Ricardo Pinto Moura

Licenciado em Matemática - Ramo Educacional

Likelihood-based Inference for Multivariate Regression Models using Synthetic Data

Dissertação para obtenção do Grau de Doutor em

Estatística e Gestão do Risco

Co-orientadores: Carlos Agra Coelho, Associate Professor with Habilitation, Faculty of Sciences and Technology NOVA University of Lisbon

Bimal K. Sinha, Presidential Research Professor, University of Maryland Baltimore County

Júri

Presidente:	Prof. Doutor José Paulo Barbosa Mota
Arguentes:	Prof.ª Doutora Maria Ivette Leal de Carvalho Gomes
	Prof. Doutor Pedro José Ramos Moreira de Campos
Vogais:	Prof. Doutor António Manuel Pacheco Pires
	Prof. Doutor Filipe José Gonçalves Pereira Marques
	Prof. Doutor Bimal K. Sinha

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

December, 2016

Likelihood-based Inference for Multivariate Regression Models using Synthetic Data

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "unlthesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2]. [1] https://github.com/joaomlourenco/unlthesis [2] http://www.di.fct.unl.pt

To Veronica and Joel
for their love and full support.

Acknowledgements

Firstly, I would like to thank my adviser Prof. Dr. Carlos Agra Coelho for giving me the honor of being his advisee, for trusting me, for his great friendship and support and for giving the brightest advices every time. I would like to also thank my co-adviser Prof. Dr. Bimal Sinha for introducing me to the SDC world and to the multiple imputation technique that would be the foundation of this thesis, for trusting me the task of developing this work, for sharing with me his expertise and also for his friendship.

To my friends in the Centre of Mathematics and Aplications of the Nova University of Lisbon, for making me feel at home in the center and for their support.

To UMBC staff and students for welcoming me to their family and to make me a member of that family.

To Fulbright, by choosing me for the Research Grant which allowed me to build the core work near the Prof. Dr. Bimal Sinha at UMBC and for allowing me to experience the culture of the U.S.A..

To Dr. Martin Klein, for letting me be in his Statistical Computing classes and for his support.

To my parents, for their sacrifice in all the rough times showing that there's no gain without work.

Most importantly, I want to thank my brother Joel, for always making me believe that I'm an hero, for looking at me with such pride that I could never stop going further, and, finally, to my girlfriend Verónica, that pushed me into making the PhD, for believing that my destiny was yet not defined and that I was destined for greater accomplishments, for cheering me every time that I seemed down and things were not running well and for making me believe that I would always surpass the incoming problems.

Abstract

Likelihood-based exact inference procedures are derived for the multivariate regression model, for singly and multiply imputed synthetic data generated via Posterior Predictive Sampling (PPS), via a newly proposed sampling method, which will be called Fixed-Posterior Predictive Sampling (FPPS), and via Plug-in sampling. By contemplating the single imputation case, the new developed procedures fill the gap in the existing literature where inferential methods are only available for multiple imputation and, by being based in exact distributions, it may even be applied to cases where the sample size is small. Simulation studies compare the results obtained from all the proposed exact inferential procedures and also compare these with the results obtained from the adaptation of Reiter's combination rule to multiply imputed synthetic datasets. An application using U.S. 2000 Current Population Survey data is discussed and measures of privacy are presented and compared among all methods.

Keywords: Finite sample inference, Maximum likelihood estimation, Pivotal quantity, Partially synthetic data, Statistical Disclosure Control, Unbiased estimators

Resumo

Procedimentos inferenciais baseados em funções de verosimilhança são deduzidos para o Modelo de Regressão Linear Multivariado, para dados sintéticos de imputação única e de imputação múltipla gerados via *Posterior Predictive Sampling* (PPS), via um novo método, que se denominará por *Fixed-Posterior Predictive Sampling* (FPPS), e via *Plug-in Sampling*. Ao contemplar o caso de imputação única, os novos procedimentos desenvolvidos preenchem um vazio na literatura existente onde métodos inferenciais estão disponíveis exclusivamente para casos de imputação múltipla e, como se baseiam em distribuições exatas, podem ainda assim ser aplicados a casos onde a dimensão da amostra é pequena. O estudo de simulações permite a comparação de todos os resultados provenientes dos procedimentos exatos propostos como também a comparação destes com os resultados obtidos aquando da aplicação da regra combinatória de Reiter a dados sintéticos de múltipla imputação. É discutida uma aplicação usando dados da *U.S. 2000 Current Population Survey* e medidas de privacidade são apresentadas e comparadas entre todos os métodos.

Palavras-chave: Inferência amostral/estatística finita, Estimação de máxima verosimilhança, Quantidades pivotais, Dados parcialmente sintetizados, Controlo da divulgação estatística, Estimadores centrados

Contents

List of Figures xv

List of Tables xvii

Listings xix

1 General Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Generating Synthetic Data . . . 4
  - 1.2.1 Posterior Predictive Sampling (PPS) . . . 5
  - 1.2.2 Fixed-Posterior Predictive Sampling (FPPS) . . . 5
  - 1.2.3 Plug-in Sampling . . . 5
- 1.3 The Multivariate Linear Regression Model . . . 5
- 1.4 An important Lemma . . . 7

2 Inference for Multivariate Regression Model based on synthetic data generated via PPS and FPPS 9
- 2.1 Posterior Predictive Sampling (PPS) . . . 9
  - 2.1.1 Single Imputation: Posterior Predictive Sampling Method . 10
- 2.2 Multiple imputation: Posterior Predictive Sampling . . . 20
  - 2.2.1 Reiter's adapted methodology . . . 21
  - 2.2.2 Exact inference for multiple imputation cases based in single imputation inference . . . 22
- 2.3 Multiple Imputation: Fixed-Posterior Predictive Sampling (FPPS) . 24
  - 2.3.1 A First Procedure . . . 25
  - 2.3.2 A Second Procedure . . . 27

3 Inference for Multivariate Regression Model based on synthetic data generated via Plug-in Sampling 35
- 3.1 Single Imputation: Plug-in Sampling . . . 36
- 3.2 Multiple imputation: Plug-in Sampling . . . 42

3.2.1 Reiter's adapted Methodology 43
3.2.2 A First New Procedure 43
3.2.3 A Second New Procedure 46

4 Radius and Accuracy **51**
4.1 Measuring the confidence sets 51
4.1.1 Volume of the confidence sets 52
4.1.2 *Radius* of the confidence sets 54
4.2 Simulation Studies 57
4.2.1 Accuracy of Procedures proposed in Chapters 2 and 3 . . . 58
4.2.2 *Radius* of the confidence sets when using PPS, FPPS and Plug-in Sampling cases 60

5 An application to Current Population Survey (CPS) data and Risk Level Comparison **63**
5.1 CPS Application 63
5.2 Privacy Protection of Singly and Multiply Imputed Synthetic Data 74

6 Final Remarks **77**

Bibliography **81**

A Appendices **85**
A.1 On the distribution of the statistics based on the original data . . . 85
A.2 Important Identities 86
A.3 Mathematica® source codes for the empirical distributions of $T_M^{\dagger}$, $T_M^{\bullet}$, $T_{comb}^{\bullet}$, T_M^{*} and T_{comb}^{*} 88

List of Figures

2.1 Smothed Empirical distributions and cut-off points ($\gamma = 0.05$) of $T_1^\bullet$, $T_2^\bullet$, $T_3^\bullet$ and $T_4^\bullet$ for $\rho = 0.2, 0.4, 0.6, 0.8$. 16

3.1 Smoothed empirical distributions and cut-off points ($\gamma = 0.05$) of T_1^*, T_2^*, T_3^* and T_4^* for $\rho = 0.2, 0.4, 0.6, 0.8$. 37

5.1 Histograms (with same vertical scale) of the empirical distributions of $T_M^\dagger$ for $M = 2$ and 5 (for $m = 3$, $p = 24$, $n = 141$, $\alpha = 8$ and 10^4 simulation sizes). 68

5.2 Histograms (with same vertical scale for each M) of the empirical distributions of both $T_M^\bullet$ and $T_{comb}^\bullet$ for $M = 1, 2$ and 5 (for $m = 3$, $p = 24$, $n = 141$, $\alpha = 8$ and 10^4 simulation sizes). 68

5.3 Histograms (with same vertical scale for each M) of the empirical distributions of both T_M^* and T_{comb}^* for $M = 1, 2$ and 5 (for $m = 3$, $p = 24$, $n = 141$ and 10^4 simulation sizes). 68

5.4 Box-plots of p-values obtained, when testing the joint significance of I(R=2), I(R=4) and I(S=2), from 100 draws of synthetic datasets using PPS, FPPS and Plug-in Sampling method as also when using Reiter's adapted combination rule for $M = 1$, $M = 2$ and $M = 5$ 70

5.5 Box-plots of p-values obtained, when testing the joint significance of A and E, from 100 draws of synthetic datasets using PPS, FPPS and Plug-in Sampling method as also when using Reiter's adapted combination rule, for $M = 1$, $M = 2$ and $M = 5$ 71

List of Tables

2.1 Cut-off points of the 95% confidence set for the regression coefficients matrix **B**. 19

3.1 Cut-off points of the 95% confidence set for the regression coefficients matrix **B**. 41

4.1 Estimated coverage probability for **vec**(**B**), **B** and **AB** under PPS. 59
4.2 Estimated coverage probability for **B** and **AB** under FPPS. 60
4.3 Estimated coverage probability for **vec**(**B**), **B** and **AB** under Plug-in Sampling. 60
4.4 Average values of the *radius* when using FPPS and Plug-in Sampling with the corresponding expected values and the values of $\Upsilon^{\dagger}_{M,min}$ and $\Upsilon^{\dagger}_{M,max}$ defined in (4.9) and (4.8) when using PPS, for the confidence set for **B**. 61
4.5 Average values of the *radius* when using FPPS and Plug-in Sampling with the corresponding expected values and the values of $\Upsilon^{\dagger}_{M,min}$ and $\Upsilon^{\dagger}_{M,max}$ defined in (4.9) and (4.8) when using PPS, for the confidence set for **C**. 61

5.1 Summary of CPS data variables. 64
5.2 Estimates of the regressor coefficients from the synthetic data and from the original data. 66
5.3 Approximated values of the cut-off points computed from the empirical distributions of $T^{\dagger}_{M}$, $T^{\bullet}_{M}$, $T^{\bullet}_{comb}$, T^{*}_{M} and T^{*}_{comb} respectively defined in subsections 2.2.2, 2.3.1, 2.3.2, 3.2.2 and 3.2.3, for $\gamma = 0.05$. 69
5.4 Power for the test to the hypothesis (5.3), with **B**(1) and **B**(2) denoting the first and second procedures developed in Chapters 2 and 3, for the FPPS, PPS (only one procedure is available) and Plug-in methods, with **vec**(**B**) denoting Reiter's adapted procedures. 73
5.5 Power for the test to the hypothesis (5.4), with **C**(1) and **C**(2) denoting the first and second procedures developed in Chapters 2 and 3, for the FPPS, PPS (only one procedure is available) and Plug-in methods, with **vec**(**C**) denoting Reiter's adapted procedures. 73
5.6 Values of $\Gamma_{1,0.01}$, $\Gamma_{2,0.01}$ and a summary of the distribution of $D_{1,0.01}$. 75

5.7 Values of $\Gamma_{3,0.1}$ and a summary of the distribution of D_3. 76

Listings

A.1 Example of source code for the empirical distribution of $T_M^{\dagger}$ defined in (2.27) used in the CPS application. 88
A.2 Example of source code for the empirical distribution of $T_M^{\bullet}$ and $T_{comb}^{\bullet}$ respectively defined in (2.32) and (2.40), used in the CPS application. 89
A.3 Example of source code for the empirical distribution of T_M^{*} and T_{comb}^{*} respectively defined in (3.11) and (3.15), used in the CPS application. 90

Chapter 1

General Introduction

«In intelligence work,[. . .] there are limits to the amount of information one can share. Confidentiality is essential[. . .].»

Gijs de Vries

1.1 Introduction

When releasing microdata to the public, methods of statistical disclosure control (SDC) are used to protect confidential data, that is "data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information" [29], without compromising an adequate and accurate statistical analysis of the data. Several SDC methods have been recently developed in order to protect the data, without changing its fundamental structure.

One may classify these methods into perturbative, which distort the original data, non-perturbative, which suppress or reduce the detail without altering the original data, and, more recently, methods of generation of synthetic microdata, which preserve some statistics or relationships of the original data [5, 9, 12]. Most of the proposed perturbative methods reduces the quality of the data as such their utility may be questionable and researchers in general tend to not trust these. Noise addition, data swapping and rounding are some examples of these methods. When non-perturbative methods are used there is an higher risk of disclosure, specially when applied to microdata on businesses, because its population is usually smaller than the population associated to microdata on individuals and the size of information may be already available in public sites.

Recoding, suppression, top and bottom coding are classified as non-perturbative [5, 9, 12].

With these methods one faces the problem that some released data guarantees respondents confidentiality but researchers may not accept it due to dubious quality or the problem that in order to have high quality data, respondents sensitive information could be put in high risk of disclosure. With the synthetic data approach, which has gained considerable popularity and importance in recent times, these problems can be overcome [21, 37].

Little [24] and Rubin [36], in 1993, first supported the use of synthetic data for SDC, using the framework of multiple imputation [35]. Rubin claimed that synthetic data so created do not correspond to any actual sampling unit, thus preserving the confidentiality of the respondents. Rubin also proposed that one could use fitted models to generate random and independent samples of the original survey data and release these synthetic versions of microdata publicly, called fully synthetic datasets. The quality of this approach is dependent on the model to impute the values, therefore, all the relationship between variables must be included and the joint distribution of these has to be specified, in order to not give biased results when using the synthetic data [5, 6]. Later that year, Little [24] proposed to only replace, with imputed values, the observed values that could contain sensitive information, leaving the rest unchanged, a proposed solution to overcome the problems inherent to the creation of fully synthetic datasets. This approach is called generation of partially synthetic datasets. This will be the context of the present work. In 1997, Kennickell [14] was the first to use multiply-imputed partially synthetic data to protect the confidentiality of respondents in the Survey of Consumer Finances. Only in 2003, inferential methods for fully synthetic data were developed by Raghunathan et al [27], while, at the same time, Reiter [30] presented the first methods for drawing inference for partially synthetic data.

In comparison with the standard SDC methods, multiple imputation techniques presents many advantages dealing with many real data problems that other methods cannot. It preserves the joint distribution of the original data offering a better quality analysis; is applicable to both categorical and continuous variables; released fully synthetic datasets gives a very small disclosure risk; with partially synthetic datasets generation one may only synthesize the records at risk, maintaining intact the records that have no need to be protected; it allows the possibility to impute missing values before generating synthetic datasets having

no need to give up on some records; preserves linear constraints; allows the analyst to decide if valid results will be given from the synthetic data based on the meta-data information. Some drawbacks exist as well. Since it is a perturbation method there is a question on the utility limit of the data and only the statistical properties gathered by the model are preserved [2, 5].

The most common methods to synthesize data are the Posterior Predictive Sampling and Plug-in Sampling. Although most inferential methods for synthetic data are based on multiple imputation, Klein and Sinha [17, 18, 19, 20] in a series of recent papers developed exact parametric inferential methods based on singly imputed synthetic data for several probability models, including the multiple linear regression model where the sole response variable is considered sensitive, thus requiring protection, while the covariates are treated as non-sensitive, having no need of confidentiality protection.

The main goal of this thesis is to extend this scenario to the multivariate linear regression model where there are multiple sensitive responses variables following a multivariate normal distribution with expected values modeled as linear combinations of multiple non-sensitive covariates. Based on the fitted multivariate linear regression model, the sensitive responses are synthesized based on the Posterior Predictive Sampling method, Plug-in Sampling method and on a new proposed sampling method that will be called Fixed-Posterior Predictive Sampling, and exact data analysis procedures are developed for both single and multiple imputation, for all methods. Reiter [30, 31] combining rules for scalar and vector parameters are the most commonly used methodologies in the analysis of released multiply imputed synthetic datasets [1, 4, 8, 13, 20], due to its easy applicability to various statistical models, as such, in this thesis, one will compare the new developed inferential procedures with the adaptations of Reiter's [31] methodology to Posterior Predictive Sampling and Plug-in Sampling multiply imputed synthetic data, under the Multivariate Linear Regression model. The contents of this thesis are as follows:

- In Chapter 2, based on singly imputed synthetic data generated via Posterior Predictive Sampling, an exact inferential procedure is developed for the matrix regression coefficients $\mathbf{B}$. Based on multiply imputed synthetic data generated via Posterior Predictive Sampling, an exact inference procedure is presented, and based on multiply imputed synthetic data generated via Fixed-Posterior Predictive Sampling, two exact inference procedures are developed. The new exact procedure for the Posterior Predictive Sampling

is contrasted with Reiter's asymptotic methodology adaptation for multiple imputation synthetic data [31]. It is also shown that pivot statistics based in the classical test statistics for **B** under Multivariate Linear Regression model are not pivotal for imputed synthetic data generated via Posterior-Predictive Sampling.

- In Chapter 3, an exact inferential procedure for the matrix of regression coefficients **B** is developed based on singly imputed synthetic data generated via Plug-in Sampling. Based on multiply imputed synthetic data, two exact inference procedures are developed and compared with Reiter's asymptotic methodology adapted to multiply imputed synthetic data. It is also shown that pivot statistics based on the classical test statistics for **B** under the Multivariate Linear Regression model are also not pivotal for imputed synthetic data generated via Plug-in Sampling.

- In Chapter 4, it is proposed a measure, the *radius*, that measures the distance between the center and the edge of the confidence sets for the regression coefficients matrix **B**. Simulation results corroborate the accuracy of the theoretically derived results for the singly imputed and multiply imputed synthetic datasets. These are compared with the results from Reiter's adapted procedures. Values for the confidence sets *radius* for all new procedures developed in Chapters 2 and 3 are also compared.

- Chapter 5 presents data analyses under the proposed methods for singly and multiply imputed synthetic data in the context of public use data from the 2000 U.S. Current Population Survey. The results are compared with those obtained from the original data. Using the same public use data, the levels of Privacy Protection for single and multiple imputation released synthetic data for all sampling methods used in Chapters 2 and 3 are compared.

- A general discussion of the main results and conclusions is presented in Chapter 6.

1.2 Generating Synthetic Data

In order to generate synthetic data for the purpose of public release, the techniques used throughout this thesis will be the Posterior Predictive Sampling (PPS), a new adapted method that we will call Fixed-Posterior Predictive Sampling (FPPS) and the Plug-in Sampling.

Brief descriptions of the three techniques are presented in the following subsections where we consider that $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ are the original data which are jointly distributed according to the probability density function (pdf) $f_{\boldsymbol{\theta}}(\mathbf{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar, vector or matrix) parameter.

1.2.1 Posterior Predictive Sampling (PPS)

A prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is assumed and then the posterior distribution of $\boldsymbol{\theta}$ is obtained as $\pi(\boldsymbol{\theta}|Y) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}$, and used to draw M independent estimates $\boldsymbol{\theta}_1^{\bullet}, \ldots, \boldsymbol{\theta}_M^{\bullet}$ of $\boldsymbol{\theta}$. Following, M replacements of $\mathbf{Y}$ are generated, namely, $\mathbf{W}_j = \left(\mathbf{w}_{j1}, \ldots, \mathbf{w}_{jn}\right)$, $j = 1, \ldots, M$, drawn all independently from the corresponding j-th pdf $f_{\boldsymbol{\theta}_j^{\bullet}}$, where $f_{\boldsymbol{\theta}_j^{\bullet}}$ is the pdf of $\mathbf{Y}$ where the original $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}_j^{\bullet}$, for $j = 1, \ldots, M$. These synthetic datasets $\mathbf{W}_j$ $(j = 1, \ldots, M)$ will be the datasets available to the general public.

1.2.2 Fixed-Posterior Predictive Sampling (FPPS)

A prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is assumed and then the posterior distribution of $\boldsymbol{\theta}$ is obtained as $\pi(\boldsymbol{\theta}|Y) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}$, and used to draw just one estimate of θ, $\boldsymbol{\theta}_f^{\bullet}$. Then, one generates M replacements of $\mathbf{Y}$, namely, $\mathbf{W}_j = \left(\mathbf{w}_{j1}, \ldots, \mathbf{w}_{jn}\right)$, $j = 1, \ldots, M$ drawn all independently from the same $f_{\boldsymbol{\theta}_f^{\bullet}}$, where $f_{\boldsymbol{\theta}_f^{\bullet}}$ is the pdf of the original $\mathbf{Y}$ where $\boldsymbol{\theta}_f^{\bullet}$ replaces the original $\boldsymbol{\theta}$. These synthetic datasets $\mathbf{W}_j$ $(j = 1, \ldots, M)$ will be the datasets available to the general public. One may observe that, for $M = 1$, both the Posterior Predictive Sampling and the new Fixed-Posterior Predictive Sampling methods concur.

1.2.3 Plug-in Sampling

We start by taking the value of a point estimator $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ of $\boldsymbol{\theta}$, and plug it into the joint pdf of $\mathbf{Y}$. The resulting pdf, with the unknown $\boldsymbol{\theta}$ replaced by the observed value of the point estimator $\hat{\boldsymbol{\theta}}(\mathbf{Y})$, is denoted by $f_{\hat{\boldsymbol{\theta}}}$. The multiply imputed synthetic datasets, denoted by $\mathbf{W}_j (j = 1, \ldots, M)$, are then generated independently by drawing $\mathbf{W}_j = \left(\mathbf{w}_{j1}, \ldots, \mathbf{w}_{jn}\right)$ from the joint pdf $f_{\hat{\boldsymbol{\theta}}}$ and these synthetic datasets $\mathbf{W}_j$ $(j = 1, \ldots, M)$ will be the datasets available to the general public.

1.3 The Multivariate Linear Regression Model

Since the inferential procedures will be developed for the Multivariate Linear Regression (MLR) model, it will be important to give a general description of

this model in the context of partially-synthetic data analysis and define the test statistics that will be used for the original data.

Consider m *sensitive* variables $y_j, j = 1,\ldots,m$, that should be replaced by their synthetic version because they present a risk to respondents confidentiality, originating the vector $\mathbf{y} = (y_1,\ldots,y_m)'$, and a set of p *non-sensitive* variables $\mathbf{x} = \left(x_1,\ldots,x_p\right)'$, that do not need to be protected.

In terms of the MLR model, $\mathbf{y} = (y_1,\ldots,y_m)'$ will be considered the vector of response variables and $\mathbf{x} = \left(x_1,\ldots,x_p\right)'$ the set of predictor variables or covariates.

We will consider that $\mathbf{y}|_{\mathbf{x}} \sim N_m(\mathbf{B}'\mathbf{x},\Sigma)$, with $\mathbf{B}$ and Σ unknown, and the original data will consist of $\mathcal{Y} = \left\{(y_{1i},\ldots,y_{mi},x_{1i},\ldots,x_{pi}), i = 1,\ldots,n\right\}$, observing that predictor variables are considered fixed. Let us write $\mathbf{Y} = (\mathbf{y}_1,\ldots,\mathbf{y}_n)$ with $\mathbf{y}_i = (y_{1i},\ldots,y_{mi})'$ and $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ with $\mathbf{x}_i = (x_{1i},\ldots,x_{pi})'$, assuming that $rank(\mathbf{X} : p\times n) = p < n$ and $n \geq m+p$. Therefore we have the following regression model

$$\mathbf{Y}_{m\times n} = \mathbf{B}'_{m\times p}\mathbf{X}_{p\times n} + \mathbf{E}_{m\times n} \tag{1.1}$$

where $\mathbf{E}_{m\times n}$ will be distributed as $N_{mn}(\mathbf{0},\mathbf{I_n}\otimes\Sigma)$.

The maximum likelihood estimator (MLE) and uniformly minimum-variance unbiased estimator (UMVUE) of $\mathbf{B}$, distributed as $N_{pm}\left(\mathbf{B},\Sigma\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$

$$\hat{\mathbf{B}} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Y}', \tag{1.2}$$

independent of

$$\hat{\Sigma} = \frac{1}{n}\left(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X}\right)\left(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X}\right)' \tag{1.3}$$

which is the MLE of Σ, with $n\hat{\Sigma} \sim W_m(\Sigma, n-p)$ [3, Chapter 8]. Therefore $\mathbf{S} = \frac{n\hat{\Sigma}}{n-p}$ will be an unbiased estimator (UE) of Σ.

Several tests for $\mathbf{B}$ based on the original data can be found in the literature [3, Chapter 8], but, as it will be shown in the next Chapter, the adaptations of this classical tests to the synthetic data cannot be used, therefore, for purposes of comparison, it is developed a new test procedure for $\mathbf{B}$ and also for $\mathbf{C} = \mathbf{A}\mathbf{B}$, where $\mathbf{A}$ is a $k\times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$. Inference based on the original data will be drawn and will be used to compare with inference drawn from the synthetic data. This test procedure will be based on

$$T_O = \frac{\left|\left(\hat{\mathbf{B}}-\mathbf{B}\right)'(\mathbf{X}\mathbf{X}')\left(\hat{\mathbf{B}}-\mathbf{B}\right)\right|}{|(n-p)\mathbf{S}|} \sim \prod_{i=1}^{m}\frac{p-i+1}{n-p-i+1}F_i \tag{1.4}$$

and

$$T_{O,\mathbf{C}} = \frac{\left|\left(\mathbf{A}\hat{\mathbf{B}}-\mathbf{C}\right)'\left(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}'\right)^{-1}\left(\mathbf{A}\hat{\mathbf{B}}-\mathbf{C}\right)\right|}{|(n-p)\mathbf{S}|} \sim \prod_{i=1}^{m}\frac{k-i+1}{n-p-i+1}F_{k,i} \tag{1.5}$$

where F_i $(i = 1,\ldots,m)$ and $F_{k,i}$ $(i = 1,\ldots,m)$ are two sets of independent random variables following respectively $F_{p-i+1,n-p-i+1}$ and $F_{k-i+1,n-p-i+1}$ distributions.

The derivation of the distributions of T_O and $T_{O,\mathbf{C}}$ can be seen in Appendix A.

1.4 An important Lemma

Concluding this Chapter, it will be important for the derivation of all the results developed in Chapters 2 and 3 to make an observation regarding the existence of *sufficient statistics*.

Suppose the original data are $\mathbf{Y} \sim f_{\hat{\theta}(\mathbf{Y})}$, and the synthetic data $\mathscr{V} = (\mathbf{V}_1,\ldots,\mathbf{V}_M)$ are generated such that $\mathbf{V}_1|\mathbf{Y},\ldots,\mathbf{V}_M|\mathbf{Y}$ are independent and identically distributed (i.i.d.) from $f_{\hat{\theta}(\mathbf{Y})}$. Suppose that $\mathbf{T}(\mathbf{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$ based on the original data. Then the pdf of the synthetic data $\mathscr{V} = (\mathbf{V}_1,\ldots,\mathbf{V}_M)$ is

$$\begin{aligned}\int\left\{\prod_{i=1}^{M} f_{\hat{\theta}(\mathbf{Y})}(\mathbf{V}_i)\right\} f_{\boldsymbol{\theta}}(\mathbf{Y})d\mathbf{Y} &= \int\left\{\prod_{i=1}^{M} g_{\hat{\theta}(\mathbf{Y})}(\mathbf{T}(\mathbf{V}_i))\,h(\mathbf{V}_i)\right\} f_{\boldsymbol{\theta}}(\mathbf{Y})d\mathbf{Y}\\ &= \left\{\prod_{i=1}^{M} h(\mathbf{V}_i)\right\}\int\left\{\prod_{i=1}^{M} g_{\hat{\theta}(\mathbf{Y})}(\mathbf{T}(\mathbf{V}_i))\right\} f_{\boldsymbol{\theta}}(\mathbf{Y})d\mathbf{Y},\end{aligned}$$

where $g_{\hat{\theta}(\mathbf{Y})}(\mathbf{T}(\mathbf{V}_i))$ and $h(\mathbf{V}_i)$ are non-negative functions, with g depending on $\mathbf{V}_i$ only through the statistic $T(\mathbf{V}_i)$ and h only depending on $\mathbf{V}_i$, which implies the following result.

Lemma 1.4.1. *Suppose that when the original data* $\mathbf{Y}$ *are observed,* $\mathbf{T}(\mathbf{Y})$ *is a sufficient statistic for* $\boldsymbol{\theta}$*. Then when the synthetic data* $\mathscr{V} = (\mathbf{V}_1,\ldots,\mathbf{V}_M)$ *are observed,* $(\mathbf{T}(\mathbf{V}_1),\ldots,\mathbf{T}(\mathbf{V}_M))$ *is jointly sufficient for* $\boldsymbol{\theta}$*. Furthermore, if* $M = 1$*, the sufficient statistic is simply* $\mathbf{T}(\mathbf{V}_1)$*, and if* $M > 1$*, then* $\sum_{i=1}^{M}\mathbf{T}(\mathbf{V}_i)$ *is sufficient if* $f_{\boldsymbol{\theta}}(\mathbf{Y}) = h(\mathbf{Y})\psi(\boldsymbol{\theta})\exp\{\gamma(\boldsymbol{\theta})'\mathbf{T}(\mathbf{Y})\}$*, i.e., if* $f_{\boldsymbol{\theta}}(\mathbf{Y})$ *belongs to the exponential family.*

Chapter 2

Inference for Multivariate Regression Model based on synthetic data generated via PPS and FPPS

The main objective of this chapter is to present the likelihood-based approach developed for the analysis of partially-synthetic data generated via Posterior Predictive Sampling (PPS) and via a new proposed sampling method called Fixed-Posterior Predictive Sampling (FPPS) which is originated from an adaptation of the PPS method.

When one uses the PPS method to generate the multiply imputed synthetic data one has to deal with the problem of obtaining the distribution of a sum of Wishart distributions with different parameters. The use of the FPPS method is suppose to overcome this problem.

Most of the content of Sections 2.1 and 2.3 is taken from [25].

2.1 Posterior Predictive Sampling (PPS)

Imputing the sensitive variables in the original data is somewhat similar as treating them as missing data. Rubin [35] was the first to propose the use of multiple imputation in order to handle the problem of missing data and Little [24] and Rubin [36], in 1993, first supported the use of synthetic data for SDC, using the framework of multiple imputation. Since then, many authors [5, 10, 28, 32, 34]

developed asymptotic based inferential procedures to analyze multiply imputed synthetic data using the PPS method to generate these synthetic data. All these inferential procedures are only suitable for the analysis of multiply imputed synthetic datasets leaving a gap in the state of art by leaving out the single imputation case. Since, in some cases [11, 15, 16] it is mandatory to only release one single synthetic data due to the high risk of disclosure, it is important to make available an inference procedure for this case.

2.1.1 Single Imputation: Posterior Predictive Sampling Method

The PPS method was generally described in subsection 1.2.1. In this subsection, we will start by describing specifically the PPS method under the MLR model case.

Let us consider the MLR model (1.1) with $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{B}$, Σ, $\hat{\mathbf{B}}$ and $\mathbf{S}$ defined in that same context. Consider the joint prior distribution

$$\pi(\mathbf{B},\Sigma) \propto |\Sigma|^{-\alpha/2}$$

and from it let us develop the posterior distributions for the unknown parameters Σ and $\mathbf{B}$ in model (1.1). Let us observe that $\mathbf{Y}|_{\mathbf{B},\Sigma} \sim N_{mn}(\mathbf{B}'\mathbf{X}, \mathbf{I_n} \otimes \Sigma)$. Hence the likelihood function for $\mathbf{Y}$ will be

$$l(\mathbf{B},\Sigma|_{\mathcal{Y}}) \propto |\Sigma|^{-n/2} e^{-\frac{1}{2}tr\left\{\Sigma^{-1}(\mathbf{Y}-\mathbf{B}'\mathbf{X})(\mathbf{Y}-\mathbf{B}'\mathbf{X})'\right\}},$$

and the joint posterior distribution for $(\mathbf{B},\Sigma)$ can be obtained from the product of the previous prior and likelihood functions

$$\pi(\mathbf{B},\Sigma|_{\mathcal{Y}}) \propto |\Sigma|^{-\frac{n+\alpha}{2}} e^{-\frac{1}{2}tr\left\{\Sigma^{-1}(\mathbf{Y}-\mathbf{B}'\mathbf{X})(\mathbf{Y}-\mathbf{B}'\mathbf{X})'\right\}}.$$

Regarding the exponent part of this joint posterior distribution, we may decompose it as

$$\begin{aligned}
tr\left\{\Sigma^{-1}(\mathbf{Y}-\mathbf{B}'\mathbf{X})(\mathbf{Y}-\mathbf{B}'\mathbf{X})'\right\} &= tr\left\{\Sigma^{-1}(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X}+\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X}+\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})'\right\} \\
&= tr\left\{\Sigma^{-1}\left[(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})'\right]\right\} \\
+ tr\left\{\Sigma^{-1}\left[(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})(\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})' + (\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})' + (\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})(\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})'\right]\right\} \\
&= tr\left\{\Sigma^{-1}\left[(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})'\right] + (\mathbf{B}-\hat{\mathbf{B}})'(\mathbf{X}\mathbf{X}')(\mathbf{B}-\hat{\mathbf{B}})\right\} \\
&\quad + 2tr\left\{\Sigma^{-1}\left[(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})(\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})'\right]\right\}.
\end{aligned}$$

Note that $2tr\left\{\Sigma^{-1}\left[(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})(\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X})'\right]\right\}$ is null, since $\hat{\mathbf{B}}' = \left[(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Y}'\right]' = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}$ and so

$$\begin{aligned}\left(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X}\right)\left(\hat{\mathbf{B}}'\mathbf{X}-\mathbf{B}'\mathbf{X}\right)' &= \mathbf{Y}\mathbf{X}'\hat{\mathbf{B}}-\mathbf{Y}\mathbf{X}'\mathbf{B}+\hat{\mathbf{B}}\mathbf{X}\mathbf{X}'\hat{\mathbf{B}}+\hat{\mathbf{B}}\mathbf{X}\mathbf{X}'\mathbf{B}\\ &= \mathbf{Y}\mathbf{X}'\hat{\mathbf{B}}-\mathbf{Y}\mathbf{X}'\mathbf{B}+\mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}'\hat{\mathbf{B}}+\mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}'\mathbf{B}\\ &= \mathbf{Y}\mathbf{X}'\hat{\mathbf{B}}-\mathbf{Y}\mathbf{X}'\mathbf{B}-\mathbf{Y}\mathbf{X}'\hat{\mathbf{B}}+\mathbf{Y}\mathbf{X}'\mathbf{B}=0,\end{aligned}$$

therefore obtaining the posterior distribution proportional to

$$|\Sigma|^{-\frac{n+\alpha-p}{2}}e^{-\frac{n-p}{2}tr\{\Sigma^{-1}\mathbf{S}\}}\,|\Sigma|^{-\frac{p}{2}}e^{-\frac{1}{2}tr\left\{\Sigma^{-1}(\mathbf{B}-\hat{\mathbf{B}})'(\mathbf{X}\mathbf{X}')(\mathbf{B}-\hat{\mathbf{B}})\right\}},$$

by recalling the definition of $\mathbf{S}$ as $\mathbf{S} = \frac{1}{n-p}(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})(\mathbf{Y}-\hat{\mathbf{B}}'\mathbf{X})'$.

