Overview of synthetic data

Jingchen (Monika) Hu

Vassar College

Statistical Data Privacy

Outline

Outline

What are synthetic data?

- To provide privacy protection of individuals in datasets
- Usually created by simulating variables of records from statistical models estimated on the confidential data
- Objective: preserve data integrity (e.g., important characteristics in the confidential data, such as means and correlations of variables)

What are synthetic data?

- To provide privacy protection of individuals in datasets
- Usually created by simulating variables of records from statistical models estimated on the confidential data
- Objective: preserve data integrity (e.g., important characteristics in the confidential data, such as means and correlations of variables)

- To do so, we start with developing suitable statistical models for the confidential data

How are synthetic created?

- If the developed statistical models are appropriate and model estimation is done properly, the estimated models can capture important characteristics in the confidential data

How are synthetic created?

- If the developed statistical models are appropriate and model estimation is done properly, the estimated models can capture important characteristics in the confidential data
- And synthetic records are simulated from these estimated models
- Then, these synthetic records could potentially preserve important features in the confidential data

How are synthetic created?

- If the developed statistical models are appropriate and model estimation is done properly, the estimated models can capture important characteristics in the confidential data

- And synthetic records are simulated from these estimated models
- Then, these synthetic records could potentially preserve important features in the confidential data

- Moreover, they can provide some levels of privacy protection, as compared to releasing the confidential data

What we will do in this lecture

- We will go over important aspects of any synthetic data approach
  - Two flavors of synthetic data: partial synthesis and full synthesis
  - Two general approaches to synthetic data creation: sequential synthesis and joint synthesis
  - Two aspects of evaluation: utility and disclosure risks

What we will do in this lecture

- We will go over important aspects of any synthetic data approach
  - Two flavors of synthetic data: partial synthesis and full synthesis
  - Two general approaches to synthetic data creation: sequential synthesis and joint synthesis
  - Two aspects of evaluation: utility and disclosure risks
- Our course focuses on Bayesian synthesis models (Lectures 3 and 4)
- There also exist non-Bayesian data synthesis models (e.g., the `synthpop` R package)

Outline

Partial synthesis

- Proposed by Little (1993)
- Some variables in the collected dataset, such as sensitive variables and key identifiers, are synthesized
- The resulting synthetic data contain sensitive variables with synthesized values while other variables remain unchanged

Full synthesis

- Proposed by Rubin (1993)
- A synthetic population is first simulated
- Then a synthetic sample is selected from the synthetic population
- The resulting synthetic data have every variable synthesized, contain no records from the confidential data, and it may even have a different sample size than the confidential data if needed

Full synthesis

- Proposed by Rubin (1993)
- A synthetic population is first simulated
- Then a synthetic sample is selected from the synthetic population
- The resulting synthetic data have every variable synthesized, contain no records from the confidential data, and it may even have a different sample size than the confidential data if needed

- One can also create fully synthetic data following the partial synthesis approach, i.e., only working on the sample
  - This approach is actually more widely used when creating fully synthetic data

Comparisons

- The choice depends on data disseminator's protection goals
- Assuming a quality synthesis
  - Utility: higher in partially synthetic data
  - Disclosure risks: higher in partially synthetic data
- Utility-risk trade-off

Example: SynLBD

SynLBD variable description. Taken from Table 1 in Kinney et al. (2011) with some modifications.

Name	Type	Notation	Action
ID	Identifier		Created
County	Categorical	x_1	Not released
SIC	Categorical	x_2	Not to synthesize
Firstyear	Categorical	y_1	To synthesize
Lastyear	Categorical	y_2	To synthesize
Year	Categorical		Created
Multiunit	Categorical	y_3	To synthesize
Employment	Continuous	y_4	To synthesize
Payroll	Continuous	y_5	To synthesize

Example: SynLBD

SynLBD variable description. Taken from Table 1 in Kinney et al. (2011) with some modifications.

Name	Type	Notation	Action
ID	Identifier		Created
County	Categorical	x_1	Not released
SIC	Categorical	x_2	Not to synthesize
Firstyear	Categorical	y_1	To synthesize
Lastyear	Categorical	y_2	To synthesize
Year	Categorical		Created
Multiunit	Categorical	y_3	To synthesize
Employment	Continuous	y_4	To synthesize
Payroll	Continuous	y_5	To synthesize

Partial synthesis or full synthesis?

