RESEARCH REPORT SERIES
(Statistics #2018-02)

Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling

Ricardo Moura
CMA, Faculty of Sciences and Technology, Nova University of Lisbon

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

John Zylstra
Department of Mathematics and Statistics,
University of Maryland, Baltimore County

Carlos Coelho
CMA and Mathematics Department, Faculty of Sciences and Technology,
Nova University of Lisbon

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

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

Report Issued: February 5, 2018

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

Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling

Ricardo Moura
CMA, Faculty of Sciences and Technology, Nova University of Lisbon

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

John Zylstra
Department of Mathematics and Statistics,
University of Maryland, Baltimore County

Carlos Coelho
CMA and Mathematics Department, Faculty of Sciences and Technology,
Nova University of Lisbon

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

Abstract

In this paper, the authors derive the likelihood-based exact inference for *singly* and *multiply* imputed synthetic data in the context of a multivariate regression model. The synthetic data are generated via the Plug-in Sampling method, where the unknown parameters in the model are set equal to the observed values of their point estimators based on the original data, and synthetic data are drawn from this estimated version of the model. Simulation studies are carried out in order to confirm the

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

theoretical results. In case multiple synthetic datasets are permissible, the authors provide an *exact* test procedure and compare their results with the asymptotic results of Reiter (2005). An application using U.S. 2000 Current Population Survey data is discussed. Furthermore, properties of the proposed methodology are evaluated in scenarios where some of the conditions that were used to derive the methodology do not hold.

Key Words: Finite sample analysis, Maximum likelihood estimators, Multivariate Regression, Pivotal quantities, Plug-in Sampling, Statistical Disclosure Control, Unbiased estimators.

1 Introduction

Methods of statistical disclosure control are used to achieve the competing goals of publishing statistical outputs from surveys, while protecting the survey respondents' confidential data from disclosure. Statistical disclosure control methods include data swapping, perturbation with randomly added or multiplied noise, and the release of synthetic data. The use of synthetic datasets has gained considerable popularity and importance in recent times (Klein et al., 2013). In this paper, we investigate some inferential aspects of statistical analysis based on synthetic data when real datasets are not released and a single synthetic dataset based on the original data is created as substitute for publication and analysis. We will also discuss the case when multiple synthetic datasets are released and analyzed. Little (1993) and Rubin (1993) first advocated use of synthetic data for statistical disclosure control, using the framework of multiple imputation (Rubin, 1987). Rubin (1993) argued that synthetic data so created do not correspond to any actual sampling unit, thus preserving the confidentiality of the respondents. Inferential methods for fully synthetic data were developed by Raghunathan et al. (2003), Reiter (2005) presented an illustration and empirical study of fully synthetic data and Reiter and Raghunathan (2007) provided an overview of multiple imputation techniques, including its use in statistical disclosure control. Reiter (2003) presented methods for drawing inference for partially synthetic data. This is exactly the context of our paper. Usually there are two ways one can generate

synthetic data: Posterior Predictive Sampling and Plug-in Sampling (Reiter and Kinney, 2012), and statistical methods of data analysis can deal with both cases.

Although most inferential methods for synthetic data are based on multiple imputation, Klein and Sinha (2015a,b,c, 2016) in a series of recent papers developed exact parametric inferential methods based on singly imputed synthetic data for several probability models, including multiple linear regression model where the sole response variable is taken as sensitive, thus requiring protection, while the covariates are treated as non-sensitive. There are cases where singly imputed synthetic data have been released (Hawala, 2008), and therefore procedures for valid data analysis are desirable.

Our main objective in this paper is to extend this scenario to the case of a multivariate linear regression model where there are multiple sensitive responses following a multivariate normal distribution with means modeled as linear combinations of multiple non-sensitive covariates. Based on the fitted multivariate linear regression model, we synthesize the sensitive responses based on the Plug-in Sampling method, and develop exact data analysis procedures for both single and multiple imputation.

A brief description of the Plug-in Sampling method, which will be used throughout the paper, follows (in a future communication we will address the case of Posterior Predictive Sampling method). Suppose that $\mathbf{Y} = (\mathbf{y_1}, ..., \mathbf{y_n})$ 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. 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 singly imputed synthetic data, denoted by $\mathbf{V}$, are then generated by drawing $\mathbf{V} = (\mathbf{v_1}, ..., \mathbf{v_n})$ from the joint pdf $f_{\hat{\boldsymbol{\theta}}}$. In case of multiple imputation, this procedure is independently repeated M times to generate M synthetic datasets.

In terms of the multivariate linear regression model, in our context, we consider several

sensitive response variables $y_j, j = 1, ..., m$, originating the vector of response variables $\mathbf{y} = (y_1, ..., y_m)'$, and a set of p non-*sensitive* predictors $\mathbf{x} = (x_1, ..., x_p)'$. We assume that $\mathbf{y}|\mathbf{x} \sim N_m(\mathbf{B}'\mathbf{x}, \mathbf{\Sigma})$, with $\mathbf{B}$ and $\mathbf{\Sigma}$ unknown, and the original data consist of $\mathcal{Y} = \{(y_{1i}, ..., y_{mi}, x_{1i}, ..., x_{pi}), i = 1, ..., n\}$. We write $\mathbf{Y} = (\mathbf{y_1}, ..., \mathbf{y_n})$ with $\mathbf{y_i} = (y_{1i}, ..., y_{mi})'$ and $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_n})$ with $\mathbf{x_i} = (x_{1i}, ..., x_{pi})'$. We also assume that $rank(\mathbf{X} : p \times n) = p < n$ and $n \geq m + p$. We are thus considering the following regression model

$$\mathbf{Y}_{m\times n} = \mathbf{B}'_{m\times p}\mathbf{X}_{p\times n} + \mathbb{E}_{m\times n} \tag{1}$$

where $\mathbb{E}_{m\times n}$ is distributed as $N_{mn}(\mathbf{0}, \mathbf{I_n} \otimes \mathbf{\Sigma})$. It is well known that based on the original data, $\hat{\mathbf{B}} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Y}'$ is the MLE and the UMVUE of $\mathbf{B}$, distributed as $N_{pm}(\mathbf{B}, \mathbf{\Sigma} \otimes (\mathbf{X}\mathbf{X}')^{-1})$, independent of $\hat{\mathbf{\Sigma}} = \frac{1}{n}(\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X})(\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X})'$ which is the MLE of $\mathbf{\Sigma}$, with $n\hat{\mathbf{\Sigma}} \sim W_m(\mathbf{\Sigma}, n - p)$. Therefore $\mathbf{S} = \frac{n\hat{\mathbf{\Sigma}}}{n-p}$ will be the UE of $\mathbf{\Sigma}$.

There are several tests for $\mathbf{B}$ based on the original data in the literature (Anderson, 2003). In this paper, the authors will develop a new procedure for using synthetic data to draw inference for $\mathbf{B}$, and also for $\mathbf{C} = \mathbf{AB}$ and $\mathbf{\Delta} = \boldsymbol{ABD}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, and $\boldsymbol{D}$ is an $m \times r$ matrix with $rank(\boldsymbol{D}) = r \leq k$. This procedure will be based on the statistics

$$T_O = \frac{|(\hat{\mathbf{B}} - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\hat{\mathbf{B}} - \mathbf{B})|}{|(n-p)\mathbf{S}|} \sim \prod_{i=1}^{m} \frac{p-i+1}{n-p-i+1} F_{p-i+1,n-p-i+1}, \tag{2}$$

$$T_{O,\mathbf{C}} = \frac{|(\mathbf{A}\hat{\mathbf{B}} - \mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\mathbf{B}} - \mathbf{C})|}{|(n-p)\mathbf{S}|} \sim \prod_{i=1}^{m} \frac{k-i+1}{n-p-i+1} F_{k-i+1,n-p-i+1}, \tag{3}$$

$$T_{O,\mathbf{\Delta}} = \frac{|(\mathbf{A}\hat{\mathbf{B}}\mathbf{D} - \mathbf{\Delta})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{D} - \mathbf{\Delta})|}{|(n-p)\mathbf{D}'\mathbf{S}\mathbf{D}|} \sim \prod_{i=1}^{r} \frac{k-i+1}{n-p-i+1} F_{k-i+1,n-p-i+1}. \tag{4}$$

The organization of the paper is as follows. In Section 2, based on singly imputed synthetic data generated via Plug-in Sampling, we develop inference for the regression coefficients $\mathbf{B}$.

This will be based on a pivot statistic which is different from the classical test statistics for $\mathbf{B}$ under this model (see Anderson (2003)). These statistics are shown to be non-pivotal in the case of singly imputed synthetic data generated via Plug-in Sampling. Section 3 presents two exact inference procedures based on multiply imputed synthetic data. These are contrasted with Reiter's asymptotic methodology for multiple imputation synthetic data (Reiter, 2005). In Section 4, we present some simulation results in order to check the accuracy of the theoretically derived results for the singly imputed and multiply imputed synthetic data, comparing the latter with the results obtained using Reiter's methodology. We also measure the *radius* (distance between the center and the edge) of the confidence sets for the regression coefficients $\mathbf{B}$, using with the original data, the singly imputed and the multiply imputed synthetic data. Section 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, and the results are compared with those obtained from the original data. In Section 6, properties of the proposed methodology are studied in scenarios where some of the conditions that were used to derive the methodology do not hold. Specifically, in Section 6.1 we study the scenario where the error term in the regression model is not normally distributed; in Section 6.2 we study the scenario where the imputer who creates the synthetic data and/or the data analyst underspecify or overspecify the regression model; and in Section 6.3 we study the scenario where the data analysis model is something other than the regression of the sensitive variables on the non-sensitive variables. Some concluding remarks are added in Section 7. Proofs of the theorems, and other technical derivations appear in Appendices A and B.

We conclude this section with an observation regarding the existence of *sufficient statistics*. Suppose the original data are $\mathbf{Y} \sim f_{\hat{\boldsymbol{\theta}}(\mathbf{Y})}$, and the synthetic data $\mathscr{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_M)$ are generated such that $\mathbf{V}_1, \ldots, \mathbf{V}_M | \mathbf{Y}$ are *iid* from $f_{\hat{\boldsymbol{\theta}}(\mathbf{Y})}$. 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

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

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

which implies the following result.

