Chapter 5
Evolution on the Generation and Analysis of Single Imputation Synthetic Datasets in Statistical Disclosure Control

Ricardo Moura, Carlos A. Coelho, and Bimal Sinha

5.1 Introduction

In today's world, simply using a smartphone leads to the creation of a vast array of data. This data is automatically saved, and more and more, various organizations, businesses, and institutions are insisting on accessing this data for research and examination purposes. However, if this data is shared without proper control, it could threaten the privacy of the individuals or groups to whom the data pertains.

To adhere to the principle of statistical confidentiality besides the physical safeguarding of data—meaning data that is stored and accessible only to authorized personnel—many national and international bodies commonly implement Statistical Disclosure Control (SDC) methods. The purpose of these methods is to protect the confidentiality of existing data deemed sensitive, minimizing the risk of individual

R. Moura (✉)
Mathematics Department, Lisbon School of Economics & Management (ISEG), Lisbon, Portugal

NOVA Math – Center for Mathematics and Applications, Caparica, Portugal

Portuguese Navy Research Center (CINAV), Alfeite, Portugal
e-mail: rp.moura@fct.unl.pt

C. A. Coelho
Mathematics Department, NOVA FCT- NOVA University of Lisbon, Caparica, Portugal

NOVA Math – Center for Mathematics and Applications, Caparica, Portugal
e-mail: cmac@fct.unl.pt

B. Sinha
Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA

Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA
e-mail: sinha@umbc.edu

C. A. Coelho, D.-G. Chen (eds.), *Statistical Modeling and Applications*, Emerging Topics in Statistics and Biostatistics, https://doi.org/10.1007/978-3-031-69622-0_5

identification.[1] This allows for the public release of such data while ensuring privacy protection.

Noise addition, data rounding, selective data suppression, and the creation of synthetic data are a few methods of Statistical Disclosure Control (SDC) employed by organizations like EUROSTAT and the US CENSUS BUREAU before releasing data to the public (Hundepool et al., 2010; Abowd et al., 2020). This text will delve into the synthetic data generation techniques, which in brief, involve replacing the actual data with its synthetic counterparts. This technique is not only relatively new but also has the significant advantage of maintaining the statistical characteristics of the original model assumed for the data, a feature not commonly found in other SDC methods (Drechsler, 2011). Consequently, this approach is being actively promoted for research by government institutions worldwide.

Little (1993) and Rubin (1993), are often credited as being the forerunners in developing this technique, being the first to propose the generation of synthetic data through multiple imputation (Rubin, 1987) as a form of SDC. This process entails substituting original data with synthetic versions, allowing for public disclosure without risking the privacy of the individuals who provided the data, as these synthetic versions lack enough detail to breach individual confidentiality.

It was only a decade later that the feasibility of procedures enabling the analysis of these synthetic data generated by multiple imputation became available. The works of Reiter (2002, 2003), Raghunathan et al. (2003), and Reiter (2005a,b) were instrumental in this development, driven by a Bayesian approach, allowing for the study of the parameter or vector of parameters involved.

However, in specific situations, where there is a significant risk of revealing a respondent's identity, it may become impractical to publish multiple iterations of the original data. Under these circumstances, the strategy shifts to the release of a single synthetic version, relying thus on data produced through single imputation (Abowd et al., 2006; Kinney et al., 2011, 2014; Bowen et al., 2020; Alam et al., 2020).

Evidently, the conventional methods for combining inferences are not suitable when just one synthetic dataset is released, leading to the question of whether valid inferences can still be drawn in these scenarios.

In the following sections, attention will be directed towards an extensive review of the literature concerning the generation and analysis of single imputation synthetic datasets. This exploration will cover a range of methods, highlight the key challenges faced, and discuss the significant advancements. The aim is to offer a thorough insight into how the field currently stands and to shed light on prospective developments in utilizing single imputation synthetic data as a tool for preserving privacy when releasing data.

[1] Regulation (European Commission) Number 223/2009, 2009.

5.2 Different Types of Synthetic Data

As mentioned before, we want to focus on the generation of synthetic data as a way of protecting the confidentiality and thus we must start to specify the different categories of synthetic data available.

5.2.1 Fully Synthetic Data

Introduced by Rubin (1993), the approach releases, instead of the original dataset, one or more synthetic datasets obtained by generating the overall data, that is, generating values for all variables, from a model that is assumed to fit the original data. If missing data is present in the original data, this process may even fill these gaps. Subsequently, only the synthetic versions of these datasets are released to the public. These publicly available datasets will not contain any of the original data. When circumstances dictate that only one synthetic dataset can be released, we will be in the single imputation scenario.

5.2.2 Partially Synthetic Data

Little (1993) suggested to partially synthesize the data. This method refers to situation where synthetic versions are synthesized only for those individual records deemed sensitive, that is, which release could compromise their confidentiality. The remaining values are left unchanged, ensuring protection without compromising the overall quality of the publicly disclosed data (Reiter, 2005c). Partially synthetic generated or imputed data may thus involve (1) the synthetic generation of some of the unit or observational records, for all the variables, or only for the sensitive variables, or (2) otherwise involve the synthetic generation of the values for all units (or only sensitive units) for given sensitive variables. Clearly, releasing some of the original data values to the public presents a greater risk of disclosure. However, as a trade-off, the dataset made available will more closely mirror the properties of the original dataset, especially when compared to a scenario where the available data is fully synthetic. Indeed, if the number of key variables or key values in the original data, which are supposed to remain unaltered, is reduced to zero, then the scenario of partially synthetic data effectively becomes equivalent to the case of fully synthetic data. However, the use of the partially synthetic data approach to create a fully imputed synthetic dataset will lead to a different inferential analysis approach.

5.3 Methods of Creating Synthetic Data

This survey presents an overview of the diverse techniques employed to generate synthetic data. We may say there are two main classes of methods for synthetic data generation, model-based and machine learning based.

5.3.1 Model-Based

The model-based approach to synthetic data generation heavily depends on assuming a specific model for the data and then synthesizing this data by sampling (randomly generating values) from the specified model. Model-based techniques include the use of various parametric models like logit models, exponential models, pareto, univariate and multivariate normal models, multiple and multivariate regression models, and sequential conditional regression models. Some references, to name just a few, include works by Kennickell (1997), Abowd et al. (2006), Reiter (2005a), An and Little (2007), Klein and Sinha (2015c,b,d,a), Moura et al. (2017, 2021), and Mishra and Barui (2022).

It is crucial to note that these model-based methods can be categorized into two distinct approaches when synthesizing multiple key variables.

