

New Measures of Data Utility

A.F. Karr*, A. Oganian[†], J.P. Reiter[‡] and Mi-Ja Woo[§]

Abstract

When data is released to the public it is important to find the data alteration method with high confidentiality that provides satisfactory data quality. This paper focuses on developing methods of measuring data quality when the distribution of data is unknown but is postulated to be multivariate non-normal. We treat the data utility as a problem of evaluating similarities of original data structure to masked data structure. The data utilities we present here are rooted in the cumulative distribute function, clustering and propensity score approaches. When the distribution is departed from normal, simulations for a wide variety of data structures show how these measures can be used for evaluating disclosure limitation procedures.

Key words and Phrases : Confidentiality; Statistical disclosure limitation; Utility; Cumulative distribute function; Clustering; Propensity score.

1 Introduction

Dissemination of microdata serves the needs of researchers and the public by allowing them to analyze or model microdata. Society profits from disseminating microdata and research advances are expedited from it in many areas of knowledge. When microdata is released, data producers are charged with collecting high quality data. However, when it is disseminated in the collected form, it is possible that external users would reveal respondents' identities or identifying information about that individual contained in the data collections. Any effort to reveal them may be violating laws recently enacted to protect confidentiality such as HIPPA and CIPSEA

*National Institute of Statistical Sciences, Research Triangle Park, NC, USA.

[†]National Institute of Statistical Sciences, Research Triangle Park, NC, USA.

[‡]Duke University, Durham, NC, USA.

[§]National Institute of Statistical Sciences, Research Triangle Park, NC, USA.

(Wallman and Harris-Kojetin, 2004) in the U.S. Besides, data producers who do not preserve confidentiality may lose trust of the public, so they may not collect accurate answers or may be hard to find respondents in survey since potential respondents are reluctant to give correct answers or to participate in a survey.

Therefore, protection of confidentiality necessitates perturbing data in such a way of removing or altering values of identifying information and sensitive attributes before releasing data. There are many statistical disclosure limitation (SDL) strategies of reducing disclosure risk, such as adding noise to numerical values; swapping data values for selected records; releasing variables in aggregated categories, and so on. (see Willenborg and de Waal, 2001 for more details.)

SDL methods can be applied with differing degrees of intensity and increase of intensity lowers the disclosure risk. Also, it makes the data become less valuable since inferences obtained from the perturbed data are not accurate, which is often referred to as data utility. (Willenborg and de Waal, 2001)

There has been ample work done on developing measures of disclosure risk, which are discussed in Duncan and Lambert (1986, 1989), Lambert (1993), Fienberg et al. (1997), Skinner and Elliot (2002), and Reiter (2005a). In other words, there has not been much work on developing measures of data utility. For categorical data, Dobra et al. (2002) and Gomatam et al. (2005b) discussed measures based on inference. For numerical data, measures are developed in terms of differences between point estimates of the first and second moments based on the observed data and on the perturbed data by Domingo-Ferrer and Torra (2001), Yancey et al. (2002) and Oganian (2003).

Recently, A. F. Karr, C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2005) adopted an inference-based approach beyond moments and constructed a utility measure in the context of linear regression models. Their utility relies on differences between inferences based on observed data and those on masked data for linear regression models when data is assumed to be distributed with multivariate normal. Also, their utility measures the effects of SDL on interval estimation. Through Monte Carlo simulations and real data, they assessed the performance of their utility .

In many applications, however, it is unrealistic to expect that the observed or released data have multivariate normal distribution. If the data are departed from multivariate normal, but they are postulated by multivariate normal, then utility measures based on moments and inferences may be misleading and they may result in incorrect decision.

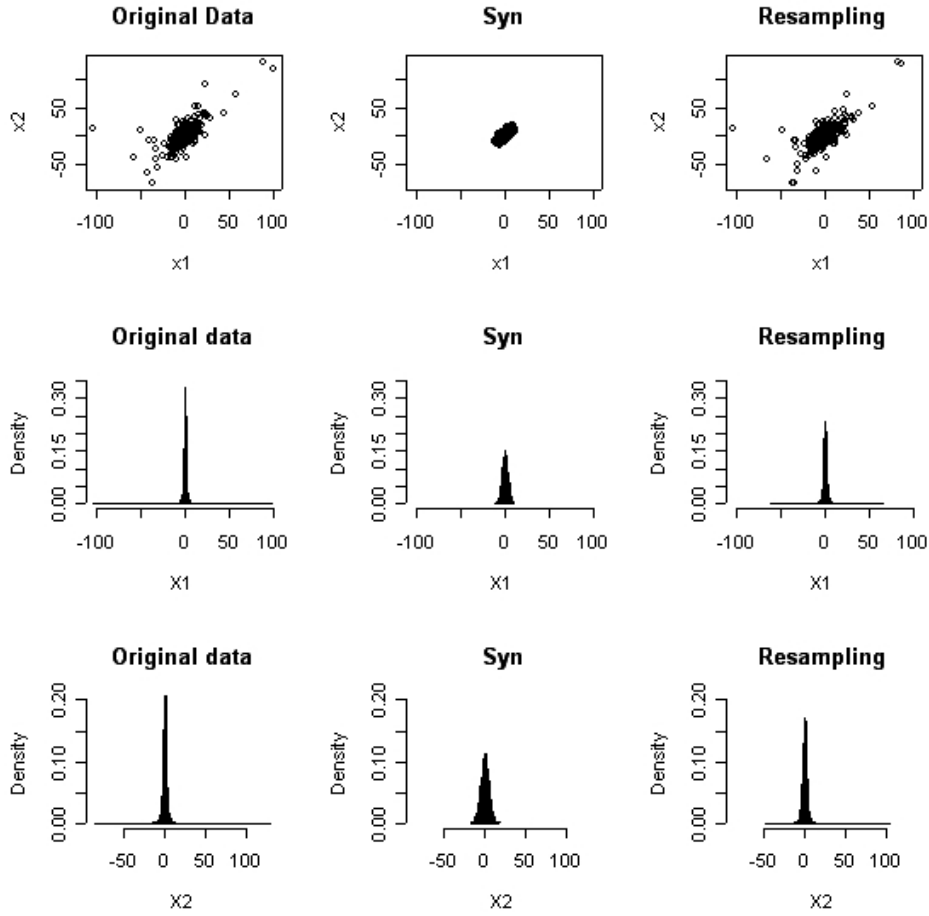


Figure 1: Scatter plots and histograms of X_1 and X_2 variables for original and two masked data.

To illustrate it, consider the two-dimensional data, where the data is not distributed with multivariate normal. To perturb this data, we employ two SDL methods, one of which is a synthetic method and the other of which is a resampling method. The masked data by synthetic method is created by generating samples from normal with empirical means and covariances of original data, where the masked data is transformed to have the same empirical means and covariances of the original data. To perturb the original data by resampling method, three bootstrap samples are used in this example. For simplicity, we abbreviate these SDL methods as Syn and Resampling. To assess utilities of two SDL strategies graphically, we examine the scatter plots and histograms of X_1 and X_2 variables, given in Figure 1. The scatter plots displayed on the first row in Figure 1 show that the masked data by resam-

Table 1: Moments and coefficient estimates for original and two masked data

		Original Data	Syn	Resampling
Moment	Mean of X_1	-0.02228	-0.022282	-0.039938
	Mean of X_2	-0.009305	-0.009305	-0.027097
	Variance of X_1	10.36007	10.36007	10.28796
	Variance of X_2	20.47026	20.47026	19.75331
	Correlation	0.73634	0.73634	0.73627
Regression	intercept	-0.01741	-0.01741	-0.02534
	Slope	0.52384	0.52384	0.53135

pling method has a more similar structure to original data than the masked data by synthetic method. Scatter plots are not sufficient to judge which one is better since they do not show how many data points are overlapped, so histograms are investigated. The histograms on the last two rows in Figure 1 also indicate that resampling method produces better utility than synthetic. Note that the data masked by synthetic method is normally distributed by the mechanism of generating the masked data by synthetic method. To compare utilities of two SDL methods in the context of moment and model-based inference, we obtained moments and inferences in the linear regression model. More specifically, means and covariance are computed for the moment-based utility, and the linear regression model is fitted to data by letting response and independent variables be X_1 and X_2 , respectively, for the model-based utility. Table 1 displays the values of mean, covariance and estimated coefficients of the linear model for each SDL strategy. The results show that masked data by synthetic method preserve the mean and covariance structures, while masked data by resampling method distort mean, covariance and coefficient estimates. These facts have an effect on the coefficients of the linear model for two SDL methods. Therefore, resampling method is dominated by synthetic method in terms of moments and inferences based on linear regression model. Even the same values are obtained for the masked data by synthetic method. As shown in Figure 1, these results are not consistent to graphical results. Also, resampling method is known to be SDL strategy which preserves the structure of original data even if its disclosure risk is high. It should be noted that assumption of multivariate normality is not satisfied except for the data masked by synthetic method. Therefore, it is doubtful that utilities gained from moments and inferences based on the linear regression model are good utility measures.