Using Corollary 2.4.6.2. in [22], the posterior distribution for Σ is given by

$$\Sigma|_{\mathbf{S}} \sim W_m^{-1}\Big((n-p)\mathbf{S}, n+\alpha-p\Big), \tag{2.1}$$

where $W_m^{-1}(\boldsymbol{\Psi}, \nu)$ denotes the Inverse Wishart distribution with $\boldsymbol{\Psi} : m\times m$ a positive definite matrix and ν degrees of freedom, and the posterior distribution for $\mathbf{B}$ is given by

$$\mathbf{B}|_{\hat{\mathbf{B}},\Sigma} \sim N_{pm}\left(\hat{\mathbf{B}}, \Sigma\otimes(\mathbf{X}\mathbf{X}')^{-1}\right) \tag{2.2}$$

assuming $n+\alpha > p+m+1$.

Now, it is possible to generate the synthetic dataset under the MLR model. Start by drawing $\tilde{\Sigma}$ from (2.1) and $\tilde{\mathbf{B}}$ from (2.2), upon replacing Σ by $\tilde{\Sigma}$ in this latter expression, and then generate one single synthetic dataset, denoted as $\mathbf{W} = (\mathbf{w}_1,\ldots,\mathbf{w}_n)$ where $\mathbf{w}_i = (w_{1i},\ldots,w_{mi})'$, will be independently distributed as

$$\mathbf{w}_i|_{\tilde{\mathbf{B}},\tilde{\Sigma}} \sim N_m\left(\tilde{\mathbf{B}}'\mathbf{x}_\mathbf{i}, \tilde{\Sigma}\right),\quad i=1,\ldots,n. \tag{2.3}$$

Let us define

$$\mathbf{B}^\bullet = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}' \tag{2.4}$$

and

$$\mathbf{S}^\bullet = \frac{1}{n-p}\left(\mathbf{W}-\mathbf{B}^{\bullet'}\mathbf{X}\right)\left(\mathbf{W}-\mathbf{B}^{\bullet'}\mathbf{X}\right)', \tag{2.5}$$

that will be the estimators of $\mathbf{B}$ and Σ, respectively. By Lemma 1.4.1 these estimators are jointly sufficient. Note that conditionally on $\left(\tilde{\mathbf{B}}, \tilde{\Sigma}\right)$, $\mathbf{B}^\bullet$ is independent of $\mathbf{S}^\bullet$, as in the original data $\hat{\mathbf{B}}$ and $\mathbf{S}$ were also independent.

With the access to the partially synthetic data one may derivate the joint pdf involving the estimators of $\mathbf{B}$ and Σ obtained from this data.

Theorem 2.1.1. *The joint pdf of* $\mathbf{B}^\bullet$, $\mathbf{S}^\bullet$ *and* $\tilde{\Sigma}^{-1}$, *for* $\mathbf{B}^\bullet$ *and* $\mathbf{S}^\bullet$ *respectively defined in (2.4) and (2.5), is proportional to*

$$e^{-\frac{1}{2}tr\left\{(2\tilde{\Sigma}+\Sigma)^{-1}(\mathbf{B}^\bullet-\mathbf{B})'\mathbf{X}\mathbf{X}'(\mathbf{B}^\bullet-\mathbf{B})+(n-p)\tilde{\Sigma}^{-1}\mathbf{S}^\bullet\right\}}$$

$$\times\frac{|\mathbf{S}^\bullet|^{\frac{(n-p)-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{(n-p)+n+\alpha}{2}-m-1}}|\Sigma|^{-\frac{n}{2}}\left|\frac{1}{2}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right|^{-p/2}\left|\tilde{\Sigma}^{-1}+\Sigma^{-1}\right|^{-\frac{2n+\alpha-2p-m-1}{2}}. \quad (2.6)$$

Proof. Given $\tilde{\Sigma}$ and $\tilde{\mathbf{B}}$ respectively from (2.1) and (2.2), we have that

$$\mathbf{W}'|_{\tilde{\mathbf{B}},\tilde{\Sigma}}\sim N_{nm}\left(\mathbf{X}'\tilde{\mathbf{B}},\tilde{\Sigma}\otimes\mathbf{I}_n\right)\implies\mathbf{B}^\bullet|_{\tilde{\mathbf{B}},\tilde{\Sigma}}=(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}'|_{\tilde{\mathbf{B}},\tilde{\Sigma}}\sim N_{pm}\left(\tilde{\mathbf{B}},\tilde{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$(n-p)\mathbf{S}^\bullet|_{\tilde{\Sigma}}\sim W_m\left(\tilde{\Sigma},n-p\right).$$

Since $\mathbf{B}^\bullet$ and $\mathbf{S}^\bullet$ are independent, the conditional joint pdf of $(\mathbf{B}^\bullet,\mathbf{S}^\bullet)$, given $\tilde{\mathbf{B}}$ and $\tilde{\Sigma}$, will be proportional to

$$e^{-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}\left[(\mathbf{B}^\bullet-\tilde{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\mathbf{B}^\bullet-\tilde{\mathbf{B}})+(n-p)\mathbf{S}^\bullet\right]\right\}}\frac{|\mathbf{S}^\bullet|^{\frac{(n-p)-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{(n-p)+p}{2}}}, \quad (2.7)$$

while, due to the independence of $\tilde{\mathbf{B}}$ and $\tilde{\Sigma}^{-1}$ drawn respectively from (2.2) and (2.1), the joint pdf of $(\tilde{\mathbf{B}},\tilde{\Sigma}^{-1})$, given $\mathbf{S}$, is proportional to

$$|\tilde{\Sigma}|^{-p/2}e^{-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}\left[(\tilde{\mathbf{B}}-\hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}}-\hat{\mathbf{B}})+(n-p)\mathbf{S}\right]\right\}}\frac{|\mathbf{S}|^{\frac{n+\alpha-p-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{n+\alpha-p}{2}-m-1}}. \quad (2.8)$$

Given the independence of $\hat{\mathbf{B}}$ and $\mathbf{S}$, defined in (1.2) and (1.3), the joint pdf of $(\hat{\mathbf{B}},\mathbf{S})$ is proportional to

$$e^{-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(\hat{\mathbf{B}}-\mathbf{B})'\mathbf{X}\mathbf{X}'(\hat{\mathbf{B}}-\mathbf{B})+(n-p)\mathbf{S}\right]\right\}}\frac{|\mathbf{S}|^{\frac{n-p-m-1}{2}}}{|\Sigma|^{\frac{n}{2}}}. \quad (2.9)$$

Thus, by multiplying the three pdf's (2.7), (2.8) and (2.9), the joint pdf of $\left(\mathbf{B}^\bullet,\mathbf{S}^\bullet,\tilde{\mathbf{B}},\tilde{\Sigma}^{-1},\hat{\mathbf{B}},\mathbf{S}\right)$ is obtained.

Since

$$tr\left\{\left(\mathbf{B}^\bullet-\tilde{\mathbf{B}}\right)'\mathbf{X}\mathbf{X}'\left(\mathbf{B}^\bullet-\tilde{\mathbf{B}}\right)\right\}=tr\left\{\left(\tilde{\mathbf{B}}-\mathbf{B}^\bullet\right)'\mathbf{X}\mathbf{X}'\left(\tilde{\mathbf{B}}-\mathbf{B}^\bullet\right)\right\},$$

and since from (A.1) in Result A.2.1,

$$\left(\tilde{\mathbf{B}}-\mathbf{B}^\bullet\right)'\mathbf{X}\mathbf{X}'\left(\tilde{\mathbf{B}}-\mathbf{B}^\bullet\right)+\left(\tilde{\mathbf{B}}-\hat{\mathbf{B}}\right)'\mathbf{X}\mathbf{X}'\left(\tilde{\mathbf{B}}-\hat{\mathbf{B}}\right)=$$

$$=2\left[\tilde{\mathbf{B}}-\frac{1}{2}\left(\mathbf{B}^\bullet+\hat{\mathbf{B}}\right)\right]'\mathbf{X}\mathbf{X}'\left[\tilde{\mathbf{B}}-\frac{1}{2}\left(\mathbf{B}^\bullet+\hat{\mathbf{B}}\right)\right]+\frac{1}{2}\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)'\mathbf{X}\mathbf{X}'\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right),$$

the joint pdf of $(\mathbf{B}^\bullet, \mathbf{S}^\bullet, \tilde{\Sigma}^{-1}, \hat{\mathbf{B}}, \mathbf{S})$ is obtained by integrating out $\tilde{\mathbf{B}}$, being proportional to

$$e^{-\frac{1}{2}tr\left\{\tilde{\Sigma}^{-1}\left[\frac{1}{2}(\mathbf{B}^\bullet-\hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\mathbf{B}^\bullet-\hat{\mathbf{B}})+(n-p)(\mathbf{S}^\bullet+\mathbf{S})\right]+\Sigma^{-1}\left[(\hat{\mathbf{B}}-\mathbf{B})'\mathbf{X}\mathbf{X}'(\hat{\mathbf{B}}-\mathbf{B})+(n-p)\mathbf{S}\right]\right\}}$$
$$\times\frac{|\mathbf{S}^\bullet|^{\frac{(n-p)-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{(n-p)+n-\alpha}{2}-m-1}}\frac{|\mathbf{S}|^{n+\frac{\alpha}{2}-p-m-1}}{|\Sigma|^{\frac{n}{2}}}. \quad (2.10)$$

Taking into account that

$$tr\left\{\frac{1}{2}\tilde{\Sigma}^{-1}\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)'\mathbf{X}\mathbf{X}'\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)+\Sigma^{-1}\left(\hat{\mathbf{B}}-\mathbf{B}\right)'\mathbf{X}\mathbf{X}'\left(\hat{\mathbf{B}}-\mathbf{B}\right)\right\}=$$
$$=tr\left\{\mathbf{X}\mathbf{X}'\left[\frac{1}{2}\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)\tilde{\Sigma}^{-1}\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)'+\left(\hat{\mathbf{B}}-\mathbf{B}\right)\Sigma^{-1}\left(\hat{\mathbf{B}}-\mathbf{B}\right)'\right]\right\}$$

and from (A.2) in Result A.2.2,

$$\frac{1}{2}\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)\tilde{\Sigma}^{-1}\left(\mathbf{B}^\bullet-\hat{\mathbf{B}}\right)'+\left(\hat{\mathbf{B}}-\mathbf{B}\right)\Sigma^{-1}\left(\hat{\mathbf{B}}-\mathbf{B}\right)'=$$
$$=\left[\hat{\mathbf{B}}-\left(\frac{1}{2}\mathbf{B}^\bullet\tilde{\Sigma}^{-1}+\mathbf{B}\Sigma^{-1}\right)\left(\frac{1}{2}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right)^{-1}\right]\left(\frac{1}{2}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right)$$
$$\left[\hat{\mathbf{B}}-\left(\frac{1}{2}\mathbf{B}^\bullet\tilde{\Sigma}^{-1}+\mathbf{B}\Sigma^{-1}\right)\left(\frac{1}{2}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right)^{-1}\right]'$$
$$+(\mathbf{B}^\bullet-\mathbf{B})\left(2\tilde{\Sigma}+\Sigma\right)^{-1}(\mathbf{B}^\bullet-\mathbf{B})',$$

integrating out $\hat{\mathbf{B}}$, the joint pdf of $\left(\mathbf{B}^\bullet, \mathbf{S}^\bullet, \tilde{\Sigma}^{-1}, \mathbf{S}\right)$ will be proportional to

$$e^{-\frac{1}{2}tr\left\{(2\tilde{\Sigma}+\Sigma)^{-1}(\mathbf{B}^\bullet-\mathbf{B})'\mathbf{X}\mathbf{X}'(\mathbf{B}^\bullet-\mathbf{B})+(n-p)\tilde{\Sigma}^{-1}(\mathbf{S}^\bullet+\mathbf{S})+(n-p)\Sigma^{-1}\mathbf{S}\right\}}$$
$$\times\frac{|\mathbf{S}^\bullet|^{\frac{(n-p)-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{(n-p)+n-\alpha}{2}-m-1}}\frac{|\mathbf{S}|^{n+\frac{\alpha}{2}-p-m-1}}{|\Sigma|^{\frac{n}{2}}}\left|\frac{1}{2}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right|^{-p/2}.$$

Consequently, integrating out $\mathbf{S}$ one ends up with the joint pdf of $\left(\mathbf{B}^\bullet, \mathbf{S}^\bullet, \tilde{\Sigma}^{-1}\right)$ proportional to (2.6), concluding the proof.

□

In expression (2.6), one may see that $\mathbf{S}^\bullet$ and $\mathbf{B}^\bullet$, conditionally on $\tilde{\Sigma}^{-1}$, are separable, with

$$\mathbf{B}^\bullet|_{\tilde{\Sigma}} \sim N_{pm}\left(\mathbf{B}, (2\tilde{\Sigma}+\Sigma)\otimes(\mathbf{X}\mathbf{X}')^{-1}\right) \quad (2.11)$$

and

$$\mathbf{S}^\bullet|_{\tilde{\Sigma}} \sim W_m\left(\frac{1}{n-p}\tilde{\Sigma}, n-p\right), \quad (2.12)$$

independent of $\mathbf{B}^\bullet$.

Immediately from pdf (2.6), we may conclude that the MLE of $\mathbf{B}$ based on the synthetic data will be $\mathbf{B}^\bullet$, with

$$E(\mathbf{B}^\bullet) = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}E(\mathbf{W}') = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}'E\left(\tilde{\mathbf{B}}\right) = E\left(\hat{\mathbf{B}}\right) = \mathbf{B},$$

therefore making $\mathbf{B}^\bullet$ an UE. Its variance may be derived from

$$Var(\mathbf{B}^\bullet) = Var\left[E(\mathbf{B}^\bullet|\tilde{\mathbf{B}},\tilde{\Sigma})\right] + E\left[Var(\mathbf{B}^\bullet|\tilde{\mathbf{B}},\tilde{\Sigma})\right],$$

where, for $n+\alpha > p+2m+2$,

$$\begin{aligned} Var\left[E(\mathbf{B}^\bullet|\tilde{\mathbf{B}},\tilde{\Sigma})\right] &= Var\left[\tilde{\mathbf{B}}\right] = Var\left[E(\tilde{\mathbf{B}}|\hat{\mathbf{B}},\tilde{\Sigma})\right] + E\left[Var(\tilde{\mathbf{B}}|\hat{\mathbf{B}},\tilde{\Sigma})\right] \\ &= Var(\hat{\mathbf{B}}) + E\left[\tilde{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\right] = \Sigma\otimes(\mathbf{X}\mathbf{X}')^{-\mathbf{1}} + \frac{n-p}{n+\alpha-p-2m-2}\Sigma\otimes(\mathbf{X}\mathbf{X}')^{-\mathbf{1}} \end{aligned}$$

and

$$E\left[Var(\mathbf{B}^\bullet|\tilde{\mathbf{B}},\tilde{\Sigma})\right] = E\left[\tilde{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\right] = \frac{n-p}{n+\alpha-p-2m-2}\Sigma\otimes(\mathbf{X}\mathbf{X}')^{-\mathbf{1}},$$

thus yielding

$$Var(\mathbf{B}^\bullet) = \frac{2(n-p-m-1)+n-p+\alpha}{n+\alpha-p-2m-2}\Sigma\otimes(\mathbf{X}\mathbf{X}')^{-\mathbf{1}},$$

under the condition that $n+\alpha > p+2m+2$.

We may also observe that $\mathbf{S}^\bullet$ is an UE of Σ, if $\alpha = 2m-2$, since

$$E(\mathbf{S}^\bullet) = E(\tilde{\Sigma}) = E\left(\frac{n-p}{n+\alpha-p-2m-2}\mathbf{S}\right) = \frac{n-p}{n+\alpha-p-2m-2}\Sigma.$$

This way, having access only to one released synthetic dataset it is simple to compute estimates for the unknown parameters from the usual estimators.

At this point one could suggest, in order to perform tests for $\mathbf{B}$, the following adaptations of the classical test criteria for the multivariate regression model (see [3, Secs 8.3 and 8.6] for the classical criteria):

(a) $T_1^\bullet = |\mathbf{S}^\bullet|\left|\mathbf{S}^\bullet + (\mathbf{B}^\bullet-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet-\mathbf{B})\right|^{-1}$ (Wilks' Lambda Criterion);

(b) $T_2^\bullet = tr\left\{(\mathbf{B}^\bullet-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet-\mathbf{B})(\mathbf{S}^\bullet)^{-1}\right\}$ (Pillai's Trace Criterion);

(c) $T_3^\bullet = tr\left\{(\mathbf{B}^\bullet-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet-\mathbf{B})\left[(\mathbf{B}^\bullet-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet-\mathbf{B})+\mathbf{S}^\bullet\right]^{-1}\right\}$ (Hotelling - Lawley Trace Criterion);

(d) $T_4^\bullet = \lambda_1$ where λ_1 denotes the largest eigenvalue of $(\mathbf{B}^\bullet - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet - \mathbf{B})(\mathbf{S}^\bullet)^{-1}$ (Roy's Largest Root Criterion).

However, these statistics are non-pivotal, that is, their distributions will be function of Σ. When using the term 'statistic' we are assuming $\mathbf{B}$ known. In fact, Lehmann in [23, Sec. 8.4] said that the distributions of these classical test statistics will depend only on the nonzero roots of the equation $|\mathbf{E} - \lambda\Sigma| = 0$ thus implying that their distribution will depend on Σ. Therefore, it is expected that the adapted statistics in (a)-(d) for the analysis of synthetic data will also have distributions that will be function of Σ. Considering this fact, let us begin by rewriting all four classical statistics $T_1^\bullet$, $T_2^\bullet$, $T_3^\bullet$ and $T_4^\bullet$, in order to make them assume the same type of form and then prove why all of them are non-pivotal.

Let us consider

$$\mathbf{H} = \left(2\tilde{\Sigma} + \Sigma\right)^{-\frac{1}{2}} (\mathbf{B}^\bullet - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet - \mathbf{B})\left(2\tilde{\Sigma} + \Sigma\right)^{-\frac{1}{2}} \tag{2.13}$$

and

$$\mathbf{G} = (n-p)\tilde{\Sigma}^{-\frac{1}{2}}\mathbf{S}^\bullet\tilde{\Sigma}^{-\frac{1}{2}}. \tag{2.14}$$

By Theorem 2.4.1 in [22], for $p \geq m$,

$$(\mathbf{B}^\bullet - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^\bullet - \mathbf{B})|_{\tilde{\Sigma}^{-1}} \sim W_m\left(2\tilde{\Sigma} + \Sigma, p\right).$$

As such from Theorem 2.4.2 in [22] and subsection 7.3.3 in [3] we have for $\mathbf{H}$ and $\mathbf{G}$ in (2.13) and (2.14)

$$\mathbf{H}|_{\tilde{\Sigma}^{-1}} \sim W_m(\mathbf{I}_m, p) \tag{2.15}$$

and

$$\mathbf{G}|_{\tilde{\Sigma}^{-1}} \sim W_m(\mathbf{I}_m, n-p), \tag{2.16}$$

whose distributions are not function of Σ.

The statistic $T_1^\bullet$ may then be rewritten as

$$T_1^\bullet = \frac{|\mathbf{G}|}{\left|\mathbf{G} + (n-p)\tilde{\Sigma}^{-1/2}\left(2\tilde{\Sigma} + \Sigma\right)^{1/2}\mathbf{H}\left(2\tilde{\Sigma} + \Sigma\right)^{1/2}\tilde{\Sigma}^{-1/2}\right|}.$$

while $T_2^\bullet$ and $T_3^\bullet$ may be rewritten as

$$T_2^\bullet = (n-p)tr\left\{\mathbf{H}\left(2\tilde{\Sigma} + \Sigma\right)^{1/2}\tilde{\Sigma}^{-1/2}\mathbf{G}^{-1}\tilde{\Sigma}^{-1/2}\left(2\tilde{\Sigma} + \Sigma\right)^{1/2}\right\}$$

and

$$T_3^\bullet = tr\left\{\mathbf{H}\left[\mathbf{H} + (n-p)\left(2\tilde{\Sigma} + \Sigma\right)^{-1/2}\tilde{\Sigma}^{1/2}\mathbf{G}\tilde{\Sigma}^{1/2}\left(2\tilde{\Sigma} + \Sigma\right)^{-1/2}\right]^{-1}\right\}.$$

Concerning $T_4^\bullet$, we have that $T_4^\bullet = \lambda_1$ where λ_1 denotes the largest eigenvalue of

$$(n-p)\mathbf{H}\left(2\tilde{\Sigma}+\Sigma\right)^{1/2}\tilde{\Sigma}^{-1/2}\mathbf{G}^{-1}\tilde{\Sigma}^{-1/2}\left(2\tilde{\Sigma}+\Sigma\right)^{1/2}.$$

Now, let us observe that a term in the denominator of the expression $T_1^\bullet$ is

$$\tilde{\Sigma}^{-1/2}\left(2\tilde{\Sigma}+\Sigma\right)^{1/2}\mathbf{H}\left(2\tilde{\Sigma}+\Sigma\right)^{1/2}\tilde{\Sigma}^{-1/2} \sim W_m\left(\left(2\mathbf{I}_m+\tilde{\Sigma}^{-1/2}\Sigma\tilde{\Sigma}^{-1/2}\right),p\right),$$

while in the expressions for the other statistics there are terms similar to

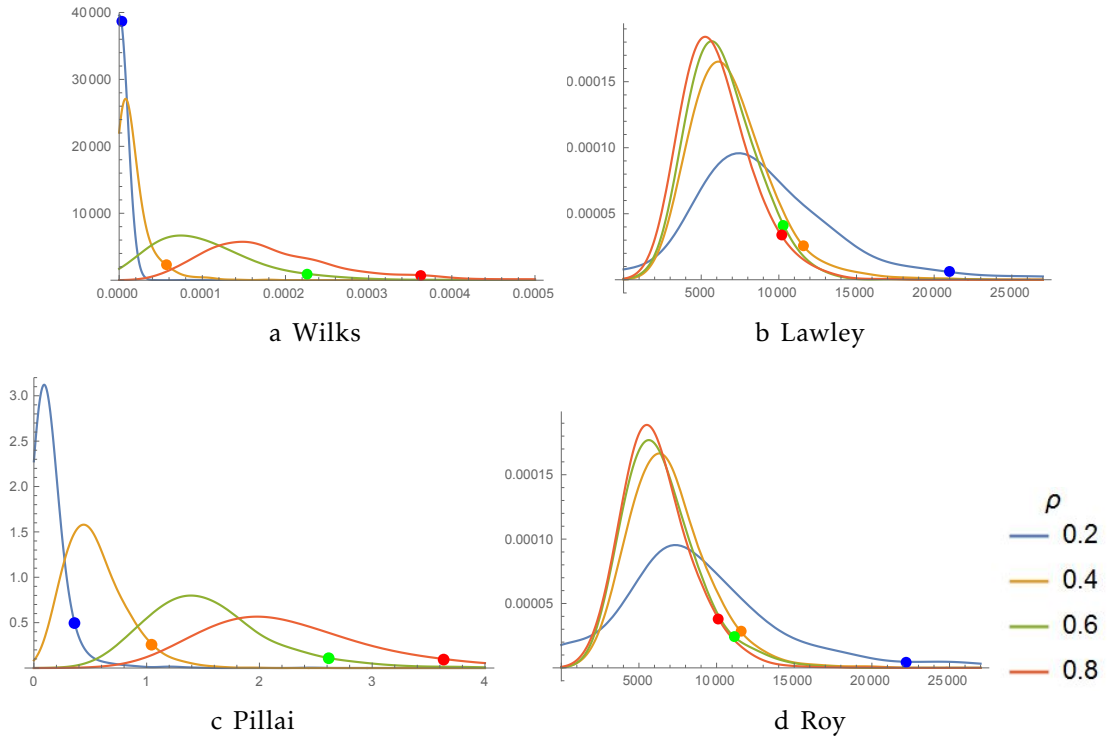
$$(n-p)\left(2\tilde{\Sigma}+\Sigma\right)^{-1/2}\tilde{\Sigma}^{1/2}\mathbf{G}\tilde{\Sigma}^{1/2}\left(2\tilde{\Sigma}+\Sigma\right)^{-1/2} \sim W_m\left(\left(2\tilde{\Sigma}+\Sigma\right)^{-1/2}\tilde{\Sigma}\left(2\tilde{\Sigma}+\Sigma\right)^{-1/2},n-p\right)$$

which have distributions that will depend on Σ. Since the remaining terms in $T_1^\bullet$ are $\mathbf{G}$ and in the other three statistics are $\mathbf{H}$, the distributions of these statistics will themselves be function of Σ, therefore making these statistics non-pivotal.

In order to illustrate how these statistics are dependent on Σ, one may analyze in Figure 2.1 the empirical distributions of $T_1^\bullet$, $T_2^\bullet$, $T_3^\bullet$ and $T_4^\bullet$ for $m = 3$, $p = 24$, $\alpha = 4$, $n = 100$ and

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

with $\rho = 0.2, 0.4, 0.6, 0.8$ for a simulation size of 10^4.



a Wilks b Lawley

c Pillai d Roy

Figure 2.1: Smothed Empirical distributions and cut-off points ($\gamma = 0.05$) of $T_1^\bullet$, $T_2^\bullet$, $T_3^\bullet$ and $T_4^\bullet$ for $\rho = 0.2, 0.4, 0.6, 0.8$.

As seen, these adaptations of the classical criteria cannot be used to make inference about the regression coefficient matrix since they will always depend on the original data, through $\tilde{\Sigma}$. Therefore there is a need to propose a different quantity which will be pivotal, not dependent on the original data.

Theorem 2.1.2 makes available a pivotal statistic that is not dependent on the original data and which may be used to draw inference for $\mathbf{B}$ from the synthetic version of the original data, which is the accessible data to general public.

Theorem 2.1.2. *Let us consider*

$$T^{\bullet} = \frac{|(\mathbf{B}^{\bullet} - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^{\bullet} - \mathbf{B})|}{|(n-p)\mathbf{S}^{\bullet}|}. \tag{2.17}$$

Its distribution can be obtained from the decomposition

$$T^{\bullet}|_{\mathbf{\Omega}} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_i \right\} |2\mathbf{I}_m + \mathbf{\Omega}|$$

where $\overset{st}{\sim}$ *means 'stochastic equivalent to' and where* $F_i \sim F_{p-i+1,(n-p)-i+1}$ *are independent random variables, themselves independent of* $\mathbf{\Omega}$*, which has the same distribution as that of* $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$ *where* $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$ *and* $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ *are two independent random variables.*

Proof. Let us recall the distributions of $\mathbf{S}^{\bullet}$ and $\mathbf{B}^{\bullet}$ in (2.11) and (2.12) and that conditionally on $\tilde{\Sigma}^{-1}$, $\mathbf{S}^{\bullet}$ is independent of $\mathbf{B}^{\bullet}$.

Let us also recall $\mathbf{H}$ and $\mathbf{G}$ defined in (2.13) and (2.14), whose distributions are given in (2.15) and (2.16). Given the independence of $\mathbf{B}^{\bullet}$ and $\mathbf{S}^{\bullet}$, conditionally on $\tilde{\Sigma}$, $\mathbf{H}$ will be independent of $\mathbf{G}$.

Since $T^{\bullet}$ in (2.17) can be written as

$$T^{\bullet} = \frac{|(\mathbf{B}^{\bullet} - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^{\bullet} - \mathbf{B})|}{|(n-p)\mathbf{S}^{\bullet}|} = \frac{\left|2\tilde{\Sigma} + \Sigma\right|}{|\tilde{\Sigma}|} \times \frac{|\mathbf{H}|}{|\mathbf{G}|},$$

where $|\mathbf{G}| \sim \prod_{i=1}^{m} \chi^2_{n-p-i+1}$ and $|\mathbf{H}| \sim \prod_{i=1}^{m} \chi^2_{p-i+1}$, with independent chi-square random variables in each product, the distribution of $|\mathbf{H}|/|\mathbf{G}|$, given $\tilde{\Sigma}^{-1}$, is that of a product of independent F-distributions, given the independence of $\mathbf{H}$ and $\mathbf{G}$. Since the distributions of $\mathbf{H}$ and $\mathbf{G}$, respectively given in (2.15) and (2.16), are not function of $\tilde{\Sigma}$ then we will have that they will be independent of $\left|2\tilde{\Sigma} + \Sigma\right|/|\tilde{\Sigma}|$, therefore making this latter ratio independent of $|\mathbf{H}|/|\mathbf{G}|$.

Thus,

$$T^{\bullet}|_{\tilde{\Sigma}^{-1}} \sim \left\{ \prod_{i=1}^{m} \frac{p-i+1}{(n-p)-i+1} F_{p-i+1,(n-p)-i+1} \right\} \times \left| \tilde{\Sigma}^{-1} \left(2\tilde{\Sigma} + \Sigma \right) \right|.$$

Noting that

$$\begin{aligned} \left| \tilde{\Sigma}^{-1} \left(2\tilde{\Sigma} + \Sigma \right) \right| &= \left| 2\mathbf{I} + \tilde{\Sigma}^{-1}\Sigma \right| = \left| 2\Sigma^{-1} + \tilde{\Sigma}^{-1} \right| |\Sigma| \\ &= \left| \Sigma^{1/2} \right| \left| 2\Sigma^{-1} + \tilde{\Sigma}^{-1} \right| \left| \Sigma^{1/2} \right| = \left| 2\mathbf{I} + \Sigma^{1/2}\tilde{\Sigma}^{-1}\Sigma^{1/2} \right|, \end{aligned}$$

from (2.6), integrating out $\mathbf{B}^{\bullet}$ and $\mathbf{S}^{\bullet}$ we end up with

$$\begin{aligned} f_{\Sigma}(\tilde{\Sigma}^{-1}) &\propto |\tilde{\Sigma}|^{\frac{n-p}{2}} \left| 2\tilde{\Sigma} + \Sigma \right|^{\frac{p}{2}} \frac{1}{|\tilde{\Sigma}|^{\frac{(n-p)+n-\alpha}{2}-m-1}} |\Sigma|^{-\frac{n}{2}} \\ &\qquad \times \left| \frac{1}{2}\tilde{\Sigma}^{-1} + \Sigma^{-1} \right|^{-p/2} |\tilde{\Sigma}^{-1} + \Sigma^{-1}|^{-\frac{2n+\alpha-2p-m-1}{2}} \\ &\propto |\tilde{\Sigma}^{-1}|^{\frac{n+\alpha-2m-2}{2}} \left| 2\tilde{\Sigma} + \Sigma \right|^{\frac{p}{2}} |\Sigma|^{-\frac{n}{2}} \left| \frac{1}{2}\tilde{\Sigma}^{-1} + \Sigma^{-1} \right|^{-p/2} |\tilde{\Sigma}^{-1} + \Sigma^{-1}|^{-\frac{2n+\alpha-2p-m-1}{2}}. \end{aligned}$$

Making the transformation $\mathbf{\Omega} = \Sigma^{\frac{1}{2}}\tilde{\Sigma}^{-1}\Sigma^{\frac{1}{2}}$, which implies that $\tilde{\Sigma}^{-1} = \Sigma^{-\frac{1}{2}}\mathbf{\Omega}\Sigma^{-\frac{1}{2}}$ (the Jacobian of the transformation of $\tilde{\Sigma}^{-1}$ to $\mathbf{\Omega}$ is $|\Sigma|^{-\frac{m+1}{2}}$) we have

$$f(\mathbf{\Omega}) \propto |\mathbf{\Omega}|^{\frac{n+\alpha-2m-2}{2}} \left| 2\mathbf{\Omega}^{-1} + \mathbf{I_m} \right|^{\frac{p}{2}} \left| \frac{1}{2}\mathbf{\Omega} + \mathbf{I_m} \right|^{-p/2} |\mathbf{\Omega} + \mathbf{I_m}|^{-\frac{2n+\alpha-2p-m-1}{2}}.$$

Since $|2\mathbf{\Omega}^{-1} + \mathbf{I_m}|^{\frac{p}{2}} = 2^{p/2} \left| \frac{1}{2}\mathbf{\Omega} + \mathbf{I_m} \right|^{\frac{p}{2}} |\mathbf{\Omega}|^{-\frac{p}{2}}$ we end up having

$$f(\mathbf{\Omega}) \propto |\mathbf{\Omega}|^{\frac{n+\alpha-p-2m-2}{2}} |\mathbf{\Omega} + \mathbf{I_m}|^{-\frac{2n+\alpha-2p-m-1}{2}}$$

not function of Σ, where from [26, Theorem 8.2.8.] $\mathbf{\Omega}$ has the same distribution as that of $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$, where $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$ and $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ are two independent random variables, and where $\mathbf{\Omega}$, being a function of $\tilde{\Sigma}$, is independent of $|\mathbf{H}|/|\mathbf{G}|$.

