Combining SDL methods to preserve relationships and data-specific constraints in survey data

Anna Oganian
National Center for Health Statistics, USA

Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.
Data Linkage and Anonymisation. New developments in Data Privacy, December 7th 2016
Motivation

- Swapping and noise are among the most widely used approaches by statistical agencies for public-use non-interactive and non-synthetic data release.

- Their core ideas are conceptually easy to understand (and to implement) and are naturally suited for masking purposes.

- They are worth revisiting with a special emphasis given to their utility aspects.

- I will describe how these methods can be combined to increase each others efficiency while protecting variables of different types.
Additive Noise

• Usually applied to continuous variables.
• Desirable properties of additive noise: preserves means and correlation structure.
• However, if noise (usually Normal) is added to the original values some data-specific constraints, such as positivity and inequality constraints will not be satisfied.

  Age, economic variables (gross income, taxes) and many “demographic” variables obey positivity constrains.
  Federal taxes ≤ gross income

• Risk aspect: the extent to which constraints are violated can be informative about the nature and intensity of SDL applied to the data.
• **We want to retain the desirable properties of additive noise, but in addition, we want to preserve these structural constraints.**
Relationship of Additive and Multiplicative noises

Additive noise usually satisfies the following properties:

\[
\begin{align*}
E[(X_o + E)] &= E[X_o] \\
\Sigma(X_o + E) &= (1 + c)\Sigma(X_o) 
\end{align*}
\]

By analogy with additive noise we can require:

\[
\begin{align*}
E[X_o \circ \exp(E)] &= E[X_o] \\
\Sigma(X_o \circ \exp(E)) &= (1 + c)\Sigma(X_o)
\end{align*}
\]
We will not release $X_o \circ \exp(E)$ but the following transformation:

$$X_m = \frac{(\sqrt{1+c} - 1)E[X_o] + [X_o \circ \exp(E)]}{\sqrt{1+c}} \quad (3)$$

It is straightforward to verify that if

$$
\begin{align*}
E[X_o \circ \exp(E)] &= E[X_o] \\
\Sigma(X_o \circ \exp(E)) &= (1+c)\Sigma(X_o)
\end{align*}
$$

then

$$
\begin{align*}
E[X_m] &= E[X_o] \\
\Sigma(X_m) &= \Sigma(X_o)
\end{align*}
$$

So masking parameter $c$ does not need to be released to adjust for variance inflation.

There are indications from the ongoing research that the release of parameter $c$ can be dangerous for multiplicative noise as it can lead to the disclosure of the original value $X_o$. 
What kind of noise would satisfy \[
\begin{align*}
E[X_o \circ \exp(E)] &= E[X_o] \\
\Sigma(X_o \circ \exp(E)) &= (1 + c)\Sigma(X_o)
\end{align*}
\]

For \( E \sim N(\mu_E, \Sigma_E) \) the solution of

\[
\begin{align*}
E[X_o \circ \exp(E)] &= E[X_o] \\
\Sigma(X_o \circ \exp(E)) &= (1 + c)\Sigma(X_o)
\end{align*}
\]

is:

\[
\begin{align*}
\Sigma_E(i, j) &= \log(1 + \frac{c\Sigma_o(i, j)}{E[X_o(i)X_o(j)]}) \\
\mu_E(i) &= -\sigma_E(i) / 2
\end{align*}
\]

These the parameters of \( E \) lead to the lognormal distribution of \( \exp(E) \) with the shape similar to Normal, so the introduced skewness in the masked data is minimal.
General case

If the data contains not only nonnegative variables but also variables with negative and positive values it may happen that

\[
1 + \frac{c \Sigma_o(i,j)}{E[X_o(i)X_o(j)]} < 0
\]

(6)

Lemma:
If data set contains variables of different types (strictly positive, non-negative, variables that can take positive and negative values) then the following transformation guarantees that expression (3) never take place.

\[
X_m = (\sqrt{1 + c - 1})E[X_o] \circ \exp(E_{z_{pos}}) + [X_o \circ \exp(E_{z_{pos}})]
\]

