ROMM Methodology for Microdata Release

Daniel Ting and Stephen E. Fienberg
Carnegie Mellon University
Pittsburgh, PA USA

and

Mario Trottini
Universidad de Alicante, Spain
Outline

• Some existing methods
• Motivating principles
• Random Orthogonal Matrix Masking (ROMM)
• Numerical implementation
• Risk-Utility tradeoff
• Simulation
Methods for Microdata

- Micro data in form of $n \times k$ data matrix $x$.
- Examples of methods in literature:
  - Adding noise
  - “Statistical obfuscation”
  - PRAM
  - Rank swapping
  - Data shuffling
  - Multiple imputation
- Many of these are examples of matrix masking:
  - $x \rightarrow y = AxB + C$. 
Motivating Principles

• Need to allow data analyst to make inferences about parameters of interest in model applicable to original unreleased data.

• One way to achieve this is to provide details of the transformation method to allow creation of a usable likelihood function for the true unreleased data, c.f. PRAM.

• Another strategy is to preserve essential features of the data as part of the transformation process, e.g., minimal sufficient statistics.
  – For multivariate normal data, MSSs are mean and covariance matrix.
Random Orthogonal Matrix Masking (ROMM)

1. Generate a random orthogonal matrix, $t$, from a distribution $G$ defined on group of $n \times n$ orthogonal matrices which keep $1_n$ invariant, i.e., $t \ 1_n = 1_n$ where $1_n$ is the column vector consisting of $n$ 1’s.

2. Apply orthogonal operator, $t$, to original data $x$ to produce perturbed microdata $y$: $y = tx$.

3. Release (a) $y$; (b) information that $y$ has been obtained applying orthogonal operator randomly generated from distribution $G$; (c) exact distribution $G$. 


ROMM Properties

Theorem 1: Let $\bar{x}$ and $\Sigma_x$ be the sample mean and sample covariance matrix of the original microdata and let $\bar{y}$ and $\Sigma_y$ be the corresponding quantities in the masked microdata produced by ROMM. Then $\bar{x} = \bar{y}$ and $\Sigma_x = \Sigma_y$.

Theorem 2: Let M be any data masking procedure that generates a random microdata, $y$, with the same sample mean and sample covariance matrix as the original microdata. Then M is a special case of ROMM for a suitable choice of the “parameter” G.
Implementation

- Co-ordinate by co-ordinate:
  - Add small amount of noise to identity matrix and then make it orthogonal.
    1. Choose a parameter $\lambda > 0$ corresponding to the magnitude of perturbation.
    2. Draw $n \times n$ random matrix $M$ with entries from $N(0,1)$.
    3. Put $P = I + \lambda M$.
    4. Apply Gram-Schmidt and normalize the columns of $P$ to obtain an orthonormal matrix $T$.
      - It is easy to see that, when $\lambda = 0$, $T$ is the identity and no perturbation has occurred. When $\lambda = 1$, $T$ is a draw from the uniform distribution on orthogonal matrices.

- Block diagonal distributions:
  - Orthogonal matrices with eigenvalues near 1.
Example: Boston Housing Data

- Extract of 13 randomly selected observations (from 506) on 4 (of 20) variables:
  - RM: average number of rooms per dwelling
  - PTRATIO: pupil-teacher ratios per town
  - LSTAT: % of lower status population
  - MEDV: med. value of owner-occupied housing in $1000

<table>
<thead>
<tr>
<th>Obs</th>
<th>RM</th>
<th>PTRATIO</th>
<th>LSTAT</th>
<th>MEDV</th>
<th>Obs</th>
<th>RM</th>
<th>PTRATIO</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.630</td>
<td>19.5</td>
<td>6.53</td>
<td>26.6</td>
<td>9</td>
<td>6.315</td>
<td>16.6</td>
<td>7.60</td>
<td>22.3</td>
</tr>
<tr>
<td>3</td>
<td>6.709</td>
<td>14.7</td>
<td>15.79</td>
<td>19.4</td>
<td>11</td>
<td>5.757</td>
<td>20.2</td>
<td>10.11</td>
<td>15.0</td>
</tr>
<tr>
<td>4</td>
<td>6.877</td>
<td>14.7</td>
<td>12.14</td>
<td>23.9</td>
<td>12</td>
<td>5.304</td>
<td>20.2</td>
<td>26.64</td>
<td>10.4</td>
</tr>
<tr>
<td>5</td>
<td>6.462</td>
<td>14.7</td>
<td>11.32</td>
<td>22.3</td>
<td>13</td>
<td>5.423</td>
<td>20.2</td>
<td>12.03</td>
<td>16.1</td>
</tr>
<tr>
<td>6</td>
<td>6.762</td>
<td>15.2</td>
<td>6.69</td>
<td>32.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6.423</td>
<td>19.1</td>
<td>9.52</td>
<td>24.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Comparison of:
  - ROMM, co-ordinate by co-ordinate, with $\lambda=1/3$.
  - Comparison with bias corrected correlated additive noise method of Kim, with $\sqrt{c}=1/2$.