□

Remark 2.1.1. *When $m = 1$ and $M = 1$, the statistic in (2.17) reduces to the statistic T^2 used in [17] whose pdf is obtained by noting that*

$$T^2|_{\Omega=\omega} \sim \frac{p}{n-p}(2+\omega)F_{p,n-p} \quad \textit{where} \quad f_{\Omega}(\omega) \propto \frac{\omega^{\frac{n+\alpha-p-4}{2}}}{(1+\omega)^{\frac{2n+\alpha-2p-2}{2}}}.$$

In Table 2.1 are listed the simulated 0.05 cut-off points for $T^{\bullet}$ for some values of p, m and n.

Table 2.1: Cut-off points of the 95% confidence set for the regression coefficients matrix **B**.

	$p=3$			
n	$m=1$		$m=3$	
	$\alpha=2$	$\alpha=4$	$\alpha=4$	$\alpha=6$
10	6.568	7.433	20.11	29.08
50	5.502E-01	5.581E-01	9.277E-03	9.691E-03
100	2.518E-01	2.542E-01	9.212E-04	9.443E-04
200	1.207E-01	1.208E-01	1.049E-04	1.064E-04
	$p=4$			
n	$m=1$		$m=3$	
	$\alpha=2$	$\alpha=4$	$\alpha=4$	$\alpha=6$
10	11.08	12.69	239.2	372.7
50	6.884E-01	6.984E-01	3.550E-02	3.697E-02
100	3.108E-01	3.128E-01	3.487E-03	3.564E-03
200	1.487E-01	1.490E-01	3.674E-04	3.723E-04

If instead of testing the regression coefficients matrix **B**, someone wants to test a linear combination of the parameters in **B**, namely, $\mathbf{C}=\mathbf{AB}$ where **A** is a $k\times p$ matrix with $rank(\mathbf{A})=k\leq p$ and $k\geq m$, one may define

$$T^{\bullet}_{\mathbf{C}}=\frac{|(\mathbf{AB}^{\bullet}-\mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{AB}^{\bullet}-\mathbf{C})|}{|(n-p)\mathbf{S}^{\bullet}|},$$

which will present a similar distribution to that of $T^{\bullet}$ in (2.17). If one notes that

$$(\mathbf{AB}^{\bullet}-\mathbf{AB})'|_{\tilde{\Sigma}^{-1}}\sim N_{mk}\Big(\mathbf{0},\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}'\otimes(2\tilde{\Sigma}+\Sigma)\Big),$$

and that

$$(\mathbf{AB}^{\bullet}-\mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}')^{-1}(\mathbf{AB}^{\bullet}-\mathbf{C})|_{\tilde{\Sigma}^{-1}}\sim W_m(2\tilde{\Sigma}+\Sigma,k),$$

easily realizing that

$$T^{\bullet}_{\mathbf{C}}|_{\mathbf{\Omega}}\overset{st}{\sim}\left\{\prod_{i=1}^{m}\frac{k-i+1}{(n-p)-i+1}F_{k,i}\right\}|2\mathbf{I}_m+\mathbf{\Omega}|, \tag{2.18}$$

with $F_{k,i}$ as independent random variables with $F_{k-i+1,(n-p)-i+1}$ distribution, themselves independent of $\mathbf{\Omega}$ defined in Theorem 2.1.2.

So, if one wants to perform the test

$$H_0:\mathbf{C}=\mathbf{C}_0\ \ versus\ \ H_1:\mathbf{C}\neq\mathbf{C}_0,$$

should reject H_0 whenever $T^{\bullet}_{\mathbf{C}_0}$ exceeds $\omega_{k,m,n,p;\gamma}$, where $\omega_{k,m,n,p;\gamma}$ satisfies $(1-\gamma)=Pr(T^{\bullet}_{\mathbf{C}_0}\leq\omega_{k,m,n,p;\gamma})$ when H_0 is true. The value of $\omega_{k,m,n,p;\gamma}$ can be obtained by

simulating the distribution in (2.18). To perform a test for $\mathbf{B} = \mathbf{B_0}$ one must take $\mathbf{A} = \mathbf{I_p}$, $\mathbf{C}_0 = \mathbf{B}_0$ and $k = p$ in (2.18).

A $(1-\gamma)$ level confidence set for $\mathbf{C}$ will be given by

$$\Delta(\mathbf{C}) = \{\mathbf{C} : T^{\bullet}_{\mathbf{C}} \leq \omega_{k,m,n,p;\gamma}\}. \tag{2.19}$$

The inference for the regression coefficients when just a single partially synthetic dataset is released is now made available fulfilling the existing gap and fulfilling the first objective of this work.

This derivation of an exact inference procedure for the singly imputed synthetic data generated via PPS, allows the derivation of an exact inferential procedure for multiply imputed synthetic datasets generated via PPS, as shown in subsection 2.2.2.

2.2 Multiple imputation: Posterior Predictive Sampling

In Chapter 1, it was referred the inclusion of multiple imputation as a SDC technique for partially synthetic data, treating the values from *sensitive* variables as missing data and replacing these by synthesized values. These values may be generated independently M times via PPS method, which is the most common method when dealing with missing data.

To specify in detail the PPS method in the MLR model context, let us consider again the model (1.1) with $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{B}$, Σ, $\hat{\mathbf{B}}$ and $\mathbf{S}$ defined in that context.

The synthetic data will consist of M synthetic versions of $\mathbf{Y}$ generated based on the PPS method.

Let us consider the joint prior distribution

$$\pi(\mathbf{B}, \Sigma) \propto |\Sigma|^{-\alpha/2},$$

as in subsection 2.1.1 leading to the same posterior distributions for Σ and $\mathbf{B}$ as in (2.1) and (2.2), assuming that $n + \alpha > p + m + 1$. Consequently, we draw $\tilde{\Sigma}_1$ from (2.1) and $\tilde{\mathbf{B}}_1$ from (2.2), generating the first synthetic dataset, denoted as $\mathbf{Z}_1 = (\mathbf{z}_{11}, \ldots, \mathbf{z}_{1n})$ where $\mathbf{z}_{1i} = (\mathbf{z}_{11i}, \ldots, \mathbf{z}_{m1i})$, are independently distributed as

$$\mathbf{z}_{1i}|_{\tilde{\mathbf{B}}_1, \tilde{\Sigma}_1} \sim N_m\left(\tilde{\mathbf{B}}_{\mathbf{1}}' \mathbf{x_i}, \tilde{\Sigma}_1\right) \ , i = 1, \ldots, n. \tag{2.20}$$

Then, we repeat the same procedure in order to generate i.i.d. $\mathbf{Z}_2, \ldots, \mathbf{Z}_M$, by drawing sequentially $\tilde{\Sigma}_2, \ldots, \tilde{\Sigma}_M$ and $\tilde{\mathbf{B}}_2, \ldots, \tilde{\mathbf{B}}_M$, in order to generate i.i.d.

$\mathbf{Z}_2,\ldots,\mathbf{Z_M}$.

2.2.1 Reiter's adapted methodology

Before presenting the development of an exact inference procedure, we will present an adaptation of Reiter's [31] methodology for drawing inference based on multiply synthetic data generated via PPS for a vector valued parameter, to the inference on a matrix value parameter.

In order to be possible to use Reiter's methodology, developed for a vector of parameters, to estimate $\mathbf{B}$, a $p\times m$ dimensional matrix parameter, let us consider $\mathbf{vec}(\mathbf{B}) = (\mathbf{B}_1' \ \mathbf{B}_2' \ \ldots \ \mathbf{B}_m')'$, a $pm\times 1$ vector parameter, where $\mathbf{B}_j\,(j=1,\ldots,m)$ denotes the j-th column of $\mathbf{B}$.

Based on the original data, $\mathbf{vec}(\hat{\mathbf{B}})$ is an estimator of $\mathbf{vec}(\mathbf{B})$ and its covariance matrix estimator is $\mathbf{U} = \mathbf{S}\otimes(\mathbf{XX}')^{-1}$, a $pm\times pm$ matrix. Let $\mathbf{Z}_1,\ldots,\mathbf{Z}_M$ be the M synthetic datasets obtained via PPS. Let $\mathbf{vec}(\mathbf{B}_j^{\dagger}) = \mathbf{vec}((\mathbf{XX}')^{-1}\mathbf{XZ}_j')$ and $\mathbf{U}_j = \mathbf{S}_j^{\dagger}\otimes(\mathbf{XX}')^{-1}$, where $\mathbf{S}_j^{\dagger} = \frac{1}{n-p}(\mathbf{Z}_j - \mathbf{B}_j^{\dagger'}\mathbf{X})(\mathbf{Z}_j - \mathbf{B}_j^{\dagger'}\mathbf{X})'$, for $j=1,\ldots,M$. Note that based on $\mathbf{Z}_j$, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$, $\mathbf{vec}(\mathbf{B}_j^{\dagger})$ is an UE of $\mathbf{vec}(\mathbf{B})$ and $\mathbf{U}_j$ is an UE of its variance. Then the following estimators

$$\mathbf{vec}(\overline{\mathbf{B}}_M^{\dagger}) = \frac{1}{M}\sum_{j=1}^{M}\mathbf{vec}(\mathbf{B}_j^{\dagger}), \qquad \overline{\mathbf{U}}_M = \frac{1}{M}\sum_{j=1}^{M}\mathbf{U}_j, \tag{2.21}$$

$$\mathbf{b}_M = \frac{1}{M-1}\sum_{j=1}^{M}(\mathbf{vec}(\mathbf{B}_j^{\dagger}) - \mathbf{vec}(\overline{\mathbf{B}}_M^{\dagger}))(\mathbf{vec}(\mathbf{B}_j^{\dagger}) - \mathbf{vec}(\overline{\mathbf{B}}_M^{\dagger}))' \tag{2.22}$$

should be Reiter's estimators to be used to draw inference about $\mathbf{B}$, where $\mathbf{vec}(\overline{\mathbf{B}}_M^{\dagger})$ is an estimator for $\mathbf{vec}(\mathbf{B})$, its variance being estimated by $\mathbf{T} = \frac{1}{M}\mathbf{b}_M + \overline{\mathbf{U}}_M$.

Let us consider

$$T_{R,M} = \frac{(\mathbf{vec}(\overline{\mathbf{B}}_M^{\dagger}) - \mathbf{vec}(\mathbf{B}))'(\overline{\mathbf{U}}_M)^{-1}(\mathbf{vec}(\overline{\mathbf{B}}_M^{\dagger}) - \mathbf{vec}(\mathbf{B}))}{pm(1+r)} \tag{2.23}$$

where $r = \frac{tr(\mathbf{b_M}\overline{\mathbf{U}}_{\mathbf{M}}^{-\mathbf{1}})}{Mpm}$. Then following Reiter [31], the distribution of $T_{R,M}$ is approximated by an $F_{pm,w(r)}$ distribution where

$$w(r) = 4 + [pm(M-1) - 4]\left[1 + \frac{1}{r} - \frac{2}{(M-1)rpm}\right]^2. \tag{2.24}$$

This result is one of the most used in the multiple imputation case [1, 4, 7, 8, 13, 20]. Nevertheless it faces some problems. First, it is not adequate for single

imputation cases, since, in fact, was developed only for multiple imputation and also if one takes $M = 1$ in the expression of $w(r)$, above, it becomes meaningless. Thus the inclusion of the single exact inference method in the previous Section. Second, it is asymptotic in nature and not exact, thus it is not fit for relatively small sample sizes.

With this second problem in mind, follows in the next subsection the development of an exact inference procedure for the multiply imputed synthetic datasets via PPS.

2.2.2 Exact inference for multiple imputation cases based in single imputation inference

In order to develop an exact inferential procedure for the multiple imputation case, one first idea could be to obtain the distribution of the mean of the M individual estimators of $\mathbf{B}$, $\mathbf{B}_j^\dagger$, and the distribution of the mean of the M individual estimators of Σ, $\mathbf{S}_j^\dagger$, defined in the previous subsection. Unfortunately, this would be too hard to materialize for the distribution of the mean of the $\mathbf{S}_j^\dagger$ estimators. Since, to obtain the exact pdf of such an estimator, under the MLR model, one would face the problem of deriving the distribution of the sum of variables that follow Wishart distributions with different parameter matrices.

Therefore, a new approach is presented where each synthetic data is seen as an individual sample.

Let

$$\mathbf{B}_j^\dagger = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Z}_j' \tag{2.25}$$

and

$$\mathbf{S}_j^\dagger = \frac{1}{n-p}(\mathbf{Z}_j - \mathbf{B}_j^{\bullet'}\mathbf{X})(\mathbf{Z}_j - \mathbf{B}_j^{\bullet'}\mathbf{X})' \tag{2.26}$$

be respectively the estimators of $\mathbf{B}$ and Σ based on the j-th synthetic dataset ($j = 1,\ldots,M$), which by Lemma 1.4.1 are jointly sufficient for $\mathbf{B}$ and Σ.

Conditionally on $(\tilde{\mathbf{B}}_j, \tilde{\Sigma}_j)$, for every $j = 1,\ldots,M$, $\mathbf{B}_j^\dagger$ will be independent of $\mathbf{S}_j^\dagger$ and $\left\{(\mathbf{B}_1^\dagger, \mathbf{S}_1^\dagger),\ldots,(\mathbf{B}_M^\dagger, \mathbf{S}_M^\dagger)\right\}$ will be jointly sufficient estimators for $\mathbf{B}$ and Σ.

For each $j = 1,\ldots,M$, individually, one may note that, from Section 2.1.1, the MLE of $\mathbf{B}$ would be $\mathbf{B}_j^\dagger$ and $\hat{\mathbf{S}}_j = \frac{n+\alpha-p-2m-2}{n-p}\mathbf{S}_j^\dagger$ would be an UE for Σ.

Then, it is proposed the following test statistic

$$T_M^\dagger = \sum_{j=1}^{M} T_j^\dagger \tag{2.27}$$

with

$$T_j^\dagger = \frac{|(\mathbf{B}_j^\dagger - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}_j^\dagger - \mathbf{B})|}{|(n-p)\mathbf{S}_j^\dagger|}.$$

Let us note that $T_M^\dagger$ will be a pivotal quantity due to the fact that, for $j = 1,\ldots,M$, all $T_j^\dagger$ are not function of any original data parameter. From Theorem 2.1.2 we have that

$$T_j^\dagger|_{\boldsymbol{\Omega}} \overset{st}{\sim} \left\{\prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_i\right\} |2\mathbf{I}_m + \boldsymbol{\Omega}|, \quad j = 1,\ldots,M$$

leading to the conclusion that

$$T_M^\dagger|_{\boldsymbol{\Omega}} \overset{st}{\sim} \sum_{j=1}^{M} \left[\left\{\prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_i\right\} |2\mathbf{I}_m + \boldsymbol{\Omega}|\right],$$

where $\boldsymbol{\Omega}$ has the same distribution as $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$, where $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$, $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ and $F_i \sim F_{p-i+1,n-p-i+1}$ $(i = 1,\ldots,m)$, all independent random variables.

To test a linear combination of the parameters in $\mathbf{B}$, namely, $\mathbf{C} = \mathbf{A}\mathbf{B}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, one defines

$$T_{M,\mathbf{C}}^\dagger = \sum_{j=1}^{M} \frac{|(\mathbf{A}\mathbf{B}_j^\dagger - \mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\mathbf{B}_j^\dagger - \mathbf{C})|}{|(n-p)\mathbf{S}_j^\dagger|}$$

and proceed by noting that

$$T_{M,\mathbf{C}}^\dagger|_{\mathbf{W}} \overset{st}{\sim} \sum_{j=1}^{M} \left[\left\{\prod_{i=1}^{m} \frac{k-i+1}{n-p-i+1} F_{k,i}\right\} |2\mathbf{I}_m + \boldsymbol{\Omega}|\right] \tag{2.28}$$

with $F_{k,i} \sim F_{k-i+1,n-p-i+1}$, all independent variables, themselves independent of $\boldsymbol{\Omega}$, which is defined as above.

In order to test

$$H_0 : \mathbf{C} = \mathbf{C}_0 \quad versus \quad H_1 : \mathbf{C} \neq \mathbf{C}_0,$$

we reject H_0 whenever $T_{M,\mathbf{C}_0}^\dagger$ exceeds $\psi_{M,k,m,n,p;\gamma}$, where $\psi_{M,k,m,n,p;\gamma}$ satisfies $(1-\gamma) = Pr(T_{M,\mathbf{C}_0}^\dagger \leq \psi_{M,k,m,n,p;\gamma})$ when H_0 is true, where the value of $\psi_{M,k,m,n,p;\gamma}$ can be obtained by simulating the distribution in (2.28). Again if one wants to perform a test for $\mathbf{B} = \mathbf{B}_0$ one must take $\mathbf{A} = \mathbf{I}_p$, $\mathbf{C}_0 = \mathbf{B}_0$ and $k = p$ in (2.28).

A $(1-\gamma)$ level confidence set for $\mathbf{C}$ will be given by

$$\Psi_M(\mathbf{C}) = \{\mathbf{C} : T^{\dagger}_{M,\mathbf{C}} \leq \psi_{M,k,m,n,p;\gamma}\}. \tag{2.29}$$

Thus, an exact procedure is made available to draw inference from the released multiply imputed synthetic datasets overcoming the problem implicit in asymptotic based procedures, that is, its inapplicability to small sample sizes.

The procedure developed in this subsection only carries with the small problem that when resorting to Monte Carlo simulations to construct the empirical distribution one may take a fair amount of time to reach a satisfying accuracy for the distribution, since the number of cycles on the simulations will be multiplied by M. Another problem that this procedure faces is the fact that it will not be possible to compute the 'radius' of the confidence sets directly, as one may see in Chapter 4, being only possible to frame this between an upper and lower bound.

2.3 Multiple Imputation: Fixed-Posterior Predictive Sampling (FPPS)

In this subsection, under the MLR model, two new exact likelihood-based procedures are presented for the analysis of synthetic data generated using the new FPPS Sampling method, for which a brief description can be found in subsection 1.2.2. The FPPS method will overcome the problem that rises when one uses the PPS method to generate multiply imputed synthetic datasets, that is, the problem of obtaining the distribution of a sum of Wishart distributions with different parameters when estimating Σ from the synthetic datasets. It is expected that this new method of generating synthetic data will offer a lower level of disclosure risk.

In order to specify in detail the new FPPS method in the MLR model context, let us consider again the model (1.1).

We consider the same joint prior distribution $\pi(\mathbf{B}, \Sigma) \sim |\Sigma|^{-\alpha/2}$, leading to the same posterior distributions for Σ and $\mathbf{B}$ as in expressions (2.1) and (2.2), respectively, assuming $n + \alpha > p + m + 1$.

Now, draw, only once, $\tilde{\Sigma}$ and $\tilde{\mathbf{B}}$ from (2.1) and (2.2), respectively, and generate the M i.i.d. synthetic datasets, denoted as $\mathbf{W}_j = (\mathbf{w}_{j1}, \ldots, \mathbf{w}_{jn}), j = 1, \ldots, M$ where $\mathbf{w}_{ji} = (\mathbf{w}_{1ji}, \ldots, \mathbf{w}_{mji})$, will be independently distributed as

$$\mathbf{w}_{ji}|_{\tilde{\mathbf{B}},\tilde{\Sigma}} \sim N_m(\tilde{\mathbf{B}}'\mathbf{x_i}, \tilde{\Sigma}), i = 1, \ldots, n, j = 1, \ldots, M. \tag{2.30}$$

Observe that now all synthetic versions are independently generated from the same distribution, instead of being generated from M different normal distributions, based on M independent draws from M different posterior distributions.

2.3.1 A First Procedure

For $j = 1,\ldots,M$, let

$$\mathbf{B}_j^\bullet = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}_j'$$

and

$$\mathbf{S}_j^\bullet = \frac{1}{n-p}(\mathbf{W}_j - \mathbf{B}_j^{\bullet'}\mathbf{X})(\mathbf{W}_j - \mathbf{B}_j^{\bullet'}\mathbf{X})'$$

be the estimators of $\mathbf{B}$ and Σ, respectively, which by Lemma 1.4.1 are jointly sufficient for $\mathbf{B}$ and Σ.

Conditionally on $(\tilde{\mathbf{B}},\tilde{\Sigma})$, for each $j = 1,\ldots,M$, $\mathbf{B}_j^\bullet$ is independent of $\mathbf{S}_j^\bullet$ and $\left\{(\mathbf{B}_1^\bullet,\mathbf{S}_1^\bullet),\ldots,(\mathbf{B}_M^\bullet,\mathbf{S}_M^\bullet)\right\}$ are jointly sufficient estimators for $\mathbf{B}$ and Σ.

Let us define also

$$\overline{\mathbf{B}}_M^\bullet = \frac{1}{M}\sum_{j=1}^M \mathbf{B}_j^\bullet \quad \text{and} \quad \overline{\mathbf{S}}_M^\bullet = \frac{1}{M}\sum_{j=1}^M \mathbf{S}_j^\bullet, \tag{2.31}$$

which, given $\tilde{\mathbf{B}}$ and $\tilde{\Sigma}$, are mutually independent. For $p \geq m$ and $n+\alpha > p+m+1$, let us consider the two following Corollaries of Theorems 2.1.1 and 2.1.2.

Corollary 2.3.1. *The joint pdf of $\overline{\mathbf{B}}_M^\bullet$, $\overline{\mathbf{S}}_M^\bullet$ and $\tilde{\Sigma}^{-1}$, for $\overline{\mathbf{B}}_M^\bullet$ and $\overline{\mathbf{S}}_M^\bullet$ defined in (2.31), is proportional to*

$$e^{-\frac{1}{2}tr\left\{\frac{M+1}{M}(\tilde{\Sigma}+\Sigma)^{-1}(\overline{\mathbf{B}}_M^\bullet-\mathbf{B})'\mathbf{X}\mathbf{X}'(\overline{\mathbf{B}}_M^\bullet-\mathbf{B})+M(n-p)\tilde{\Sigma}^{-1}\overline{\mathbf{S}}_M^\bullet\right\}}$$
$$\times\frac{|\overline{\mathbf{S}}_M^\bullet|^{\frac{M(n-p)-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{M(n-p)+n+\alpha}{2}-m-1}}|\Sigma|^{-\frac{n}{2}}\left|\frac{M}{M+1}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right|^{-p/2}|\tilde{\Sigma}^{-1}+\Sigma^{-1}|^{-\frac{2n+\alpha-2p-m-1}{2}}.$$

Proof. The proof is identical to the proof of Theorem 2.1.1, replacing the joint pdf of $(\mathbf{B}^\bullet,\mathbf{S}^\bullet)$ by the joint pdf of $(\overline{\mathbf{B}}_M^\bullet,\overline{\mathbf{S}}_M^\bullet)$, noting that we have

$$\overline{\mathbf{B}}_M^\bullet|_{\tilde{\mathbf{B}},\tilde{\Sigma}} \sim N_{pm}\left(\tilde{\mathbf{B}},\frac{1}{M}\tilde{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$M(n-p)\overline{\mathbf{S}}_M^\bullet|_{\tilde{\Sigma}} \sim W_m(\tilde{\Sigma},M(n-p)),$$

independent of $\overline{\mathbf{B}}_M^\bullet$. □

Corollary 2.3.2. *Let us consider*

$$T_M^{\bullet} = \frac{|(\overline{\mathbf{B}}_M^{\bullet} - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}_M^{\bullet} - \mathbf{B})|}{|M(n-p)\overline{\mathbf{S}}_M^{\bullet}|}. \tag{2.32}$$

Its distribution can be obtained from the decomposition

$$T_M^{\bullet}|\boldsymbol{\Omega} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{p-i+1}{M(n-p)-i+1} F_i \right\} \left| \frac{M+1}{M} \mathbf{I}_m + \boldsymbol{\Omega} \right|$$

where $F_i \sim F_{p-i+1,M(n-p)-i+1}$ *are independent random variables, themselves independent of* $\boldsymbol{\Omega}$*, which has the same distribution as* $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$ *where* $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$ *and* $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ *are two independent random variables.*

Proof. The proof is identical to the proof of Theorem 2.1.2, replacing $\mathbf{B}^{\bullet}$ by $\overline{\mathbf{B}}_M^{\bullet}$ and $\mathbf{S}^{\bullet}$ by $\overline{\mathbf{S}}_M^{\bullet}$, noting that, from the distribution in Corollary 2.3.1,

$$\overline{\mathbf{B}}_M^{\bullet}|_{\tilde{\Sigma}} \sim N_{pm}\left(\mathbf{B}, \left(\frac{M+1}{M}\tilde{\Sigma} + \Sigma\right) \otimes (\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$\overline{\mathbf{S}}_M^{\bullet}|_{\tilde{\Sigma}} \sim W_m\left(\frac{1}{M(n-p)}\tilde{\Sigma}, M(n-p)\right),$$

independent of $\overline{\mathbf{B}}_M^{\bullet}$. □

From the two corollaries above and proceeding similarly as in Section 2.1.1, it is possible to conclude that the MLE of $\mathbf{B}$ is $\overline{\mathbf{B}}_M^{\bullet}$, which will also be unbiased for $\mathbf{B}$, with

$$Var(\overline{\mathbf{B}}_M^{\bullet}) = N_{M,n,m,p,\alpha}\Sigma \otimes (\mathbf{X}\mathbf{X}')^{-1},$$

where

$$N_{M,n,m,p,\alpha} = \frac{2M(n+\frac{\alpha}{2}-p-m-1)+n-p}{M(n+\alpha-p-2m-2)},$$

for $n+\alpha > p+2m+2$. An UE of Σ will be $\hat{\mathbf{S}}_M = \frac{n+\alpha-p-2m-2}{n-p}\overline{\mathbf{S}}_M^{\bullet}$.

From Corollary 2.3.2, we have that $T_M^{\bullet}$ defined in (2.32) is a pivotal quantity for the synthetic datasets since it does not depend on any parameter from the original data in its definition and in its distribution.

In order to perform a test to a linear combination of the parameters in $\mathbf{B}$, namely, $\mathbf{C} = \mathbf{A}\mathbf{B}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$ we define

$$T_{M,\mathbf{C}}^{\bullet} = \frac{|(\mathbf{A}\overline{\mathbf{B}}_M^{\bullet} - \mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}_M^{\bullet} - \mathbf{C})|}{|M(n-p)\overline{\mathbf{S}}_M^{\bullet}|}$$

and proceed by noting that

$$T^{\bullet}_{M,\mathbf{C}}|\boldsymbol{\Omega} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{M(n-p)-i+1} F_{k,i} \right\} \left| \frac{M+1}{M} \mathbf{I}_m + \boldsymbol{\Omega} \right| \tag{2.33}$$

with $F_{k,i} \sim F_{k-i+1,M(n-p)-i+1}$ and $\boldsymbol{\Omega}$ defined as in Corollary 2.3.2.

Therefore, when testing

$$H_0 : \mathbf{C} = \mathbf{C}_0 \quad versus \quad H_1 : \mathbf{C} \neq \mathbf{C}_0$$

we should reject H_0 whenever $T^{\bullet}_{\mathbf{C}_0}$ exceeds $\omega_{M,k,m,n,p;\gamma}$ where $\omega_{M,k,m,n,p;\gamma}$ satisfies $(1-\gamma) = Pr(T^{\bullet}_{M,\mathbf{C}_0} \leq \omega_{M,k,m,n,p;\gamma})$ when H_0 is true, where the value of $\omega_{M,k,m,n,p;\gamma}$ is obtained by simulating the distribution in (2.28). To perform a test for $\mathbf{B} = \mathbf{B}_0$ one must take $\mathbf{A} = \mathbf{I}_p$, $\mathbf{C}_0 = \mathbf{B}_0$ and $k = p$ in (2.33).

A $(1-\gamma)$ level confidence set for $\mathbf{C}$ is given by

$$\Delta^{\bullet}_{M}(\mathbf{C}) = \{\mathbf{C} : T^{\bullet}_{M,\mathbf{C}} \leq \omega_{M,k,m,n,p;\gamma}\}. \tag{2.34}$$

2.3.2 A Second Procedure

One may use more information from the released synthetic data if more information about Σ is included in the test statistic used to perform inference about $\mathbf{B}$. Therefore, we propose a second likelihood-based approach for exact inference about $\mathbf{B}$, which is expected to offer more precision in the inference analysis than the previous procedure.

Let us recall that $\mathbf{W}_j\,(j = 1,\ldots,M)$, is a $m \times n$ matrix formed by the vectors $(\mathbf{w}_{j1},\ldots,\mathbf{w}_{jn})$ as columns, generated from (2.30) and note that, conditionally on $\tilde{\mathbf{B}}$ and $\tilde{\Sigma}$, $(\mathbf{w}_{1i},\ldots,\mathbf{w}_{Mi})$ is a random sample from $N_m(\tilde{\mathbf{B}}'\mathbf{x}_i, \tilde{\Sigma})$, for each $i = 1,\ldots,n$.

Let us consider

$$\overline{\mathbf{w}}_i = \frac{1}{M}\sum_{j=1}^{M} \mathbf{w}_{ji} \quad \text{and} \quad \mathbf{S}_{\mathbf{w}i} = \sum_{j=1}^{M} (\mathbf{w}_{ji} - \overline{\mathbf{w}}_i)(\mathbf{w}_{ji} - \overline{\mathbf{w}}_i)'$$

which are sufficient statistics for Σ, based on the i-th covariate vector.

Defining

$$\mathbf{S}_{\mathbf{w}} = \sum_{i=1}^{n} \mathbf{S}_{\mathbf{w}i}, \tag{2.35}$$

one has $(\overline{\mathbf{w}}_1,\ldots,\overline{\mathbf{w}}_n,\mathbf{S}_{\mathbf{w}})$ as the joint sufficient statistics for $(\mathbf{B},\Sigma)$.

Conditionally on $\tilde{\mathbf{B}}$ and $\tilde{\Sigma}$, we have that

$$\overline{\mathbf{w}}_i \sim N_m(\tilde{\mathbf{B}}'\mathbf{x}_i, \frac{1}{M}\tilde{\Sigma}) \quad \text{and} \quad \mathbf{S}_{\mathbf{w}i} \sim W_m(\tilde{\Sigma}, M-1).$$

From the M released synthetic data matrices $\mathbf{W}_j, j = 1,\ldots,M$, let us define

$$\overline{\mathbf{W}}_M = \frac{1}{M}\sum_{j=1}^{M}\mathbf{W}_j \tag{2.36}$$

and for $\mathbf{B}$ its estimator

$$\overline{\mathbf{B}}^{\bullet}_M = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\overline{\mathbf{W}}'_M, \tag{2.37}$$

and for Σ its estimator

$$\mathbf{S}^{\bullet}_{comb} = \frac{\mathbf{S}_{\mathbf{w}} + M \times \mathbf{S}^{\bullet}_{mean}}{Mn - p} \tag{2.38}$$

where we take

$$\mathbf{S}^{\bullet}_{mean} = \left(\overline{\mathbf{W}}_M - \overline{\mathbf{B}}^{\bullet'}_M\mathbf{X}\right)'\left(\overline{\mathbf{W}}_M - \overline{\mathbf{B}}^{\bullet'}_M\mathbf{X}\right). \tag{2.39}$$

We use again the notation $\overline{\mathbf{B}}^{\bullet}_M$ for the present estimator of $\mathbf{B}$ since it is indeed the same estimator used in the first procedure, since for $\overline{\mathbf{B}}^{\bullet}_M$ in (2.37)

$$\overline{\mathbf{B}}^{\bullet}_M = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\overline{\mathbf{W}}'_M = \frac{1}{M}\sum_{j=1}^{M}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}'_j = \frac{1}{M}\sum_{j=1}^{M}\mathbf{B}^{\bullet}_j.$$

It is with the estimator $\mathbf{S}^{\bullet}_{comb}$ that a surplus of information about Σ is acquired, when compared with the estimator used in the first procedure.

In fact, if the M synthetic datasets are treated as a single big synthetic sample of size nM, the estimators obtained for $\mathbf{B}$ and Σ would actually be the same as those in (2.37) and (2.38).

With the purpose of proving this fact, let us start by considering the synthetic datasets as one only sample of size nM arranged as

$$\left(\begin{array}{c} \mathbf{W}_a \\ \mathbf{X}_a \end{array}\right) = \left(\begin{array}{c|c|c|c} \mathbf{W}_1 & \mathbf{W}_2 & \ldots & \mathbf{W}_M \\ \hline \mathbf{X} & \mathbf{X} & \ldots & \mathbf{X} \end{array}\right),$$

where $\mathbf{W}_a = (\mathbf{W}_1|\ldots|\mathbf{W}_M)$ is the $m \times nM$ matrix of the synthesized data under FPPS and $\mathbf{X}_a = (\mathbf{X}|\ldots|\mathbf{X})$ the $p \times nM$ matrix of the M repeated 'fixed' sets of covariates, from the original data.

