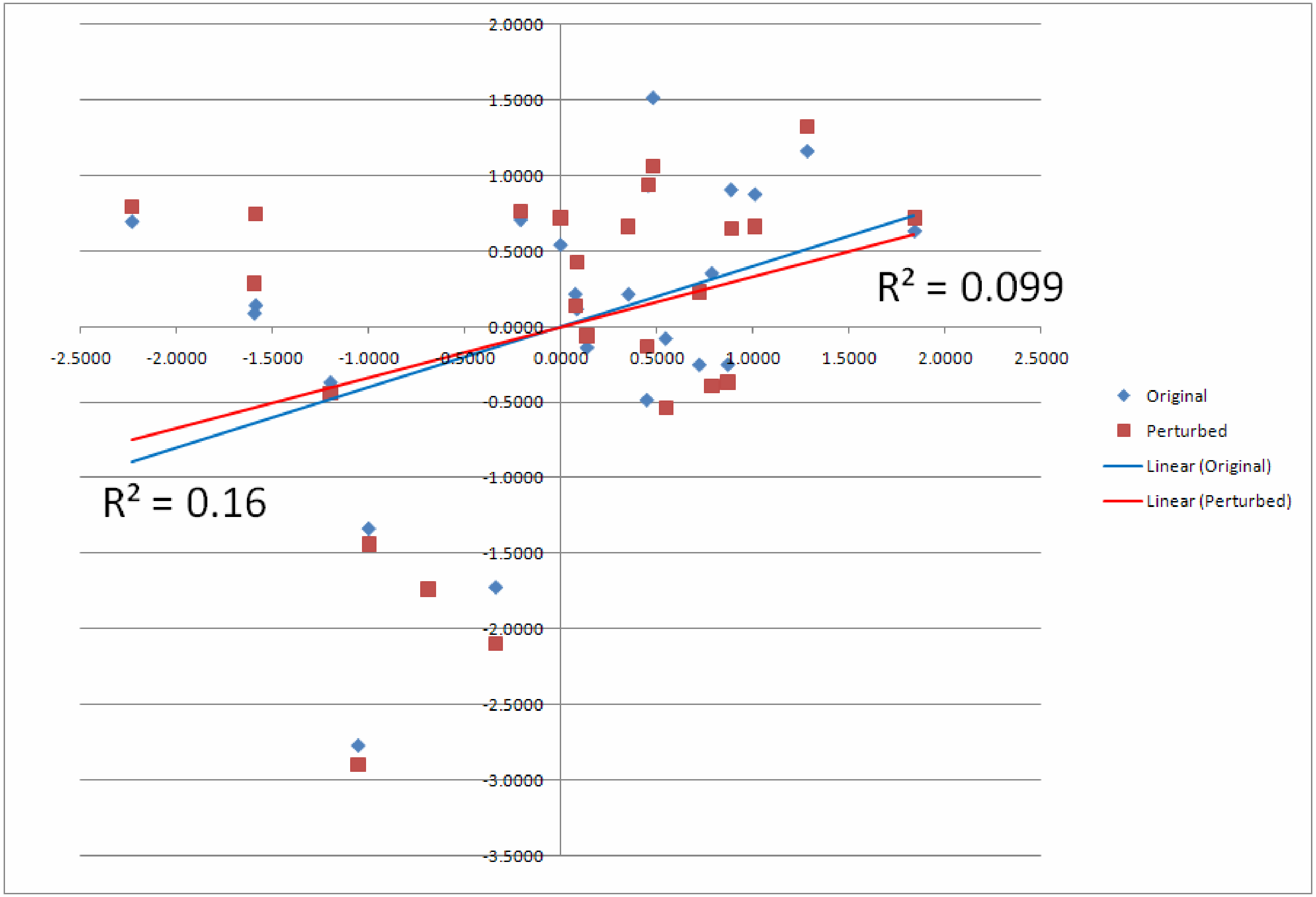
	S	X	Y_1	Y_2	Y_3	Y_4
	-2.2314	0.6972	0.7913	0.8459	0.9075	0.9947
	-0.6940	-1.7339	-1.7386	-1.7413	-1.7444	-1.7488
	1.2790	1.1636	1.3211	1.4127	1.5158	1.6617
	0.4442	-0.4836	-0.1338	0.0695	0.2987	0.6227
	-1.5884	0.1432	0.7453	1.0952	1.4895	2.0472
	1.0069	0.8774	0.6616	0.5362	0.3948	0.1949
	-0.2114	0.7093	0.7665	0.7997	0.8371	0.8901
	0.8827	0.9078	0.6490	0.4986	0.3291	0.0893
	0.4523	0.9337	0.9351	0.9359	0.9367	0.9380
	-1.0557	-2.7687	-2.8979	-2.9730	-3.0577	-3.1773
	0.0808	0.1185	0.4247	0.6026	0.8032	1.0868
	0.0729	0.2172	0.1377	0.0916	0.0395	-0.0340
	-0.3407	-1.7221	-2.1020	-2.3228	-2.5717	-2.9236
	0.7820	0.3549	-0.3915	-0.8253	-1.3141	-2.0054
	0.4765	1.5159	1.0617	0.7978	0.5003	0.0796
	0.8657	-0.2492	-0.3693	-0.4392	-0.5179	-0.6292
	-0.0043	0.5429	0.7193	0.8218	0.9374	1.1008
	0.5420	-0.0771	-0.5383	-0.8063	-1.1083	-1.5354
	-1.1997	-0.3667	-0.4418	-0.4855	-0.5347	-0.6042
	1.8372	0.6342	0.7205	0.7707	0.8272	0.9072
	-1.0015	-1.3335	-1.4381	-1.4989	-1.5674	-1.6643
	0.7178	-0.2504	0.2281	0.5061	0.8195	1.2626
	0.3491	0.2160	0.6636	0.9236	1.2167	1.6313
	0.1329	-0.1370	-0.0596	-0.0147	0.0360	0.1076
	-1.5950	0.0904	0.2857	0.3992	0.5271	0.7080
Variance	1.0000	1.0000	1.1203	1.2828	1.5473	2.0688
Correlation with S		0.4000	0.3157	0.2612	0.2031	0.1332

Result

- The result of the application of noise addition is obvious
  - As the noise variance increases, the variance of the perturbed variable increases and the correlation between S and Y is attenuated
  - The results are asymptotic. The variance of the perturbed data will approach (Var of X + Var of e) as the size of the data set increases. For small data sets such as this one, we will see small differences between the expected variance and the actual variance

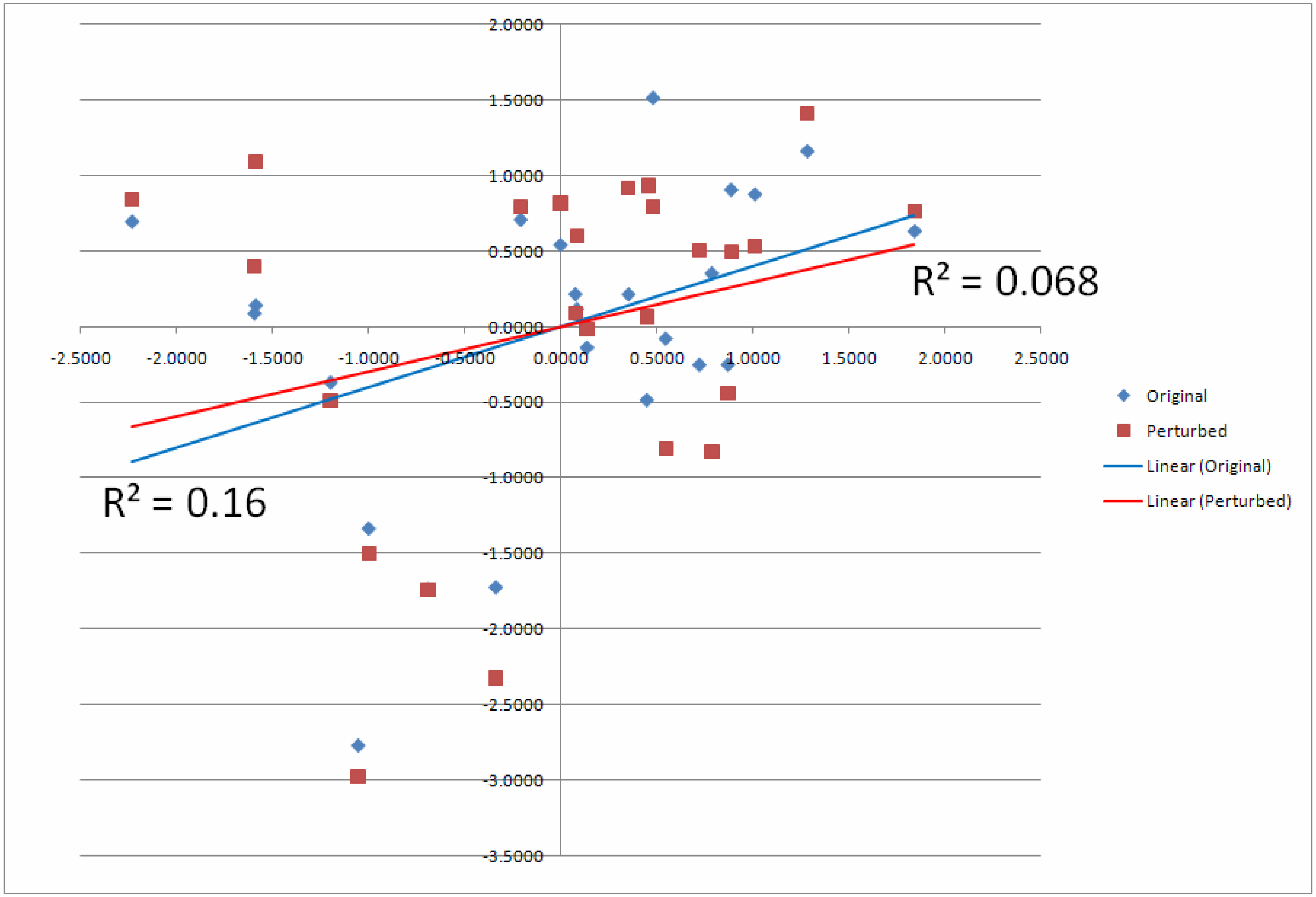
Impact on Data Analysis

(Var(e) = 0.10) (S on x-axis (X or Y) on y-axis)



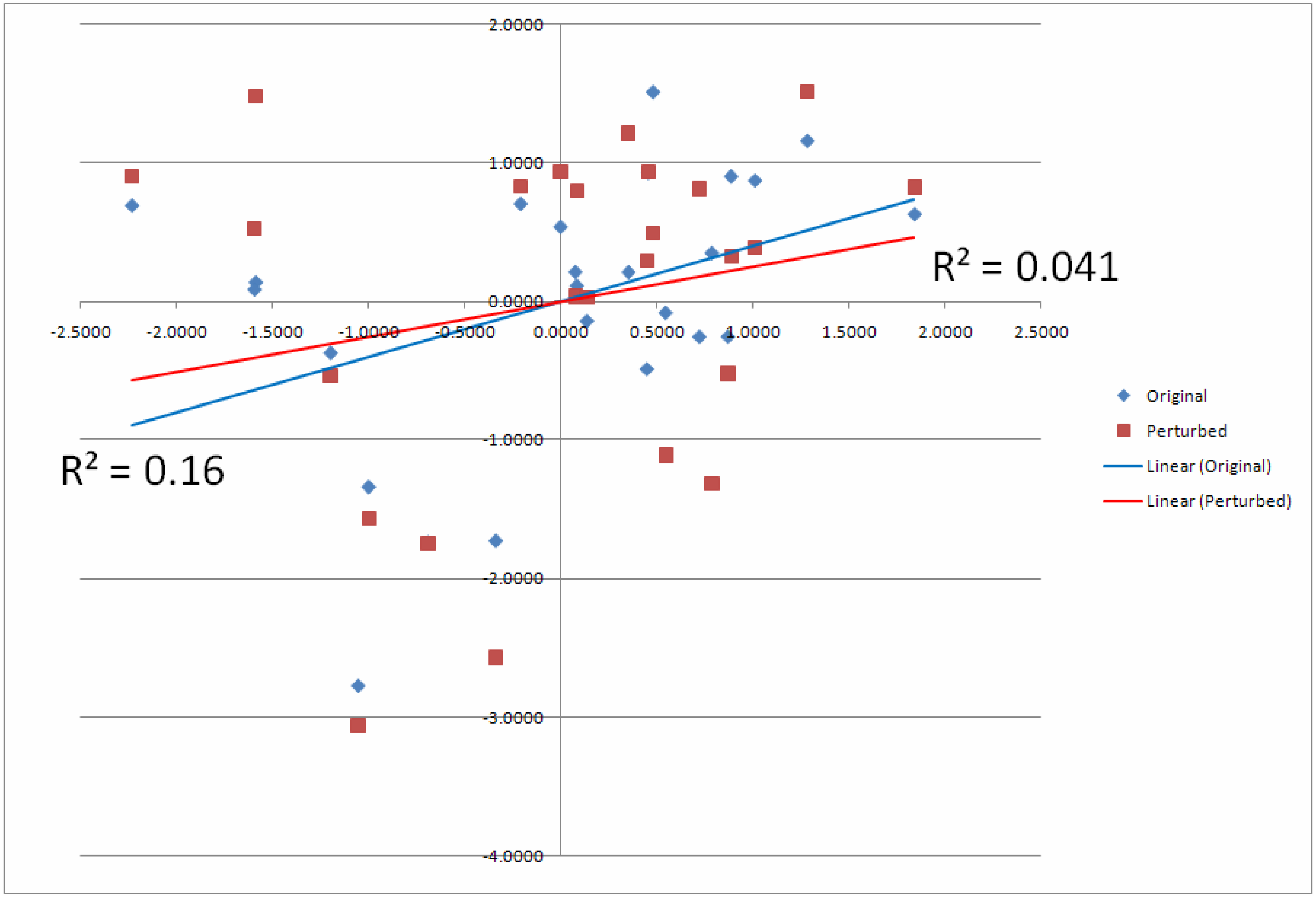
Impact on Data Analysis

(Var(e) = 0.25) (S on x-axis (X or Y) on y-axis)



