RESEARCH REPORT SERIES
(Statistics #2019-06)

Multivariate Normal Inference based on Singly Imputed Synthetic Data under Plug-in Sampling

Martin Klein,
Ricardo Moura[1],
Bimal Sinha [2]

[1]CINAV, Portuguese Naval Academy; and CMA, Faculty of Sciences and Technology, Nova University of Lisbon
[2]Center for Statistical Research and Methodology, U.S. Census Bureau; and Department of Mathematics and Statistics University of Maryland, Baltimore County

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: May 7, 2019

Multivariate Normal Inference based on Singly Imputed Synthetic Data under Plug-in Sampling

Martin Klein, Ricardo Moura, Bimal Sinha

Abstract

In this paper we consider singly imputed synthetic data generated via plug-in sampling under the multivariate normal model. Based on the observed synthetic dataset, we derive a statistical test for the generalized variance, the sphericity test, a test for independence between two subsets of variables, and a test for the regression of one set of variables on the other. The procedures are based on finite sample theory. Some simulation studies are presented which confirm that the proposed procedures perform as expected.

Keywords: Multivariate normal, Pivotal quantity, Plug-in sampling, Statistical disclosure control, Tests for covariance structure.

1 Introduction

The release of synthetic data is a form of statistical disclosure control methodology that is based on principles of multiple imputation for missing data (Rubin, 1993). The fully synthetic data approach was originally proposed by Rubin (1993), and methodology for

Martin Klein, Center for Statistical Research and Methodology, U.S. Census Bureau.

Ricardo Moura, CINAV, Portuguese Naval Academy; and CMA, Faculty of Sciences and Technology, Nova University of Lisbon.

Bimal Sinha, Department of Mathematics and Statistics, University of Maryland, Baltimore County; and Center for Statistical Research and Methodology, U.S. Census Bureau.

Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

drawing inference based on fully synthetic data was developed by Raghunathan et al. (2003). The partially synthetic data approach was originally proposed by Little (1993), and methodology for drawing valid inference based partially synthetic data was developed by Reiter (2003) and Reiter (2005). Drechsler (2011) provides a detailed account of synthetic data.

The state of the art synthetic data methodology discussed above, uses concepts of multiple imputation, and hence requires the release of multiple synthetic datasets in order to obtain valid statistical inference. However, there are situations where it is desired to release only a single synthetic dataset; see for example Hawala (2008), Kinney et al. (2011), and Kinney et al. (2014). In a series of papers, Klein and Sinha (2015a,b,c,d); Moura et al. (2017, 2018) derived model-based inferential procedures that can accommodate the single imputation scenario. In this paper we extend this line of research by deriving procedures for drawing valid inference about some features of the underlying covariance structure of a multivariate normal population using a singly imputed synthetic dataset. It is assumed that the synthetic dataset is generated via plug-in sampling (PLS).

The outline of the rest of this paper is as follows. In Section 2 we review the basic modeling assumptions, and the PLS approach to generating synthetic data. In Section 3 we derive some inferential procedures for the covariance matrix based on the synthetic data. Specifically, in Section 3.1 we consider inference for the generalized variance; in Section 3.2 we consider the sphericity test; in Section 3.3 we consider testing independence of two subsets of variables; and in Section 3.4 we consider a test of the regression of one set of variables on the other. Section 4 contains some simulation studies to evaluate the proposed methodology, and Section 5 contains some concluding remarks.

2 Plug-in Sampling dataset generation

Before presenting the inferential procedures, a brief description is presented below on how to create a synthetic dataset via PLS method, under the multivariate normal model.

Consider $\boldsymbol{x} = (x_1, \ldots, x_p)'$ as the vector of variables assumed to be *sensitive*, i.e., which cannot be released to the public. Consequently, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, with $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$, $i = 1, \ldots, n$, will be the original dataset which will be considered confidential. The dataset is assumed to be normally distributed, that is,

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

where $n > p$.

From the original data, $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, the sample mean, and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}/(n-1)$, the sample covariance matrix, are computed, where $\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$ is the sample Wishart matrix, that is,

$$\boldsymbol{S}|_{\boldsymbol{\Sigma}} \sim W_p(n-1, \boldsymbol{\Sigma}). \tag{1}$$

In order to generate one synthetic dataset via PLS (Klein and Sinha, 2015b) one can obtain $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$, the new synthetic version of $\boldsymbol{X}$, by drawing

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

Analogous to $\bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}$, we define $\bar{\boldsymbol{v}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i$ as the PLS mean, and $\hat{\boldsymbol{\Sigma}}^{\star} = \boldsymbol{S}^{\star}/(n-1)$ as the PLS covariance matrix, where

$$\boldsymbol{S}^{\star} = \sum_{i=1}^{n}(\boldsymbol{v}_i - \bar{\boldsymbol{v}})(\boldsymbol{v}_i - \bar{\boldsymbol{v}})'. \tag{2}$$

Note that $(\bar{\boldsymbol{v}}, \boldsymbol{S}^{\star})$ are jointly sufficient for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and that $\bar{\boldsymbol{v}}$ and $\boldsymbol{\Sigma}^{\star} = \boldsymbol{S}^{\star}/(n-1)$ are the unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively (Klein and Sinha, 2015b).