In this paper, our objective is to develop a utility measure which can be used when the multivariate normality is not assumed. To evaluate the data utility caused by an SDL method on a microdata, we propose to assess how similar the structure of the perturbed data is to that of the original data. In section 2, we present three measures of data utility which measure the preservation of the structure of the original data. In section 3, our utilities described in section 2 are employed to assess the features of the utilities for wide types of data. Finally, summary and conclusions are given in section 4.

2 Methods of Utility Measures

In this section, we present three measures of data utility which are not affected greatly by assumption of multivariate normality. The first utility considered is the extension of two sample tests based on empirical cumulative distribute function (CDF) for univariate data. More specifically, it measures the differences between empirical CDF of original and that of masked data. Next, we consider measuring data utility by using a clustering approach. As the last utility, we describe the utility which measures differences between estimated propensity scores of original data and those of masked data, where propensity score is defined as the conditional probability of assignment to a particular treatment given covariates. The properties of propensity scores and its application to utility measure are illustrated in the following subsection.

2.1 CDF Utility

CDF completely describes probability function. Therefore, the degree of differences between empirical distributions obtained from original and masked data can be appropriate for measuring data utility.

Let X_1, \dots, X_n be independent and identically distributed (iid) univariate random variables with distribute functions $F(x)$ and let Y_1, \dots, Y_m be iid univariate random variables with $G(y)$. We now consider testing the difference between two sample distributions. To this end, Kolmogorov statistics is calculated below.

$$D(S_1, S_2) = \sup_{1 \leq j \leq n+m} |S_1(z_j) - S_2(z_j)|, \quad (2.1)$$

where S_1 is the empirical CDF of the X sample, S_2 is that of the Y sample, the empirical distribute function $S(x)$ is defined by the fraction of x_i 's such that is less than

or equal to x , $-\infty < x < \infty$, and z_j is the j -th observation of the data pooling X and Y samples. Under the independence of X and Y samples, a variety of algorithms are used to calculate p-values based on D statistics. When the two sample sizes are equal, Lehmann (1975) proposed an algorithm to calculate the exact distribution of the two sided test. For unequal sample sizes, Kim and Jennrich (1973) give an algorithm to compute it for various sample sizes. In other cases, they recommend approximations using Smirnov's asymptotic distribution, after applying continuity corrections to the scaled test statistic. For more details on the empirical distribution function statistics, see Hollander and Wolfe (1973) and Gibbons and Chakraborti (1992). In our applications, original and masked data are not independent, so approximations mentioned above can not be adopted in our cases. However, it is clear that large D statistics imply that two samples X and Y are distributed differently without assuming independency of two samples.

We turn to CDF utility for the multivariate data. Let X_1, \dots, X_n be independent and identically distributed p -dimensional random variables with distribute function $F(\mathbf{x})$, and let Y_1, \dots, Y_m be independent and identically distributed p -dimensional random variables with distribute function $G(\mathbf{y})$. Also, let $S_X(\mathbf{x})$ and $S_Y(\mathbf{y})$ be the empirical distributions obtained from X and Y , respectively, and let $\hat{F}_X(\mathbf{x})$ be the empirical distributions. For multivariate case, $\hat{F}_X(\mathbf{x})$ is defined by

$$\hat{F}_X(x_1, \dots, x_p) = \frac{1}{n} \sum_{i=1}^n I(X_{i1} \leq x_1, \dots, X_{ip} \leq x_p)$$

where X_{ij} is the j -th variable of i -th observation, $i = 1, \dots, n$ and $j = 1, \dots, p$. Motivated by the univariate D statistics (2.1), we consider a CDF utility measure of the forms:

$$MD(S_X, S_Y) = \sup_{\mathbf{z}_i, 1 \leq i \leq n+m} |S_X(\mathbf{z}_i) - S_Y(\mathbf{z}_i)|, \quad (2.2)$$

and additionally, another CDF utility can be measured as following:

$$MCM(S_X, S_Y) = \sum_{i=1}^{n+m} (S_X(\mathbf{z}_i) - S_Y(\mathbf{z}_i))^2, \quad (2.3)$$

where \mathbf{z}_i is the i -th record of the data combining original data and masked data. Here, large MD and MCM statistics tell that two samples X and Y are distributed differently. MD takes on values of zero to one, and MCM has the values of zero to $(n+m)(2nm+1)/(6nm)$. For example, when two samples are perfectly same, MD and MCM have values of zero, while when all records from one sample are greater

or less than those from the other sample, MD and MCM have the values of one and $(n + m)(2nm + 1)/(6nm)$.

2.2 Cluster Utility

A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. Therefore, a cluster can be regarded as a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.

Note that a data set can be said to be randomly assigned when the proportion of observations from original data for each cluster is constant. In other words, two data sets can be said to have the same distributions when proportion of observations from original or masked data for each cluster is constant. Motivated by it, we test that the distribution of original data is the same as that of masked data by assigning observations in pooled data to clusters and computing differences between the number of observations from original and masked data for each cluster. Let g be the number of clusters. Then, cluster utility is defined by

$$U_{cluster} = \sum_{j=1}^g w_i \left(\frac{n_{i1}}{n_i} - c \right)^2, \quad (2.4)$$

where n_i is the total number of observations grouped in i -th cluster, n_{i1} is the number of observations from original data for i -th cluster, c is a constant and w_i is the weight assigned to i -th cluster. In cluster utility, one value is obtained for each cluster, so w_i in (2.4) makes cluster utility explain how much each cluster contributes to the utility. Clustering methods can be clustered in many different ways. In our computations, weight is selected by n_i , and average linkage method is used to cluster records in terms of squared distances. (Gordon, 1999, Kaufman and Rousseeuw, 1990)

2.3 Propensity Score Utility

Consider the N units $\mathbf{x}_1, \dots, \mathbf{x}_N$, and two treatments 1 and 0. Let $r_i = 1$ if unit i is assigned to the treatment 1 and $r_i = 0$ if unit i is assigned to the treatment 0. Let \mathbf{x}_i be an observed covariate for the i -th unit. Also, consider the conditional probability of assignment to treatment 1, given the covariate \mathbf{x} , that is, $e(\mathbf{x}) = P(r = 1|\mathbf{x})$. The function $e(\mathbf{x})$ is called the propensity score in that it is the probability of being exposed toward treatment 1 given the observed covariate \mathbf{x} .

Assuming propensity scores are known, Rosenbaum and Rubin (1983) showed that treatment assignment and the observed covariates are conditionally independent given the propensity score, that is,

$$\mathbf{x} \perp r \mid e(\mathbf{x}).$$

It implies that if a subclass of units is homogeneous in $e(\mathbf{x})$, then the treated and control units in that subclass will have the same distribution of \mathbf{x} . It is practically important to assess the utility measure. Consider the class of \mathbf{x} , $\lambda(e) = \{\mathbf{x} : e(\mathbf{x}) = e\}$. The conditional distribution of \mathbf{x} in $\lambda(e)$ is the same for treated and control units. In order to test that the distribution of \mathbf{x} is the same for treated and control units for all \mathbf{x} , the class of \mathbf{x} , $\lambda(e)$ should be equal to the class of all \mathbf{x}_i 's. Therefore, testing the differences of two sample distributions is equivalent to testing $e(\mathbf{x}) = c$ for all \mathbf{x} , where c is a fixed number. Also, we can show that c is $P(r = 1)$.

In practice, propensity scores are unknown, so they should be estimated by modelling propensity scores. There are several ways of modelling propensity scores such as logistic model, tree model, and so on. In our computations, we consider slight modification of logistic model, as well as logistic model and tree model. The procedure of modified logistic model starts by classifying pooled data points into groups for a fixed number of groups, which are regarded as partitions of whole space such that \mathbf{x}_i 's are homogeneous in each partition. After classifying data points into groups, propensity score estimates are obtained by modeling linear logistic function to data fallen in partitions. Finally, propensity score utility is measured by

$$\sum_{i=1}^N (\hat{e}(\mathbf{x}_i) - c)^2, \tag{2.5}$$

where $\hat{e}(\mathbf{x}_i)$ is the estimated propensity score at x_i , and c is estimated by the proportion of treated units in pooled data.

Either logistic model or clustering is adopted to measure data utility as described previously. There are several reasons of modifying logistic model instead of using either one. Two methods possess conflicting properties. Estimation of propensity scores is very crucial in measuring propensity score utility. When we choose logistic model, it is a polynomial in \mathbf{x} of degree k , and in setting degree k we wish to consider the largest k since the larger k yields more exact estimates of propensity scores. In high-dimensional data, however, it is not easy to fit logistic model to data since the number of terms in the model increase exponentially to dimension p . Besides, there

is no guarantee that degree k is enough to estimate propensity scores. In propensity score utility with logistic model, each covariate contributes to utility measure since each one has its own propensity score. In cluster utility, the covariates assigned to the same cluster produce the same value. Cluster utility does not measure the differences between two structures of original and masked data within a cluster, that is, within-cluster variation, since the proportion n_{i1}/n_i and the number of covariates w_i assigned to the cluster are considered to obtain the cluster utility.

