Advances in Mathematical Sciences and Applications
Vol. 34, No. 2 (2025), pp. 779–801

GAKKOTOSHO
TOKYO JAPAN

TUNABLE CORRELATION RETENTION: A STATISTICAL METHOD FOR GENERATING SYNTHETIC DATA,

Nicklas Jävergård
Department of Mathematics and Computer Science, Karlstad University, Sweden
(E-mail: nicklas.javergard@kau.se)

Adrian Muntean
Department of Mathematics and Computer Science, Karlstad University, Sweden
(E-mail: adrian.muntean@kau.se)

Rainey Lyons
Department of Applied Mathematics, CU Boulder, Colorado, U.S.
(E-mail: Rainey.Lyons@colorado.edu)

and

Jonas Forsman
CGI, Karlstad, Sweden
(E-mail: jonas.forsman@cgi.com)

Abstract. We propose a method to generate statistically representative synthetic data from a given dataset. The main goal of our method is for the created data set to mimic the inter–feature correlations present in the original data, while also offering a tunable parameter to influence the privacy level. In particular, our method constructs a statistical map by using the empirical conditional distributions between the features of the original dataset. Part of the tunability is achieved by limiting the depths of conditional distributions that are being used. We describe in detail our algorithms used both in the construction of a statistical map and how to use this map to generate synthetic observations. This approach is tested in three different ways: with a hand calculated example; a manufactured dataset; and a real world energy-related dataset of consumption/production of households in Madeira Island. We evaluate the method by comparing the datasets using the Pearson correlation matrix with different levels of resolution and

Communicated by Editors; Received June 17, 2025
This work is supported by Swedish Energy Agency's project Solar Electricity Research Centre (SOLVE) with grant number 52693-1.
AMS Subject Classification: 62-04, 62-08, 62P30, 62H99.
Keywords: synthetic data generation, computational statistics.

depths of correlation. These two considerations are being viewed as tunable parameters influencing the resulting datasets fidelity and privacy.

The proposed methodology is general in the sense that it does not rely on the used test dataset. We expect it to be applicable in a much broader context than indicated here.