Lemma 1.1. *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.*

2 Analysis under Single Imputation

In this section, a likelihood-based approach for analysis of synthetic data generated from a multivariate regression model is presented based on the Plug-in Sampling method.

Consider the multivariate linear regression model (1) with $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{B}$, $\boldsymbol{\Sigma}$, $\hat{\mathbf{B}}$ and $\mathbf{S}$ defined in that same context.

The synthetic data consist of a single synthetic version of $\mathbf{Y}$ generated based on the Plug-in method as described below. From the original data $(y_{i1}, ..., y_{im}, x_{1i}, ..., x_{pi}), i = 1, ..., n$, after estimating $\mathbf{B}$ and $\boldsymbol{\Sigma}$ by $\hat{\mathbf{B}}$ and $\mathbf{S}$, respectively, we generate the synthetic data, denoted as $\mathbf{V} = (\mathbf{v_1}, ..., \mathbf{v_n})$ where $\mathbf{v_i} = (v_{1i}, ..., v_{mi})'$, are independently distributed as

$$\mathbf{v_i} | \hat{\mathbf{B}}, \mathbf{S} \sim N_m(\hat{\mathbf{B}}'\mathbf{x_i}, \mathbf{S}), i = 1, ..., n. \tag{5}$$

Our goal is to draw inference on $\mathbf{B}$ based on the synthetic data $(v_{1i}, ..., v_{mi}, x_{1i}, ..., x_{pi})$, for $i = 1, ..., n$. Towards this end, let us define

$$\mathbf{B}^* = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{V}' \quad \text{and} \quad \mathbf{S}^* = \frac{1}{n-p}(\mathbf{V} - \mathbf{B}^{*\,\prime}\mathbf{X})(\mathbf{V} - \mathbf{B}^{*\,\prime}\mathbf{X})', \tag{6}$$

which, by Lemma 1.1, are jointly sufficient for $(\mathbf{B}, \mathbf{\Sigma})$. The main *inferential* results we derive are, for $p \geq m$,:

1. the MLE of $\mathbf{B}$ is $\mathbf{B}^*$, which is unbiased for $\mathbf{B}$, with $Var(\mathbf{B}^*) = 2\mathbf{\Sigma} \otimes (\mathbf{X}\mathbf{X}')^{-1}$ (see Appendix B.3);
2. an unbiased estimator of $\mathbf{\Sigma}$ is $\mathbf{S}^*$(see Appendix B.3);
3. we prove in Theorem 2.2 (see below) that

$$T = \frac{|(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|}{|(n-p)\mathbf{S}^*|} \tag{7}$$

   is a pivotal quantity and, for $\mathbf{W} \sim W_m(\mathbf{I}, n-p)$ and $F_i \sim F_{p-i+1,n-p-i+1}$,

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

   where $\overset{st}{\sim}$ means 'stochastic equivalent to';
4. if one wants to test the significance of a set of regression coefficients or more generally, a linear combination of $\mathbf{B}$, 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 $T_{\mathbf{C}} = |(\mathbf{A}\mathbf{B}^* - \mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\mathbf{B}^* - \mathbf{C})|/|(n-p)\mathbf{S}^*|$ and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}, n-p)$ and $F_{k,i} \sim F_{k-i+1,n-k-i+1}$,

$$T_{\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{n-p-i+1} F_{k,i} \right\} \left| (n-p)\mathbf{W}^{-1} + \mathbf{I}_m \right|; \tag{8}$$

   (i) *Test for the significance of* $\mathbf{C}$*:* In order to test $H_0 : \mathbf{C} = \mathbf{C}_0$ versus $H_1 : \mathbf{C} \neq \mathbf{C}_0$, we reject H_0 whenever $T_{\mathbf{C}_0}$ exceeds $\delta_{k,m,p,n;\gamma}$ where $\delta_{k,m,p,n;\gamma}$ satisfies $(1-\gamma) = Pr(T_{\mathbf{C}_0} \leq \delta_{k,m,p,n;\gamma})$ when H_0 is true; In particular, a test for $\mathbf{B} = \mathbf{B_0}$ follows upon taking $\mathbf{A} = \mathbf{I_p}$;

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

$$\Delta(\mathbf{C}) = \{\mathbf{C} : T_{\mathbf{C}} \leq \delta_{k,m,n,p;\gamma}\}, \tag{9}$$

where the value of $\delta_{k,m,n,p;\gamma}$ can be obtained by simulating the distribution of $T_{\mathbf{C}}$ using the distribution in (8);

5. to infer about $\mathbf{\Delta} : k \times r = \mathbf{ABD}$ where $\mathbf{A} : k \times p$, $\mathbf{B} : p \times m$, $\mathbf{D} : m \times r$ with $r \leq k$, we start from its natural point estimator $\mathbf{\Delta}^* = \mathbf{AB}^*\mathbf{D}$ and propose to use the pivotal quantity
$$T_{\mathbf{\Delta}} = \frac{|(\mathbf{\Delta}^* - \mathbf{\Delta})'[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}^* - \mathbf{\Delta})|}{|(n-p)\mathbf{D}'\mathbf{S}^*\mathbf{D}|},$$
whose distribution is given by
$$T_{\mathbf{\Delta}} \overset{st}{\sim} \left\{ \prod_{i=1}^{r} \frac{k-i+1}{n-p-i+1} F_i \right\} \frac{|\mathbf{W}^* + (n-p)\mathbf{I}_r|}{|\mathbf{W}^*|}$$
where $F_i \sim F_{k-i+1,n-p-i+1}$ and $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, all independently. Taking $r = 1$ and $k = 1$, and making $\mathbf{A} : 1 \times p$ a matrix of zeros except for $\mathbf{A}_{1,i} = 1$, and $\mathbf{D} : m \times 1$ a matrix of zeros except for $\mathbf{D}_{j,1} = 1$, for $i = 1, ..., p$ and $j = 1, ..., p$ we may observe that
$$T_{\mathbf{\Delta}} = T_{\mathbf{B}_{(i,j)}} = \frac{(\mathbf{B}^*_{(i,j)} - \mathbf{B}_{(i,j)})[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\mathbf{B}^*_{(i,j)} - \mathbf{B}_{(i,j)})}{(n-p)\mathbf{D}'\mathbf{S}^*\mathbf{D}}$$
therefore concluding that the $(1-\alpha)$ confidence interval for $\mathbf{B}_{(i,j)}$ will be given by
$$\mathbf{B}^*_{(i,j)} \pm \sqrt{q^*_{1-\alpha}(n-p)\mathbf{D}'\mathbf{S}^*\mathbf{DA}(\mathbf{XX}')^{-1}\mathbf{A}'}$$
where in fact, $\mathbf{D}'\mathbf{S}^*\mathbf{D} = \mathbf{S}^*_{(j,j)}$ and $\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}' = (\mathbf{XX}')^{-1}_{(i,i)}$, with $q^*_{1-\alpha}$ being the value of the $1-\alpha$ cut-off point of the distribution of $T_{\mathbf{\Delta}}$, for $i = 1, .., p$ and $j = 1, ..., m$.

Results in 1-5 are derived based on Theorems 2.1 and 2.2.

Theorem 2.1. *The joint pdf of* $(\mathbf{B}^*, \mathbf{S}^*)$ *defined in (6) is proportional to*

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

Proof. See Appendix A □

Theorem 2.2. *The distribution of* T *defined in (7) can be obtained from the decomposition*

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

where $F_i \sim F_{p-i+1,n-p-i+1}$ *and* $\mathbf{W} \sim W_m(I, n-p)$.

Proof. See Appendix A □

Remark 2.1. *When* $m=1$*, the statistic* T *in (7) reduces to the statistic* T^2 *used in (Klein and Sinha, 2015a) which has a pdf obtained from the fact that*

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

Remark 2.2. *In Table 1, we list the simulated* γ *cut-off points* $d_{m,p,n;\gamma}$ *for* T *for some values of* p*,* m *and* n *for* $\gamma = 0.05$.

Table 1: Cut-off points of the 95% confidence set for the regression coefficient $\mathbf{B}$

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

In our context, one could suggest the following adaptations of the classical test criteria for the multivariate regression model (see Anderson (2003) for the classical test criteria)

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

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

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

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

but, unfortunately, these test statistics are non-pivotal statistics (see Appendix B.2).

3 Analysis under Multiple Imputation

In this section we first provide two new and exact inferential procedures based on the likelihood principle for multiple imputation synthetic data and then clarify Reiter's approximate analysis for our setup.

3.1 A First New Procedure

In this subsection, we present a likelihood-based approach for exact inference about $\mathbf{B}$ in case one has access to multiple released synthetic data sets. Let us consider that we generate $M(>1)$ i.i.d. synthetic data sets $\mathbf{V_1}, ..., \mathbf{V_M}$. Let $\mathbf{B_i^*} = (\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{V_i'}$ and $\mathbf{S_i^*} = \frac{1}{n-p}(\mathbf{V_i} - \mathbf{B_i'}^*\mathbf{X})(\mathbf{V_i} - \mathbf{B_i'}^*\mathbf{X})'$ be the estimators of $\mathbf{B}$ and $\mathbf{\Sigma}$ based on $\mathbf{V}_i$. Conditionally on $(\hat{\mathbf{B}}, \mathbf{S})$, for any $i = 1, ..., M$, $\mathbf{B_i^*}$ is independent of $\mathbf{S_i^*}$ and $\{(\mathbf{B}_1^*, \mathbf{S}_1^*), \ldots, (\mathbf{B}_M^*, \mathbf{S}_M^*)\}$ are jointly sufficient estimators for $\mathbf{B}$ and $\mathbf{\Sigma}$. Let us also define

$$\overline{\mathbf{B}}_M^* = \frac{1}{M}\sum_{i=1}^{M}\mathbf{B_i^*} \quad \text{and} \quad \overline{\mathbf{S}}_M^* = \frac{1}{M}\sum_{i=1}^{M}\mathbf{S_i^*}, \tag{10}$$

which are mutually independent, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$. Analogous to what was done in Section 2, one can derive the following inferential results, for $p \geq m$:

1. $\overline{\mathbf{B}}_M^*$ is an unbiased estimator for $\mathbf{B}$, with $Var(\overline{\mathbf{B}}_M^*) = (M+1)/M \times \mathbf{\Sigma} \otimes (\mathbf{X}\mathbf{X}')^{-1}$ (see Appendix B.3);

2. an unbiased estimator of $\mathbf{\Sigma}$ is $\overline{\mathbf{S}}^*_M$(see Appendix B.3);
3. we prove in Corollary 3.2 (see below) that $T_M = \left|(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{XX}')(\overline{\mathbf{B}}^*_M - \mathbf{B})\right| / \left|(n-p)\overline{\mathbf{S}}^*_M\right|$ is a pivotal quantity and, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_i \sim F_{p-i+1, M(n-p)-i+1}$,

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

4. if one wants to test the significance of a set of regression coefficients or more generally of a linear combination of these regression coefficients, $\mathbf{AB} = \mathbf{C}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, one may define $T_{M,\mathbf{C}} = |(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})|/|(n-p)\overline{\mathbf{S}}^*_M|$ and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{k,i} \sim F_{k-i+1, M(n-p)-i+1}$,

$$T_{M,\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{\prod_{i=1}^{m} \frac{k-i+1}{M(n-p)-i+1} F_{k,i}\right\} |M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m|; \tag{11}$$

(i) *Test for the significance of* $\mathbf{C}$*:* In order to test $H_0 : \mathbf{C} = \mathbf{C}_0$ versus $H_1 : \mathbf{C} \neq \mathbf{C}_0$, we reject H_0 whenever $T_{M,\mathbf{C}_0}$ exceeds $\delta_{M,k,m,p,n;\gamma}$ where $\delta_{M,k,m,p,n;\gamma}$ satisfies $(1-\gamma) = Pr(T_{M,\mathbf{C}_0} \leq \delta_{M,k,m,p,n;\gamma})$ when H_0 is true; in particular a test for $\mathbf{B} = \mathbf{B_0}$ follows upon taking $\mathbf{A} = \mathbf{I_p}$;

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

$$\Delta(\mathbf{C}) = \{\mathbf{C} : T_{M,\mathbf{C}} \leq \delta_{M,k,m,n,p;\gamma}\}; \tag{12}$$

the value of $\delta_{M,k,m,n,p;\gamma}$ above can be obtained by simulating the distribution of $T_{\mathbf{C}}$, by first generating $\mathbf{W} \sim W_m(I, n-p)$ and then generating

$$T_{M,\mathbf{C}}|\mathbf{W} \sim \left\{\prod_{i=1}^{m} \frac{k-i+1}{M(n-p)-i+1} F_{k-i+1, M(n-p)-i+1}\right\} \left|M(n-p)\mathbf{W}^{-1} + I_m\right|; \tag{13}$$

5. to infer about $\mathbf{\Delta} : k \times r = \mathbf{ABD}$ where $\mathbf{A} : k \times p$, $\mathbf{B} : p \times m$, $\mathbf{D} : m \times r$ with $r \leq k$, we start from its natural point estimator $\mathbf{\Delta}^* = \mathbf{AB}^*\mathbf{D}$ and propose to use the pivotal quantity

$$T_{M,\mathbf{\Delta}} = \frac{|(\mathbf{\Delta}_M^* - \mathbf{\Delta})'[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}_M^* - \mathbf{\Delta})|}{|(n-p)\mathbf{D}'\overline{\mathbf{S}}_M^*\mathbf{D}|},$$

whose distribution is given by

$$T_{M,\mathbf{\Delta}} \overset{st}{\sim} \left\{ \prod_{i=1}^{r} \frac{k-i+1}{M(n-p)-i+1} F_i \right\} \frac{|\mathbf{W}^* + M(n-p)\mathbf{I}_r|}{|\mathbf{W}^*|}$$

where $F_i \sim F_{k-i+1,n-p-i+1}$ and $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, all independently. Taking $r = 1$ and $k = 1$, and making $\mathbf{A} : 1 \times p$ a matrix of zeros except for $\mathbf{A}_{1,i} = 1$, and $\mathbf{D} : m \times 1$ a matrix of zeros except for $\mathbf{D}_{j,1} = 1$, for $i = 1, ..., p$ and $j = 1, ..., m$ we may observe that the $(1-\alpha)$ confidence interval for $\mathbf{B}_{(i,j)}$ will be given by

$$\overline{\mathbf{B}}_{M(i,j)}^* \pm \sqrt{q_{M,1-\alpha}^*(n-p)\overline{\mathbf{S}}_{M(j,j)}^*(\mathbf{XX}')_{(i,i)}^{-1}}.$$

The above results are derived based on the facts that $\overline{\mathbf{B}}_M^*|_{\hat{\mathbf{B}},\mathbf{S}} \sim N_{pm}(\hat{\mathbf{B}}, \frac{1}{M}\mathbf{S} \otimes (\mathbf{XX}')^{-1})$ and $M(n-p)\overline{\mathbf{S}}_M^*|_{\mathbf{S}} \sim W_m(\mathbf{S}, M(n-p))$, and based on the following two Corollaries, of Theorems 2.1 and 2.2, whose proofs are provided in Appendix A.

Corollary 3.1. *The joint pdf of* $(\overline{\mathbf{B}}_M^*, \overline{\mathbf{S}}_M^*)$ *defined in (10) is proportional to*

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

Corollary 3.2. *The pdf of* T_M *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* $F_i \sim F_{p-i+1,M(n-p)-i+1}$.

3.2 A Second New Procedure

Noting that it will be possible to gather more information from the released synthetic data we propose, in this sub-section, another likelihood-based approach for exact inference about $\mathbf{B}$. Let us start to recall that $\mathbf{V}$ is a $m \times n$ matrix constituted with the vectors $(\mathbf{v}_1, ..., \mathbf{v}_n)$, generated from $\mathbf{v_i}|\hat{\mathbf{B}}, \mathbf{S} \sim N_m[\hat{\mathbf{B}}'\mathbf{x_i}, \mathbf{S}], i = 1, ..., n$. Having access to the M imputations $\mathbf{V}_1, ..., \mathbf{V}_M$, with $\mathbf{V}_j = (\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)})$, $j = 1, ..., M$, and note that, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$, $(\mathbf{v}_i^{(1)}, ..., \mathbf{v}_i^{(M)})$ is a random sample from $N_m[\hat{\mathbf{B}}'\mathbf{x_i}, \mathbf{S}]$, for any $i = 1, ..., n$. Let us consider $\overline{\mathbf{v}}_i = \frac{1}{M}\sum_{j=1}^{M} \mathbf{v}_i^{(j)}$ and $\mathbf{S}_{\mathbf{v}i} = \sum_{j=1}^{M}(\mathbf{v}_i^{(j)} - \overline{\mathbf{v}}_i)(\mathbf{v}_i^{(j)} - \overline{\mathbf{v}}_i)'$ which are the sufficient statistics for $\mathbf{\Sigma}$, based on the i-th covariate vector. Defining $\mathbf{S_v} = \sum_{i=1}^{n} \mathbf{S}_{\mathbf{v}i}$, we have $(\overline{\mathbf{v}}_1, ..., \overline{\mathbf{v}}_n, \mathbf{S_v})$ as the joint sufficient statistics for $(\mathbf{B}, \mathbf{\Sigma})$. Conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$, we have $\overline{\mathbf{v}}_i \sim N_m(\hat{\mathbf{B}}'\mathbf{x_i}, \frac{1}{M}\mathbf{S})$ and $\mathbf{S_v} \sim W_m(\mathbf{S}, n(M-1))$ since $\mathbf{S}_{\mathbf{v}i} \sim W_m(\mathbf{S}, M-1)$.

From the M released synthetic data matrices $\mathbf{V}_j, j = 1, ..., M$, we may define $\overline{\mathbf{V}}_M = \frac{1}{M}\sum_{j=1}^{M} \mathbf{V}_j$ and then define for $\mathbf{B}$ the estimator

$$\overline{\mathbf{B}}_M^* = (\mathbf{XX}')^{-\mathbf{1}}\mathbf{X}\overline{\mathbf{V}}_M', \tag{14}$$

which ends up being the same estimator defined in subsection 3.1. Furthermore, we may obtained additional information about $\mathbf{\Sigma}$ as $\mathbf{S}_{mean} = (\overline{\mathbf{V}}_M - \overline{\mathbf{B}}_M^{*'}\mathbf{X})'(\overline{\mathbf{V}}_M - \overline{\mathbf{B}}_M^{*'}\mathbf{X})$ which can be combined with the previous estimator $\mathbf{S_v}$ as

$$\mathbf{S}_{comb} = \frac{\mathbf{S_v} + M \times \mathbf{S}_{mean}}{Mn - p}. \tag{15}$$

Analogous to what was done in Section 2, one can derive the following inferential results, for $p \geq m$:

1. an unbiased estimator of $\mathbf{\Sigma}$ is $\mathbf{S}_{comb}$ (see Appendix B.4);
2. we prove in Corollary 3.4 (see below) that

$$T_{comb} = \frac{\left|(\overline{\mathbf{B}}_M^* - \mathbf{B})'(\mathbf{XX}')(\overline{\mathbf{B}}_M^* - \mathbf{B})\right|}{\left|(n - \frac{p}{M})\mathbf{S}_{comb}\right|}, \tag{16}$$

is a pivotal quantity and that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_i \sim F_{p-i+1, Mn-p-i+1}$,

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

3. if one wants to test the significance of a set of regression coefficients or more generally, a linear combination of $\mathbf{B}$, 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 $T_{comb,\mathbf{C}} = |(\mathbf{A}\overline{\mathbf{B}}_M^* - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}_M^* - \mathbf{C})|/|(n - \frac{p}{M})\mathbf{S}_{comb}|$ and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{k,i} \sim F_{k-i+1, Mn-p-i+1}$,

$$T_{comb,\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{Mn-p-i+1} F_{k,i} \right\} \left| M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m \right|; \tag{17}$$