By combining two approaches, we can reconcile the conflicting properties of two methods. By partitioning space and fitting linear logistic model for each partition separately, we can borrow strength of much higher degree k since a function can be approximated by linear function. It can be used to high-dimensional data without much difficulty. Also, we can measure within-cluster variation by fitting logistic model for each cluster.

Notice that the cluster utility actually defines one way of propensity score utility since the value of proportion can be regarded as the propensity score estimates corresponding to all observations within a cluster. As in propensity score utility, constant c in (2.4) is used by the proportion of observations from original data.

3 Simulation Study

In this section, we examine the performance of our several utilities using simulated data. We carry out these studies in various types of data, but in all data the multivariate normal is not postulated. The eight types of two-dimensional data are created by crossing the following three cases: (i) when data are generated from a non-normal where distributions are symmetrically and non-symmetrically departed from multivariate normal distribution, (ii) when data are generated from non-normal distributions where variables are highly and lowly correlated to each other, and (iii) when data are generated from non-normal distributions where variables are negatively and positively correlated to each other. These data span over symmetric and non-symmetric departures from normality along with four different types of correlation. By differing data structures, we can illustrate the features of our utilities for a variety of characteristics of data. The samples are drawn from various data structures with $n = 10,000$.

For our applications, we use a variety of SDL methods based on SDL investigated by Oganian (2003); microaggregation using z-score projection to group observations

together with three observations per group ($k = 3$); rank swapping where each ranked value of a variable is swapped with another ranked value randomly chosen within a $p = 0.15$ of the total number records; resampling where bootstrap samples are drawn independently $t = 3$ times from the data with replacement for each variable; synthetic where normal samples are generated from empirical mean and covariance of original data; and microaggregation using z score projection with $k = 3$ followed by normal noise, where noises are generated from normal with mean zero and empirical covariance of data consisting in differences between original and microaggregated data.

Let X_1, \dots, X_n be random samples drawn from the eight different types of data structures previously specified. Also, let Y_1, \dots, Y_n be the masked data of X_1, \dots, X_n using SDL methods described above. For calculations of our utilities, we combine original data with masked data, and denote pooled data by Z_1, \dots, Z_{2n} . Then, we create a new variable R , where the $R_i = 0$ if record i is from original data, and $R_i = 1$ otherwise, $i = 1, \dots, 2n$. Given a data set, Z_1, \dots, Z_{2n} , the procedures of computing three utilities are implemented for each SDL method. There are several computational details in the course of carrying out the propensity score and clustering utilities.

The first issue is the choice of model when estimating propensity score for each covariate. In our simulation, we consider logistic model, tree model and modified logistic model. In the propensity score utility using logistic model, the selection of model is critical for the estimation of propensity score. We wish to take as high degree of polynomial in \mathbf{x} as we can since fitness is better as the degree increases. For our simulations, we consider the following two logistic models to Z_1, \dots, Z_{2n} :

$$\text{Model I} : \text{logit}(p_i) = \beta_0 + \beta_1 Z_{i,1} + \beta_2 Z_{i,2} + \beta_3 Z_{i,1} Z_{i,2} + \beta_4 Z_{i,1}^2 + \beta_5 Z_{i,2}^2 \quad (3.1)$$

$$\begin{aligned} \text{Model III} : \text{logit}(p_i) = & \beta_0 + \beta_1 Z_{i,1} + \beta_2 Z_{i,2} + \beta_3 Z_{i,1} Z_{i,2} + \beta_4 Z_{i,1}^2 + \beta_5 Z_{i,2}^2 \\ & + \beta_6 Z_{i,1}^2 Z_{i,2}^2 + \beta_7 Z_{i,1}^3 + \beta_8 Z_{i,2}^3, \end{aligned} \quad (3.2)$$

where $p_i = P(R_i = 1|Z_i)$ and $Z_{i,j}$ is the j -th variable of i -th records in pooled data.

When we perform the propensity score utility using tree model, the size of tree is crucial for measuring utility, and complexity parameter (cp) is considered as one way of specifying the size of tree. (See W. N. Venables and B. D. Ripley, 2002.) For example, any split that does not decrease the overall lack of fit by a factor of 'cp'

is not attempted. In our simulations, we select two values of cp by $cp=0.001$ and $cp=0.0001$.

Next, in computing cluster utility, the number of clusters has an effect on the measurement. Small number of clusters makes it hard to take account on differences between original data structure and masked data structure since cluster utility does not evaluate within-cluster variation. On the other hands, large number of clusters makes utility too sensitive to small differences since it does not account for distances between clusters. In our simulation, we consider $g=500$ (5%) and $g=1000$ (10%).

The last issue is the choice of the number clusters in the propensity score utility using modified logistic model. Unlike cluster utility, the propensity score utility measures the within-cluster variation by performing logistic model to data for each cluster. We want to consider both rough and small differences in original and masked data structures, simultaneously. In order to do it, we wish to pick the smaller number of groups than in cluster utility, and to fit linear logistic model to the data for each group. In our simulations, we choose the number of clusters as $g = 100$ (1%), and in addition to linear logistic function, we also fit quadratic logistic function to look at the effect of degree on propensity score utility.

We implement our utilities for eight different type of data structures, and their results are reported in Tables 2 to 9, where MD and MCM correspond to the CDF utilities given by (2.2) and (2.3), Logistic I and Logistic II to the propensity score utilities given by logistic model (3.1) and (3.2); Tree model I and II to the propensity score utilities using tree model given by $cp=0.001$ and $cp=0.0001$; Modified Logistic I and Modified Logistic II to the propensity score utilities fitting linear and quadratic logistic functions; and Clustering $g=500$ and $g=1,000$ to the cluster utility with the number of cluster as $g=500$ and $g=1,000$. Also, Syn is denoted by synthetic method; Micz03 is denoted by microaggregation; Micz03+Noise is denoted by Micz03 followed by normal noise; Rank is denoted by rank swapping; and Resampling is denoted by resampling.

Our results from these studies are as follows. For symmetric cases, except Logistic I and Logistic II, all utilities show that Rank and Resampling are better than other three SDL methods. For Logistic II, Rank and Resampling are better for some data structures, and Rank is worse than Micz03+Noise for some. When we compare Syn, Micz03 and Micz03+Noise, the utilities indicate that Micz03 is the best and Syn is the worst but for Logistic I, Logistic II, and Tree Model II. More specifically, contrary to other utilities, Logistic I yields very small values of utility for Syn and Micz03+Noise,

whereas Logistic II and Tree model II produce smaller value of Micz03+Noise than Micz03. Especially, Tree Model II gives excessive differences in estimated propensity scores to Micz03. When we compare Rank and Resampling, Resampling is better than Rank in terms of all utilities other than CDF and Tree model I. As shown in Tables 2 to 9, CDF does not judge which one is better than the other, and Tree model I can not distinguish utilities of two SDL methods.

For non-symmetric cases, except Logistic I and Logistic II, all utilities show that Rank and Resampling are better than other three SDL methods, which are similar to symmetric cases. When we compare Syn, Micz03 and Micz03+Noise, except Logistic I, all utilities indicate that Micz03+Noise is the best out of three SDL methods for high-negative, high-positive and low-negative data structures, while Micz03 is the best for low-positive data structure. However, Tree Model II and Cluster utilities yield excessive differences in estimated propensity scores to Micz03 for most data structures as in symmetric cases, and it is more severe than for symmetric cases. Also, Micz03 is not differentiable from Syn for high-negative, high-positive and low-negative cases. When we compare Rank and Resampling, Resampling is better than Rank with respect to all utility measures other than Tree model I, Tree model II and CDF. As shown in symmetric cases, CDF and Tree model I can not make clear which utility is better than the other of two SDL methods, but unlike symmetric cases, even Tree model II gives zero utilities to Rank.

Notice that there are three replicates in Micz03 masked data. When we use tree model with large size of tree, it is easier to separate original data points from masked data points by Micz03 than by the other SDL methods listed in Tables 2 to 9. Tree model with large size of tree gives more opportunity of disparities for microaggregation, which results in the worse utility of Micz03. The same behaviors happen when we apply for cluster utility with large number of clusters since we group records into clusters. Also, it is more severe for non-symmetric cases than for symmetric cases.

In Rank swapping, the ranked value is swapped with next ordered value, so marginal distribution of each variable is preserved, but covariance structure is distorted. The characteristic of Rank swapping, that is, preserving marginal distribution of original data may be one of reasons that CDF is favorable to Rank swapping and that tree model with small size of tree can not prune the whole space of data combining original data and masked data by Rank swapping. Besides, tree model with large size of tree suffers this problem for non-symmetric data structure.