(7)

where \( E_{z_{pos}} \sim N(\mu_{E_{z_{pos}}}, \Sigma_{E_{z_{pos}}}) \) with \( \mu_{E_{z_{pos}}} \) and \( \Sigma_{E_{z_{pos}}} \) computed using (5) but as applied to the Z-scores of the original data plus \( \sqrt{1 + c E(X_o) / \sigma_o} \).
Preservation of inequality constraints

If in addition to positivity constraints we have also inequality constraints \( X > Y \),

<table>
<thead>
<tr>
<th>Original data</th>
<th>Mult. noise transform</th>
<th>Masked data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>( Y^* )</td>
<td>( Y^* + [X - Y]^* )</td>
</tr>
<tr>
<td>( Y )</td>
<td>( [X - Y]^* )</td>
<td>( Y^* )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>( y_1 )</td>
<td>( y_1^* + [x_1 - y_1]^* )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2 )</td>
<td>( y_2^* + [x_2 - y_2]^* )</td>
</tr>
<tr>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
</tr>
</tbody>
</table>

\((Y^*, [X - Y]^*)\) - result of the application of multiplicative noise (described in the previous section) to the bivariate data \((Y, X-Y)\).

This masking scheme have the following properties:

1) \( \text{E}(X_m, Y_m) = \text{E}(X, Y) \)
2) \( \text{cov}(X_m, Y_m) = \text{cov}(X, Y) \)
Example with multiple variables related by inequality constraints:

\[ X_{O_1} > X_{O_2} > X_{O_3} \quad X_{O_4} > X_{O_5} \quad X_{O_6} \quad X_{O_d} \]

Masking scheme will be the following:

\[
\begin{align*}
X_{m_1} &= X_3^* + [X_{O_2} - X_{O_3}]^* + [X_{O_1} - X_{O_2}]^* \\
X_{m_2} &= X_3^* + [X_{O_2} - X_{O_3}]^* \\
X_{m_3} &= X_3^* \\
X_{m_4} &= X_5^* + [X_{O_4} - X_{O_5}]^* \\
X_{m_5} &= X_5^* \\
X_{m_6} &= X_6^* \\
&\vdots \\
X_{m_d} &= X_d^*
\end{align*}
\]
Swapping

• Swapping may significantly affect the relationships between the variables. It is especially problematic when these variables are design variables or the variables defining important subgroups in the population.

• I will present an approach similar to swapping which takes into account relationships between the variables, thus reducing the impact of disclosure limitation.

• I will show how the two methods - our versions of multiplicative noise and swapping - can be used together so that the relationships between swapped categorical variables and continuous masked variables (which were applied noise) be preserved.
Conditional Group Swapping (CGS)

1. **CGS moves the records** from one strata to another in such a way so that the within-strata statistics are affected as little as possible. As a consequence the overall relationships between the variables should be better preserved.

2. The records are moved between the closest strata only (closest from the point of view of distributions of the strata).

3. PPS sampling is used to chose the records that should be moved to another stratum. The **probability of selection is proportional** to the **probability of classification of this record to the stratum where it is being moved**. In this way only the records that fit the best the distributional properties of the stratum where they are being moved are selected.
CGS algorithm

1. Compute pairwise distances between all the strata. We use the propensity score metric to measure the distance (Woo, et al., 2009).

2. For all the records in two closest strata A and B compute the swapping probabilities in the following way:
   a. Combine together all the records from A and B and add an indicator variable T: T=1 for all the records from stratum B and T=0 for the records from A.
   b. For every record i in stratum A compute $P_{i}^{AB} = \Pr(i \rightarrow B | x_i)$ - conditional probability that i is assigned to stratum B given the values of the variables for this record, $x_i$, that is, compute the propensity score.
CGS algorithm (continued)

3. Swap the records:
   a. Select $n_s$ records from stratum A ($n_s$ is the desired swapping rate) with the probabilities proportional to $P_{iAB}$. For the selected records change their stratum indicator to B.
   b. Select $n_s$ records from stratum B with probabilities proportional to $P_{iBA} = \Pr(i \rightarrow A \mid x_i)$ and ``move'' them to stratum A. (The records that arrived from stratum A on the previous step will be excluded from the selection).