Let

$$\mathbf{B}_a = (\mathbf{X}_a\mathbf{X}_a')^{-1}\mathbf{X}_a\mathbf{W}_a'$$

be the estimator of $\mathbf{B}$, based on the dataset of size nM, obtained by joining all the M synthetic datasets in one only dataset. Consequently one has that

$$\begin{aligned}\mathbf{B}_a &= (\mathbf{X}_a\mathbf{X}_a')^{-1}\mathbf{X}_a\mathbf{W}_a' = (M(\mathbf{X}\mathbf{X}'))^{-1}\mathbf{X}_a\mathbf{W}_a' = \frac{1}{M}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}_a\mathbf{W}_a' \\ &= \frac{1}{M}(\mathbf{X}\mathbf{X}')^{-1}\Big(\underbrace{\mathbf{X}|\ \ldots\ |\mathbf{X}}_{M\ times}\Big)\mathbf{W}_a' = \frac{1}{M}\left((\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}_1 + \ldots + (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}_M\right) \\ &= \frac{1}{M}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\left(\mathbf{W}_1 + \ldots + \mathbf{W}_M\right) = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\overline{\mathbf{W}}_M,\end{aligned}$$

which is same estimator for $\mathbf{B}$ as that in (2.37).

Now let

$$\mathbf{S}_a = \frac{1}{nM-p}(\mathbf{W}_a - \mathbf{B}_a'\mathbf{X}_a)(\mathbf{W}_a - \mathbf{B}_a'\mathbf{X}_a)'$$

be the estimator for Σ, based on the dataset of size nM, obtained by joining the M synthetic datasets in one only dataset.

Observe that $\overline{\mathbf{W}}_M$, defined in (2.36), can be written as

$$\overline{\mathbf{W}}_M = \frac{1}{M}\mathbf{W}_a\mathbf{R}$$

with $\mathbf{R} = \left(\overrightarrow{\mathbf{1}}_M \otimes \mathbf{I}_n\right)$ where $\overrightarrow{\mathbf{1}}_M$ is a vector of 1's of size M.

Now let us consider the estimator $\mathbf{S}_{\mathbf{w}}$ of Σ, defined in (2.35). This estimator may be written as

$$\mathbf{S}_{\mathbf{w}} = \sum_{i=1}^{n}\sum_{j=1}^{M}(\mathbf{w}_{ji} - \overline{\mathbf{w}}_i)(\mathbf{w}_{ji} - \overline{\mathbf{w}}_i)',$$

where $\mathbf{w}_{ji}$ is the i-th column of $\mathbf{W}_j$ $(i = 1,\ldots,n; j = 1,\ldots,M)$. We may thus write

$$\begin{aligned}\mathbf{S}_{\mathbf{w}} &= \left(\mathbf{W}_a - \overrightarrow{\mathbf{1}}'_M \otimes \overline{\mathbf{W}}_M\right)\left(\mathbf{W}_a - \overrightarrow{\mathbf{1}}'_M \otimes \overline{\mathbf{W}}_M\right)' \\ &= \left(\mathbf{W}_a - \frac{1}{M}\overrightarrow{\mathbf{1}}'_M \otimes (\mathbf{W}_a\mathbf{R})\right)\left(\mathbf{W}_a - \frac{1}{M}\overrightarrow{\mathbf{1}}'_M \otimes (\mathbf{W}_a\mathbf{R})\right)' \\ &= \left(\mathbf{W}_a - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\right)\left(\mathbf{W}_a - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\right)',\end{aligned}$$

while the estimator $\mathbf{S}_{mean}$ of Σ, defined in (2.39), may be written as

$$\mathbf{S}_{mean} = \left(\frac{1}{M}\mathbf{W}_a\mathbf{R} - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\right)\left(\frac{1}{M}\mathbf{W}_a\mathbf{R} - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\right)'.$$

We may therefore write the combination estimator $\mathbf{S}_{comb}$ defined in (2.38) as

$$\mathbf{S}_{comb} = \frac{1}{nM-p}\left[\left(\mathbf{W}_a - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\right)\left(\mathbf{W}_a - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\right)'\right] + \frac{1}{nM-p}\left[M \times \left(\frac{1}{M}\mathbf{W}_a\mathbf{R} - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\right)\left(\frac{1}{M}\mathbf{W}_a\mathbf{R} - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\right)'\right].$$

To prove that $\mathbf{S}_{comb}$ is equal to $\mathbf{S}_a$ it will only be necessary to focus on

$$\begin{aligned}&\left(\mathbf{W}_a - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\right)\left(\mathbf{W}_a - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\right)' \\ &\qquad + M \times \left(\frac{1}{M}\mathbf{W}_a\mathbf{R} - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\right)\left(\frac{1}{M}\mathbf{W}_a\mathbf{R} - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\right)' \\ &= \mathbf{W}_a\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' + \frac{1}{M^2}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{R}\mathbf{R}'\mathbf{W}_a' \\ &\quad + \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{X}_a'\mathbf{B}_a + \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\mathbf{R}'\mathbf{X}_a'\mathbf{B}_a,\end{aligned}$$

which, using the fact that $\frac{1}{M}\mathbf{X}_a\mathbf{R}\mathbf{R}' = \mathbf{X}_a$ and $\frac{1}{M}\mathbf{R}\mathbf{R}'\mathbf{R}\mathbf{R}' = \mathbf{R}\mathbf{R}'$, may be written as

$$\begin{aligned}&\mathbf{W}_a\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' + \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' \\ &\quad + \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \mathbf{B}_a'\mathbf{X}_a\mathbf{W}_a' - \mathbf{W}_a\mathbf{X}_a'\mathbf{B}_a + \mathbf{B}_a'\mathbf{X}_a\mathbf{X}_a'\mathbf{B}_a \\ &= \mathbf{W}_a\mathbf{W}_a' - \mathbf{B}_a'\mathbf{X}_a\mathbf{W}_a' - \mathbf{W}_a\mathbf{X}_a'\mathbf{B}_a + \mathbf{B}_a'\mathbf{X}_a\mathbf{X}_a'\mathbf{B}_a \\ &= (\mathbf{W}_a - \mathbf{B}_a'\mathbf{X}_a)(\mathbf{W}_a - \mathbf{B}_a'\mathbf{X}_a)' = (nM-p)\mathbf{S}_a.\end{aligned}$$

Therefore, $\mathbf{S}_{comb} = \mathbf{S}_a$ as it was referred.

In future derivations it will be used $\mathbf{S}_{comb}$ instead of $\mathbf{S}_a$ to be easier to recall that this estimator contains a *combination* of information gathered about Σ.

Following, important Corollaries of Theorems 2.1.1 and 2.1.2 in order to develop inference analysis for the matrix of regressor coefficients $\mathbf{B}$ are presented.

Corollary 2.3.3. *For $p \geq m$, $n+\alpha > p+m+1$, the joint pdf of $\overline{\mathbf{B}}_M^{\bullet}$, $\mathbf{S}_{comb}^{\bullet}$, $\tilde{\Sigma}^{-1}$, for $\overline{\mathbf{B}}_M^{\bullet}$ and $\mathbf{S}_{comb}^{\bullet}$ defined in (2.37) and (2.38), is proportional to*

$$e^{-\frac{1}{2}tr\left\{\left(\frac{\mathbf{M}+1}{\mathbf{M}}\tilde{\Sigma}+\Sigma\right)^{-1}(\overline{\mathbf{B}}_M^{\bullet}-\mathbf{B})'\mathbf{X}\mathbf{X}'(\overline{\mathbf{B}}_M^{\bullet}-\mathbf{B})+(Mn-p)\tilde{\Sigma}^{-1}\mathbf{S}_{comb}^{\bullet}\right\}} \times \frac{|\mathbf{S}_{comb}^{\bullet}|^{\frac{Mn-p-m-1}{2}}}{|\tilde{\Sigma}|^{\frac{Mn-p+n+\alpha}{2}-m-1}}|\Sigma|^{-\frac{n}{2}}\left|\frac{M}{M+1}\tilde{\Sigma}^{-1}+\Sigma^{-1}\right|^{-p/2}|\tilde{\Sigma}^{-1}+\Sigma^{-1}|^{-\frac{2n+\alpha-2p-m-1}{2}}.$$

Proof. The proof is identical to the proof of Theorem 2.1.1 replacing the joint pdf of $(\mathbf{B}^{\bullet}, \mathbf{S}^{\bullet})$ by the joint pdf of $(\overline{\mathbf{B}}_M^{\bullet}, \mathbf{S}_{comb}^{\bullet})$, noting that we have

$$\overline{\mathbf{B}}_M^{\bullet}|_{\tilde{\mathbf{B}},\tilde{\Sigma}} \sim N_{pm}\left(\tilde{\mathbf{B}}, \frac{1}{M}\tilde{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$(Mn-p)\mathbf{S}^{\bullet}_{comb}|_{\tilde{\Sigma}} \sim W_m(\tilde{\Sigma}, Mn-p),$$

independent of $\overline{\mathbf{B}}^{\bullet}_{M}$.

□

Corollary 2.3.4. *Let us consider*

$$T^{\bullet}_{comb} = \frac{|(\overline{\mathbf{B}}^{\bullet}_{M}-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^{\bullet}_{M}-\mathbf{B})|}{|(Mn-p)\mathbf{S}^{\bullet}_{comb}|}. \tag{2.40}$$

For $p \geq m$, $n+\alpha > p+m+1$, its distribution can be obtained from the decomposition

$$T^{\bullet}_{comb}|_{\mathbf{\Omega}} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{p-i+1}{Mn-p-i+1} F_i \right\} \left| \frac{M+1}{M}\mathbf{I}_m + \mathbf{\Omega} \right|$$

where $F_i \sim F_{p-i+1,Mn-p-i+1}$ are independent random variables, themselves independent of $\mathbf{\Omega}$, which has the same distribution as $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$ where $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$ and $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ are two independent random variables.

Proof. The proof is identical to the proof of Theorem 2.1.2 replacing $\mathbf{B}^{\bullet}$ by $\overline{\mathbf{B}}^{\bullet}_{M}$ and $\mathbf{S}^{\bullet}$ by $\mathbf{S}^{\bullet}_{comb}$, noting that from the distribution of Corollary 2.3.3,

$$\overline{\mathbf{B}}^{\bullet}_{M}|_{\tilde{\Sigma}} \sim N_{pm}\left(\mathbf{B}, \left(\frac{M+1}{M}\tilde{\Sigma}+\Sigma\right)\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$\mathbf{S}^{\bullet}_{comb}|_{\tilde{\Sigma}} \sim W_m\left(\frac{1}{Mn-p}\tilde{\Sigma}, Mn-p\right),$$

independent of $\overline{\mathbf{B}}^{\bullet}_{M}$.

□

From the two Corollaries above, one easily concludes that an UE of Σ will be $\hat{\mathbf{S}}_M = \frac{n+\alpha-p-2m-2}{n-p}\mathbf{S}^{\bullet}_{comb}$.

Similar to what happened in subsection 2.3.1, we have that $T^{\bullet}_{comb}$ defined in (2.40) is also a pivotal quantity for the synthetic datasets since it does not depend on any parameter from the original data in its definition and in its distribution.

If one wants to test a linear combination of the parameters in $\mathbf{B}$, namely, $\mathbf{C} = \mathbf{A}\mathbf{B}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$ one defines

$$T^{\bullet}_{comb,\mathbf{C}} = \frac{|(\mathbf{A}\overline{\mathbf{B}}^{\bullet}_{M}-\mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}^{\bullet}_{M}-\mathbf{C})|}{|(Mn-p)\overline{\mathbf{S}}^{\bullet}_{comb}|}$$

and proceeds by noting that

$$T^{\bullet}_{comb,\mathbf{C}}|_{\mathbf{\Omega}} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{Mn-p-i+1} F_{k,i} \right\} \left| \frac{M+1}{M} \mathbf{I}_m + \mathbf{\Omega} \right| \tag{2.41}$$

with $F_{k,i} \sim F_{k-i+1,Mn-p-i+1}$ and $\mathbf{\Omega}$ defined in Corollary 2.3.3, all independent variables.

In order to perform the test

$$H_0 : \mathbf{C} = \mathbf{C}_0 \quad \text{versus} \quad H_1 : \mathbf{C} \neq \mathbf{C}_0$$

one will reject H_0 whenever $T^{\bullet}_{comb,\mathbf{C}_0}$ exceeds $\omega_{comb,M,k,m,n,p;\gamma}$ where $\omega_{comb,M,k,m,n,p;\gamma}$ satisfies $(1-\gamma) = Pr(T^{\bullet}_{comb,\mathbf{C}_0} \leq \omega_{comb,k,m,n,p;\gamma})$ when H_0 is true, where the value of $\omega_{comb,M,k,m,n,p;\gamma}$ can be obtained by simulating the distribution in (2.41).

A $(1-\gamma)$ level confidence set for $\mathbf{C}$ is given by

$$\Delta^{\bullet}_{comb}(\mathbf{C}) = \{\mathbf{C} : T^{\bullet}_{comb,\mathbf{C}} \leq \omega_{comb,M,k,m,n,p;\gamma}\}. \tag{2.42}$$

It is expected that this second exact procedure of analyzing the synthetic data generated via FPPS will offer a better precision than the first one, at least for microdata with small sample sizes, that is, originating smaller confidence sets; in other words, confidence sets with smaller *radius*. The main difference between the distributions in (2.33) and (2.41) is observed in the denominator degrees of freedom of the F distributions. As n and M increases these degrees of freedom become closer and closer and one may see that these two methods will become identical.

In fact, making a simple scale change, the distributions from both procedures converge in distribution to the same distribution. Concerning the first procedure, making a scale change, in $T^{\bullet}_{M}$ defined in (2.33) one will have that

$$(M(n-p))^m \, T^{\bullet}_{M}|_{\mathbf{\Omega}} \xrightarrow[n\to\infty]{d} \left\{ \prod_{i=1}^{m} \chi^2_{p-i+1} \right\} \left| \frac{M+1}{M} \mathbf{I}_m + \mathbf{\Omega} \right|$$

and

$$(M(n-p))^m \, T^{\bullet}_{M}|_{\mathbf{\Omega}} \xrightarrow[M\to\infty]{d} \left\{ \prod_{i=1}^{m} \chi^2_{p-i+1} \right\} |\mathbf{I}_m + \mathbf{\Omega}|,$$

where $\xrightarrow{d}$ represents convergence in distribution. This is due to the fact that

$$\frac{\chi^2_b}{b} \xrightarrow[b\to\infty]{d} 1 \implies a\, F_{a,b} \xrightarrow[b\to\infty]{d} \chi^2_a.$$

On the other hand, relatively to the second procedure, making a similar scale change in $T^{\bullet}_{comb}$ defined in (2.41), one will also have

$$(Mn-p)^m\, T^{\bullet}_{comb}|\boldsymbol{\Omega} \xrightarrow[n\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} \left|\frac{M+1}{M}\mathbf{I}_m + \boldsymbol{\Omega}\right|$$

and

$$(Mn-p)^m\, T^{\bullet}_{comb}|\boldsymbol{\Omega} \xrightarrow[M\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} |\mathbf{I}_m + \boldsymbol{\Omega}|\,.$$

Therefore concluding that the two FPPS inferential procedures become approximately equal for large values of M, number of synthetic datasets released, and for large values of n, the datasets sample size.

Chapter 3

Inference for Multivariate Regression Model based on synthetic data generated via Plug-in Sampling

In the previous Chapter, we presented likelihood-based inference procedures to draw inference about $\mathbf{B}$ when synthetic datasets generated via PPS method are released. Nevertheless, this is not the only method which one can use to generate replications of the original data and draw inference from them. We can generate synthetic versions of the original data by plugging in the estimators $\hat{\mathbf{B}}$ of $\mathbf{B}$ and $\mathbf{S}$ of Σ, obtained directly from this original data, in the MLR model. This method is called the Plug-in method. Reiter and Kinney [33] showed that using the methodology developed in [31] one can draw inference about the unknown parameters when having access to multiple synthetic datasets generated via Plug-in sampling. This recent method of generating synthetic data is simpler and since it uses directly the original data point estimators plugged in the model one expects to generate datasets with 'better' quality than the synthetic data generated via PPS, thus representing a good alternative to the PPS method. However, Reiter's combination rules are not applicable to single imputation cases and are asymptotic in nature, that is, are not applicable to datasets with small sample size. With this fact in mind, in this Chapter, one will develop likelihood-based exact inferential procedures of analyzing partially-synthetic datasets generated via Plug-in Sampling Method, for the single and multiple imputation case under the MLR model.

Concerning other models, one would like to point out that Klein and Sinha in [18, 19, 20] already developed exact inferential procedures for Plug-in generated synthetic datasets when assuming that the original data follows an exponential or multivariate normal distributions and when a single sensitive variable is modeled using a linear regression model, for the single imputation case, and under the normal model, for both single and multiple imputation.

In this chapter, one will extend their work by presenting likelihood-based exact inferential procedures to analyze partially-synthetic datasets generated via Plug-in Sampling Method, for the single and multiple imputation case under the MLR model, that is, when not only one but several sensitive variables are considered and these are used as a set of response variables in an MLR model. Using this method, a synthetic version of the original data is generated by plugging in directly the original data estimators $\hat{\mathbf{B}}$ of $\mathbf{B}$ and $\mathbf{S}$ of Σ into the MLR model. Since it is a more direct approach of generating synthetic data, it will be expected to present data with 'better' quality than the synthetic data generated via PPS or via FPPS.

3.1 Single Imputation: Plug-in Sampling

Since literature does not contemplate methods of inference for cases when only one synthetic dataset is released by statistical agencies, it will be presented in first place the inferential procedure analyses for the single imputation case.

The released synthetic data will consist of a single synthetic version of $\mathbf{Y}$ generated from the original data $(y_{i1},\dots,y_{im},x_{1i},\dots,x_{pi})$, $i=1,\dots,n$. Considering the MLR model (1.1) and the corresponding point estimators of $\mathbf{B}$ and Σ, $\hat{\mathbf{B}}$ and $\mathbf{S}$, respectively, one plugs these estimators into the joint pdf of $\mathbf{Y}$. The synthetic data, denoted as $\mathbf{V}=(\mathbf{v_1},\dots,\mathbf{v_n})$, is then generated, noting that $\mathbf{v_i}=(v_{1i},\dots,v_{mi})'$ are independently distributed as

$$\mathbf{v_i}|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_m\left(\hat{\mathbf{B}}'\mathbf{x_i},\mathbf{S}\right), i=1,\dots,n. \tag{3.1}$$

Based on the released $\mathbf{V}$ and $\mathbf{X}$, let us define

$$\mathbf{B}^* = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{V}' \quad \text{and} \quad \mathbf{S}^* = \frac{1}{n-p}(\mathbf{V}-\mathbf{B}^{*\prime}\mathbf{X})(\mathbf{V}-\mathbf{B}^{*\prime}\mathbf{X})', \tag{3.2}$$

as the estimators of $\mathbf{B}$ and Σ, respectively. By Lemma 1.4.1 these estimators are jointly sufficient for $(\mathbf{B},\Sigma)$.

In order to perform tests for $\mathbf{B}$ one could propose to adapt the classical test criteria for the MLR model (see [3, Secs 8.3 and 8.6]) by using the estimators

defined in (3.2). Nevertheless, these adaptations would face the same problem as for the PPS case, that is, they are not pivotal, due to the fact that their distributions will be function of Σ, as illustrated in Figure 3.1, where the empirical distributions of

(a) $T_1^* = |\mathbf{S}^*||\mathbf{S}^* + (\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|^{-1}$ (Wilks' Lambda Criterion);

(b) $T_2^* = tr\left[(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})(\mathbf{S}^*)^{-1}\right]$ (Pillai's Trace Criterion);

(c) $T_3^* = tr\left[(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})[(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B}) + \mathbf{S}^*]^{-1}\right]$ (Hotelling - Lawley Trace Criterion);

(d) $T_4^* = \lambda_1$ where λ_1 denotes the largest eigenvalue of $(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})(\mathbf{S}^*)^{-1}$ (Roy's Largest Root Criterion);

are presented considering the case where $m = 2$, $p = 3$, $n = 100$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with $\rho = 0.2, 0.4, 0.6, 0.8$ for a simulation size of 1000.

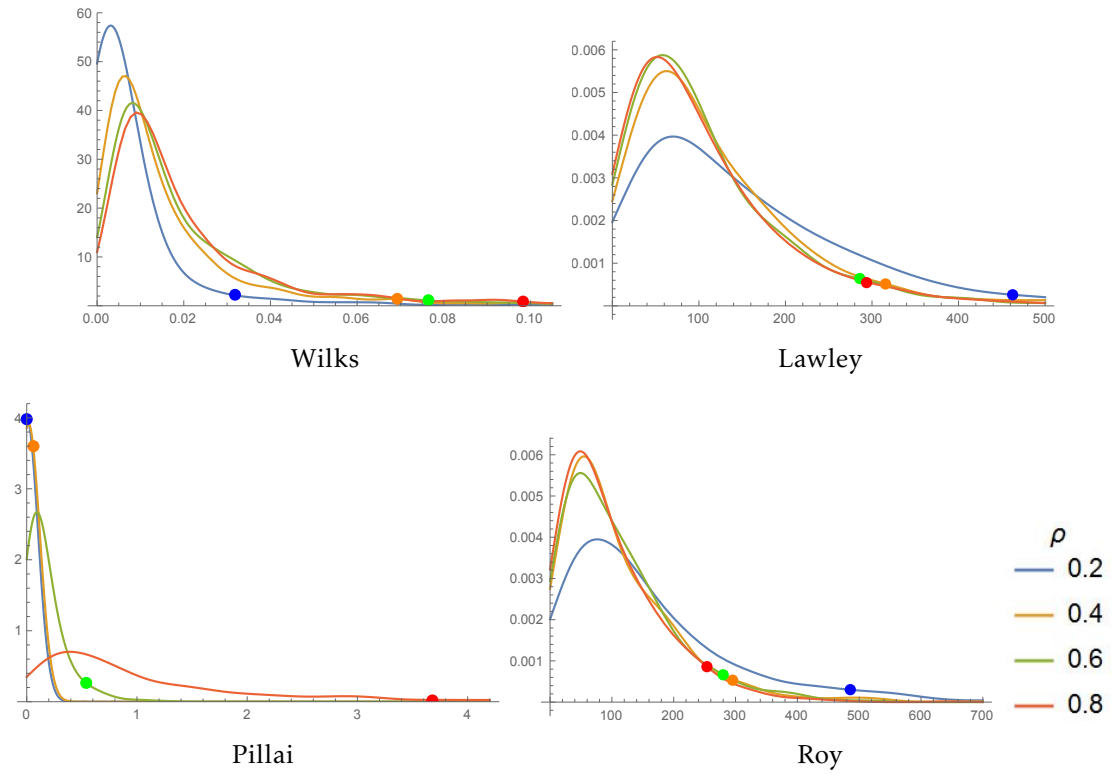


Figure 3.1: Smoothed empirical distributions and cut-off points ($\gamma = 0.05$) of T_1^*, T_2^*, T_3^* and T_4^* for $\rho = 0.2, 0.4, 0.6, 0.8$.

Since it is not possible to use the adaptations of the referred classical test criteria, it is required the introduction of a new pivotal statistic, which we propose to be somewhat similar to the statistic found in Theorem 2.1.2. This proposed

statistic will use the estimators $\mathbf{B}^*$ and $\mathbf{S}^*$ defined in (3.2), thus it will be necessary their joint pdf, in first place. As in Chapter 2, when using the term 'statistic' we are assuming $\mathbf{B}$ known.

Theorem 3.1.1. *The joint pdf of* $\mathbf{B}^*$ *and* $\mathbf{S}^*$ *defined in (3.2), is proportional to*

$$\int e^{-\frac{1}{2}tr\left\{(\Sigma(\mathbf{I}+\mathbf{\Psi}))^{-1}(\mathbf{B}^*-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^*-\mathbf{B})+(n-p)\mathbf{\Psi}^{-1}\Sigma^{-1}\mathbf{S}^*+(n-p)\mathbf{\Psi}\right\}} \times |\mathbf{S}^*|^{\frac{n-p-m-1}{2}} \frac{|\mathbf{\Psi}|^{-\frac{p+m+1}{2}}}{|\Sigma|^{\frac{n-m+1}{2}}} \left|\mathbf{I}_m+\mathbf{\Psi}^{-1}\right|^{-p/2} d\mathbf{\Psi}.$$

Proof. Given $(\hat{\mathbf{B}},\mathbf{S})$, one has for $\mathbf{B}^*$ and $\mathbf{S}^*$ defined in (3.2),

$$\mathbf{V}'|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_{nm}(\mathbf{X}'\hat{\mathbf{B}},\mathbf{S}\otimes\mathbf{I}_n) \implies \mathbf{B}^*|_{\hat{\mathbf{B}},\mathbf{S}} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{V}'|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_{pm}(\hat{\mathbf{B}},\mathbf{S}\otimes(\mathbf{X}\mathbf{X}')^{-1})$$

and

$$(n-p)\mathbf{S}^*|_{\mathbf{S}} \sim W_m(\mathbf{S},n-p).$$

Given the independence of $\mathbf{B}^*$ and $\mathbf{S}^*$, the conditional joint pdf of $(\mathbf{B}^*,\mathbf{S}^*)$ is proportional to

$$e^{-\frac{1}{2}tr\left\{\mathbf{S}^{-1}\left[(\mathbf{B}^*-\hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\mathbf{B}^*-\hat{\mathbf{B}})+(n-p)\mathbf{S}^*\right]\right\}} \frac{|\mathbf{S}^*|^{\frac{n-p-m-1}{2}}}{|\mathbf{S}|^{\frac{n}{2}}}, \tag{3.3}$$

while, given the independence of $\hat{\mathbf{B}}$ and $\mathbf{S}$, defined in (1.2) and (1.3), the joint pdf of $(\hat{\mathbf{B}},\mathbf{S})$ is proportional to

$$e^{-\frac{1}{2}tr\left\{\Sigma^{-1}\left[(\hat{\mathbf{B}}-\mathbf{B})'\mathbf{X}\mathbf{X}'(\hat{\mathbf{B}}-\mathbf{B})+(n-p)\mathbf{S}\right]\right\}} \frac{|\mathbf{S}|^{\frac{n-p-m-1}{2}}}{|\Sigma|^{\frac{n}{2}}}. \tag{3.4}$$

Therefore, the joint pdf of $(\mathbf{B}^*,\mathbf{S}^*,\hat{\mathbf{B}},\mathbf{S})$ will be obtained by multiplying the two joint pdf's (3.3) and (3.4).

Since

$$tr\left\{\mathbf{S}^{-1}(\mathbf{B}^*-\hat{\mathbf{B}})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^*-\hat{\mathbf{B}})+\Sigma^{-1}(\hat{\mathbf{B}}-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\hat{\mathbf{B}}-\mathbf{B})\right\} = tr\left\{(\mathbf{B}^*-\hat{\mathbf{B}})\mathbf{S}^{-1}(\mathbf{B}^*-\hat{\mathbf{B}})'(\mathbf{X}\mathbf{X}')+(\hat{\mathbf{B}}-\mathbf{B})\Sigma^{-1}(\hat{\mathbf{B}}-\mathbf{B})'(\mathbf{X}\mathbf{X}')\right\},$$

where, from (A.2) in Result A.2.2,

$$\begin{aligned}&(\mathbf{B}^*-\hat{\mathbf{B}})\mathbf{S}^{-1}(\mathbf{B}^*-\hat{\mathbf{B}})'+(\hat{\mathbf{B}}-\mathbf{B})\Sigma^{-1}(\hat{\mathbf{B}}-\mathbf{B})' = \\ &= \left[\hat{\mathbf{B}}-(\mathbf{B}^*\mathbf{S}^{-1}+\mathbf{B}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\right](\mathbf{S}^{-1}+\Sigma^{-1})\left[\hat{\mathbf{B}}-(\mathbf{B}^*\mathbf{S}^{-1}+\mathbf{B}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\right]' \\ &\qquad +(\mathbf{B}^*-\mathbf{B})(\mathbf{S}+\Sigma)^{-1}(\mathbf{B}^*-\mathbf{B})',\end{aligned}$$

integrating out $\hat{\mathbf{B}}$, one obtains the joint pdf of $(\mathbf{B}^*, \mathbf{S}^*, \mathbf{S})$ proportional to

$$e^{-\frac{1}{2}tr\left\{(\Sigma+\mathbf{S})^{-1}(\mathbf{B}^*-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^*-\mathbf{B})+(n-p)\mathbf{S}^{-1}\mathbf{S}^*+(n-p)\Sigma^{-1}\mathbf{S}\right\}}$$
$$\times|\mathbf{S}^*|^{\frac{n-p-m-1}{2}}\frac{|\mathbf{S}|^{-\frac{p+m+1}{2}}}{|\Sigma|^{\frac{n}{2}}}\left|\Sigma^{-1}+\mathbf{S}^{-1}\right|^{-p/2}. \quad (3.5)$$

By making the transformation $\mathbf{\Psi} = \Sigma^{-1}\mathbf{S}$, where the Jacobian is $|\Sigma|^m$, and integrating out $\mathbf{\Psi}$, the desired result is obtained. □

From (3.5), the MLE of $\mathbf{B}$ based on the synthetic data is $\mathbf{B}^*$, with

$$E(\mathbf{B}^*) = \mathbf{B}$$

which is therefore an UE of $\mathbf{B}$, with

$$Var(\mathbf{B}^*) = 2\Sigma \otimes (\mathbf{X}\mathbf{X}')^{-1}.$$

It is also possible to conclude that an UE of Σ is $\mathbf{S}^*$, since $E(\mathbf{S}^*) = \Sigma$.

After deriving the joint pdf of $\mathbf{B}^*$ and $\mathbf{S}^*$ in Theorem 3.1.1, it is now possible to propose a pivotal statistic which is a function of these estimators with the purpose of making available a procedure that may be used to draw inference for $\mathbf{B}$.

Theorem 3.1.2. *Let us consider*

$$T^* = \frac{|(\mathbf{B}^*-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^*-\mathbf{B})|}{|(n-p)\mathbf{S}^*|}. \quad (3.6)$$

Its distribution can be obtained from the decomposition

$$T^*|_{\mathbf{W}} \overset{st}{\sim} \left\{\prod_{i=1}^{m}\frac{p-i+1}{n-p-i+1}F_i\right\}\left|(n-p)\mathbf{W}^{-1}+\mathbf{I}_m\right|$$

where $F_i \sim F_{p-i+1,n-p-i+1}$ are independent random variables, themselves independent of $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$.

Proof. From (3.5), we may observe that $\mathbf{S}^*$ and $\mathbf{B}^*$, conditional on $\mathbf{S}$, are separable, with

$$\mathbf{B}^* \sim N_{pm}\left(\mathbf{B}, (\Sigma+\mathbf{S})\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$(n-p)\mathbf{S}^* \sim W_m(\mathbf{S}, n-p),$$

independent of $\mathbf{B}^*$.

Then, we have that

$$(\mathbf{B}^* - \mathbf{B})'|_{\mathbf{S}} \sim N(\mathbf{0}, (\mathbf{X}\mathbf{X}')^{-\mathbf{1}} \otimes (\Sigma + \mathbf{S})),$$

and by Theorem 2.4.1 in [22] one has that, for $p \geq m$,

$$(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|_{\mathbf{S}} \sim W_m(\Sigma + \mathbf{S}, p).$$

From Theorem 2.4.2 in [22] and subsection 7.3.3 in [3] we have

$$\mathbf{H}|_{\mathbf{S}} = (\Sigma + \mathbf{S})^{-\frac{1}{2}}(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})(\Sigma + \mathbf{S})^{-\frac{1}{2}} \sim W_m(\mathbf{I}_m, p)$$

and

$$\mathbf{G}|_{\mathbf{S}} = (n-p)\mathbf{S}^{-\frac{1}{2}}\mathbf{S}^*\mathbf{S}^{-\frac{1}{2}} \sim W_m(\mathbf{I}_m, n-p),$$

where $\mathbf{H}$ and $\mathbf{G}$ are two independent random variables, given the independence of $\mathbf{B}^*$ and $\mathbf{S}^*$.

Considering T^* defined in (3.6), one may write it as

$$T^*|_{\mathbf{S}} = \frac{|(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|}{|(n-p)\mathbf{S}^*|} = \frac{|\Sigma + \mathbf{S}|}{|\mathbf{S}|} \times \frac{|\mathbf{H}|}{|\mathbf{G}|},$$

where$|\mathbf{G}| \sim \prod_{i=1}^{m} \chi^2_{n-p-i+1}$ and $|\mathbf{H}| \sim \prod_{i=1}^{m} \chi^2_{p-i+1}$, with the chi-square random variables in each product independent, ending up with a product of independent F-distributions, given the independence of $\mathbf{H}$ and $\mathbf{G}$. The distribution of $|\mathbf{H}|/|\mathbf{G}|$ will be independent of $|\Sigma + \mathbf{S}|/|\mathbf{S}|$, due to the fact that both $\mathbf{H}$ and $\mathbf{G}$ have distributions which are not function of $\mathbf{S}$.