Overall, according to the utilities considered in the studies, SDL methods are ordered from the highest to the lowest as follows: for symmetric data structure, Resampling, Rank, Micz03, Micz03+Noise, and Syn; for non-symmetric data structure with highly negative, highly positive, and lowly negative, Resampling, Rank, and Micz03+Noise, but Micz03 and Syn are similar; and for non-symmetric data structure with lowly positive, Resampling, Rank, Micz03, Micz03+Noise, and Syn. To show that the conclusions obtained from our utilities are made appropriately, scatter plots and histograms of X_1 are displayed in Figures 2 to 10. As mentioned above, there are three replicates in masked data by the microaggregation method, but the scatter plots can not show overlapping. Therefore, We can not figure out how many data points are departed from original data by looking at scatter plot, and we can not judge correctly which one is better or worse by using scatter plots. The scatter plots of masked data by Rank say it distorts the original data a lot, but the histograms indicate that masked data by Rank is similarly distributed with original data. Our utilities conclude that the masked data by Rank are not much differently distributed from original data, which are consistent with the results of histograms. In most cases, graphical consequences support the results which are obtained from our utility measures. All measures such as plots and utilities have weaknesses and strengths, and one utility does not outperform others for any cases. Therefore, all possible ways should be considered to investigate data utilities of SDL methods.

Our findings obtained from these simulation studies are summarized as follows.

- CDF: It does not involve any parameters to carry out CDF utility. It tends to be favorable to Rank SDL method.
- Clustering: It does not measure the differences in original and masked data within a cluster. In general, it is consistent to overall results, but for non-symmetric cases, large number of clusters have a tendency to produce worse utility for the masked data by microaggregation.
- Propensity score with logistic model: The choice of degree is very crucial. Sometimes, even degree of three is not likely to be enough to model data, so it is difficult in dealing with high-dimensional data.
- Propensity score with tree model: Small size of tree can not distinguish utility of Rank from that of Resample, while large size of tree leads to bad utility for the microaggregation method. For some cases, large size of tree can not partition space for Rank method. It is also favorable to Rank SDL method.

- Propensity score with modified logistic model: It is the combination of cluster and propensity score utilities, so it possesses both advantages and disadvantages of logistic model and clustering. However, it is likely to be consistent to overall results for all data structures with $p = 1\%$.

Modified logistic model compromises logistic with clustering. It improves both logistic model and clustering in terms of choosing degree in logistic model and measuring within-cluster variation in cluster utility. When we consider linear and quadratic terms to calculate propensity score utility based on modified logistic model, Tables 2 to 9 show that there are no changes to make decision according to degree of logistic function for symmetric and non-symmetric data.

There are several ways of partitioning data space such as splitting data space evenly and using tree models, instead of using clustering. When we split data space evenly for each covariate, the number of partitions increase much faster as the number of dimension becomes larger. Suppose we partition the space into 10 for each variable and we cross 10 partitions by all variables to create groups. If we have p -dimensional data, there are 10^p partitions in the dataset. Also, it can create many cells which do not have any classified data points.

When we classify observations into groups by using tree model, responses as well as covariates are used as a prior information, while responses are not used as a information when we classify using clustering. We want to group similar observations with ignoring the sources of original and masked data when we classify, and then we take consider in responses to fit propensity scores to data. Besides, tree model can not sometimes partition the whole space as shown in the Tables 2 to 9.

Table 2: Data Utility for Symmetric Highly Negative Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.3204	0.1447	0.2746	0.0089	0.0058
	MCM	117.399	35.736	69.331	0.109	0.029
Clustering	$g = 500$	1318.57	397.393	739.922	140.075	45.616
	$g = 1000$	1448.20	588.027	876.646	247.219	86.533
Propensity	Logistic I	0.0000	27.5052	0.0031	16.1094	0.4074
	Logistic II	491.340	60.010	40.128	24.377	0.474
Score	Tree Model I	1247.756	250.822	632.424	0.000	0.000
Utility	Tree Model II	2161.905	2304.599	1668.949	155.936	46.543
	Modified Logistic I	1123.41	281.51	607.61	84.39	17.25
	Modified Logistic II	1292.81	305.40	678.33	93.06	17.44

Table 3: Data Utility for Symmetric Highly Positive Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.2340	0.0507	0.1568	0.009	0.008
	MCM	282.491	12.194	89.209	0.127	0.122
Clustering	$g = 500$	1525.89	359.629	877.46	154.386	39.378
	$g = 1000$	1651.50	555.973	1055.22	253.018	92.608
Propensity	Logistic I	0.000	16.0487	0.0020	9.9971	0.08135
	Logistic II	596.97	32.89	1.06	18.2919	0.1287
Score	Tree Model I	1445.943	204.192	843.333	0.000	0.000
Utility	Tree Model II	2306.393	2319.399	1889.630	220.581	163.555
	Modified Logistic I	1202.40	128.14	508.89	90.42	17.76
	Modified Logistic II	1527.14	264.43	765.81	102.85	23.44

Table 4: Data Utility for Symmetric Lowly Negative Ddata

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.2942	0.112	0.242	0.005	0.007
	MCM	188.743	46.568	114.844	0.0315	0.0754
Clustering	$g = 500$	1310.23	379.188	732.766	127.278	66.931
	$g = 1000$	1433.15	560.940	852.404	232.161	119.079
Propensity	Logistic I	0.000	22.7232	0.0093	1.7916	0.9926
	Logistic II	475.749	40.499	28.336	22.3447	1.4391
Score	Tree Model I	1265.156	255.322	610.325	0.000	0.000
Utility	Tree Model II	2126.655	2222.952	1746.085	55.597	68.890
	Modified Logistic I	935.283	233.397	575.274	59.989	25.371
	Modified Logistic II	1299.38	279.26	666.65	66.82	27.21

Table 5: Data Utility for Symmetric Lowly Positive Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.2463	0.0701	0.1899	0.005	0.005
	MCM	238.679	27.559	110.506	0.026	0.064
Clustering	$g = 500$	1388.38	357.074	695.288	125.322	42.748
	$g = 1000$	1512.55	555.627	866.178	230.854	86.052
Propensity	Logistic I	0.000	36.529	0.0045	1.3335	0.1478
	Logistic II	465.330	40.437	1.611	16.855	0.5036
Score	Tree Model I	1340.6174	228.463	614.056	0.000	0.000
Utility	Tree Model II	2221.793	2260.441	1778.514	156.984	79.011
	Modified Logistic I	1032.65	138.48	989.89	62.31	13.36
	Modified Logistic II	1417.31	168.28	1151.85	69.37	20.04

Table 6: Data Utility for Non-symmetric Highly Negative Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.1782	0.1657	0.1394	0.0113	0.0107
	MCM	61.908	64.404	44.428	0.095	0.2007
Clustering	$g = 500$	690.760	702.619	533.736	139.408	20.729
	$g = 1000$	796.510	856.033	642.149	237.940	53.596
Propensity	Logistic I	0.000	247.869	0.0965	44.153	0.525
	Logistic II	506.15	478.16	120.47	68.239	0.772
Score	Tree Model I	644.525	566.008	443.068	0.000	0.000
Utility	Tree Model II	1704.461	2359.788	1524.031	0.000	51.560
	Modified Logistic I	666.930	639.419	490.235	106.099	25.571
	Modified Logistic II	702.504	675.245	532.743	119.678	33.817

Table 7: Data Utility for Non-symmetric Highly Positive Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.1367	0.0642	0.0276	0.0105	0.0075
	MCM	85.433	20.002	1.986	0.0834	0.1740
Clustering	$g = 500$	710.982	671.061	162.902	129.157	26.189
	$g = 1000$	811.539	846.778	288.807	231.317	42.3527
Propensity	Logistic I	0.000	314.734	0.0045	35.294	0.105
	Logistic II	511.651	398.395	0.385	54.939	0.6045
Score	Tree Model I	653.244	565.869	57.574	0.000	0.000
Utility	Tree Model II	1541.373	2420.578	872.952	0.000	110.440
	Modified Logistic I	660.733	592.077	116.057	82.957	19.071
	Modified Logistic II	703.541	676.615	154.467	100.005	30.457

Table 8: Data Utility for Non-symmetric Lowly Negative Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.1440	0.1420	0.1031	0.0060	0.0110
	MCM	61.888	53.830	33.011	0.029	0.2105
Clustering	$g = 500$	677.951	698.289	441.349	103.477	15.233
	$g = 1000$	779.863	888.560	558.011	218.239	31.263
Propensity	Logistic I	0.000	166.989	0.0245	8.003	0.403
	Logistic II	475.817	396.989	6.616	15.933	1.001
Score	Tree Model I	631.762	585.561	367.839	0.000	0.000
Utility	Tree Model II	1768.438	2345.072	1552.975	0.000	230.3188
	Modified Logistic I	647.683	639.038	382.881	53.324	19.183
	Modified Logistic II	690.294	685.234	449.454	73.587	25.613

Table 9: Data Utility for Non-symmetric Lowly Positive Data

		SDL method				
		Syn	Micz03	Micz03+Noise	Rank	Resampling
CDF	MD	0.3540	0.1282	0.3193	0.0045	0.0086
	MCM	559.005	29.374	387.413	0.0263	0.2318
Clustering	$g = 500$	2913.04	661.19	2014.19	112.842	65.974
	$g = 1000$	3153.95	1097.21	2569.32	233.712	171.051
Propensity	Logistic I	0.000	63.434	0.009	6.558	0.374
	Logistic II	2132.36	95.205	205.10	7.038	0.431
Score	Tree Model I	3267.603	1151.447	2721.885	0.000	0.000
Utility	Tree Model II	3665.534	2784.064	3306.998	0.000	190.083
	Modified Logistic I	2413.24	463.78	1275.19	37.27	20.99
	Modified Logistic II	2657.47	721.77	1733.46	50.99	32.64

4 Summary and Conclusions

Approaches based on CDF, clustering and propensity scores are used to measure the utility of SDL method, when the distribution is not postulated to be multivariate normal. The cluster utility can be regarded to as propensity score utility since the proportion of original dataset is considered as the estimated propensity score. Estimation of propensity score is very important to perform propensity score utilities, and it can be carried out using various modeling. The choice of degree in polynomial of logistic model undoubtedly has an impact on the propensity score utility, whereas the selection of the number of partitions are critical in using clustering, tree model, and modified logistic model. When samples are drawn from the variety of data structure, the simulations show that CDF utility and propensity score utility with tree model tend to yield good utility of Rank SDL method, and the large number of partitions in cluster and tree model is against microaggregation SDL method. Also, they show that in general, propensity score utility using modified logistic model with $p = 1\%$ conforms overall results. Our utilities can be performed for any data where normality assumption is not known to hold. However, when we focus on the differences in results of some specific analysis, model-based utilities should be considered to measure data utility.

