Sequential and Hybrid Frameworks for Enhancing the Iterative Proportional Fitting (IPF) Algorithm in Population Synthesis

William O. Agyapong[1], Amy Wagler[2]

[1]Department of Mathematical Sciences, The University of Texas at El Paso

[2]Department of Public Health Sciences, The University of Texas at El Paso

Abstract

High-dimensional data poses significant challenges for the Iterative Proportional Fitting (IPF) algorithm, a method commonly used in population synthesis (synthetic population generation) and survey analysis for post-stratification. Despite its simplicity and computational efficiency in low-dimensional settings, IPF struggles with zero-cell issues in sparse contingency tables and faces substantial computational costs as dimensionality increases. Additionally, concerns exist regarding its ability to preserve the dependence structure of the reference joint distribution and to generate realistic, out-of-sample instances, particularly in population synthesis. To address these limitations, we propose a Sequential IPF approach that applies standard IPF to subgroups of correlated features. We evaluate this method using synthetic population generation experiments based on data from the American Community Survey (ACS) for the State of Maryland. Preliminary results demonstrate that Sequential IPF significantly enhances computational efficiency and scalability compared to traditional IPF. Furthermore, we introduce a hybrid framework that integrates IPF-synthesized data with machine learning–based generative models such as variational autoencoders, Bayesian networks, and generative adversarial networks. This hybrid approach not only satisfies marginal constraints but also generates novel samples that better capture multivariate dependencies, enhancing the realism and utility of the synthesized populations. These contributions offer practical solutions for high-dimensional data applications, improving upon standard IPF methods and advancing the field of synthetic population generation.

Keywords: Iterative Proportional Fitting (IPF); Sequential IPF; dependence structure; high-dimensional data; sparse contingency tables; zero-cell problems; synthetic population generation; machine learning generative models; multivariate dependencies, computational efficiency.