(i) *Test for the significance of* $\mathbf{C}$*:* In order to test $H_0 : \mathbf{C} = \mathbf{C}_0$ versus $H_1 : \mathbf{C} \neq \mathbf{C}_0$, we reject H_0 whenever $T_{comb,\mathbf{C}_0}$ exceeds $\omega_{M,k,m,p,n;\gamma}$ where $\omega_{M,k,m,p,n;\gamma}$ satisfies $(1-\gamma) = Pr(T_{comb,\mathbf{C}_0} \leq \omega_{M,k,m,p,n;\gamma})$ when H_0 is true; in particular a test for $\mathbf{B} = \mathbf{B_0}$ follows upon taking $\mathbf{A} = \mathbf{I_p}$;

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

$$\Delta(\mathbf{C}) = \{\mathbf{C} : T_{comb,\mathbf{C}} \leq \omega_{M,k,m,n,p;\gamma}\}, \tag{18}$$

where the value of $\omega_{M,k,m,n,p;\gamma}$ may be obtained by simulating the distribution of $T_{comb,\mathbf{C}}$, by first generating $\mathbf{W} \sim W_m(I, n-p)$ and then generating

$$T_{comb,\mathbf{C}}|\mathbf{W} \sim \left\{ \prod_{i=1}^{m} \frac{k-i+1}{Mn-p-i+1} F_{k-i+1, Mn-p-i+1} \right\} \left| M(n-p)W^{-1} + I_m \right|.$$

4. to infer about $\mathbf{\Delta} : k \times r = \mathbf{ABD}$ where $\mathbf{A} : k \times p$, $\mathbf{B} : p \times m$, $\mathbf{D} : m \times r$ with $r \leq k$, we start from its natural point estimator $\mathbf{\Delta}^* = \mathbf{AB}^*\mathbf{D}$ and propose to use the pivotal quantity

$$T_{comb,\mathbf{\Delta}} = \frac{|(\mathbf{\Delta}_M^* - \mathbf{\Delta})'[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}_M^* - \mathbf{\Delta})|}{|(n - \frac{p}{M})\mathbf{D}'\mathbf{S}_{comb}^*\mathbf{D}|},$$

whose distribution is given by

$$T_{comb,\mathbf{\Delta}} \overset{st}{\sim} \left\{ \prod_{i=1}^{r} \frac{k-i+1}{Mn-p-i+1} F_i \right\} \frac{|\mathbf{W}^* + M(n-p)\mathbf{I}_r|}{|\mathbf{W}^*|},$$

where $F_i \sim F_{k-i+1,n-p-i+1}$ and $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, all independently. Taking $r = 1$ and $k = 1$, and making $\mathbf{A} : 1 \times p$ a matrix of zeros except for $\mathbf{A}_{1,i} = 1$, and $\mathbf{D} : m \times 1$ a matrix of zeros except for $\mathbf{D}_{j,1} = 1$, for $i = 1, ..., p$ and $j = 1, ..., m$ we may observe that the $(1-\alpha)$ confidence interval for $\mathbf{B}_{(i,j)}$ will be given by

$$\overline{\mathbf{B}}^*_{M(i,j)} \pm \sqrt{q^*_{comb,1-\alpha}\left(n - \frac{p}{M}\right) \mathbf{S}^*_{comb,(j,j)} (\mathbf{X}\mathbf{X}')^{-1}_{(i,i)}}.$$

The above results are derived based on the observation 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 $(Mn-p)\mathbf{S}_{comb}|_{\mathbf{S}} \sim W_m(\mathbf{S}, Mn-p)$, noting that $\mathbf{S}_{mean}|_{\mathbf{S}} \sim W_m(\frac{\mathbf{S}}{M}, n-p)$, and based on the following two Corollaries of Theorems 2.1 and 2.2, whose proofs are provided in Appendix A.

Corollary 3.3. *The joint pdf of* $(\overline{\mathbf{B}}^*_M, \mathbf{S}_{comb})$ *defined in (14) and (15) is proportional to*

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

Corollary 3.4. *The pdf of* T_{comb} *defined in (16) 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* $F_i \sim F_{p-i+1,Mn-p-i+1}$.

Remark 3.1. *It is the case that* $Var(\overline{\boldsymbol{S}}^*_M) > Var(\boldsymbol{S}_{comb})$*, and therefore the second new procedure is expected to outperform the first new procedure.*

Proof. Recalling the definition of $\overline{\boldsymbol{S}}^*_M$*, we get* $M(n-p)\overline{\boldsymbol{S}}^*_M|\boldsymbol{S} \sim W_m(\boldsymbol{S}, M(n-p))$ *and* $(n-p)\boldsymbol{S} \sim W_m(\boldsymbol{\Sigma}, n-p)$*. We can write* $Var(\overline{\boldsymbol{S}}^*_M) = E[Var(\overline{\boldsymbol{S}}^*_M)|\boldsymbol{S}] + Var[E(\overline{\boldsymbol{S}}^*_M)|\boldsymbol{S}] = E[\frac{Var(W_m(\boldsymbol{S}, M(n-p)))}{M^2(n-p)^2}] + Var[\boldsymbol{S}]$*...(*)*

Likewise, we can write $Var(\boldsymbol{S}_{comb}) = E[Var(\boldsymbol{S}_{comb})|\boldsymbol{S}] + Var[E(\boldsymbol{S})_{comb}|\boldsymbol{S}] = E[\frac{Var(W_m(\boldsymbol{S}, Mn-p))}{(Mn-p)^2}] + Var(\boldsymbol{S})$*...(**)*

To compare (*) and (**), obviously the 2nd terms are identical, and we show that the first term in (*) is bigger than that in (**). We now use a very general result (Muirhead, 1982; page 90) that for any Wishart matrix $W_p(\Sigma, \nu)$*,* $Var[vec(W)] = \nu[\mathbf{I}_{p^2} + \mathbf{K}][\boldsymbol{\Sigma} \times \boldsymbol{\Sigma}]$ *where* $\mathbf{K}$ *is a matrix of constants of order* $p^2 \times p^2$*. Using the fact that* $\nu = M(n-p)$ *in (*) and* $\nu = Mn - p$ *in (**), and rest of the expression in Var(.) is the same for both, it readily follows that* $(*) > (**)$*.*

3.3 Reiter's (2005) Methodology Under Multiple Imputation

Now we present a review of the methodology of Reiter (2005) for drawing inference based on multiply synthetic data for a vector valued parameter and adapt it to our model. Originally developed for synthetic data generated by Posterior Predictive Sampling, but Reiter and Kinney (2012) indicate that it is also valid for synthetic data generated via Plug-in Sampling.

In order to be possible to use Reiter's methodology to estimate $\mathbf{B}$, a $p \times m$ dimensional matrix parameter, we consider $\mathbf{vec}(\mathbf{B}) = (\mathbf{B}'_1\ \mathbf{B}'_2\ ...\ \mathbf{B}'_m)'$, a $pm \times 1$ vector parameter. Based on the original data, $\mathbf{vec}(\hat{\mathbf{B}})$ is an estimator of $\mathbf{vec}(\mathbf{B})$ and its covariance matrix estimator is $\mathbf{U} = \mathbf{S} \otimes (\mathbf{X}\mathbf{X}')^{-1}$ a $pm \times pm$ matrix. Generating M synthetic data sets instead of just one we end up with $\mathbf{V}_1, ..., \mathbf{V}_M$ as the synthetic data sets. Let $\mathbf{vec}(\mathbf{B}^*_i) = \mathbf{vec}((\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\mathbf{V}'_i)$ and $\mathbf{U}_i = \mathbf{S}^*_i \otimes (\mathbf{X}\mathbf{X}')^{-1}$, where $\mathbf{S}^*_i = \frac{1}{n-p}(\mathbf{V}_i - \mathbf{B}^{*'}_i\mathbf{X})(\mathbf{V}_i - \mathbf{B}^{*'}_i\mathbf{X})'$. Note that based on $\mathbf{V}_i$, conditionally in $\hat{\mathbf{B}}$ and $\mathbf{S}$, $\mathbf{vec}(\mathbf{B}^*_i)$ is an unbiased estimator of $vec(\mathbf{B})$ and $\mathbf{U}_i$ is an

unbiased estimator of its variance. Then the following estimators

$$\mathbf{vec}(\overline{\mathbf{B}}^*_M) = \frac{1}{M}\sum_{i=1}^{M}\mathbf{vec}(\mathbf{B}^*_i), \qquad \overline{\mathbf{U}}_M = \frac{1}{M}\sum_{i=1}^{M}\mathbf{U}_i,$$

$$\mathbf{b}_M = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{vec}(\mathbf{B}^*_i) - \mathbf{vec}(\overline{\mathbf{B}}^*_M))(\mathbf{vec}(\mathbf{B}^*_i) - \mathbf{vec}(\overline{\mathbf{B}}^*_M))'$$

should be Reiter's estimators to be used to draw inference about $\mathbf{B}$, where $\mathbf{vec}(\overline{\mathbf{B}}^*_M)$ 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 the statistic

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

where $r = \frac{tr(\mathbf{b_M}\overline{\mathbf{U}}_\mathbf{M}^{-1})}{Mpm}$. The distribution of $T_{R,M}$ is approximated by an $F_{pm,w(r)}$ distribution where $w(r) = 4 + [pm(M-1) - 4]\left[1 + 1/r - 2(rpm)^{-1}(M-1)^{-1}\right]^2$.