One approach, known as *joint modeling*, involves specifying the joint distribution of the data. This method focuses on clearly defining a joint model that captures the existing relations among all variables, both those that are to be synthesized and those that may remain unchanged (Drechsler & Haensch, 2023).

The other approach is *sequential modeling*, where each sensitive variable is synthesized in sequence using models that are conditioned on the original variables that do not need to be replaced and on any of the remaining variables that need to be synthesized. To better understand the process of sequential modeling let us consider that we have a set of p variables from which $Y_1, \ldots, Y_{p_1}$ are considered to be sensitive, and which thus need to be replaced by the synthetic versions, and variables $X_{p_1+1}, \ldots, X_p$, which are the non-sensitive variables. One approach, is the sequential regression multivariate imputation (SRMI) where a sequence of conditional distributions $f(Y_j|Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_{p_1}, X_{p_1+1}, \ldots, X_p)$ is used to generate synthetic versions of the original Y_j, for $j = 1, \ldots, p_1$ (Raghunathan et al., 2001). However, this process may be very complex and may require a large amount of time. Thus, a simplified algorithm commonly used considers the distributions conditioned only on the non-sensitive variables, $X_{p_1+1}, \ldots, X_p$ and the variables that have already been synthesized. That is, generates values sequentially by using the distributions $f(Y_1|X_{p_1+1}, \ldots, X_p)$, $f(Y_2|Y_1, X_{p_1+1}, \ldots, X_p)$, and so on, until we generate the values for the last variable from $f(Y_{p_1}|Y_1, \ldots, Y_{p_1-1}, X_{p_1+1}, \ldots, X_p)$ (Drechsler, 2011). Due to its simpler application, this is the algorithm that is often used in practice.

Rubin (1987) suggested that for an approximately unbiased estimation of variance in multiple imputation of missing data, the replacement data should be drawn from posterior predictive distributions. This idea formed the basis for generating synthetic data through *Posterior Predictive Sampling* (PPS). However, Reiter and Kinney (2012) established that this is not strictly necessary. They showed that one can simply use sample estimates, such as posterior modes or maximum likelihood estimates, and plug them in the model, drawing then the synthetic data from this model. This method of generation is commonly referred to as the *Plug-in Sampling* (PS) method.

In the PPS method, the synthetic data is essentially seen as predictions coming from the posterior distribution of the model, whose parameters are estimated from the confidential dataset. On the other hand, in the PS method, synthetic data is created directly from the parametric model by plugging in the sample estimates, which are also based on the confidential data. This approach differs from PPS in that it uses same sample estimates to generate the different imputations of synthetic data rather than using several posterior distributions to generate those imputations.

In the model-based approach to synthetic data generation, it is generally advised to use a large set of predictor variables for generating the synthetic variables. This way, when someone wants to analyze the data, using a simpler model than the one used to synthesized the data, it is still ensured a good degree of congruence between the two models.

5.3.2 Machine Learning Based

Contrary to the notion of using a parametric model to generate the data, one can consider adopting a non-parametric approach. This alternative methodology does not rely on predefined parametric models but instead uses methods that are more flexible and less dependent on assumptions about the data's underlying distribution. Reiter (2005c) proposed the use of Classification and Regression Trees (CART) for synthetic data generation. When there are multiple key variables to be replaced with synthetic values, a sequential regression approach, as mentioned earlier, is recommended. This method offers more flexibility compared to the model-based approach, allowing for the accommodation of non-linear relationships between variables and non-standard variable distributions. However, due to the sequential nature of generation in this method, the first variables are generated independently of those generated later. As a result, if there are any relationships between these early and later variables, such interdependencies may not be fully captured in the synthetic data creation process. We may see extensions of the very popular CART approach in the use of random forests (Caiola & Reiter, 2010), support vector machines (Drechsler, 2010) and even into the use of deep learning methods such as Generative adversarial networks (GANs) (Park et al., 2018; Xu et al., 2019; Koivu et al., 2020), and Variational Autoencoders (VAEs) (Srivastava et al., 2017; Vardhan & Kok, 2020; Ma et al., 2020).

5.4 Inferential Analysis of Single Imputation Synthetic Datasets

5.4.1 Origins of Inferential Analysis for Synthetic Datasets

Before delving into the inference analysis of the single imputation case, it is essential to first discuss the origins of inference analysis in this context. The concept of synthetic data, originating from the work of Rubin (1987), naturally led to the development of inferential procedures initially based on scenarios involving both fully and partially generated data, related to the multiple imputation framework. The first comprehensive rules for analyzing multiply imputed full synthetic data were derived by Raghunathan et al. (2003). In a similar way, methodologies for partially synthetic data were introduced by Reiter (2003). Further advancements were made by Reiter (2005c), who expanded these concepts to include multivariate techniques, thereby extending the application of Wald tests and Likelihood ratio tests.

To gain a deeper insight into the essence of these rules, consider an analyst who is focused on a specific parameter Q from the true model. When M synthetic datasets are released, the procedure entails gathering the point estimate q_j $(j = 1, \ldots, M)$ and its corresponding variance estimate u_j, from every synthetic dataset as though it were the original data. These rules, commonly designated as combination rules, as formulated by the previously mentioned authors, involve calculating the variance of the point estimate by combining the within-variance of the point estimates q_j with the average of the variance estimates u_j. Consequently, the analyst can use the mean of q_j as an unbiased estimator for Q and then use either a Student's t-distribution or an F-distribution, to infer about Q. For a more comprehensive understanding, readers are directed to the work of Reiter and Raghunathan (2007).

As the reader might have surmised, these methodologies and associated distributions need that the number of synthetic versions released to the public exceed 1. This requirement is critical, as a single dataset or a very small number of synthetic datasets would render the distributions used for inference impractical. In such scenarios, one would face the predicament of not being able to calculate the degrees of freedom, or, in cases where the number is very small, extremely limited degrees of freedom in the distributions.

As highlighted in the Introduction, certain scenarios (Abowd et al., 2006; Kinney et al., 2011, 2014; Bowen et al., 2020; Alam et al., 2020) need the release of a single synthetic dataset to mitigate the risk of data disclosure.

5.4.2 The Single Imputation Scenario

The inability to effectively handle single imputation requirements or a limited number of releases of synthetic versions of data was what prompted Klein and Sinha (2015d) to address this specific challenge. Beginning in 2013, these authors

embarked on compiling a collection of papers focused on this issue. In their work, Klein and Sinha (2015c), they considered both Posterior Predictive Sampling (PPS) and Plug-in Sampling (PS) generation methods. They developed exact inference procedures for analyzing data, based on sufficient statistics, when it is presumed that the original data follows an exponential model. These exact procedures are applicable not only to the multiple imputation scenario but also to single imputation cases. For both PPS and PS cases unbiased estimators of the true parameters can be derived.

