Obfuscating Continuous Data

Group Project

Diganta Bhattacharya
Muntasir Sheikh
Archi De

26 August 2020

Introduction

Many real-life data-sets like income data, medical data need to be secured before making it public. However, security comes at the cost of losing some useful statistical information about the data-set. Data obfuscation deals with this problem of masking a data-set in such a way that

- the utility of the data is maximized.
- the risk of the disclosure of sensitive information is minimised.

Aim

In this project, we have tried to obfuscate a data set with two attributes recorded per individual, both of which are sensitive data and continuous. We want to modify this dataset so that

- the values of this attributes for a particular individual is not disclosed.
- the important statistical information about the distribution of the data, such as, mean, moments of both the attributes, quantiles of the distributions can be estimated.
- The correlation between the two attributes can be estimated.

The Data

We will work with some large data set with 2 variables, corresponding to n individuals. Both these variables are numeric and sensitive and need to be protected. Let $\{X_i, 1 \le i \le n\}$ and $\{U_i, 1 \le i \le n\}$ be the data values corresponding to the variables. We assume

- (X_i, U_i) are i.i.d for $i = 1, 2, \ldots, n$.
- Both the variables are coming from some unknown continuous distribution.
- All the moments of both these distribution are finite.
- The correlation of these two variables is ρ.

,

Masking the data

The masked data (Z, W) is given by the following steps-

1. We simulate $\{B_i, 1 \leq i \leq n\}$, which are iid from $Binomial(1, p)$ and independent of $\{X_i, 1 \leq i \leq n\}$ and $\{U_i, 1 \leq i \leq n\}$.
2. Then, for $1 \leq i \leq n$
   - If $B_i = 1$, we draw a random number j uniformly from $\{1, \ldots, i-1, i+1, \ldots, n\}$ and set
   $$Z_i = X_j$$
   $$W_i = U_j$$
   - If $B_i = 0$,
   $$Z_i = X_i + Y_i$$
   $$W_i = U_i + V_i$$
   such that $(Y, V) \sim N(0, 0, \sigma_Y^2, \sigma_V^2, \rho_{YV})$ independent of X and U.

Disclosure risk

The disclosure risk for any estimator τ for X_i can be measured by,

$$P(|\tau - X_i| < d),\ d > 0$$

i.e., the probability that X_i lies within a d-boundary of its estimator.
For S simulations, an estimate of risk is given by,

$$\frac{\sum_{s=1}^{S} I(\tau_s \in (X_i - d, X_i + d))}{S}$$

where τ_s is the estimate of X_i for the sth simulation and $I(.)$ is the indicator function.

Estimation of Raw Moments

Theorem: If $\{X_i, 1 \leq i \leq n\}$ is assumed to be an i.i.d. sample from some unknown distribution function $G(x)$ (G is a continuous function) with finite absolute raw moments, i.e., $E(|X_i|^k) < \infty \forall k \in N$ and $\{Z_i, 1 \leq i \leq n\}$ is obtained using above method, then an unbiased estimator for the kth raw moment of $X, (X \sim G(.), k \in N)$ is obtained from the recursion relation given by $\hat{\mu}_{(X,k)}$

$$= \hat{\mu}_{(Z,k)} - (1-p).(\mu_{(Y,k)} + \binom{k}{1}\mu_{(Y,k-1)}\hat{\mu}_{(X,1)} + \cdots + \binom{k}{k-1}\mu_{(Y,1)}\hat{\mu}_{(X,k-1)})$$

where,
$\hat{\mu}_{(X,1)} = \bar{Z},\ \hat{\mu}_{(Z,k)} = \frac{1}{n}\sum_{j=1}^{n} Z_j^k$ and $\mu_{(Y,k)} = kth\ raw\ moment\ of\ Y.$

Specifically, we have,

- Unbiased estimate of $\mu_X = \bar{Z}$.
- Unbiased estimate of $\mu_U = \bar{W}$.
- Also,

$$\hat{\mu}_{(X,2)} = \hat{\mu}_{(Z,2)} - (1-p).(\mu_{(Y,2)} + \binom{2}{1}\mu_{(Y,1)}\hat{\mu}_{(X,1)})$$

$$= \hat{\mu}_{(Z,2)} - (1-p).\sigma_Y^2$$

  Hence, unbiased estimate for $Var(X)$ is $\hat{S}_X^2 = \hat{S}_Z^2 - (1-p)\sigma_Y^2$.
- Similarly, unbiased estimate for $Var(U)$ is $\hat{S}_U^2 = \hat{S}_W^2 - (1-p)\sigma_V^2$