Based on $(\bar{\boldsymbol{v}}, \boldsymbol{S}^\star)$, Klein and Sinha (2015b) discussed inferential procedures regarding the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and also the inferential test and construction of confidence sets for $\boldsymbol{\mu}$. In this paper, we develop appropriate inferential procedures for the generalized variance $|\boldsymbol{\Sigma}|$, sphericity test, test for independence of two subsets, and test for matrix of regression coefficients of one subset on the other, all based on synthetic data generated under plug-in sampling.

3 Tests for covariance structure

3.1 The Generalized Variance

The generalized variance is defined as $|\boldsymbol{\Sigma}|$ (Wilks, 1932) and it can be seen as a measure of the scatter of a dataset (Anderson, 1984). The following theorem will be the basis for the inferential procedures about the generalized variance.

Theorem 3.1. *Define*

$$T_1^\star = (n-1)^p \frac{|\boldsymbol{S}^\star|}{|\boldsymbol{\Sigma}|}. \tag{3}$$

where $\boldsymbol{S}^\star$ is defined as in (2). $T_1^\star$ is a pivotal random variable and its distribution can be obtained from the decomposition

$$T_1^\star \overset{st}{\sim} \left(\prod_{i=1}^{p} A_i\right)\left(\prod_{i=1}^{p} B_i\right) \tag{4}$$

where $\overset{st}{\sim}$ means 'stochastic equivalent to' and where $A_1, \ldots, A_p$, $B_1, \ldots, B_p$ are independently distributed such that $A_i \sim \chi^2_{n-i}$ and $B_i \sim \chi^2_{n-i}$ for $i = 1, \ldots, p$.

Proof. See Appendix A. □

The $(1-\alpha)$ level confidence interval for $|\boldsymbol{\Sigma}|$ is given by

$$\left(\frac{(n-1)^p \, |\boldsymbol{S}^\star|}{t^\star_{1,1-\alpha/2}}, \frac{(n-1)^p \, |\boldsymbol{S}^\star|}{t^\star_{1,\alpha/2}} \right)$$

where $t^\star_{1,\gamma}$ is the γth percentile of $T_1^\star$ in (3). For a given value $\delta_0 > 0$, a level α test for

$$H_0 : |\boldsymbol{\Sigma}| = \delta_0 \;\; vs. \;\; H_1 : |\boldsymbol{\Sigma}| \neq \delta_0$$

is to reject the H_0 if

$$\frac{(n-1)^p \left| \tilde{\boldsymbol{S}}^\star \right|}{\delta_0} \geq t^\star_{1,1-\alpha/2} \quad \text{or} \quad \frac{(n-1)^p \left| \tilde{\boldsymbol{S}}^\star \right|}{\delta_0} \leq t^\star_{1,\alpha/2},$$

where $\tilde{\boldsymbol{S}}$ is the observed value of $\tilde{\boldsymbol{S}}$.

To obtain the values of $t^\star_{1,\gamma}$, one may use Monte Carlo simulation as follows:

Given n and p

1. Generate $A_i \sim \chi^2_{n-i}$, $B_i \sim \chi^2_{n-i}$, $i = 1, ..., p$, independently.
2. Calculate $T_1^\star = \left(\prod_{i=1}^p A_i\right) \left(\prod_{i=1}^p B_i\right)$.
3. Repeat steps 1-2 M times and obtain M values of (3), which can be used to empirically determine the cut-off value.

3.2 The Sphericity Test

The sphericity test consists of testing if the population covariance matrix $\boldsymbol{\Sigma}$ is a diagonal matrix with all diagonal elements equal to some unknown value σ^2, that is, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p$. In practical terms, one may test if a set of random variables are all independent and share the same population variance, important condition for the analysis of variance scenario. Similar to the criterion found in Muirhead (1982) which was first derived by Mauchly (1940), we have the following theorem.

Theorem 3.2. *Define*

$$T_2^\star = \frac{\left|\boldsymbol{S}^\star\right|^{1/p}}{tr(\boldsymbol{S}^\star)} \tag{5}$$

where $\boldsymbol{S}^\star$ is defined as in (2). Under the assumption that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p$, the distribution of the pivotal random variable $T_2^\star$ can be obtained from the decomposition

$$T_2^\star \overset{st}{\sim} \frac{\left|\boldsymbol{\Omega}_1 \boldsymbol{\Omega}_2\right|^{1/p}}{tr\left(\boldsymbol{\Omega}_1 \boldsymbol{\Omega}_2\right)} \tag{6}$$

where $\overset{st}{\sim}$ means 'stochastic equivalent to' and $\boldsymbol{\Omega}_1 \sim W_p\left(n-1, \frac{\boldsymbol{I}_p}{n-1}\right)$ is independent of $\boldsymbol{\Omega}_2 \sim W_p\left(n-1, \boldsymbol{I}_p\right)$.

Proof. See Appendix B. □

A level α test for

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

reject the null hypothesis if, for an observed value $\tilde{\boldsymbol{S}}^\star$ of $\boldsymbol{S}^\star$, we have

$$\frac{\left|\tilde{\boldsymbol{S}}^\star\right|^{1/p}}{tr(\tilde{\boldsymbol{S}}^\star)} < t_{2,\alpha}^\star$$

where $\tilde{\boldsymbol{S}}^\star$ is an observed value of $\boldsymbol{S}^\star$, and $t_{2,\gamma}^\star$ is the γth percentile of $T_2^\star$ in (5).

To obtain the values of $t_{2,\gamma}^\star$, one may use Monte Carlo simulation as follows.