4 Simulation Studies

In this section we present results of some simulations. The objectives of these simulations are to demonstrate that (i) the inference methods used in Sections 2 and 3 perform as we predicted for our proposed methodology for singly and multiply imputed synthetic data generated via Plug-in Sampling, and (ii) to compare the accuracy of our proposed methodology with the accuracy of Reiter (2005) methodology for multiply imputed partially synthetic data. All simulations were carried out using the software Mathematica®. To conduct the simulation, we take the population distribution as a multivariate normal distribution with expected value given by the right hand side of (1), with matrix of regressor coefficients $\mathbf{B}$, and covariance matrix $\mathbf{\Sigma}$, given by

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

for $m = 2$ and $p = 3$. The regressor variables $x_{1i}, x_{2i}, x_{3i}, i = 1, ..., n$ are generated as iid $N(1, 1)$ and held fixed for the entire simulation. Based on a Monte Carlo simulation with 10^5 iterations, we compute an estimate of the coverage probability (percentage of observed values of the statistics smaller than the respective theoretical cut-off points) of the following confidence regions, where in all cases, the level of the confidence region is set to 0.95:

1. the confidence sets for $\mathbf{B}$ and for $\mathbf{C} = \mathbf{AB}$, given by (9), respectively with $\mathbf{A} = \mathbf{I}_3$ and $\mathbf{A} = (\ \mathbf{0}_{2\times 1} |\ \mathbf{I}_2)$, based on a single synthetic dataset generated as in (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 2;
2. for the two new procedures in subsections 3.1 and 3.2, based on multiple synthetic data, the confidence sets for $\mathbf{B}$ and for $\mathbf{C} = \mathbf{AB}$, given by (12) and (18), respectively with $\mathbf{A} = \mathbf{I}_3$ and $\mathbf{A} = (\ \mathbf{0}_{2\times 1} |\ \mathbf{I}_2)$, using the methodology described in the two subsections referred above; for $M = 2, 5, 10, 20$ synthetic datasets, the estimated coverage probabilities of the confidence sets are shown in Table 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;
3. the confidence set for $\mathbf{B}$ obtained using the methodology of Reiter (2005), for $M(> 1)$ synthetic datasets, in subsection 3.3; for each of the cases $M = 2, 5, 10, 20$, the estimated coverage probabilities of the confidence sets are shown in Table 2 under the column $\mathbf{vec}(\mathbf{B})$.

The results reported in Table 2 for sample sizes n=10, 20, 50, 100, 200, show that, based on singly imputed and multiply imputed synthetic data, the 0.95 confidence sets for $\mathbf{B}$ and $\mathbf{AB}$ have an estimated coverage probability approximately equal to 0.95, confirming that the confidence sets perform as predicted. As supported by the theory, using the adapted methodology of Reiter (2005) for multiply imputed partially synthetic data the estimated coverage probabilities fall short of the stipulated level of 0.95 for very small sample sizes,

Table 2: Average coverage for **B** and **AB**

	$M=1$	$M=2$			$M=5$			$M=10$			$M=20$		
n		**vec**(**B**)	**B**(1)	**B**(2)	**vec**(**B**)	**B**(1)	**B**(2)	**vec**(**B**)	**B**(1)	**B**(2)	**vec**(**B**)	**B**(1)	**B**(2)
10	0.951	0.830	0.950	0.950	0.754	0.949	0.947	0.748	0.950	0.949	0.753	0.948	0.947
20	0.953	0.919	0.955	0.953	0.874	0.947	0.948	0.867	0.948	0.949	0.870	0.950	0.950
50	0.953	0.955	0.950	0.951	0.924	0.949	0.948	0.921	0.949	0.948	0.924	0.949	0.948
100	0.946	0.957	0.946	0.947	0.934	0.946	0.946	0.931	0.948	0.948	0.935	0.948	0.948
200	0.949	0.964	0.953	0.952	0.943	0.950	0.951	0.943	0.949	0.949	0.944	0.949	0.950

(a) Average coverage for **B**

	$M=1$	$M=2$			$M=5$			$M=10$			$M=20$		
n		**vec**	(1)	(2)	**vec**	(1)	(2)	**vec**	(1)	(2)	**vec**	(1)	(2)
10	0.950	0.968	0.949	0.950	0.791	0.947	0.946	0.799	0.948	0.946	0.791	0.944	0.947
20	0.952	0.994	0.950	0.951	0.888	0.948	0.947	0.891	0.949	0.949	0.888	0.950	0.950
50	0.954	0.999	0.953	0.954	0.931	0.950	0.949	0.927	0.948	0.946	0.928	0.949	0.949
100	0.946	1.000	0.946	0.948	0.940	0.948	0.948	0.937	0.949	0.950	0.939	0.948	0.948
200	0.951	1.000	0.953	0.952	0.946	0.949	0.950	0.948	0.951	0.949	0.948	0.950	0.949

(b) Average coverage for **AB**

but quickly attain the desired level even for moderate sample sizes for the cases where $M \geq 5$.

In order to measure the *radius* (distance between the center and the edge) of the confidence sets, we propose the measure

$$\Upsilon_M = d_{M,m,n,p,\gamma} \times |(n-p)\tilde{\mathbf{S}}_M|,$$

where $d_{M,m,n,p,\gamma}$ is the cut-off point, where we take $M = 0$ for the original data with $\tilde{\mathbf{S}}_0 = \mathbf{S}$, $M = 1$ for singly imputed synthetic data with $\tilde{\mathbf{S}}_1 = \mathbf{S}^*$, and $M = 2, 5, 10, 20$ for multiply imputed synthetic data with $\tilde{\mathbf{S}}_M = \overline{\mathbf{S}}^*_M$ for the first new procedure, and $\tilde{\mathbf{S}}_M = (n - \frac{p}{M})/(n-p)\mathbf{S}_{comb}$ for the second new procedure. The expected value of this measure will be

$$E(\Upsilon_M) = d_{M,m,n,p,\gamma} \times \frac{(n-p)!}{(n-p-m)!} \times K_{M,n,p,m}|\mathbf{\Sigma}|$$

where $K_{0,n,p,m} = 1$ for the original data and, for $M \geq 1$,

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

for the first new procedure and

$$K_{M,n,p,m} = \frac{1}{M^m (n-p)^m} \frac{(Mn - p)!}{(Mn - p - m)!}$$

for the second new procedure, where when $M = 1$ it will refer the single imputed synthetic data. For more details about these expected values see Appendix B.5.

For $M = 0, 1, 2, 5, 10, 20$ and sample sizes $n = 10, 20, 50, 100, 200$, we present the average (avg) of simulated values of Υ_M and its expected value (exp) $E(\Upsilon_M)$ for **B** in Table 3 and for **C** in Table 4.

Table 3: Average values of Υ_M and the values of $E(\Upsilon_M)$ for the confidence set for **B**.

n	Original	$M = 1$		$M = 2$				$M = 5$			
		avg	exp	1st Approach		2nd Approach		1st Approach		2nd Approach	
				avg	exp	avg	exp	avg	exp	avg	exp
10	37.55	214.29	219.96	90.62	92.84	85.74	87.56	54.83	55.28	51.61	51.94
20	22.76	102.91	103.29	53.50	53.79	52.20	52.47	32.58	32.69	32.50	32.57
50	18.73	78.10	77.43	42.58	42.34	43.13	42.86	27.34	27.17	27.11	26.94
100	17.74	73.44	73.34	40.11	40.09	40.27	40.24	25.28	25.27	25.54	25.53
200	17.41	69.81	69.72	39.07	39.01	38.62	38.56	25.09	25.07	25.08	25.06

n	$M = 10$				$M = 20$			
	1st Approach		2nd Approach		1st Approach		2nd Approach	
	avg	exp	avg	exp	avg	exp	avg	exp
10	44.57	44.78	43.86	43.95	41.04	41.08	40.24	40.27
20	27.06	27.13	27.56	27.61	25.25	25.31	25.03	25.08
50	22.98	22.82	22.66	22.50	20.90	20.77	20.66	20.53
100	21.51	21.51	21.61	21.61	19.50	19.50	19.29	19.28
200	20.78	20.77	20.79	20.78	18.94	18.94	19.15	19.14

Table 4: Average values of Υ_M and the values of $E(\Upsilon_M)$ for the confidence set for **AB**.

n	Original	$M=1$		$M=2$				$M=5$			
		avg	exp	1st Approach		2nd Approach		1st Approach		2nd Approach	
				avg	exp	avg	exp	avg	exp	avg	exp
10	13.36	72.57	74.49	32.62	33.42	30.61	31.26	19.56	19.72	19.00	19.12
20	8.66	39.00	39.15	20.12	20.23	20.17	20.27	12.71	12.75	12.45	12.48
50	7.44	31.23	30.97	16.67	16.58	16.79	16.69	10.70	10.63	10.55	10.48
100	7.14	28.30	28.27	15.89	15.88	16.12	16.11	10.07	10.07	10.10	10.10
200	6.92	27.76	27.72	15.80	15.78	15.56	15.54	9.72	9.72	9.85	9.84

n	$M=10$				$M=20$			
	1st Approach		2nd Approach		1st Approach		2nd Approach	
	avg	exp	avg	exp	avg	exp	avg	exp
10	16.07	16.15	15.61	15.64	14.44	14.45	14.49	14.50
20	10.38	10.41	10.41	10.43	9.55	9.58	9.54	9.56
50	8.94	8.88	8.76	8.70	8.25	8.20	8.10	8.04
100	8.42	8.43	8.63	8.63	7.68	7.68	7.77	7.77
200	8.40	8.40	8.16	8.16	7.64	7.63	7.56	7.56

Observing Tables 3 and 4, we conclude that as the number M of released synthetic datasets increases, Υ_M decreases and eventually coincides with the value of Υ_0, indeed as expected, since as M increases, the amount of information about the original data released increases, leading us closer to the inference drawn from the original data . We also observe that the values of $\Upsilon_M, M > 1$, for both procedures become identical for larger sample sizes.

