A Method for Avoiding Data Disclosure While Automatically Preserving Multivariate Relations

Norman Matloff[*] Patrick Tendick[†]

October 14, 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. In addition, operational methods for assessing data quality and risk will be explored.

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 [12],

> ...[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 [10] [8] [14]. 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 [9] and the citations therein for some existing methodology. To broaden the methods available to Data Stewardship Organizations (DSOs), 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 DSO 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. Moreover, 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 DSO 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 implement 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 DSO 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 Selection of Tuning Parameters

In some modern statistical methods, the user is faced with selection of a large number of tuning parameters, both numeric and policy-oriented, such as in the SIS package [7]. The user may find the task of setting those parameters daunting and bewildering.

In SDL settings, though, the DSO may *welcome* the selection of tuning parameters. The goal is achieving a good balance between statistical accuracy of the released data and disclosure risk, so from the DSO's point of view, the more tuning parameters the better.

5.1 Choices

For a given set of tuning parameters, the DSO wishes to assess

(a) whether the results of statistical analyses on the 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.

Regarding (a), we propose an operational approach.[1] Though many authors have proposed global measures of distance between the original and released data sets, we suggest gauging the statistical accuracy of the latter in a more directmanner, motivted by the intended usage of the data, namely statistical analyses.

[1]We have not seen this in the literature, though it is likely that some DSOs have experimented with this approach.

In other words, under this approach the DSO would run several representative statistical analyses, say regression and principle components, on both the original and released data sets. The DSO would then compare the resultss

Our approach to issue (b) is similarly practical. The DSO identifies some representative (sample) uniques, and then tracks what happens to them in the released data. Have they been hidden sufficiently well?

We advocate these methods (which of course can be used in conjunciton with other methods) because they expose the system in ways that *directly* address the goals (a) and (b):

- No matter what SDL method is used – noise addition, cell suppression, data swapping, our method introduced in this paper, etc. — will necessarily result in some distortion to statistical analyses. The fact that two (empirical) distributions are close of course does not imply that a given functional will have similar values on those two distributions.

  Thus is vital to get an idea *how much* distortion the statistical users of the data may need to tolerate. This is what our approach addresses.

- An example in some of the SDL literature has involved preserving the privacy of the lone female electrical engineer in a company employee database. The DSO can pose questions like this for their given data set, and find that, say, while the female EE was hidden, the lone programmer over age 50 was not.

5.2 The Roles of n and p

In setting these parameters, the DSO must take into account not only the desired balance between (a) and (b) above, but also the values of n and p. For fixed p, the larger n is, the fewer the number of uniquely identifiable individuals in the data, and thus the decreased need for privacy measures.[2] On the other hand, for fixed n, the larger the value of p, the more potential identifiable uniques.

[2]As noted, we are treating the data as a sample from some (tangible or conceptual) population. As such, the notion of a *population unique*, seen in some of the SDL literature, doesn't apply. If a combination of the categorical variables appears in our data, then by definition that combination has nonzero probability in the population, and we'll get more and more individuals of that type as n grows. For continuous variables, a similar statement holds in the sense that as n grows, we will have more and more individuals near the given value.

6 Example

We used the Census data set in the package **regtools** (*https://github.com/matloff/regtools*) to simulate an employee database, sampling 5000 records from this data.[3]

The call used was

```
> p1p <- nbrs(p1,eps=0.3,wts=c(2,4,5,rep(0.05,3)))
```

To gauge how close this new version of the data was to the original, we ran a linear regression analysis, predicting WageIncome from Age, Gender, WeeksWorked, MSDegree and PhD. The estimated coefficients for the original and modified data were

data	Age	Gender	WeeksWorked	MS	PhD
original	447.2	-9591.7	1286.4	17333.0	21291.3
released	466.1	-8423.2	1270.7	18593.9	22161.4

The results fairly good. What about disclosure risk?

In the original data set, there was one female worker with age under 31:

```
> p1[p1$sex==2 & p1$phd==1 & p1$age < 31,]
          age sex wkswrkd ms phd wageinc
7997 30.79517   2      52  0   1  100000
```

How well was she hidden in the modified data? Quite well, it turns out:

```
> p1pc <- na.omit(p1p)
> p1pc[p1pc$sex==2 & p1pc$phd==1 & p1pc$age < 31,]
           age sex wkswrkd ms phd wageinc
12522 30.5725   2      52  0   1   50000
```

There is one person listed in the released data of the given description (female, PhD, age $<$ 31). But she is listed as having an income of \$50000 rather than \$10000. In fact, it is a different person, worker number 12522, not 7997.[4] Where is the latter now?

```
> which(rownames(p1p) == 7997)
[1] 3236
> p1p[3236,]
```

[3]Since this is just an illustration, the data were not cleaned, and some WageIncome values were 0 that probably should have been designated as missing.

[4]Of course, ID numbers would be suppressed.

```
         age  sex  wkswrkd  ms  phd  wageinc
7997  31.9746    1       52   0    1   100000
```

Ah, she became a man! That certainly hides her.

There are now four women fitting the given conditions, none of which was the one we highlighted in the original data, worker number 7997

Next, we tried $\epsilon = 0.2$. The new regression coefficients were

data	Age	Gender	WeeksWorked	MS	PhD
relased	520	-9063	1278	18342	23532

Now there were no workers in the modified data set satisfying the given conditions, so again worker 7997 is protected:

```
> p1pc[p1pc$sex==2 & p1pc$phd==1 & p1pc$age < 31,]
[1] age      sex      wkswrkd ms      phd      wageinc
<0 rows> (or 0-length row.names)
```

This of course just barely scratches the surface of the various tuning parameter values that the DSO could experiment with, in addition to doing so on other types of analyses, say principle components analysis.

7 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 DSO may choose to use a weighted distance metric, with the weights going to 0 as the variable index goes to infinity [11].

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.

8 Work to Be Done

The presentation here is of course preliminary, and many aspects need to be explored. The method will be tried on a wide variety of data sets; effects of varying the tuning parameters will be explored; the possible usefulness of making the values of the tuning parameters vary from one record to another will be investigated; and so on.

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. Fan, Y. Feng, D. Saldana, R. Samworth and Y. Wu (2015). “Package ‘SIS”’, CRAN, *https://cran.r-project.org/web/packages/SIS/index.html*,

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

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

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