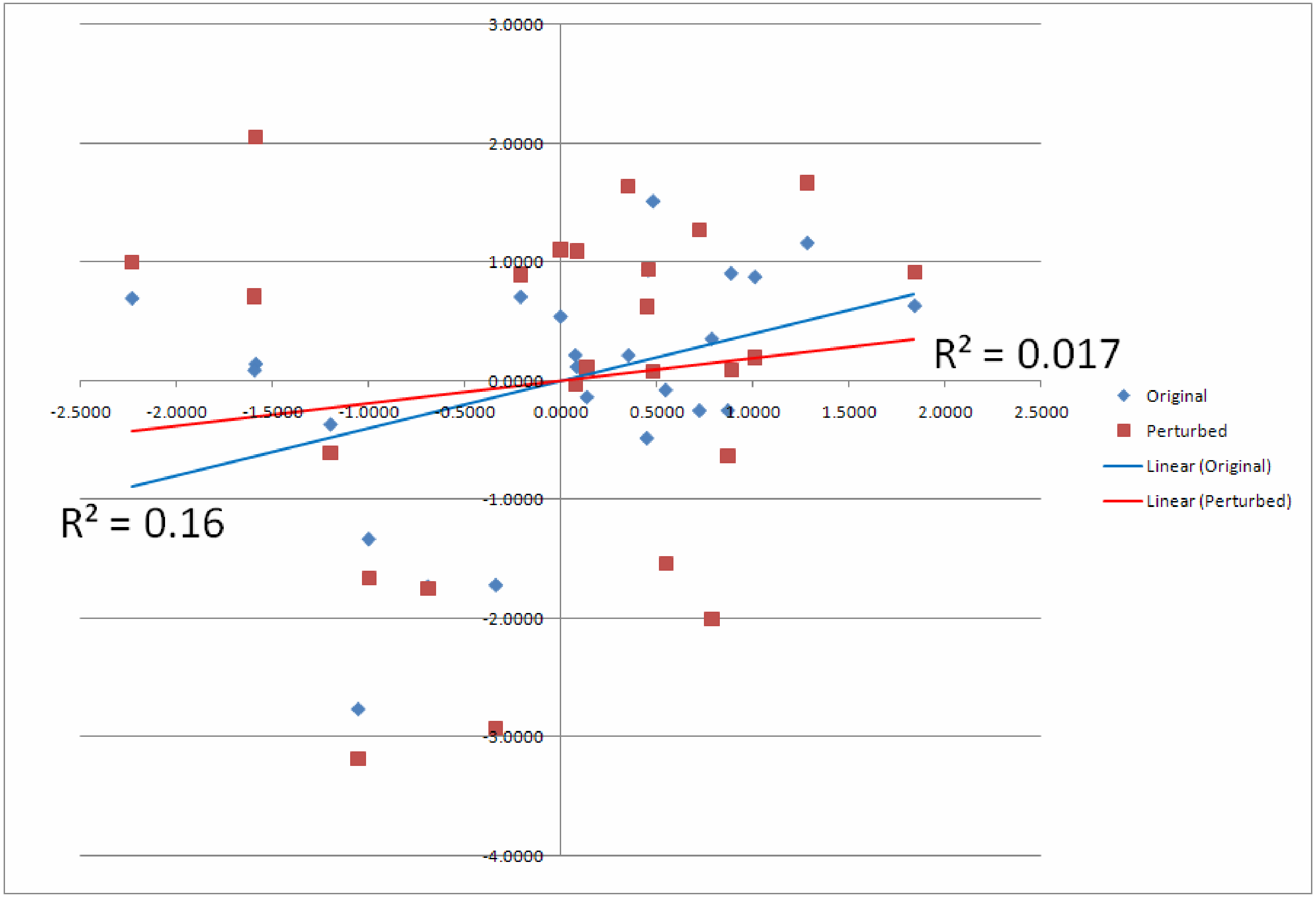
Impact on Data Analysis

(Var(e) = 0.50) (S on x-axis (X or Y) on y-axis)



Impact on Data Analysis

(Var(e) = 1.00) (S on x-axis (X or Y) on y-axis)



Summary

- Not surprisingly, when the level of noise is small, the results using the perturbed data are very similar to that using the original data.
- As the level of noise increases, the results using the perturbed data diverge from the results using the original data.

Kim's Method

- For a single confidential variable, the application of Kim's method is the same as simple additive noise.
  - When there are multiple confidential variables, Kim's method uses a noise term with covariance matrix of the form $d\Sigma_{XX}$ where Σ_{XX} is the covariance matrix of the confidential variables
  - By contrast, simple additive noise terms would have a covariance matrix that is diagonal

General Additive Data Perturbation (GADP) ***

- In GADP, the perturbed values are generated as follows:
  - $y_i = \beta_0 + \beta_1 S + \varepsilon$
  - The values of β_0, β_1, and the characteristics of the noise ε are estimated from the original data
  - This is the equivalent of generating a set of values Y from a linear approximation to the conditional expectation of X|S. Asymptotically, the mean vector and covariance matrix of (S, Y) is the same as (S,X)
  - Asymptotically, this model maximizes security
    - $f(X|S,Y,T) = f(X|S,T)$ where T represents information available from the release of summary information on the first 2 moments regarding X

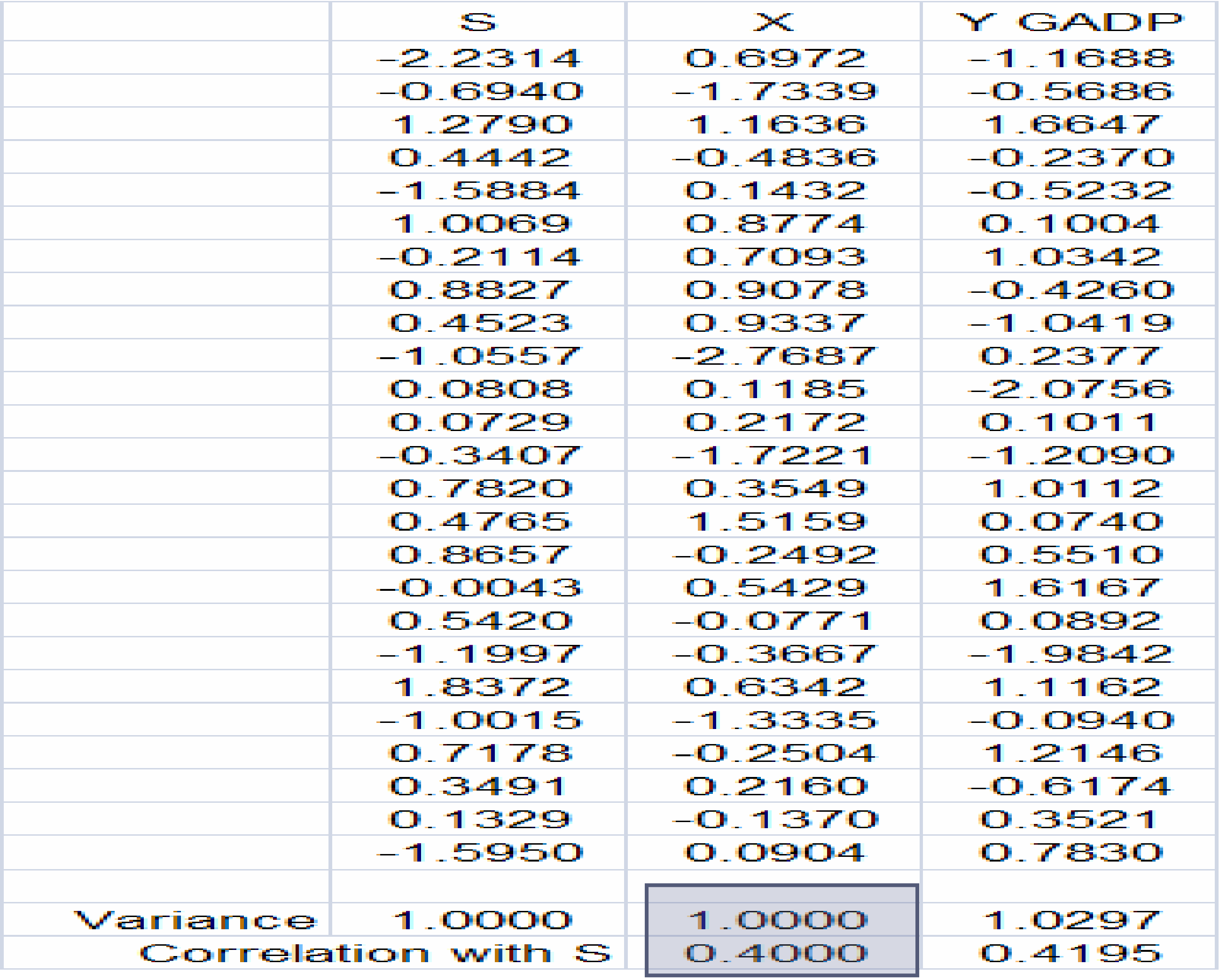
Perturbation Parameter

- In noise addition, the variance of the noise term represents the “perturbation parameter” since it dictates the extent of perturbation
- No such parameter is required in the GADP method. The entire perturbation is based on the relationships between the variables in the data set

Estimating the model parameters

- $\beta_0 = \mu_X - \Sigma_{XS}(\Sigma_{SS})^{-1}\mu_S$
- $\beta_1 = \Sigma_{XS}(\Sigma_{SS})^{-1}$
- $\varepsilon \sim MVN(0, \Sigma_{XX} - \Sigma_{XS}(\Sigma_{SS})^{-1}\Sigma_{SX})$
- For the current example
  - $\beta_0 = 0$
  - $\beta_1 = 0.4000$
  - $\varepsilon \sim \text{Normal}(0, 0.84)$

GADP Applied to Example



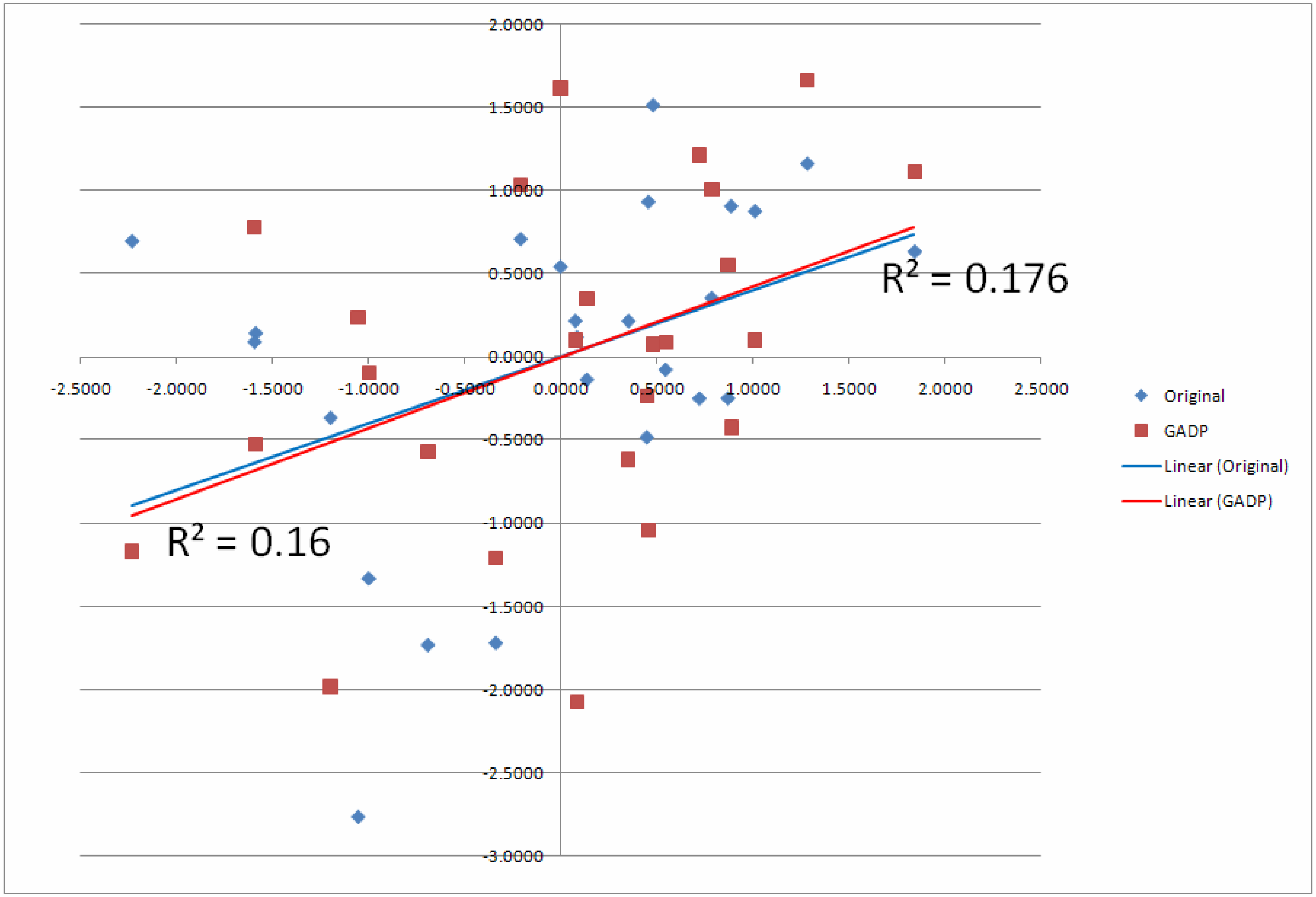
	S	X	Y GADP
	-2.2314	0.6972	-1.1688
	-0.6940	-1.7339	-0.5686
	1.2790	1.1636	1.6647
	0.4442	-0.4836	-0.2370
	-1.5884	0.1432	-0.5232
	1.0069	0.8774	0.1004
	-0.2114	0.7093	1.0342
	0.8827	0.9078	-0.4260
	0.4523	0.9337	-1.0419
	-1.0557	-2.7687	0.2377
	0.0808	0.1185	-2.0756
	0.0729	0.2172	0.1011
	-0.3407	-1.7221	-1.2090
	0.7820	0.3549	1.0112
	0.4765	1.5159	0.0740
	0.8657	-0.2492	0.5510
	-0.0043	0.5429	1.6167
	0.5420	-0.0771	0.0892
	-1.1997	-0.3667	-1.9842
	1.8372	0.6342	1.1162
	-1.0015	-1.3335	-0.0940
	0.7178	-0.2504	1.2146
	0.3491	0.2160	-0.6174
	0.1329	-0.1370	0.3521
	-1.5950	0.0904	0.7830
Variance	1.0000	1.0000	1.0297
Correlation with S		0.4000	0.4195

Results

- For small data sets, the results from the application of the GADP procedure would be slightly different from the original data. As the size of the data set increases, the values of the perturbed data approach the values of the original data
  - We see a small difference in the mean, variance, and correlation when GADP is applied to this relatively small data set

Impact on Data Analysis

(S on x-axis (X or Y) on y-axis)



Summary

- Compared to the noise added methods, the results from GADP more closely resemble the original data
  - As sample size increases, the estimates from the GADP data will approach the original data

Information Preserving Statistical Obfuscation (IPSO) (Burridge 2003) ***

- IPSO can be viewed as a GADP model with the additional requirement that the mean vector and covariance matrix of the perturbed data are *exactly* the same as the original data even for small datasets
  - Ensures that the perturbed data has the some "sufficient statistics" as the original data for a sample of any size
  - The sufficient statistics that are being maintained are the mean vector and covariance matrix

Sufficient Statistics

- A sufficient statistic is a particular kind of statistic that contains all the information about θ, the population parameter for which it is the estimator, that is contained in the sample.
  - It is important to note that it is assumed that the density of the distribution is known. The sufficient statistic cannot be used to check the validity of the assumption that the density is of a particular form
  - (Mood, Graybill, Boes, “Introduction to the Theory of Statistics,” McGraw Hill, New York, 2001)