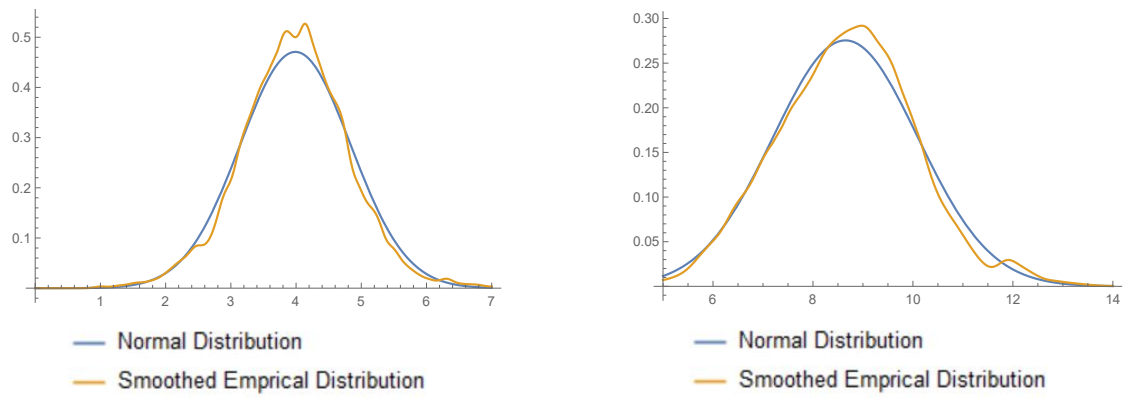
5 An Application Using Current Population Survey Data

In this section we provide an application based on some real data and compare the inference based on the original data with the inference based on the synthetic data, according the procedures developed in sections 2 and 3 and also the method of Reiter (2005). The data are from the U.S. 2000 Current Population Survey (CPS) March supplement, available online from

http://www.census.gov.cps/. We will only focus on the household records. The full data has seventeen variables measured on 51,016 heads of households and it includes the variables age, race, sex and marital status as key identifiers and a mix of other categorical and numerical variables. For the vector $\mathbf{y}$ of response sensitive variables, we have selected two numerical variables, namely, *total household income*(I) and *household property tax*(PT). After deleting all entries where at least one of these variables are reported as 0, we were left with a sample size of 32923. The example addressed below, based on a sample of size $n = 32923$, using the proposed exact methods developed in Section 2 and Subsections 3.1 and 3.2, illustrates the capabilities of this methods. Moreover, in situations where the number of response sensitive variables is larger that the number of non-sensitive predictors being analyzed, only the procedure of Reiter (2005) can be used. We will use the assumption of the normality of the fifth root of the response variables, and as such we will use the fifth root of the original variables. As we may observe in Figure 1, the marginal distribution of the transformed variables is approximately normal.

We proceed as if these $n = 32923$ households are a random sample, and that these two variables are confidential. We treat these public use data as the original data. Although in the data file a large number of variables is available, we will only use the following set of covariates:

N: number of people in household;



(a) household property tax (b) total household income

Figure 1: Smothed Empirical distributions of variable responses PT and I

L: number of people in the household who are less than 18 years old;

A: age for the head of household;

E: education level for the head of the household(coded to take values 31, 34-37, 39-46);

M: marital status for the head of the household (coded to take values 1,3-7);

R: race of the head of the household (coded to take values 1,2,4);

S: sex of the head of the household (coded to take values 1,2).

We refer to the Current Population Survey March 2000 technical documentation (available at http://www.census.gov/prod/techdoc/cps/cpsmar00.pdf) and Klein and Sinha (2015a) for details.

As such, in this application, $\mathbf{x}$, the vector of regressor variables, is defined as

$$\mathbf{x} = \Big(1, \mathrm{N}, \mathrm{L}, \mathrm{A}, \mathrm{I}(\mathrm{E}=32), ..., \mathrm{I}(\mathrm{E}=46), \mathrm{I}(\mathrm{M}=2), ..., \mathrm{I}(\mathrm{M}=7), \\ \mathrm{I}(\mathrm{R}=2), ..., \mathrm{I}(\mathrm{R}=4), \mathrm{I}(\mathrm{S}=2)\Big)', \tag{19}$$

where $\mathrm{I}(\mathrm{E}=31)$ is the indicator variable for $\mathrm{E}=31$, i.e. for individuals that have completed less than 1st grade, $\mathrm{I}(\mathrm{E}=32)$ is the indicator variable for $\mathrm{E}=32$, i.e. for individuals that have completed 1st,2nd,3rd,or 4th grade, and so on, and where the indicator variables for the first code present in the sample for each variable is taken out in order to make the

model matrix full rank. The model matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ has $p = 29$ rows and $n = 32923$ columns, with rank equal to 29. Using the plug-in sampling method, we generate a single synthetic dataset. The realizations of the unbiased estimators $\mathbf{B}^*$ and $\mathbf{S}^*$ of $\mathbf{B}$ and $\mathbf{\Sigma}$, are respectively shown in Table 5 and in expression (20), along with the realizations of the original data estimators $\hat{\mathbf{B}}$ and $\mathbf{S}$. These estimates are respectively denoted by $\widetilde{\mathbf{B}}^*$, $\widetilde{\mathbf{S}}^*$, $\widetilde{\hat{\mathbf{B}}}$ and $\widetilde{\mathbf{S}}$, with

$$\widetilde{\mathbf{S}}^* = \begin{pmatrix} 0.6576 & 0.2090 \\ 0.2090 & 1.2905 \end{pmatrix}, \quad \widetilde{\mathbf{S}} = \begin{pmatrix} 0.6626 & 0.2130 \\ 0.2130 & 1.2898 \end{pmatrix}. \tag{20}$$

We see that the point estimates of $\hat{\mathbf{B}}$ based on the synthetic data and the original data tend to be in agreement. We also find that the two estimates of $\mathbf{\Sigma}$, $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{S}}^*$, tend to have a general agreement.

We now present inferences on regression coefficients obtained by applying the methodology from Section 2 and Section 3 to analyze the singly imputed synthetic data and multiply synthetic data, considering $M = 2$ and $M = 5$. For this purpose, we use the statistics T, T_M, T_{comb} and $T_{R,M}$ defined in sections 2 and 3 and their empirical distributions (10^5 simulation size) to test the significance of the model, for $\gamma = 0.05$. For $M = 1$ we obtained $T_{calc} = 4.96468$ that is larger than the determined cut-off point for this case, $d_{2,32923,29;0.05} = 5.14914 \times 10^{-6}$ with a corresponding p-value approximately equal to 0, therefore, rejecting the non-significance of the model, that is, assuming that the explanatory variables in $\mathbf{x}$ have a significant role in determining the values of the response variables in $\mathbf{y}$. For $M = 2$ and $M = 5$, using both new procedures, one finds a similar p-value, with the values of the cut-off points as 4.94839 and 5.06947, for the first procedure, and cut-off points of 4.94420 and 5.06190, for the second procedure, respectively. If we perform the same test in the original data using (2), we obtain for T_O in (2) the computed value $T_{O,calc} = 4.93432$ that is also larger than the determined cut-off point 1.27984×10^{-6} with a the p-value approximately equal to 0, also rejecting the non-significance of the model.

In figure 2, one can see a histogram associated with the empirical distribution of T (10^5 simulation size).

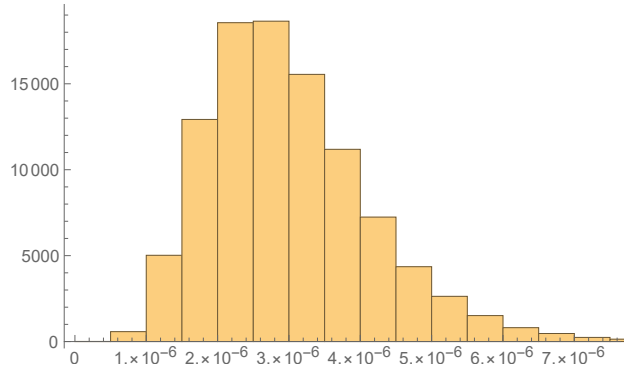


Figure 2: Histogram of the Empirical values of T

Generating datasets for $M = 1$, $M = 2$ and $M = 5$ synthetic datasets where we gather the different p-values obtained using (8), (11), (17) and also the adapted procedure of Reiter (2005). In the latter we replaced $\mathbf{vec}(\mathbf{B}_i^*)$ by $\mathbf{vec}(\mathbf{A}\mathbf{B}_i^*)$, $\mathbf{vec}(\mathbf{B})$ by $\mathbf{vec}(\mathbf{AB})$ and took $\mathbf{U}_i = \mathbf{S}_i^* \otimes (\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')$. From the synthetic datasets the p-values gathered for $M = 1$, $M = 2$ first, second and adapted Reiter's procedures, $M = 5$ first, second and adapted Reiter's procedures where respectively 0.00408, 0.00001, 0.00001, 0.00016, 0, 0 and 0. we may note that for all procedures the p-values are very close to zero as also was the original p-value. It is interesting to observe that the gathered p-values for $M=1$ are not very far from the ones obtained for $M=2$, leading all to the same conclusion, the rejection of the non-significance of the set of regressor coefficients. Comparing the two multiple imputation procedures developed we observe that they present very similar p-values, with the second new procedure having a slightly better approximation to the p-values obtained from the original data. Also, with the increase of the value of M the p-values gets smaller, which although it may be seen as an advantage, it comes at the expenses of a decrease in confidentiality.

Alternatively it is possible to construct the individual confidence intervals of all regres-

sion coefficients by using the procedures in Sections 2 and 3. In Appendices B.6, B.7, B.8, B.9, and B.10 are shown the confidence intervals for all regression coefficients deriving from the original data and the synthetic datasets for M=1,2,5. From these confidence intervals one may observe that for increasing value of M, the confidence intervals becomes smaller and smaller becoming closer and closer to the size of the one derived from the original data. This fact concurs with the study of the radius done in Section 4.

6 Evaluations Under Non-Ideal Conditions

6.1 Non-normal error distribution

In this section we briefly discuss the issue of robustness of our proposed synthetic data analysis methods when errors are non-normal, but the regression is still multivariate linear. In other words, the error term $\mathbb{E}$ in (1) is not normally distributed. In the sequel we consider two types of deviations from normality: t-type (keeping symmetry) and skew-normal type.

Under t-type, we generate $\boldsymbol{y}_i : m \times 1$ as

$$\boldsymbol{y}_i = \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{\Sigma}^{1/2}\boldsymbol{t}_i \left[\sqrt{\frac{\nu-2}{\nu}}\right], \text{ for } i = 1, 2, \ldots, n$$

where $\mathbf{t}_i = \boldsymbol{z}_i \left[\sqrt{\frac{\nu}{\eta}}\right]$ and $\boldsymbol{z}_i$'s are iid with each component distributed as $N(0,1)$ independent of $\eta \sim \chi^2_\nu$. This results in the $\boldsymbol{y}_i$'s being independent multivariate t-distributed vectors. In the Table 5 we have displayed the results of a simulation study under a similar scenario as in Section 4, except that the original data are now generated from the regression model above whose error term has the multivariate t-distribution. We observe in Table 5 that the coverage probability of the confidence regions of our proposed procedures is approximately equal to the nominal value of 0.95. In the table we observe that for sufficiently large n the coverage probability of the procedures of Reiter (2005) is also approximately equal to 0.95

in the $M \geq 5$ cases considered.

