A Method for Avoiding Data Disclosure While Automatically Preserving Multivariate Relations

Norman Matloff* Patrick Tendick†

October 2, 2015

Abstract

Statistical disclosure limitation (SDL) methods aim to provide analysts general access to a data set while limiting the risk of disclosure of individual records. Many methods in the existing literature are aimed only at the case of univariate distributions, but the multivariate case is crucial, since most statistical analyses are multivariate in nature. Yet preserving the multivariate structure of the data can be challenging, especially when both continuous and categorical variables are present. Here we present a new SDL method that automatically attains the correct multivariate structure, regardless of whether the data are continuous, categorical or mixed.

1 Introduction

Statistical disclosure limitation (SDL) methods aim to provide analysts general access to a data set while limiting the risk of disclosure of individual records. Common methods include noise addition, swapping of parts of records, replacing data by synthetic equivalents, suppression of small cells in contingency tables, and so on [6].

Long the field of statistical research, in recent years SDL issues have attracted the interest of computer scientists [4]. There has been a marked contrast in the approaches taken by the two communities: The statistical view is that of serving

*Dept. of Computer Science, University of California, Davis
†Avaya

research analysts who wish to do classical inference from samples, while the computer scientists, coming from a cryptographic background, have viewed the data itself as the primary focus. In other words, in the computer science approach, the 'S' in SDL has perhaps had lesser attention, compared to the statisticians' view of things. However, there is some indication of increasing interaction between the two groups [1].

For an overview of how methodology has been refined and expanded over time, compare a 1989 survey paper [2], a 2002 Census Bureau viewpoint [5], the current statistical view [6], and the more recent computer science approach [4].

Whatever approach is taken, a primary goal remains statistical analysis by the end user. And in order to perform meaningful statistical analysis on the data, **one's methods must at least approximately preserve multivariate structure**. Most ststistical analysis — linear regression, logistic models, princple components analysis, the log-linear model and so on – are inherently multivariate. Unfortunately, many existing SDL methods place little or no emphasis on this aspect, and this is an absolutely central issue. Regression coefficient estimates, for instance, can turn out substantially biased as a result. As noted in [11],

> ...[in using] noise addition techniques...the original data suffers loss of some of its statistical properties even while confidentiality is granted, thus making the dataset almost meaningless to the user of the published dataset.

The above statement applies only to independent noise variables. Noise addition methods can preserve the multivariate structure of continuous variables, if the data come from an approximate multivariate normal distribution, by adding correlated noise [9] [7] [13]. However, this does not apply to the discrete-variable case, and moreover, the same problems apply to most if not all of the other major classes of SDL methods.

Developing methodology for the mixed continuous/discrete case is a difficult problem; see [8] and the citations therein for some existing methodology. To broaden the methods available to DBAs, a new method is proposed in this paper to deal with the multivariate structure preservation problem. Our method has several important advantages:

- The method works on general data, i.e. continuous, discrete or mixed.
- The method does not require the database administrator (DBA) to estimate

the dependency structure between the variables, or make assumptions regarding that structure.

- The method has several tuning parameters, affording database administrator broad flexibility in attaining the desired balance between privacy and statistical usability.

2 Overview of the Method

Let $W_{ij}, i = 1, ...n, j = 1, ..., p$ denote our original data on n individuals and p variables. Choose $\epsilon > 0$ and $0 < q \leq 1$. Then we form our released data W'_{ij} as follows:

For $i = 1, ...n$:

- Consider record i in the data base:

$$r_i = (W_{i1}, ..., W_{ip}) \tag{1}$$

- With probability $1 - q$, skip the next steps.
- Find the set S of points in the data set within ϵ distance of r_i.
- Draw a random sample (with replacement) of p items from S, resulting in values $a_{km}, k = 1, ..., p, m = 1, ..., p$.
- For $j = 1, ..., p$, set

$$W'_{ij} = a_{jj} \tag{2}$$

3 Theoretical Justification

Theorem: Consider a bivariate random vector (X, Y) and $\epsilon > 0$. For any t in R^2, let $A_{t,\epsilon}$ denote the ϵ neighborhood of t, defined by some metric. Let F denote the cdf of (X, Y), and define $G_{t,\epsilon}$ to be the conditional cdf of (X, Y), given that that vector is in $A_{t,\epsilon}$. Finally, given (X, Y), define *independent* random variables U and V to be drawn randomly from the first- and second-coordinate marginal distributions of $G_{(X,Y),\epsilon}$, respectively. Then

$$\lim_{\epsilon \to 0} P\left(U \leq a \text{ and } V \leq b\right) = F(a,b) \tag{3}$$

for all $-\infty < a, b < \infty$.