References

- Aitkin, M., and Wilson, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325-331.
- DeGroot and Schervish (2002), *Probability and Statistics*, Addison Wesley.
- Dobra, A., Fienberg, S. E., Karr, A. F., and Sanil, A. P. (2002), "Software systems for tabular data releases," *Int. J. Uncertainty, Fussiness and knowledge Based Systems*, 10(5), 529-544.
- Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. (2001), "Comparing SDL methods for microdata on the basis of information loss and disclosure risk," *In Proc ETK-NTTS 2001*, pages 807-825, Luxembourg. Eurostat.
- Duncan, G. T. and Lambert, D. (1986). "Disclosure-limited data dissemination," *Journal of the American Statistical Association*," 81,10-28.
- Duncan, G. T. and Lambert, D. (1989), "The risk of disclosure for microdata," *Journal of Business and Economic Statistics*, 7, 207-217.
- Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997), "A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data," *Journal of Official Statistics*, 13, 75-89.
- Gibbons, J.D. and Chakraborti, S. (1992), "Nonparametric Statistical Inference," Third Edition, New York: Marcel Dekker, Inc.
- Gomantam, S., Karr, A. F., and Sanil, A. P. (2005), "Data swapping as a decision problem," *Journal of Official Statistics*, To appear, available on-line at WWW.niss.org/dgii/technicalreports.html.
- Gordon, A. D. (1999), "Classification," London: Chapman & Hall, / CRC.
- Hollander, M. and Wolfe, D.A. (1973), "Nonparametric Statistical Methods," New York: John Wiley & Sons, Inc.
- Karr, A. F., Kohonen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P., "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality," available on-line at WWW.niss.org/dgii/techreports.html.
- Kaufman, L. and Rouseeuw, P. J. (1990) "Finding Groups in Data. An introduction to Cluster Analysis," New York: John Wiley and Sons.
- Kim, P.J. and Jennrich, R.I. (1970), "Tables of the Exact Sampling Distribution of the Two-sample Kolmogorov-Smirnov Criterion," Selected Tables in Mathematical Statistics (Harter and Owen, eds.), Chicago: Markham Publishing Co.
- Lehmann, E.L. (1975), "Nonparametrics: Statistical Methods Based on Ranks," San Francisco: Holden-Day.
- Lamber, D. (1993), "Measures of disclosure risk and harm," *Journal of Official Statistics*, 9, 313-331.
- Oganian, A. (2003), "Security and Information Loss in Statistical Database Protection," PhD theses, University Politècnica de Catalunya.
- Reiter, J. P. (2005), "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study," *Journal of Royal Statistical Society, Series*

- A,” 168, 185-205.
- Rosenbaum, P. R. and Rubin, D. B (1983), “The Central Role of the propensity score in observational studies for Causal Effects,” *Biometrika*, 70, 41-55.
- Skinner, C. J. and Ellior, M. J. (2002), “A measure of disclosure risk for microdata,” *Jornal of the Royal Statistical Society, Series B*,” 64, 855-867.
- Wallman, K. K. and Harris-Kojetin, B. A. (2004), “Implementing the confidential information protection and statistical efficiency act of 2002,” *Chance*, 17(3),21-25.
- Venables, W. N. and Ripley, B. D. (2002), “Modern Applied Statistics with S,” *Springer-Verlag*, New York.
- Willenborg, L. C. R. J. and de Waal, T. (2001), “Elements of Statistical Disclosure Control,” *Springer-Verlag*, New York.
- Yancey, W. E., Winkler, W. E., and Creecy, R. H. (2002), “Disclosure risk assessment in perturbative microdata protection,” In Domingo-Ferrer, J. editor, *Inference Control in Statistical Databases*, pages 135-152, Berlin: Springer-Verlag.

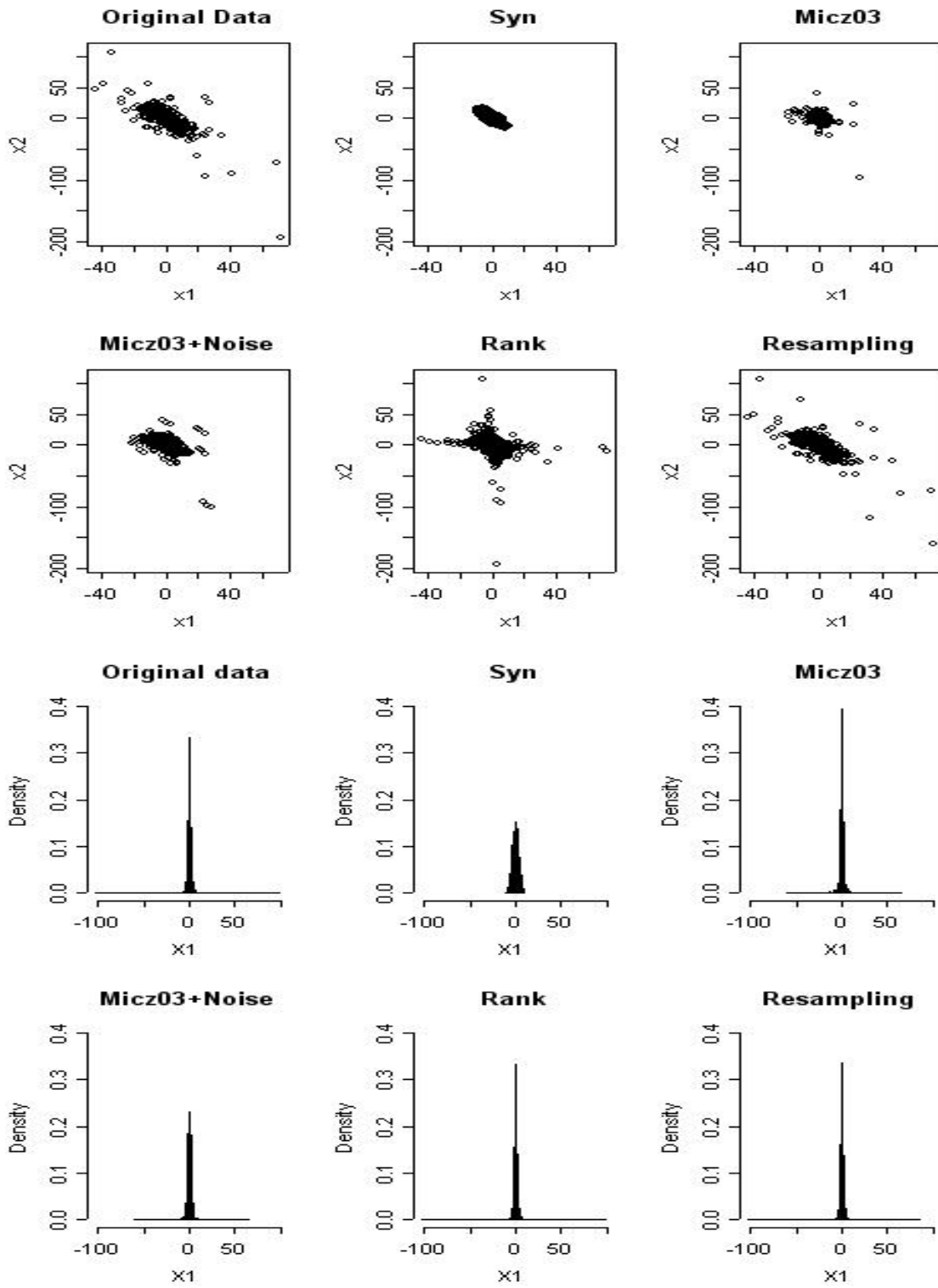


Figure 2: Scatter plots and histograms of X_1 for symmetric high negative data.