In the same year, the same authors, published three additional works (Klein & Sinha, 2015b,d,a) focusing on single imputation synthetic data for confidentiality purposes. These papers expanded the results of exact inference procedures for both PPS and PS methods, considering scenarios where the original data is assumed to follow univariate or multivariate normal, and for cases of partially generated data through multiple linear regression (MultiLR) models. In these works, the authors successfully derive the exact distributions for key sample measures in the synthetic data: the mean, variance, and regression coefficients.

All these studies are based on the assumption of normality in the data, but they differ in the type of synthetic data used. For the univariate and multivariate normal cases, it is assumed that the synthetic data will be fully synthesized. In contrast, for the MultiLR model, the released data is partially synthesized; here, the response variables are considered sensitive, and the predictor variables remain unchanged.

In the latter scenario, those authors explored cases where the imputer and/or analyst either overfits or underfits the regression model. They also examined cases where the analyst's regression model is the regression of the unsynthesized variable on the synthesized variable, which is the opposite of the generation case. These explorations were conducted under the PS method.

In the following year, Moura (2016) delved into the multivariate linear regression model (MLR), resulting in two more publications (Moura et al., 2017, 2021). In Moura et al. (2017), attention was given to the PPS method, whereas Moura et al. (2021) focused on the PS method. A major challenge tackled in these studies was the measurement of the size of confidence sets based on the test statistics employed in drawing inference about the matrix of regression coefficients. To tackle this problem, they introduced a new measure for the multivariate scenario, termed "*radius*", which may be seen as the distance between the centre and the edge of the confidence set. This measure, based on the quantiles of the distribution of the test statistic used for inference on the matrix of regression coefficients and on the determinant of the estimator of the variance-covariance matrix of the variables based on the synthetic data provides an useful approach to evaluating confidence in multivariate contexts, in situations where the confidence sets may have an infinite volume.

A critical aspect of Moura et al. (2021) was their approach to scenarios involving the release of synthetic data that was either non-normal or involved non-continuous response variables. Despite these challenges, the inferential procedures that were based on normality assumptions continued to perform effectively. The authors proposed adaptations to the PS method to suit these more complex data scenarios.

This showed the method's adaptability and efficacy in handling a diverse array of data types, highlighting the wide applicability of this synthetic data generation methods.

It's important to recognize that these studies did not just develop exact inference procedures for multiple imputation synthetic data generated by PPS and PS methods, but also extended these procedures to single imputation approaches. Additionally, they adapted Reiter (2005b) vector methodology, known for its combination rules, to a matrix parameter, as it is relevant in the context of the MLR. This adaptation is a significant advancement, as it extends the applicability of Reiter's methodology to more complex statistical models.

In 2022, Guin et al. (2022) addressed the inferential analysis for the MultiLR model under PPS and PS, but from a Bayesian inference perspective. They employed a non-informative diffuse prior, in contrast to the frequentist approach commonly seen in the literature. This study extended the work of Klein and Sinha (2015d,a) into the Bayesian realm.

The paper by Guin et al. (2022) explored two distinct methodologies for generating data. The first method focuses exclusively on using the sensitive data to estimate the model parameters. The rationale behind this approach is to ensure that in the released dataset, the synthetic portion, which represents the sensitive data, remains independent of the non-synthetic and non-sensitive part. This separation is key for maintaining the integrity of the non-sensitive data while protecting the confidentiality of the sensitive information. The second method involves utilizing the entire original dataset for parameter estimation. This comprehensive approach can potentially offer a more holistic view of the data's underlying structure, albeit with increased difficulty in maintaining data privacy.

Interestingly, in terms of estimating the regression coefficients, both approaches in this paper mirror the frequentist inference method. The estimator for the vector of regression coefficients, in the MultiLR model, is computed in the same way as the usual unbiased estimator used with the original data. This parallelism simplifies the computation of estimates when only one dataset is available.

In 2023, the same authors turned their focus to the Multivariate Normal Model, examining both PPS and PS scenarios through the lens of Bayesian inference (Guin et al., 2023). Once again, opting for a non-informative prior, they obtained Bayes estimators for the mean, the covariance matrix, and the generalized covariance. Notably, the Bayes estimator for the population mean mirrors precisely the estimator derived from traditional frequentist inference methods. Furthermore, they outlined methods for constructing credibility sets for both the generalized variance and the population mean, providing valuable insights into the convergence of Bayesian and frequentist methodologies in this context.

Klein et al. (2021) addressed a gap in previous research concerning inference about the variance structure under the Multivariate Normal assumption, particularly in the context of fully synthetic data generation. Their work significantly expanded the scope of inferential procedures to include more complex statistical assessments, especially when dealing with PS synthetic data.

Specifically, they developed inference procedures for several key statistical measures: the generalized variance, the sphericity test, and tests for the independence of two subsets of variables and for the regression of one subset on the other. This extension allows for a more comprehensive analysis of synthetic data, encompassing a broader range of inferential needs.

A notable feature of the proposed test statistics is their similarity to the conventional statistics commonly found in existing literature. This familiarity facilitates analysts' quick identification and computation of the desired test statistic values. The only challenge for analysts lies in utilizing the 'R' package 'PSinference' (Moura et al., 2023), which provides access to the Monte Carlo simulated distributions of these test statistics. This package represents a valuable tool, enabling analysts to accurately apply these advanced statistical procedures to synthetic data.

Under the same assumption of normal distribution, Basak and Sinha (2024) tackled the univariate case for fully synthetic data generated by both PS and PPS. Their work focused on inferential procedures for testing the equality of means across several populations, akin to ANOVA (Analysis of Variance). They provided insights into the distribution properties not just under the null hypothesis but also under the alternative hypothesis.

An important aspect of their study was examining the robustness of these proposed tests when the assumption of normality in the original data was violated. The results were quite noteworthy. When the original data is generated from distributions like Double Exponential, Student's t, Exponential, or Lognormal, the Type 1 error rates are significantly impacted in the traditional ANOVA F-test. However, this is not the case with their proposed tests for PS and PPS synthetic data. This finding indicates that the tests developed for synthetic data are more robust than the traditional tests, offering greater reliability when the underlying assumptions about the data distribution are not met.

Venturing beyond the normality assumption, Mishra and Barui (2022) explored the Pareto Model (specifically the two-parameter model) in the context of synthetic data generation. Their study encompassed both types of data synthesis: PS and PPS. In the case of PS, the authors developed inferential procedures for all possible scenarios: whether one of the parameters is known and the other unknown, or both parameters are unknown. This comprehensive approach in the PS framework allows for a broad application of the Pareto Model in various data contexts. On the other hand, for PPS, the focus was narrower. The inferential procedures were presented only for the shape parameter, under the assumption that the other parameter (scale parameter) is known.

