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 $\sum_x$ be the sample mean and sample covariance matrix of the original microdata and let $\bar{y}$ and $\sum_y$ be the corresponding quantities in the masked microdata produced by ROMM. Then $\bar{x} = \bar{y}$ and $\sum_x = \sum_y$.**

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 occured. 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

Obs	RM	PTRATIO	LSTAT	MEDV
1	6.630	18.5	6.53	26.6
2	5.986	19.1	14.81	21.4
3	5.709	14.7	15.79	19.4
4	5.877	14.7	12.14	23.8
5	6.402	14.7	11.32	22.3
6	6.782	15.2	6.68	32.0
7	6.433	19.1	9.52	24.5

Obs	RM	PTRATIO	LSTAT	MEDV
8	6.315	16.6	7.60	22.3
9	6.023	18.4	11.72	19.4
10	6.251	20.2	14.19	19.9
11	5.757	20.2	10.11	15.0
12	5.304	20.2	26.64	10.4
13	6.425	20.2	12.03	16.1

- **Comparison of:**
  - ROMM, co-ordinate by co-ordinate, with λ=1/3.
  - Comparison with bias corrected correlated additive noise method of Kim, with $\sqrt{c}$= 1/2.
- **Regression of MEDV on other variables.**

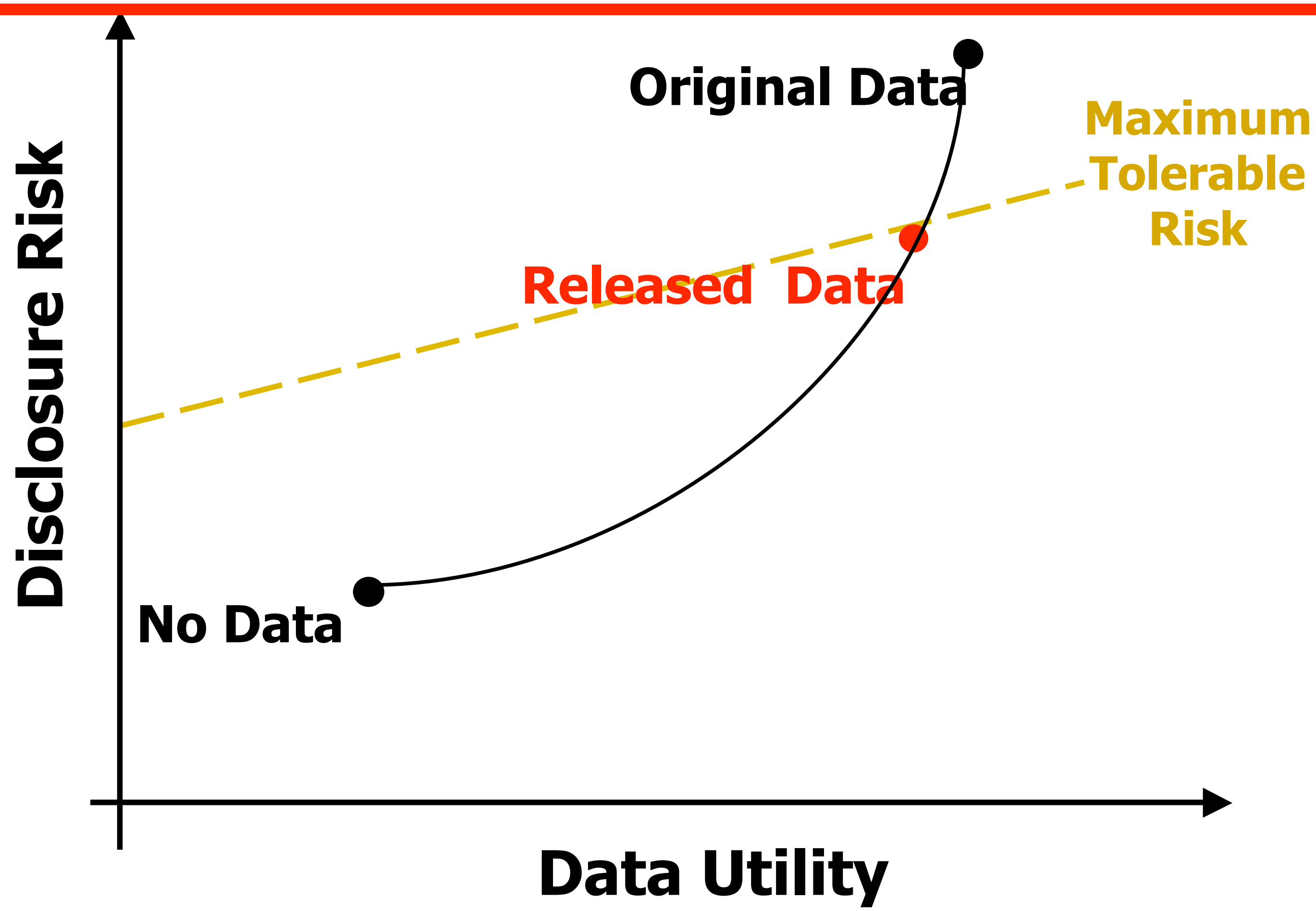
Example Results

ROMM				Additive Noise			
RM	PTRATIO	LSTAT	MEDV	RM	PTRATIO	LSTAT	MEDV
-0.253	3.725	4.721	-7.681	-0.337	1.509	1.848	-2.173