1. Generate $\boldsymbol{W}_1 \sim W_p(n-1, \frac{\boldsymbol{I}_p}{n-1})$, and $\boldsymbol{W}_2 \sim W_p(n-1, \boldsymbol{I}_p)$, independently.
2. Calculate $T_2^\star = \frac{|\boldsymbol{W}_1 \boldsymbol{W}_2|^{1/p}}{tr(\boldsymbol{W}_1 \boldsymbol{W}_2)}$
3. Repeat steps 1-2 M times and obtain M values of (5), which can be used to empirically determine the cut-off value of (5).

3.3 Testing the independence of two subsets of variables

Another feature of the covariance matrix that one might be interested in analyzing is if a set of variables is independent of other set, which is in fact the same as analyzing if the regression of one set on the other is equal to zero (Anderson, 1984).

For the construction of the procedures for the independence of subsets test, it is important to begin by considering partitions of $\boldsymbol{\Sigma}$, $\boldsymbol{S}$ and $\boldsymbol{S}^\star$ as follows:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \quad \boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix}, \quad \boldsymbol{S}^\star = \begin{bmatrix} \boldsymbol{S}^\star_{11} & \boldsymbol{S}^\star_{12} \\ \boldsymbol{S}^\star_{21} & \boldsymbol{S}^\star_{22} \end{bmatrix} \tag{7}$$

where $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{S}_{11}$ and $\boldsymbol{S}^\star_{11}$ are $p_1 \times p_1$ $(p_1 < p)$ matrices and let us consider $p_2 = p - p_1$.

We consider the problem of testing if the partition matrix $\boldsymbol{\Sigma}_{12}$ is a null matrix, and have the following theorem.

Theorem 3.3. *Define*

$$T_3^\star = \frac{|\boldsymbol{S}^\star|}{|\boldsymbol{S}^\star_{11}|\,|\boldsymbol{S}^\star_{22}|} \tag{8}$$

where $\boldsymbol{S}^\star$ is defined as in (2) and partitioned as in (7). Under the assumption that $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, the distribution of the pivotal random variable $T_3^\star$ can be obtained from the decomposition

$$T_3^\star \overset{st}{\sim} \frac{|\boldsymbol{\Omega}|}{|\boldsymbol{\Omega}_{11}|\,|\boldsymbol{\Omega}_{22}|} \tag{9}$$

where $\overset{st}{\sim}$ means 'stochastic equivalent to', for $\boldsymbol{\Omega} \sim W_p\left(n-1, \frac{\boldsymbol{W}}{n-1}\right)$ and $\boldsymbol{W} \sim W_p(n-1, \boldsymbol{I}_p)$, and where

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

partitioned in the same way as $\boldsymbol{\Sigma}$.

Proof. See Appendix C. □

The assumption of independence between two sets of variables, that is, the test of hypothesis

$$H_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0} \;\; vs. \;\; H_1 : \boldsymbol{\Sigma}_{12} \neq \mathbf{0},$$

will be rejected if for an observed value $\tilde{\boldsymbol{S}}^\star$ of $\boldsymbol{S}^\star$ we have

$$\frac{\left|\tilde{\boldsymbol{S}}^\star\right|}{\left|\tilde{\boldsymbol{S}}^\star_{11}\right|\left|\tilde{\boldsymbol{S}}^\star_{22}\right|} < t^\star_{3,\alpha}$$

for α-significance level, where $t^\star_{3,\gamma}$ is the γth percentile of $T^\star_3$ in (8).

To obtain the values of $t^\star_{3,\gamma}$, one may use Monte Carlo simulation as follows:

Given n, p and p_1

1. Generate $\boldsymbol{W}_1 \sim W_p(n-1, \boldsymbol{I}_p)$, and $\boldsymbol{W}_2 \sim W_p(n-1, \frac{\boldsymbol{W}_1}{n-1})$, independently.

2. Obtain

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

   where $\boldsymbol{W}_{11}$ is a $p_1 \times p_1$ matrix.

3. Calculate $T^\star_3 = \frac{|\boldsymbol{W}_2|}{|\boldsymbol{W}_{11}||\boldsymbol{W}_{22}|}$

4. Repeat steps 1-3 M times and obtain M values of (8), which can be used to empirically determine the cut-off value of (8).

3.4 Test for the regression of one set of variables on the other

In the previous section, one may find the procedure to test a very particular case of the regression of one set of variates on the other, but one might be interested in characterizing the dependence between two sets of variables, that is, if a subset of variables are constrained

to have a linear relation, in some way, to the other subset. (Anderson, 1984; Muirhead, 1982). Thus, one is interested in testing if $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ is equal to some $\boldsymbol{\Delta_0}$.

The following Theorem will be important to construct an inferential procedure for the above matrix parameter.