Thus, conditionally on $\mathbf{S}$,

$$T^*|_{\mathbf{S}} \sim \left\{ \prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_{p-i+1, n-p-i+1} \right\} \times \left| \mathbf{S}^{-\mathbf{1}}(\Sigma + \mathbf{S}) \right|.$$

Noting that we have $(n-p)\mathbf{S} \sim W_m(\Sigma, n-p)$, thus implying $\frac{1}{n-p}\mathbf{S}^{-1} \sim W_m^{-1}\left(\Sigma^{-\mathbf{1}}, n-p+m+1\right)$, one has that, for $\mathbf{W} = (n-p)\Sigma^{-1/2}\mathbf{S}\Sigma^{-1/2}$,

$$\mathbf{W}^{-1} = \frac{1}{n-p}\Sigma^{1/2}\mathbf{S}^{-1}\Sigma^{1/2} \sim W_m^{-1}(\mathbf{I}_m, n-p+m+1),$$

and that the distribution of $|\mathbf{S}^{-\mathbf{1}}(\Sigma + \mathbf{S})| = |(n-p)\mathbf{W}^{-1} + \mathbf{I}_m|$ will not depend on the parameter Σ, concluding the proof. □

One may use T^* defined in Theorem 3.1.2 and its distribution to draw inference about $\mathbf{B}$ from a single synthetic version of the original data under the Plug-in Sampling method. For instance, it can be used to perform the test of significance

of the matrix of regression coefficients. But, if instead of performing a test on the full matrix of regression coefficients, $\mathbf{B}$, one wants to test the significance of a set of regression coefficients, or more generally, of a linear combination of the parameters in $\mathbf{B}$, namely, $\mathbf{C} = \mathbf{AB}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, one may define

$$T^*_{\mathbf{C}} = \frac{|(\mathbf{AB}^* - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}')^{-1}(\mathbf{AB}^* - \mathbf{C})|}{|(n-p)\mathbf{S}^*|}$$

and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}, n-p)$ and $F_{k,i} \sim F_{k-i+1,n-p-i+1}$, all independent random variables,

$$T^*_{\mathbf{C}}|_{\mathbf{W}} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{n-p-i+1} F_{k,i} \right\} \left| (n-p)\mathbf{W}^{-1} + \mathbf{I}_m \right|. \tag{3.7}$$

Thus, in order to test $H_0 : \mathbf{C} = \mathbf{C}_0$ versus $H_1 : \mathbf{C} \neq \mathbf{C}_0$, we should reject H_0 whenever $T^*_{\mathbf{C}_0}$ exceeds $\delta_{k,m,n,p;\gamma}$, where $\delta_{k,m,n,p;\gamma}$ satisfies $(1-\gamma) = Pr(T^*_{\mathbf{C}_0} \leq \delta_{k,m,n,p;\gamma})$ when H_0 is true and where the value of $\delta_{k,m,n,p;\gamma}$ can be obtained by simulating the distribution in (3.7). In particular, a test for $\mathbf{B} = \mathbf{B}_0$ follows upon taking $\mathbf{A} = \mathbf{I}_p$, $\mathbf{C}_0 = \mathbf{B}_0$ and $k = p$ in (3.7).

A $(1-\gamma)$-level confidence set for $\mathbf{C}$ is given by

$$\Delta^*(\mathbf{C}) = \{\mathbf{C} : T^*_{\mathbf{C}} \leq \delta_{k,m,n,p;\gamma}\}. \tag{3.8}$$

In Table 3.1 we list the simulated 0.05 cut-off points for T^*_C for some values of p, m and n, for $\gamma = 0.05$, $k = p$ and $\mathbf{C} = \mathbf{B}$.

Table 3.1: Cut-off points of the 95% confidence set for the regression coefficients matrix $\mathbf{B}$.

p	n	$m=1$	$m=2$	$m=3$
3	10	4.667	8.033	8.108
	20	1.234	5.419E-01	1.083E-01
	50	3.698E-01	4.922E-02	2.849E-03
	100	1.697E-01	1.044E-02	2.749E-04
	200	8.212E-02	2.418E-03	3.040E-05
p	n	$m=1$	$m=2$	$m=3$
4	10	7.693	29.22	106.1
	20	1.652	1.165	5.356E-01
	50	4.621E-01	9.248E-02	1.115E-01
	100	2.089E-01	1.903E-02	1.034E-02
	200	9.997E-02	4.339E-03	1.113E-03

Remark 3.1.1. *When $m = 1$, T^* in (3.6) reduces to the statistic T^2 used in [20] which has a pdf obtained from the fact that*

$$T^2|_{W=w} \sim \frac{p}{n-p}\left(1+\frac{n-p}{w}\right)F_{p,n-p} \quad where \quad f_W(w) = \frac{1}{2^{\frac{n-p}{2}}\Gamma\left(\frac{n-p}{2}\right)}e^{-\frac{w}{2}}w^{\frac{n-p}{2}-1}.$$

This way, with the development of an exact inferential procedure for the matrix of the regression coefficients for the single imputation case, one fulfills another existing gap in the literature, in this case, for partially synthetic datasets generated via Plug-in sampling.

Regarding the multiple imputation case, since the original data estimators are directly plugged in to the generating distribution, one does not have to develop an exact inferential procedure similar to the one developed for the PPS case. It is possible to derive exact inferential procedures somewhat similar to the ones developed for the new method of generating synthetic datasets, the FPPS method, as it is shown in subsections 3.2.2 and 3.2.3. These procedures will not have the constraints presented by the PPS method in subsection 2.2.2, which are the necessity of dealing with the distribution of the sum of Wishart distributions with different parameters and the consequent fact of not being able to use the proposed *radius* measure to evaluate the extent of the confidence sets.

3.2 Multiple imputation: Plug-in Sampling

In this section, an adaptation of Reiter's combination rule [31] to be used when dealing with multiple synthetic datasets generated via Plug-in Sampling is presented and two new exact likelihood-based procedures for the analysis of these same synthetic datasets are developed.

For this purpose, let us recall again the MLR model (1.1). To generate synthetic versions of $\mathbf{Y}$ based on the Plug-in Sampling method one takes the original data $(y_{i1},\ldots,y_{im},x_{1i},\ldots,x_{pi})$, $i = 1,\ldots,n$, and after estimating $\mathbf{B}$ and Σ by $\hat{\mathbf{B}}$ and $\mathbf{S}$, respectively, generates the M synthetic datasets, denoted as $\mathbf{V}_j = (\mathbf{v}_{1j},\ldots,\mathbf{v}_{nj})$, $j = 1,\ldots,M$ where $\mathbf{v}_{ij} = (v_{1ij},\ldots,v_{mij})'$, are independently distributed as

$$\mathbf{v}_{ij}|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_m(\hat{\mathbf{B}}'\mathbf{x}_{ij},\mathbf{S}), i = 1,\ldots,n, j = 1,\ldots,M, \tag{3.9}$$

that is, by plugging in the estimators of the original data into the model in order to draw synthetic data from it.

3.2.1 Reiter's adapted Methodology

In this subsection, an adaptation of Reiter [31] methodology to a matrix value parameter is formulated for multiply imputed synthetic data generated via Plug-in Sampling. As it was referred in subsection 2.2.1, Reiter's methodology was originally developed for the analysis of synthetic datasets generated via Posterior Predictive Sampling, but Reiter and Kinney [33], in 2012, argued that it is also valid when synthetic datasets are generated via the Plug-in Sampling method.

Therefore, having access to M synthetic data sets $\mathbf{V}_1,\ldots,\mathbf{V}_M$ as the synthetic data sets generated via Plug-in Sampling , let us define $\mathbf{vec}(\mathbf{B}_j^*) = \mathbf{vec}((\mathbf{XX}')^{-1}\mathbf{XV}_j')$ and $\mathbf{U}_j = \mathbf{S}_j^* \otimes (\mathbf{XX}')^{-1}$, where $\mathbf{S}_j^* = \frac{1}{n-p}(\mathbf{V}_j - \mathbf{B}_j^{*'}\mathbf{X})(\mathbf{V}_j - \mathbf{B}_j^{*'}\mathbf{X})'$, for $j = 1,\ldots,M$. Based on $\mathbf{V}_j$, given $\hat{\mathbf{B}}$ and $\mathbf{S}$, $\mathbf{vec}(\mathbf{B}_j^*)$ will be an UE of $\mathbf{vec}(\mathbf{B})$ and $\mathbf{U}_j$ will be an UE of its variance. Then if we consider the estimators in (2.21) and (2.22), upon replacing $\mathbf{B}_j^\dagger$ by $\mathbf{B}_j^\bullet$, $T_{R,M}$ given by (2.23) will yet be approximated by an $F_{pm,w(r)}$ distribution, with $w(r)$ defined in (2.24).

However, we should recall that this methodology is based on an asymptotic combination rule, and therefore it is a methodology which is inadequate for cases where the sample size of the synthesized datasets is small. This leads to the need of the development of exact inference procedures for the analysis of multiple synthetic datasets generated under Plug-in Sampling method which are developed in the next sections. We will then use this Reiter's adapted methodology with the purpose of comparing results.

3.2.2 A First New Procedure

Let us start by developing the first exact inference procedure for the analysis of multiple synthetic datasets generated under Plug-in Sampling method where the estimators of $\mathbf{B}$ and Σ will be the mean of the estimators for each synthetic dataset, as it was similarly done in subsection 2.3.1.

Let

$$\mathbf{B}_\mathbf{j}^* = (\mathbf{XX}')^{-1}\mathbf{XV}_\mathbf{j}'$$

and

$$\mathbf{S}_\mathbf{j}^* = \frac{1}{n-p}(\mathbf{V}_\mathbf{j} - \mathbf{B}_\mathbf{j}'^*\mathbf{X})(\mathbf{V}_\mathbf{j} - \mathbf{B}_\mathbf{j}'^*\mathbf{X})'$$

be the estimators of $\mathbf{B}$ and Σ based on $\mathbf{V}_j$, for $j = 1,\ldots,M$. By Lemma 1.4.1, $\mathbf{B}_\mathbf{j}^*$ and $\mathbf{S}_\mathbf{j}^*$ will be jointly sufficient for $\mathbf{B}$ and Σ.

Conditionally on $(\hat{\mathbf{B}},\mathbf{S})$, for each $j = 1,\ldots,M$, $\mathbf{B}_\mathbf{j}^*$ is independent of $\mathbf{S}_\mathbf{j}^*$ and $\left\{(\mathbf{B}_1^*,\mathbf{S}_1^*),\ldots,(\mathbf{B}_M^*,\mathbf{S}_M^*)\right\}$ are jointly sufficient estimators for $\mathbf{B}$ and Σ.

Let us also define

$$\overline{\mathbf{B}}^*_M = \frac{1}{M}\sum_{j=1}^{M}\mathbf{B}^*_{\mathbf{j}} \quad \text{and} \quad \overline{\mathbf{S}}^*_M = \frac{1}{M}\sum_{i=1}^{M}\mathbf{S}^*_{\mathbf{j}}, \tag{3.10}$$

which are mutually independent, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$.

For $p \geq m$, let us consider the two following two Corollaries of Theorems 3.1.1 and 3.1.2 which will be important to draw inference about $\mathbf{B}$ from a pivotal statistic.

Corollary 3.2.1. *The joint pdf of $\overline{\mathbf{B}}^*_M$ and $\overline{\mathbf{S}}^*_M$ defined in (3.10) is proportional to*

$$\int e^{-\frac{1}{2}tr\left\{(\boldsymbol{\Sigma}(\mathbf{I}_m+\frac{1}{M}\boldsymbol{\Psi}))^{-1}(\overline{\mathbf{B}}^*_M-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M-\mathbf{B})+M(n-p)\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}^{-1}\overline{\mathbf{S}}^*_M+(n-p)\boldsymbol{\Psi}\right\}}$$
$$\times|\overline{\mathbf{S}}^*_M|^{\frac{M(n-p)-m-1}{2}}\frac{|\boldsymbol{\Psi}|^{-\frac{M(n-p)-n+2p+m+1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{M(n-p)+p-m+1}{2}}}|\mathbf{I}+M\boldsymbol{\Psi}^{-1}|^{-p/2}d\boldsymbol{\Psi}.$$

Proof. The proof is identical to the proof of Theorem 3.1.1, replacing the joint pdf of $(\mathbf{B}^*,\mathbf{S}^*)$ by the joint pdf of $(\overline{\mathbf{B}}^*_M,\overline{\mathbf{S}}^*_M)$ and observing that

$$\overline{\mathbf{B}}^*_M|_{\hat{\mathbf{B}},\mathbf{S}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{B}^*_j|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_{pm}(\hat{\mathbf{B}},\frac{1}{M}\mathbf{S}\otimes(\mathbf{X}\mathbf{X}')^{-1}),$$

and

$$M(n-p)\overline{\mathbf{S}}^*_M|_{\mathbf{S}} = (n-p)\sum_{j=1}^{M}\mathbf{S}^*_j|_{\mathbf{S}} \sim W_m(\mathbf{S},M(n-p)),$$

are independent. □

The following Corollary makes available pivotal statistic that may be used to make inference about $\mathbf{B}$ based on multiply imputed synthetic datasets generated via Plug-in Sampling.

Corollary 3.2.2. *The pdf of T^*_M defined as*

$$T^*_M = \frac{\left|(\overline{\mathbf{B}}^*_M-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M-\mathbf{B})\right|}{\left|(n-p)\overline{\mathbf{S}}^*_M\right|} \tag{3.11}$$

can be obtained from the decomposition

$$T_M|_{\mathbf{W}} \overset{st}{\sim} \left\{\prod_{i=1}^{m}\frac{p-i+1}{M(n-p)-i+1}F_i\right\}\left|M(n-p)\mathbf{W}^{-1}+\mathbf{I}_m\right|$$

where $\mathbf{W} \sim W_m(I,n-p)$ and $F_i \sim F_{p-i+1,M(n-p)-i+1}$, all independent random variables.

Proof. The proof is identical to the proof of Theorem 3.1.2, replacing $\mathbf{B}^*$ and $\mathbf{S}^*$ by $\overline{\mathbf{B}}^*_M$ and $\overline{\mathbf{S}}^*_M$, and noting that from the distribution in Corollary 3.2.1 one has that

$$\overline{\mathbf{B}}^*_M|_{\mathbf{S}} \sim N_{pm}(\mathbf{B}, (\Sigma + \frac{1}{M}\mathbf{S}) \otimes (\mathbf{XX}')^{-\mathbf{1}})$$

and

$$M(n-p)\overline{\mathbf{S}}^*_M|_{\mathbf{S}} \sim W_m(\mathbf{S}, M(n-p)),$$

independent of $\overline{\mathbf{B}}^*_M$. □

From the above Corollaries and proceeding similarly as in Section 3.1 one may conclude, for $p \geq m$, that $\overline{\mathbf{B}}^*_M$ is the unbiased MLE of $\mathbf{B}$, with

$$Var(\overline{\mathbf{B}}^*_M) = \frac{M+1}{M} \times \Sigma \otimes (\mathbf{XX}')^{-1},$$

and that $\overline{\mathbf{S}}^*_M$ is an UE of Σ.

To test the significance of a set of regression coefficients or more generally of a linear combination of these regression coefficients, $\mathbf{C} = \mathbf{AB}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, we define

$$T^*_{M,\mathbf{C}} = \frac{|(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})|}{|(n-p)\overline{\mathbf{S}}^*_M|}$$

and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{k,i} \sim F_{k-i+1,M(n-p)-i+1} (i = 1, \ldots, m)$, all independent random variables,

$$T^*_{M,\mathbf{C}}|_{\mathbf{W}} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{M(n-p)-i+1} F_{k,i} \right\} |M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m|. \tag{3.12}$$

In order to test

$$H_0 : \mathbf{C} = \mathbf{C}_0 \quad \text{versus} \quad H_1 : \mathbf{C} \neq \mathbf{C}_0,$$

we should reject H_0 whenever $T^*_{M,\mathbf{C}_0}$ exceeds $\delta_{M,k,m,n,p;\gamma}$, where $\delta_{M,k,m,n,p;\gamma}$ satisfies $(1-\gamma) = Pr(T^*_{M,\mathbf{C}_0} \leq \delta_{M,k,m,n,p;\gamma})$ when H_0 is true and where the value of $\delta_{M,k,m,n,p;\gamma}$ can be obtained by simulating the distribution of $T^*_{\mathbf{C}}$, by first generating $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and then generating the distribution in (3.12).

A $(1-\gamma)$-level confidence set for $\mathbf{C}$ is given by

$$\Delta^*_M(\mathbf{C}) = \{\mathbf{C} : T^*_{M,\mathbf{C}} \leq \delta_{M,k,m,n,p;\gamma}\}. \tag{3.13}$$

3.2.3 A Second New Procedure

As it was done in subsection 2.3.2, we propose in this subsection, a second likelihood-based approach for exact inference about $\mathbf{B}$, by treating the M synthetic datasets as a big synthetic data sample of size nM.

Let us consider the M synthetic datasets as a single sample of size nM arranged as

$$\begin{pmatrix} \mathbf{V}_a \\ \mathbf{X}_a \end{pmatrix} = \left(\begin{array}{c|c|c|c} \mathbf{V}_1 & \mathbf{V}_2 & \dots & \mathbf{V}_M \\ \hline \mathbf{X} & \mathbf{X} & \dots & \mathbf{X} \end{array}\right),$$

where $\mathbf{V}_a = (\mathbf{V}_1|\dots|\mathbf{V}_M)$ and $\mathbf{X} = (\mathbf{X}|\dots|\mathbf{X})$. Let us also consider

$$\mathbf{B}_a^* = (\mathbf{X}_a\mathbf{X}_a')^{-1}\mathbf{X}_a\mathbf{V}_a'$$

and

$$\mathbf{S}_a^* = \frac{1}{nM-p}(\mathbf{V}_a - \mathbf{B}_a'\mathbf{X}_a)(\mathbf{V}_a - \mathbf{B}_a'\mathbf{X}_a)'$$

as the estimators of $\mathbf{B}$ and Σ, respectively. Using a procedure similar to the one employed in subsection 2.3.2, one can conclude that $\mathbf{B}_a^*$ will be exactly the same estimator as $\overline{\mathbf{B}}_M^*$, defined in (3.10), and $\mathbf{S}_a^*$ will be exactly the same as the estimator

$$\mathbf{S}_{comb}^* = \frac{\mathbf{S}_{\mathbf{v}} + M \times \mathbf{S}_{mean}}{Mn-p}, \tag{3.14}$$

where

$$\mathbf{S}_{\mathbf{v}} = \sum_{i=1}^{n}\sum_{j=1}^{M}(\mathbf{v}_{ji} - \bar{\mathbf{v}}_i)(\mathbf{v}_{ji} - \bar{\mathbf{v}}_i)',$$

with $\bar{\mathbf{v}}_i = \frac{1}{M}\sum_{j=1}^{M}\mathbf{v}_{ji}$ and $\mathbf{v}_{ji}$ as the column vectors of $\mathbf{V}_j\,(j = 1,\dots,M)$, and

$$\mathbf{S}_{mean} = \left(\overline{\mathbf{V}}_M - \overline{\mathbf{B}}_M^{*'}\mathbf{X}\right)'\left(\overline{\mathbf{V}}_M - \overline{\mathbf{B}}_M^{*'}\mathbf{X}\right),$$

with $\overline{\mathbf{V}}_M = \frac{1}{M}\sum_{i=1}^{M}\mathbf{V}_j$, for $j = 1,\dots,M$.

With the estimator $\mathbf{S}_a^* = \mathbf{S}_{comb}^*$ we end up using more information about Σ than when $\mathbf{S}_M^*$ defined in (3.10) is used. In fact, we may observe that the estimator $\mathbf{S}_{comb}^*$, which is the same as $\mathbf{S}_a^*$, is a combination of two estimators of Σ, $\mathbf{S}_v$ and $\mathbf{S}_{mean}$. To recall that we have indeed a combination of estimators, we will proceed the development of this second procedure using from now on the notation $\mathbf{S}_{comb}^*$.

Next, we present a Corollary of Theorem 3.1.1 where it is derived the joint pdf of $\overline{\mathbf{B}}_M^*$ and $\mathbf{S}_{comb}^*$, estimators of $\mathbf{B}$ and Σ, respectively. These estimators will then be used to define a pivotal statistic which allows us to draw inference about $\mathbf{B}$ based on multiply imputed synthetic datasets generated via Plug-in Sampling.

Corollary 3.2.3. *For $p \geq m$, the joint pdf of $\overline{\mathbf{B}}^*_M$ and $\mathbf{S}^*_{comb}$ defined in (3.10) and (3.14) is proportional to*

$$\int e^{-\frac{1}{2}tr\left\{(\Sigma(\mathbf{I}+\frac{1}{M}\mathbf{\Psi}))^{-1}(\overline{\mathbf{B}}^*_M-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M-\mathbf{B})+(Mn-p)\mathbf{\Psi}^{-1}\Sigma^{-1}\mathbf{S}^*_{comb}+(n-p)\mathbf{\Psi}\right\}}$$
$$\times|\mathbf{S}^*_{comb}|^{\frac{Mn-p-m-1}{2}}\frac{|\mathbf{\Psi}|^{-\frac{Mn-p-n+2p+m+1}{2}}}{|\Sigma|^{\frac{Mn-p+p-m+1}{2}}}|\mathbf{I}+M\mathbf{\Psi}^{-1}|^{-p/2}d\mathbf{\Psi}.$$

Proof. The proof is identical to the proof of Theorem 3.1.1 replacing the joint pdf of $(\mathbf{B}^*,\mathbf{S}^*)$ by the joint pdf of $(\overline{\mathbf{B}}^*_M,\mathbf{S}^*_{comb})$, respectively, and observing that

$$\overline{\mathbf{B}}^*_M|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_{pm}(\hat{\mathbf{B}},\frac{1}{M}\mathbf{S}\otimes(\mathbf{X}\mathbf{X}')^{-1})$$

and

$$(Mn-p)\mathbf{S}^*_{comb}|_{\mathbf{S}} \sim W_m(\mathbf{S},Mn-p)$$

independent of $\overline{\mathbf{B}}^*_M$. □

Now that the joint pdf of $\overline{\mathbf{B}}^*_M$ and $\mathbf{S}^*_{comb}$ has been given, in the following Corollary of Theorem 3.1.2, we make available a pivotal statistic along with its distribution allowing us to draw inference about $\mathbf{B}$ when the M Plug-in generated synthetic datasets are available.

Corollary 3.2.4. *Let us consider*

$$T^*_{comb} = \frac{\left|(\overline{\mathbf{B}}^*_M-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M-\mathbf{B})\right|}{\left|(n-\frac{p}{M})\mathbf{S}^*_{comb}\right|}. \tag{3.15}$$

For $p \geq m$, its distribution can be obtained from the decomposition

$$T_{comb}|_{\mathbf{W}} \overset{st}{\sim} \left\{\prod_{i=1}^{m}\frac{p-i+1}{Mn-p-i+1}F_i\right\}\left|M(n-p)\mathbf{W}^{-1}+\mathbf{I}_m\right|$$

where $\mathbf{W} \sim W_m(I,n-p)$ and $F_i \sim F_{p-i+1,Mn-p-i+1}$, all independent random variables.

Proof. The proof is identical to the proof of Theorem 3.1.2 replacing $\mathbf{B}^*$ and $\mathbf{S}^*$ by $\overline{\mathbf{B}}^*_M$ and $\mathbf{S}^*_{comb}$, respectively, noting that from the distribution in Corollary 3.2.1,

$$\overline{\mathbf{B}}^*_M|_{\mathbf{S}} \sim N_{pm}\left(\mathbf{B},\left(\Sigma+\frac{1}{M}\mathbf{S}\right)\otimes(\mathbf{X}\mathbf{X}')^{-1}\right)$$

and

$$(Mn-p)\mathbf{S}^*_{comb}|_{\mathbf{S}}\, W_m(\mathbf{S},Mn-p),$$

independent of $\overline{\mathbf{B}}^*_M$. □

From the two Corollaries above, one observes that an UE of Σ will be $\mathbf{S}^*_{comb}$.

If one wants to test the significance of a set of regression coefficients or more generally, a linear combination of the parameters in $\mathbf{B}$, namely, $\mathbf{C} = \mathbf{AB}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, we define

$$T^*_{comb,\mathbf{C}} = \frac{|(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})|}{|\left(n - \frac{p}{M}\right)\mathbf{S}^*_{comb}|}$$

and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{k,i} \sim F_{k-i+1,Mn-p-i+1}$, all independent variables,

$$T^*_{comb,\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{\prod_{i=1}^{m} \frac{k-i+1}{Mn-p-i+1} F_{k,i}\right\} \left|M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m\right|. \tag{3.16}$$

In order to test

$$H_0 : \mathbf{C} = \mathbf{C}_0 \text{ versus } H_1 : \mathbf{C} \neq \mathbf{C}_0,$$

we reject H_0 whenever $T^*_{comb,\mathbf{C}_0}$ exceeds $\delta_{comb,M,k,m,n,p;\gamma}$ where $\delta_{comb,M,k,m,n,p;\gamma}$ satisfies $(1-\gamma) = Pr(T^*_{comb,\mathbf{C}_0} \leq \delta_{comb,M,k,m,n,p;\gamma})$ when H_0 is true, where the value of $\delta_{comb,M,k,m,n,p;\gamma}$ may be obtained by simulating the distribution in (3.16).

A $(1-\gamma)$ level confidence set for $\mathbf{C}$ is given by

$$\Delta^*_{comb}(\mathbf{C}) = \{\mathbf{C} : T^*_{comb,\mathbf{C}} \leq \delta_{comb,M,k,m,n,p;\gamma}\}. \tag{3.17}$$

As in the FPPS case in subsection 2.3, the second exact procedure developed in this Chapter is expected to offer better precision, originating confidence sets which will have smaller *radius* than the ones obtained from the first procedure. It is also expected that the two procedures will come closer together for larger values of n and M.

In fact, making a simple change of scale on T^*_{comb}, defined in (3.15), the distributions of the statistics proposed in the two Plug-in procedures developed in this Chapter, will converge in distribution to the same distribution. For T^*_M given by (3.11) in the first procedure, we have

$$T^*_M|\mathbf{W} \xrightarrow[M\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} \left|\mathbf{W}^{-1}\right|$$

and

$$T^*_M|\mathbf{W} \xrightarrow[n\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} \left|\mathbf{W}^{-1} + \frac{1}{M(n-p)}\mathbf{I}_m\right|,$$

and for a scale change of T^*_{comb} given by (3.15) in the second procedure, we have

$$\frac{\left(n-\frac{p}{M}\right)^m}{(n-p)^m} T^*_{comb}|\mathbf{w} \xrightarrow[M\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} \left|\mathbf{W}^{-1}\right|$$

and

$$\frac{\left(n-\frac{p}{M}\right)^m}{(n-p)^m} T^*_{comb}|\mathbf{w} \xrightarrow[n\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} \left|\mathbf{W}^{-1} + \frac{1}{M(n-p)}\mathbf{I}_m\right|.$$

As such, for large values of M and large values of n both procedures become identical.

Chapter 4

Radius and Accuracy

The two previous chapters presented the development of several procedures that enable an analyst to draw inference about the regression coefficients matrix when he or she has access to synthetic datasets generated by replicating the original data via PPS, FPPS or Plug-in Sampling methods. In this Chapter, simulation studies are undertaken to show that the inference methods developed in this work perform as predicted, to compare the accuracy of these with the accuracy of Reiter's adapted procedure, for the multiple imputation case, and to measure the extent of the confidence sets obtained from all exact procedures developed. With this last objective in mind, before presenting the simulation studies it is important to define a measure that will evaluate the extent of the referred confidence sets.

Most of the content associated to the FPPS method in Sections 4.1 and 4.2 is taken from [25].

4.1 Measuring the confidence sets

In order to evaluate the 'size' of the confidence sets defined in Chapters 2 and 3 it is usual to calculate its volume. Unfortunately, in our case this volume is infinite as shown in the next subsection. Therefore, a measure that will be called *radius*, which measures the distance between the center and the edge of the confidence sets, is proposed and used in the simulation studies to illustrate the differences between methods.

4.1.1 Volume of the confidence sets

In this subsection, it will be proved that the volume cannot be used to measure the confidence sets for the Plug-in Sampling case under single imputation using, without any loss of generality, (3.13), with $\mathbf{A} = \mathbf{I}_p$. It is easy to observe that similar proofs for all of the other cases can be made.

Let us recall that $p \geq m$, $\mathbf{X}$ is a $p \times n$ matrix with rank equal to $p < n$, $\mathbf{B}$ is a $p \times m$ matrix, and $\mathbf{S}$ is a $m \times m$ matrix. The confidence set (3.13) in the referred case will be

$$\Delta(\mathbf{B}) = \left\{ \mathbf{B} : \frac{|(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|}{|(n-p)\mathbf{S}^*|} \leq \delta_{M,m,n,p;\gamma} \right\}. \tag{4.1}$$

The volume of this confidence set will be given by

$$\int \cdots \int_{\Delta(\mathbf{B})} (d\mathbf{B}) = \int \cdots \int_{\Delta(\tilde{\mathbf{B}})} (d\tilde{\mathbf{B}})$$

where $\tilde{\mathbf{B}} = (\mathbf{B} - \mathbf{B}^*)$, with $J(\mathbf{B} \to \tilde{\mathbf{B}}) = 1$, and

$$\Delta(\tilde{\mathbf{B}}) = \left\{ \tilde{\mathbf{B}} : |\tilde{\mathbf{B}}'(\mathbf{X}\mathbf{X}')\tilde{\mathbf{B}}| \leq d_{m,n,p;\gamma} \times |(n-p)\mathbf{S}^*| \right\}.$$

Let us consider the transformation $\tilde{\tilde{\mathbf{B}}} = (\mathbf{X}\mathbf{X}')^{1/2}\tilde{\mathbf{B}}$. By Theorem 2.1.4 in [26] we have that $J(\tilde{\tilde{\mathbf{B}}} \to \tilde{\mathbf{B}}) = |\mathbf{X}\mathbf{X}'|^p$ thus implying that $J(\tilde{\mathbf{B}} \to \tilde{\tilde{\mathbf{B}}}) = |\mathbf{X}\mathbf{X}'|^{-p}$. Therefore, the volume of the confidence set $\Delta(\mathbf{B})$ will be given by

$$|\mathbf{X}\mathbf{X}'|^{-p} \int \cdots \int_{\Delta(\tilde{\tilde{\mathbf{B}}})} (d\tilde{\tilde{\mathbf{B}}})$$

where

$$\Delta(\tilde{\tilde{\mathbf{B}}}) = \left\{ \tilde{\tilde{\mathbf{B}}} : \left\| \tilde{\tilde{\mathbf{B}}}' \tilde{\tilde{\mathbf{B}}} \right\| \leq d_{m,n,p;\gamma} \times |(n-p)\mathbf{S}^*| \right\}.$$

Since $\tilde{\tilde{\mathbf{B}}}' \tilde{\tilde{\mathbf{B}}}$ is a positive definite symmetric square matrix it can be represented as $\mathbf{T}'\mathbf{T}$ where $\mathbf{T} = (t_{ij})$ is an upper-triangular matrix $m \times m$, with positive diagonal elements.

Let us take $\tilde{\tilde{\mathbf{B}}} = \mathbf{H}_1\mathbf{T}$, where $\mathbf{H}_1$ is a $p \times m$ matrix with $\mathbf{H}_1'\mathbf{H}_1 = \mathbf{I}_m$ and $\mathbf{T}$ is a $m \times m$ upper-triangular matrix with positive diagonal elements.

Let $\mathbf{H}_2$ (a function of $\mathbf{H}_1$) be a $p \times (p-m)$ matrix such that $\mathbf{H} = [\mathbf{H}_1 : \mathbf{H}_2]$ is an orthogonal $p \times p$ matrix and let us write $\mathbf{H} = [\mathbf{h}_1 ... \mathbf{h}_m : \mathbf{h}_{m+1} ... \mathbf{h}_p]$, where $\mathbf{h}_1, ..., \mathbf{h}_m$ are the columns of $\mathbf{H}_1$ and $\mathbf{h}_{m+1}, ..., \mathbf{h}_p$ are the columns of $\mathbf{H}_2$.

All this decomposition of $\tilde{\tilde{\mathbf{B}}}' \tilde{\tilde{\mathbf{B}}}$ allows the use of Theorem 2.1.13 in [26] making the volume equal to

$$|\mathbf{X}\mathbf{X}'|^{-p} \int \cdots \int_{\Delta(\mathbf{T})} \int_{V_{m,p}} \prod_{i=1}^{m} t_{ii}^{p-i} (d\mathbf{T})(\mathbf{H}_1' d\mathbf{H}_1)$$

with

$$(\mathbf{H}_1' d\mathbf{H}_1) \equiv \bigwedge_{i=1}^{m} \bigwedge_{j=i+1}^{p} \mathbf{h}_j' d\mathbf{h}_i$$

and $(d\mathbf{T}) = \bigwedge_{i\leq j}^{m} dt_{ij}$ where

$$\Delta(\mathbf{T}) = \left\{\mathbf{T} : |\mathbf{T}'\mathbf{T}| \leq d_{m,n,p;\gamma} \times |(n-p)\mathbf{S}^*|\right\}$$

and $V_{m,p} = \{\mathbf{H}_1 : \mathbf{H}_1'\mathbf{H}_1 = \mathbf{I}_m\}$, with $\bigwedge$ denoting the exterior or wedge product.