4. Repeat steps 2 and 3 for another pair of strata with the next closest distance.

5. Repeat step 4 until there is no more strata that have not been swapped.
Combination of CGS and multiplicative noise

1. Divide data set in strata.
2. Compute pairwise distances between the strata (Step 1 of CGS method).
3. For the two closest strata A and B estimate the model parameters to compute propensity scores, $P_{i}^{AB}$:

$$\log \left( \frac{P_{i}^{AB}}{1 - P_{i}^{AB}} \right) = \beta_{0}^{or} + \beta_{1}^{or}V_{1}^{or} + \beta_{2}^{or}V_{2}^{or} + \cdots + \beta_{k}^{or}V_{k}^{or}$$

where terms $V_{i}^{or}$ are original variables or their functions, and coefficients $\beta_{i}^{or}$ are estimated based on the original data.
4. Apply multiplicative noise described above to continuous variables in strata A and B.
5. Compute swapping probabilities for categorical variables using model (9) estimated from the original data, but substitute $V_{i}^{or}$ by $V_{i}^{mask}$ (masked values of $V_{i}^{or}$) when computing propensity scores.

6. Swap the selected $n_s$ records between strata A and B using swapping probabilities computed on the previous step (Step 3 of CGS).

7. Repeat steps 3, 4, 5, 6 for the next closest strata until there is no more strata that have not been perturbed.
Numerical experiments

Titanic data set: 889 passengers; 8 variables of different types.

Survived - survival status (0=No; 1=Yes)
Pclass - passenger class (1=1st; 2=2nd; 3=3rd),
Sex - sex,
Age - age in years,
SibSp - number of siblings/spouses aboard,
Parch - number of parents/children aboard,
Fare - passenger fare,
Embarked - port of embarkation (C= Cherbourg; Q=Queenstown; S= Southampton).
Relevant analyses and important subpopulations.

**Analyses:** What sorts of people were likely to survive?

Since sinking of Titanic there was a widespread belief of the social norm “Women and children first”. However, (Elinder and Erixson, 2011) study present a very different picture of gender-specific death rates in maritime disasters. According to (Elinder and Erixson, 2011) Titanic events (20% of men and 70% of women survived) were highly unusual, if not unique!

Besides gender, the class of the Titanic passengers was also related to their survival status.

**Six strata:** 1) first class male, 2) first class female, 3) second class male, 4) second class female, 5) third class male, 6) third class female.
Goals of numerical experiments

1) **Explore potential benefits of using conditional swapping probabilities for data utility.**

   Compare Conditional Group Swapping (CGS) to Random Group Swapping (RGS) characterized by uniform swapping probabilities. RGS reflects the idea of the traditional approach for swapping.

   Details of Swapping:
   RGS and CGS were implemented in the same way, except for the way how the probabilities of swapping were computed.
   Swapping rates: $n_s = 20$ and $n_s = 40$ records exchanged between the strata.
   For each swapping rate, we generated 100 realizations of swapped data using Random and Conditional Groups Swapping.

2) **Test the combination of multiplicative noise and swapping.**
Analysis-specific utility: comparison of regression coefficients’ confidence intervals

Titanic data regressions:
1. Logistic regression on the complete Titanic data, Reg1:
   \( \text{Survived} \sim \text{Pclass} + \text{Sex} + \text{Age} \)
2. Logistic regressions within each stratum, Reg2:
   \( \text{Survived} \sim \text{Age} + \text{Fare} \).

Metric: relative overlap of the confidence intervals for regression coefficients:

\[
J_k = \frac{1}{2} \left[ \frac{U_{\text{over},k} - L_{\text{over},k}}{U_{\text{orig},k} - L_{\text{orig},k}} + \frac{U_{\text{over},k} - L_{\text{over},k}}{U_{\text{mask},k} - L_{\text{mask},k}} \right]
\]  
(10)
Conditional Group Swapping alone  
Regression-specific utility

<table>
<thead>
<tr>
<th></th>
<th>CGS</th>
<th>RGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rate</td>
<td>Average</td>
</tr>
<tr>
<td>Reg 1</td>
<td>20</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.65</td>
</tr>
<tr>
<td>Reg 2</td>
<td>20</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 1. Original and masked confidence interval overlap (larger overlap – better the regression-specific utility).