Theorem 3.4. *Define*

$$T_4^\star = \frac{\left|\left(\boldsymbol{S}_{12}^\star \left(\boldsymbol{S}_{22}^\star\right)^{-1} - \boldsymbol{\Delta}\right) \boldsymbol{S}_{22}^\star \left(\boldsymbol{S}_{12}^\star \left(\boldsymbol{S}_{22}^\star\right)^{-1} - \boldsymbol{\Delta}\right)'\right|}{|\boldsymbol{S}_{11.2}^\star|} \tag{10}$$

where $\boldsymbol{S}^\star$ is defined as in (2), partitioned as in (7) and

$$\boldsymbol{S}_{11.2}^\star = \boldsymbol{S}_{11}^\star - \boldsymbol{S}_{12}^\star \left(\boldsymbol{S}_{22}^\star\right)^{-1} \boldsymbol{S}_{21}^\star. \tag{11}$$

For $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ where $\boldsymbol{\Sigma}$ partitioned as in (7), the distribution of the pivotal random variable $T_4^\star$ can be obtained from the decomposition, for $p_1 \leq p_2$,

$$T_4^\star \overset{st}{\sim} \frac{\left|\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\right|}{\left|\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\right|} \tag{12}$$

where $\overset{st}{\sim}$ means 'stochastic equivalent to', for $\boldsymbol{\Omega} \sim W_p\left(n-1, \frac{\boldsymbol{W}}{n-1}\right)$ and $\boldsymbol{W} \sim W_p(n-1, \boldsymbol{I}_p)$, and where

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

partitioned in the same way as $\boldsymbol{\Sigma}$.

Proof. See Appendix D. □

In order to test

$$H_0 : \boldsymbol{\Delta} = \boldsymbol{\Delta}_0 \;\; vs. \;\; H_1 : \boldsymbol{\Delta} \neq \boldsymbol{\Delta}_0$$

one should reject the null hypothesis if for an observed value $\tilde{\boldsymbol{S}}^{\star}$ of $\boldsymbol{S}^{\star}$ we have

$$\frac{\left|\left(\tilde{\boldsymbol{S}}_{12}^{\star}\left(\tilde{\boldsymbol{S}}_{22}^{\star}\right)^{-1}-\boldsymbol{\Delta}_0\right)\tilde{\boldsymbol{S}}_{22}^{\star}\left(\tilde{\boldsymbol{S}}_{12}^{\star}\left(\tilde{\boldsymbol{S}}_{22}^{\star}\right)^{-1}-\boldsymbol{\Delta}_0\right)'\right|}{\left|\tilde{\boldsymbol{S}}_{11.2}^{\star}\right|} > t_{4,1-\alpha}^{\star}$$

for α-significance level, where $t_{4,\gamma}^{\star}$ is the γth percentile of $T_4^{\star}$ in (10).

To obtain the values of $t_{3,\gamma}^{\star}$, one may use Monte Carlo simulation as follows:

Given n, p and p_1

1. Generate $\boldsymbol{W}_1 \sim W_p(n-1, \boldsymbol{I}_p)$, and $\boldsymbol{W}_2 \sim W_p(n-1, \frac{\boldsymbol{W}_1}{n-1})$, independently.

2. Obtain

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

   where $\boldsymbol{W}_{11}$ is a $p_1 \times p_1$ matrix.

3. Calculate $D = \boldsymbol{W}_{12}\boldsymbol{W}_{22}^{-1}\boldsymbol{W}_{21}$

4. Calculate $T_4^{\star} = \frac{|D|}{|\boldsymbol{W}_{11}-D|}$

5. Repeat steps 1-4 M times and obtain M values of (10), which can be used to empirically determine its cut-off value.

4 Simulations

We now provide a set of simulations to evaluate the performance of the four tests presented in this work. All the simulations were performed using software Python® and codes can be accessed at `https://github.com/ricardomourarpm/PLS_VarianceStructure`.

Under the multivariate normal model with $p = 4$ $\boldsymbol{\mu} = (1, 2, 3, 4)'$ we consider four different covariance matrices

$$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_4, \quad \boldsymbol{\Sigma}_2 = 5\boldsymbol{I}_4, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \quad \boldsymbol{\Sigma}_4 = \begin{pmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0.2 \\ 0 & 0 & 0.2 & 4 \end{pmatrix} \tag{13}$$

chosen this way to illustrate the performance for all tests presented in the previous section. We used Monte Carlo simulations with 10^5 iterations and estimated probability of not rejecting the null hypothesis when the null hypothesis is true ($1-P$(Type I error)) for nominal $\alpha = 0.05$ significance level. For every iteration the PLS single imputed dataset is created using the method in Section 2 and the sample sizes considered are $n = 10, 20, 100, 500$.

Table 1 shows for any value of n and for selected covariance matrices defined in (13) the estimated coverage values for:

- test for the Generalized Variance found in Section 3.1 under the column *Generalized Variance* and selected $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$;
- Sphericity test found in Section 3.2 under the column *Sphericity* and selected $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$;
- Independence test found in Section 3.3 under the column *Independence* and selected $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_4$, for $p_1 = 1$ and $p_1 = 2$, respectively;
- Test for the regression of one set of variables on the other found in Section 3.4 under the column *Regression* and selected $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$, for $p_1 = 2$ and $p_1 = 1$, respectively.

We may observe that all values in the table are approximately equal to the nominal value 0.95, as expected.