From Theorem 2.1.15 [26] the volume will be given by

$$|\mathbf{X}\mathbf{X}'|^{-p} \frac{2^m \pi^{mn/2}}{\Gamma_m(\frac{1}{2}n)} \int \cdots \int_{\Delta(\mathbf{T})} \prod_{i=1}^{m} t_{ii}^{p-i} (d\mathbf{T}).$$

By definition, $\mathbf{T}$ is an upper triangular matrix therefore allowing the conclusion that $|\mathbf{T}'\mathbf{T}|$ will be the product of its diagonal elements, that is, $|\mathbf{T}'\mathbf{T}| = \prod_{i=1}^{m} t_{ii}^2 \leq d_{m,n,p;\gamma} \times |(n-p)\mathbf{S}^*|$.

Let us define $C^2 = d_{m,n,p;\gamma} \times |(n-p)S^*|$, making the domain of integration to be given by

$$\Delta(\mathbf{T}) = \left\{\mathbf{T} : |\mathbf{T}'\mathbf{T}| \leq C^2\right\}.$$

Observing the integral that remains unsolved

$$I = \int \cdots \int_{\Delta(\mathbf{T})} \prod_{i=1}^{m} t_{ii}^{p-i} (d\mathbf{T}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int \cdots \int_{\Delta(\mathbf{T})} \prod_{i=1}^{m} t_{ii}^{p-i} \left(\bigwedge_{i=1}^{m} dt_{ii} \right) \right] \left(\bigwedge_{i<j}^{m} dt_{ij} \right),$$

it is possible to obtain the value of the integral containing the diagonal elements since it is separable from the off-diagonal elements in I.

Let us define

$$I_1 = \int \cdots \int_{\Delta(\mathbf{T})} \prod_{i=1}^{m} t_{ii}^{p-i} \left(\bigwedge_{i=1}^{m} dt_{ii} \right).$$

From $\prod_{i=1}^{m} t_{ii}^2 \leq C^2$, we have that for any $j = 1, ..., m$,

$$t_{jj} \leq \frac{C}{\prod_{i\neq j}^{m} t_{ii}}.$$

Leaving out t_{mm} and dt_{mm}, and integrating sequentially in order to $t_{m-1m-1}, ..., t_{11}$, one can easily verify that

$$I_1 = \frac{C^p}{p-m+1} \prod_{i=1}^{m} t_{ii}^{1-i},$$

thus having that

$$I = \frac{C^p}{p-m+1} \prod_{i=1}^{m} t_{ii}^{1-i} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\bigwedge_{i<j}^{m} dt_{ij}).$$

But, since $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\bigwedge_{i<j}^{m} dt_{ij})$ is infinite, then I will also be infinite, as we wanted to prove.

4.1.2 *Radius* of the confidence sets

As such, it is not possible to use the volume to determine the extent of each confidence set defined in Chapters 2 and 3, but anyway it would be very important to have a measure that would allow us to compare the precision of the inference procedures developed for the analysis of PPS, FPPS and Plug-in synthetic datasets by measurig the extent of the corresponding confidence sets.

Observing that the confidence set defined in (2.34) for the first FPPS procedure can be written as

$$\Delta_M^{\bullet}(\mathbf{C}) = \left\{ \mathbf{C} : |(\mathbf{A}\overline{\mathbf{B}}_M^{\bullet} - \mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}_M^{\bullet} - \mathbf{C})| \leq \omega_{M,k,m,n,p;\gamma} |M(n-p)\overline{\mathbf{S}}_M^{\bullet}| \right\},$$

it will be possible to use $\omega_{M,k,m,n,p;\gamma}|M(n-p)\overline{\mathbf{S}}_M^{\bullet}|$ as a boundary for the confidence sets. Let us also observe that we may also rewrite the confidence sets of the FPSS second procedure and of both Plug-in Sampling procedures in a similar way, but that the same cannot be done for the PPS confidence set. With this fact in mind, it will be proposed a measure that evaluates the distance between the center and the edge of a confidence set which we will call *radius*. This *radius* will then be defined for the FPPS and Plug-in Sampling methods, while for the PPS method, where this *radius* cannot be used, we will propose two measures which will be an upper and a lower bound for the real distance between the center and the edge.

We propose

$$\Upsilon_M^{\bullet} = d_{M,m,n,p,\alpha,\gamma}^{\bullet} \times |\tilde{\mathbf{S}}_M^{\bullet}| \tag{4.2}$$

as the radius of the confidence sets when using synthetic data generated under FPPS method, and

$$\Upsilon_M^{*} = d_{M,m,n,p,\gamma}^{*} \times |\tilde{\mathbf{S}}_M^{*}| \tag{4.3}$$

as the radius of the confidence sets when using synthetic data generated under Plug-in Sampling.

For $M = 0$ the two measures proposed in (4.2) and (4.3) will be equal and will refer to the original data where $\tilde{\mathbf{S}}_0^{\bullet} = \tilde{\mathbf{S}}_0^{*} = (n-p)\mathbf{S}$ and $d_{0,m,n,p,\alpha,\gamma}^{\bullet} = d_{0,m,n,p,\alpha,\gamma}^{*}$ will be the γ cut-off point for the original data.

For $M > 0$, $d^{\bullet}_{M,m,n,p,\alpha,\gamma}$ will be the cut-off point in (2.34) or (2.42) and $d^{*}_{M,m,n,p,\gamma}$ will be the cut-off point in (3.13) or (3.17), with $\tilde{\mathbf{S}}^{\bullet}_{M} = M(n-p)\overline{\mathbf{S}}^{\bullet}_{M}$ for the first FPPS new procedure, $\tilde{\mathbf{S}}^{\bullet}_{M} = (Mn-p)\mathbf{S}^{\bullet}_{comb}$ for the second FPPS new procedure, $\tilde{\mathbf{S}}^{*}_{M} = (n-p)\overline{\mathbf{S}}^{*}_{M}$ for the first Plug-in new procedure and $\tilde{\mathbf{S}}^{*}_{M} = (n-\frac{p}{M})\mathbf{S}^{*}_{comb}$ for the second Plug-in new procedure, recalling that for $M = 1$ the two procedures coincide in both FPPS and Plug-in methods.

We may observe that this *radius* is function of the matrix of variances computed from the synthetic data thus it would also be important to derivate the expectation of these *radius*.

Let us start with the FPPS method. Recalling that for the original data, $(n-p)\mathbf{S} \sim W_m(\Sigma, n-p)$, thus implying that

$$E(|(n-p)\mathbf{S}|) = |\Sigma| E\left(\prod_{i=1}^{m} \chi^2_{n-p-i+1}\right) = \frac{(n-p)!}{(n-p-m)!}|\Sigma|, \tag{4.4}$$

where $\prod_{i=1}^{m} \chi^2_{n-p-i+1}$ will be a product of independent chi-square variables, since $(n-p)\Sigma^{-1/2}\mathbf{S}\Sigma^{-1/2} \sim W_m(\mathbf{I}_m, n-p)$, and recalling that

$$\tilde{\Sigma}|_{\mathbf{S}} \sim W_m^{-1}((n-p)\mathbf{S}, n+\alpha-p) \implies \tilde{\Sigma}^{-1}|_{\mathbf{S}} \sim W_m\left(\frac{1}{n-p}\mathbf{S}^{-1}, n+\alpha-p-m-1\right)$$

therefore, taking $\kappa_{n,\alpha,p,m} = n+\alpha-p-m-1$, given $\mathbf{S}$, one has that

$$E(|\tilde{\Sigma}|) = E(|\tilde{\Sigma}^{-1}|^{-1}) = |(n-p)\mathbf{S}| E\left(\frac{1}{\prod_{i=1}^{m} \chi^2_{\kappa_{n,\alpha,p,m}-i+1}}\right) = |(n-p)\mathbf{S}| \frac{(-2+\kappa_{n,\alpha,p,m}-m)!}{(-2+\kappa_{n,\alpha,p,m})!}, \tag{4.5}$$

where $\prod_{i=1}^{m} \chi^2_{\kappa_{n,\alpha,p,m}-i+1}$ is a product of independent chi-square variables. Also let us recall that, given $\tilde{\Sigma}$, $M(n-p)\overline{\mathbf{S}}^{\bullet}_{M} \sim W_m(\tilde{\Sigma}, M(n-p))$ and $(Mn-p)\mathbf{S}^{\bullet}_{comb} \sim W_m(\tilde{\Sigma}, Mn-p)$, thus concluding that, given $\tilde{\Sigma}$,

$$E(|M(n-p)\overline{\mathbf{S}}^{\bullet}_{M}|) = \frac{(Mn-Mp)!}{(Mn-Mp-m)!} \times |\tilde{\Sigma}|$$

and

$$E(|(Mn-p)\mathbf{S}^{\bullet}_{comb}|) = \frac{(Mn-p)!}{(Mn-p-m)!} \times |\tilde{\Sigma}|.$$

Combining the results for $E(|(n-p)\mathbf{S}|)$ in (4.4) and $E(|\tilde{\Sigma}||_{\mathbf{S}})$ in (4.5), respectively with the corresponding expected values of $|M(n-p)\mathbf{S}^{\bullet}_{M}|$ and $|(Mn-p)\mathbf{S}^{\bullet}_{comb}|$, given $\tilde{\Sigma}$, we end up with the expression for $E(\Upsilon^{\bullet}_{M})$ as

$$E(\Upsilon^{\bullet}_{M}) = d^{\bullet}_{M,m,n,p,\gamma} \times \frac{(n-p)!}{(n-p-m)!} \times K^{\bullet}_{M,n,p,m}|\Sigma|$$

where

$$K^{\bullet}_{M,n,p,m} = \frac{(-2+\kappa_{n,\alpha,p,m}-m)!}{(-2+\kappa_{n,\alpha,p,m})!}\frac{(Mn-Mp)!}{(Mn-Mp-m)!}$$

for the procedure in subsection 2.3.1 and

$$K^{\bullet}_{M,n,p,m} = \frac{(-2+\kappa_{n,\alpha,p,m}-m)!}{(-2+\kappa_{n,\alpha,p,m})!}\frac{(Mn-p)!}{(Mn-p-m)!}$$

for the procedure in subscection 2.3.2, with $\kappa_{n,\alpha,p,m} = n+\alpha-p-m-1$, assuming $n+\alpha > p+2m+2$. For the original data we take $K^{\bullet}_{0,n,p,m} = 1$.

For the Plug-in method, we have

$$E(|(n-p)\mathbf{S}|) = \frac{(n-p)!}{(n-p-m)!}|\Sigma|,$$

and, conditionally on $\mathbf{S}$, $M(n-p)\overline{\mathbf{S}}^{*}_{M} \sim W_m(\mathbf{S}, M(n-p))$ and $(Mn-p)\mathbf{S}^{*}_{comb} \sim W_m(\mathbf{S}, Mn-p)$, thus concluding that, conditionally on $\mathbf{S}$,

$$E(|(n-p)\overline{\mathbf{S}}^{*}_{M}|) = \frac{1}{M^m(n-p)^m} \times \frac{(Mn-Mp)!}{(Mn-Mp-m)!} \times |(n-p)\mathbf{S}| \tag{4.6}$$

and

$$E(|(n-p/M)\mathbf{S}_{comb}|) = \frac{1}{M^m(n-p)^m} \times \frac{(Mn-p)!}{(Mn-p-m)!} \times |(n-p)\mathbf{S}|. \tag{4.7}$$

Combining the result of $E(|(n-p)\mathbf{S}|)$ defined in (4.4) with each of the expected values in (4.6) and (4.7), conditionally on $\mathbf{S}$, we end up with the expression for $E(\Upsilon^{*}_{M})$ as

$$E(\Upsilon^{*}_{M}) = d^{*}_{M,m,n,p,\gamma} \times \frac{(n-p)!}{(n-p-m)!} \times K^{*}_{M,n,p,m}|\Sigma|$$

where $K^{*}_{0,n,p,m} = 1$ for the original data,

$$K^{*}_{M,n,p,m} = \frac{1}{M^m(n-p)^m}\frac{(Mn-Mp)!}{(Mn-Mp-m)!}$$

for the procedure in subsection 3.2.2 and

$$K^{*}_{M,n,p,m} = \frac{1}{M^m(n-p)^m}\frac{(Mn-Mp)!}{(Mn-Mp-m)!}$$

for the procedure in subscection 3.2.3.

For the PPS method, in the multiple imputation case, the *radius* cannot be used directly, due to the fact that it involves a sum of ratios where the denominators are the different estimators $\mathbf{S}^{\dagger}_{j}\,(j = 1,\ldots,M)$, being only possible to frame the 'real' *radius* between an upper and a lower bound.

Let us then recall $T_M^\dagger$ defined in (2.27) from Subsection 2.2.2, that can be written as

$$T_M^\dagger = \sum_{j=1}^{M} \frac{|(\mathbf{B}_j^\dagger - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}_j^\dagger - \mathbf{B})|}{|(n-p)\mathbf{S}_j^\dagger|}.$$

One may see that it is possible to delimitate the values of $T_M^\dagger$ by considering the maximum and the minimum of the denominators sum.

If one takes $\chi_{max} = \max_{j=1,\dots,M}\left\{|(n-p)\mathbf{S}_j^\dagger|\right\}$ and $\chi_{min} = \min_{j=1,\dots,M}\left\{|(n-p)\mathbf{S}_j^\dagger|\right\}$ then

$$T_M^\dagger \geq \sum_{j=1}^{M} \frac{|(\mathbf{B}_j^\dagger - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}_j^\dagger - \mathbf{B})|}{\chi_{max}}$$

and

$$T_M^\dagger \leq \sum_{j=1}^{M} \frac{|(\mathbf{B}_j^\dagger - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}_j^\dagger - \mathbf{B})|}{\chi_{min}}$$

thus framing the values of $T_M^\dagger$ by an upper bound and a lower bound.

Therefore, let us propose the following two measures, for $M > 1$,

$$\Upsilon_{M,max}^\dagger = \frac{1}{M} d_{M,m,n,p,\alpha;\gamma}^\dagger \chi_{max} \tag{4.8}$$

and

$$\Upsilon_{M,min}^\dagger = \frac{1}{M} d_{M,m,n,p,\alpha;\gamma}^\dagger \chi_{min} \tag{4.9}$$

that will make possible to frame the value of the *radius* for the PPS case. For the PPS minimum (4.9) and maximum (4.8), $d_{M,m,n,p,\alpha;\gamma}^\dagger$ will be the cut-off point in (2.29). The actual *radius* will be delimited by these boundaries and can be, for example, estimated by means of these. One could think on using the mean of the $\mathbf{S}_j^\dagger$, $(j = 1,\dots,M)$, to define an approximate value for the *radius*. However, this would entail problems trying to obtain its expected value due to the problem of dealing with the distribution of the sum of Wishart distributions with different parameter matrices.

The proposed *radius* measures will allow us to compare the precision of the three methods of synthesizing data, PPS, FPPS and Plug-in Sampling methods.

4.2 Simulation Studies

In this section, we will perform some simulations in order to show that the inference methods developed in Chapters 2 and 3 perform as predicted, as well as in order to compare the *radius* of the confidence sets defined for our exact procedures, which will allow us to compare the precision of the proposed methods.

For the entire simulations, the population distribution is taken as a multivariate normal distribution with expected value given by the right hand side of (1.1), for $m = 2$ and $p = 3$, with matrix of regression coefficients

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & 1 \end{pmatrix}$$

and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

All simulations were carried out using the software Mathematica® version 9.

For the FPPS and PPS methods, we set $\alpha = 6$, in order to have, $\mathbf{S}_j^{\dagger}, j = 1,...,M$, $\overline{\mathbf{S}}_M^{\bullet}$ and $\mathbf{S}_{comb}^{\bullet}$ as UEs of Σ. The regression variables x_{1i}, x_{2i}, x_{3i}, $i = 1,...,n$ are generated as i.i.d. $N(1,1)$ and held fixed for the entire simulation.

4.2.1 Accuracy of Procedures proposed in Chapters 2 and 3

The first objective of these simulations is to show that the new exact inference methods developed in subsections 2.2.2, 2.3.1, 2.3.2, 3.2.2 and 3.2.3 perform as predicted in terms of the confidence sets coverage, as well as to compare the accuracy of our proposed methodologies with the accuracy of the adapted Reiter methodology for multiply imputed partially synthetic data.

Based on Monte Carlo simulation with 10^5 iterations, an estimate of the coverage probability (percentage of observed values of the statistics smaller than the respective theoretical cut-off points) is computed for the following confidence regions, where in all cases, the confidence level is set to 0.95 ($\gamma = 0.05$):

1. the confidence sets for $\mathbf{B}$ and for $\mathbf{C} = \mathbf{AB}$, given by (2.29), respectively with $\mathbf{A} = \mathbf{I}_3$ and $\mathbf{A} = (\,\mathbf{0}_{2\times 1} | \,\mathbf{I}_2)$, based on single and multiple synthetic datasets generated via PPS, for $M = 1, 2, 5$; the estimated coverage probability of the confidence set for $\mathbf{B}$ and the estimated coverage probability of the confidence set for $\mathbf{AB}$ are shown in Table 4.1;

2. the confidence set for $\mathbf{B}$ obtained using Reiter's adapted methodology, for $M(> 1)$ synthetic datasets generated via PPS, as described in subsection 2.2.1; for each of the cases $M = 2$ and $M = 5$, the estimated coverage probabilities of the confidence sets are shown in Table 4.1 under the column $\mathbf{vec}(\mathbf{B})$;

3. the confidence sets for $\mathbf{B}$ and for $\mathbf{C} = \mathbf{AB}$, given by (2.34), for the first procedure, and (2.42), for the second procedure, with $\mathbf{A} = \mathbf{I}_3$ and $\mathbf{A} = (\,\mathbf{0}_{2\times 1}|\,\mathbf{I}_2)$, based on multiple synthetic dataset generated via FPPS as in (2.30), for $M = 2, 5$; the estimated coverage probabilities of the confidence sets are shown in Table 4.2 under the columns $\mathbf{B}(1)$ and $\mathbf{AB}(1)$ for the 1st new procedure, and under the columns $\mathbf{B}(2)$ and $\mathbf{AB}(2)$ for the 2nd new procedure;

4. the confidence set for $\mathbf{B}$ obtained using Reiter's adapted methodology, for $M(> 1)$ synthetic datasets generated via Plug-in Sampling, in subsection 3.2.1; for each of the cases $M = 2$ and $M = 5$, the estimated coverage probabilities of the confidence sets are shown in Table 4.3 under the column $\mathbf{vec}(\mathbf{B})$;

5. the confidence sets for $\mathbf{B}$ and for $\mathbf{C} = \mathbf{AB}$, given by (3.13), for the first procedure, and (3.17), for the second procedure, with $\mathbf{A} = \mathbf{I}_3$ and $\mathbf{A} = (\,\mathbf{0}_{2\times 1}|\,\mathbf{I}_2)$, based on single and multiple synthetic dataset generated via Plug-in as in (3.9), for $M = 1, 2, 5$; for $M = 1$ the estimated coverage probabilities of the confidence sets are shown in Table 4.3 under the columns $\mathbf{B}$ and $\mathbf{AB}$, and for $M = 2$ and $M = 5$ the estimated coverage probabilities of the confidence sets are shown in Table 4.3 under the columns $\mathbf{B}(1)$ and $\mathbf{AB}(1)$ for the 1st new procedure, and under the columns $\mathbf{B}(2)$ and $\mathbf{AB}(2)$ for the 2nd new procedure.

Table 4.1: Estimated coverage probability for $\mathbf{vec}(\mathbf{B})$, $\mathbf{B}$ and $\mathbf{AB}$ under PPS.

n	$M = 1$		$M = 2$			$M = 5$		
	$\mathbf{B}$	$\mathbf{AB}$	$\mathbf{vec}(\mathbf{B})$	$\mathbf{B}$	$\mathbf{AB}$	$\mathbf{vec}(\mathbf{B})$	$\mathbf{B}$	$\mathbf{AB}$
10	0.951	0.949	0.856	0.948	0.950	0.749	0.950	0.949
50	0.949	0.951	0.939	0.950	0.949	0.931	0.951	0.949
100	0.950	0.949	0.955	0.951	0.950	0.943	0.950	0.951
200	0.949	0.950	0.956	0.950	0.950	0.945	0.951	0.950

In Table 4.2, the values for $M = 1$ of the estimated coverage probability for $\mathbf{B}$ and $\mathbf{AB}$ are not included since the FFPS concurs with the PPS method.

The results in Tables 4.1, 4.2 and 4.3, for $n = 10, 50, 100, 200$, show that, based on singly or multiply imputed synthetic data, the confidence sets for $\mathbf{B}$ and $\mathbf{AB}$ when $\gamma = 0.05$ have an estimated coverage probability approximately equal to 0.95 for all the exact likelihood based procedures developed in this work. One may also observe, that when using Reiter's adapted methodology the estimated

Table 4.2: Estimated coverage probability for **B** and **AB** under FPPS.

n	$M=2$				$M=5$			
	1st Proc.		2nd Proc.		1st Proc.		2nd Proc.	
	B(1)	**AB**(1)	**B**(2)	**AB**(2)	**B**(1)	**AB**(1)	**B**(2)	**AB**(2)
10	0.948	0.950	0.951	0.952	0.950	0.949	0.948	0.950
50	0.950	0.949	0.951	0.949	0.951	0.949	0.949	0.949
100	0.951	0.950	0.949	0.950	0.950	0.951	0.950	0.951
200	0.950	0.950	0.951	0.951	0.951	0.950	0.951	0.951

Table 4.3: Estimated coverage probability for **vec**(**B**), **B** and **AB** under Plug-in Sampling.

n	$M=1$		$M=2$					$M=5$				
	B	**AB**	Reiter	1st Proc.		2nd Proc.		Reiter	1st Proc.		2nd Proc.	
			vec(**B**)	**B**(1)	**AB**(1)	**B**(2)	**AB**(2)	**vec**(**B**)	**B**(1)	**AB**(1)	**B**(2)	**AB**(2)
10	0.951	0.949	0.828	0.948	0.950	0.951	0.952	0.748	0.950	0.949	0.948	0.950
50	0.949	0.951	0.951	0.950	0.949	0.951	0.949	0.918	0.951	0.949	0.949	0.949
100	0.950	0.949	0.955	0.951	0.950	0.949	0.950	0.941	0.950	0.951	0.950	0.951
200	0.949	0.950	0.958	0.950	0.950	0.951	0.951	0.943	0.951	0.950	0.951	0.951

coverage probabilities fall short of the stipulated level of 0.95 for very small sample sizes, converging to the desired level when increasing the sample size, due to the fact that Reiter's combination rule is asymptotic in nature. Thus, with these results we show that, in fact, the exact procedures developed in this thesis can be applied to synthetic datasets even when the sample size is small.

4.2.2 *Radius* of the confidence sets when using PPS, FPPS and Plug-in Sampling cases

The second objective of these simulations is to compare the *radius* of the confidence sets when inference is made about **B** using the PPS, FPPS and Plug-in Sampling methods. In the PPS case, considering $M > 1$, one faces a problem, which is, as referred at the end of subsection 4.1.2 the impossibility to compute the exact expectation of both χ_{max} and χ_{min} defined in (4.8) and (4.9). Therefore, for the purpose of comparison of the *radius* between all the methods, for the PPS case we will only use the average values of $\Upsilon^{\dagger}_{M,min}$ and $\Upsilon^{\dagger}_{M,max}$ simulated from the synthetic data.

In Table 4.4 are presented the average of the simulated values of the *radius* $\Upsilon^{\bullet}_{M}$, Υ^{*}_{M}, in (4.2) and (4.3), for the confidence sets $\Delta^{\bullet}_{M}(C)$ and $\Delta^{*}_{M}(C)$ (FPPS and Plug-in first procedures) and for the confidence sets $\Delta^{\bullet}_{comb}(C)$ and $\Delta^{*}_{comb}(C)$ (FPPS and Plug-in second procedures) with their corresponding expected values, when taking $\mathbf{A} = \mathbf{I}_p$. Also in Table 4.4 are presented the average of the simulated values of $\Upsilon^{\dagger}_{M,max}$ and $\Upsilon^{\dagger}_{M,min}$, defined respectively in (4.8) and (4.9) for the confidence set

$\Psi_M(\mathbf{C})$ defined in (2.29), also taking $\mathbf{A} = \mathbf{I}_p$. Under the columns Orig are shown the expected values of the *radius* concerning the original data. In Table 4.5 we present the same quantities as in Table 4.4 but when taking $\mathbf{C} = \mathbf{AB}$ with $\mathbf{A} = (\mathbf{0}_{2\times 1} | \mathbf{I}_2)$. These values are based on a Monte Carlo simulation with 10^5 iterations.

Table 4.4: Average values of the *radius* when using FPPS and Plug-in Sampling with the corresponding expected values and the values of $\Upsilon^{\dagger}_{M,min}$ and $\Upsilon^{\dagger}_{M,max}$ defined in (4.9) and (4.8) when using PPS, for the confidence set for $\mathbf{B}$.

n	Orig	$M=1$		$M=2$				$M=5$			
		avg	exp	1st Proc.		2nd Proc.		1st Proc.		2nd Proc.	
				avg	exp	avg	exp	avg	exp	avg	exp
Plug-in Sampling											
10	37.0	210.3	216.9	91.5	93.1	85.3	88.0	53.8	54.7	51.3	52.5
50	19.1	78.6	78.1	42.9	42.7	42.8	42.7	27.1	27.0	27.1	27.0
200	17.5	69.5	69.7	39.1	39.2	39.1	39.2	25.0	25.0	25.0	25.0
Fixed-Posterior Predictive Sampling											
10	37.0	507.3	512.2	251.6	252.6	237.6	238.7	175.3	176.2	163.8	168.9
50	19.1	176.4	176.5	121.2	121.5	121.2	121.5	92.3	92.8	92.3	92.8
200	17.5	154.9	156.1	105.8	106.6	105.9	106.7	81.9	82.4	81.9	82.4
Posterior Predictive sampling											
		avg	exp	min		max		min		max	
10	37.0	507.3	512.2	206.9		728.8		78.5		1004.9	
50	19.1	176.4	176.5	111.9		178.9		66.4		172.8	
200	17.5	154.9	156.1	108.8		136.0		76.7		122.5	

Table 4.5: Average values of the *radius* when using FPPS and Plug-in Sampling with the corresponding expected values and the values of $\Upsilon^{\dagger}_{M,min}$ and $\Upsilon^{\dagger}_{M,max}$ defined in (4.9) and (4.8) when using PPS, for the confidence set for $\mathbf{C}$.

n	Orig	$M=1$		$M=2$				$M=5$			
		avg	exp	1st Proc.		2nd Proc.		1st Proc.		2nd Proc.	
				avg	exp	avg	exp	avg	exp	avg	exp
Plug-in Sampling											
10	13.4	72.8	75.1	32.5	33.1	30.9	31.8	19.0	19.4	18.6	19.0
50	7.3	30.7	30.5	16.7	16.7	16.7	16.6	10.5	10.5	10.5	10.5
200	7.1	27.5	27.6	15.5	15.5	15.5	15.5	9.9	9.9	9.9	9.9
Fixed-Posterior Predictive Sampling											
10	13.4	172.6	172.3	92.2	92.4	86.2	86.6	63.1	63.4	61.3	61.7
50	7.3	68.9	69.0	47.8	47.9	47.5	47.6	35.3	35.5	35.1	35.3
200	7.1	60.7	61.1	41.7	42.1	41.7	42.1	32.5	32.5	32.5	32.5
Posterior Predictive sampling											
		avg	exp	min		max		min		max	
10	13.4	172.6	172.3	73.2		257.7		28.8		368.5	
50	7.3	68.9	69.0	46.4		74.2		27.1		70.6	
200	7.1	60.7	61.1	47.8		59.8		32.5		51.9	

Observing Tables 4.4 and 4.5 and comparing the entries for the FPPS and for the Plug-in Sampling, we may see that when synthetic data are generated under FPPS, larger *radius* are obtained, for the same sample sizes. In the singly

imputed case, one can observe that the FPPS synthetic datasets will lead to a *radius* that is approximately two and half times that of the *radius* under Plug-in Sampling. Comparing the PPS bound values of $\Upsilon^{\dagger}_{M,min}$ and $\Upsilon^{\dagger}_{M,max}$ with the values obtained when using FPPS, one may note that the 'real' PPS *radius* points out to be indeed larger than the FPPS *radius*, by observing that the mean of the minimum and maximum values is always larger than the FPPS *radius*, and that for some cases the PPS minimum value is even larger than the FPPS *radius*. This may be explained by the fact that with the proposed PPS statistic, which is the sum of statistics associated to each single imputation data analysis used to perform the multiple imputation data analysis, we are not collecting as much information across the synthetic datasets as we do when we use the statistics in the FPPS and Plug-in methods.

We may observe that, for $M > 1$, the values of the *radius*, for both procedures in each FPPS and Plug-in methods become identical for larger sample sizes, as theoretically predicted at the end of Chapters 2 and 3.

As the number M of released synthetic datasets increases, the *radius* decreases in all methods. Eventually for the FPPS case, one may need very large values of M, in order to have values of the *radius* close to the value of the original data's *radius*. Although one may look at this fact as a drawback of the FPPS method, as we will see in the next Chapter, FPPS is the method of generating synthetic data that offers the highest level of privacy protection. We are indeed always dealing with the inevitability of having to balance the quality of the generated synthetic data and the level of disclosure risk.

Chapter 5

An application to Current Population Survey (CPS) data and Risk Level Comparison

5.1 CPS Application

To compare the original data inference with the inferential results obtained from the methods developed in Chapters 2 and 3 for the cases where synthetic datasets are generated via PPS, FPPS and Plug-in Sampling and also to compare these with the inferential results obtained using Reiter's adapted methodology, we provide an application of these procedures to a public use data from the 2000 Current Population Survey (CPS) March supplement conducted by the Census for the Bureau of Labor Statistics based on the civilian non-institutional population of the United States. The full data are available online from http://www.census.gov/cps/. These data was previously used by Reiter [31, 32] and Drechsler and Reiter [7] to illustrate various properties of multiple imputation sampling. The complete data comprises household, family and individual records, but for our study we will focus solely on the household records.

Most of the content associated to the FPPS method in this Section is taken from [25].

The CPS data file contains statistical records on 51,016 households and has a set of seventeen categorical and numerical variables which are shown in Table 5.1.

For the application of our methods to the CPS data, three numerical variables I, AP and PT were selected to form the vector $\mathbf{y}$ of response variables, which will be

Table 5.1: Summary of CPS data variables.

Variable	Label	Range/Category Code
Row name	RN	Row Name
Household alimony payments	AP	Numerical (0 – 54,008)
Household child support payment	CS	Numerical (0 – 23,917)
Household property tax	PT	Numerical (0 – 99,997)
Household income	I	Numerical (-21,011 – 768,742)
Household ID number	ID	Numerical (1 – 64,994)
Household survey weight	SW	Numerical (98.5 – 12,904.7)
Number of people in household	N	Numerical (1 – 16)
Number of people in under 18	L	Numerical (0 – 11)
Number of people married	HM	Numerical (0 – 8)
Child Support Payment for head	CP	Numerical (0 – 23,917)
Age (Years)	A	Numerical (0 – 90)
Highest Level of Education attained	E	31 – Less than 1st grade 32 – 1st to 4th grade 33 – 5th or 6th grade 34 – 7th or 8th grade 35 – 9th grade 36 – 10th grade 37 – 11th grade 38 – 12th grade 39 – High School graduate 40 – Some college but no degree 41 – Associate degree in college (occupation/vocation program) 42 – Associate degree in college (academic program) 43 – Bachelor's degree 44 – Master's degree 45 – Professional school degree 46 – Doctorate degree
Marital status	M	1 – Married 2 – Married armed forces spouse present 3 – Married armed forces spouse absent 4 – Widowed 5 – Divorced 6 – Separated 7 – Single
Race	R	1 – White 2 – Black 3 – Native American 4 – Asian/Pacific Islander
Sex	S	1 – Male 2 – Female
Social Security Payments	SS	Numerical (0 – 50,000)

considered the sensitive variables. The set of regression variables (N,L,A,E,M,R,S) were selected as the non-sensitive variables. After deleting all entries where at least one of the variables I, AP and PT is reported as 0, the sample size was reduced to 141 households.

The assumption of the log-normality of the response variables is used instead of normality and therefore the logarithm of the selected response variables is used. To check the assumed multivariate normality of logarithm of the set of response variables a set of a goodness of fit tests for the multivariate normality was perfomed on the logarithm of the vector **y** of sensitive response variables, using the software Mathematica® version 9. The p-values obtained when using the Anderson-Darling, the Baringhaus-Henze, the Cramér-von Mises, the Jarque-Bera ALM, the Kolmogorov-Smirnov, the Kuiper, the Mardia Kurtosis, the Pearson χ^2 and the Watson U^2 test statistics were larger than 0.05. When using the Mardia Combined, the Mardia Skewness and the Shapiro-Wilk test statistics the p-values obtained were smaller than 0.05, but if the goodness of fit test is computed for the normality of the response variables I, AP and PT, considered separately, the p-values obtained using those three test statistics will be larger than 0.05, except the one obtained from the Shapiro-Wilk test statistic for the variable I, which returned a p-value approximately equal to 0.024. Thus, with all these results in mind, we decided to not reject the assumption that the logarithm of the vector of response variables **y** comes from a multivariate normal distribution.

Even if these CPS data are public use data we will consider the values corresponding to the variables I, AP and PT as the set of values that should not be released to the general public.