-0.529	2.006	3.843	-9.182	0.182	2.249	-1.479	-1.238
0.494	0.738	-12.337	5.930	-0.036	2.335	-0.761	-3.027
0.045	0.880	1.249	-0.095	0.235	-0.132	-3.451	0.908
0.223	1.313	-1.086	3.740	-0.057	1.274	3.057	-3.099
-0.183	1.841	0.338	-3.079	-0.324	0.380	1.999	-3.797
-0.269	2.093	1.883	-4.906	-0.281	-1.957	3.303	0.980
0.139	-0.867	0.348	3.018	0.175	-0.046	-0.831	0.765
0.414	0.367	-1.688	0.357	0.053	-1.423	-2.049	1.165
0.212	-3.778	-5.139	4.602	0.086	-0.116	0.367	0.883
-0.437	-3.218	10.437	-2.359	-0.107	0.195	2.658	0.143
0.601	-2.851	-6.872	12.014	0.390	-0.581	-8.411	4.243
-0.457	-2.249	4.302	-2.361	-0.468	0.254	6.544	-5.830

Variances	RM	PTRATIO	LSTAT	MEDV
ROMM	0.146	5.618	33.307	34.327
AddNoise	0.066	1.641	13.950	7.348

Original Data (ROMM)			Additive Noise		
Variable	Estimate	SE	Variable	Estimate	SE
(intercept)	-5.5641	23.6617	(Intercept)	-2.4696	25.1829
RM	7.4488	3.3663	RM	6.2655	3.5843
PTRATIO	-0.9557	0.3691	PTRATIO	-0.3930	0.4329
LSTAT	-0.1770	0.2741	LSTAT	-0.6780	0.2995

R-U Confidentiality Map



(Duncan, et al. 2004)

Risk Utility Tradeoff

- **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 λ.**
- ***Method*: Acceptance-rejection sampling to sample from posterior.**