While much of the work in synthetic data generation focuses on methods like Plug-in Sampling (PS) or Posterior Predictive Sampling (PPS), it's important to acknowledge alternative approaches. Klein and Datta (2018) explored a method based on the principle of sufficiency, particularly under the assumption of a normal MultiLR model. In this method, partial synthetic data is generated by sampling from the conditional distribution of the data, given the sufficient statistics which are free from unknown parameters.

This approach is particularly advantageous for several reasons. Firstly, when the regression model is correctly specified, the values of the sufficient statistics can be directly calculated from the synthetic version of the data. This allows the analyst to use standard regression methodologies. Furthermore, if the imputer overfits the regression model, valid inferences can still be obtained by the analyst, even when using an underfitted model. However, the same is not true if the imputer underfits the model and the analyst overfits model.

It's also crucial to note that this type of synthetic data generation and the subsequent release of synthetic data is contingent upon the institution holding the original data agreeing to release the sufficient statistics of those data, since with access to these sufficient statistics and the specified model, additional versions of the original data can be generated.

Finally, we should highlight the work of Raab et al. (2018), which introduced a novel variance estimator suitable for use with large synthetic datasets. This estimator is applicable regardless of whether the datasets are generated through non-parametric or parametric methods, and whether they are derived from the posterior predictive distribution or not. A key feature of this new variance estimator is that it does not account for variance between multiple synthetic datasets. As a result, it can be used in the inferential analysis of datasets that have undergone single imputation, using procedures somewhat akin to those found in Raghunathan et al. (2003).

Despite the promising nature of this development, there remains a keen interest in seeing its application in real-world scenarios, particularly through simulations and practical applications on actual datasets. Specifically, it would be informative to observe the performance of this variance estimate in cases where the number of synthetic versions is reduced to one, a scenario that was not examined in their work. Such investigations could provide valuable insights into the effectiveness and reliability of this variance estimator in different contexts of synthetic data analysis.

Table 5.1 presents a summary of the references in this section pertaining to singly imputation analysis, where the works are ordered chronologically. We may note that all works except the one by Raab et al. (2018) address exact inferential procedures for the analysis of synthetic data.

It is important to emphasize the significance of exact inferential methods, which are not only applicable in both single and multiple imputation scenarios but also serve as a vital tool for assessing the quality and effectiveness of asymptotic procedures in these contexts.

These exact methods provide a robust framework for researchers and practitioners, ensuring the reliability and accuracy of statistical inferences drawn from synthetic data, thereby playing a crucial role in advancing the methodologies and applications of SDC. As such, these exact inferential methods are an essential component to consider in the ongoing evolution and refinement of SDC techniques.

Table 5.1 Summary of the work done for the single imputation case

Author(s)	Year	Model	Generation method	Type of synthetic generation
Klein and Sinha (2015c)	2015	Exponential	PPS/PS	Fully/Partially
Klein and Sinha (2015b)	2015	Univariate normal	PPS/PS	Fully
Klein and Sinha (2015d)	2015	Multivariate normal and MultiLR	PPS	Fully/Partially
Klein and Sinha (2015a)	2015	Multivariate normal and MultiLR	PS	Fully/Partially
Moura et al. (2017)	2017	MLR	PPS	Partially
Klein and Datta (2018)	2018	MultiLR	Sufficiency principle	Partially
Raab et al. (2018)	2018	Generic	PPS/PS	Fully/Partially
Moura et al. (2021)	2021	MLR	PS	Partially
Klein et al. (2021)	2021	Multivariate normal (Tests for the covariance structure)	PS	Fully
Guin et al. (2022)	2023	MultiLR (Bayesian inference)	PPS/PS	Partially
Mishra and Barui (2022)	2022	Pareto	PPS/PS	Fully
Guin et al. (2023)	2023	Multivariate normal (Bayesian inference)	PPS/PS	Fully
Basak and Sinha (2024)	2024	ANOVA	PPS/PS	Fully

5.5 Multivariate Normal Model: A Theoretical Overview

Given the range of theoretical results surrounding single imputation, our discussion will focus specifically on the Multivariate Normal Model. This model is selected for its comprehensive suite of inferential procedures and its broad applicability. Our focus will be on the PS generation method due to its straightforward implementation and the fact that a single synthetic dataset may suffice for drawing inferences. This approach not only simplifies the process but also ensures an adequate level of privacy protection, making it a highly suitable choice for our analysis.

To start, let us delve into the generation of synthetic data using the PS method, within the context of the Multivariate Normal Model. For that, we begin to consider the original data as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where

$$\mathbf{x}_i \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i = 1, \ldots, n \ , \tag{5.1}$$

for $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})'$, where n is the sample size and p the number of variables involved. The initial phase is estimating the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$, from the original dataset, respectively by the sample mean, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, and the sample variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{\mathbf{S}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})\,(\mathbf{x}_i - \bar{\mathbf{x}})'\,. \tag{5.2}$$

The second step, is to plug-in the estimates in to the model and generate the synthetic versions of $\mathbf{x}_i$, $\mathbf{v}_i$, by randomly drawing

$$\mathbf{v}_i \overset{i.i.d.}{\sim} N_p\left(\bar{\mathbf{x}}, \hat{\boldsymbol{\Sigma}}\right), i = 1, \ldots, n \ .$$

Thus, we end up with one fully synthetic dataset, $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$. If someone wishes to reproduce more than one fully synthetic dataset, it has only to multiply create $\mathbf{V}_1, \ldots, \mathbf{V}_M$ $(M > 1)$, independently from each other. As one may see, the process is very straightforward, since we draw from a multivariate Normal model where the sample estimates are plugged-in.

The idea of releasing a single synthetic version of the original dataset and not the sample mean and sample variance is due to the fact that institutions are usually not comfortable in doing so, since that way someone may produce multiple synthetic versions of the original dataset and come up closer to the original dataset, possibly compromising the confidentiality of the respondents (Reiter et al., 2014). As such, in this section we will only address the case of singly imputed datasets.

5.5.1 Synthetic Mean and Variance-Covariance Matrix

Now that the synthetic version of the dataset is available, obtaining estimators for the population mean and covariance matrix becomes a crucial next step. Fortunately, this process is straightforward.

