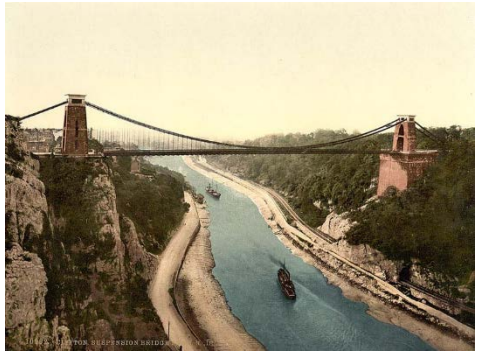
Anonymisation of Public Use Data Sets

Methods for Reducing Disclosure Risk and the Analysis of Perturbed Data

Harvey Goldstein
University of Bristol and University College London
and
Natalie Shlomo
University of Manchester



The problem and some solutions

- Release of large (pseudonymised) datasets for analysis potentially allows ‘statistical attack’ via searching for records satisfying certain constraints (e.g. age, location, medication..)
- Standard solution is to degrade data values to make it ‘unlikely’ that an attacker could correctly identify individuals. Typically judge using k-anonymity
- Two types of disclosure control methods under ‘safe data’ approach
    1. Non-perturbative methods – reduce information content
    2. Perturbative methods – alters the data to increase uncertainty of identification

The problem and some solutions

- Non-perturbative methods:
  1. Remove ‘cells’ with small counts if data in tabular form, preserving margins
  2. Delete sensitive variables
  3. Group categories or categorise continuous variables of disclosive variables such as postcode, age….
  4. Sub-sample
- Perturbative methods:
  1. Add random ‘noise’ to increase uncertainty around correct identification (this includes random misclassification for categorical variables)
  2. micro-aggregation of similar cases (effectively reduces variation)
  3. Create ‘synthetic’ data values while preserving data structure

Effects on statistical analysis – a key concern

1) Cell removal: may over - coarsen data and in particular remove interesting interaction effects

2) Grouping: like (1) may smooth over complex relationships

3) Addition of random noise will lead to incorrect standard errors and also biased coefficients in generalised linear models unless properly adjusted for

4) Synthetic data may lead to severely biased coefficients if analysis models do not include variables used in the synthesis

Synthetic data

- Synthetic data: relies on assumed or modelled data relationships to simulate (impute) new data that approximates real data. This can be done for all data or a subset.
- Producing multiply imputed datasets allows corrections to be made for imputation variance or approximations available.
- Few would advocate that such data should be used for a final analysis: rather they can provide an indication for a small set of final models that can then simply be fitted (in a secure environment) to produce required model estimates.

Synthetic data

- This poses particular problems:
  1. There is a strong reliance on producing the right structure, typically via a series of conditional models.
  2. Even using synthetic data in ‘exploratory mode’ can lead users astray, where their models based upon an approximation to the true structure become biased, and lead to the selection of inappropriate final models to be estimated using the real data.

Adding random noise in general

- Adding random 'noise' is less extreme than synthetic data.
- We suppose that the attacker has available a set of *q* values for y (the variables to be used), say y^* that she intends to match against records in the data set. We propose to construct a new set of variables, *z*, which is what the attacker will see

$$z = y + m$$

where *m* has a predefined (normal) distribution (other distributions are possible e.g. differential privacy techniques often use a double exponential distribution)

- For simplicity, assume independence across variables to be 'disturbed' or we might consider the case of 'correlated noise' (correlated with true values) to preserve the correlation structure and sufficient statistics . Note that *y* can be continuous or discrete (categories numbered 1,…,p)

Adding random noise in general

- The value of the variance (σ_m^2) will determine the strength of the resistance to attack and can be a function of the 'true' variability of each variable.
- We now form a measure of the distance between the y^* and each *z* and then rank these distances.
- A general distance measure can be written in the form

