Creating synthetic data with `mice`

Volker, Vink, Van Buuren

2022-07-13

What is synthetic data

"The concept of synthetic data generation is to take an original data source (dataset) and create new, artificial data, with similar statistical properties from it."
- EUROPEAN DATA PROTECTION SUPERVISOR

- Synthetic data is fake data, generated from some (statistical) model.
- https://this-person-does-not-exist.com

Why do we want synthetic data?

Access to actual data is often restricted!

- Do research without sharing data of real people.

But also:

- Can save time and effort.
- What if the data would have looked differently?

Use cases

- Governments making synthetic data available (US Census, DUO).
- Researchers enhancing reproducibility by sharing code that runs on synthetic data (see here).
- Teaching purposes.
- Getting to know characteristics of the data.
- Allows to write analysis scripts before having actual data.

Privacy-utility trade-off

Privacy

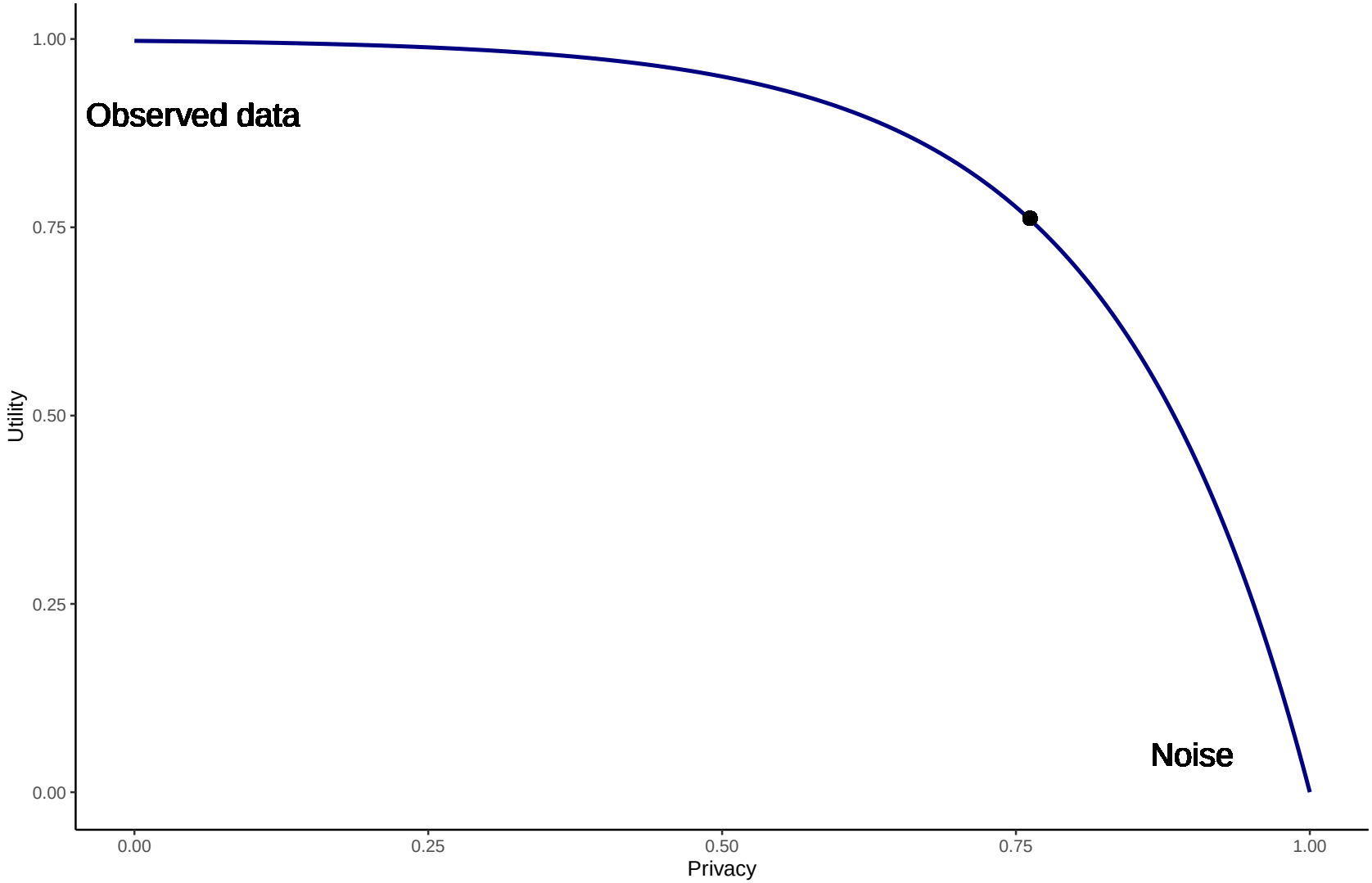
- Can we reproduce the original data on the basis of the synthetic data?
- Can we infer whether someone was part of the original data?
- Open question: how do we quantify this?

Utility

- Same estimates (univariate, multivariate (relationships))?
- Do identical analyses give identical results (as compared to observed data)?

Privacy-utility trade-off

Privacy decreases as the utility increases.



Partially versus fully synthetic data

Fully synthetic data

- Impute (and sample from) the population (extends beyond the sample).

Partially synthetic data

- Overimpute (part of) the sample at hand.
- Approach in `mice` (see here).

Creating synthetic data

The `where`-matrix

```
where <- make.where(boys, "all")
head(where, 4)
```

```
##      age  hgt  wgt  bmi   hc  gen  phb   tv  reg
## 3   TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## 4   TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## 18 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## 23 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Creating synthetic data

Imputing

```
syn <- mice(boys,
            m = 5,
            maxit = 1,
            where = where,
            method = method,
            predictorMatrix = pred)
```

Note: missing entries in the observed data complicate the synthesis (see, e.g., Drechsler)

Analyzing synthetic data

Different pooling rules!

Between-imputation variance is less important (if you have no missings).

```
syn_fit <- syn %$% lm(DV ~ IV_1 + ... + IV_k)

pool(syn_fit, rule = "reiter2003")
```

But, be cautious

The utility of the synthetic data is much easier to evaluate than the level of privacy protection.

Synthetic data might contain actual observations.

Work in progress!