Using the synthetic version of the data, we can calculate the synthetic sample mean as

$$\bar{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i,$$

and the synthetic variance-covariance matrix as

$$\hat{\boldsymbol{\Sigma}}_{PS} = \frac{\mathbf{S}_{PS}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{v}_i - \bar{\mathbf{v}})(\mathbf{v}_i - \bar{\mathbf{v}})'.$$

Some key properties of $\bar{\mathbf{v}}$ and $\hat{\boldsymbol{\Sigma}}_{PS}$ are:

1. They are jointly sufficient statistics for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, ensuring efficient use of data for estimation (Klein & Sinha, 2015b).
2. $\bar{\mathbf{v}}$ is the maximum likelihood estimator of $\boldsymbol{\mu}$, providing an unbiased estimate with a variance of $\frac{2\boldsymbol{\Sigma}}{n}$ (Klein & Sinha, 2015b).
3. $\hat{\boldsymbol{\Sigma}}_{PS}$ offers an unbiased estimation of $\boldsymbol{\Sigma}$, enhancing the reliability of the synthetic data (Klein & Sinha, 2015b).
4. With a non-informative joint prior, $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\gamma/2}$, the Bayes estimator for $\boldsymbol{\mu}$ aligns with $\bar{\mathbf{v}}$, and the Bayes estimator for $\boldsymbol{\Sigma}$ is determined as $\frac{(n-1)^2 \hat{\boldsymbol{\Sigma}}_{PS}}{(n-2p+\gamma-3)^2}$, with $n > \max\{p, 2p - \gamma + 3\}$ (Guin et al., 2023).

These properties offer a framework for analyzing synthetic data, since they ensure that the estimates of the mean and covariance matrix are both statistically sound and practical for further analysis.

5.5.2 *Confidence and Credibility Sets for $\boldsymbol{\mu}$*

A fundamental step in conducting inference on a dataset involves making inferences about the mean. The synergistic contributions of Klein et al. (2021) and Guin et al. (2023) methodologies facilitates the creation of both confidence and credibility sets for $\boldsymbol{\mu}$. This dual approach encompasses the principles of frequentist and Bayesian inference, offering two ways to draw inference about the mean in the context of normal distribution.

5.5.2.1 Confidence Ellipsoid for $\boldsymbol{\mu}$

To infer about $\boldsymbol{\mu}$ using synthetic data, consider the random variable

$$T^2 = \frac{n}{n-1} (\bar{\mathbf{v}} - \boldsymbol{\mu})' \hat{\boldsymbol{\Sigma}}_{PS}^{-1} (\bar{\mathbf{v}} - \boldsymbol{\mu}), \tag{5.3}$$

as defined by Klein and Sinha (2015b). Its distribution can be simulated through a series of steps to generate a random value of T^2:

- Generate a $p \times p$ random matrix $\boldsymbol{\Sigma}^*$ from a Wishart distribution, with parameter $\mathbf{I}_p$ and $n - 1$ degrees of freedom
- Compute $\lambda_1, \ldots, \lambda_p$, as the solution of $\left|(n-1)\mathbf{I}_p + (1-\lambda)\boldsymbol{\Sigma}^*\right| = 0$
- Generate χ^2_{1j}, $j = 1, \ldots, p$, independently, from a χ^2 distribution with 1 degree of freedom
- Compute $T_2 = \sum_{j=1}^{p} \lambda_j \chi^2_{1j}$

- Generate T_1, independently of T_2, from a inverse χ^2 distribution with $n - p$ degrees of freedom
- Compute $T^2 = T_1 \times T_2$.

To construct a $(1 - \alpha)$ confidence ellipsoid for $\boldsymbol{\mu}$, calculate the $(1 - \alpha)$ percentile, $c_{n,p;\alpha}$ from the empirically simulated values from the distribution of T^2. This ellipsoid is given by

$$\Delta(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : T^2 \leq c_{n,p,\alpha}\}$$

with a volume of

$$V_{\boldsymbol{\mu}}(\mathbf{V}) = \frac{(n-1)^{p/2}\pi^{p/2}}{n^{p/2}\Gamma(p/2+1)}(c_{n,p;\alpha})^{p/2}|\hat{\boldsymbol{\Sigma}}_{PS}|^{1/2},$$

and corresponding expected volume

$$E[V_{\boldsymbol{\mu}}(\mathbf{V})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma(p/2+1)}(c_{n,p;\alpha})^{p/2}\prod_{j=1}^{p}\left[\frac{2^{1/2}\Gamma\left(\frac{n-j+1}{2}\right)}{\Gamma\left(\frac{n-j}{2}\right)}\right]|\boldsymbol{\Sigma}|^{1/2}.$$

5.5.2.2 Credibility Set for $\boldsymbol{\mu}$

From a Bayesian perspective, the random variable defined in (5.3) is also considered, but taking into account the posterior distribution of the synthetic data. Guin et al. (2023), proposed the following steps to repeatedly generate a value of its distribution:

- Generate a $p \times p$ random matrix $\mathbf{B}$ from an inverse Wishart distribution, with parameter $\mathbf{I}_p$ and $n - p + \gamma - 2$. Here one must establish one value for γ, abiding by the restriction in Sect. 5.1. In their simulations study, Guin et al. (2023) suggest setting the value of γ at 10 for the PS scenario
- Generate a $p \times p$ random matrix $\mathbf{A}^*$ from an inverse Wishart distribution, with parameter $(n-1)\mathbf{B}$ and $n - p + \gamma - 2$ degrees of freedom
- Compute $\mathbf{A} = \mathbf{A}^* + \mathbf{B}$
- Compute $\lambda_1, \ldots, \lambda_p$, as the solution of $\left|\mathbf{A} - \lambda\mathbf{I}_p\right| = 0$
- Generate χ^2_{1j}, $j = 1, \ldots, p$, independently, from a χ^2 distribution with 1 degree of freedom
- Compute $T^2 = \sum_{j=1}^{p} \lambda_j \chi^2_{1j}$.

To construct a $(1 - \alpha)$ credibility ellipsoid for $\boldsymbol{\mu}$, obtain the $(1 - \alpha)$ percentile, $k_{n,p,\gamma;\alpha}$, from the empirically simulated distribution of T^2. This ellipsoid is given by

$$\Delta_{Bayes}(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : T^2 \leq k_{n,p,\gamma;\alpha}\}$$

with volume

$$V_{Bayes,\boldsymbol{\mu}}(\mathbf{V}) = \frac{(n-1)^{p/2}\pi^{p/2}}{n^{p/2}\Gamma(p/2+1)}(k_{n,p,\gamma;\alpha})^{p/2}|\hat{\boldsymbol{\Sigma}}_{PS}|^{1/2}$$

and corresponding expected volume

$$E[V_{Bayes,\boldsymbol{\mu}}(\mathbf{V})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma(p/2+1)}(k_{n,p,\gamma;\alpha})^{p/2}\prod_{j=1}^{p}\left[\frac{2^{1/2}\Gamma\left(\frac{n-j+1}{2}\right)}{\Gamma\left(\frac{n-j}{2}\right)}\right]|\boldsymbol{\Sigma}|^{1/2}.$$

It's noteworthy that the expected volumes for the confidence and credibility sets, derived from the frequentist and Bayesian approaches respectively, exhibit a high degree of similarity. This resemblance arises primarily due to the analogous nature of their construction, with the only distinction being the specific cut-off points chosen from the distributions of T^2 for each approach.