In other words, as ϵ goes to 0, the distribution of (U,V) goes to that of (X,Y), *even though U and V are conditionally independent*.

Proof:

Given $(X,Y) = t = (t_1, t_2)$,

$$\lim_{\epsilon \to 0} U = t_1 \tag{4}$$

and

$$\lim_{\epsilon \to 0} V = t_2 \tag{5}$$

Then by bounded convergence,

$$\begin{aligned}
\lim_{\epsilon \to 0} P\left(U \leq a \text{ and } V \leq b\right) &= \lim_{\epsilon \to 0} E\left[P\left(U \leq a \text{ and } V \leq b \mid X,Y\right)\right] && (6)\\
&= \lim_{\epsilon \to 0} E\left[P(U \leq a \mid X,Y) \cdot P(V \leq b \mid X,Y)\right] && (7)\\
&= E\left[1_{X \leq a} \cdot 1_{Y \leq b}\right] && (8)\\
&= E\left[1_{X \leq a \text{ and } Y \leq b}\right] && (9)\\
&= P(U \leq a \text{ and } V \leq b) && (10)\\
&= F(a,b) && (11)
\end{aligned}$$

■

The key word *independent* in the above theorem has a major implication: We can make our released data approximate the multivariate distribution of the original data (or the population from which the latter are drawn), **without knowing or even estimating the multivariate relationship of our variables**. We simply sample *independently* from S, yet attain the correct *dependency* relationship among the variables.

The bit of seeming similarity between this new method and data swapping is largely deceiving. Clearly our method does do swapping of values, and in some sense our neighborhood approach relates somewhat to the fact that data swapping is typically conducted on a within-stratum basis, such as strata defined by age and race; a stratum then has some similarity to our neighborhoods.

But actually the two methods are quite different. First, with data swapping, records from one stratum are switched with those in *another* stratum, whereas in our method everything stays within the same neighborhood. This is very important, because with data swapping, choosing the stratifying variables precludes analysts doing statistical analyses that including those variables.

Moreover, our swapping is done on individual variables, not entire records, and our neighborhoods can grow or shrink in size, as opposed to the fixed stratum size in data swapping.

4 Code and Tuning Parameters

The method provides the DBA with excellent flexibility in achieving the desired balance between privacy and accurate multivariate structure, via the following tuning tuning parameters:

- The neighborhood radius, ϵ.
- The distance metric.
- The proportion of modified records.

Code implementing the method is provided on GitHub (*https://github.com/matloff/statdb*) to implemeent the method. The call form is

```
nbrs(z, eps, modprop = 1, wts = NULL)
```

where **eps** is ϵ, **modprop** is q in the algorithm in Section 2, and the **wts** argument controls the distance metric, to be explained shortly. The return value is the released data set.

It is assumed that all categorical variables have been converted to dummy variables. Ordinary Euclidean distance is used on the scaled data, including any dummy variables. Scaling places all the variables on the same footing — all now have standard

deviation 1 — but there is still a difference between the continuous variables and the dummies and other discrete variables, as follows.

As sample size n grows (treating the original data as a sample from some population), one would want ϵ to become smaller, but this would not work well for the discrete variables. With large n, the latter would come to dominate the distance metric, and one could not drop ϵ below some minimum threshhold. Thus the **wts** argument provides the DBA with a tool to reduce that dominance, by allowing the weights of the discrete variables (or others) to decrease as n increases.

If for example we set **wts = c(5,12,13,rep(0.6,3))**. then in computing distances the variables in columns 5, 12 and 13 of the data matrix are reduced in weight by a factor of 0.6.

5 Example

We used the Census data set in the package **regtools** (*https://github.com/matloff/regtools*) to simulate an employee database, sampling 1000 records from this data.

To set the value of ϵ and the other tuning parameters, the DBA may devise a few representative statistical analyses, and then assess whether (a) the results of the analysis on the our released data set are reasonbly close to those of the original data and (b) whether records that were at risk in the original data are masked sufficiently well in the released data.

We ran a linear regression analysis, predicting WageIncome from Age, Gender, WeeksWorked, MSDegree and PhD. Suppose this (simulated) firm is concerned about possible gender discrimination withim the firm. Then they might focus on the estimated regression coefficient for Gender.[1].