Reg1: Survived ~ Pclass + Sex + Age (on the whole data)
Reg2: Survived ~ Age + Fare (within each stratum)
Generic propensity score utility measure

- This is a scalar that can be used as a relative distance measure, i.e. to compare the difference between the distributions of data set A and data set B, with the difference between the distributions of A and C, A and D and so on...
- Based on the idea of discrimination: if it is difficult to discriminate between the records from data sets A and B then the two data sets have similar distributions.
- Discrimination is based on propensity scores, \( e(x_i) \).
- Main idea: Data sets A and B have approx. the same distributions if the propensity scores for all the records in the combined set are approx. the same (Woo et al., 2009).

\[
PS = \sum_i (\hat{e}(x_i) - c)^2 \tag{11}
\]

where \( c \) is the proportion of records in the combined data set which are from B.
### Conditional Group Swapping alone

**Generic utility**

<table>
<thead>
<tr>
<th>Rate</th>
<th>CGS Average</th>
<th>CGS Range</th>
<th>RGS Average</th>
<th>RGS Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.25</td>
<td>(0.02, 0.63)</td>
<td>4.22</td>
<td>(3.11, 5.86)</td>
</tr>
<tr>
<td>40</td>
<td>2.22</td>
<td>(1.30, 1.47)</td>
<td>14.53</td>
<td>(11.38, 17.50)</td>
</tr>
</tbody>
</table>

Table 2. Generic propensity score data utility/information loss (smaller values – better the overall utility).
Multiplicative noise and CGS
Regression-specific utility

<table>
<thead>
<tr>
<th></th>
<th>CGS(MN(original)) – as described on Slides 14-15 (Variant 1)</th>
<th>MN(CGS(original)) independently for each stratum (Variant 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rate</td>
<td>Average</td>
<td>Range</td>
</tr>
<tr>
<td>Reg 1</td>
<td>20</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.66</td>
</tr>
<tr>
<td>Reg 2</td>
<td>20</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Table 3. Original and masked confidence interval overlap (larger overlap – better the regression-specific utility). Noise parameter c=0.15

Reg1: Survived ~ Pclass + Sex + Age (on the whole data)
Reg2: Survived ~ Age + Fare (within each stratum)
### Generic propensity score data utility

<table>
<thead>
<tr>
<th></th>
<th>(Variant 1)</th>
<th>(Variant 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Range</td>
</tr>
<tr>
<td>rate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.53</td>
<td>(0.98, 2.40)</td>
</tr>
<tr>
<td>40</td>
<td>3.22</td>
<td>(2.04, 4.49)</td>
</tr>
</tbody>
</table>

Within strata generic utility varied from the stratum to stratum.

MN + CGS dependent on MN worked better than RGS + independent CGS for those strata where continuous variables were significant predictors of strata-defining variables which were swapped.
Notes:

• An overall advantage in utility of Variant 1 combination over Variant 2 combination is not “dramatic” for this data set.

• Nevertheless, when big values of noise were generated for continuous variables, Variant 1 led to the selection of “more appropriate” records for swapping.

Example: record i got a large value of noise for variable $X_j$ in stratum A, as a consequence the probability of being moved to stratum B increased, because the mean of $X_j$ is larger in stratum B. The masked record now “fits better to the (distribution of) stratum B” comparative to a similar record which got a smaller value of noise.

Swapping in the combination CGS(MN(original)) (Variant 1) has a potential to “counteract” the effects of other methods.
Conclusions and future research

- Multiplicative noise scheme is a good choice for masking continuous variables that obey sign constraints and inequality constraints, and for the analyses when it is important to preserve the first two moments.

- Conditional Group Swapping may be a good choice for the stratified data when strata-defining variables need to be masked and it is important to preserve relationships between these variables and other variables in the data.

- Combining multiplicative noise and conditional group swapping has a potential to improve the distributional characteristics of the data when the methods are applied sequentially and conditionally one on other.
• We carried out an initial risk assessment of the combination of CGS and multiplicative noise.

• The re-identification disclosure risk, defined as an average percentage of correctly identified records when record linkage techniques are used to match the original and masked data, was low: about 4% of all records were correctly identified.

• Thorough investigation of the disclosure risk for the combinations of Conditional Group Swapping together with other SDL methods is the topic of our future research.