In this application, **x**, the vector of regressor variables, will be defined as

$$\begin{aligned}\mathbf{x} = \Big(&1, \mathrm{N}, \mathrm{L}, \mathrm{A}, \cancel{\mathrm{I(E{=}31)}}, \cancel{\mathrm{I(E{=}32)}}, \cancel{\mathrm{I(E{=}33)}}, \mathrm{I(E{=}34)}, \mathrm{I(E{=}35)}, \mathrm{I(E{=}36)}, \mathrm{I(E{=}37)}, \cancel{\mathrm{I(E{=}38)}},\\ &\mathrm{I(E{=}39)}, \mathrm{I(E{=}40)}, \mathrm{I(E{=}41)}, \mathrm{I(E{=}42)}, \mathrm{I(E{=}43)}, \mathrm{I(E{=}44)}, \mathrm{I(E{=}45)}, \mathrm{I(E{=}46)}, \cancel{\mathrm{I(M{=}1)}},\\ &\cancel{\mathrm{I(M{=}2)}}, \mathrm{I(M{=}3)}, \mathrm{I(M{=}4)}, \mathrm{I(M{=}5)}, \mathrm{I(M{=}6)}, \mathrm{I(M{=}7)}, \cancel{\mathrm{I(R{=}1)}}, \mathrm{I(R{=}2)}, \cancel{\mathrm{I(R{=}3)}}, \mathrm{I(R{=}4)},\\ &\cancel{\mathrm{I(S{=}1)}}, \mathrm{I(S{=}2)}\Big)', \end{aligned} \tag{5.1}$$

where I(E=31) will be the indicator variable for E=31, i.e., for individuals that not have completed the 1st grade, I(E=32) will be the indicator variables for E=32, i.e, for individuals that have completed the 1st to 4th grade, and so on, and where the indicator variables for the first code present in the sample for each variable, $\cancel{\mathrm{I(E{=}31)}}$, $\cancel{\mathrm{I(M{=}1)}}$, $\cancel{\mathrm{I(R{=}1)}}$ and $\cancel{\mathrm{I(S{=}1)}}$, are taken out in order to make the model matrix full rank. The x-canceled indicator variables $\cancel{\mathrm{I(E{=}32)}}$, $\cancel{\mathrm{I(E{=}33)}}$, $\cancel{\mathrm{I(E{=}38)}}$, $\cancel{\mathrm{I(M{=}2)}}$ and $\cancel{\mathrm{I(R{=}3)}}$ correspond to categories that were not found in the

141 households sample, thus being also taken out. Therefore, the model matrix $\mathbf{X} = [\mathbf{x}_1 ... \mathbf{x}_n]$ will have $p = 24$ rows and $n = 141$ columns, with rank equal to 24. Using the PPS and Plug-in sampling methods, a single synthetic dataset is generated via each method, assuming $\alpha = 8$ when using PPS method in order to have $\mathbf{S}_1^\dagger$ as an UE of Σ (recalling that the FPPS method concurs with the PPS method for singly imputed synthetic data). For each case, one has in expression (5.2) the realizations of the UEs of Σ, $\mathbf{S}_1^\dagger$ and $\mathbf{S}_1^*$ for the two synthetic datasets generated, and $\mathbf{S}$ for the original data, respectively denoted by $\widetilde{\mathbf{S}}_1^\dagger$, $\widetilde{\mathbf{S}}_1^*$ and $\widetilde{\mathbf{S}}$

$$\widetilde{\mathbf{S}}_1^\dagger = \begin{pmatrix} 1.58572 & -0.20443 & 0.27981 \\ -0.20443 & 1.61395 & 0.16089 \\ 0.27981 & 0.16089 & 0.34648 \end{pmatrix}, \quad \widetilde{\mathbf{S}}_1^* = \begin{pmatrix} 1.2929 & -0.2195 & 0.2518 \\ -0.2195 & 1.3983 & 0.1176 \\ 0.2518 & 0.1176 & 0.4490 \end{pmatrix},$$

$$\widetilde{\mathbf{S}} = \begin{pmatrix} 1.1980 & -0.0375 & 0.2970 \\ -0.0375 & 1.0699 & 0.1175 \\ 0.2970 & 0.1175 & 0.4045 \end{pmatrix}. \tag{5.2}$$

In Table 5.2 are presented the realizations of the UEs of $\mathbf{B}$, $\mathbf{B}_1^\dagger$ and $\mathbf{B}_1^*$ for the synthetic datasets, and $\hat{\mathbf{B}}$ for the original data, respectively denoted by $\widetilde{\mathbf{B}}_1^\dagger$, $\widetilde{\mathbf{B}}_1^*$ and $\widetilde{\hat{\mathbf{B}}}$.

Table 5.2: Estimates of the regressor coefficients from the synthetic data and from the original data.

regress	*SyntheticData* ($\widetilde{\mathbf{B}}^\dagger$)			*SyntheticData* ($\widetilde{\mathbf{B}}^*$)			*OriginalData* ($\widetilde{\hat{\mathbf{B}}}$)		
	I	AP	PT	I	AP	PT	I	AP	PT
Interc	11.4996	3.3381	8.1713	10.1829	3.7094	10.9787	9.8339	4.6663	10.1095
N	0.2801	−0.2562	0.6317	−0.0938	0.1435	0.6189	0.0457	0.0375	0.4585
L	−0.3996	0.4960	−0.6017	0.0812	0.0163	−0.5932	0.0186	0.1310	−0.3851
A	−0.0061	0.0223	0.0018	0.0075	0.0285	−0.0097	0.0118	0.0181	−0.0020
I(E=34)	−4.7732	0.3476	−0.4662	−6.6680	1.2055	−2.0664	−4.4348	0.5944	−1.2291
I(E=35)	−5.5990	2.8081	1.9914	−1.2231	−0.0154	−0.7091	−1.4060	0.9188	−0.1468
I(E=36)	−4.2467	2.2712	0.6907	−0.4478	2.1718	−0.9172	−2.3100	1.0416	−0.5002
I(E=37)	−3.5281	0.7339	1.4653	−1.1547	1.3009	−1.0659	−2.0490	0.7410	0.2335
I(E=39)	−3.3369	1.5590	1.0109	−2.5737	0.7234	−1.1346	−2.2208	0.4054	−0.4136
I(E=40)	−2.8766	1.7608	1.2350	−1.8032	1.0617	−0.6940	−1.8834	0.8519	0.0852
I(E=41)	−2.8266	2.7954	2.3165	−1.5615	1.6881	−0.0291	−1.9468	1.4222	0.1094
I(E=42)	−3.5901	2.3990	0.7908	−2.4543	2.0378	−1.1494	−2.3381	1.3840	−0.0808
I(E=43)	−1.9852	2.1149	1.9765	−1.7090	1.1722	−0.4341	−1.5057	1.0766	0.5309
I(E=44)	−3.2012	2.0495	1.7665	−2.2668	1.5629	−0.2140	−1.8082	1.1301	0.4936
I(E=45)	0.1813	1.1103	1.7535	−1.8984	2.1024	−0.4636	−0.9893	0.7958	0.3057
I(E=46)	0.5791	2.3091	3.5534	0.4558	1.4836	1.1497	−0.6198	1.0766	1.0624
I(M=3)	−2.3691	0.8545	−0.3594	−1.9077	−0.4988	−0.4836	−2.7258	0.0964	−0.2156
I(M=4)	−4.4234	2.2640	−1.2282	−0.0088	0.5609	−0.2349	−0.0134	0.5887	0.3864
I(M=5)	−1.0787	1.5611	0.1170	0.3767	0.6729	0.1184	0.1455	0.4770	0.1558
I(M=6)	−0.8300	−0.2358	−0.2713	0.3948	−0.3092	−0.1046	−0.7122	−0.4448	−0.4025
I(M=7)	−2.8242	2.9533	0.5456	1.0576	0.5476	0.5187	−0.1990	1.1750	0.6685
I(R=2)	0.3378	3.8443	1.4196	−1.0805	3.0078	−0.1619	−0.9205	1.3432	0.4696
I(R=4)	0.0340	1.9168	−0.4519	0.6883	−0.3211	0.3639	−0.7040	0.0975	−0.1618
I(S=2)	1.3582	−0.4793	−0.1588	0.0564	−0.2309	−0.2849	0.1236	−0.1355	−0.4025

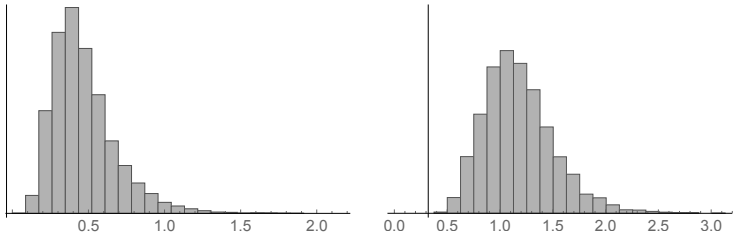
It can be observed, at a first glance, that the point estimates originated via Plug-in Sampling seem to agree more with the original data point estimates than the ones drawn from the PPS method. Nevertheless, this might be due just to a matter of chance when generating the single synthetic dataset for each case. It might have happened that one of the generated datasets resulted from a more biased draw than the other synthetic datasets.

Thus, to be able to have a non-biased analysis of the inferential results obtained for the several inferential procedures developed for the three sampling methods, it is suggested that one conducts inferences on the regression coefficients based on multiple draws instead of one unique draw, thus having a set of values gathered for each method which may help us understand and analyze the differences between the methods proposed in this work. We have therefore decided to generate 100 synthetic datasets for each sampling method and conducted inference on each of these datasets.

Applying methodologies found in subsections 2.2.1, 2.2.2, 2.3.1, 2.3.2, 3.2.1, 3.2.2 and 3.2.3, inferential results on regression coefficients will be obtained, under the form of p-values, with the purpose of analyzing singly and multiply imputed synthetic datasets, considering $M = 1$, $M = 2$ and $M = 5$. The statistics $T_M^{\dagger}$, $T_M^{\bullet}$, $T_{comb}^{\bullet}$, T_M^{*} and T_{comb}^{*} and corresponding empirical distributions, based on simulations with 10^4 iterations, will be used to test the fit of the model and the significance of some regressors. In each inference analysis, one will compute the p-values as the fraction of values of the empirical distribution of the corresponding statistic that are larger than the computed value of the statistic.

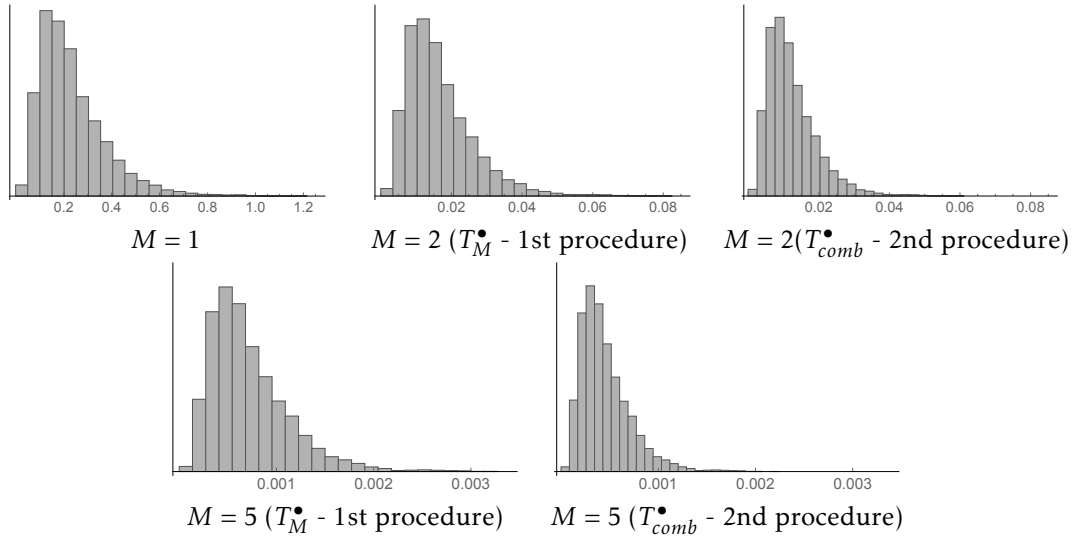
Regarding the test of fit of the model, for all values of M, the results found in every draw of synthetic datasets lead all to the same conclusion, that is, the explanatory variables in $\mathbf{x}$ have a significant role in determining the values of the response variables in $\mathbf{y}$, since the computed p-values were all approximately zero, for all sampling methods developed in this work, for Reiter's adaptations and as well for the original data. As such, there is not much to compare methods and inferential procedures concerning this test.

Remark 5.1.1. *In Figures 5.1, 5.2 and 5.3, one may see the histograms associated with the empirical distributions of* $T_M^{\dagger}$, $T_M^{\bullet}$, $T_{comb}^{\bullet}$, T_M^{*} *and* T_{comb}^{*} *for* $M = 1, 2$ *and* 5 *(for* $m = 3$, $p = 24$, $n = 141$, $\alpha = 8$ *and* 10^4 *simulation sizes).*



$M = 2$ $M = 5$

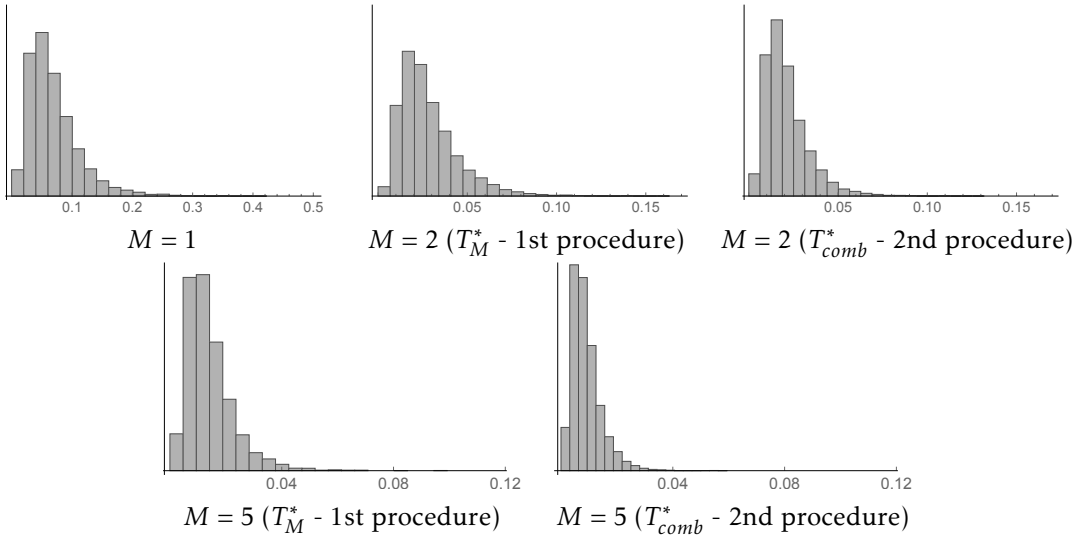
Figure 5.1: Histograms (with same vertical scale) of the empirical distributions of $T_M^{\dagger}$ for $M = 2$ and 5 (for $m = 3$, $p = 24$, $n = 141$, $\alpha = 8$ and 10^4 simulation sizes).



$M = 1$ $M = 2$ ($T_M^{\bullet}$ - 1st procedure) $M = 2$($T_{comb}^{\bullet}$ - 2nd procedure)

$M = 5$ ($T_M^{\bullet}$ - 1st procedure) $M = 5$ ($T_{comb}^{\bullet}$ - 2nd procedure)

Figure 5.2: Histograms (with same vertical scale for each M) of the empirical distributions of both $T_M^{\bullet}$ and $T_{comb}^{\bullet}$ for $M = 1, 2$ and 5 (for $m = 3$, $p = 24$, $n = 141$, $\alpha = 8$ and 10^4 simulation sizes).



$M = 1$ $M = 2$ (T_M^{*} - 1st procedure) $M = 2$ (T_{comb}^{*} - 2nd procedure)

$M = 5$ (T_M^{*} - 1st procedure) $M = 5$ (T_{comb}^{*} - 2nd procedure)

Figure 5.3: Histograms (with same vertical scale for each M) of the empirical distributions of both T_M^{*} and T_{comb}^{*} for $M = 1, 2$ and 5 (for $m = 3$, $p = 24$, $n = 141$ and 10^4 simulation sizes).

Remark 5.1.2. *In Table 5.3 are presented the approximated values of the $\gamma = 0.05$ cut-off points computed from the empirical distributions used in the test of fit of the model for the PPS, FPPS and Plug-in Sampling imputed data (for $M = 1$, it was already ascertained that $T_M^{\dagger} = T_M^{\bullet} = T_{comb}^{\bullet}$ and $T_M^{*} = T_{comb}^{*}$).*

Table 5.3: Approximated values of the cut-off points computed from the empirical distributions of $T_M^{\dagger}$, $T_M^{\bullet}$, $T_{comb}^{\bullet}$, T_M^{*} and T_{comb}^{*} respectively defined in subsections 2.2.2, 2.3.1, 2.3.2, 3.2.2 and 3.2.3, for $\gamma = 0.05$.

values of M	PPS	FPPS		Plug-in	
1		$T_M^{\bullet}$ 5.04E-01		T_M^{*} 1.48E-01	
2	$T_M^{\dagger}$ 8.69E-01	$T_M^{\bullet}$ 3.46E-02	$T_{comb}^{\bullet}$ 2.57E-02	T_M^{*} 6.02E-02	T_{comb}^{*} 4.39E-02
5	$T_M^{\dagger}$ 1.77E00	$T_M^{\bullet}$ 1.49E-03	$T_{comb}^{\bullet}$ 9.35E-04	T_M^{*} 3.02E-02	T_{comb}^{*} 1.92E-02

The study of two different cases is now proposed, that may enable a comparison between methods. Firstly, it is proposed the test of the significance of regressor variables R and S and, secondly, the significance of regressor variables A and E.

For the first case, we consider the 3×24 matrix

$$\mathbf{A} = \left(\mathbf{0}_{3\times 21} \,\middle|\, \mathbf{I}_3 \right)$$

that isolates the indicator regressor values corresponding to the variables R and S. It is intended to test the hypothesis $H_0 : \mathbf{AB} = \mathbf{C}_0$, where C_0 is a 3×3 matrix consisting of only zeros. Performing the test with the original data using (1.5), the p-value computed was approximately equal to 0.249.

We now generate 100 draws of $M = 1$, $M = 2$ and $M = 5$ synthetic datasets via PPS, FPPS and Plug-in Sampling methods. For each draw the corresponding p-values are computed using the empirical distributions of the statistics in (2.28), (2.33), (2.41), (3.12) and (3.16) and also Reiter's adapted procedures for the PPS and Plug-in cases. In these latter cases of Reiter's adapted procedures, for the procedure in subsection 2.2.1, $\mathbf{vec}(\mathbf{B}_i^{\dagger})$ is replaced by $\mathbf{vec}(\mathbf{AB}_i^{\dagger})$, $\mathbf{vec}(\mathbf{B})$ by $\mathbf{vec}(\mathbf{AB})$ and we take $\mathbf{U}_i = \mathbf{S}_i^{\dagger} \otimes (\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}')$, while for procedure in subsection 3.2.1, $\mathbf{vec}(\mathbf{B}_i^{*})$ is replaced by $\mathbf{vec}(\mathbf{AB}_i^{*})$, $\mathbf{vec}(\mathbf{B})$ by $\mathbf{vec}(\mathbf{AB})$ and we take $\mathbf{U}_i = \mathbf{S}_i^{*} \otimes (\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}')$. In Figure 5.4 are presented the box-plots of the referred p-values, with a line marking the original data p-value 0.249.

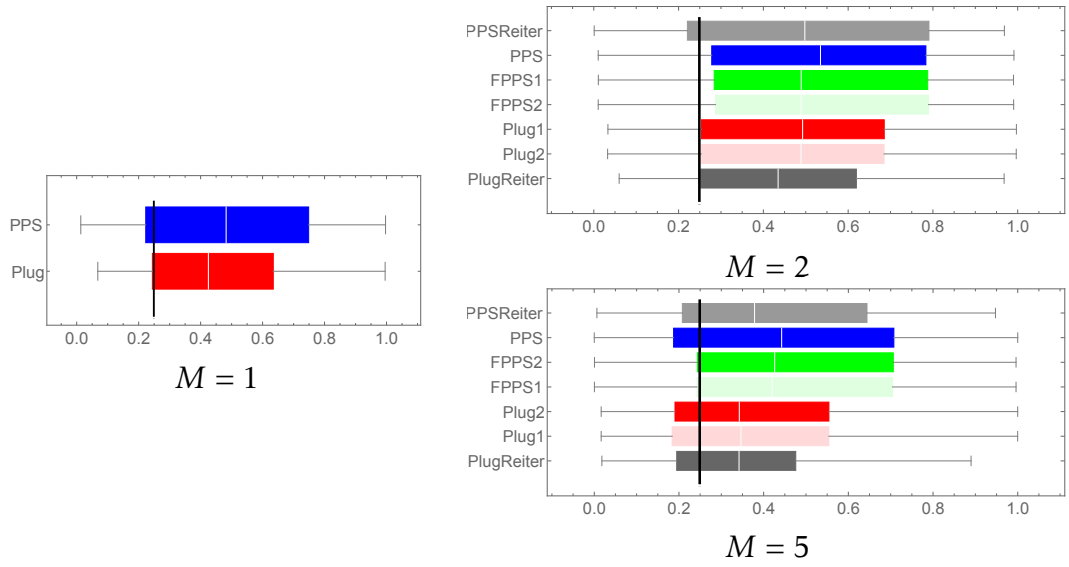


Figure 5.4: Box-plots of p-values obtained, when testing the joint significance of I(R=2), I(R=4) and I(S=2), from 100 draws of synthetic datasets using PPS, FPPS and Plug-in Sampling method as also when using Reiter's adapted combination rule for $M = 1$, $M = 2$ and $M = 5$.

Before making any observations of the results we should note that in general, in cases where the p-value obtained from the original data is rather low, we expect to obtain larger p-values for the synthetic data, given the inherent variability of these synthetic data and the "need" of the inferential exact methods to preserve the $1 - \gamma$ coverage level, and impossibility of compressing the synthetic data p-values towards zero.

Observing Figure 5.4, note that for all procedures the p-values perform as expected, that is, the majority of the p-values obtained from the synthetic data are larger than the ones obtained from the original data. In this case, the estimated coverage probability of Reiter's adapted procedures when using PPS and Plug-in generated data, are respectively approximately equal to 0.938 and to 0.94 leading to performances very similar to that of our procedures. For $M = 1$, where only two box-plots are presented due to the concurrence of methods, the PPS and FPPS methods, and because of the inapplicability of Reiter's adapted procedures to singly imputed synthetic datasets, the gathered p-values do not differ that much from the ones obtained for $M = 2$. When comparing all methods developed in this work we may observe that the spread of p-values is larger for the FPPS method and smaller for the Plug-in method with the p-values obtained for the PPS method having a spread of p-values in between. Using all developed methods and Reiter's adaptations, the majority of the p-values lead to similar conclusions as those obtained from the original data for $\gamma = 0.05$, that is, to not reject that variables R and S do not have significant influence on the response variables.

For the case where we want to test the joint significance of variables A and E, we consider the hypothesis $H_0 : \mathbf{AB} = \mathbf{C}_0$, where C_0 is a 13×13 matrix consisting of only zeros, and

$$\mathbf{A} = \left(\mathbf{0}_{13\times 3} \,\middle|\, \mathbf{I}_{13} \,\middle|\, \mathbf{0}_{13\times 8} \right).$$

The p-value obtained for the original data, based on (1.5), was approximately 0.033, thus rejecting their non-significance for $\gamma = 0.05$, but not rejecting for $\gamma = 0.01$. As in the previous case, in Figure 5.5 are presented the box-plots obtained for the PPS, FPPS, Plug-in Sampling and Reiter's adapted procedures obtained by generating 100 draws of synthetic datasets, for $M = 1$, $M = 2$ and $M = 5$. The vertical line represents again the original data p-value.

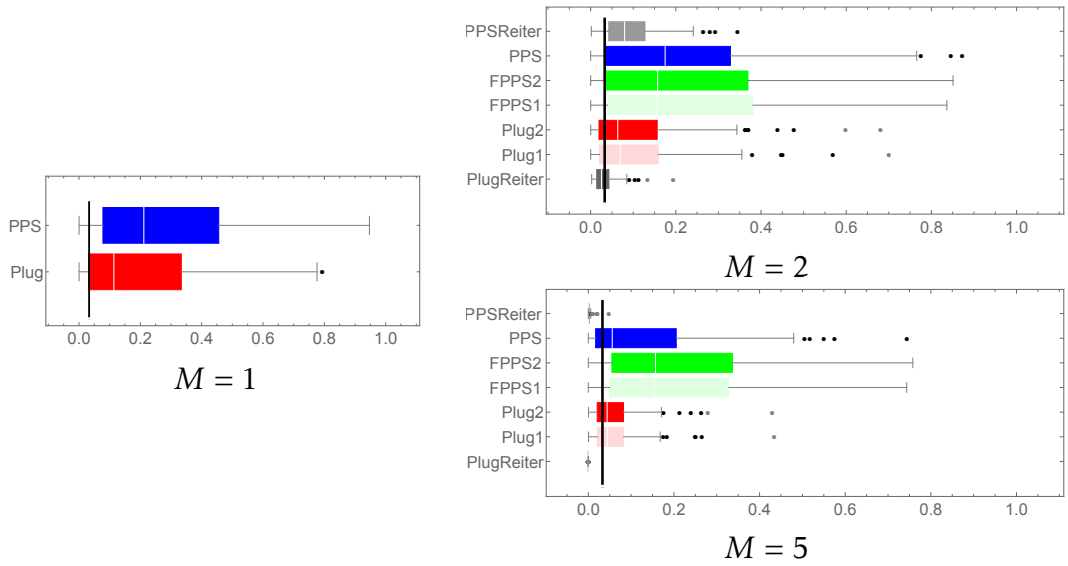


Figure 5.5: Box-plots of p-values obtained, when testing the joint significance of A and E, from 100 draws of synthetic datasets using PPS, FPPS and Plug-in Sampling method as also when using Reiter's adapted combination rule, for $M = 1$, $M = 2$ and $M = 5$.

From Figure 5.5, one may see clearer differences among all methods than for the previous case. The spread of p-values is larger for the FPPS method than for the PPS method, and this latter one has a larger spread of p-values than the Plug-in method, mainly when $M = 5$ datasets are available.

In the $M = 2$ box-plots, if one considers the test for $\gamma = 0.05$, the obtained p-values lead to split decisions, even so, leading in majority to the non-rejection of the null hypothesis except for Reiter's adapted procedures applied to Plug-in Sampling that majorly would reject this null hypothesis. If one considers $\gamma = 0.01$, the p-values obtained when using all inference procedures majorly lead to the same conclusion as the original data p-value.

For the $M = 5$ box-plots, when using Reiter's adapted procedures, the conclusions are in majority to reject the null hypothesis when all other procedures majorly lead to the non-rejection, for $\gamma = 0.05$ or $\gamma = 0.01$.

The different results found with Reiter's adapted procedures, may be explained by the fact that the estimated coverage probability, when considering $\gamma = 0.05$, are approximately 0.932 for $M = 5$ synthetic datasets generated via PPS, and approximately 0.938 for $M = 2$ and 0.922 for $M = 5$ for synthetic datasets generated via Plug-in Sampling, falling short of the stipulated level of 0.95. This may be due to the fact that now a larger number of predictor variables is being used, with a rather small sample size. The estimated coverage probability when Reiter's adapted procedures are applied to $M = 2$ synthetic datasets generated via PPS is 0.955, thus giving results closer to the ones obtained from the new procedures developed in this work. We recall that for procedures with coverage probability approximately equal to 0.95, for $\gamma = 0.05$, the p-values are expected to be majorly larger than the original p-values.

For the two cases studied, the two FPPS multiple imputation procedures presented have very similar p-values, as well as the two Plug-in Sampling procedures. As M increases the spread of the p-values from PPS, FPPS and Plug-in becomes smaller and closer to the original data's p-value, but the FPPS and PPS spread of p-values becomes smaller at a smaller rate than that for the p-values from the Plug-in Sampling.

Another way of illustrating the quality of every method analyzed that is by estimating the power for a given test. For that purpose, let us consider the tests

$$H_0 : \mathbf{B} = \mathbf{B_0} (\neq \mathbf{0}) \; vs \;\; H_1 : \mathbf{B} = \mathbf{B_1} \tag{5.3}$$

and

$$H_0 : \mathbf{AB} = \mathbf{C_0} (\neq \mathbf{0}) \; vs \;\; H_1 : \mathbf{AB} = \mathbf{C_1} \tag{5.4}$$

for $\mathbf{B_0}$ equal to $\tilde{\tilde{\mathbf{B}}}$, rounded to two decimal places,

$$\mathbf{A} = \left(\begin{array}{c|c|c} \mathbf{0}_{12\times 4} & \mathbf{I}_{12} & \mathbf{0}_{12\times 8} \end{array} \right),$$

a 12×12 matrix defined appropriately in order to isolate the indicator variables associated with the variable E, and $\mathbf{C}_1 = \mathbf{AB}_1$ where $\mathbf{B}_1$ takes different values, found in Tables 5.4 and 5.5, with $\mathbf{D}$ a $p \times m$ matrix of 1's. The power is then simulated for the original data, for the synthetic data as well as the power for the case when these synthetic datasets are treated as if they were the original data.

In Tables 5.4 and 5.5, these values are displayed except those for the power when synthetic datasets are treated as if they were the original data, since in these cases the estimated power obtained was always approximately equal to 1, thus having no need to present it. This value is obviously misleading due to the

fact that the estimated coverage probability for these cases will be approximately equal to 0.0198 for testing $\mathbf{B} = \mathbf{B}_0$ and 0.144 for testing $\mathbf{AB} = \mathbf{C}_0$, when synthetic datasets are generated via Plug-in method, and approximately equal to 0 for testing $\mathbf{B} = \mathbf{B}_0$ and 0.006 for testing $\mathbf{AB} = \mathbf{C}_0$, when synthetic datasets are generated via PPS method.

Table 5.4: Power for the test to the hypothesis (5.3), with $\mathbf{B}(1)$ and $\mathbf{B}(2)$ denoting the first and second procedures developed in Chapters 2 and 3, for the FPPS, PPS (only one procedure is available) and Plug-in methods, with $\mathbf{vec}(\mathbf{B})$ denoting Reiter's adapted procedures.

Power $\mathbf{B}_1 =$	orig $\mathbf{B}$	Method	M=1 $\mathbf{B}$	M=2 $\mathbf{B}(1)$	M=2 $\mathbf{B}(2)$	M=2 $\mathbf{vec}(\mathbf{B})$	M=5 $\mathbf{B}(1)$	M=5 $\mathbf{B}(2)$	M=5 $\mathbf{vec}(\mathbf{B})$
$\mathbf{B}_0 + 0.005\mathbf{D}$	0.537	PPS	0.215	0.259		0.433	0.455		0.697
		FPPS		0.252	0.253	N/A	0.275	0.279	N/A
		Plug	0.279	0.382	0.385	0.599	0.471	0.472	0.768
$\mathbf{B}_0 * 0.95$	0.945	PPS	0.535	0.712		0.932	0.903		0.998
		FPPS		0.634	0.637	N/A	0.700	0.700	N/A
		Plug	0.679	0.840	0.841	0.988	0.906	0.909	0.999

Table 5.5: Power for the test to the hypothesis (5.4), with $\mathbf{C}(1)$ and $\mathbf{C}(2)$ denoting the first and second procedures developed in Chapters 2 and 3, for the FPPS, PPS (only one procedure is available) and Plug-in methods, with $\mathbf{vec}(\mathbf{C})$ denoting Reiter's adapted procedures.

Power $\mathbf{C}_1 =$	orig $\mathbf{C}$	Method	M=1 $\mathbf{C}$	M=2 $\mathbf{C}(1)$	M=2 $\mathbf{C}(2)$	M=2 $\mathbf{vec}(\mathbf{C})$	M=5 $\mathbf{C}(1)$	M=5 $\mathbf{C}(2)$	M=5 $\mathbf{vec}(\mathbf{C})$
$\mathbf{A}(\mathbf{B}_0 + 3\mathbf{D})$	0.465	PPS	0.185	0.236		0.388	0.402		0.602
		FPPS		0.202	0.207	N/A	0.245	0.246	N/A
		Plug	0.284	0.334	0.343	0.650	0.416	0.418	0.792
$\mathbf{A}(\mathbf{B}_0 * 0.5)$	0.393	PPS	0.136	0.175		0.265	0.314		0.424
		FPPS		0.160	0.161	N/A	0.179	0.181	N/A
		Plug	0.197	0.271	0.279	0.370	0.326	0.327	0.483

From the power values in Tables 5.4 and 5.5 we may observe that tests based on the synthetic data via FPPS show lower values for its power than the ones based on PPS generation, and this latter show lower values than the ones based on Plug-in generation, as expected, since multiple FPPS was supposed to generate more perturbed data than multiple PPS, and PPS more perturbed data than Plug-in Sampling, even in the single imputation case. These values increase along with the value of M, but with a smaller rate for FPPS synthetic datasets. The huge gains in power based on Reiter's adapted procedures can be explained by the fact that the estimated coverage probability for the tests (5.3) and (5.4) are, in fact, for the PPS case respectively 0.912 and 0.938 for $M = 2$, and 0.912 and 0.925 for $M = 5$, and for the Plug-in case respectively 0.906 and 0.932 for $M = 2$, and 0.908 and 0.921 for $M = 5$, never reaching the nominal value 0.95, being therefore again misleading.