In the original data, this is -10795.4, suggesting that women are on average paid about $11,000 less than men of the same age, number of weeks worked, and educational level.

To produce the released data, the call used was

```
p1p <- nbrs(p1,eps=0.3,wts=c(2,4,5,rep(0.2,3)))
```

The estimated Gender coefficient now became -10591.9, a change of about 1.9%, presumably acceptable to the DBA. What about disclosure avoidance?

[1] In a real study many more variables would need to be included. Note too that the data was not cleaned before use; for instance, some values for WageIncome are 0 but clearly should not be.

In the original data, there had been exactly one record, with a WeeksWorked value of 43, employee number 5016:[2]

```
           age  sex  wkswrkd  ms  phd  wageinc
5016  33.97025    1       43   0    0        0
```

Suppose, just as a simple example, that this employee happened to mention to an intruder that he had worked 43 weeks. Would that expose his salary?

In the released data, there is again one such record — but now for employee number 3208:

```
           age  sex  wkswrkd  ms  phd  wageinc
3208  36.46872    1       43   0    0        0
```

So, an intruder who knew that only one employee had worked 43 weeks would not be able to identify the new record for employee number 5016, which happens to be this:

```
           age  sex  wkswrkd  ms  phd  wageinc
5016  36.46872    1       40   0    0        0
```

This employee's value of WeeksWorked changed from 43 to 40. On the other hand his (Gender = 1 meant male, 2 for female) age changed only slightly. Might this have helped an intruder? Possibly, but the intruder also knows that in the released data this particular employee may have been listed as female. With several tuning paramenters available, the DBA has many ways in which to exploit this uncertainty in the intruder's mind.

6 Discussion

Note that "a little bit of privacy can go a long way": As long as the intruder knows that the data have been modified (even for the nonsensitive variables), there may be enough doubt in his/her mind as to make the data useless for nefarious purposes (while still being very useful for legitimate purposes).

In databases with large p, one must take into account the Curse of Dimensionality [3]. The DBA may choose to use a weighted distance metric, with the weights going to 0 as the variable index goes to infinity [10].

[2]The employee numbers would be suppressed in actual usage.

In general, the choice of ϵ must also be made carefully This approach does require fairly large data sets, so that for example the set S contains some female workers. One might even allow the value of ϵ to vary from record to record.

References

[1] J. Abowd (2015). Comments as the Discussant in a session at JSM 2015.

[2] N.R. Adam and J.C. Wortmann (1989). Security-Control Methods for Statistical Databases: A Comparative Study, *ACM Computing Surveys*, 21(1989).

[3] K. Beyer, J. Goldstein, R. Ramakrishnan (1999). When Is "Nearest Neighbor" Meaningful?, *Lecture Notes in Computer Science*, Volume 1540, 1999, 217-235.

[4] C. Dwork (2008). Theory and Applications of Models of Computation *Lecture Notes in Computer Science, Vol. 4978*, M. Agrawal *et al* (es.), 1-19.

[5] U.S. Census Bureau (2002). *Census Confidentiality and Privacy: 1790 - 2002*, *http://www.census.gov/prod/2003pubs/conmono2.pdf*.

[6] Duncan, G., Elliot, M., Salazar, G. (2011). *Statistical Confidentiality: Principles and Practice*, Springer.

[7] J. Kim (1986). A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, *Proceedings of ASA Section on Survey Research Methods*, 370-374.

[8] Manrique-Vallier, D., Reiter, J. (2012). "Estimating Identification Disclosure Risk Using Mixed Membership Models," *JASA*, 107, 1385-1394.

[9] N. Matloff (1986). Another Look at the Use of Noise Addition for Database Security. *Proceedings of the 1986 IEEE Symposium on Security and Privacy*, April 1986, pp. 173-180.

[10] N. Matloff (2015). Big-n vs. Big-p in Big Data, in *Handbook of Big Data*, Buhlmann and Kane (eds.), Chapman and Hall, to appear.

[11] K. Mivule (2011). *Utilizing Noise Addition for Data Privacy, an Overview*, *http://arxiv.org/pdf/1309.3958.pdf*.

[12] Shlomo, N., Skinner, C. (2008). "Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata," *Annals of Applied Statistics*, 4,3, 1291-1310.

[13] P. Tendick and N. Matloff (1994). A Modified Random Perturbation Method for Database Security, *ACM Transactions on Database Systems*, 19, 47-63.

[14] W. Winkler (2005). *Microdata Confidentiality References*, *https://www.census.gov/srd/sdc/Winkler.List.May.2005.pdf*.