General Linear Models

- Assume that we are attempting to estimate the parameters of a linear model.
- It is well known that the sufficient statistics for estimating the parameters of the General Linear Model with normality assumptions, are the mean vector and covariance matrix.
- These two statistics contain as much information regarding the parameters of the GLM as the entire sample data

Mean vector and Covariance Matrix as Sufficient Statistics

- The mean vector and covariance matrix serve as the sufficient statistics for many parametric statistical analysis including but not limited to
  - Simple hypothesis testing
  - Analysis of variance
  - Regression analysis
  - Multivariate analysis (MANOVA, principal components analysis, canonical correlations)

IPSO

- Burridge (2003) essentially argues that if we are able to generate a perturbed data set with exactly the same mean vector and covariance matrix as the original data set, for statistical analyses for which the mean vector and covariance matrix are sufficient statistics such as the GLM, the results of the analysis using the perturbed data will be **exactly the same** as that using the original data

But …

- It is important to remember that there is no guarantee regarding the results for other types of analyses (non-parametric, data mining, etc.)
- In addition, the results of tests of underlying assumption (such as normality, outliers, etc.) will not be the same using the perturbed data as they are using the original data

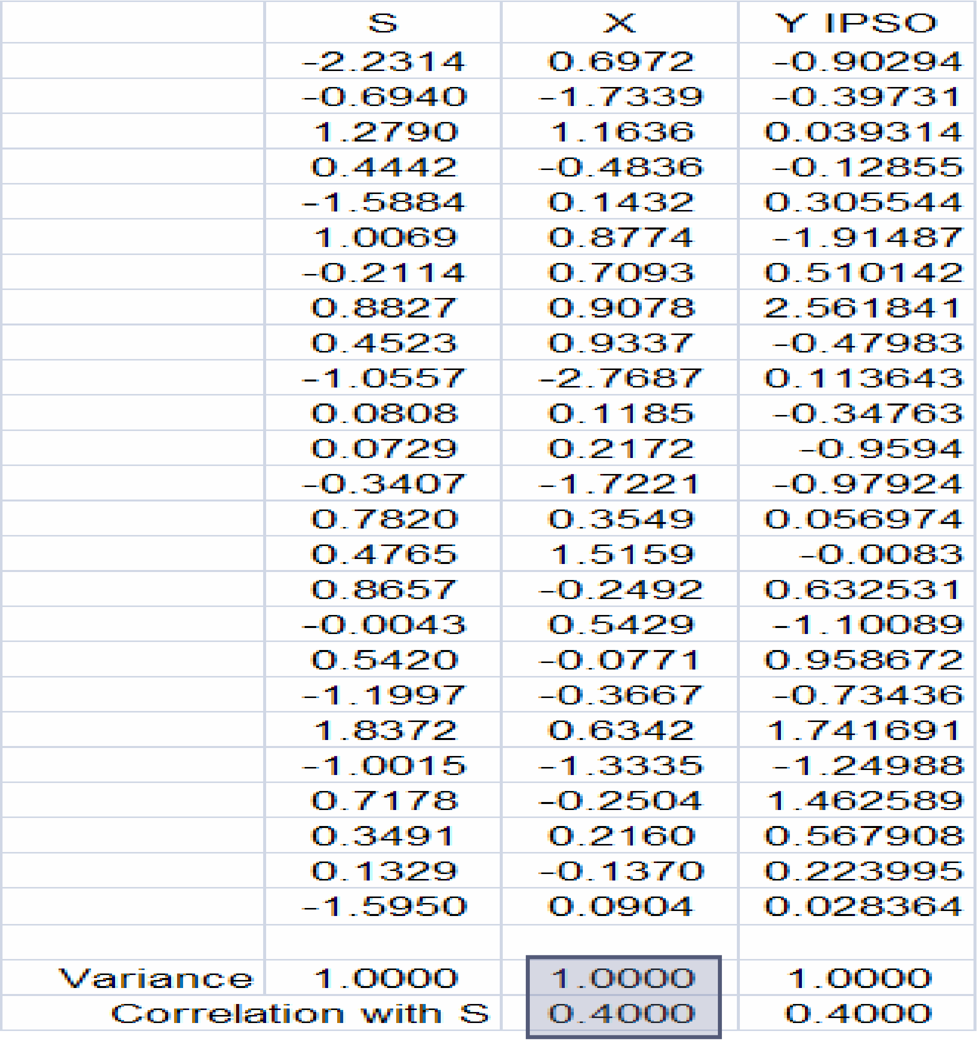
Implementing IPSO

- The IPSO model is the same as the GADP model
  - $y_i = \beta_0 + \beta_1 S + \varepsilon$
  - The values of β_0, β_1, and the characteristics of the noise ε are estimated from the original data

- The only difference between GADP and IPSO is in the generation of the noise terms
  - Generate the noise terms orthogonal to the original data
  - Standardize the noise term to have mean vector **exactly** 0 and covariance matrix **exactly** equal to $\Sigma_{XX} - \Sigma_{XS}(\Sigma_{SS})^{-1}\Sigma_{SX}$

(Note that this is a refinement of the IPSO presented by Burridge)

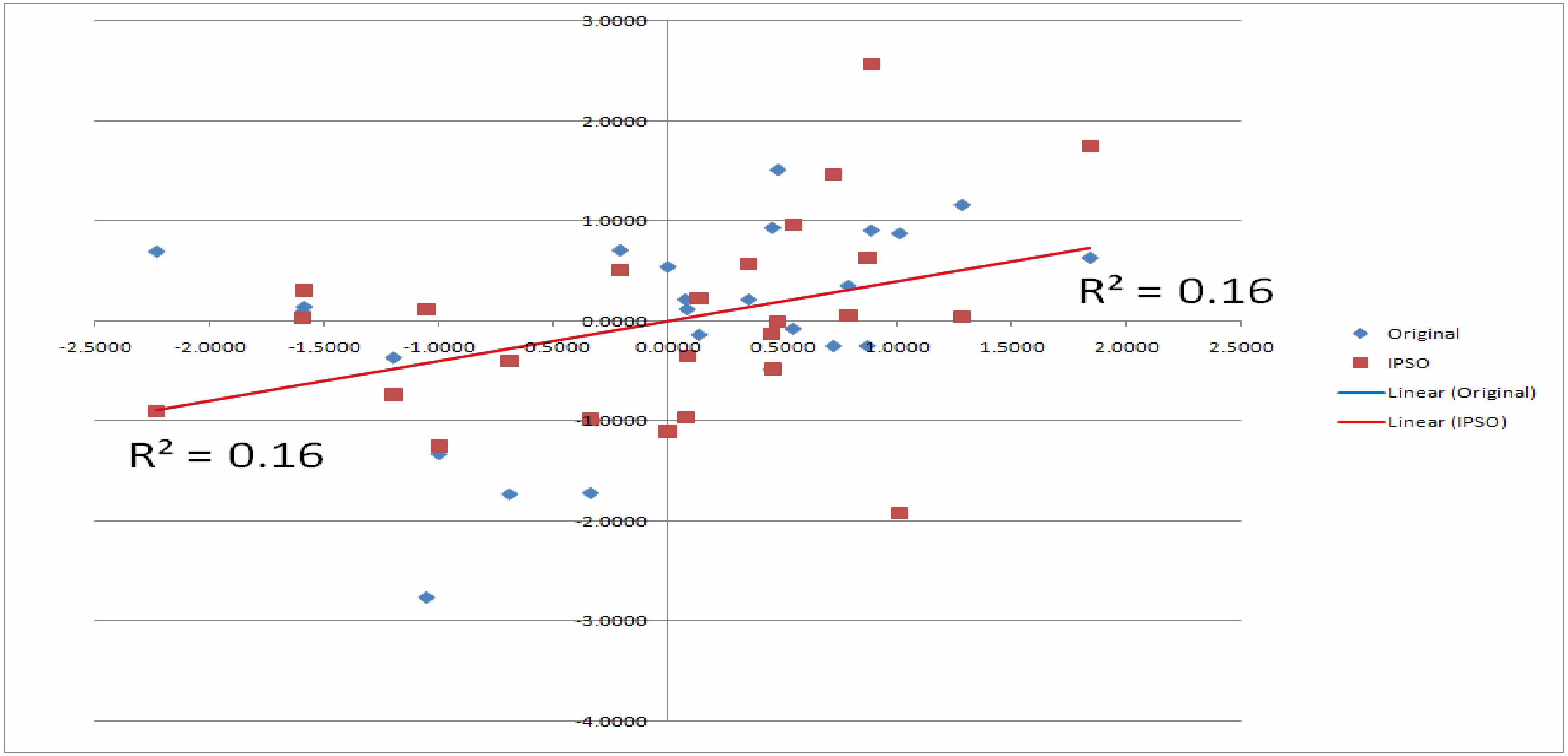
IPSO Applied to Example



	S	X	Y IPSO
	-2.2314	0.6972	-0.90294
	-0.6940	-1.7339	-0.39731
	1.2790	1.1636	0.039314
	0.4442	-0.4836	-0.12855
	-1.5884	0.1432	0.305544
	1.0069	0.8774	-1.91487
	-0.2114	0.7093	0.510142
	0.8827	0.9078	2.561841
	0.4523	0.9337	-0.47983
	-1.0557	-2.7687	0.113643
	0.0808	0.1185	-0.34763
	0.0729	0.2172	-0.9594
	-0.3407	-1.7221	-0.97924
	0.7820	0.3549	0.056974
	0.4765	1.5159	-0.0083
	0.8657	-0.2492	0.632531
	-0.0043	0.5429	-1.10089
	0.5420	-0.0771	0.958672
	-1.1997	-0.3667	-0.73436
	1.8372	0.6342	1.741691
	-1.0015	-1.3335	-1.24988
	0.7178	-0.2504	1.462589
	0.3491	0.2160	0.567908
	0.1329	-0.1370	0.223995
	-1.5950	0.0904	0.028364
Variance	1.0000	1.0000	1.0000
Correlation with S		0.4000	0.4000

Impact on Data Analysis

(S on x-axis (X or Y) on y-axis)



- Only one line is visible in the above chart because the regression lines using the original and perturbed data are **exactly** the same as are all other estimates from this regression analysis

Disclosure Risk

- We will provide an extensive discussion of disclosure risk later in the presentation.
- At this time, suffice it to say that IPSO produces microdata with the lowest level of disclosure risk.

IPSO versus Noise Addition

- Data utility
  - For many traditional statistical analyses, IPSO provides exactly the same estimation results as the original data. Noise addition methods do not.
- Disclosure risk
  - Releasing IPSO perturbed microdata results in the lowest level of disclosure risk. Noise addition methods do not.
- IPSO provides better utility and lower disclosure risk than noise addition methods

IPSO – A Complete Solution?

- IPSO is a complete solution to the perturbation problem if the joint distribution of the original data set is multivariate normal
  - It is an almost complete solution even when the non-confidential variables are categorical and the joint distribution of the confidential variables is multivariate normal

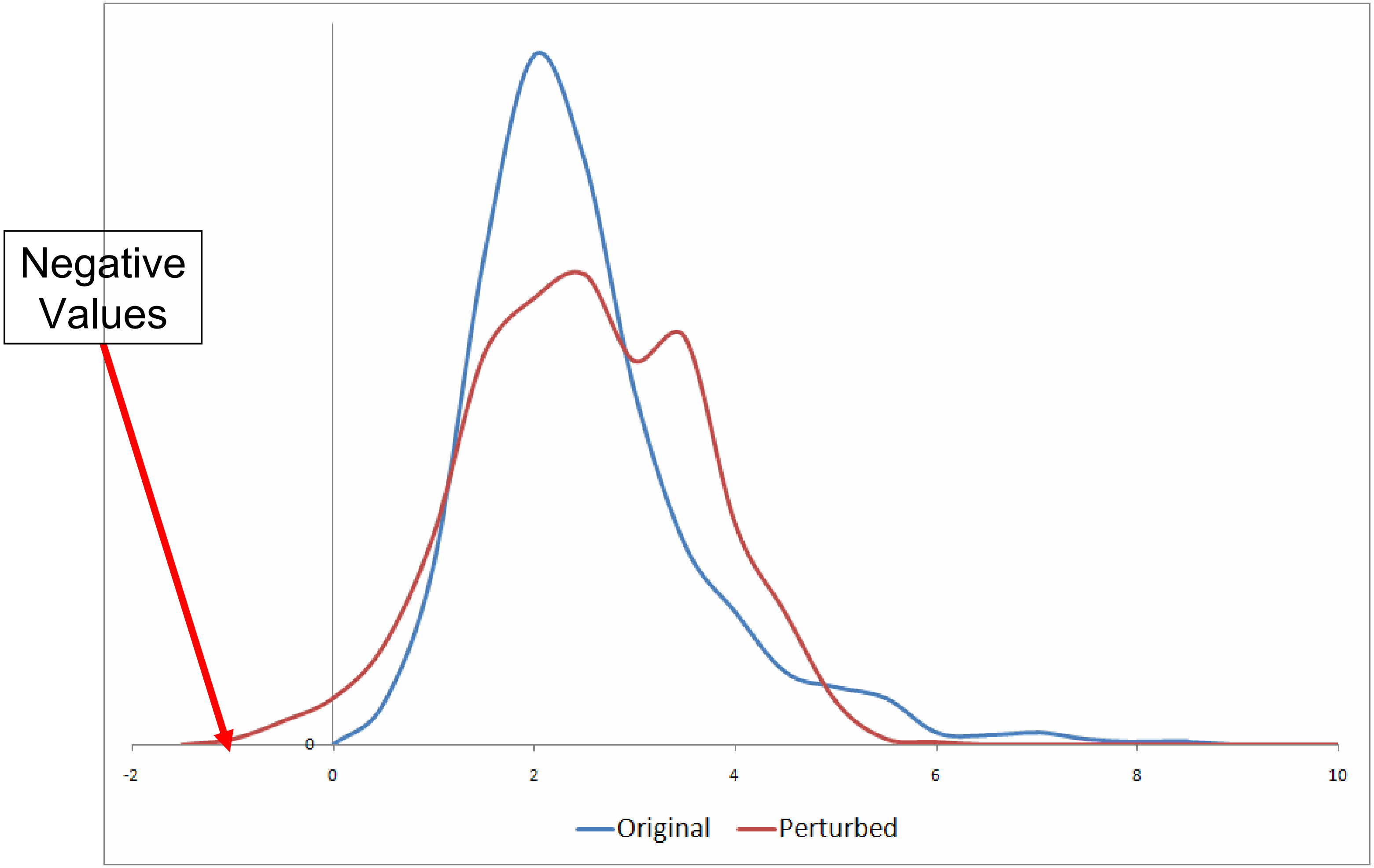
Issues with IPSO

- The IPSO approach uses a linear model to generate perturbed data. Consequently, when the confidential variables do not have a joint multivariate normal distribution
  - Marginal distribution is significantly altered
    - May see negative values in the perturbed data when the original data is all positive
  - Non-linear relationships are not maintained
  - The perturbed data may be considered "synthetic"