- Regression of MEDV on other variables.
Example Results

<table>
<thead>
<tr>
<th></th>
<th>RM</th>
<th>PTRATIO</th>
<th>LSTAT</th>
<th>MEDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROMM</td>
<td>0.146</td>
<td>5.618</td>
<td>33.307</td>
<td>34.327</td>
</tr>
<tr>
<td>AddNoise</td>
<td>0.066</td>
<td>1.641</td>
<td>13.950</td>
<td>7.348</td>
</tr>
</tbody>
</table>

### Original Data (ROMM)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(intercept)</td>
<td>-5.5641</td>
<td>23.6617</td>
</tr>
<tr>
<td>RM</td>
<td>7.4488</td>
<td>3.3663</td>
</tr>
<tr>
<td>PTRATIO</td>
<td>-0.9557</td>
<td>0.3691</td>
</tr>
<tr>
<td>LSTAT</td>
<td>-0.1770</td>
<td>0.2741</td>
</tr>
</tbody>
</table>

### Additive Noise

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.4696</td>
<td>25.1829</td>
</tr>
<tr>
<td>RM</td>
<td>6.2655</td>
<td>3.5843</td>
</tr>
<tr>
<td>PTRATIO</td>
<td>-0.3930</td>
<td>0.4329</td>
</tr>
<tr>
<td>LSTAT</td>
<td>-0.6780</td>
<td>0.2995</td>
</tr>
</tbody>
</table>
R-U Confidentiality Map

(Duncan, et al. 2004)
• A rigorous assessment of disclosure risk and utility requires:
  – Model for users’ behaviors when the output of ROMM is released.
  – Assessment of agency uncertainty about this model's inputs (users' targets, prior information, estimation procedure, etc.).
  – Formalization of agency's perception of the consequences of data users' actions and of agency's preference structure for consequences of users’ actions.

• We consider a simplified scenario where:
  – Modeling of users and agency's behaviors does not take explicitly into account some relevant aspects of the problem.
  – Agency has no uncertainty about the users’ model inputs.
Technical Details

• In the paper:
  – Utility
    • Under normality.
    • Under non-normality.
  – Disclosure Risk
Simulation Setup

• **Data:** We used same 13 values from Boston Housing Data in Example.

• **Assumptions:**
  – Intruder’s prior on the original data is uniform.
  – External information available to intruder is values of RM with i.i.d. $N(0,(0.6)^2)$ error.
  – Agency uses coordinate-by-coordinate approach to perturb data. It releases (a) perturbed data and (b) value of parameter $\lambda$.

• **Method:** Acceptance-rejection sampling to sample from posterior.
For $\lambda=0.2$, 0.4, and 0.6, we generated a sequence of $N=80$ (perturbed data, intruder data) pairs and sampled from the posterior of the original data given each pair.

We then calculated point estimates for each “individual.”

We then looked at the behavior of the distribution of these point estimates which allows us to evaluate risk.
Results (small $\lambda$)

• For small $\lambda$ (i.e., $\lambda = 0.2$), the mean point estimates were close to the true values.

• Average MSE of point estimates for confidential variable LSTAT was 7.99 (variance of LSTAT is 27.38).

• Compared to just using the values in the perturbed data, the MSE of the posterior point estimates were substantially reduced
  – Average difference in MSE for LSTAT was 2.18.
Results (all $\lambda$)

- For all $\lambda$, posterior point estimates shrunk towards the mean for each variable with “shrinkage” increasing with $\lambda$.
- In general, MSE of point estimates increased with $\lambda$ in concave fashion as expected.
- Outliers:
  - 12\textsuperscript{th} observation has an unusually high value for LSTAT and posterior point estimates for this tended to be biased towards the mean.
  - MSE of posterior point estimates for this observation was substantially higher than estimates from the perturbed data for $\lambda = 0.2$ and 0.4. For $\lambda=0.6$, the MSE was slightly lower.
Plot illustrating the disclosure protection increasing with $\lambda$ and the marginal protection diminishing.
Summary

• Motivating principles
• Random Orthogonal Matrix Masking (ROMM)
  – Preserves means and covariance matrix.
  – Includes “statistical obfuscation” as a special case
• Numerical implementation
  – Comparison with additive noise: more variability (protection) but superior regression estimates.
• Risk-Utility tradeoff
• Preliminary Simulation
  – Can use RU idea to pick a suitable value of $\lambda$.  


The End