Estimation of quantiles

Theorem: In the given method, if $G(x)$ is the cdf of X, for $p > 0.5$,

$$T_1(x) = \frac{1}{np} \sum_{j=1}^{n} \sum_{t=0}^{\infty} \lambda^t . \Phi_{\sigma\sqrt{t}}(x - Z_j)$$

is an unbiased estimator for $G(x)\ \forall x \in R$, where $\lambda = -\frac{1-p}{p}$, and $\Phi_m(i)$ is the cumulative distribution function of a normal variable at i with mean 0 and standard deviation m for $m > 0$, and for $m = 0$, $\Phi_0(i) = I(i > 0)$ where $I(.)$ is the indicator function.

Theorem: In the given method, if $G(x)$ is the cdf of X, for $p > 0.5$,

$$T_b(x) = \frac{1}{np} \sum_{j=1}^{n} \sum_{t=0}^{\infty} \lambda^t . \Phi_{b_t}(x - Z_j)$$

is another estimator for $G(x),\ x \in R$, where $\lambda = -\frac{1-p}{p}$, $b_t = \sqrt{tb^2 + \sigma^2}$ and $\Phi_m(i)$ is the cumulative distribution function of a normal variable at i with mean 0 and standard deviation m.

Covariance of masked variables

Note that

$$Cov(Z, W) = E(ZW) - E(Z)E(W)$$

$$= pE(ZW|B = 1) + (1 - p).E(ZW|B = 0) - E(Z)E(W)$$

$$= pE(XU) + (1 - p).E((X + Y)(U + V)) - E(X)E(U)$$

$$= pE(XU)+(1-p).[E(XU)+E(X)E(V)+E(Y)E(U)+E(YV)]-E(X)E(U)$$

$$= \rho\sigma_X\sigma_U + (1 - p).\rho_{YV}\sigma_Y\sigma_V$$

Estimating correlation

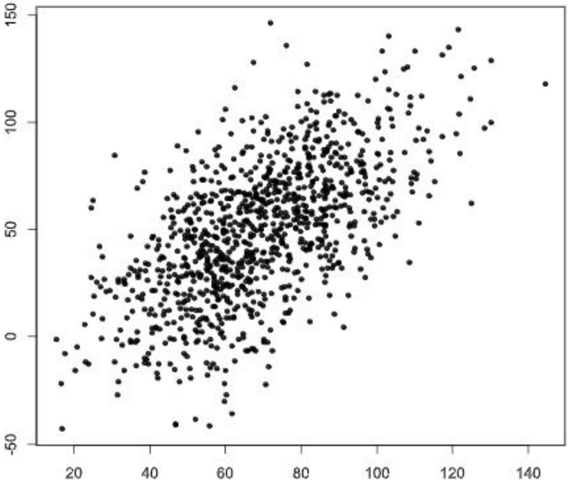
From the above we can see that, an estimate of $\rho = Corr(X, U)$ can be given by

$$\hat{\rho} = \frac{\hat{Cov}(Z, W) - (1 - p).\rho_{YV}\sigma_Y\sigma_V}{\sqrt{\hat{Var}(X)\hat{Var}(U)}}$$

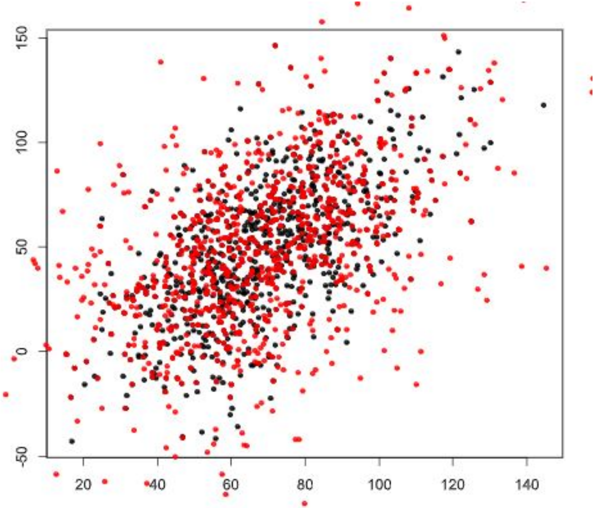
Note that we already have a form for $\hat{Var}(X)$ and $\hat{Var}(U)$.