	Gener. Variance		*Sphericity*		*Independence*		*Regression*	
	$\boldsymbol{\Sigma}_3$	$\boldsymbol{\Sigma}_4$	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_2$	$\boldsymbol{\Sigma}_1$	$\boldsymbol{\Sigma}_4$	$\boldsymbol{\Sigma}_3$	$\boldsymbol{\Sigma}_4$
n					$p_1 = 1$	$p_1 = 2$	$p_1 = 2$	$p_1 = 1$
10	0.949	0.950	0.950	0.950	0.951	0.951	0.949	0.952
20	0.948	0.949	0.951	0.950	0.950	0.950	0.950	0.951
100	0.950	0.951	0.949	0.950	0.951	0.951	0.950	0.951
500	0.950	0.950	0.949	0.950	0.949	0.950	0.950	0.951

Table 1: Estimates of $1 - P(\text{Type I error})$ for the tests of Sections 3.1, 3.2, 3.3, and 3.4 for $n = 10, 20, 100, 500$, $p_1 = 1, 2$, $\boldsymbol{\mu} = (1, 2, 3, 4)'$ and $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_3$ and $\boldsymbol{\Sigma}_4$ defined in (13).

5 Concluding Remarks

Under the assumption of a multivariate normal distribution on the original data, we derived appropriate likelihood-based exact inference procedures using singly imputed synthetic data generated under plug-in sampling method. In particular, inference procedures have been developed for the Generalized Variance, the Sphericity test, the test of independence of one subset of variables from other subset, and also for the regression vector of one subset on the other. Simulation studies demonstrate that the four procedures perform as expected, for all sample sizes, even for small sample sizes.

Acknowledgments. Ricardo Moura thanks FCT (Portuguese Foundation for Science and Technology) project UID/MAT/00297/2013 awarded through CMA/UNL. Martin Klein and Bimal Sinha thank Eric Slud, William Winkler, and Tommy Wright for encouragement. The authors thank Thomas Mathew for some helpful comments.

Appendix A Proof of Theorem 3.1

From the proof of Theorem 3.1 in Klein and Sinha (2015b), we have that

$$\boldsymbol{S}^{\star} \sim W_p\left(n-1, \frac{\boldsymbol{S}}{n-1}\right) \tag{14}$$

which by Theorem 3.2.15 in Muirhead (1982) leads to the fact that

$$(n-1)^p \frac{|\boldsymbol{S}^{\star}|}{|\boldsymbol{S}|}|_{\boldsymbol{S}} \sim \prod_{i=1}^{p} \chi^2_{n-i}. \tag{15}$$

The above distributional result is independent of $\boldsymbol{S}$. In turn, from Theorem 7.5.3 in Anderson (1984), it follows that

$$\frac{|\boldsymbol{S}|}{|\boldsymbol{\Sigma}|} \sim \prod_{i=1}^{p} \chi^2_{n-i} \tag{16}$$

where the χ^2_{n-i} variables are all independent. Therefore it is easy to observe that $T_1^{\star}$ in (3) will have the distribution of the independent product of (15) and (16), completing the proof.

Appendix B Proof of Theorem 3.2

From (1), (14) and from Theorem 3.2.5. in Muirhead (1982), we observe that

$$\tilde{\boldsymbol{S}}^{\star} = \boldsymbol{S}^{-1/2}\boldsymbol{S}^{\star}\boldsymbol{S}^{-1/2}|_{\boldsymbol{S}} \sim W_p\left(n-1, \frac{\boldsymbol{I}_p}{n-1}\right) \tag{17}$$

and

$$\tilde{\tilde{\boldsymbol{S}}}^{\star} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}\boldsymbol{\Sigma}^{-1/2}|_{\boldsymbol{\Sigma}} \sim W_p\left(n-1, \boldsymbol{I}_p\right). \tag{18}$$

Considering (5), we will have

$$T_2^\star = \frac{|\boldsymbol{S}^\star|^{1/p}}{tr\left(\boldsymbol{S}^\star\right)} = \frac{\left|\boldsymbol{S}^{1/2}\tilde{\boldsymbol{S}}^\star\boldsymbol{S}^{1/2}\right|^{1/p}}{tr\left(\boldsymbol{S}^{1/2}\tilde{\boldsymbol{S}}^\star\boldsymbol{S}^{1/2}\right)} = \frac{\left|\tilde{\boldsymbol{S}}^\star\boldsymbol{S}\right|^{1/p}}{tr\left(\tilde{\boldsymbol{S}}^\star\boldsymbol{S}\right)} = \frac{\left|\tilde{\boldsymbol{S}}^\star\boldsymbol{\Sigma}^{1/2}\ \tilde{\tilde{\boldsymbol{S}}}^\star\boldsymbol{\Sigma}^{1/2}\right|^{1/p}}{tr\left(\tilde{\boldsymbol{S}}^\star\boldsymbol{\Sigma}^{1/2}\ \tilde{\tilde{\boldsymbol{S}}}^\star\boldsymbol{\Sigma}^{1/2}\right)}$$
$$= \frac{\left|\tilde{\boldsymbol{S}}^\star\tilde{\tilde{\boldsymbol{S}}}^\star\boldsymbol{\Sigma}\right|^{1/p}}{tr\left(\tilde{\boldsymbol{S}}^\star\tilde{\tilde{\boldsymbol{S}}}^\star\boldsymbol{\Sigma}\right)}.$$

Under the assumption that $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_p$, one may conclude that $T_2^\star$ will have the same distribution as

$$\frac{|\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2|^{1/p}}{tr\left(\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2\right)}$$

where $\boldsymbol{\Omega}_1 \sim W_p\left(n-1, \frac{\boldsymbol{I}_p}{n-1}\right)$ is independent of $\boldsymbol{\Omega}_2 \sim W_p\left(n-1, \boldsymbol{I}_p\right)$.