5.5.3 Variance-Covariance Structure

Based on the result found in Klein and Sinha (2015b), conditionally on $\mathbf{S}$, defined in (5.2),

$$\mathbf{S}_{PS} \sim W_p\left(n-1, \frac{\mathbf{S}}{n-1}\right),$$

Klein et al. (2021) presented four inferential procedures regarding the structure of the variance-covariance matrix, which are presented in the following subsections.

5.5.3.1 Generalized Variance

We have seen, previously, that both the expected volumes for the confidence set and credibility set for $\boldsymbol{\mu}$ are a function of $|\boldsymbol{\Sigma}|$, the generalized variance. The generalized variance is an important measure, used not only to determine the volume of those sets but also to measure the spread of the dataset.

Consider the random variable

$$T_1^{PS} = (n-1)^p\frac{|\mathbf{S}_{PS}|}{|\boldsymbol{\Sigma}|}.$$

The distribution of T_1^{PS} can be obtained using Monte Carlo simulation, following these steps, to generate one of its values:

- Generate p independent random variables A_j and, independently from these, p independent random variables B_j, for $j = 1, \ldots, p$, all from a chi-square distribution with $n - j$ degrees of freedom
- Calculate $T_1^{PS} = \left(\prod_{j=1}^{p} A_j\right)\left(\prod_{j=1}^{p} B_j\right)$.

Consequently, we may use the empirical distribution of T_1^{PS} to calculate the αth percentile, $t_{1;\alpha}^{PS}$, and use it to construct the $(1-\alpha)$ level confidence interval for $|\boldsymbol{\Sigma}|$ which is given by

$$\left(\frac{(n-1)^p\,|\mathbf{S}_{PS}|}{t_{1;1-\alpha/2}^{PS}}, \frac{(n-1)^p\,|\mathbf{S}_{PS}|}{t_{1;\alpha/2}^{PS}}\right).$$

Remarkably, for those preferring the Bayesian methodology over the frequentist approach, the test statistic suggested by Guin et al. (2023), denoted as T_1^{PS}, remains the same. However, the distribution of T_1^{PS} differs due to a modification in the generation process of the independent B_j random variables. Specifically, when generating these variables, the degrees of freedom should be adjusted from $n - j$ to $n - p + \gamma - j$, reflecting the Bayesian framework's unique considerations, considering the restriction for γ in Sect. 5.1.

5.5.3.2 Sphericity Test

In multivariate statistics, the assumption of sphericity represents a critical condition for the validity of certain inferential tests, related to MANOVA, repeated measures ANOVA, and multivariate analysis of covariance (MANCOVA).

The sphericity test is designed to assess this assumption, providing a means to test the hypotheses

$$H_0 : \boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p \;\; versus \;\; H_1 : \boldsymbol{\Sigma} \neq \sigma^2\mathbf{I}_p.$$

Defining

$$T_2^{PS} = \frac{|\mathbf{S}_{PS}|^{1/p}}{tr\left(\mathbf{S}_{PS}\right)/p}, \tag{5.4}$$

Klein et al. (2021) demonstrated that its distribution, assuming that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p$, can be empirically obtained through Monte Carlo simulation by iteratively executing the steps:

- Generate $\mathbf{W}_1 \sim W_p\left(n-1, \mathbf{I}_p/(n-1)\right)$ and $\mathbf{W}_2 \sim W_p\left(n-1, \mathbf{I}_p\right)$, independently

- Calculate $T_2^{PS} = \frac{|W_1 W_2|^{1/p}}{tr(W_1 W_2)/p}$.

With the empirical distribution we may conduct a sphericity test, that is, the hypotheses in (5.4), rejecting the null hypothesis, for a level of significance α, if the observed value of T_2^{PS} falls bellow $t_{2;\alpha}^{PS}$, the αth percentile of T_2^{PS}.

5.5.3.3 Test for the Independence of Two Subsets of Variables

Assessing the independence between two subsets of variables is crucial for uncovering complex relationships within a dataset. This analysis is key to understand the variables' structure, aiding in feature selection, dimensionality reduction, and other processes. The independence test aims to clarify whether variations in one variable set are statistically related to variations in another set or if the two sets behave independently.

Considering the vectors $\mathbf{x}_i$ in (5.1) split into two subvectors or sets of variables $\mathbf{x}_i = [\mathbf{x}'_{1i}, \mathbf{x}'_{2i}]'$ where the first set has size p_1, and the second set has size p_2, with $p = p_1 + p_2$, to carry out this test, the matrices $\mathbf{\Sigma}$ and $\mathbf{S}_{PS}$ are partitioned as follows:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}, \quad \mathbf{S}_{PS} = \begin{bmatrix} \mathbf{S}_{PS,11} & \mathbf{S}_{PS,12} \\ \mathbf{S}_{PS,21} & \mathbf{S}_{PS,22} \end{bmatrix},$$

where both $\mathbf{\Sigma}_{11}$ and $\mathbf{S}_{PS,11}$ are square matrices of dimension p_1.

Once assumed the normality of the $\mathbf{x}_i$, to test the hypothesis of independence of the two subsets of variables is equivalent to test the hypotheses

$$H_0 : \mathbf{\Sigma}_{12} = \mathbf{0} \ \ versus \ \ H_1 : \mathbf{\Sigma}_{12} \neq \mathbf{0}. \tag{5.5}$$

Once defined the random variable

$$T_3^{PS} = \frac{|\mathbf{S}_{PS}|}{|\mathbf{S}_{PS,11}|\,|\mathbf{S}_{PS,22}|}$$

it is feasible to simulate its distribution, under H_0 in (5.5), by iteratively following the following steps (Klein et al., 2021):

- Generate $\mathbf{\Omega}_1$ from a Wishart distribution with parameters $n-1$ and $\mathbf{I}_p$
- Generate $\mathbf{\Omega}_2$ from a Wishart distribution with parameters $n-1$ and $\frac{\mathbf{\Omega}_1}{n-1}$
- Partition $\mathbf{\Omega}_2$ as

$$\mathbf{\Omega}_2 = \begin{bmatrix} \mathbf{\Omega}_{2,11} & \mathbf{\Omega}_{2,12} \\ \mathbf{\Omega}_{2,21} & \mathbf{\Omega}_{2,22} \end{bmatrix}$$

  where $\mathbf{\Omega}_{2,11}$ is a square matrix with dimension p_1
- Compute $T_3^{PS} = \frac{|\mathbf{\Omega}_2|}{|\mathbf{\Omega}_{2,11}||\mathbf{\Omega}_{2,22}|}$.