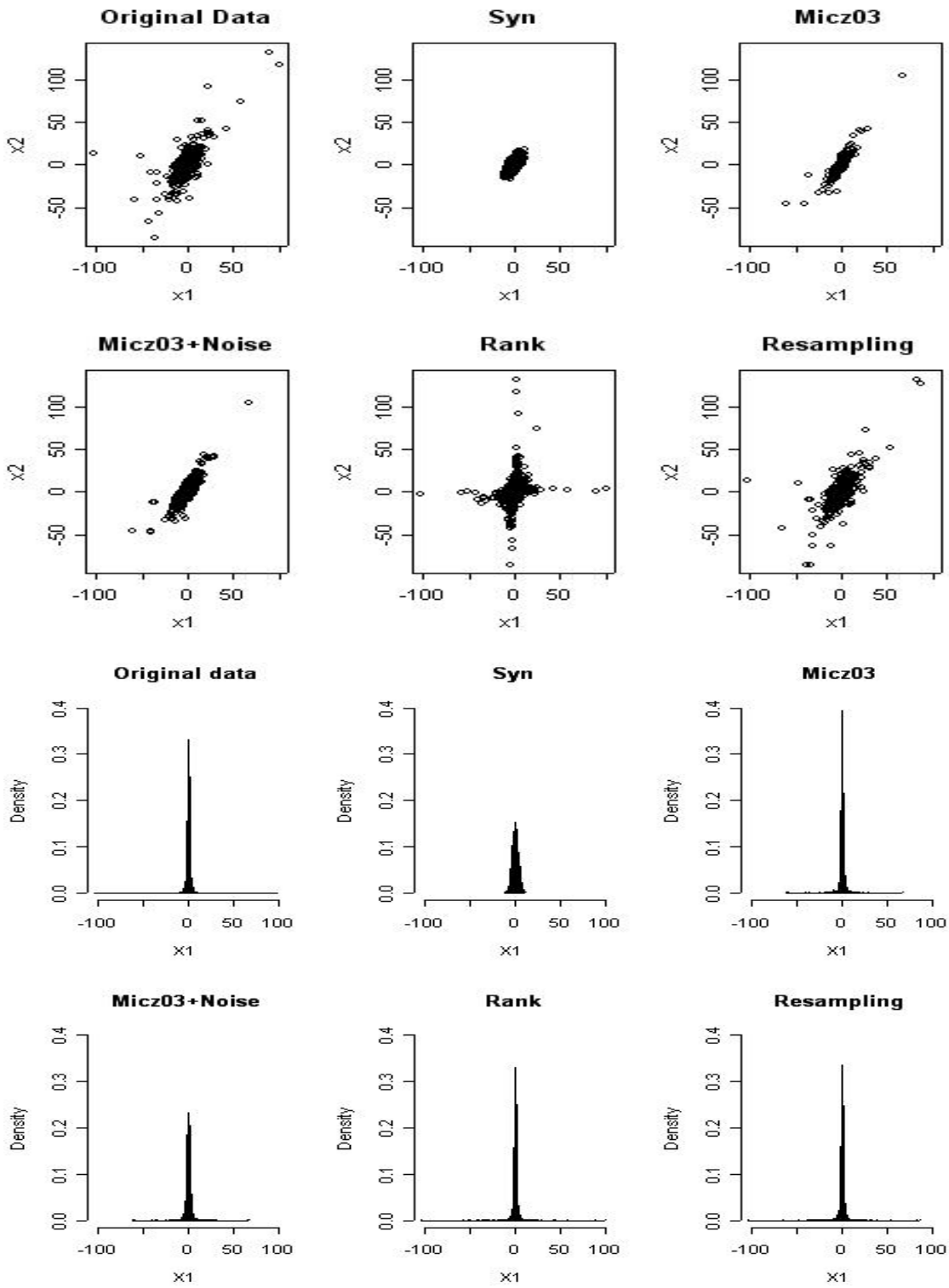


Figure 3: Scatter plots and histograms of X_1 for symmetric high positive data.

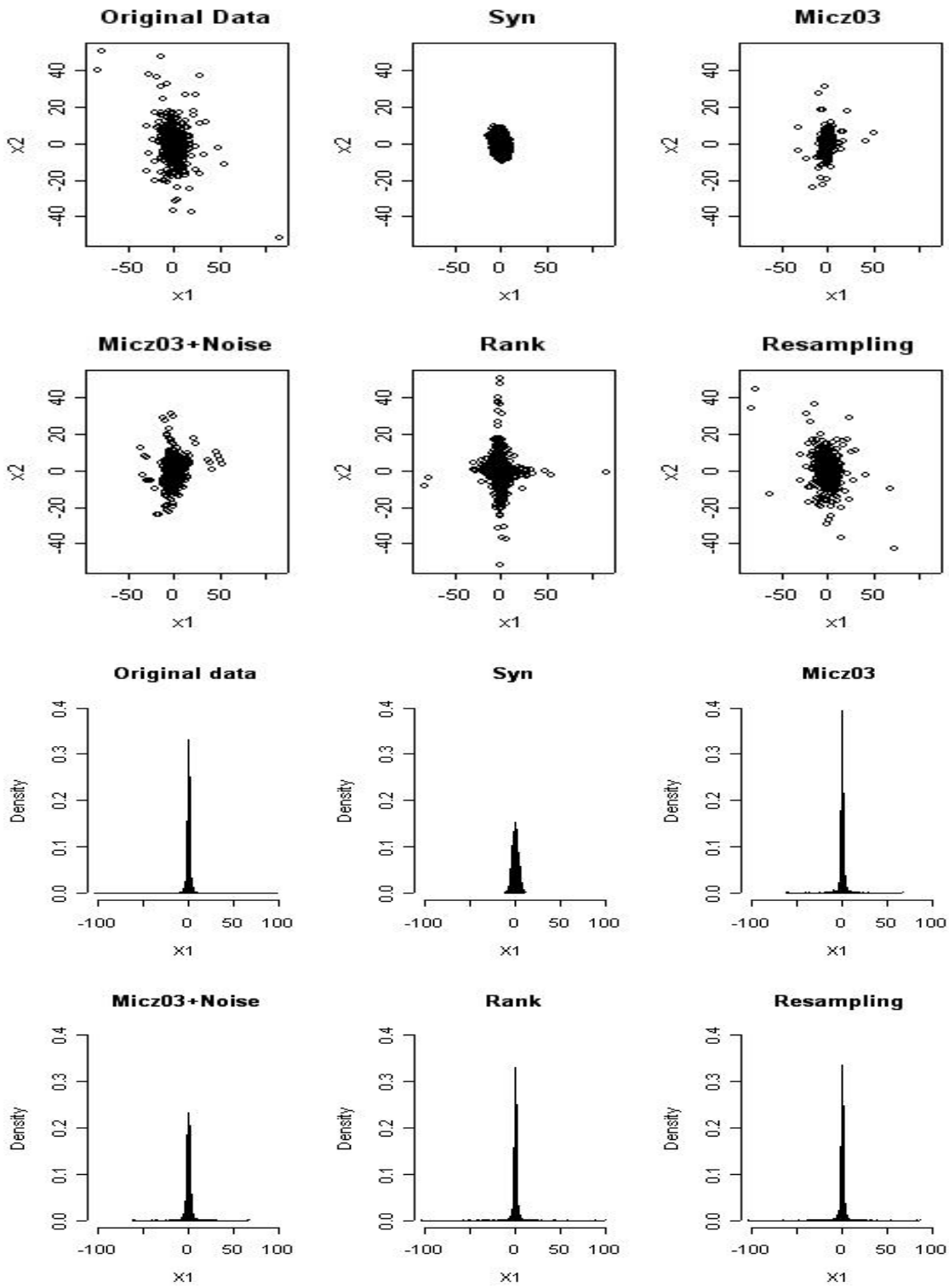


Figure 4: Scatter plots and histograms of X_1 for symmetric low negative data.