A Different Example

- In this example, we consider a situation where the marginal distribution of X is skewed. In addition, the relationship between X and S is non-linear
- We have perturbed this data using the IPSO approach

Impact on Marginal Distribution of the Confidential Variable

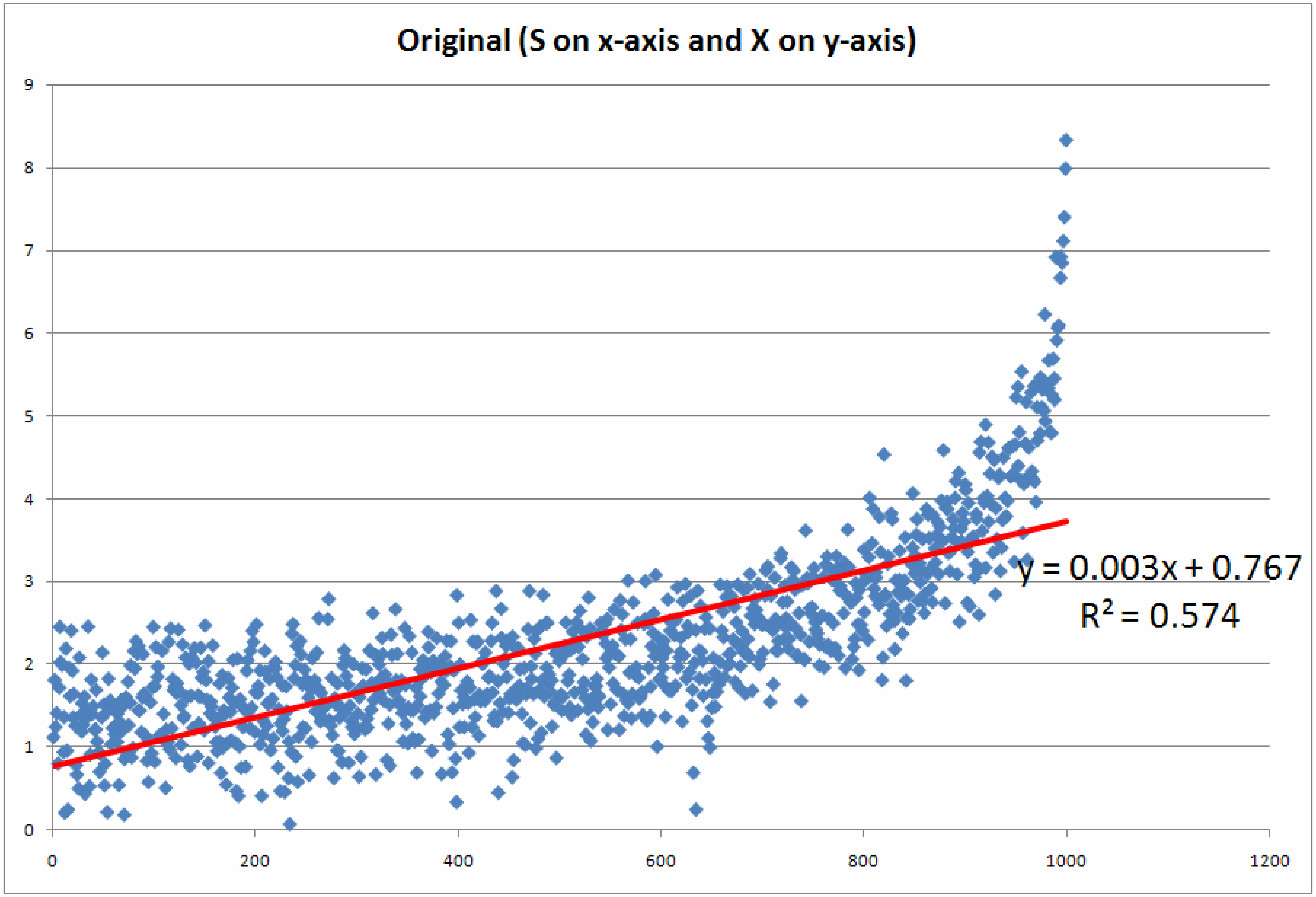


Negative Values

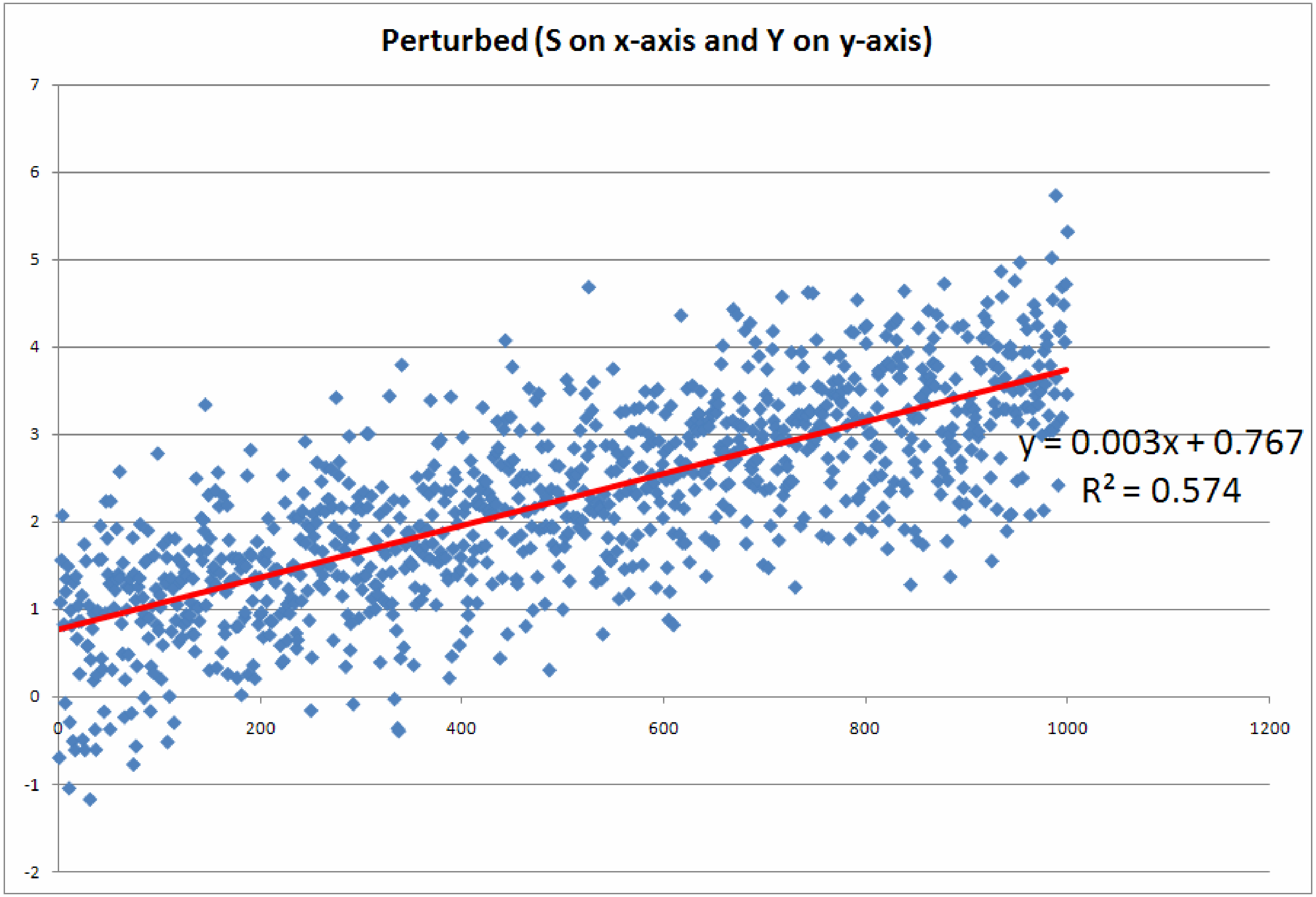
Impact on Marginal Distribution of X

- The marginal distribution of the perturbed variable is significantly different (less skewed) from the original variable
- The original variable did not have negative values while the perturbed variable has negative values

Original Relationship between S and X



Perturbed Relationship between S and Y



Relationships

- Linear regression results using the perturbed data are exactly the same as the original data
- But …
  - Perturbation results in completely changing the relationship between the variables. The original relationship is non-linear while after perturbation, the relationship is almost linear

“Synthetic” Data

- The IPSO perturbed values are generated only as a function of the non-confidential variables S and independent of the values of X … often referred to as “synthetic” data
- Some researchers and users are hesitant to use synthetic data because it is not “related” to the original data

Statistical versus Practical Perspective

- From a statistical perspective, we can argue that IPSO delivers exactly what it promises. We can also argue that most users will be analyzing the data only using traditional statistics and so IPSO offers a good solution
- From a practical perspective, the problem is that if providers and users are not willing to use the procedure, then its statistical effectiveness is moot

What is the alternative?

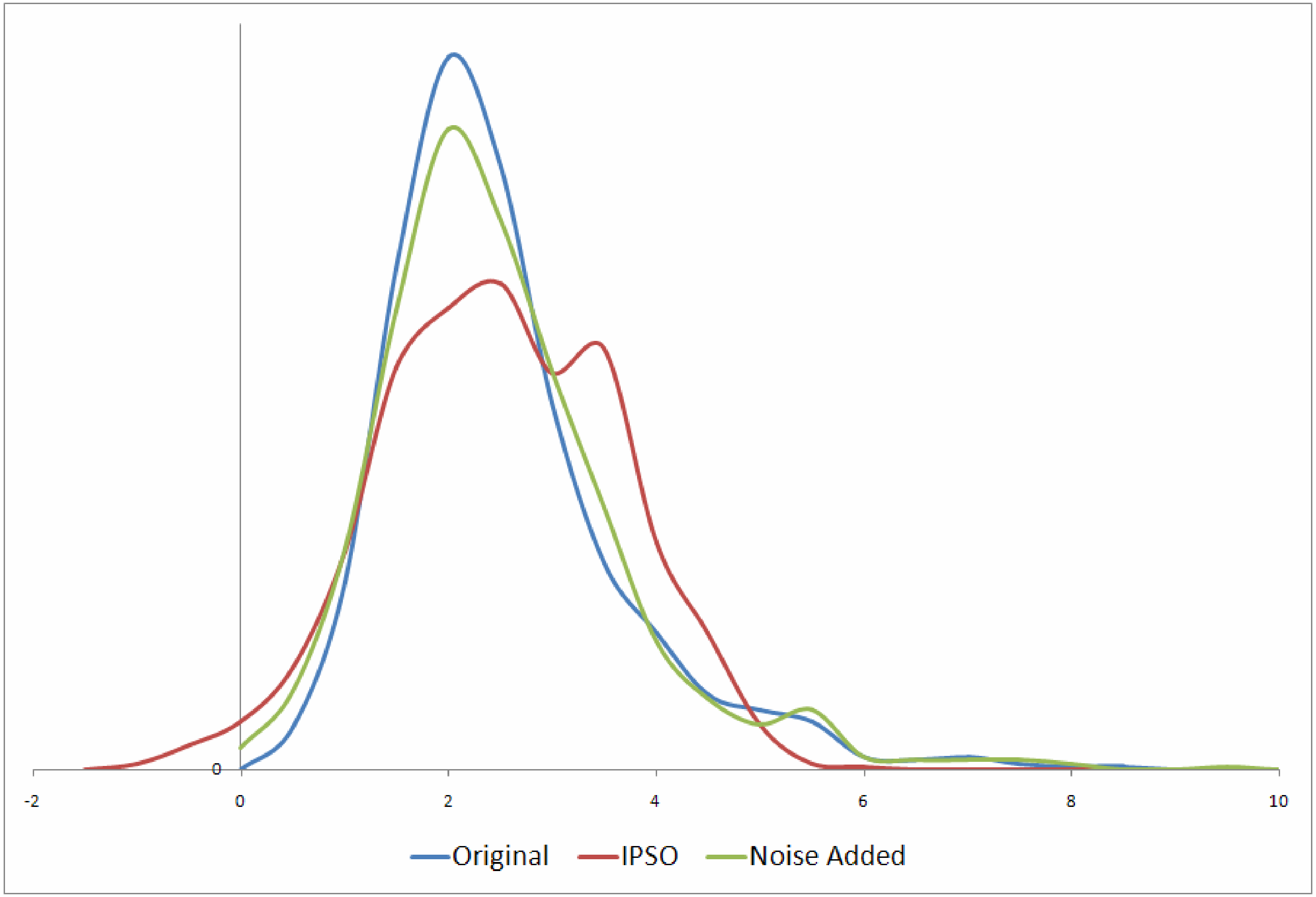
- The alternative that is often suggested is to use simple noise addition
  - Alternative approaches such as copula based perturbation and data shuffling can also be considered.

- With simple noise addition, when the perturbed values are in “proximity” to the original values we can
  - maintain the marginal distribution to be “close” to the original
  - maintain most relationships to be “close” to the original

Perturbing the data with noise addition

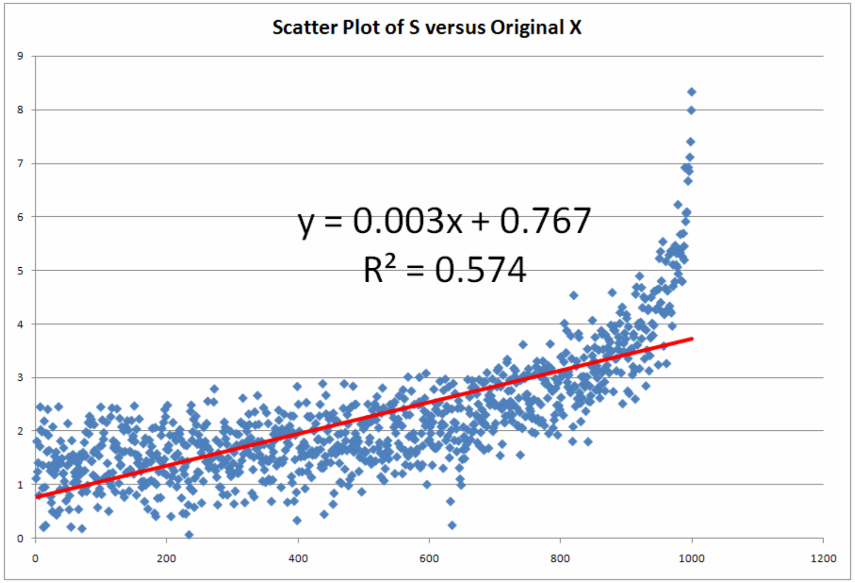
- Assume that the confidential variable X was perturbed using noise addition with variance of noise term equal to 10% of the variance of the confidential variable.
- What is the impact?