Appendix C Proof of Theorem 3.3

Considering $\boldsymbol{\Sigma}$, $\boldsymbol{S}$ and $\boldsymbol{S}^\star$ partitioned as in (7) and from Proposition 1.3.2 in Kollo and Rosen (2005), we have that

$$|\boldsymbol{S}^\star| = |\boldsymbol{S}_{22}^\star|\,|\boldsymbol{S}_{11.2}^\star|$$

where $\boldsymbol{S}_{11.2}^\star = \boldsymbol{S}_{11}^\star - \boldsymbol{S}_{12}^\star\left(\boldsymbol{S}_{22}^\star\right)^{-1}\boldsymbol{S}_{21}^\star$. Therefore, we will have $T_3^\star$ in (8) as

$$T_3^\star = \frac{|\boldsymbol{S}_{11.2}^\star|}{|\boldsymbol{S}_{11}^\star|} = \frac{|\boldsymbol{S}_{11.2}^\star|}{\left|\boldsymbol{S}_{11.2}^\star + \boldsymbol{S}_{12}^\star\left(\boldsymbol{S}_{22}^\star\right)^{-1}\boldsymbol{S}_{21}^\star\right|}$$
$$= \frac{|\boldsymbol{S}_{11.2}^\star|\,|\boldsymbol{S}_{11.2}|^{-1}}{\left|\boldsymbol{S}_{11.2}^\star + \boldsymbol{S}_{12}^\star\left(\boldsymbol{S}_{22}^\star\right)^{-1}\boldsymbol{S}_{21}^\star\right|\,|\boldsymbol{S}_{11.2}|^{-1}}$$
$$= \frac{\left|\tilde{\boldsymbol{S}}_{11.2}^\star\right|}{\left|\tilde{\boldsymbol{S}}_{11.2}^\star + \boldsymbol{S}_{11.2}^{-1/2}\boldsymbol{S}_{12}^\star\left(\boldsymbol{S}_{22}^\star\right)^{-1}\boldsymbol{S}_{21}^\star\boldsymbol{S}_{11.2}^{-1/2}\right|}.$$

where $\tilde{\boldsymbol{S}}^{\star}_{11.2} = \boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}^{\star}_{11.2}\boldsymbol{S}^{-1/2}_{11.2}$ and $\tilde{\boldsymbol{S}}_{11.2} = \boldsymbol{S}_{11} - \boldsymbol{S}_{12}\left(\boldsymbol{S}_{22}\right)^{-1}\boldsymbol{S}_{21}$. From Theorems 3.2.5 and 3.2.10 in Muirhead (1982), we have

$$\tilde{\boldsymbol{S}}^{\star}_{11.2}|_{\boldsymbol{S}} \sim W_{p_1}\left(n-1-p_2, \frac{\boldsymbol{I}_{p_1}}{n-1}\right). \tag{19}$$

From the previous result we conclude that we only need to focus on the random variable

$$\boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}^{\star}_{12}\left(\boldsymbol{S}^{\star}_{22}\right)^{-1}\boldsymbol{S}^{\star}_{21}\boldsymbol{S}^{-1/2}_{11.2} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$$

where $\Gamma = \boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}^{\star}_{12}\left(\boldsymbol{S}^{\star}_{22}\right)^{-1/2}$.

From Theorem 3.2.10 in Muirhead (1982), we have that

$$\boldsymbol{S}^{\star}_{12}|_{\boldsymbol{S},\boldsymbol{S}^{\star}_{22}} \sim N_{p_1,p_2}\left(\boldsymbol{S}_{12}\boldsymbol{S}^{-1}_{22}\boldsymbol{S}^{\star}_{22}, \frac{\boldsymbol{S}_{11.2}}{n-1}\otimes\boldsymbol{S}^{\star}_{22}\right) \tag{20}$$

a matrix normal distribution, which leads to the fact that

$$\boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}^{\star}_{12}|_{\boldsymbol{S},\boldsymbol{S}^{\star}_{22}} \sim N_{p_1,p_2}\left(\boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}_{12}\boldsymbol{S}^{-1}_{22}\boldsymbol{S}^{\star}_{22}, \frac{\boldsymbol{I}_{p1}}{n-1}\otimes\boldsymbol{S}^{\star}_{22}\right)$$

and, consequently to

$$\boldsymbol{\Gamma} = \boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}^{\star}_{12}(\boldsymbol{S}^{\star}_{22})^{-1/2}|_{\boldsymbol{S},\boldsymbol{S}^{\star}_{22}} \sim N_{p_1,p_2}\left(\boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}_{12}\boldsymbol{S}^{-1}_{22}(\boldsymbol{S}^{\star}_{22})^{-1/2}, \frac{\boldsymbol{I}_{p1}}{n-1}\otimes\boldsymbol{I}_{p_2}\right).$$