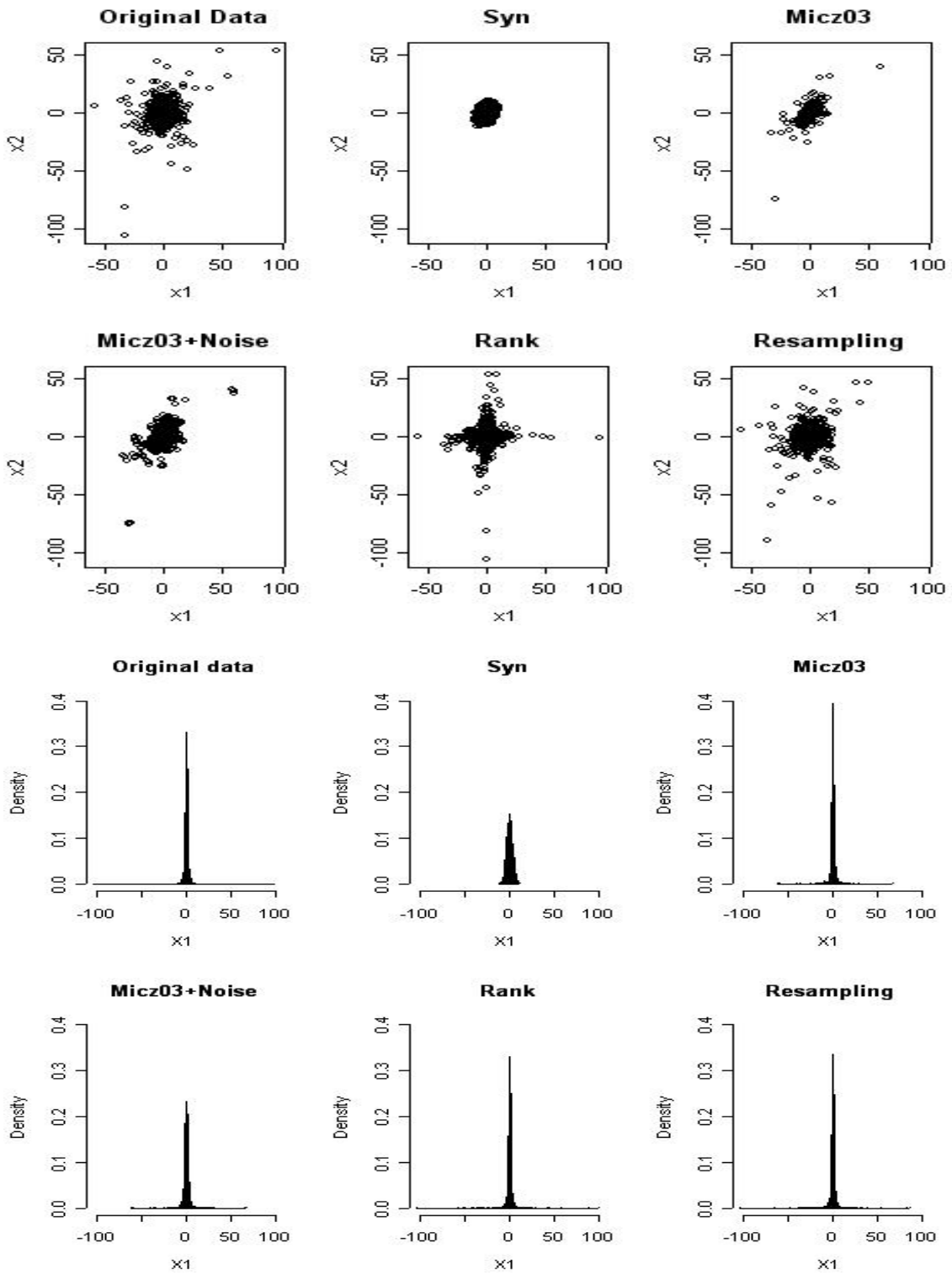


Figure 5: Scatter plots and histograms of X_1 for symmetric low positive data.

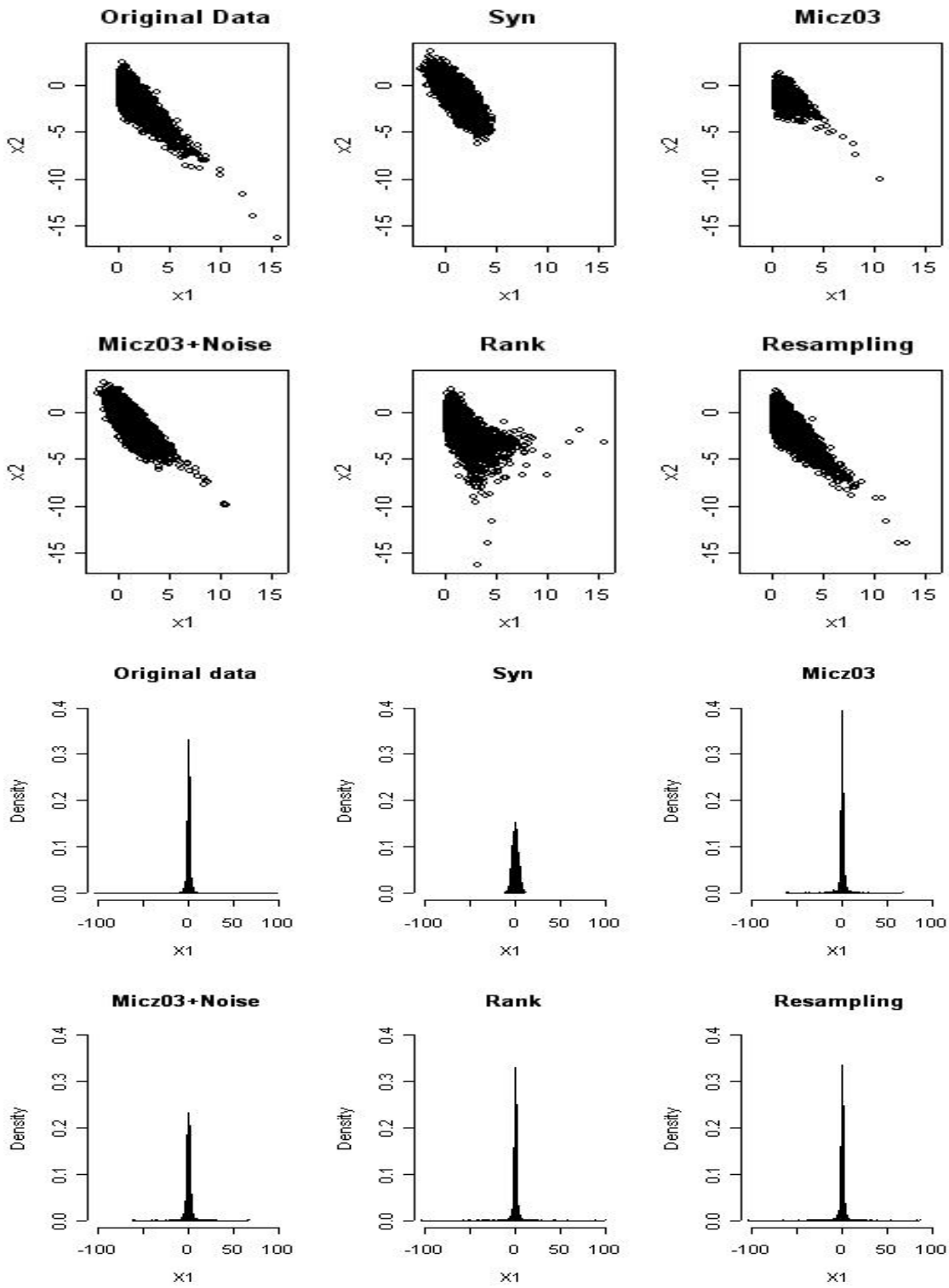


Figure 6: Scatter plots and histograms of X_1 for non-symmetric high negative data.

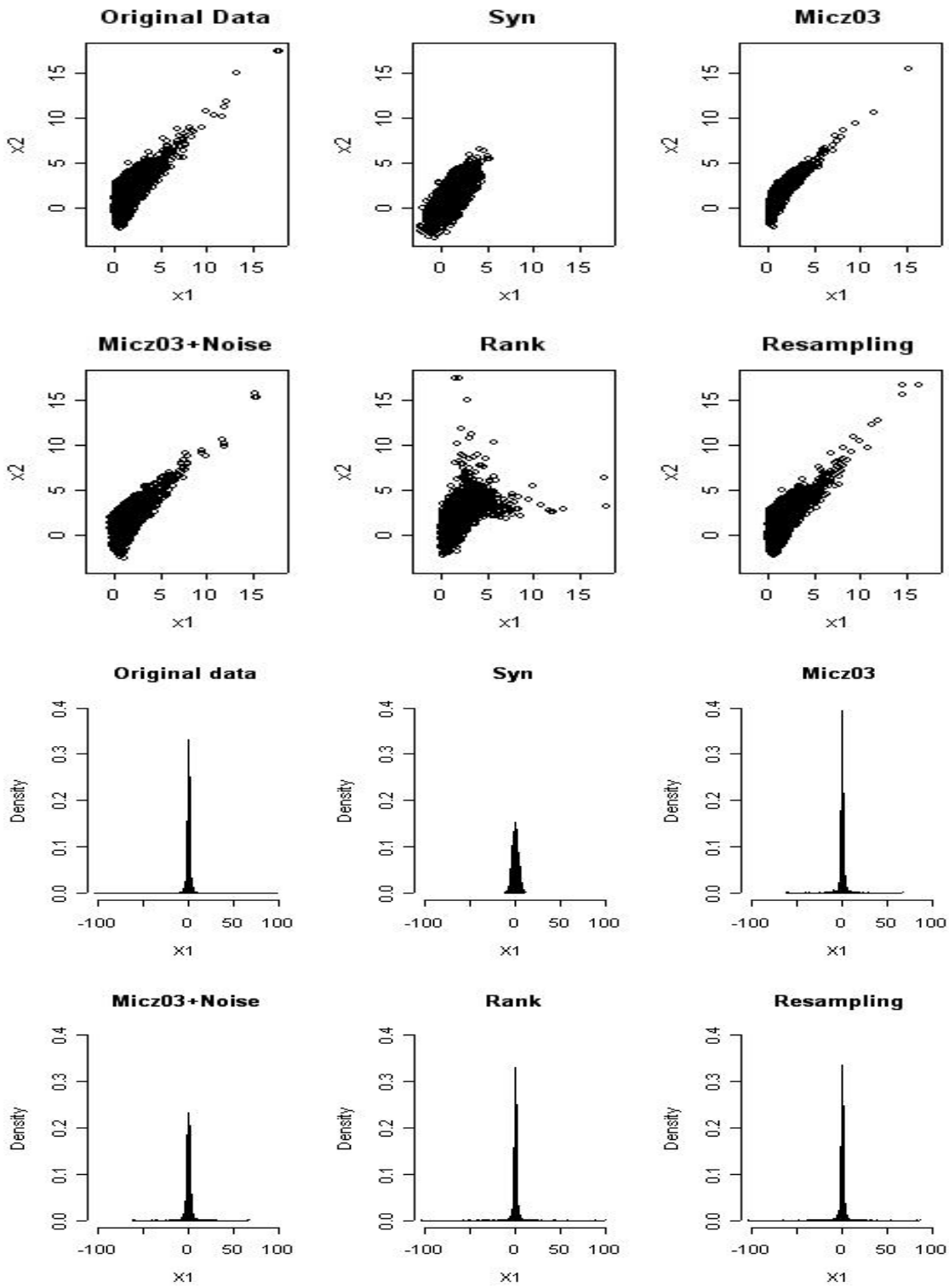


Figure 7: Scatter plots and histograms of X_1 for non-symmetric high positive data.