The null hypothesis in (5.5) should then be rejected, for a level of significance α, if the observed value of T_3^{PS} falls bellow $t_{3;\alpha}^{PS}$, the αth percentile of T_3^{PS}.

The strategic choice for $\boldsymbol{\Omega}_1$, whose Wishart distribution parameter matrix is an identity matrix, makes the derivation of the empirical values of T_3^{PS} easier, and also laid the groundwork for the next result (for details see Klein et al. 2021).

5.5.3.4 Test for the Regression of One Set of Variables on the Other

The regression of one set of variables on another emerges as a powerful tool, enabling researchers to explore and quantify the relationships between two distinct sets of variables. It enables the understanding of how variations in one group of variables can explain or predict variations in another group, uncovering the underlying dynamics and dependencies within the data. Tests for the matrix of regression coefficients of one set of variables on another set involves the test of the hypotheses

$$H_0 : \boldsymbol{\Delta} = \boldsymbol{\Delta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\Delta} \neq \boldsymbol{\Delta}_0, \tag{5.6}$$

where $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$, and where $\boldsymbol{\Delta}_0$ represents any specified $p_1 \times (p - p_1)$ matrix.

Consider the test statistic

$$T_4^{PS} = \frac{\left|\left(\mathbf{S}_{PS,12}\mathbf{S}_{PS,22}^{-1} - \boldsymbol{\Delta}_0\right)\mathbf{S}_{PS,22}\left(\mathbf{S}_{PS,12}\mathbf{S}_{PS,22}^{-1} - \boldsymbol{\Delta}_0\right)'\right|}{\left|\mathbf{S}_{PS,11} - \mathbf{S}_{PS,12}\mathbf{S}_{PS,22}^{-1}\mathbf{S}_{PS,21}\right|}.$$

To empirically derive the distribution of T_4^{PS}, one can utilize Monte Carlo simulation. This involves the following set of iterative steps:

- Generate $\boldsymbol{\Omega}_1$ from a Wishart distribution with parameters $n - 1$ and $\mathbf{I}_p$
- Generate $\boldsymbol{\Omega}_2$ from a Wishart distribution with parameters $n - 1$ and $\frac{\boldsymbol{\Omega}_1}{n-1}$
- Partition $\boldsymbol{\Omega}_2$ as

$$\boldsymbol{\Omega}_2 = \begin{bmatrix} \boldsymbol{\Omega}_{2,11} & \boldsymbol{\Omega}_{2,12} \\ \boldsymbol{\Omega}_{2,21} & \boldsymbol{\Omega}_{2,22} \end{bmatrix}$$

  where $\boldsymbol{\Omega}_{2,11}$ is a square matrix with dimension p_1
- Compute $T_4^{PS} = \frac{\left|\boldsymbol{\Omega}_{2,12}\boldsymbol{\Omega}_{2,22}^{-1}\boldsymbol{\Omega}_{2,21}\right|}{\left|\boldsymbol{\Omega}_{2,11} - \boldsymbol{\Omega}_{2,12}\boldsymbol{\Omega}_{2,22}^{-1}\boldsymbol{\Omega}_{2,21}\right|}$.

For a given significance level α, we should reject the null hypothesis in (5.6), if the observed value of T_4^{PS} exceeds the $(1 - \alpha)$th percentile, $t_{4,1-\alpha}^{PS}$, of the T_4^{PS} distribution.

5.5.4 As a Summary

As shown, in the context of the fully synthetic PS method under the assumption of a normal distribution, exact inferential procedures for the mean, generalized variance, and tests for the covariance structure are readily available. Although comprehensive validation across all scenarios is pending, Moura et al. (2021) have explored the implications of assuming non-normal and even non-continuous distributions for the data, particularly for response variables, within the context of partially synthetic datasets. Their results indicate that the normality assumption holds up well, even when not fully met, showing its practical effectiveness. This paper posits that such effectiveness is likely to extend to fully synthetic datasets, highlighting the adaptability and utility of these inferential approaches even when faced with deviations from normality.

Moreover, because these procedures are exact and adaptable to multiple imputation contexts for synthetic data, they establish a benchmark for evaluating asymptotic inferential techniques tailored to both single (Raab et al., 2018) and multiple imputation frameworks (Raghunathan et al., 2003). Contrasting confidence sets generated from synthetic data using these exact procedures against those derived from asymptotic approaches offers a way to gauge the effectiveness and precision of asymptotic methods. Traditionally, the evaluation of inferential methods' quality involved comparisons with results from the original dataset. However, the exact nature of the procedures discussed in this manuscript provides an opportunity to compare expected values of sample estimators obtained from these exact procedures against the findings from asymptotic methods.

5.6 Conclusion and Last Remarks

This manuscript has traveled across the comprehensive landscape of generating and analyzing single imputation synthetic datasets within the domain of Statistical Disclosure Control (SDC). Through a detailed analysis of the literature on the development of inferential procedures for single imputation, this work gives a description of the evolution and the current state of methodologies for creating and utilizing synthetic datasets to uphold individual privacy while enabling statistical analysis.

The exact inferential procedures for analyzing synthetic data, as showcased by the contributions of several authors, show the potential of synthetic datasets in maintaining statistical integrity, while protecting the privacy. These procedures not only provide a basis for making robust inferences from synthetic data but also offer a benchmark for assessing the quality of asymptotic inferential methods in both single and multiple imputation scenarios. The adaptability of these exact methods to the Bayesian framework further extends their applicability, offering a versatile toolset for researchers and practitioners in SDC.

One of the key insights from this review is the effectiveness of the normality assumption in synthetic data analysis. Despite potential deviations from normality in real-world datasets, the procedures developed under this assumption can still show remarkable robustness, suggesting that their utility extends beyond strictly normal conditions. Broader studies of these methodologies, in the presence of non-normal or non-continuous data types, would be desirable and worthy.

Furthermore, the exploration into the variance-covariance structure and the development of tests for sphericity, independence, and regression analysis within the synthetic data context have enriched the toolkit available for SDC. These advances not only facilitate a deeper understanding of the data underlying structure but also enable analyses that were previously challenging in the realm of synthetic data.