First simulation

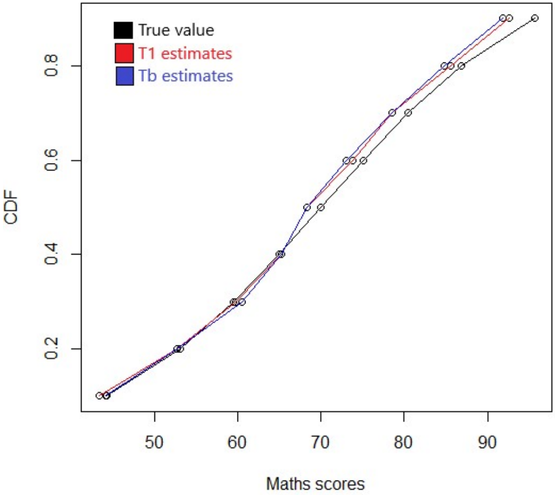
We simulated a data for marks of students in two subjects, say, maths and physics to test the masking procedure.
Here, $(X_i, U_i), i = 1, \ldots, 1000$ is from BVN.



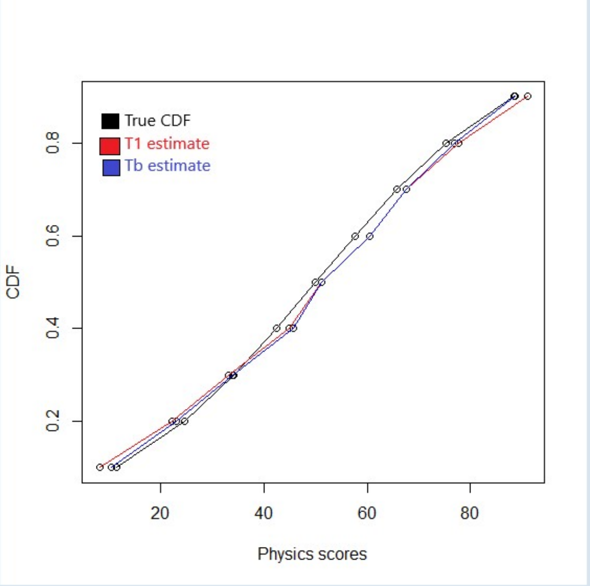
We mask the dataset with $p = 0.6$. The masked data appears as below



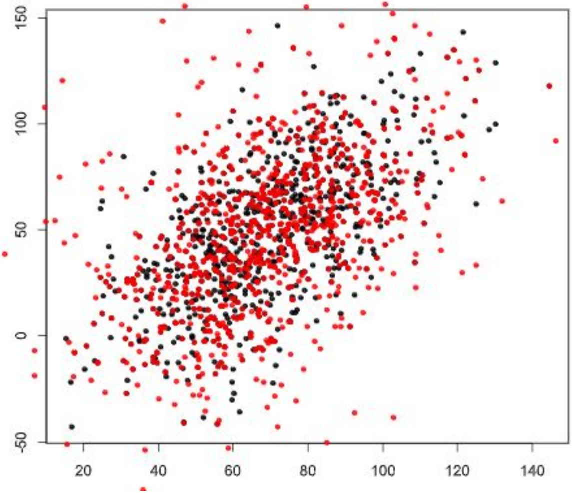
The CDF for the first subject is estimated as given



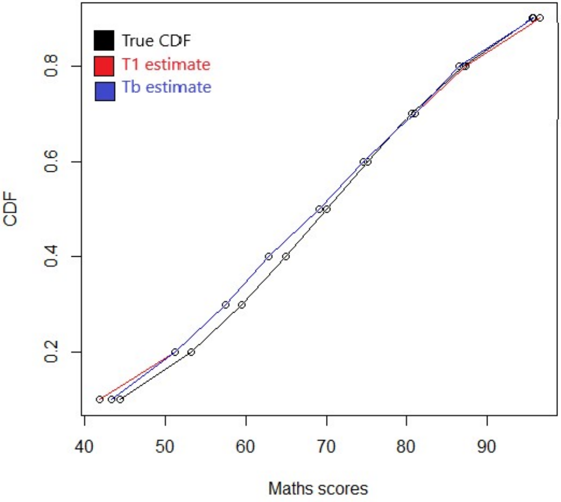
The CDF for the second subject is estimated as given