  $D^* \propto (z - y^*)^T W (z - y^*)$, where, for example, $W^{-1} \propto \Omega_k$
- But, more simply we can choose the Euclidean distance for each comparison record *i*

$$D_i^* = \sum_{j=1}^{q} (z_{ij} - y_j^*)^2, \quad D_i = \sum_{j=1}^{q} (z_{ij} - y_j)^2, \quad i = 1, \ldots, n$$

Ranking the distances

- A 'rational' attacker chooses closest record(s) to their own as the correct one(s).
- Form $R_i^* = Rank(D_i^*), \quad R_i = Rank(D_i),$
- Define $i^* = \text{value of } i \text{ for } R_i^* = 1.$
- Define $h = R_{i^*} - 1,$ Thus if *h*=0 we have the correct match.
- For example:

i	R_i^*	R_i
1	3	2
2	1	3
3	2	1

$i^* = 2, \text{ so } h = 3 - 1 = 2$ if attacker chooses closest record.

h measures difference between chosen and 'correct' method

So choose noise added large enough so that, say, $\Pr(h < p) < \epsilon$ (say, $p = 3, \epsilon = 0.1$)

A simulation

- Generate 10^3 records with 5 normal variables and $\sigma_m^2 = 0.1$
- All variances =1 and covariances = 0.25.
- For each true value record (attacker's *y**) generate D, D^*
- The following table gives some estimates of disclosiveness in terms of *h* for a range of individuals at different distances from the median.

Distribution for *h*

$h \ll$	Cumulative percentile of *D* distribution				
	10	**20**	**30**	**40**	**50**
0	52.2	49.4	43.9	41.3	41.7
1	62.9	60.7	56.1	53.1	53.1
2	70.0	65.3	62.0	61.2	60.8
3	74.7	70.2	68.6	66.0	65.8
4	78.5	74.4	72.8	70.1	68.7
5	80.8	77.5	76.5	72.7	71.5

More results

Lowest decile Pr(h>5). For combinations of Ω and σ_m^2 where Ω always has unit diagonal elements and equal off-diagonal elements (given by columns 0.1 – 0.5) are shown. Sample size =1000.					
σ_m^2.	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**
0.1	0.15	0.16	0.19	0.23	0.24
0.2	0.45	0.43	0.46	0.50	0.54
0.3	0.58	0.63	0.63	0.65	0.70
0.4	0.73	0.74	0.74	0.76	0.77

We see that the procedure is readily 'tuned' simply by changing the variance of the noise.

We are also studying the possibility of a more sophisticated attack that uses vales of *y* predicted from the perturbed dataset rather than the *z* themselves.

The h-index and *k*-anonymisation

- If we have, say, 2-anonymity this implies that an attacker is able to identify two individual records matching her own information, so choosing either of them at random means that there is a probability of 0.5 that it is the correct one.
- The *h*-index, however, only yields a single individual as the closest, for example with a probability about 0.5 and thus provides less information to the attacker than in the case of 2-anonymity.

The h-index and *k*-anonymization II

- For k-anonymity an attacker may be quite content that they can access 2 or perhaps even 5 records containing the one that is sought.
- By contrast, with the *h*-index procedure, in our most favourable case, the probability of the sought-for individual being one of the two nearest is just over 60% and one of the five nearest just under 80%
- Thus it could be argued that this is sufficient to deter an attacker and hence suitable in terms of disclosiveness. In practice careful attention needs to be paid to the amount of noise required to satisfy disclosure concerns.

How to remove the noise

Assume noise $\eta_i \sim iid(0, \sigma_\eta^2)$ add to continuous variable

We get unbiased totals and means but larger variance and biases where predictors incorporate noise

How to make correct inferences in a general modelling framework?

Assume a simple regression model with a dependent variable y_i that has been subjected to Gaussian additive noise η_i with a mean of 0 and a positive variance σ_η^2

The predictor variable x_i is error free we assume.