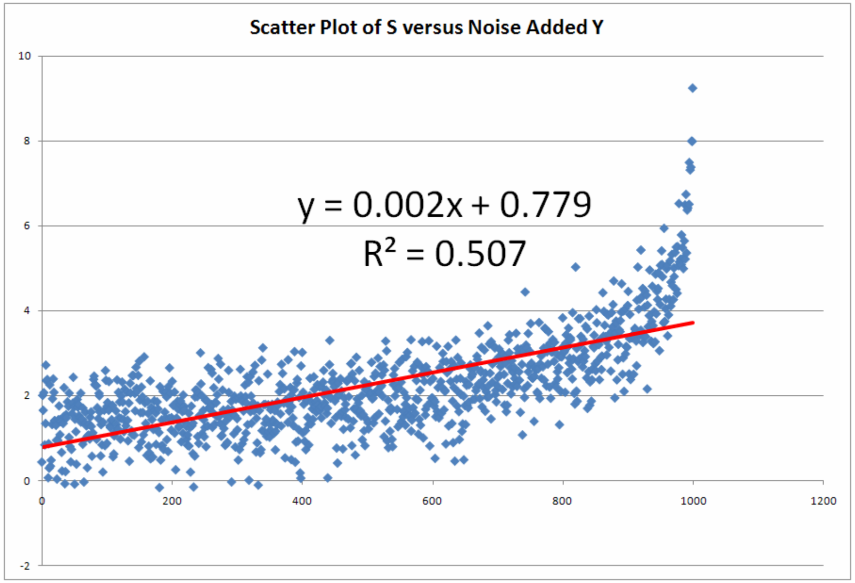
Impact on Marginal Distribution



Impact on Relationships

Even with small noise addition, the results are different





Comparison

- With noise addition
  - The marginal distribution of the perturbed variable is better than that of the IPSO variable
  - The (non-linear) relationship is maintained better than IPSO
- But
  - The variance of the noise added variable (1.4285) is higher than the variance of the original variable (1.2762) and the IPSO perturbed data
  - The correlation between the noise added variable and the non-confidential variable (0.7122) is lower than the correlation between the original variables (0.7578) and the IPSO perturbed variables
  - The results of linear regression using the noise added data is different than that using the original data and the IPSO perturbed data

Can we …

- Add noise but also maintain the mean vector and covariance matrix of the perturbed data to be the same as the original data?
- In other words, can we use the noise addition approach while also retaining the benefits of IPSO?

YES!

Sufficiency-Based Noise Addition

- Model:

$$y_i = \gamma + \alpha x_i + \beta s_i + \varepsilon_i$$

- The only parameter that must selected is the "proximity parameter" α.
- All other parameters are dictated by the selection of this parameter
- "Sufficiency-based" means that we will maintain the mean and covariance matrix exactly, and therefore any linear model constructed with the perturbed data will be exactly the same as the original data.

The Proximity Parameter

- The parameter α ($0 \leq \alpha \leq 1$) dictates the strength of the relationship between X and Y.
  - When $\alpha = 1$, $Y = X$.
  - When $\alpha = 0$, the perturbed variable is generated independent of X (the IPSO model)
  - We provide the ability to specify α to achieve any degree of proximity between these two extremes

Other Model Parameters

- $\beta = (1 - \alpha)(\sigma_{XS}/\sigma^2_{SS})$
- $\gamma = (1 - \alpha)\overline{X} - \beta\overline{S}$
- $\varepsilon \sim \text{Normal}(0, (1 - \alpha^2)(\sigma_{XS})^2/\sigma^2_{SS}$
- ε orthogonal to X and S

Note that …

- In order to maintain the mean and covariance exactly, it is NECESSARY that the model for generating the perturbed values MUST be specified in this manner
  - Other “model based” approaches CANNOT maintain the mean and covariance exactly

Univariate Case

- For this case, $0 \leq \alpha \leq 1$

- When $\alpha = 1$, the model will reduce to $y_i = x_i$

- When $\alpha = 0$, the model will reduce to
  $y_i = \gamma + \beta s_i + \varepsilon_i$ … the IPSO model

- $0 < \alpha < 1$ represents intermediate cases
  - When α is close to 1, the correlation between X and Y will be close to 1
  - When α is close to 0, the correlation between X and Y are conditionally independent given S.

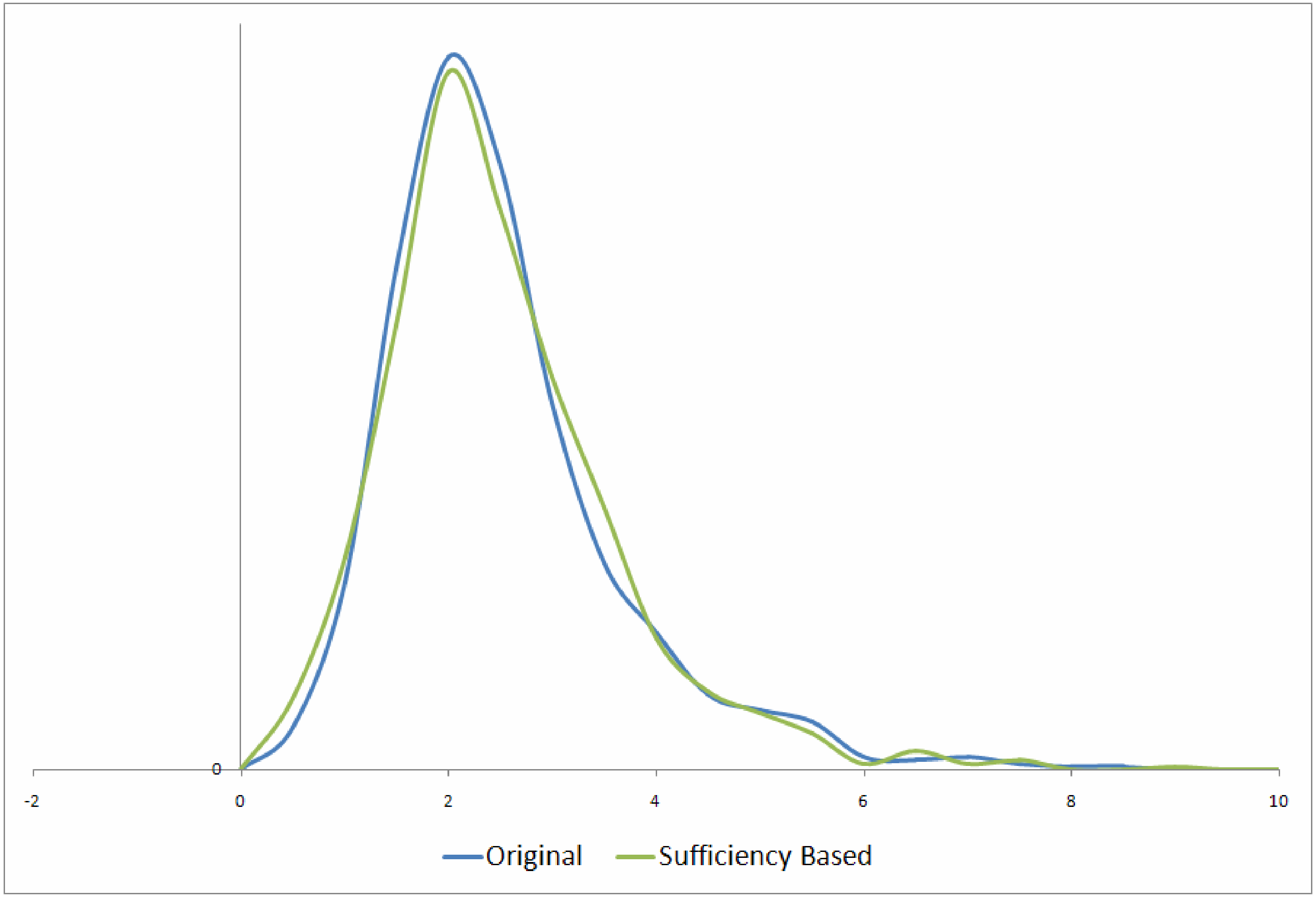
Equivalence to Noise Addition

- The selection of α is the same as selecting the variance of the noise term in noise addition
- For any given data set and a specific value of α, we can derive the effective level of noise that is added
- Conversely, if we wish to add a specified level of noise, we can determine the value of α based on this

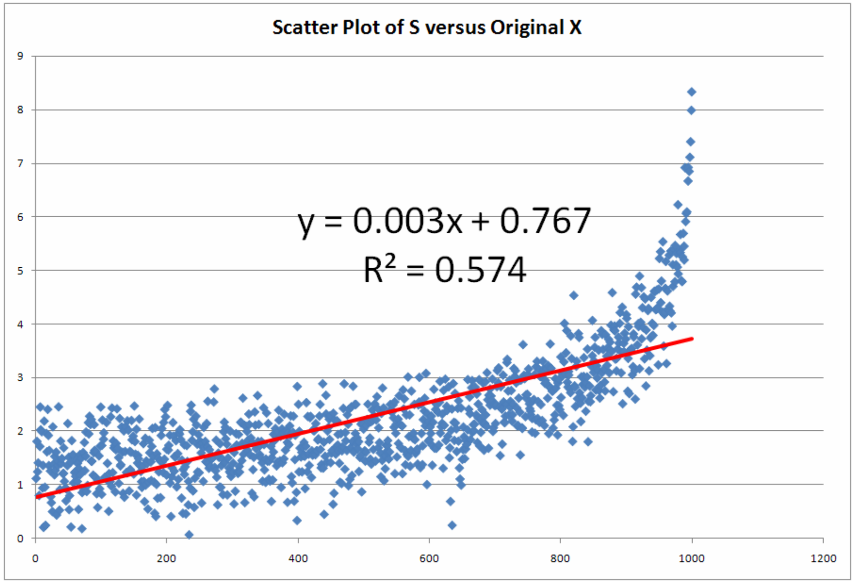
Example Data

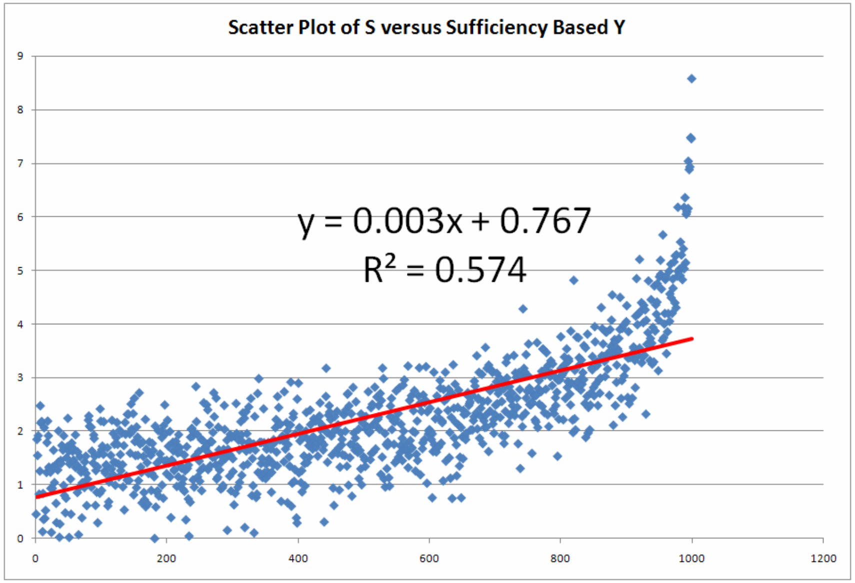
- α = 0.90
- From this specification
  - γ = 0.0767
  - β = 0.000296
  - Var(ε) = 0.10322
- This is the equivalent of implementing noise addition with approximately 8% noise
  - Var(ε)/Var(X) = 0.10322/1.2761 = 0.0809
- To get exactly 10% noise, α = 0.8747

Marginal Distribution



Relationships


Scatter Plot of S versus Original X


Scatter Plot of S versus Sufficiency Based Y

Summary

- The sufficiency based approach allows the data provider all the flexibility of noise addition while also preserving the mean vector and covariance matrix
  - There is no reason to implement simple noise addition

Multivariate Case

- The sufficiency based approach can be extended to multiple confidential variables
- Two possible cases
  - Equal proximity
  - Unequal proximity

Equal proximity

- In this case, we assume that the proximity of every perturbed variable to the original variable is the same
- $\alpha_i = \alpha$ for i = 1, 2, …, k
  - Where k is the number of confidential variables

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha & 0 & 0 & \ldots & 0 \\ 0 & \alpha & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \alpha \end{bmatrix}$$

$$0 \le \alpha \le 1$$

Unequal Proximity

- In this case, the proximity of one or more confidential variables is different from others

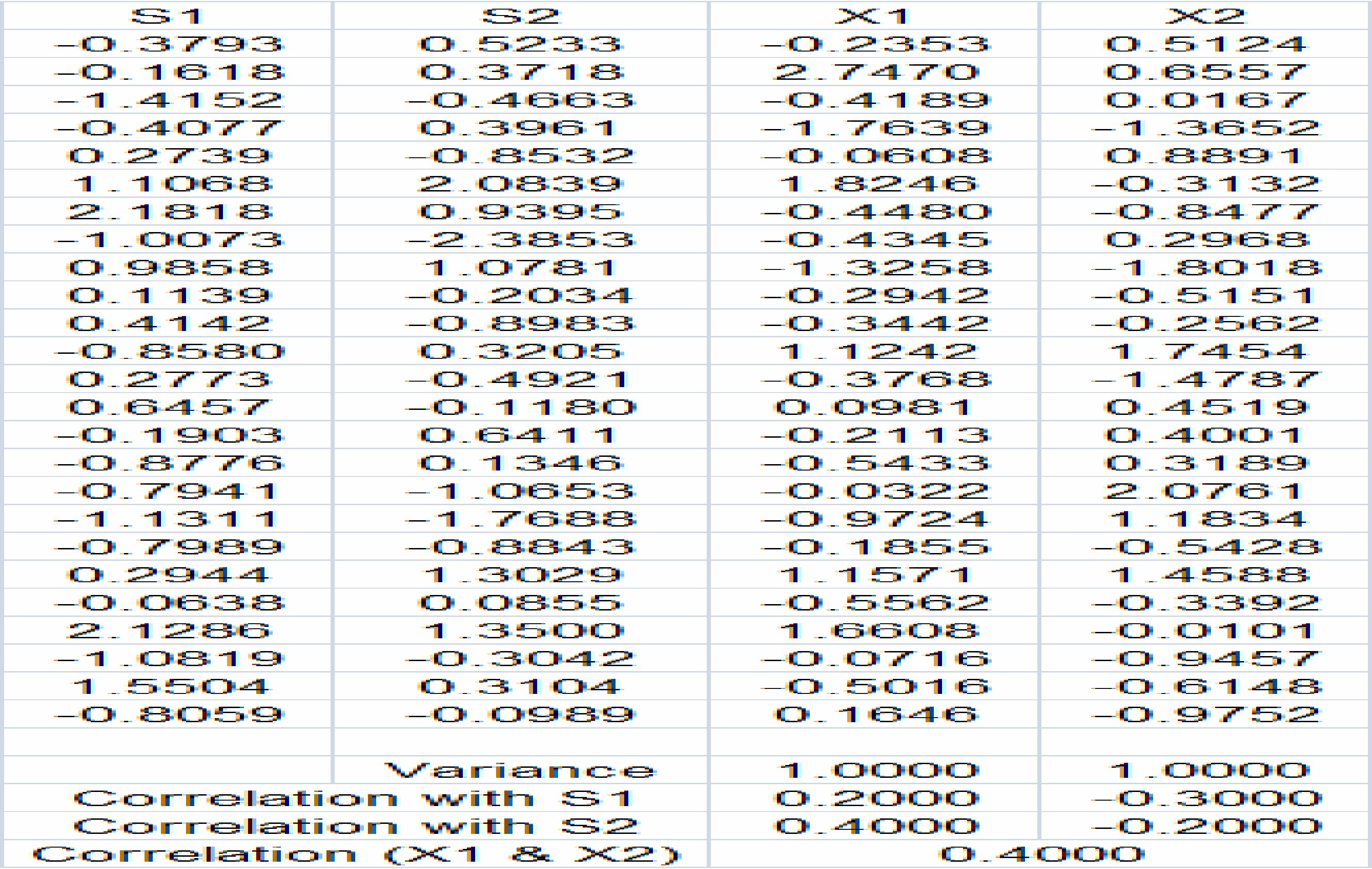
- $$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & 0 & 0 & \ldots & 0 \\ 0 & \alpha_2 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \alpha_k \end{bmatrix}$$ $0 \leq \alpha_i \leq 1$