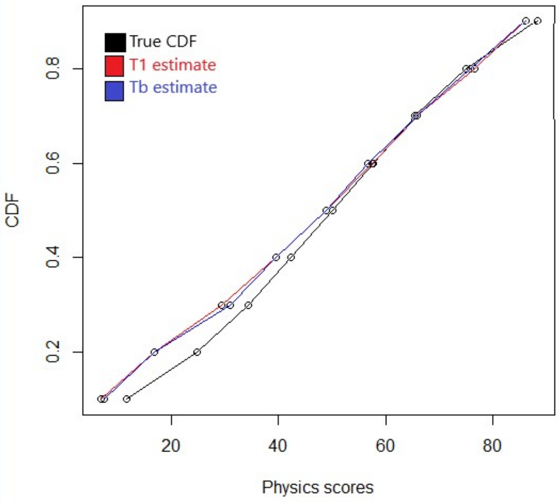
We mask the dataset with $p = 0.7$. The masked data appears as below



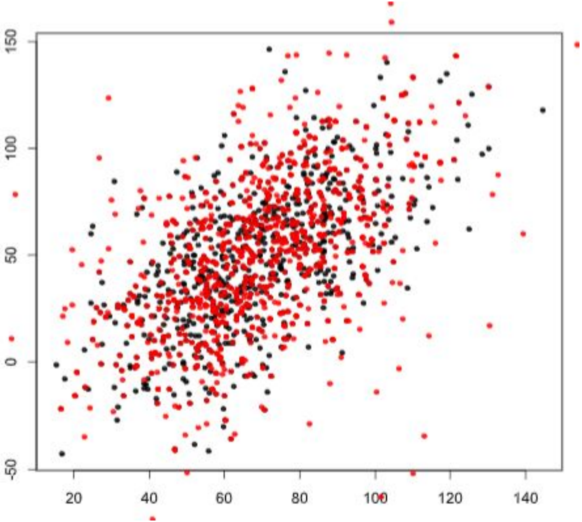
The CDF for the first subject is estimated as given



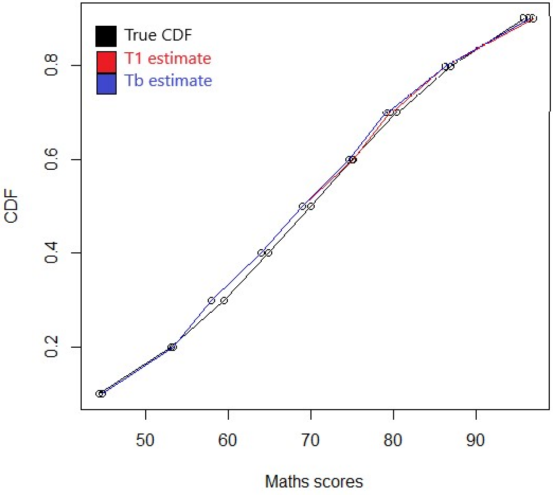
The CDF for the second subject is estimated as given



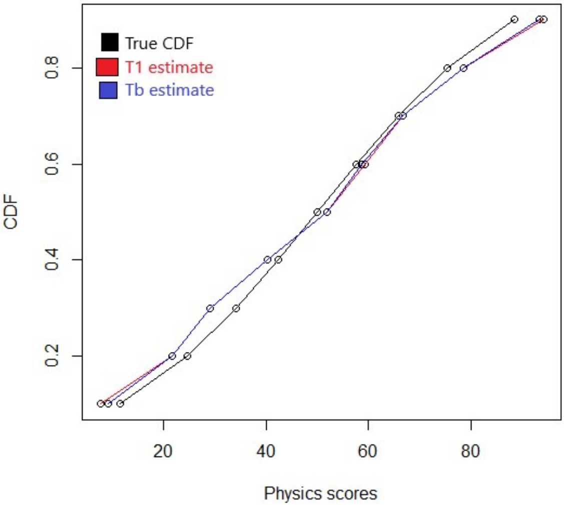
We mask the dataset with $p = 0.8$. The masked data appears as below