In conclusion, the work presented in this manuscript aims at being a valid contribution to the field of SDC by providing a detailed overview of the generation and analysis of single imputation synthetic datasets. As SDC continues to adapt to the challenges posed by modern data environments, the insights and approaches detailed herein may play a role in shaping the future of privacy-preserving statistical analysis.

Acknowledgments The research conducted by Ricardo Moura and Carlos Agra Coelho was funded by national funds through the *Fundação para a Ciência e a Tecnologia* (FCT), I.P., Center for Mathematics and Applications (NOVA Math) under the scope of the projects UIDB/00297/2020 (https://doi.org/10.54499/UIDB/00297/2020) and UIDP/00297/2020 (https://doi.org/10.54499/UIDP/00297/2020). Bimal Sinha is thankful to Tommy Wright at the Center for Statistical Research and Methodology (CSRM) of the US Census Bureau for his kind support and encouragement.

References

Abowd, J. M., Stinson, M. H., & Benedetto, G. (2006). Final report to the social security administration on the SIPP/SSA/IRS public use file project. https://ecommons.cornell.edu/items/d7cf980e-5ec3-4322-aea2-a58b344c3ac6

Abowd, J., Benedetto, G., Garfinkel, S., Dahl, S., Dajani, A., Graham, M., Hawes, M., Karwa, V., Kifer, D., Kim, H., Leclerc, P., Machanavajjhala, A., Reiter, J., Rodrigues, R., Schmutte, I., Sexton, W., Singer, P., & Vilhuber, L. (2020). The modernization of statistical disclosure limitation at the U.S. Census Bureau. Working Paper Number CED-WP-2020-009.

Alam, M. J., Dostie, B., Drechsler, J., & Vilhuber, L. (2020). Applying data synthesis for longitudinal business data across three countries. *Statistics in Transition New Series*, *21*(4), 212–236.

An, D., & Little, R. J. A. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, *170*, 923–940.

Basak, B. and Sinha, B. (2024). Analysis of one-way anova model using synthetic data. *Sankhya B*, pages 1–27.

Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., Ueyama, K., & Williams, A. R. (2020). A synthetic supplemental public use file of low-income information return data: Methodology, utility, and privacy implications. In *Privacy in statistical databases: UNESCO chair in data privacy, international conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings* (pp. 257–270). Springer International Publishing.

Caiola, G., & Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, *3*(1), 27–42.

Drechsler, J. (2010). Using support vector machines for generating synthetic datasets. In *Privacy in statistical databases: UNESCO chair in data privacy, international conference, PSD 2010, Corfu, Greece, September 22–24, 2010. Proceedings* (pp. 148–161). Springer Berlin Heidelberg.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: Theory and implementation* (Vol. 201). Springer Science & Business Media.

Drechsler, J., & Haensch, A. C. (2023). 30 years of synthetic data. Preprint, arXiv:2304.02107.

Guin, A., Roy, A., & Sinha, B. (2022). Bayesian analysis of multiply imputed synthetic data under the multiple linear regression model. *International Journal of Statistical Sciences*, *22*(2), 25–38.

Guin, A., Roy, A., & Sinha, B. (2023). Bayesian analysis of singly imputed synthetic data under the multivariate normal model. *International Journal of Statistical Sciences*, *23*(2), 1–18.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E. S., Seri, G., & Wolf, P. (2010). Handbook on statistical disclosure control. *ESSnet on Statistical Disclosure Control*. https://cros.ec.europa.eu/system/files/2023-12/SDC_Handbook.pdf

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances. *Record Linkage Techniques*, *1997*, 248–267.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, *79*(3), 362–384.

Kinney, S. K., Reiter, J. P., & Miranda, J. (2014). *Improving the synthetic longitudinal business database*. US Census Bureau Center for Economic Studies.

Klein, M., & Sinha, B. (2015a). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models. *Sankhya B*, *77*, 293–311.

Klein, M., & Sinha, B. (2015b). Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models. *Journal of Privacy and Confidentiality*, *7*(1), 43–98.

Klein, M., & Sinha, B. (2015c). Likelihood-based finite sample inference for synthetic data based on exponential model. *Thailand Statistician*, *13*(1), 33–47.

Klein, M., & Sinha, B. (2015d). Likelihood-based inference for singly and multiply imputed synthetic data under a normal model. *Statistics & Probability Letters*, *105*, 168–175.

Klein, M. D., & and Datta, G. S. (2018). Statistical disclosure control via sufficiency under the multiple linear regression model. *Journal of Statistical Theory and Practice*, *12*, 100–110.

Klein, M., Moura, R., & Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, *83*, 273–287.

Koivu, A., Sairanen, M., Airola, A., & Pahikkala, T. (2020). Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association*, *27*(11), 1667–1674.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*, 407–426.

Ma, C., Tschiatschek, S., Turner, R., Hernández-Lobato, J. M., & Zhang, C. (2020). Vaem: A deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, *33*, 11237–11247.

Mishra, N., & Barui, S. (2022). Likelihood-based finite sample inference for synthetic data from pareto model. *REVSTAT-Statistical Journal*, *20*(5), 655–676.

Moura, R. P. (2016). *Likelihood-based inference for multivariate regression models using synthetic data*. Ph.D. Thesis, Universidade NOVA de Lisboa, Portugal.

Moura, R., Klein, M., Coelho, C. A., & Sinha, B. (2017). Inference for multivariate regression model based on synthetic data generated under fixed-posterior predictive sampling. *REVSTAT: Statistical Journal*, *15*(2), 155–186.

Moura, R., Klein, M., Zylstra, J., Coelho, C. A., & Sinha, B. (2021). Inference for multivariate regression model based on synthetic data generated using plug-in sampling. *Journal of the American Statistical Association*, *116*(534), 720–733.

Moura, R., Norouzirad, M., & Mazarei, D. (2023). Psinference: Inference for released plug-in sampling single synthetic dataset. *Computer software*. https://github.com/ricardomourarpm/PSinference.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment, 11*(10), 1071–1083.

Raab, G. M., Nowok, B., & Dibben, C. (2018). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, *7*(3), 67–97.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P., et al. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, *18*(4), 531.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, *29*, 181–188.

Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(1), 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, *131*(2), 365–377.

Reiter, J. P. (2005c). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.

Reiter, J. P., & Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics*, *28*(4), 583.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 1462–1471.

Reiter, J. P., Wang, Q., & Zhang, B. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, *6*(1).

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, *9*, 461–468.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M. U., & Sutton, C. (2017). Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in Neural Information Processing Systems, 30*.

Vardhan, L. V. H., & Kok, S. (2020). Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders. In *Proceedings of the workshop on economics of privacy and data labor at the 37th international conference on machine learning (ICML)*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.