- In this case, it would be necessary to verify that the resulting covariance matrix of the noise term ε is positive definite

Multivariate Example

- A data set with 2 non-confidential and 2 confidential variables. All variables have mean 0 and variance 1. The data set has 25 observations
  - We chose a small data set to show that regardless of the size of the data set, the procedure will work effectively

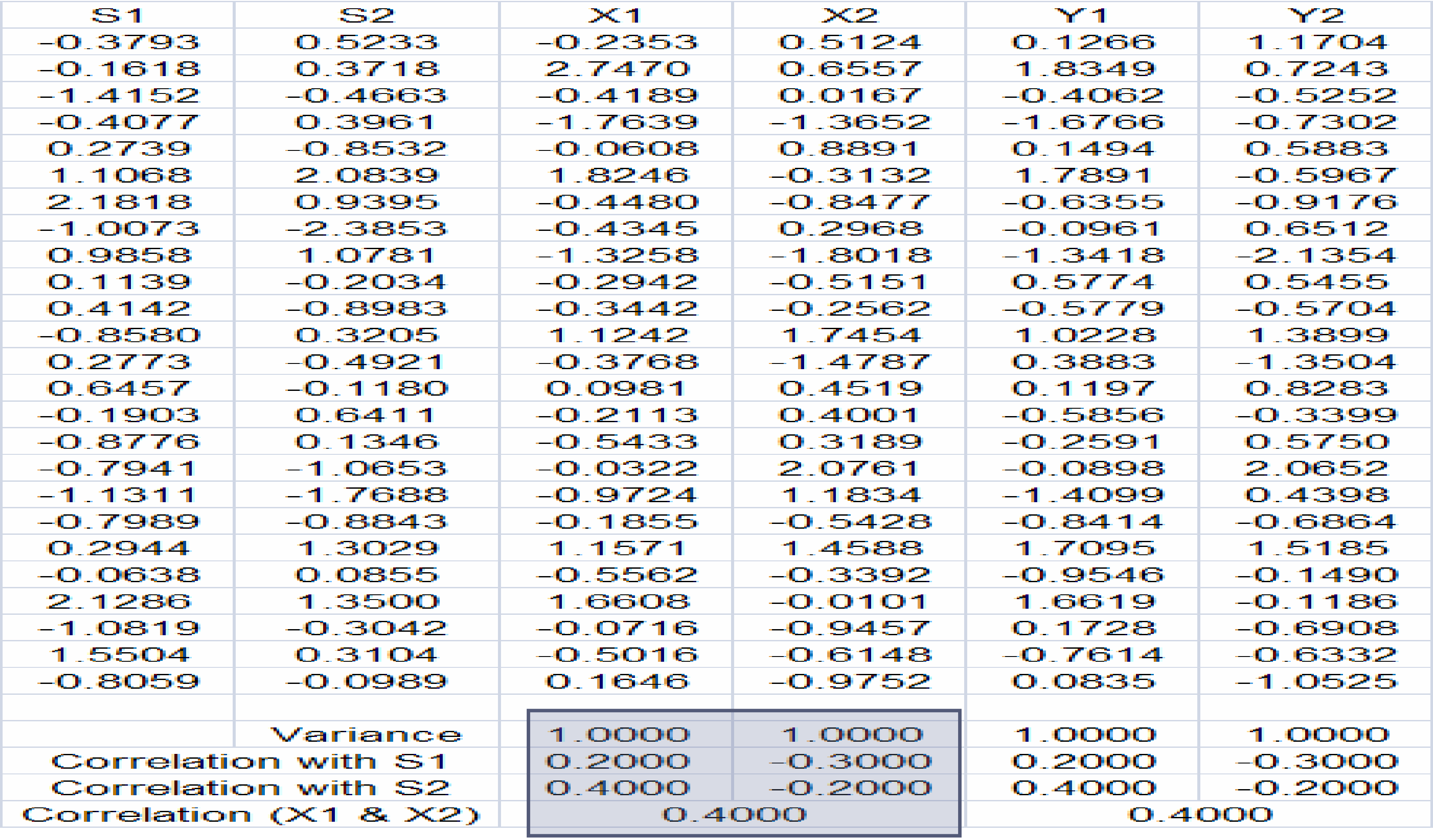
Original data



S1	S2	X1	X2
-0.3793	0.5233	-0.2353	0.5124
-0.1618	0.3718	2.7470	0.6557
-1.4152	-0.4663	-0.4189	0.0167
-0.4077	0.3961	-1.7639	-1.3652
0.2739	-0.8532	-0.0608	0.8891
1.1068	2.0839	1.8246	-0.3132
2.1818	0.9395	-0.4480	-0.8477
-1.0073	-2.3853	-0.4345	0.2968
0.9858	1.0781	-1.3258	-1.8018
0.1139	-0.2034	-0.2942	-0.5151
0.4142	-0.8983	-0.3442	-0.2562
-0.8580	0.3205	1.1242	1.7454
0.2773	-0.4921	-0.3768	-1.4787
0.6457	-0.1180	0.0981	0.4519
-0.1903	0.6411	-0.2113	0.4001
-0.8776	0.1346	-0.5433	0.3189
-0.7941	-1.0653	-0.0322	2.0761
-1.1311	-1.7688	-0.9724	1.1834
-0.7989	-0.8843	-0.1855	-0.5428
0.2944	1.3029	1.1571	1.4588
-0.0638	0.0855	-0.5562	-0.3392
2.1286	1.3500	1.6608	-0.0101
-1.0819	-0.3042	-0.0716	-0.9457
1.5504	0.3104	-0.5016	-0.6148
-0.8059	-0.0989	0.1646	-0.9752
	Variance	1.0000	1.0000
Correlation with S1		0.2000	-0.3000
Correlation with S2		0.4000	-0.2000
Correlation (X1 & X2)		0.4000	

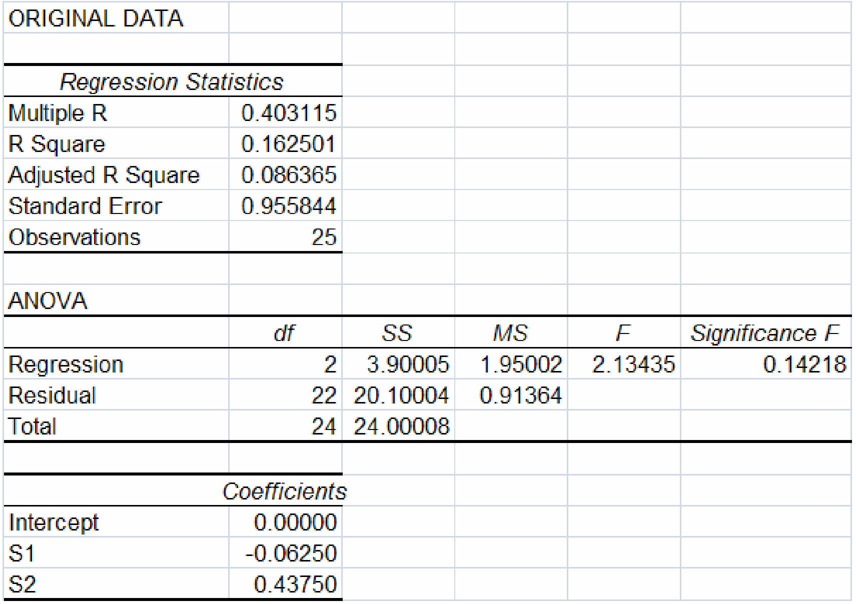
First Perturbed Example:

α = 0.90 for both variables



S1	S2	X1	X2	Y1	Y2
-0.3793	0.5233	-0.2353	0.5124	0.1266	1.1704
-0.1618	0.3718	2.7470	0.6557	1.8349	0.7243
-1.4152	-0.4663	-0.4189	0.0167	-0.4062	-0.5252
-0.4077	0.3961	-1.7639	-1.3652	-1.6766	-0.7302
0.2739	-0.8532	-0.0608	0.8891	0.1494	0.5883
1.1068	2.0839	1.8246	-0.3132	1.7891	-0.5967
2.1818	0.9395	-0.4480	-0.8477	-0.6355	-0.9176
-1.0073	-2.3853	-0.4345	0.2968	-0.0961	0.6512
0.9858	1.0781	-1.3258	-1.8018	-1.3418	-2.1354
0.1139	-0.2034	-0.2942	-0.5151	0.5774	0.5455
0.4142	-0.8983	-0.3442	-0.2562	-0.5779	-0.5704
-0.8580	0.3205	1.1242	1.7454	1.0228	1.3899
0.2773	-0.4921	-0.3768	-1.4787	0.3883	-1.3504
0.6457	-0.1180	0.0981	0.4519	0.1197	0.8283
-0.1903	0.6411	-0.2113	0.4001	-0.5856	-0.3399
-0.8776	0.1346	-0.5433	0.3189	-0.2591	0.5750
-0.7941	-1.0653	-0.0322	2.0761	-0.0898	2.0652
-1.1311	-1.7688	-0.9724	1.1834	-1.4099	0.4398
-0.7989	-0.8843	-0.1855	-0.5428	-0.8414	-0.6864
0.2944	1.3029	1.1571	1.4588	1.7095	1.5185
-0.0638	0.0855	-0.5562	-0.3392	-0.9546	-0.1490
2.1286	1.3500	1.6608	-0.0101	1.6619	-0.1186
-1.0819	-0.3042	-0.0716	-0.9457	0.1728	-0.6908
1.5504	0.3104	-0.5016	-0.6148	-0.7614	-0.6332
-0.8059	-0.0989	0.1646	-0.9752	0.0835	-1.0525
	Variance	1.0000	1.0000	1.0000	1.0000
Correlation with S1		0.2000	-0.3000	0.2000	-0.3000
Correlation with S2		0.4000	-0.2000	0.4000	-0.2000
Correlation (X1 & X2)		0.4000		0.4000	

Results of Regression to Predict X_1 given S_1 & S_2



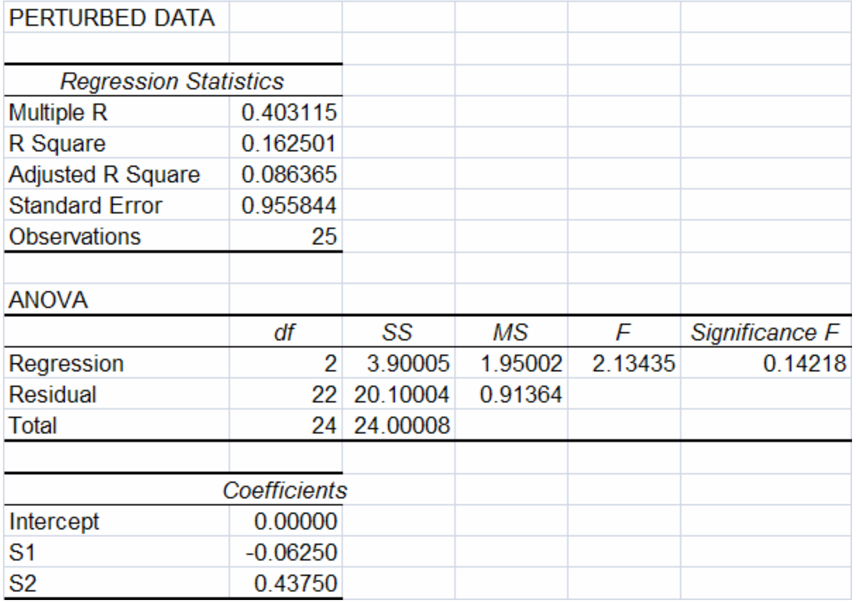
ORIGINAL DATA

Regression Statistics	
Multiple R	0.403115
R Square	0.162501
Adjusted R Square	0.086365
Standard Error	0.955844
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3.90005	1.95002	2.13435	0.14218
Residual	22	20.10004	0.91364		
Total	24	24.00008			

	Coefficients
Intercept	0.00000
S1	-0.06250
S2	0.43750



PERTURBED DATA

Regression Statistics	
Multiple R	0.403115
R Square	0.162501
Adjusted R Square	0.086365
Standard Error	0.955844
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3.90005	1.95002	2.13435	0.14218
Residual	22	20.10004	0.91364		
Total	24	24.00008			

	Coefficients
Intercept	0.00000
S1	-0.06250
S2	0.43750

First Perturbed Example:

α = 0.90 for first variable and α = 0.20 for second variable

- The resulting covariance matrix of the noise terms is shown.
- We can show that this covariance matrix is not positive definite and hence we must re-specif the values of α

$$\Sigma_{ee} = \begin{bmatrix} 0.1591 & 0.3844 \\ 0.3844 & 0.8730 \end{bmatrix}.$$