Outline

The joint distribution of synthesized variables

- This classification is based on what strategy is used to estimate the joint distribution of the variables to be synthesized
- Variables to be synthesized: $\{y_1, y_2, y_3\}$
- Insensitive variables: $\{x_1, x_2\}$
- The joint distribution of synthesized variables

$$f(y_1, y_2, y_3 \mid x_1, x_2) \tag{1}$$

The joint distribution of synthesized variables

- This classification is based on what strategy is used to estimate the joint distribution of the variables to be synthesized
- Variables to be synthesized: $\{y_1, y_2, y_3\}$
- Insensitive variables: $\{x_1, x_2\}$
- The joint distribution of synthesized variables

$$f(y_1, y_2, y_3 \mid x_1, x_2) \tag{1}$$

- How to estimate this joint distribution? Two general approaches

Sequential synthesis overview

- This approach specifies a sequence of univariate synthesis models for the sensitive variables
- This sequence ultimately gives a joint model of all sensitive variables

Sequential synthesis overview

- This approach specifies a sequence of univariate synthesis models for the sensitive variables
- This sequence ultimately gives a joint model of all sensitive variables

- Variables to be synthesized: $\{y_1, y_2, \cdots, y_{p_1}\}$
- Insensitive variables: $\{x_1, x_2, \cdots, x_{p_2}\}$
- The joint model can be expressed in a sequence of univariate model as:

$$
\begin{aligned}
f(y_1, \cdots, y_{p_1} \mid x_1, \cdots, x_{p_2}) = \quad & f(y_1 \mid x_1, \cdots, x_{p_2}) \times \\
& f(y_2 \mid y_1, x_1, \cdots, x_{p_2}) \times \\
& \cdots \\
& f(y_{(p_1-1)} \mid y_1, \cdots, y_{(p_1-2)}, x_1, \cdots, x_{p_2}) \\
& f(y_{p_1} \mid y_1, \cdots, y_{(p_1-1)}, x_1, \cdots, x_{p_2})
\end{aligned}
$$

Sequential synthesis procedure

1. Specify a synthesis model for $y_1 \mid x_1, \cdots, x_{p_2}$. Estimate this model on the confidential data, and generate synthetic y_1^* using confidential $(x_1, \cdots, x_{p_2})$.
2. Specify a synthesis model for $y_2 \mid y_1, x_1, \cdots, x_{p_2}$. Estimate this model on the confidential data, and generate synthetic y_2^* using synthetic y_1^* from step 1 and confidential $(x_1, \cdots, x_{p_2})$.
3. Repeat Step 2 for each of the variables of $\{y_3, \cdots, y_{(p_1-1)}\}$.
4. Specify a synthesis model for $y_{p_1} \mid y_1, \cdots, y_{(p_1-1)}, x_1, \cdots, x_{p_2}$. Estimate this model on the confidential data, and generate synthetic $y_{p_1}^*$ using synthetic $(y_1^*, y_2^*, \cdots, y_{(p_1-1)}^*)$ from previous steps and confidential $(x_1, \cdots, x_{p_2})$.

Joint synthesis

- The joint distribution: $f(y_1, \cdots, y_{p_1} \mid x_1, \cdots, x_{p_2})$
- Directly specify a joint model for these sensitive variables
- For example, if $\{y_1, y_2, \cdots, y_{p_1}\}$ are all continuous (and marginally normal after transformation), we can use a multivariate normal distribution:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{p_1} \end{bmatrix} \sim \mathrm{MVN}_{p_1}\left(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{p_1} \end{bmatrix} \Sigma\right), \tag{2}$$

  - MVN_{p_1} stands for multivariate normal distribution of dimension p_1
  - $\mu_1, \cdots, \mu_{p_1}$ are the mean parameters (conditional on $x_1, \cdots, x_{p_2}$)
  - Σ is covariance matrix