Therefore, we will have that, for $p_1 \leq p_2$,

$$\boldsymbol{\Gamma}\boldsymbol{\Gamma}'|_{\boldsymbol{S},\boldsymbol{S}^{\star}_{22}} \sim W_{p_1}(p_2, \frac{\boldsymbol{I}_{p_1}}{n-1}, \boldsymbol{\Delta})$$

with non-central matrix $\boldsymbol{\Delta} = \boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}_{12}\boldsymbol{S}^{-1}_{22}(\boldsymbol{S}^{\star}_{22})^{-1}\boldsymbol{S}^{-1}_{22}\boldsymbol{S}_{21}\boldsymbol{S}^{-1/2}_{11.2}$. Let us now observe that we may decompose the non-central matrix as $\boldsymbol{\Delta} = \boldsymbol{\Delta}_1\boldsymbol{\Delta}_2\boldsymbol{\Delta}'_1$, for $\boldsymbol{\Delta}_1 = \boldsymbol{S}^{-1/2}_{11.2}\boldsymbol{S}_{12}\boldsymbol{S}^{-1/2}_{22}$ and $\boldsymbol{\Delta}_2 = \boldsymbol{S}^{-1/2}_{22}(\boldsymbol{S}^{\star}_{22})^{-1}\boldsymbol{S}^{-1/2}_{22}$. Since, again from Theorem 3.2.10 in Muirhead (1982), one has

$$\boldsymbol{S}_{12}|_{\boldsymbol{\Sigma},\boldsymbol{S}_{22}} \sim N_{p_1,p_2}\left(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}^{-1}_{22}\boldsymbol{S}_{22}, \boldsymbol{\Sigma}_{11.2}\otimes\boldsymbol{S}_{22}\right) \tag{21}$$

we conclude that, under the hypothesis of $\boldsymbol{\Sigma}_{12} = 0$,

$$\boldsymbol{\Delta}_1|_{\boldsymbol{\Sigma},\boldsymbol{S}_{22}} \sim N_{p_1,p_2}\left({}_{p_1}\mathbf{0}_{p_2}, \boldsymbol{S}_{11.2}^{-1/2}\boldsymbol{\Sigma}_{11.2}\boldsymbol{S}_{11.2}^{-1/2} \otimes \boldsymbol{I}_{p_2}\right).$$

Since $\boldsymbol{S}_{11.2}|\boldsymbol{\Sigma}_{11.2} \sim W_{p_1}(\boldsymbol{\Sigma}_{11.2}, n-1-p_2)$, we observe that the first element of the Kronecker product, namely $\boldsymbol{S}_{11.2}^{-1/2}\boldsymbol{\Sigma}_{11.2}\boldsymbol{S}_{11.2}^{-1/2}$ is the inverse of a Wishart matrix, independent of $\boldsymbol{\Sigma}, \boldsymbol{S}_{22}$.

Concerning $\boldsymbol{\Delta}_2$, it is easy to show that it is independent of $\boldsymbol{S}$ or $\boldsymbol{\Sigma}$. From Theorem 3.2.10, we have

$$\boldsymbol{S}_{22}^{\star}|_{\boldsymbol{S}_{22}} \sim W_{p_2}(n-1, \frac{\boldsymbol{S}_{22}}{n-1}) \tag{22}$$

which leads to

$$\boldsymbol{\Delta}_2 \sim W_{p_2}(n-1, \frac{\boldsymbol{I}_{p_2}}{n-1}).$$

In conclusion, $T_3^{\star}$ under the hypothesis $\boldsymbol{\Sigma}_{12} = 0$ is a decomposition of random variables whose distributions are independent from the original dataset and from the PLS single synthetic dataset.

One may note that the proof for this theorem was executed considering $p_1 \leq p_2$. For $p_1 > p_2$ it is just a matter of considering initially the decomposition

$$|\boldsymbol{S}^{\star}| = |\boldsymbol{S}_{11}^{\star}|\,|\boldsymbol{S}_{22.1}^{\star}|$$

where $\boldsymbol{S}_{22.1}^{\star} = \boldsymbol{S}_{22}^{\star} - \boldsymbol{S}_{21}^{\star}(\boldsymbol{S}_{11}^{\star})^{-1}\boldsymbol{S}_{12}^{\star}$, and proceed analogously.

Appendix D Proof of Theorem 3.4

Let us consider, one more time, $\boldsymbol{\Sigma}$, $\boldsymbol{S}$ and $\boldsymbol{S}^{\star}$ partitioned as in (7). Let

$$\boldsymbol{\Delta}^{\star} = \boldsymbol{S}_{12}^{\star}\left(\boldsymbol{S}_{22}^{\star}\right)^{-1}, \quad \hat{\boldsymbol{\Delta}} = \boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1} \quad \text{and} \quad \boldsymbol{\Delta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}.$$

From (20) and (21), we immediately obtain

$$\boldsymbol{\Delta}^{\star}|_{\boldsymbol{S},\boldsymbol{S}_{22}^{\star}} \sim N_{p_1,p_2}\left(\hat{\boldsymbol{\Delta}}, \frac{\boldsymbol{S}_{11.2}}{n-1} \otimes (\boldsymbol{S}_{22}^{\star})^{-1}\right)$$

and

$$\hat{\boldsymbol{\Delta}}|_{\boldsymbol{S}_{22}} \sim N_{p_1,p_2}\left(\boldsymbol{\Delta}, \boldsymbol{\Sigma}_{11.2} \otimes \boldsymbol{S}_{22}^{-1}\right).$$