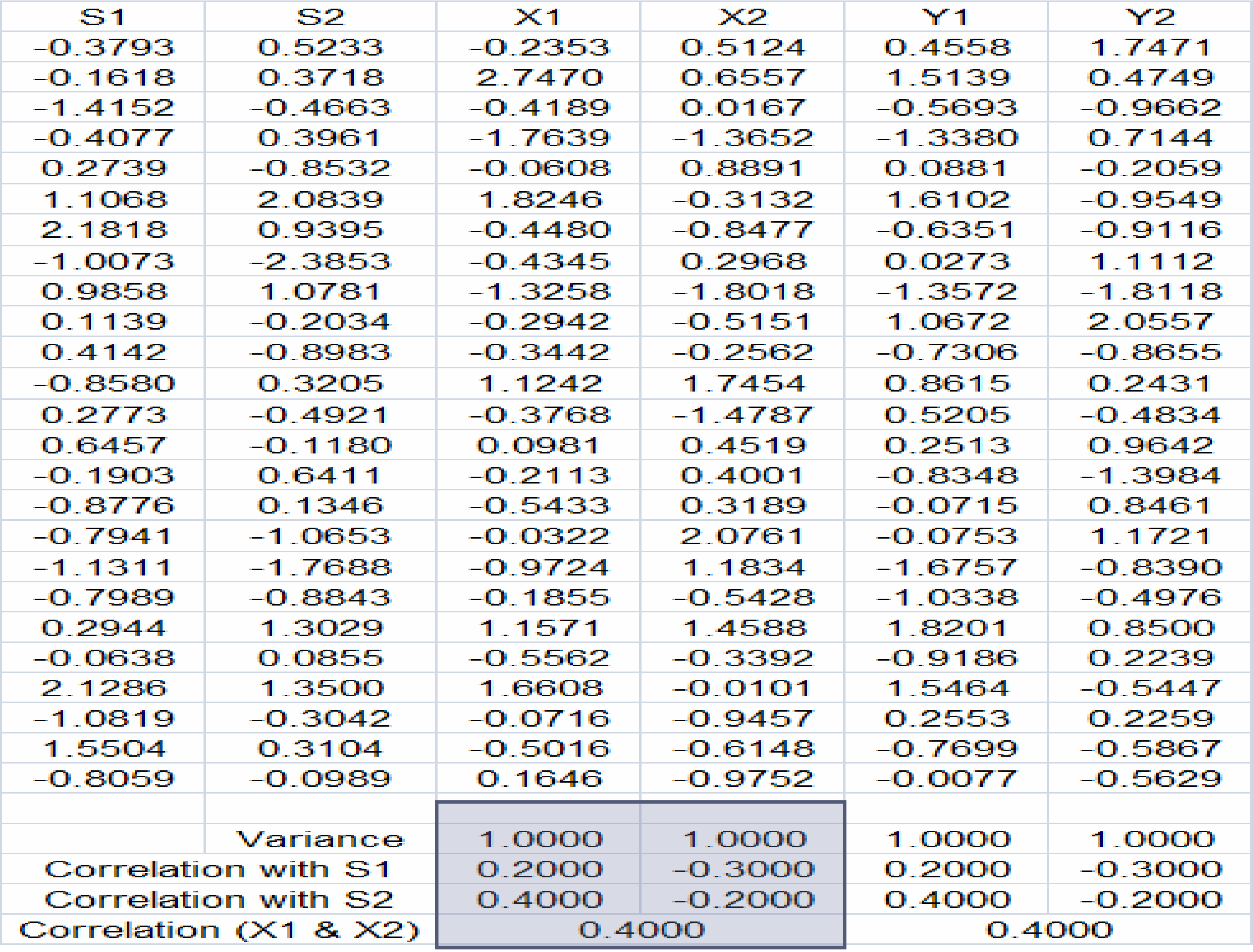
Alternative Specification

α = 0.80 for first variable and α = 0.30 for second variable

- We can verify that this covariance matri positive definite
- The perturbation performed with th values

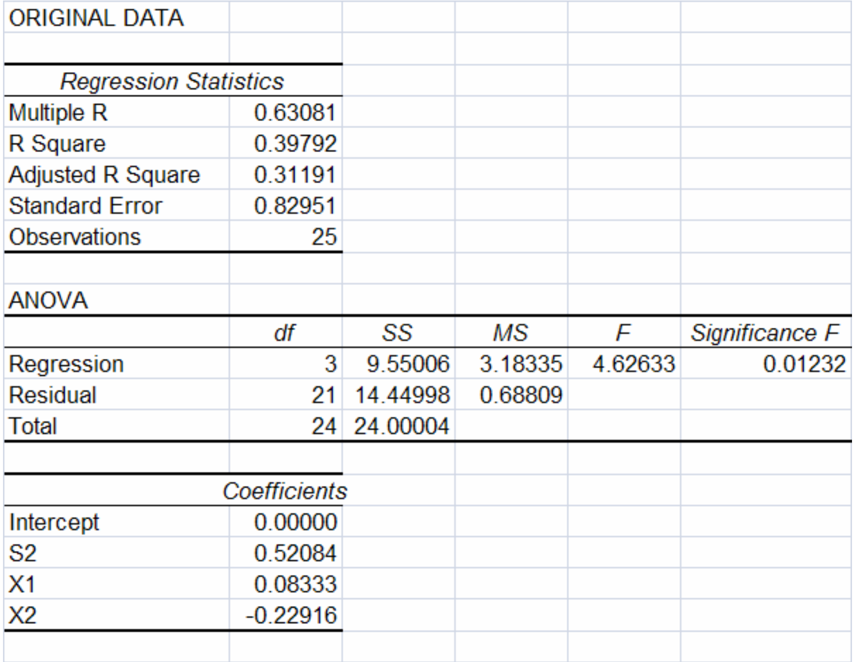
$$\Sigma_{ee} = \begin{bmatrix} 0.3015 & 0.3563 \\ 0.3563 & 0.8275 \end{bmatrix}.$$

Perturbed data



S1	S2	X1	X2	Y1	Y2
-0.3793	0.5233	-0.2353	0.5124	0.4558	1.7471
-0.1618	0.3718	2.7470	0.6557	1.5139	0.4749
-1.4152	-0.4663	-0.4189	0.0167	-0.5693	-0.9662
-0.4077	0.3961	-1.7639	-1.3652	-1.3380	0.7144
0.2739	-0.8532	-0.0608	0.8891	0.0881	-0.2059
1.1068	2.0839	1.8246	-0.3132	1.6102	-0.9549
2.1818	0.9395	-0.4480	-0.8477	-0.6351	-0.9116
-1.0073	-2.3853	-0.4345	0.2968	0.0273	1.1112
0.9858	1.0781	-1.3258	-1.8018	-1.3572	-1.8118
0.1139	-0.2034	-0.2942	-0.5151	1.0672	2.0557
0.4142	-0.8983	-0.3442	-0.2562	-0.7306	-0.8655
-0.8580	0.3205	1.1242	1.7454	0.8615	0.2431
0.2773	-0.4921	-0.3768	-1.4787	0.5205	-0.4834
0.6457	-0.1180	0.0981	0.4519	0.2513	0.9642
-0.1903	0.6411	-0.2113	0.4001	-0.8348	-1.3984
-0.8776	0.1346	-0.5433	0.3189	-0.0715	0.8461
-0.7941	-1.0653	-0.0322	2.0761	-0.0753	1.1721
-1.1311	-1.7688	-0.9724	1.1834	-1.6757	-0.8390
-0.7989	-0.8843	-0.1855	-0.5428	-1.0338	-0.4976
0.2944	1.3029	1.1571	1.4588	1.8201	0.8500
-0.0638	0.0855	-0.5562	-0.3392	-0.9186	0.2239
2.1286	1.3500	1.6608	-0.0101	1.5464	-0.5447
-1.0819	-0.3042	-0.0716	-0.9457	0.2553	0.2259
1.5504	0.3104	-0.5016	-0.6148	-0.7699	-0.5867
-0.8059	-0.0989	0.1646	-0.9752	-0.0077	-0.5629
	Variance	1.0000	1.0000	1.0000	1.0000
Correlation with S1		0.2000	-0.3000	0.2000	-0.3000
Correlation with S2		0.4000	-0.2000	0.4000	-0.2000
Correlation (X1 & X2)		0.4000		0.4000	

Results of Regression to predict S1 using the other 3 variables



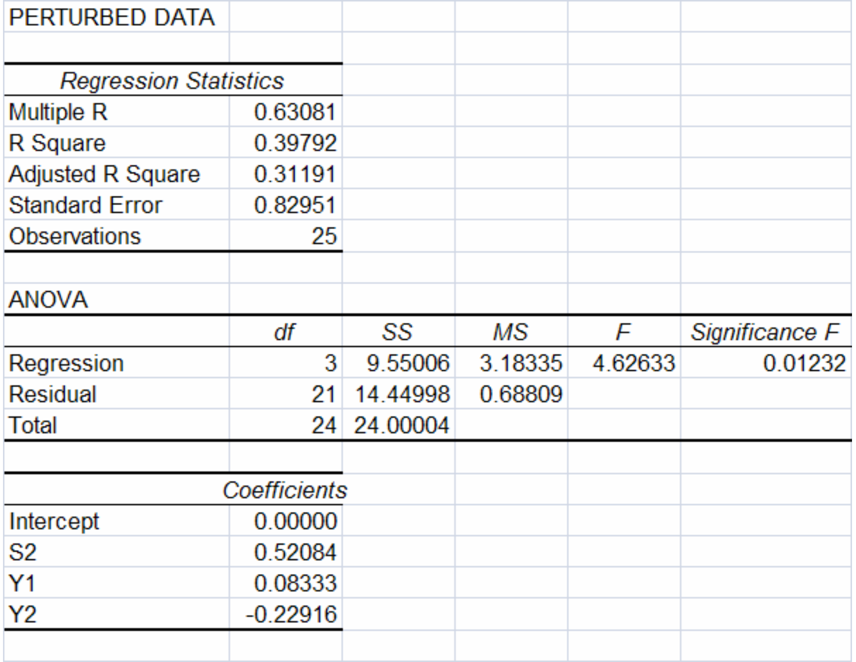
ORIGINAL DATA

Regression Statistics	
Multiple R	0.63081
R Square	0.39792
Adjusted R Square	0.31191
Standard Error	0.82951
Observations	25

ANOVA

	df	SS	MS	F	Significance F
Regression	3	9.55006	3.18335	4.62633	0.01232
Residual	21	14.44998	0.68809		
Total	24	24.00004			

	Coefficients
Intercept	0.00000
S2	0.52084
X1	0.08333
X2	-0.22916



PERTURBED DATA

Regression Statistics	
Multiple R	0.63081
R Square	0.39792
Adjusted R Square	0.31191
Standard Error	0.82951
Observations	25

ANOVA

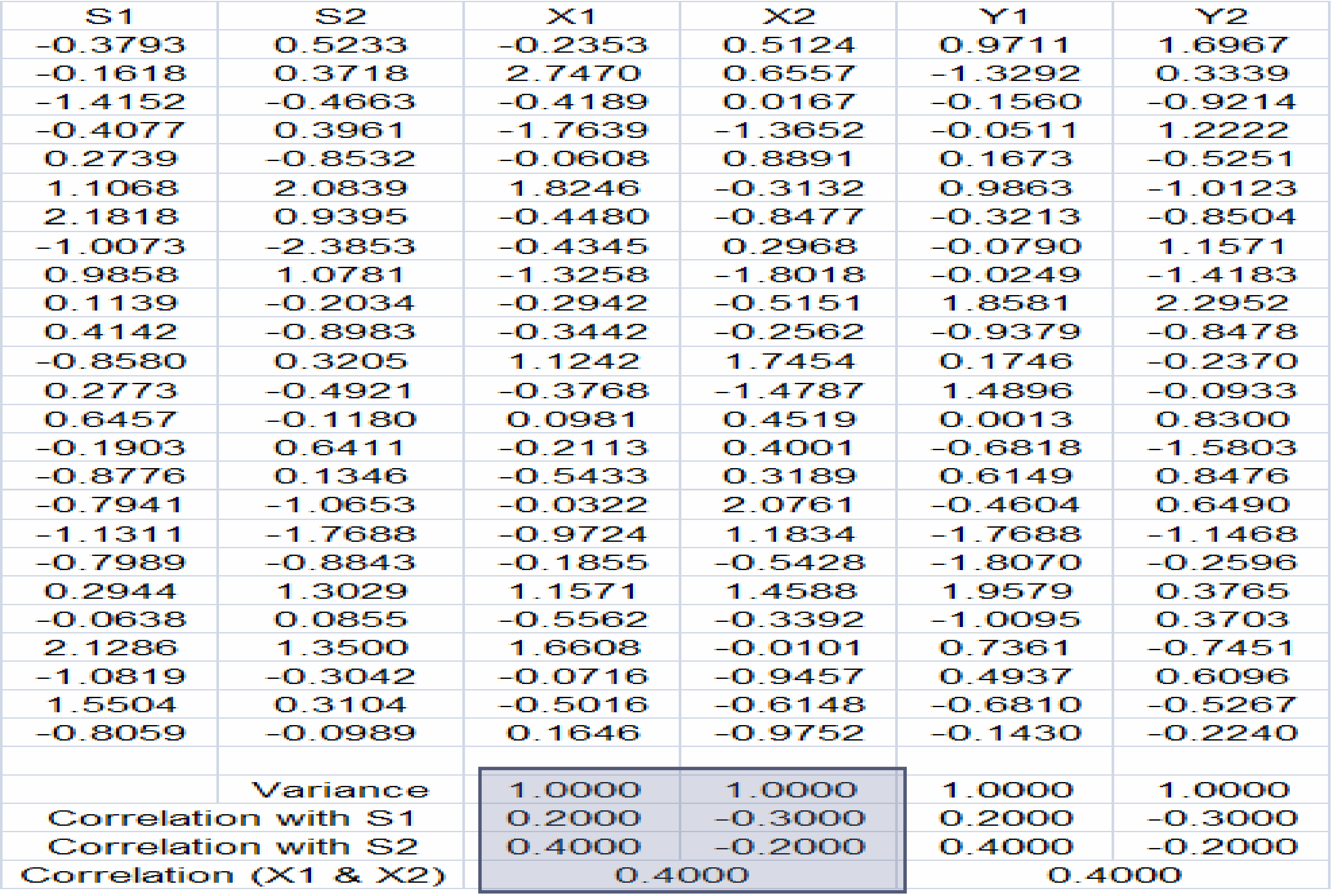
	df	SS	MS	F	Significance F
Regression	3	9.55006	3.18335	4.62633	0.01232
Residual	21	14.44998	0.68809		
Total	24	24.00004			

	Coefficients
Intercept	0.00000
S2	0.52084
Y1	0.08333
Y2	-0.22916

Third Example

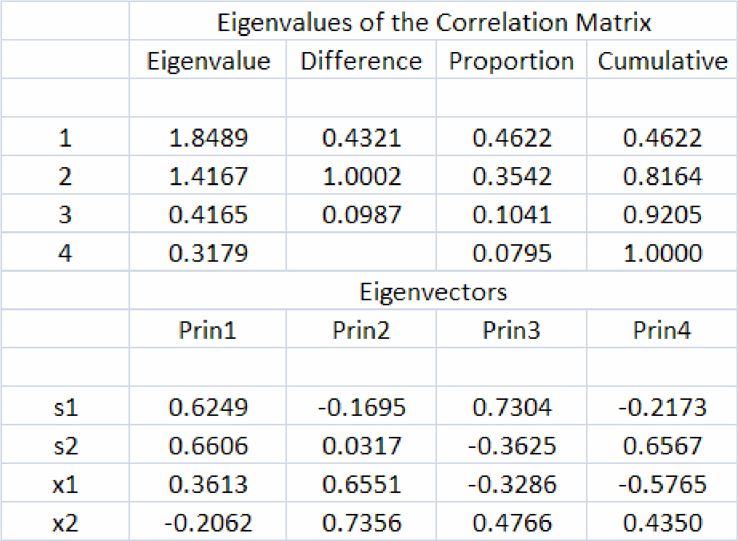
- In this perturbation, the value of α was specified as 0 for both variables. The resulting model is the IPSO model.