15

How to remove the noise

The model is:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, & i = 1,\ldots,n \\ y_i = y_i^* + \eta_i \end{cases}$$

where y_i^* denotes the true but unobserved value of the dependent variable y_i

If we regress y_i on x_i then

$$\beta = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(y^* + \eta, x)}{Var(x)} = \frac{Cov(y^*,x) + Cov(\eta,x)}{Var(x)} = \frac{Cov(y^*,x)}{Var(x)}$$

since $Cov(\eta, x) = 0$

16

How to remove the noise

Additive noise on the dependent variable thus does not bias slope coefficient but increases standard errors due to the increase in variance

$$Var(y) = Var(y^*) + Var(\eta)$$

Now add noise η_i to predictor variable

The model is now:

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, & i = 1, \ldots, n \\ x_i = x_i^* + \eta_i \end{cases}$$

where x_i^* denotes true but unobserved value of x_i

How to remove the noise

If we regress y_i on x_i then for the least squares scope coefficient:

$$\beta = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(y,x^* + \eta)}{Var(x^*) + Var(\eta)} = \frac{Cov(y,x^*) + Cov(y,\eta)}{Var(x^*) + Var(\eta)} = \frac{Cov(y,x^*)}{Var(x^*) + Var(\eta)}$$

since $Cov(y,\eta) = 0$

Additive noise on predictor variable biases slope coefficient downwards (attenuation)

Thus we need suitable methodology to deal with these ‘measurement errors’

How to remove the noise

For the lease squares slope coefficient in a simple linear regression:

$$\hat{\beta} \xrightarrow{p} \frac{Cov(y, x^*)}{Var(x^*) + Var(\eta)} = \frac{\beta \sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\eta}} = \beta(1 + \sigma^2_{\eta} / \sigma^2_{x^*})^{-1}$$

We define $\lambda = (1 + \sigma^2_{\eta} / \sigma^2_{x^*})^{-1}$ as the reliability ratio

A consistent estimate of the slope coefficient is obtained by dividing least squares estimate by λ

To calculate λ we assumes that σ^2_{η} is released and known to the researcher.

19

How to remove the noise in general

- Noise is random with known properties so a measurement error model is required
- This requires that the parameters used to generate the noise are known to the researchers.
- Current work (using a CLOSER grant at Bristol) is underway to develop software to show how the noise should be generated in such a way that the parameters can be released under a predetermined *h*-index to protect against attribute disclosure whilst preserving utility

How to remove the noise

- In simple linear regression, ‘correlated noise’ can be added which produce unbiased estimates of slope coefficients by using standard regression techniques.
- Current work at Bristol is developing algorithms incorporating measurement error models that will handle generalized linear models and multilevel data of different types.
- Specialisation to anonymisation with user software currently funded through ESRC (via Closer) at Bristol (Boyd, Goldstein and Burton)
- Can be combined with handling missing data values.
- Some loss of statistical efficiency but enables underlying ‘signal’ to be extracted and thus provides unbiased parameter estimates.

Further thoughts

- Often, a data attacker will have no pre-existing individual data and may trawl the dataset to discover an ‘interesting’ record, for example an individual with an unusual combination of values. They may then attempt to identify the real person using other variables in the data record. Our procedure is also relevant to such an attack so long as the noise has been applied to the variables in question.
- How to ‘tune’ the noise and differential noise related to ‘identifiability’ of variables is an area for further research. For example we might wish to add relatively more noise to a variable such as ‘height’ than ‘hair colour’.
- Now, it may well be the case that, conditional on the data available to the attacker, a variable such as income can be predicted with sufficient accuracy *within this dataset,* and if the data structure is well approximated – either by removing noise or via synthesis then income could be fairly accurately predicted and this may be sufficient for an attacker’s purpose. Needs further consideration.
- Provision of suitable analysis tools and training for data analysts is important – discussions are underway with Government departments and agencies through ADRN.

Thank you for your attention