5.2 Privacy Protection of Singly and Multiply Imputed Synthetic Data

After the comparison of 'precision' of all procedures present in this work it will be also important to analyze the level of disclosure risk that each of the FPPS, PPS and Plug-in methods of generating synthetic data offers. Most of the content associated to the FPPS method in this Section is taken from [25].

It is anticipated that singly imputed synthetic data will offer bigger protection than multiply imputed synthetic data and that synthetic data generated via FPPS and PPS will offer bigger protection than synthetic data generated via Plug-in Sampling, with a higher level of protection when using the FPPS method. In this section, this evaluation of risk is estimated using the same CPS data used in the previous section.

Let us consider $\mathbf{V}_l = (\mathbf{v}_{1l},...,\mathbf{v}_{nl}), (l = 1,...,M)$, as the M synthetic datasets generated by any of the sampling methods, FPPS, PPS or Plug-in Sampling, where $\mathbf{v}_{il} = (v_{1il},...,v_{mil})', i = 1,...,n$. Assume that after having access to the released synthetic data an 'intruder' tries to estimate the original values $\mathbf{y_i} = (y_{1i},...,y_{mi})'$ by $\hat{\mathbf{y}}_i = \frac{1}{M}\sum_{l=1}^{M}\mathbf{v}_{il}$. Consequently, the following three criteria are proposed as measures of the level of privacy protection

$$\Gamma_{1,\epsilon} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} Pr\left[\left|\frac{\hat{y_{ji}} - y_{ji}}{y_{ji}}\right| < \epsilon|\mathbf{Y}\right];$$

$$\Gamma_{2,\epsilon} = \frac{1}{n}\sum_{i=1}^{n} Pr\left[\sqrt{\frac{1}{m}\sum_{j=1}^{m}\frac{(\hat{y_{ji}} - y_{ji})^2}{y_{ji}^2}} < \epsilon|\mathbf{Y}\right];$$

$$\Gamma_{3,\epsilon} = Pr\left[\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\frac{\hat{y_{ji}} - y_{ji}}{y_{ji}}\right| < \epsilon|\mathbf{Y}\right],$$

where lower values mean more privacy protection (less disclosure risk) and higher values mean less privacy protection (more disclosure risk).

Let us also consider from $M_{1,\epsilon}$ the following quantity, for $i = 1,...n$ and $j = 1,..,m$,

$$D_{1,\epsilon} = Pr\left[\left|\frac{\hat{y_{ji}} - y_{ji}}{y_{ji}}\right| < \epsilon|\mathbf{Y}\right]$$

and from $M_{3,\epsilon}$ the

$$D_3 = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\frac{\hat{y_{ji}} - y_{ji}}{y_{ji}}\right|.$$

One will use Monte Carlo simulations with 10^4 iterations to estimate all the above measures for each of the $n = 141$ households in the CPS dataset.

In Table 5.6, are shown the values of $\Gamma_{1,0.01}$ and $\Gamma_{2,0.01}$ and for $D_{1,\epsilon}$ its minimum, 1st quartile (Q_1), median, 3rd quartile (Q_3) and maximum. In Table 5.7, are shown for the values of $\Gamma_{3,0.01}$, $\Gamma_{3,0.1}$ and the minimum, Q_1, median, Q_3 and maximum of D_3.

Table 5.6: Values of $\Gamma_{1,0.01}$, $\Gamma_{2,0.01}$ and a summary of the distribution of $D_{1,0.01}$.

M	Method	$\Gamma_{1,0.01}$	$\Gamma_{2,0.01}$	Min	Q_1	Median	Q_3	Max
$M = 1$	PPS	0.0602	0.0005	0	0.0385	0.0507	0.0784	0.1455
	Plug-in	0.0631	0.0006	0	0.0398	0.0552	0.0854	0.1491
$M = 2$	PPS	0.0724	0.0009	0	0.0353	0.0649	0.0911	0.2000
	FPPS	0.0702	0.0009	0	0.0357	0.0624	0.0910	0.1945
	Plug-in	0.0754	0.0010	0	0.0331	0.0697	0.0954	0.2134
$M = 5$	PPS	0.0853	0.0015	0	0.0136	0.0776	0.1268	0.2983
	FPPS	0.0797	0.0012	0	0.0214	0.0711	0.1136	0.2785
	Plug-in	0.0879	0.0018	0	0.0110	0.0792	0.1284	0.3279

Looking at Tables 5.6 and 5.7, one observes that the values of the measures $\Gamma_{1,\epsilon}$, $\Gamma_{2,\epsilon}$ and $\Gamma_{3,\epsilon}$ increase as M increases, showing that the disclosure risk increases with the increase in the number of released synthetic datasets. It is also observed that even for $M = 5$, the maximum value of $D_{1,0.01}$ is 0.3279 when synthetic data has provenience from Plug-in Sampling, thus already indicating a substantial

Table 5.7: Values of $\Gamma_{3,0.1}$ and a summary of the distribution of D_3.

M	Method	$\Gamma_{3,0.1}$	Min	Q_1	Median	Q_3	Max
$M = 1$	PPS	0	0.1091	0.1248	0.1287	0.1325	0.1544
	Plug-in	0	0.1050	0.1202	0.1233	0.1264	0.1379
$M = 2$	PPS	0.0104	0.0906	0.1060	0.1084	0.1110	0.1253
	FPPS	0.0021	0.0960	0.1088	0.1116	0.1145	0.1324
	Plug-in	0.0694	0.0948	0.1026	0.1051	0.1072	0.1159
$M = 5$	PPS	0.9934	0.0840	0.0922	0.0939	0.0956	0.1050
	FPPS	0.5008	0.0896	0.0980	0.1000	0.1020	0.1131
	Plug-in	1	0.0846	0.0905	0.0920	0.0936	0.0992

disclosure risk compared to 0.1455 from the singly imputed case when synthetic data is originated from PPS. Likewise, we may observe that from Table 5.7, one has $\Gamma_{3,\epsilon} = 0$ for $M = 1$ in both Sampling methods but $\Gamma_{3,\epsilon} = 1$ for $M = 5$ in the Plug-in case, the worst case scenario. When looking to the values of $\Gamma_{3,\epsilon}$ for the different cases, we may note that when using FPPS and the measure $\Gamma_{3,\epsilon}$ one can maintain the level of disclosure risk at approximately equal to $\Gamma_{3,\epsilon} = 0.5008$ while with the other two methods this valuable may reach approximately 1.0000. Concluding, the FPPS method of generating synthetic data offers the lowest level of disclosure risk and the Plug-in method the highest level, with the PPS method offering a level of disclosure that is in between of the levels of the other two Sampling methods, nevertheless getting nearer to the values obtained for the Plug-in method as the number of synthetic datasets available increases.

Chapter 6

Final Remarks

The generation of imputed datasets as a Statistical Disclosure Control technique is a relatively recent technique, but it has rapidly become more and more popular and data dissemination agencies already started to use this technique in order to protect and release data. Being a rather recent technique, there is still the need of fulfilling some gaps in the existing literature. One existing problem is the nonexistence of inferential procedures for the analysis of singly imputed datasets, namely under the Multivariate Linear Regression Model. With the intent of solving this problem, we developed a likelihood-based exact inference procedure for the regression coefficients matrix when only one partially synthetic dataset generated via PPS is released, under the MLR model. This way, if agencies require the release of only one synthetic data, perhaps due to privacy concerns, inferential procedures are now made available to analyze this dataset, thus satisfying even the most demanding agencies.

One other issue in the existing literature is that of the inapplicability of the available procedures to samples with rather small size, due to the asymptotic nature of this procedures. It was with this issue in mind, that a likelihood-based exact inference procedure for the usual PPS multiple imputation case was also developed based on our inferential procedure for the single imputation case. Since this procedure was developed based on an exact distribution, it is then possible to apply it even when the synthetic datasets sample size is very small, overcoming the problem that usual procedures face.

With the purpose of simplifying the process of generating datasets, and the inferential analysis of these datasets and to offer a higher level of privacy, in this

thesis, we introduced a new method of synthesizing the datasets from the original data, the FPPS method, and developed two exact inferential procedures to draw inference about the matrix of regression coefficients, under the MLR model. When applying this FPPS method, instead of using a set of different posterior predictive estimators each time a synthetic dataset is generated as in the PPS method, we draw just one set and held it fixed, generating all multiple imputed datasets from the same generating model. The estimators that are used in the generating model of both PPS and FPPS methods are drawn from the same Posterior Predictive distributions, thus, for the single imputation case where only one set of these estimators is needed, these two methods coincide.

The use of the Plug-in Sampling method to generate synthetic data from the original data instead of the PPS method is a very recent multiple imputation technique of generating data for disclosure control purpose. Even if the Plug-in Sampling method has shown to be a very good alternative to the PPS method, it still faces the same problems of the latter, which are the nonexistence of inferential procedures for the single imputation case and the fact that the available inferential procedures for the multiple imputation case are not adequate for the analysis of synthetic datasets with small sample sizes. As such, we also developed likelihood-based exact inference procedures for the single and multiple imputation cases when the partially synthetic datasets released are generated via the Plug-in Sampling method, under the MLR model. One of the advantages of this technique when compared with the FPPS and PPS methods is that the original data estimators are used directly in the generating model and therefore in order to generate partially synthesized datasets via this method one does not need to any knowledge about Bayesian statistics, as it happens when the PPS or the FPPS methods are used, making this method the easiest multiple imputation generation method to use when generating synthetic data.

Regarding the complexity and expertise needed to make inferential analysis from released synthetic datasets, we may note that all procedures developed in this thesis to draw inference for the regression coefficient matrix, considering any of the three methods of generating datasets, are very easy to implement. For instance, to employ the second procedure developed for the FPPS case or the second procedure developed for the Plug-in case, one just needs the point estimates of the regression coefficients matrix and of the covariance matrix which can be easily computed as the usual point estimates considering all multiple synthetic datasets as a unique big dataset.

In order to investigate the precision/accuracy provided by our inferential procedures we performed some simulation studies and also applied these procedures

to the CPS data. From these studies and this application, we observed that the second exact procedures provided for the analysis of data synthesized via FPPS or Plug-in Sampling methods developed are more precise than the first exact procedures, mainly for smaller sample sizes and they become approximately equal as the samples increase in size. When the same number of synthetic datasets is considered, the synthetic datasets generated via the Plug-in Sampling method present better quality than synthetic datasets created via the other two methods, being the FPPS method the one where we will generate more perturbed datasets, does giving a higher level of privacy protection. When the number of multiple imputed datasets increases we also observe an increase of the analysis precision, as one would expect. Nevertheless we should note that the precision obtained from a single imputed dataset when applying our inferential procedures is not that different than the one obtained from the inferential procedures applied when two synthetic datasets are available. We also used the CPS data to investigate the level of privacy offered when releasing replications of the original data created via the FPPS, the PPS or the Plug-in Sampling methods and concluded that the FPPS method is the one that offers more protection to respondents records, followed by the PPS method. As the number of synthetic datasets increases we observe an increase of the disclosure risk.

With the availability of three methods to generate synthetic data and with their corresponding exact inferential procedures to analyze these synthetic datasets, even for the single imputation case, agencies may choose the level of quality versus the level of confidentiality of the data they want to release. If one agency demands the highest level of privacy, disregarding the level of quality, the release of a single synthetic dataset generated via PPS method should be the chosen method. Nevertheless, one should note that if the Plug-in method were to be used in the generation of the singly imputed dataset with the purpose of public availability, the level of privacy would not decrease excessively, with an increase of the data quality. On the other hand, if quality is the main focus of the data disseminators one should release multiply imputed datasets generated via Plug-in Sampling, since with this method one does not need to release a large number of synthetic datasets to respect the data quality demanded by the statistical agency. But, if it is demanded by the agencies the release of multiple imputed synthetic datasets instead of just one single imputed synthetic dataset, the FPPS method is the one that offers the highest protection of the respondents.

Despite the contributions made in this thesis, there is still margin for future research. There is the need of developing an exact inferential procedure to use when the number of tested regressors is smaller than the number of response

variables, since for now Reiter's adapted procedure is the only method to be used in that case. There is also the need towards obtaining expressions for the exact or approximate pdf's and cdf's of the statistics developed for the inferential methods, in such a way that one would not need to resort to the use of empirical distributions and Monte Carlo Simulations. One other research goal may be in the direction of the development of exact inferential procedures to test the covariance structure of the MLR model.

With the development of exact inferential procedures under the MLR model for the PPS and Plug-in Sampling cases and by presenting a new method of generating synthetic datasets, the FPPS method, we are enriching, promoting and making users, analysts and agencies less reluctant to choose the use of single and multiple imputation as a disclosure control technique, by showing its potentiality and ease of application. By overcoming some of the obstacles existing in the literature, the present work may help to call the attention of future researchers towards this area and hopefully will help in expanding the use of single and multiple imputation, making it one of the preferred SDC techniques worldwide.

Bibliography

[1] J. M. Abowd, M. Stinson, and G. Benedetto. "Final report to the social security administration on the SIPP/SSA/IRS public use file project". In: *Suitland, MD: Census Bureau, Longitudinal Employer-Household Dynamics Program* (2006).

[2] D. An and R. J. Little. "Multiple imputation: an alternative to top coding for statistical disclosure control". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.4 (2007), pp. 923–940.

[3] T. W. Anderson. *An Introduction To Multivariate Statistical Analysis*. 3rd. Wiley, 2003.

[4] J. Domingo-Ferrer. "A survey of inference control methods for privacy-preserving data mining". In: *Privacy-preserving data mining*. Springer, 2008, pp. 53–80.

[5] J. Drechsler. "Generating multiply imputed synthetic datasets: theory and implementation". PhD thesis. Bamberg, Univ., Diss., 2009, 2010.

[6] J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control*. Vol. 201. Springer-Verlag New York, 2011.

[7] J. Drechsler and J. P. Reiter. "Sampling with synthesis: A new approach for releasing public use census microdata". In: *Journal of the American Statistical Association* 105.492 (2010), pp. 1347–1357.

[8] J. Drechsler, S. Bender, and S. Rässler. "Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel." In: *Transactions on Data Privacy* 1.3 (2008), pp. 105–130.

[9] S. E. Fienberg. "Confidentiality and disclosure limitation". In: *Encyclopedia of Social Measurement* 1 (2005), pp. 463–469.

[10] S. E. Fienberg and A. P. Sanil. "A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data". In: *Journal of Official Statistics* 13.1 (1997), p. 75.

[11] S. Hawala. "Producing partially synthetic data to avoid disclosure". In: *Proceedings of the Joint Statistical Meetings. Alexandria, VA: American Statistical Association*. 2008.

[12] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. S. Nordholt, G. Seri, and P Wolf. "Handbook on statistical disclosure control". In: *ESSnet on Statistical Disclosure Control* (2010).

[13] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.

[14] A. B. Kennickell. "Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances". In: *Record Linkage Techniques* 1997 (1997), pp. 248–267.

[15] S. K. Kinney, J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd. "Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database". In: *International Statistical Review* 79.3 (2011), pp. 362–384.

[16] S. K. Kinney, J. P. Reiter, and J. Miranda. "Improving the Synthetic Longitudinal Business Database. US Census Bureau". In: *Center for Economic Studies* (2014), pp. 12–14.

[17] M. Klein and B. Sinha. "Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models". In: *Sankhya B* 77.2 (2015), pp. 293–311.

[18] M. Klein and B. Sinha. "Likelihood-Based Finite Sample Inference for Synthetic Data Based on Exponential Model". In: *Thailand Statistician* 13.1 (2015), pp. 33–47.

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

[20] M. Klein and B. Sinha. "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models". In: *Journal of Privacy and Confidentiality* 7.1 (2016), pp. 43–98.

[21] M. Klein, T. Mathew, and B. Sinha. "A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication". In: *U.S. Census Bureau Research Report Series* #2013-02 (2013).

[22] T. Kollo and D. von Rosen. *Advanced Multivariate Statistics with Matrices*. Springer, 2005.

[23] E. L. Lehmann. *Testing statistical hypotheses*. 2nd. John Wiley & Sons, Inc., 1986.

[24] R. Little. "Statistical analysis of masked data". In: *Journal of Official Statistics* 9 (1993), pp. 407–426.

[25] R. Moura, M. Klein, C. A. Coelho, and B. Sinha. "Inference for Multivariate Regression Model based on Synthetic Data generated under Fixed-Posterior Predictive Sampling: Comparison with Plug-in Sampling". In: *Revstat* (2016), (accepted for publication).

[26] T. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., 1982.

[27] T. Raghunathan, J. P. Reiter, and D. Rubin. "Multiple Imputation for Statistical Disclosure Limitation". In: *Journal of Official Statistics* 19 (2003), pp. 1–16.

[28] T. E. Raghunathan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. "A multivariate technique for multiply imputing missing values using a sequence of regression models". In: *Survey methodology* 27.1 (2001), pp. 85–96.

[29] *REGULATION (EC) No 223/2009 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 March 2009*. EN.

[30] J. P. Reiter. "Inference for Partially Synthetic Public Use Microdata Sets". In: *Survey Methodology* 29 (2003), pp. 181–188.

[31] J. P. Reiter. "Releasing multiply-imputed synthetic public use microdata: an illustration and empirical study". In: *Journal of Royal Statistical Society*. A 168 (2005), pp. 185–205.

[32] J. P. Reiter. "Using CART to generate partially synthetic public use microdata". In: *Journal of Official Statistics* 21.3 (2005), p. 441.

[33] J. P. Reiter and S. K. Kinney. "Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions not Necessary". In: *Journal of Official Statistics* 28 (2012), pp. 583–590.

[34] J. P. Reiter and T. Raghunathan. "The Multiple Adaptations of Multiple Imputation". In: *Journal of American Statistical Association* 102 (2007), pp. 1462–1471.

[35] D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

[36] D. Rubin. "Discussion: Statistical Disclosure Limitation". In: *Journal of Official Statistics* 9 (1993), pp. 461–468.

[37] S. D. Woodcock and G. Benedetto. "Distribution-preserving statistical disclosure limitation". In: *Computational Statistics & Data Analysis* 53.12 (2009), pp. 4228–4242.

APPENDICES

A.1 On the distribution of the statistics based on the original data

From the MLR model in (1.1), recall that

$$\hat{\mathbf{B}} \sim N\left(\mathbf{B}, \Sigma \otimes (\mathbf{X}\mathbf{X}')^{-1}\right)$$

and that

$$(n-p)\mathbf{S} \sim W_m(\Sigma, n-p).$$

By Theorem 2.4.1 in [22],

$$\left(\hat{\mathbf{B}} - \mathbf{B}\right)' (\mathbf{X}\mathbf{X}') \left(\hat{\mathbf{B}} - \mathbf{B}\right) \sim W_m(\Sigma, p).$$

By considering,

$$\mathbf{H} = \Sigma^{-1/2} \left(\hat{\mathbf{B}} - \mathbf{B}\right)' (\mathbf{X}\mathbf{X}') \left(\hat{\mathbf{B}} - \mathbf{B}\right) \Sigma^{-1/2} \sim W_m(\mathbf{I}_m, p)$$

and

$$\mathbf{G} = \Sigma^{-1/2} \mathbf{S} \Sigma^{-1/2} \sim W_m(\mathbf{I}_m, n-p),$$

it is known that $|\mathbf{H}|$ will be a product of independent random chi-square variables with $p-i+1$ degrees of freedom, and $|\mathbf{G}|$ will be a product of independent random chi-square variables with $n-p-i+1$ degrees of freedom, for $i = 1, \ldots, m$. Therefore,

$$T_O = \frac{\left|\left(\hat{\mathbf{B}} - \mathbf{B}\right)'(\mathbf{X}\mathbf{X}')\left(\hat{\mathbf{B}} - \mathbf{B}\right)\right|}{|(n-p)\mathbf{S}|} \sim \prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_i$$

where F_i are independent random variables whose distributions are $F_{p-i+1,n-p-i+1}$, for $i = 1,...,m$. Analogously, for any $k \times p$ matrix $\mathbf{A}$ with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$,

$$T_{O,\mathbf{C}} = \frac{\left|\left(\mathbf{A}\hat{\mathbf{B}} - \mathbf{A}\mathbf{B}\right)'(\mathbf{A}(\mathbf{X}\mathbf{X}')\mathbf{A}')^{-1}\left(\mathbf{A}\hat{\mathbf{B}} - \mathbf{A}\mathbf{B}\right)\right|}{|(n-p)\mathbf{S}|} \sim \prod_{i=1}^{m} \frac{k-i+1}{n-p-i+1} F_{k,i}$$

where $F_{k,i}$ are independent random variables whose distributions are $F_{k-i+1,n-p-i+1}$, for $i = 1,...,m$.

A.2 Important Identities

Result A.2.1: Considering $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ as three $p \times m$ matrices, considering $\mathbf{D}$ as a square $p \times p$ matrix and $k \in \mathbb{N}$, we have that

$$\begin{aligned}
k(\mathbf{A}-\mathbf{B})'\mathbf{D}(\mathbf{A}-\mathbf{B}) + (\mathbf{A}-\mathbf{C})'\mathbf{D}(\mathbf{A}-\mathbf{C}) = \\
&= (k+1)\mathbf{A}'\mathbf{D}\mathbf{A} - k\mathbf{B}'\mathbf{D}\mathbf{A} - k\mathbf{A}'\mathbf{D}\mathbf{B} + k\mathbf{B}'\mathbf{D}\mathbf{B} - \mathbf{A}'\mathbf{D}\mathbf{C} - \mathbf{C}'\mathbf{D}\mathbf{A} + \mathbf{C}'\mathbf{D}\mathbf{C} \\
&= (k+1)\mathbf{A}'\mathbf{D}\mathbf{A} - \mathbf{A}'\mathbf{D}(k\mathbf{B}+\mathbf{C}) - (k\mathbf{B}+\mathbf{C})'\mathbf{D}\mathbf{A} + k\mathbf{B}'\mathbf{D}\mathbf{B} + \mathbf{C}'\mathbf{D}\mathbf{C} \\
&= (k+1)\left[\mathbf{A} - \frac{1}{k+1}(k\mathbf{B}+\mathbf{C})\right]'\mathbf{D}\left[\mathbf{A} - \frac{1}{k+1}(k\mathbf{B}+\mathbf{C})\right] \\
&\qquad + k\mathbf{B}'\mathbf{D}\mathbf{B} + \mathbf{C}'\mathbf{D}\mathbf{C} - \frac{1}{k+1}(k\mathbf{B}+\mathbf{C})'\mathbf{D}(k\mathbf{B}+\mathbf{C}).
\end{aligned}$$

Since

$$\begin{aligned}
k\mathbf{B}'\mathbf{D}\mathbf{B} + \mathbf{C}'\mathbf{D}\mathbf{C} - \frac{1}{k+1}(k\mathbf{B}+\mathbf{C})'\mathbf{D}(k\mathbf{B}+\mathbf{C}) \\
&= k\mathbf{B}'\mathbf{D}\mathbf{B} + \mathbf{C}'\mathbf{D}\mathbf{C} - \frac{k^2}{k+1}\mathbf{B}'\mathbf{D}\mathbf{B} - \frac{1}{k+1}\mathbf{C}'\mathbf{D}\mathbf{C} - \frac{k}{k+1}\mathbf{B}'\mathbf{D}\mathbf{C} - \frac{k}{k+1}\mathbf{C}'\mathbf{D}\mathbf{B} \\
&= \frac{k}{k+1}\mathbf{B}'\mathbf{D}\mathbf{B} + \frac{k}{k+1}\mathbf{C}'\mathbf{D}\mathbf{C} - \frac{k}{k+1}\mathbf{B}\mathbf{D}\mathbf{C} - \frac{k}{k+1}\mathbf{C}'\mathbf{D}\mathbf{B} \\
&= \frac{k}{k+1}(\mathbf{B}-\mathbf{C})'\mathbf{D}(\mathbf{B}-\mathbf{C}),
\end{aligned}$$

we may write

$$\begin{aligned}
k(\mathbf{A}-\mathbf{B})'\mathbf{D}(\mathbf{A}-\mathbf{B}) + (\mathbf{A}-\mathbf{C})'\mathbf{D}(\mathbf{A}-\mathbf{C}) = \\
&= (k+1)\left[\mathbf{A} - \frac{1}{k+1}(k\mathbf{B}+\mathbf{C})\right]'\mathbf{D}\left[\mathbf{A} - \frac{1}{k+1}(k\mathbf{B}+\mathbf{C})\right] \\
&\qquad + \frac{k}{k+1}(\mathbf{B}-\mathbf{C})'\mathbf{D}(\mathbf{B}-\mathbf{C}). \qquad \text{(A.1)}
\end{aligned}$$

Result A.2.2: Considering $\mathbf{C}$, $\mathbf{D}$ and $\mathbf{X}$ any three $p \times m$ matrices and $\mathbf{S}$ and Σ two $m \times m$ symmetric positive definite matrices we have that

$$\begin{aligned}
&(\mathbf{C}-\mathbf{X})\mathbf{S}^{-1}(\mathbf{C}-\mathbf{X})' + (\mathbf{X}-\mathbf{D})\Sigma^{-1}(\mathbf{X}-\mathbf{D})' \\
&= (\mathbf{X}-\mathbf{C})\mathbf{S}^{-1}(\mathbf{X}-\mathbf{C})' + (\mathbf{X}-\mathbf{D})\Sigma^{-1}(\mathbf{X}-\mathbf{D})' \\
&= \mathbf{X}(\mathbf{S}^{-1}+\Sigma^{-1})\mathbf{X}' - \mathbf{X}\mathbf{S}^{-1}\mathbf{C}' - \mathbf{X}\Sigma^{-1}\mathbf{D}' - \mathbf{C}\mathbf{S}^{-1}\mathbf{X}' - \mathbf{D}\Sigma^{-1}\mathbf{X}' + \mathbf{C}\mathbf{S}^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}\mathbf{D}' \\
&= \mathbf{X}(\mathbf{S}^{-1}+\Sigma^{-1})\mathbf{X}' - \mathbf{X}(\mathbf{S}^{-1}\mathbf{C}' + \Sigma^{-1}\mathbf{D}') - (\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})\mathbf{X}' + \mathbf{C}\mathbf{S}^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}\mathbf{D}' \\
&= \left[\mathbf{X} - (\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\right](\mathbf{S}^{-1}+\Sigma^{-1})\left[\mathbf{X} - (\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\right]' \\
&\qquad + \mathbf{C}\mathbf{S}^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}\mathbf{D}' - (\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})'.
\end{aligned}$$

Taking the last three terms of the previous sum, we have the following equalities

$$\begin{aligned}
&\mathbf{C}\mathbf{S}^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}\mathbf{D}' - (\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})' \\
&\quad = \mathbf{C}\mathbf{S}^{-1}\mathbf{C}' - \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}\mathbf{D}' - \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mathbf{D}' \\
&\qquad\qquad - \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mathbf{D}' - \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1}\mathbf{C}' \\
&\quad = \mathbf{C}(\mathbf{S}^{-1} - \mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1})\mathbf{C}' + \mathbf{D}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1})\mathbf{D}' \\
&\qquad\qquad - \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mathbf{D}' - \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1}\mathbf{C}'.
\end{aligned}$$

Considering the fact that for any two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1} + \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1}) = \mathbf{A}^{-1},$$

then we may use the identity

$$\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}$$

to conclude that

$$\begin{aligned}
&\mathbf{C}\mathbf{S}^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}\mathbf{D}' - (\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})' \\
&\quad = \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mathbf{C}' + \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1})\mathbf{D}' \\
&\qquad\qquad - \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}\mathbf{D}' - \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1}\mathbf{C}' \\
&\quad = \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}(\Sigma^{-1}\mathbf{C}' - \Sigma^{-1}\mathbf{D}') + \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}(\mathbf{S}^{-1}\mathbf{C}' - \mathbf{S}^{-1}\mathbf{D}') \\
&\quad = \mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\Sigma^{-1}(\mathbf{C}' - \mathbf{D}') + \mathbf{D}\Sigma^{-1}(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\mathbf{S}^{-1}(\mathbf{C}' - \mathbf{D}').
\end{aligned}$$

Finally, if one considers the fact that, for any two positive definite matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1} = \left(\mathbf{A}^{-1}(\mathbf{I_m}+\mathbf{A}\mathbf{B}^{-1})^{-1}\right)^{-1} = \left(\mathbf{A}^{-1}(\mathbf{B}+\mathbf{A})\mathbf{B}^{-1}\right)^{-1} = \mathbf{B}(\mathbf{B}+\mathbf{A})^{-1}\mathbf{B},$$

we end up having

$$\begin{aligned}&\mathbf{C}\mathbf{S}^{-1}\mathbf{C}'+\mathbf{D}\Sigma^{-1}\mathbf{D}'-(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})'\\&\quad=\mathbf{C}(\mathbf{S}+\Sigma)^{-1}(\mathbf{C}'-\mathbf{D}')+\mathbf{D}(\mathbf{S}+\Sigma)^{-1}(\mathbf{D}'-\mathbf{C}')=(\mathbf{C}-\mathbf{D})(\mathbf{S}+\Sigma)^{-1}(\mathbf{C}-\mathbf{D})'.\end{aligned}$$

Thus, in conclusion we have the following equality

$$\begin{aligned}&(\mathbf{C}-\mathbf{X})\mathbf{S}^{-1}(\mathbf{C}-\mathbf{X})'+(\mathbf{X}-\mathbf{D})\Sigma^{-1}(\mathbf{X}-\mathbf{D})'=\\&\left[\mathbf{X}-(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\right](\mathbf{S}^{-1}+\Sigma^{-1})\left[\mathbf{X}-(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\Sigma^{-1})(\mathbf{S}^{-1}+\Sigma^{-1})^{-1}\right]'+\\&\qquad(\mathbf{C}-\mathbf{D})(\mathbf{S}+\Sigma)^{-1}(\mathbf{C}-\mathbf{D})'.\end{aligned}\tag{A.2}$$

A.3 Mathematica® source codes for the empirical distributions of $T_M^{\dagger}$, $T_M^{\bullet}$, $T_{comb}^{\bullet}$, T_M^{*} and T_{comb}^{*}

Listing A.1: Example of source code for the empirical distribution of $T_M^{\dagger}$ defined in (2.27) used in the CPS application.

```
rfish[a_] := RandomVariate[FRatioDistribution[p - a, n - p - a]]*
(p - a)/(n - p - a);

Needs["MultivariateStatistics`"]
A1[M_] := RandomReal[WishartDistribution[M, n + alpha - p - m - 1]];

A2[M_] := RandomReal[WishartDistribution[M, n - p]];

M = 2;
m = 3;
p = 24;
alpha = 8;
n = 141;
sim = 10000;
Id = IdentityMatrix[m];
dist = ConstantArray[{1}, sim];

Timing[Do[AA1 = MatrixPower[A1[Id], 1/2];
dist[[i]] = (Product[rfish[a], {a, 0, m - 1}]*
  Re[Det[AA1.Inverse[A2[Id]].AA1 + 2*Id]])
    + (Product[rfish[a], {a, 0, m - 1}]*
  Re[Det[AA1.Inverse[A2[Id]].AA1 + 2*Id]]),
{i, sim}];]
```

Listing A.2: Example of source code for the empirical distribution of $T_M^{\bullet}$ and $T_{comb}^{\bullet}$ respectively defined in (2.32) and (2.40), used in the CPS application.

```
rfish1st[a_] := RandomVariate[
FRatioDistribution[p - a, M (n - p) - a]]*(p - a)/(M (n - p) - a);

rfish2nd[a_] := RandomVariate[
FRatioDistribution[p - a, M*n - p - a]]*(p - a)/(M*n - p - a);


Needs["MultivariateStatistics`"]
A1[M_] := RandomReal[WishartDistribution[M, n + alpha - p - m - 1]];

A2[M_] := RandomReal[WishartDistribution[M, n - p]];

M = 2;
m = 3;
p = 24;
alpha = 8;
n = 141;
sim = 10000;
Id = IdentityMatrix[m];
dist = ConstantArray[{1}, sim];

Timing[Do[AA1 = MatrixPower[A1[Id], 1/2];
dist[[i]] = Product[rfish1st[a], {a, 0, m - 1}]*
Re[Det[AA1.Inverse[A2[Id]].AA1 + (M + 1)/M*Id]],
{i, sim}];]

Timing[Do[AA1 = MatrixPower[A1[Id], 1/2];
dist[[i]] = Product[rfish2nd[a], {a, 0, m - 1}]*
Re[Det[AA1.Inverse[A2[Id]].AA1 + (M + 1)/M*Id]],
{i, sim}];]
```

Listing A.3: Example of source code for the empirical distribution of T_M^* and T_{comb}^* respectively defined in (3.11) and (3.15), used in the CPS application.

```
m = 3;
p = 24;
n = 141;
M= 2;

Needs["MultivariateStatistics`"]
Id = IdentityMatrix[m];
Invw[M_] := Inverse[RandomReal[WishartDistribution[M, n - p]]];

rfish21st[a_] := RandomVariate[FRatioDistribution[p - a,
M*(n - p) - a]]*(p - a)/(M*(n - p) - a);

rfish22nd[a_] := RandomVariate[FRatioDistribution[p - a,
M*n - p - a]]*(p - a)/(M*n - p - a);

sim = 10000;
dist21st = ConstantArray[{1}, sim];
dist22nd = ConstantArray[{1}, sim];

Timing[Do[
dist21st[[i]] = Product[rfish21st[i], {i, 0, m - 1}]*
Det[M*(n - p)*Invw[Id] + Id];

dist22nd[[i]] = Product[rfish22nd[i], {i, 0, m - 1}]*
Det[M*(n - p)*Invw[Id] + Id],
{i, sim}];];
```