Resulting Data

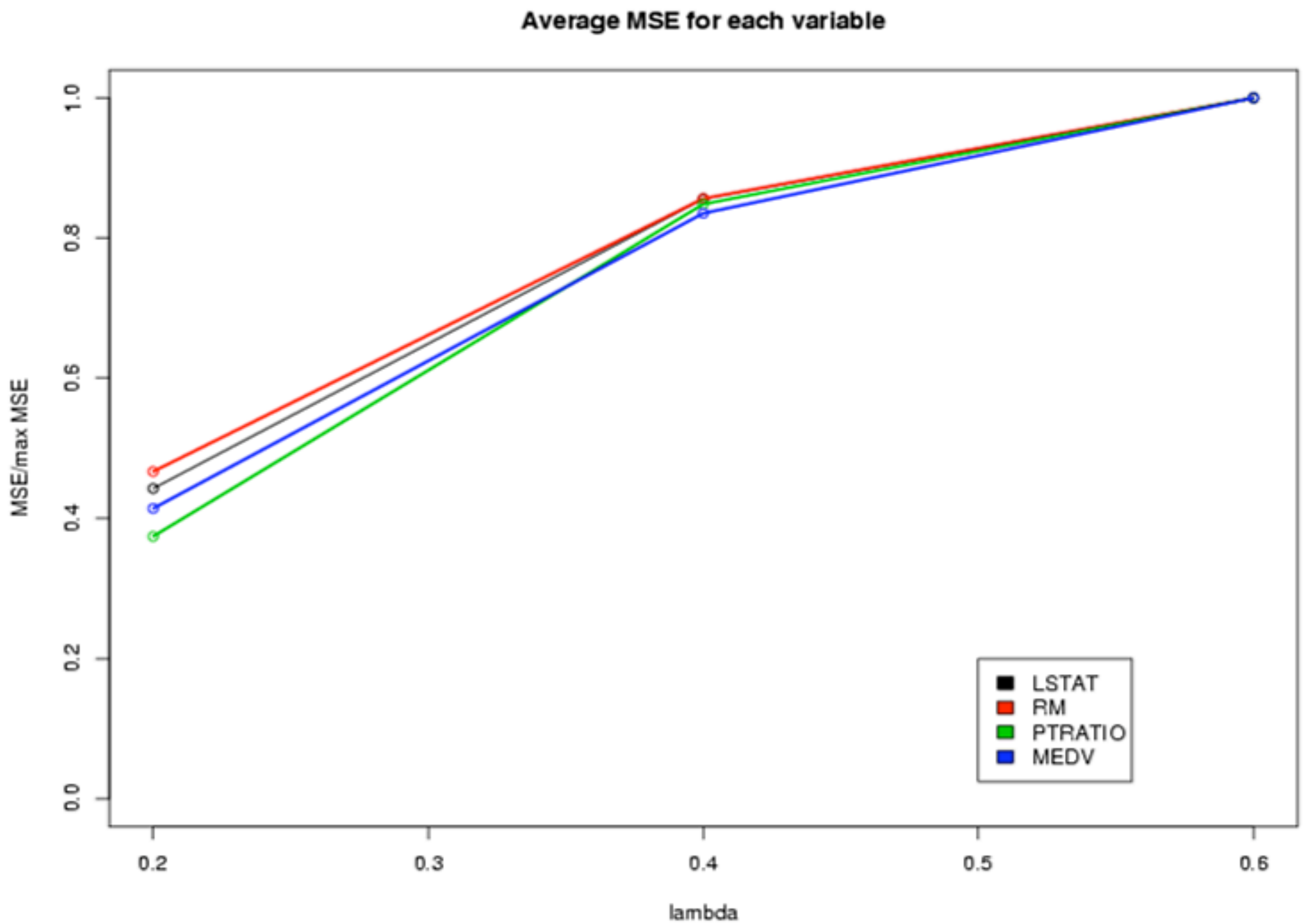
- **For λ=0.2, 0.4, and 0.6, we generated a sequence of N=80 (perturbed data, intruder data) pairs and sampled from the posterior of the original data given each pair.**
- **We then calculated point estimates for each "individual."**
- **We then looked at the behavior of the distribution of these point estimates which allows us to evaluate risk.**

Results (small λ)

- For small λ (i.e., λ =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 λ)

- **For all λ, posterior point estimates shrunk towards the mean for each variable with "shrinkage" increasing with λ.**
- **In general, MSE of point estimates increased with λ in concave fashion as expected.**
- **Outliers:**
  - **12th 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 λ = 0.2 and 0.4. For λ=0.6, the MSE was slightly lower.**



Plot illustrating the disclosure protection increasing with λ 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 λ.**

The End


Google 2084
Google Search
I'm Feeling Lucky
I'm Feeling Paranoid
Your Brain
Your Home
Family
Friends
Ex-friends
Relatives
Co-workers
Ex-spouse(s)
Enemies
Satellite Photos of People You Want to Spy On
Satellite Photos of People Spying on You
Credit Reports
Tax Records
Phone Records
Court Documents
Other People's Conversations
Books
Movies
TV Shows
Music
Pornography
Your Past
Your Present
Your Future