gdpR: An R Package for studying differentially private algorithms

by Jordan A. Awan, Kevin Eng, Robin Gong, Nianqiao Phyllis Ju, and Vinayak A. Rao

Abstract This paper serves as a reference and introduction on using the *gdpR* R package. The goal of this package is to provide some tools for exploring the impact of different privacy regimes on a Bayesian analysis. A strength of this framework is the ability to target the exact posterior in settings where the likelihood is too complex to analytically express.

1 Introduction

The ease and pervasiveness of modern data collection technologies has raised concerns about data privacy. (Dwork and Roth 2013) introduced the differential privacy framework as a means to rigorously define privacy. The framework has lead to the development of many "privitized'' versions of existing statistical methods. The process of privitizing usually consist of introducing random noise in someway using a known distribution.

2 Using gdpR

Sampling

The main function in **gplsim** is the gdp_sample function. The call signature of the function is:

```
gdp_sample(data_model, sdp, nobs, init_par, niter = 2000, warmup = floor(niter / 2),
           chains = 1, varnames = NULL)
```

The three required inputs into gdp_sample function are the privacy model (data_model), the value of the observed privatized statistic (sdp), and the total number of observations in the complete data (nobs) [MAKE SURE NOTATION IS INTRODUCED]. The gdpR package is best suited for problems where the complete data can be represented in tabular form. This is because internally, it is represented as a matrix.

The optional arguments are the number of mcmc draws (niter), the burn in period (warmup), number of chains (chains) and character vector that names the parameters. Running multiple chains can be done in parallel using the furrr package. Additionally, progress can be monitored using the progressr package.

The data_model input is a privacy object that can be constructed using the new_privacy constructor. The process of constructing a privacy object will be discussed in the next section.

Privacy Model

Creating a privacy model is done using the new_privacy constructor. The main arguments consist of the four components as outlined in the methodology section.

```
new_privacy(post_smpl = NULL, lik_smpl = NULL, ll_priv_mech = NULL,
            st_calc = NULL, add = FALSE, npar = NULL)
```

The internal implementation of the DA algorithm in gdp_sample requires some care in how each component in constructed.

- lik_smpl is an R function that samples from the likelihood. Its call signature should be lik_smpl(theta) where theta is a vector representing the likelihood model parameters being estimated. This function must work with the supplied initial parameter provide in the init_par argument of gdp_sample. The sampler need not be vectorized and vectorizing the sampler will not add any speed benefits.
- post_smpl is a function which represents the posterior sampler. It should have the call signature post_smpl(dmat, theta). Where dmat is the complete data. This sampler can be generated by wrapping mcmc samplers generated from other R packages (e.g. rstan, fmcmc, adaptMCMC).

> If using this approach, it is recommended to avoid using packages such as **mcmc** whose implementation clashes with `gdp_sample`. In the case of **mcmc**, the Metropolis-Hastings loop is implemented in C which incurs a very large overhead in `gdp_sample` since it is reinitialized every iteration. In general, repeatedly calling an R function that hooks into C code is slow. (NOT QUITE ACCURATE FIX LATER)

3 Example

4 Background

Some packages on interactive graphics include **plotly** (Sievert 2020) that interfaces with Javascript for web-based interactive graphics, **crosstalk** (Cheng and Sievert 2021) that specializes cross-linking elements across individual graphics. The recent R Journal paper **tsibbletalk** (Wang and Cook 2021) provides a good example of including interactive graphics into an article for the journal. It has both a set of linked plots, and also an animated gif example, illustrating linking between time series plots and feature summaries.

5 Customizing tooltip design with ToOoOlTiPs

6 Summary

We have displayed various tooltips that are available in the package **ToOoOlTiPs**.

References

Cheng, Joe, and Carson Sievert. 2021. *crosstalk: Inter-Widget Interactivity for HTML Widgets*. https://CRAN.R-project.org/package=crosstalk.

Dwork, Cynthia, and Aaron Roth. 2013. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9 (3-4): 211–407. https://doi.org/10.1561/0400000042.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Wang, Earo, and Dianne Cook. 2021. "Conversations in Time: Interactive Visualisation to Explore Structured Temporal Data." *The R Journal*. https://doi.org/10.32614/RJ-2021-050.

Jordan A. Awan
Purdue University
Department of Statistics
West Lafayette, IN 47907
https://www.britannica.com/animal/quokka
jawan@purdue.edu

Kevin Eng
Rutgers University
Department of Statistics
Piscataway, NJ 08854
https://www.britannica.com/animal/quokka
ke157@stat.rutgers.edu

Robin Gong
Rutgers University
Department of Statistics
Piscataway, NJ 08854
https://www.britannica.com/animal/quokka
ruobin.gong@rutgers.edu

Nianqiao Phyllis Ju
Purdue University

Department of Statistics
West Lafayette, IN 47907
https://www.britannica.com/animal/quokka
nianqiao@purdue.edu

Vinayak A. Rao
Purdue University
Department of Statistics
West Lafayette, IN 47907
https://www.britannica.com/animal/quokka
varao@purdue.edu