The CDF for the first subject is estimated as given



The CDF for the second subject is estimated as given



	True Value	$p = 0.6$	$p = 0.7$	$p = 0.8$
$\hat{\mu}_X$	70	68.6672	68.9667	69.4946
$\hat{\mu}_U$	50	50.6042	48.0966	50.1882
$\hat{\sigma}_X$	20	19.1048	20.5200	19.9102
$\hat{\sigma}_U$	30	32.2142	31.1282	33.4273
$\hat{\rho}$	0.6	0.619	0.638	0.6250
Disclosure risk				
$d = 5$	Maths	0.166	0.162	0.158
$d = 5$	Physics	0.108	0.103	0.100

Table: Estimations

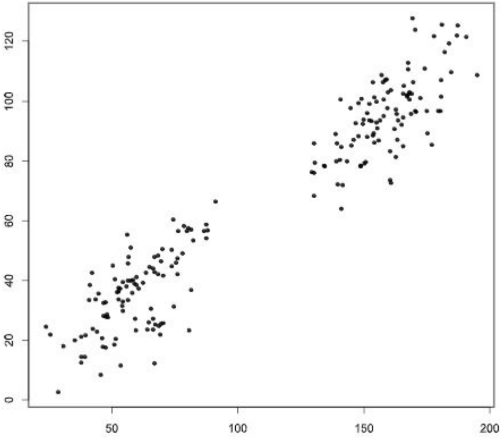
Comments

- The estimation works fairly well for all three values of p.
- For high values of n, the estimates by $T_1(x)$ and $T_b(x)$ almost coincide.
- The mean and s.d estimations are precise for all three.
- The correlation of the two variables was estimated to a value close to the true value after masking in all three cases.

The First Limitation

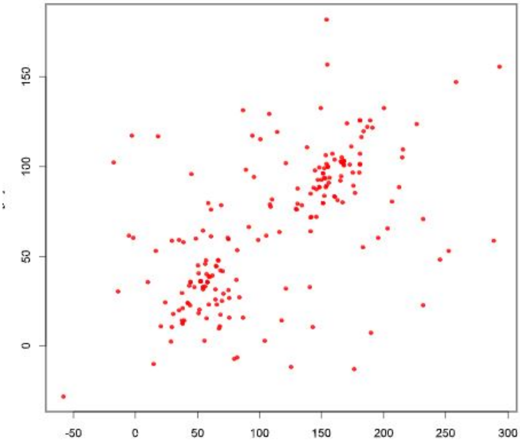
- Most of the data that we found regarding marks and salaries were a bit clustered.
- Now if we follow our masking procedure as before it is probable that we might swap two values from different clusters thus eventually resulting in a completely differently shaped masked data.
- If the data has such cluster then the variance of the data may appear high although individual clusters have lower variance. In this case, adding error with variance equally high will lead to losing the original properties of the data.
- We might lose some of the properties of the original data-set.

The First Limitation(cont)



This is the original data.

The First Limitation(cont)



Here the original data was clustered but the masked data lost this property.

Intuition for the tweak

We will try to prepare the masked data (Z, W) now, by first regrouping the data into clusters such that in each cluster the data points are more or less close. Now in each of those clusters we apply the previous masking algorithm.

We estimate that since actual marks distribution of a group of students would have clusters in the data, so the masking should become a bit more precise in this case.

How do we form the clusters?

We will use the K-means clustering algorithm and we will define it briefly,

- The main idea is to define k centers, one for each cluster.Let $X = x_1, x_2, x_3, \ldots\ldots.., x_n$ be the set of data points and $V = v_1, v_2, \ldots\ldots., v_c$ be the set of centers. At first randomly select c cluster centers.
- Then we calculate the distance between each data point and cluster centers. Now we assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
- Now we recalculate the new cluster center using: $v_i = (\frac{1}{c_i}) \sum_{j=1}^{c_i} x_j$ where c_i represents the number of data points in i^{th} cluster.

How do we form the clusters?(cotd.)

- Now we again recalculate the distance between each data point and new obtained cluster centers. We continue till the data points stop reassigning.
- Essentially we are minimizing $J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2$. Where,
  1. $||x_i - v_j||$ is the Euclidean distance between $x_i and v_j$.
  2. c_i is the number of data points in i^{th} cluster.
  3. c is the number of cluster centers.

A small tweak to the Algorithm

In order to avoid the first limitation we can device a small tweak. If we first cluster the data suitably using the K-means algorithm, and then if we apply the masking algortihm we can minimize the error that is described in step 2.