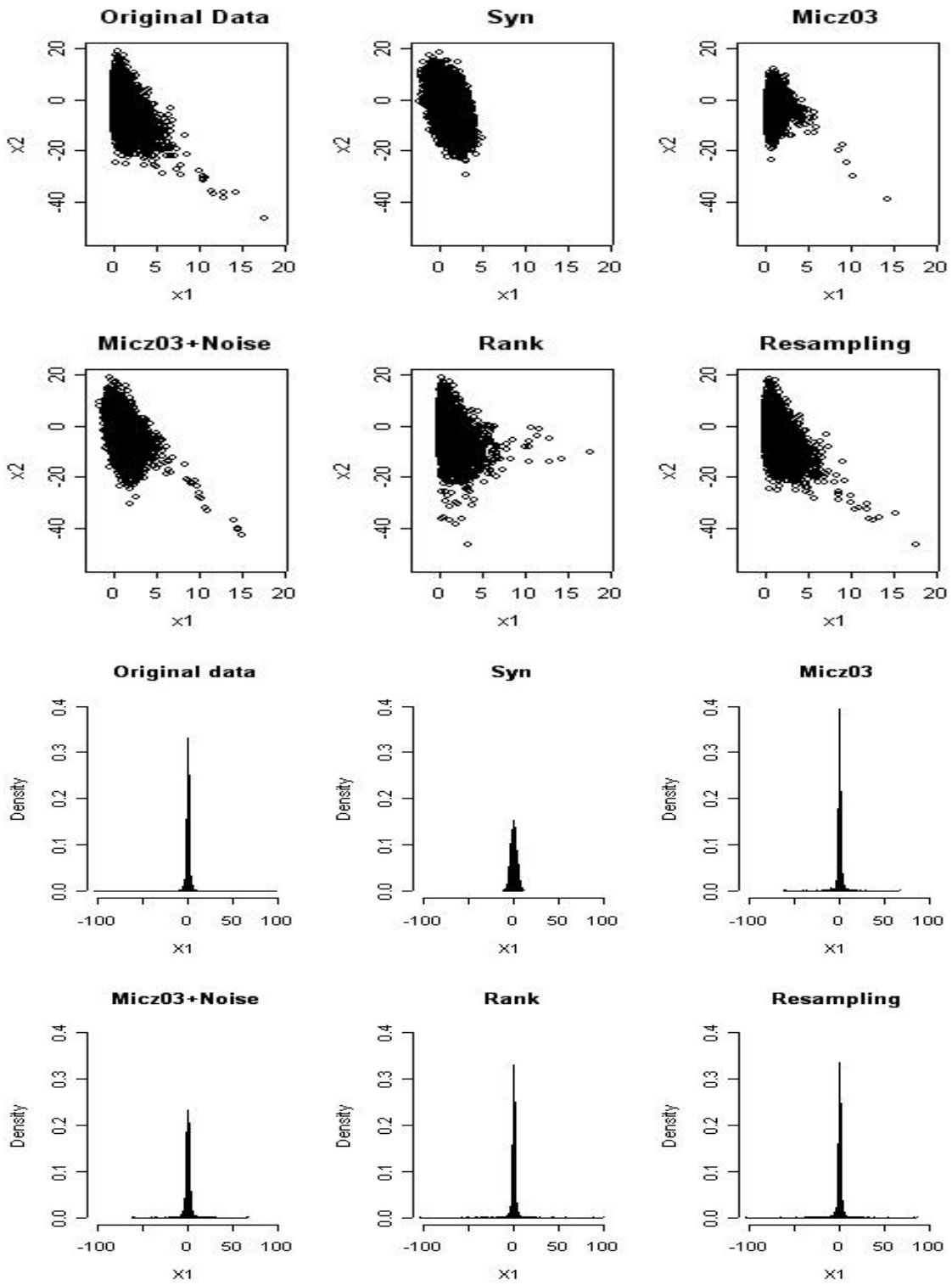


Figure 8: Scatter plots and histograms of X_1 for non-symmetric low negative data.

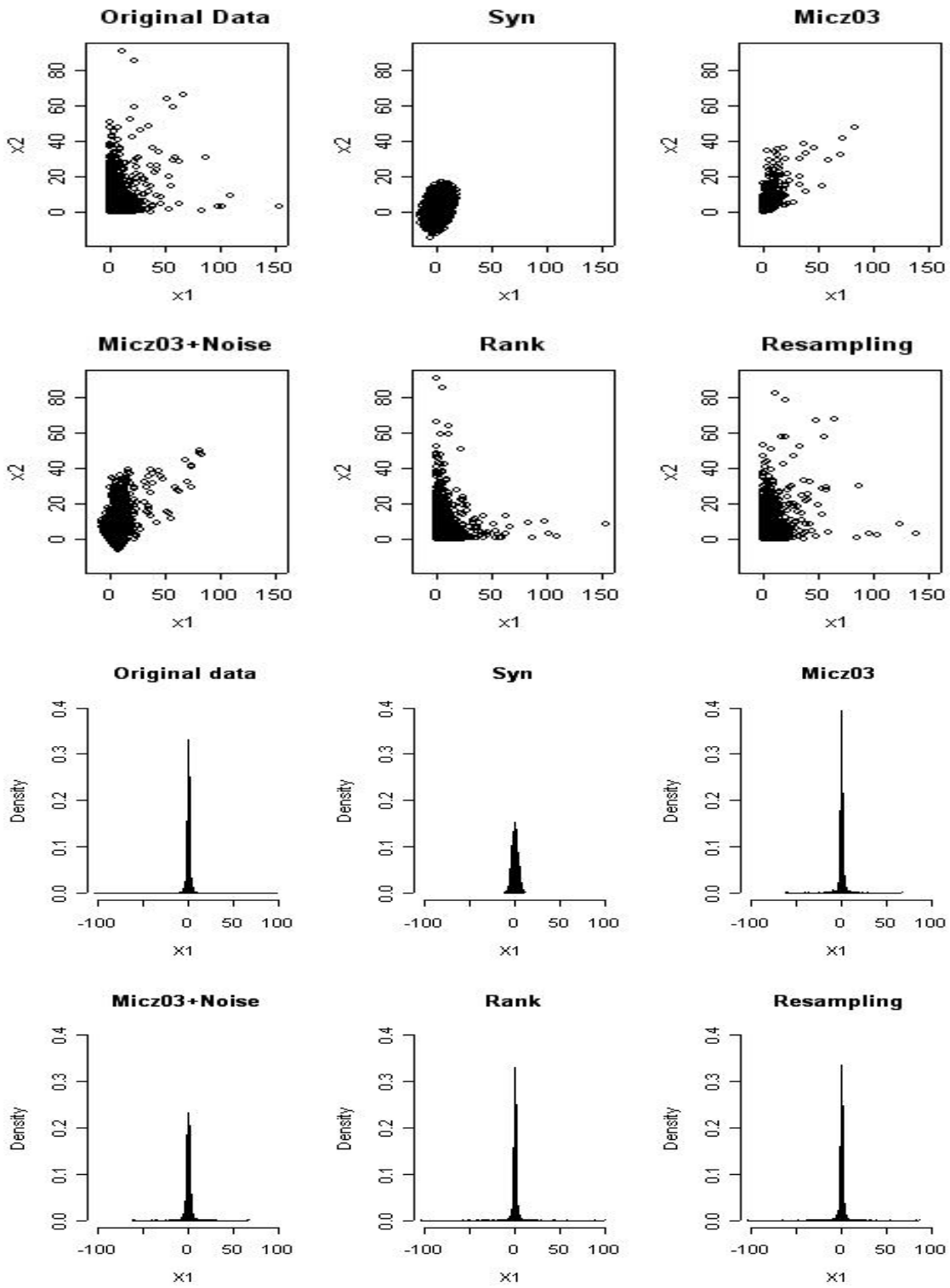


Figure 9: Scatter plots and histograms of X_1 for non-symmetric low positive data.