Since

$$\left(\boldsymbol{\Delta}^{\star} - \boldsymbol{\Delta}\right)|_{\boldsymbol{S},\boldsymbol{S}^{\star}} \sim N_{p_1,p_2}\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}, \frac{\boldsymbol{S}_{11.2}}{n-1} \otimes (\boldsymbol{S}_{22}^{\star})^{-1}\right)$$

we get

$$\boldsymbol{\Gamma}^{\star} = \boldsymbol{S}_{11.2}^{-1/2}\left(\boldsymbol{\Delta}^{\star} - \boldsymbol{\Delta}\right)(\boldsymbol{S}_{22}^{\star})^{1/2}|_{\boldsymbol{S},\boldsymbol{S}^{\star}} \sim N_{p_1,p_2}\big(\boldsymbol{S}_{11.2}^{-1/2}\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right)(\boldsymbol{S}_{22}^{\star})^{1/2}, \frac{\boldsymbol{I}_{p_1}}{n-1} \otimes \boldsymbol{I}_{p_2}\big) \quad (23)$$

and conclude that, for $p_1 \leq p_2$,

$$\boldsymbol{\Gamma}^{\star}\boldsymbol{\Gamma}^{\star\prime}|_{\boldsymbol{S},\boldsymbol{S}^{\star}} \sim W_{p_1}\left(p_2, \frac{\boldsymbol{I}_{p_1}}{n-1}, \boldsymbol{\Omega}\right)$$

with non-central matrix $\boldsymbol{\Omega} = \boldsymbol{S}_{11.2}^{-1/2}\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right)\boldsymbol{S}_{22}^{\star}\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right)'\boldsymbol{S}_{11.2}^{-1/2}$. Let us now consider the decomposition $\boldsymbol{\Omega} = \boldsymbol{\Omega}_1\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1'$ where

$$\boldsymbol{\Omega}_1 = \boldsymbol{S}_{11.2}^{-1/2}\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right)\boldsymbol{S}_{22}^{1/2} \text{ and } \boldsymbol{\Omega}_2 = \boldsymbol{S}_{22}^{-1/2}\boldsymbol{S}_{22}^{\star}\boldsymbol{S}_{22}^{-1/2}$$

Since

$$\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right)|_{\boldsymbol{S},\boldsymbol{S}^{\star}} \sim N_{p_1,p_2}\left(\boldsymbol{0}_{p_1,p_2}, \boldsymbol{\Sigma}_{11.2} \otimes \boldsymbol{S}_{22}^{-1}\right)$$

we get

$$\boldsymbol{\Omega}_1|_{\boldsymbol{S},\boldsymbol{S}^{\star}} \sim N_{p_1,p_2}\left(\boldsymbol{0}_{p_1,p_2}, \boldsymbol{S}_{11.2}^{-1/2}\boldsymbol{\Sigma}_{11.2}\boldsymbol{S}_{11.2}^{-1/2} \otimes \boldsymbol{I}_{p_2}\right)$$

which is in fact independent of $\boldsymbol{S}^{\star}$, $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$ due to the independence of $\boldsymbol{S}_{11.2}^{-1/2}\boldsymbol{\Sigma}_{11.2}\boldsymbol{S}_{11.2}^{-1/2}$ as seen in the proof of Theorem 3.3 in Appendix C. From (22), we immediately observe that

$$\boldsymbol{\Omega}_2 \sim W_{p_2}(n-1, \frac{\boldsymbol{I}_{p_2}}{n-1})$$

also independent of $\boldsymbol{S}^{\star}$, $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$. Thus the distribution of $\boldsymbol{\Omega}$ is independent of $\boldsymbol{S}^{\star}$, $\boldsymbol{S}$ and $\boldsymbol{\Sigma}$, and consequently is also the distribution of $\boldsymbol{\Gamma}^{\star}\boldsymbol{\Gamma}^{\star'}$.

Let us now consider

$$\tau^{\star} = \frac{\left|\left(\boldsymbol{\Delta}^{\star} - \boldsymbol{\Delta}\right) \boldsymbol{S}_{22}^{\star} \left(\boldsymbol{\Delta}^{\star} - \boldsymbol{\Delta}\right)'\right|}{\left|\boldsymbol{S}_{11.2}^{\star}\right|} = \frac{\left|\boldsymbol{\Gamma}^{\star}\boldsymbol{\Gamma}^{\star'}\right|}{\left|\tilde{\boldsymbol{S}}_{11.2}^{\star}\right|}.$$

Recalling $\boldsymbol{\Gamma}^{\star}$ as defined in (23) and $\tilde{\boldsymbol{S}}_{11.2}^{\star}$ as defined and distributed as in (19), it is obvious that $\tau^{\star}$ will have a distribution independent of any parameter, besides n and p.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (3rd ed. ed.). Wiley.

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

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. *Proceedings of the Joint Statistical Meetings, American Statistical Association*, 1345–1350.

Kinney, S., J. Reiter, and J. Miranda (2014). Synlbd 2.0: Improving the synthetic longitudinal business database. *Statistical Journal of the International Association for Official Statistics 30*, 129–135.

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

Klein, M. and B. Sinha (2015a). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models. *Sankhya B 77-B*(2), 293–311.

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

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

Kollo, T. and D. Rosen (2005). *Advanced Multivariate Statistics with Matrices*. Springer.

Little, R. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics 11*(2), 204–209.

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

Moura, R., M. Klein, J. Zylstra, C. Coelho, and B. Sinha (2018). Inference for multivariate regression model based on synthetic data generated using plug-in sampling. *Center for Statistical Research and Methodology (CSRM) Research Report Series (#2018-02). U.S. Census Bureau*.

Muirhead, T. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc.

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

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

Reiter, J. (2005). Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of Royal Statistical Society 168*, 185–205.

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

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 471–494.