1. That is, when we had $B_i = 1$, we drew a random number j uniformly from $1, \ldots, n$ excluding i and set

$$Z_i = X_j$$

$$W_i = U_j$$

2. Now the problem with this step in clustered data is that we might swap the values of two different clusters and eventually destroying the pattern of the data.
3. If we first cluster the data, and then within each cluster we restrict the swapping, we can preserve the original shape of the data.

A small tweak to the Algorithm(cotd.)

4. So now if we have $B_i = 1$, we draw a random number j uniformly from $(n_j, \ldots, n_k)$ excluding i and set

$$Z_i = X_j$$

$$W_i = U_j$$

where, $(n_j, \ldots, n_k)$ denotes the indices of the points in the cluster where i^{th} point belongs.

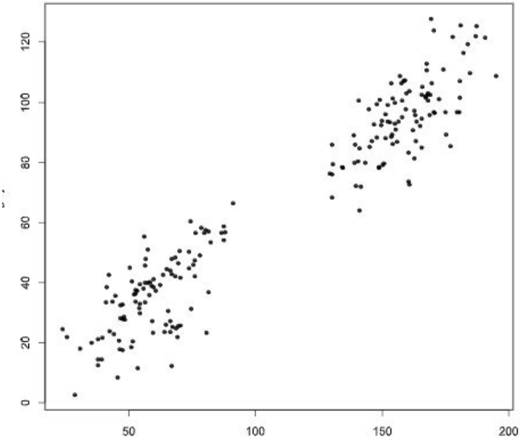
5. The remaining procedure remains the same , with

$$\sigma_X^2 = \sigma_Z^2 - (1-p).\sigma_Y^2$$

where Z was the masked data and Y was the error added.

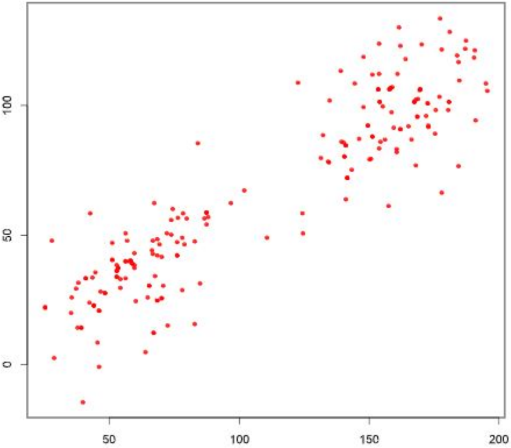
Simulations and results after the tweak

We take the following data for testing this method.



Simulations and results after the tweak

The masked data is shown below.



Simulations and results after the tweak

Properties of the first cluster estimated as below.

Statistics	true		T1		tb	
	x	y	x	y	x	y
.1	40.61340	18.47523	40.79700	20.76799	42.13426	21.74559
.2	46.37419	23.33148	46.14605	24.67839	46.14605	24.67839
.3	51.20806	26.77727	52.83238	29.56639	52.83238	29.56639
.4	54.16163	32.10846	56.17554	33.47680	56.17554	33.47680
.5	58.07539	36.22332	58.85007	37.38720	58.85007	37.38720
.6	64.15178	39.34763	66.87365	40.32000	66.87365	40.32000
.7	67.46237	42.53114	68.21092	42.27520	68.21092	42.27520
.8	73.63100	47.46381	73.55997	48.14081	73.55997	48.14081
.9	79.60559	55.46507	82.92083	56.93921	78.24040	56.93921
Mean	59.15551	35.48207	60.31051	36.18954	60.31051	36.18954
s.d	14.98289	13.51362	13.69708	13.69679	13.69708	13.69679
cor	0.6700317		0.8478699		0.8478699	

Simulations and results after the tweak

Properties of the second cluster estimated as below.