Third Example Data

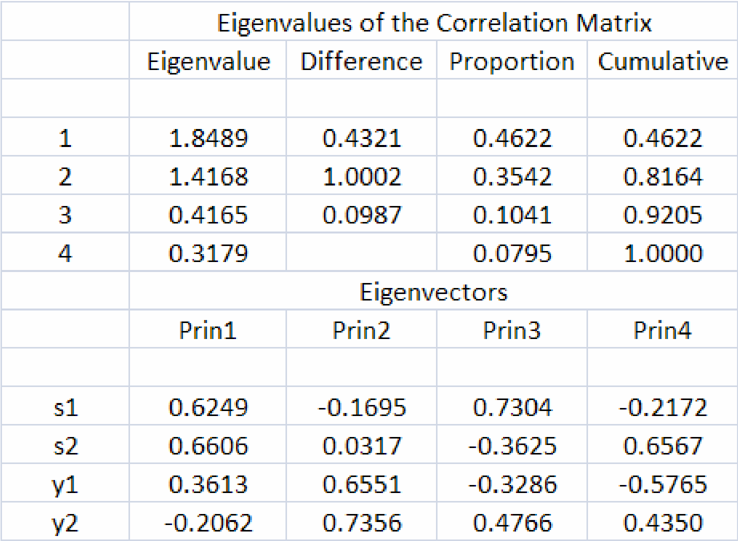


S1	S2	X1	X2	Y1	Y2
-0.3793	0.5233	-0.2353	0.5124	0.9711	1.6967
-0.1618	0.3718	2.7470	0.6557	-1.3292	0.3339
-1.4152	-0.4663	-0.4189	0.0167	-0.1560	-0.9214
-0.4077	0.3961	-1.7639	-1.3652	-0.0511	1.2222
0.2739	-0.8532	-0.0608	0.8891	0.1673	-0.5251
1.1068	2.0839	1.8246	-0.3132	0.9863	-1.0123
2.1818	0.9395	-0.4480	-0.8477	-0.3213	-0.8504
-1.0073	-2.3853	-0.4345	0.2968	-0.0790	1.1571
0.9858	1.0781	-1.3258	-1.8018	-0.0249	-1.4183
0.1139	-0.2034	-0.2942	-0.5151	1.8581	2.2952
0.4142	-0.8983	-0.3442	-0.2562	-0.9379	-0.8478
-0.8580	0.3205	1.1242	1.7454	0.1746	-0.2370
0.2773	-0.4921	-0.3768	-1.4787	1.4896	-0.0933
0.6457	-0.1180	0.0981	0.4519	0.0013	0.8300
-0.1903	0.6411	-0.2113	0.4001	-0.6818	-1.5803
-0.8776	0.1346	-0.5433	0.3189	0.6149	0.8476
-0.7941	-1.0653	-0.0322	2.0761	-0.4604	0.6490
-1.1311	-1.7688	-0.9724	1.1834	-1.7688	-1.1468
-0.7989	-0.8843	-0.1855	-0.5428	-1.8070	-0.2596
0.2944	1.3029	1.1571	1.4588	1.9579	0.3765
-0.0638	0.0855	-0.5562	-0.3392	-1.0095	0.3703
2.1286	1.3500	1.6608	-0.0101	0.7361	-0.7451
-1.0819	-0.3042	-0.0716	-0.9457	0.4937	0.6096
1.5504	0.3104	-0.5016	-0.6148	-0.6810	-0.5267
-0.8059	-0.0989	0.1646	-0.9752	-0.1430	-0.2240
	Variance	1.0000	1.0000	1.0000	1.0000
Correlation with S1		0.2000	-0.3000	0.2000	-0.3000
Correlation with S2		0.4000	-0.2000	0.4000	-0.2000
Correlation (X1 & X2)		0.4000		0.4000	

Principal Components Analysis using original and perturbed data



	Eigenvalues of the Correlation Matrix			
	Eigenvalue	Difference	Proportion	Cumulative
1	1.8489	0.4321	0.4622	0.4622
2	1.4167	1.0002	0.3542	0.8164
3	0.4165	0.0987	0.1041	0.9205
4	0.3179		0.0795	1.0000
	Eigenvectors			
	Prin1	Prin2	Prin3	Prin4
s1	0.6249	-0.1695	0.7304	-0.2173
s2	0.6606	0.0317	-0.3625	0.6567
x1	0.3613	0.6551	-0.3286	-0.5765
x2	-0.2062	0.7356	0.4766	0.4350



	Eigenvalues of the Correlation Matrix			
	Eigenvalue	Difference	Proportion	Cumulative
1	1.8489	0.4321	0.4622	0.4622
2	1.4168	1.0002	0.3542	0.8164
3	0.4165	0.0987	0.1041	0.9205
4	0.3179		0.0795	1.0000
	Eigenvectors			
	Prin1	Prin2	Prin3	Prin4
s1	0.6249	-0.1695	0.7304	-0.2172
s2	0.6606	0.0317	-0.3625	0.6567
y1	0.3613	0.6551	-0.3286	-0.5765
y2	-0.2062	0.7356	0.4766	0.4350

Assessment of Disclosure Risk

- The disclosure risk resulting from the sufficiency based approach is directly proportional to α
  - When α = 1, disclosure risk = 1
  - When α = 0, the disclosure risk is that resulting from the use of the non-confidential variables only

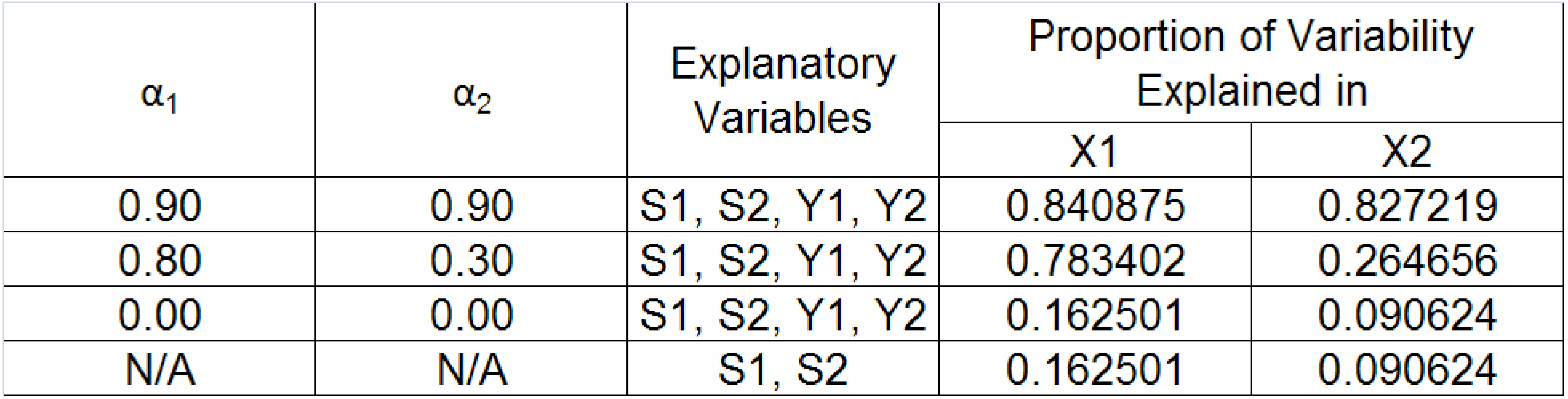
Information loss versus Disclosure Risk

- As α approaches 1, information loss decreases and disclosure risk increases
- As α approaches 0, information loss increases and disclosure risk decreases

An Assessment of Disclosure Risk

- We assessed disclosure risk as the proportion of variability in the original confidential variables using the non-confidential variables and the perturbed variable for the 3 models described above

Value Disclosure Risk



α_1	α_2	Explanatory Variables	Proportion of Variability Explained in	
			X1	X2
0.90	0.90	S1, S2, Y1, Y2	0.840875	0.827219
0.80	0.30	S1, S2, Y1, Y2	0.783402	0.264656
0.00	0.00	S1, S2, Y1, Y2	0.162501	0.090624
N/A	N/A	S1, S2	0.162501	0.090624

Summary

- The table shows that disclosure risk is directly proportional to the value of α
- The table also shows that the IPSO model offers the highest level of protection
  - All the information regarding the confidential variables is available from the non-confidential variables
  - The perturbed values do not provide any additional information

Conclusions

- We can implement additive noise methods while simultaneously preserving sufficient statistics
- There is no reason to implement noise addition without this important enhancement

Assurance to Users

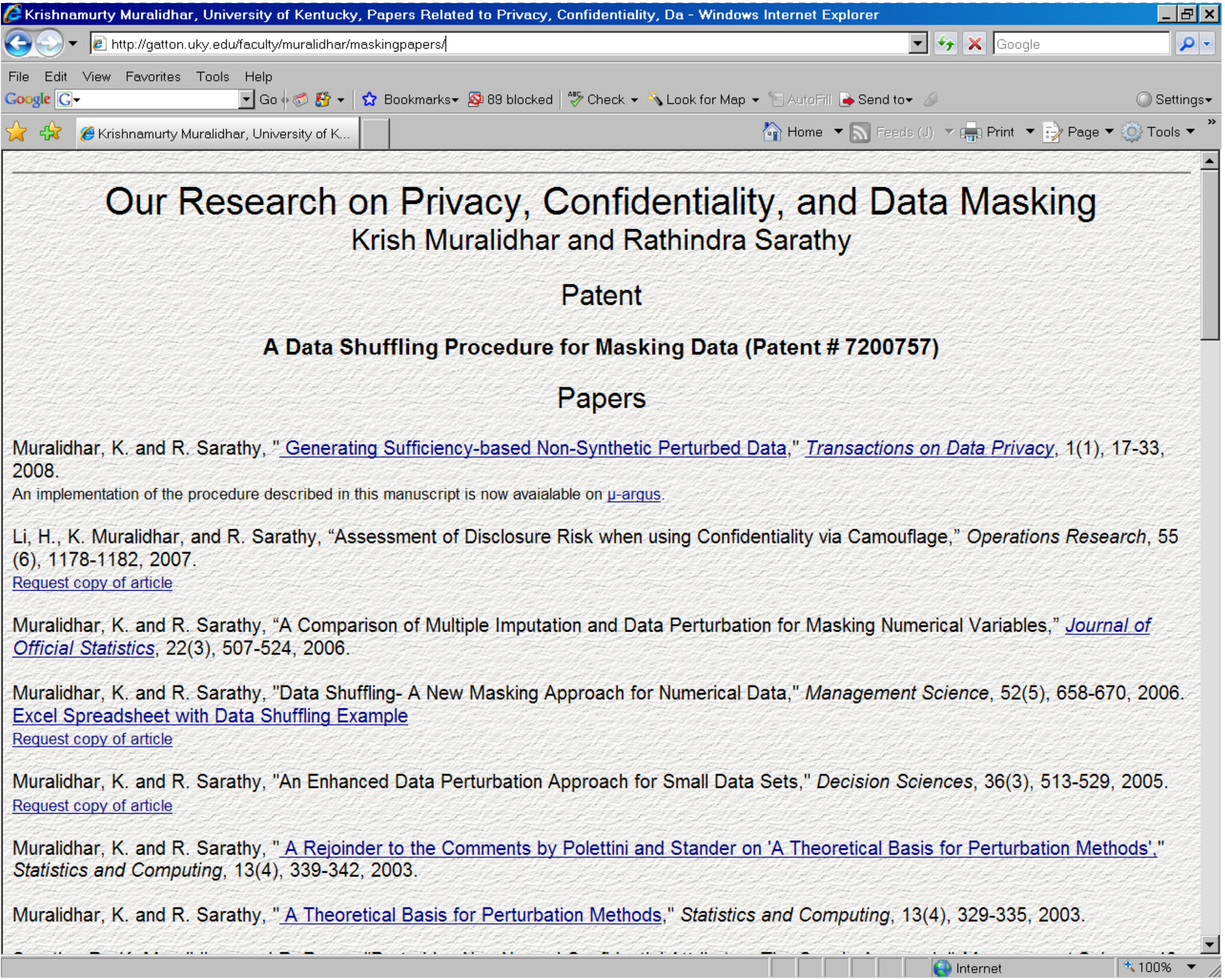
“The data has been perturbed to preserve the privacy and confidentiality of individual data points. The perturbation has been performed in such a manner that, for most traditional statistical analyses (see list below), the results of the analyses performed on this data will be **exactly the same as that using the original data**. It is important to note however that form (marginal distribution) of the confidential variables will be different from that of the original variables. In addition, this assurance does not extend to non-parametric analysis and non-traditional (data mining) analyses. If you have any questions regarding this data, please contact Mr. Joerg Drechsler for further information.”

Implementation in µ-Argus

- The sufficiency based noise addition approach is available on µ-Argus (or will soon be)

Our Web Site

http://gatton.uky.edu/faculty/muralidhar/maskingpapers/


Krishnamurty Muralidhar, University of Kentucky, Papers Related to Privacy, Confidentiality, Da - Windows Internet Explorer
http://gatton.uky.edu/faculty/muralidhar/maskingpapers/
Google
File Edit View Favorites Tools Help
Google G Go Bookmarks 89 blocked Check Look for Map AutoFill Send to Settings
Krishnamurty Muralidhar, University of K...
Home Feeds (J) Print Page Tools
Internet
100%

Our Research on Privacy, Confidentiality, and Data Masking

Krish Muralidhar and Rathindra Sarathy

Patent

A Data Shuffling Procedure for Masking Data (Patent # 7200757)

Papers

Muralidhar, K. and R. Sarathy, " Generating Sufficiency-based Non-Synthetic Perturbed Data," *Transactions on Data Privacy*, 1(1), 17-33, 2008.
An implementation of the procedure described in this manuscript is now avaialable on μ-argus.

Li, H., K. Muralidhar, and R. Sarathy, "Assessment of Disclosure Risk when using Confidentiality via Camouflage," *Operations Research*, 55 (6), 1178-1182, 2007.
Request copy of article

Muralidhar, K. and R. Sarathy, "A Comparison of Multiple Imputation and Data Perturbation for Masking Numerical Variables," *Journal of Official Statistics*, 22(3), 507-524, 2006.

Muralidhar, K. and R. Sarathy, "Data Shuffling- A New Masking Approach for Numerical Data," *Management Science*, 52(5), 658-670, 2006.
Excel Spreadsheet with Data Shuffling Example
Request copy of article

Muralidhar, K. and R. Sarathy, "An Enhanced Data Perturbation Approach for Small Data Sets," *Decision Sciences*, 36(3), 513-529, 2005.
Request copy of article

Muralidhar, K. and R. Sarathy, " A Rejoinder to the Comments by Polettini and Stander on 'A Theoretical Basis for Perturbation Methods'," *Statistics and Computing*, 13(4), 339-342, 2003.

Muralidhar, K. and R. Sarathy, " A Theoretical Basis for Perturbation Methods," *Statistics and Computing*, 13(4), 329-335, 2003.