Under the skew-normal distribution (Azzalini, 1985; Henze, 1986), we generate $\boldsymbol{y}_i$ as

$$\boldsymbol{y}_i = \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$$

where the m components of $\boldsymbol{z}_i$ are iid with each component distributed as $(sk - \mu_{sk})/\sqrt{var(sk)}$. Here $sk \sim$ skew-normal$[0, 1, \lambda]$ with $\mu_{sk} = E(sk) = \left[\sqrt{\frac{2}{\pi}}\right]\left[\frac{\lambda}{\sqrt{1+\lambda^2}}\right]$ and $\text{var}(sk) = \text{Var}(sk) = [1 - \left(\frac{2}{\pi}\right)\left(\frac{\lambda^2}{1+\lambda^2}\right)]$. Under this data generation scheme, $\boldsymbol{y}_i$'s will have a skew-normal distribution with linear regression as before. The parameter λ represents the extent of deviation from symmetry. In the Table 6 we have displayed the results of a simulation study under a similar scenario as in Section 4, except that the original data are now generated from the regression model above whose error term has the skew-normal distribution. We observe in Table 6 that the coverage probability of the confidence regions of our proposed procedures is approximately equal to the nominal value of 0.95. In the table we observe that for sufficiently large n the coverage probability of the procedures of Reiter (2005) is also approximately equal to 0.95 in the $M \geq 5$ cases considered.

6.2 Regression model is overspecified or underspecified by imputer or data analyst

In this section, which is patterned after Klein and Sinha (2016), we discuss in details the consequences of overspecification or underspecification of the regression model in the analysis of synthetic data as developed in the previous sections. This is rather crucial because the data analysis carried out by the data users which happens at the very last stage obviously depends on the nature of data released to them, which in turn depends on the underlying model which produced the observed data and the imputer's role in creating synthetic data out of it.

To be concrete, we define the following models:

- *Data-generating model* (DM): The true population model that generates the original data

- *Imputation model* (IM): The statistical agency's assumed model for the original data. Based on this model, the statistical agency creates synthetic data.

- *Analysis model* (AM): The data analyst's assumed model for the original data. Based on this model the data analyst, who only has access to the synthetic data, uses the synthetic data for inference.

Referring to our basic multivariate regression model (1), we consider the situation when the matrix $\mathbf{X}$ of covariates can be decomposed into two parts $\mathbf{X_1}$ and $\mathbf{X_2}$, and our primary interest lies in the regression coefficients $\mathbf{B_1}$ associated with $\mathbf{X_1}$, and $\mathbf{X_2}$ may or may not be relevant for the matrix response $\mathbf{Y}$. We thus consider the cases where each of these models (DM,IM,AM) takes one of two forms:

$$\text{Full Model (F)} : \boldsymbol{Y} = \mathbf{B}_1'\boldsymbol{X}_1 + \mathbf{B}_2'\boldsymbol{X}_2 + \mathbb{E}$$
$$\text{Reduced Model (R)} : \boldsymbol{Y} = \mathbf{B}_1'\boldsymbol{X}_1 + \mathbb{E}$$

where $\mathbb{E} \sim N_{mn}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{\Sigma})$. This leads to eight combinations of models which we will denote by the juxtaposition of letters associated with the appropriate model. For example, 'RFR' is the case where the Data follow the reduced model (R), the Imputer infers and imputes according to the full model (F) and the Analyst infers according the reduced model (R). We list these combinations as $(FFF, FFR, FRF, FRR, RFF, RFR, RRF, RRR)$ and use 'Case i' in reference to the model combination described by the ith term of this list.

The following matrices are used in the sequel:

$$\mathbf{P} = \boldsymbol{X}_1' \left(\boldsymbol{X}_1 \boldsymbol{X}_1'\right)^{-1} \boldsymbol{X}_1$$
$$\mathbf{R} = \mathbf{I} - \mathbf{P}$$
$$\mathbf{L} = \left(\boldsymbol{X}_1 \boldsymbol{X}_1'\right)^{-1} \boldsymbol{X}_1 \boldsymbol{X}_2'$$
$$\mathbf{M} = \left(\boldsymbol{X}_2 \mathbf{R} \boldsymbol{X}_2'\right)^{-1}$$
$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$$

Procedure Under Reduced Model

Under the reduced model, the imputer generates $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_M \overset{\text{iid}}{\sim} \mathrm{N}_{mn}\left(\widehat{\mathbf{B}}_1' \boldsymbol{X}_1,\ \mathbf{I}_n \otimes \widehat{\boldsymbol{\Sigma}}\right)$ independently given $\left(\widehat{\mathbf{B}}\,,\, \widehat{\boldsymbol{\Sigma}}\right)$. Imputer and analysts estimates under the reduced model have the form:

$$\widehat{\mathbf{B}}_1 = \left(\mathbf{X}_1 \mathbf{X}_1'\right)^{-1} \mathbf{X}_1 \boldsymbol{Y}'$$
$$\widehat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{Y}\left[\mathbf{I} - \boldsymbol{X}_1' \left(\boldsymbol{X}_1 \boldsymbol{X}_1'\right)^{-1} \boldsymbol{X}_1\right] \boldsymbol{Y}'}{n - p_1}$$

The analyst computes these estimates with $\boldsymbol{Y}$ replaced by the jth synthetic dataset $\boldsymbol{V}_j$ to produce a set of $2M$ estimates $\left\{\left(\widehat{\mathbf{B}}_{1Aj}\,,\, \widehat{\boldsymbol{\Sigma}}_{Aj}\right) \middle| j = 1, \ldots, M\right\}$.

Procedure Under Full Model

Under the full model, the imputer generates $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_M \overset{\text{iid}}{\sim} \mathrm{N}_{mn}\left(\widehat{\mathbf{B}}_1' \boldsymbol{X}_1 + \widehat{\mathbf{B}}_2' \boldsymbol{X}_2,\ \mathbf{I}_n \otimes \widehat{\boldsymbol{\Sigma}}\right)$ independently given $\left(\widehat{\mathbf{B}}\,,\, \widehat{\boldsymbol{\Sigma}}\right)$. Imputer and analyst estimates under the full model have

the form:

$$\widehat{\mathbf{B}}_1 = (\mathbf{X}_1\mathbf{X}_1')^{-1}\mathbf{X}_1\left[\mathbf{I} - \mathbf{X}_2'(\mathbf{X}_2\mathbf{R}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{R}\right]\boldsymbol{Y}'$$
$$\widehat{\mathbf{B}}_2 = (\mathbf{X}_2\mathbf{R}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{R}\boldsymbol{Y}'$$
$$\widehat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{Y}\left[\mathbf{R} - \mathbf{R}\mathbf{X}_2'(\mathbf{X}_2\mathbf{R}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{R}\right]\boldsymbol{Y}'}{n - p_1 - p_2}$$

Under the full model, the analysts computes these estimates with $\boldsymbol{Y}$ replaced by the jth synthetic dataset $\boldsymbol{V}_j$ to produce a set of $3M$ estimates $\left\{\left(\widehat{\mathbf{B}}_{1Aj}, \widehat{\mathbf{B}}_{2Aj}, \widehat{\boldsymbol{\Sigma}}_{Aj}\right)\middle| j = 1, \ldots, M\right\}$.

Analyst Model Assumptions and Estimates

The analyst is interested in estimating $\boldsymbol{B}_1$ only. When the analyst uses the full model the estimates used in the combination rules of Reiter (2005) are

$$\mathbf{q}_j = \widehat{\mathbf{B}}_{1Aj} = (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}\boldsymbol{X}_1\left[\boldsymbol{I} - \boldsymbol{X}_2'(\boldsymbol{X}_2\boldsymbol{R}\boldsymbol{X}_2')^{-1}\boldsymbol{X}_2\boldsymbol{R}\right]\boldsymbol{V}_j'$$
$$\mathbf{u}_j = \widehat{\boldsymbol{\Sigma}}_j \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right] \text{ where}$$
$$\widehat{\boldsymbol{\Sigma}}_j = \frac{\boldsymbol{V}_j\left[\boldsymbol{R} - \boldsymbol{R}\boldsymbol{X}_2'(\boldsymbol{X}_2\boldsymbol{R}\boldsymbol{X}_2')^{-1}\boldsymbol{X}_2\boldsymbol{R}\right]\boldsymbol{V}_j'}{n - p_1 - p_2}$$

When the analyst uses the reduced model the estimates used in the combination rules of Reiter (2005) are

$$\mathbf{q}_j = (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}\boldsymbol{X}_1\boldsymbol{V}_j'$$
$$\mathbf{u}_j = \widehat{\boldsymbol{\Sigma}}_j \otimes (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} \text{ where}$$
$$\widehat{\boldsymbol{\Sigma}}_j = \frac{\boldsymbol{V}_j\left[\boldsymbol{I} - \boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}\boldsymbol{X}_1\right]\boldsymbol{V}_j'}{n - p_1}$$

We will derive, for each case, the expectation of the covariance matrix estimate of Reiter (2005) for $\mathbf{B}_1$ in the case of multiply imputed synthetic data under PPS. For this,

we denote by $\boldsymbol{q}_j$ an estimate of $\mathbf{B}_1$ using the jth synthetic data set and by $\mathbf{u}_j$ an estimate of the covariance matrix of $\boldsymbol{q}_j$, for $j = 1, \ldots, M$.