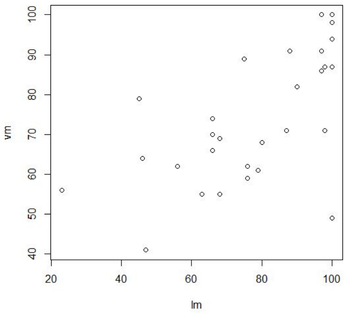
Statistics	true		T1		Tb	
	x	y	x	y	x	y
.1	139.4334	78.46023	140.4232	78.44643	141.0918	79.42403
.2	146.5080	84.52669	145.7723	85.28963	149.7840	85.28963
.3	150.9837	87.80759	153.7958	88.22243	153.7958	88.22243
.4	154.6282	92.60448	153.7958	92.13283	154.4645	92.13283
.5	158.5090	95.76052	158.4763	100.93124	158.4763	100.93124
.6	162.7737	98.35094	165.8312	101.90884	161.8194	101.90884
.7	167.2797	101.50168	168.5057	106.79684	168.5057	106.79684
.8	170.2898	106.34020	172.5175	106.79684	172.5175	106.79684
.9	180.3151	111.20021	180.5411	118.52805	180.5411	118.52805
Mean	158.8678	95.35943	159.6997	96.94615	159.6997	96.94615
s.d	14.96091	13.65429	14.68642	15.07792	14.68642	15.07792
cor	0.710895		0.7441762		0.7441762	

Comments

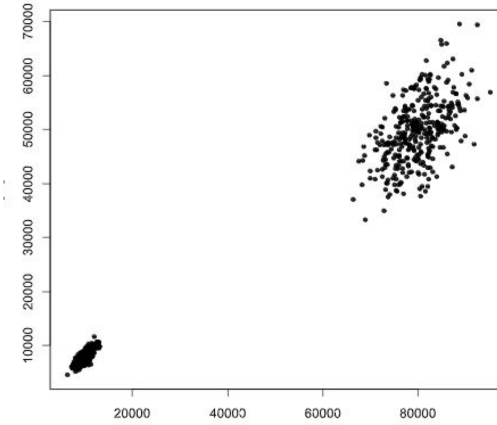
- The nature of the plot is prominent even after masking.
- The properties of both the clusters when estimated are found to be pretty close to the original clusters.

The Second Limitation

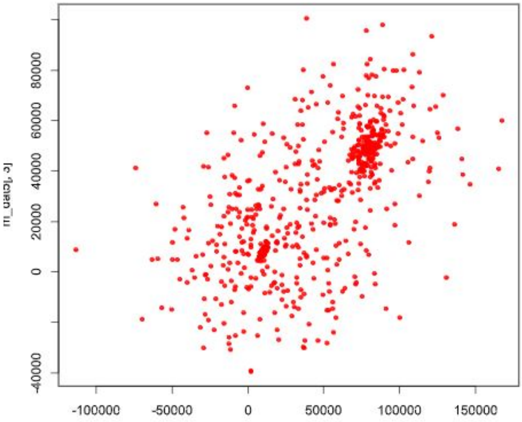
Some of the data sets that we came across did not even remotely fit the normal distribution and were visibly clustered with the variability differing in different clusters. This is a plot of marks obtained in 2 different subjects from our course.



For a visual demonstration, consider a data as below.



The masked data with fixed error variance destroys the pattern.



Alternative Approach

Now, like our previous tweak we will cluster the data at first. But, this time for the different clusters, we will allow different variances unlike the tweak where we only restricted the swapping.
We will basically divide the data set into several clusters and on each cluster we will do the masking with different variances.In this case we will not disclose the individual variances but we will report the value of a ratio r indicative of the variance. We will discuss the method in details in the next slide.

Alternative Approach(cotd.)

Notice that, in the initial method and also the tweak, we had,

$$\sigma_X^2 = \sigma_Z^2 - (1-p).\sigma_Y^2$$

where Z was the masked data and Y was the error added.
Now, let us mask the data in such a way that $\sigma_Y^2 = r^2\sigma_X^2$ for each cluster, where we fix r. Then,

$$\sigma_X^2 = \sigma_Z^2 - (1-p).r^2\sigma_X^2$$

$$\implies \sigma_Z^2 = \sigma_X^2(1 + (1-p).r^2)$$

Hence, if the value of r is disclosed instead of the value of σ_Y^2, we get,

$$\hat{\sigma}_X^2 = \frac{\hat{\sigma}_Z^2}{1 + (1-p).r^2}$$

Hence, in order to mask this type of clustered data, we follow the given steps:

1. As before, we cluster the data into k groups.
2. Now if the error added to all the clusters have the same variance, the clusters with smaller variances will end up having many outliers.
3. If we allow different error variance for different clusters, with the condition that $\sigma_Y^2 = r^2\sigma_X^2$, with some fixed value of r, this problem can be avoided.
4. We will disclose this value r.