Joint synthesis cont'd

- If sensitive variables are all categorical. . .
- A well research model is the Dirichlet Process mixture of products of multinomials (DPMPM) (Hu, Reiter, and Wang (2014)); we will introduce it in Lecture 4
- Joint synthesis model estimation is usually more challenging than sequential synthesis

Joint synthesis cont'd

- If sensitive variables are all categorical. . .
- A well research model is the Dirichlet Process mixture of products of multinomials (DPMPM) (Hu, Reiter, and Wang (2014)); we will introduce it in Lecture 4
- Joint synthesis model estimation is usually more challenging than sequential synthesis
- Bayesian networks are good approaches (Young, Graham, and Penny (2009), Kaur et al. (2021)); this could be an interesting project

Example: SynLBD

- Details are in Kinney et al. (2011) Section 3
- The SynLBD uses the sequential synthesis approach

Example: SynLBD

- Details are in Kinney et al. (2011) Section 3
- The SynLBD uses the sequential synthesis approach

The sequential synthesis procedure for the SynLBD follows the workflow below:

1. Synthesize `Firstyear` using the Dirichlet-multinomial approach and draw from the following estimated model to obtain y_1^*
$$f(y_1 \mid x_1, x_2). \tag{3}$$
2. Synthesize `Lastyear` using the Dirichlet-multinomial approach and approximate a draw from the following estimated model to obtain y_2^*
$$f(y_2 \mid y_1, x_1, x_2). \tag{4}$$

Example: SynLBD

3. Synthesize `Multiunit` using the Dirichlet-multinomial approach and approximate a draw from the following estimated model to obtain y_3^*
$$f(y_3 \mid y_2, y_1, x_1, x_2). \tag{5}$$
4. Synthesize `Employment` using the normal approach and approximate a draw from the following estimated model to obtain $y_4^{(t)*}$
$$f(y_4^{(t)} \mid y_4^{(t-1)}, y_3, y_2, y_1, x_1, x_2). \tag{6}$$
5. Synthesize `Payroll` using the normal approach and approximate a draw from the following estimated model to obtain $y_5^{(t)*}$
$$f(y_5^{(t)} \mid y_4^{(t)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2). \tag{7}$$

Outline

Utility evaluation

- Two general types of utility: global and analysis-specific

1. Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
2. Analysis-specific utility: Evaluate whether synthetic data users can obtain statistical inferences on the synthetic data that are similar to those obtained on the confidential data

Utility evaluation

- Two general types of utility: global and analysis-specific

1. Global utility: Evaluate the closeness between the confidential data distribution and the synthetic data distribution
2. Analysis-specific utility: Evaluate whether synthetic data users can obtain statistical inferences on the synthetic data that are similar to those obtained on the confidential data

- How to capture the uncertainty in the synthetic data generation process? Create multiple synthetic datasets

Example: SynLBD

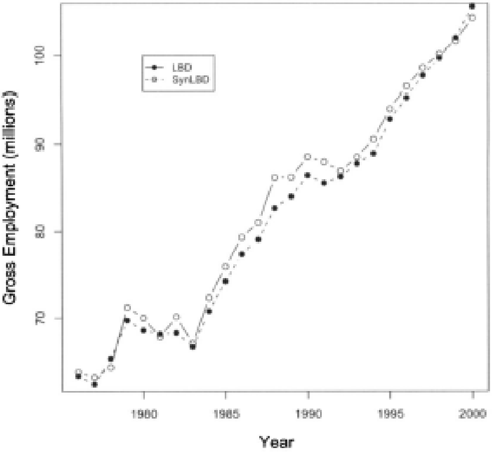


Figure 1. *Gross employment level by year, LBD versus Synthetic.*

Global utility or analysis-specific utility?

Example: SynLBD

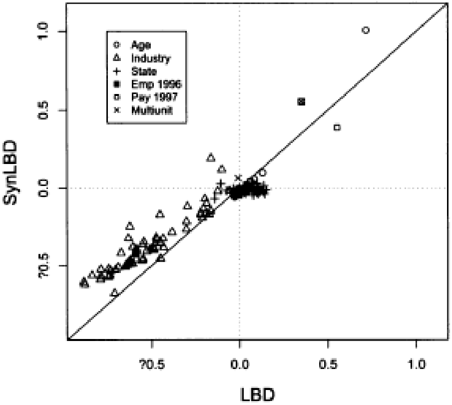


Figure 11. *Regression coefficients, LBD versus Synthetic.*

Global utility or analysis-specific utility?