5. So now if we have $B_i = 1$, the swapping procedure is same as before i.e., we swap inside the clusters.
6. If $B_i = 0$, for we set

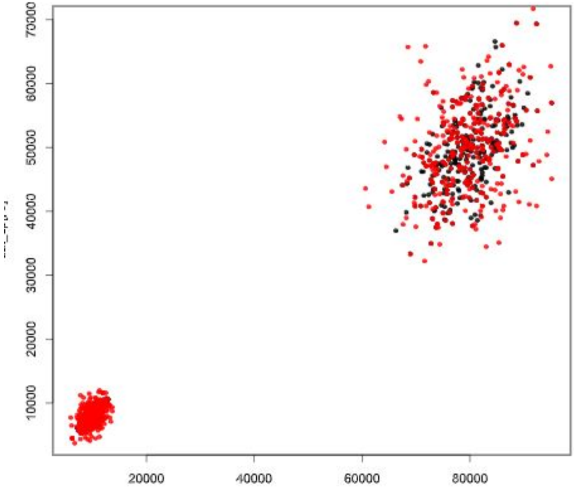
$$Z_i = X_i + Y_i$$

$$W_i = U_i + V_i$$

where, $Y_i \sim N(0, r^2\sigma^2_{Xj}), V_i \sim N(0, r^2\sigma^2_{Uj})$ independently, where σ^2_{Xj} and σ^2_{Uj} are variances of the jth cluster.
7. The estimation procedure is same within the clusters with $\hat{\sigma}^2_{Xj} = \frac{\hat{\sigma}^2_{Zj}}{1+(1-p).r^2}$ and error variance replaced with $\hat{\sigma}^2_{Yj} = r^2\hat{\sigma}^2_{Xj}$.

Simulation Results

We have implemented the above with a dataset with 2 clusters. The masked data appears as given.



Here, $r = 1$.

The estimation of the first cluster appears as given:

Statistics	true		T1		tb	
	x	y	x	y	x	y
.1	8643.576	6610.559	8652.798	6454.444	8680.037	6558.677
.2	9061.039	7053.652	8993.757	6973.626	9011.603	7031.202
.3	9444.296	7455.085	9457.761	7437.216	9517.875	7445.158
.4	9742.017	7754.899	9746.120	7707.230	9751.756	7708.223
.5	10028.494	8001.633	10057.022	7998.091	10057.022	7998.091
.6	10262.229	8192.130	10346.320	8170.821	10345.381	8166.850
.7	10525.626	8444.369	10573.626	8442.820	10564.233	8429.915
.8	10797.726	8881.497	10976.577	9010.644	10949.338	8920.309
.9	11315.559	9351.696	11392.678	9400.775	11375.771	9374.965
Mean	9976.744	7989.239	10005.94	7975.528	10005.94	7975.528
s.d	1050.973	1045.228	1078.329	1110.815	1078.329	1110.815
cor	0.8031327		0.7527303		0.7527303	

The estimation of the second cluster appears as given:

Statistics	true		T1		tb	
.1	73276.76	42126.34	73048.70	73250.94	41551.18	41656.75
.2	75446.19	44695.51	74959.47	75010.55	44335.85	44470.98
.3	76961.70	46792.74	76695.38	76696.43	47122.94	47136.21
.4	78557.16	48195.73	78450.25	78516.09	48649.20	48649.20
.5	79780.68	49623.49	79894.39	79861.21	49859.35	49859.35
.6	81074.60	50806.06	81310.61	81292.70	50974.78	50974.78
.7	82253.27	52357.11	82381.33	82381.33	52688.66	52628.33
.8	84124.81	54430.66	85212.72	85098.44	54455.02	54430.89
.9	86457.76	57542.65	86279.24	86216.56	57654.74	57395.34
Mean	79818.64	49699.23	79742.75	49891	79742.75	49891
s.d	5043.409	5938.391	5327.751	5805.605	5327.751	5805.605
cor	0.5216502		0.4594414		0.4594414	

Disclosure Risk

d	Risk
Cluster 1	
$\sigma_X/2$	0.321
$\sigma_Y/2$	0.310
Cluster 2	
$\sigma_X/2$	0.313
$\sigma_Y/2$	0.318

Table: Disclosure Risk

Comments

- The masked data clearly reserves the pattern in the original data.
- The quantile estimations are very close to the true values of quantiles.
- The mean and variances are estimated fairly close to the true values for both the variables.
- The estimate of correlation is close but has a higher margin of error than the previous cases due to the fact that the error variance is estimated instead of being known.