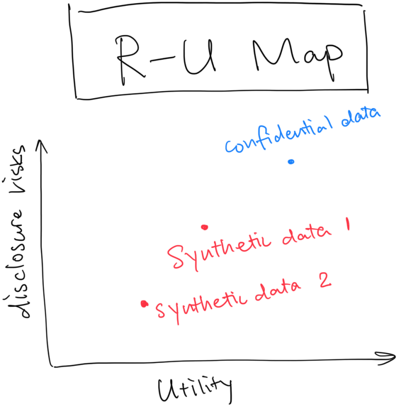
Disclosure risks evaluation

- Assuming the intruder has access to external data, two common disclosure risks: identification and attribute (Hu (2019))

1. Identification disclosure: The intruder correctly identifies records of interest in the released synthetic data
2. Attribute disclosure: The intruder correctly infers the true confidential values of the synthetic records using information from the released synthetic data

Utility-risk trade-off

- Ideally, the released synthetic data have high utility and low disclosure risks
- However this is usually not the case, due to the utility-risk trade-off (Duncan, Keller-McNulty, and Stokes (2001))



Outline

Summary

- Synthetic data
  - Partial synthesis and full synthesis
  - Sequential synthesis and joint synthesis
  - Evaluations of synthetic data: utility and disclosure risks
  - The example of SynLBD

Summary

- Synthetic data
  - Partial synthesis and full synthesis
  - Sequential synthesis and joint synthesis
  - Evaluations of synthetic data: utility and disclosure risks
  - The example of SynLBD
- Homework 1: Read Ros, Olsson, and Hu (2020) and answer a few questions regarding the discussed aspects of synthetic data
  - Submission on Moodle and prepare to discuss next time

Summary

- Synthetic data
  - Partial synthesis and full synthesis
  - Sequential synthesis and joint synthesis
  - Evaluations of synthetic data: utility and disclosure risks
  - The example of SynLBD
- Homework 1: Read Ros, Olsson, and Hu (2020) and answer a few questions regarding the discussed aspects of synthetic data
  - Submission on Moodle and prepare to discuss next time
- Lecture 2: Introduction to Bayesian modeling
  - Chapter 7 of Albert and Hu (2019): https://bayesball.github.io/BOOK/proportion.html

References I

Albert, J., and J. Hu. 2019. Probability and Bayesian Modeling. Texts in Statistical Science, Chapman Hall CRC Press.

Duncan, G. T., S. A. Keller-McNulty, and S. L. Stokes. 2001. "Disclosure Risk Vs Data Utility: The R-U Confidentiality Map." National Institute of Statistical Sciences.

Hu, J. 2019. "Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data." Transactions on Data Privacy 12 (1): 61–89.

Hu, J., J. P. Reiter, and Q. Wang. 2014. "Disclosure Risk Evaluation for Fully Synthetic Categorical Data." Privacy in Statistical Databases, 185–99.

References II

Kaur, D., M. Sobiesk, S. Patil, J. Liu, P. Bhagat, A. Gupta, and N. Markuzon. 2021. "Application of Bayesian Networks to Generate Synthetic Health Data." Journal of the American Medical Informatics Association 28 (4): 801–11.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." International Statistical Review 79 (3): 362–84.

Little, R. J. A. 1993. "Statistical Analysis of Masked Data." Journal of Official Statistics 9: 407–26.

Ros, K., H. Olsson, and J. Hu. 2020. "Two-Phase Data Synthesis for Income: An Application to the Nhis." Privacy in Statistical Databases (E-Proceedings).

References III

Rubin, D. B. 1993. "Discussion Statistical Disclosure Limitation." Journal of Official Statistics 9: 461–68.

Young, J., P. Graham, and R. Penny. 2009. "Using Bayesian Networks to Create Synthetic Data." Journal of Official Statistics 25: 549–67.