With this notation, we further define (Reiter, 2005):

$$\begin{aligned}
\overline{\boldsymbol{q}}_M &= \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{q}_j \\
\mathbf{b}_M &= \frac{1}{M-1}\sum_{j=1}^{M} \left(\boldsymbol{q}_j - \overline{\boldsymbol{q}}_M\right)\left(\boldsymbol{q}_j - \overline{\boldsymbol{q}}_M\right)' = \frac{\sum_{j=1}^{M} \boldsymbol{q}_j \boldsymbol{q}_j' - \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} \boldsymbol{q}_i \boldsymbol{q}_j'}{M-1} \\
\overline{\mathbf{u}}_M &= \frac{1}{M}\sum_{j=1}^{M} \mathbf{u}_j \\
\mathbf{T_M} &= \overline{\mathbf{u}}_M + \frac{\mathbf{b_M}}{M}
\end{aligned}$$

where $\mathbf{T}_M$ is the covariance matrix estimate of Reiter (2005). If $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_M$ are identically distributed, we have

$$\begin{aligned}
\mathrm{E}\left[(M-1)\,\mathbf{b}_M\right] &= \sum_{i=1}^{M} \mathrm{E}\left(\boldsymbol{q}_i \boldsymbol{q}_i'\right) - \frac{\sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{E}\left(\boldsymbol{q}_i \boldsymbol{q}_j'\right)}{M} \\
&= M\,\mathrm{Var}\left(\boldsymbol{q}_1\right) - \frac{1}{M}\left[M\,\mathrm{Var}\left(\boldsymbol{q}_1\right) + M(M-1)\mathrm{Cov}\left(\boldsymbol{q}_1, \boldsymbol{q}_2\right)\right] \\
&= M\,\mathrm{Var}\left(\boldsymbol{q}_1\right) - \mathrm{Var}\left(\boldsymbol{q}_1\right) - (M-1)\mathrm{Cov}\left(\boldsymbol{q}_1, \boldsymbol{q}_2\right) \\
&= (M-1)\left[\mathrm{Var}\left(\boldsymbol{q}_1\right) - \mathrm{Cov}\left(\boldsymbol{q}_1, \boldsymbol{q}_2\right)\right]
\end{aligned}$$

Under the full model in the case $m = 1$, Seber and Lee (2003) give the covariance matrix

$$\mathrm{Var}\begin{pmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{pmatrix} = \boldsymbol{\Sigma} \otimes \left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1} = \boldsymbol{\Sigma} \otimes \begin{pmatrix} \left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}' & -\boldsymbol{L}\boldsymbol{M} \\ -\boldsymbol{M}\boldsymbol{L}' & \boldsymbol{M} \end{pmatrix}$$

Thus,

$$\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}\boldsymbol{X} = \boldsymbol{X}_1'\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}\boldsymbol{X}_1' + \boldsymbol{X}_1'\boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\boldsymbol{X}_1 - \boldsymbol{X}_1'\boldsymbol{L}\boldsymbol{M}\boldsymbol{X}_2 - \boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{L}'\boldsymbol{X}_1 + \boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2$$

If the analyst instead uses our approach as presented in Section 3, then the point estimator of $\boldsymbol{B}_1$ is also $\bar{q}_M$. Under our proposed 1st procedure of Section 3, the analysts estimate of the covariance matrix of the point estimator of $\boldsymbol{B}_1$ will be $\frac{M+1}{M}\bar{\boldsymbol{u}}_M$.

Inference on B_1 under the eight Overspecification/Underspecification Scenarios

Case 1: FFF

The analysts aggregated estimate of $\mathbf{B}_1$ is $\widehat{\mathbf{B}}_{1A} = \frac{\sum_{j=1}^{M} \widehat{\mathbf{B}}_{1Aj}}{M}$ which is unbiased since:

$$\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}\right) = \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\right)\right] = \mathrm{E}\left(\widehat{\mathbf{B}}_1\right) = \mathbf{B}_1$$

for $j = 1, \ldots, M$. By similar arguments the analysts estimate is also unbiased in Cases 5, 6, 7 and 8.

Variance of $\widehat{\mathrm{B}}_{1A}$

Using the results of Seber and Lee (2003), we have:

$$\begin{aligned}
\mathrm{Var}\left(\sum_{j=1}^{M} \widehat{\mathbf{B}}_{1Aj}\right) &= \mathrm{E}\left[\mathrm{Var}\left(\sum_{j=1}^{M} \widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\sum_{j=1}^{M} \widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\
&= \mathrm{E}\left\{M\widehat{\boldsymbol{\Sigma}} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{LML}'\right]\right\} + \mathrm{Var}\left(M\widehat{\mathbf{B}}_1\right) \\
&= M\boldsymbol{\Sigma} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{LML}'\right] + M^2\boldsymbol{\Sigma} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{LML}'\right]
\end{aligned}$$

Thus, $\mathrm{Var}\left(\widehat{\mathbf{B}}_{1A}\right) = \left(1 + \frac{1}{M}\right)\boldsymbol{\Sigma} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{LML}'\right]$.

Expectation of $\mathbf{T_M}$

We first observe

$$\begin{aligned}
\operatorname{Var}(\boldsymbol{q}_j) = \operatorname{Var}\left(\widehat{\mathbf{B}}_{1Aj}\right) &= \mathrm{E}\left[\operatorname{Var}\left(\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] + \operatorname{Var}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\
&= \mathrm{E}\left\{\widehat{\boldsymbol{\Sigma}} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]\right\} + \operatorname{Var}\left(\widehat{\mathbf{B}}_1\right) \\
&= \boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right] + \boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right] \\
&= 2\boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]
\end{aligned}$$

and for $i \neq j$

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{q}_i\boldsymbol{q}_j'\right) &= \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Ai}\widehat{\mathbf{B}}_{1Aj}'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\
&= \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Ai}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\cdot\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] = \mathrm{E}\left(\widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1'\right) \\
&= \operatorname{Var}\left(\widehat{\mathbf{B}}_1\right) + \operatorname{vec}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_1\right)\right]\cdot\operatorname{vec}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_1\right)\right]' \\
&= \boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right] + \operatorname{vec}(\mathbf{B}_1)\cdot\operatorname{vec}(\mathbf{B}_1)'
\end{aligned}$$

Thus,

$$\mathrm{E}\left[(M-1)\,\mathbf{b}_M\right] = (M-1)\left[\operatorname{Var}(\boldsymbol{q}_1) - \operatorname{Cov}(\boldsymbol{q}_1, \boldsymbol{q}_2)\right] = (M-1)\boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]$$

In this case $\mathbf{u}_j = \widehat{\boldsymbol{\Sigma}}_{Aj} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}'\right]$ where

$$\widehat{\boldsymbol{\Sigma}}_{Aj} = \frac{\boldsymbol{V}_j\left[\mathbf{R} - \mathbf{R}\mathbf{X}_{\mathbf{2}}'\left(\mathbf{X_2}\mathbf{R}\mathbf{X}_{\mathbf{2}}'\right)^{-\mathbf{1}}\mathbf{X_2}\mathbf{R}\right]\boldsymbol{V}_j'}{n - p_1 - p_2}$$

and it can be shown that $\mathrm{E}\left(\widehat{\boldsymbol{\Sigma}}_{Aj}\right) = \boldsymbol{\Sigma}$, so that $\mathrm{E}\left(\overline{\mathbf{u}}_M\right) = \boldsymbol{\Sigma} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}'\right]$. Finally, this yields

$$\mathrm{E}\left(\mathbf{T}_M\right) = \mathrm{E}\left(\frac{\mathbf{b}_M}{M} + \overline{\mathbf{u}}_M\right) = \left(1 + \frac{1}{M}\right)\boldsymbol{\Sigma} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}'\right]$$

Cases 2, 3 and 4: FFR, FRF and FRR

In Case 2, the analysts aggregated estimate is $\widehat{\mathbf{B}}_{1A} = (\mathbf{X}_1\mathbf{X}_1')^{-1}\mathbf{X}_1\left(\frac{\sum_{j=1}^{M}\boldsymbol{V}_j'}{M}\right)$ with expectation $\mathbf{B}_1 + \mathbf{L}\mathbf{B}_2$ since:

$$\mathrm{E}\left(\boldsymbol{V}_j\right) = \mathrm{E}\left[\mathrm{E}\left(\boldsymbol{V}_j\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] = \mathrm{E}\left(\widehat{\mathbf{B}}_1'\boldsymbol{X}_1 + \widehat{\mathbf{B}}_2'\boldsymbol{X}_2\right) = \mathbf{B}_1'\boldsymbol{X}_1 + \mathbf{B}_2'\boldsymbol{X}_2$$

The estimators in Cases 3 and 4 are biased by similar arguments, so we do not pursue these three cases any further.

Case 5: RFF

Remark. Case 5 and Case 1 differ only in their assumptions about the value of $\mathbf{B}_2$ in the data model, which effect neither the variance of the data nor of the imputers estimate of $(\mathbf{B}_1, \mathbf{B}_2)$ nor of the analysts estimate of $\mathbf{B}_1$. Thus, all expressions in Case 5 are identical to those in Case 1. The estimates of Reiter (2005) also inherit this invariance to the data model since they are composed of similarly invariant estimates.

Case 6: RFR

Variance of $\widehat{\mathbf{B}}_{1A}$

By arguments similar to that of Case 1 above, we have

$$\mathrm{Var}\left(\sum_{j=1}^{M}\boldsymbol{V}_j'\right) = M\,\boldsymbol{\Sigma}\otimes\boldsymbol{I}_n + M^2\,\mathrm{Var}\left(\widehat{\mathbf{B}}_1'\boldsymbol{X}_1 + \widehat{\mathbf{B}}_2'\boldsymbol{X}_2\right) = M\,\boldsymbol{\Sigma}\otimes\boldsymbol{I}_n + M^2\,\boldsymbol{\Sigma}\otimes\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}\boldsymbol{X}$$

$$= M\,\boldsymbol{\Sigma}\otimes\boldsymbol{I}_n + M^2\boldsymbol{\Sigma}\otimes\left[\mathbf{X}_1'\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\mathbf{X}_1 - \mathbf{X}_1'\mathbf{L}\mathbf{M}\mathbf{X}_2 - \mathbf{X}_2'\mathbf{M}\mathbf{L}'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{M}\mathbf{X}_2 + \mathbf{X}_1'\mathbf{L}\mathbf{M}\mathbf{L}'\mathbf{X}_1\right]$$

Thus,

$$\begin{aligned}
\operatorname{Var}\left(\widehat{\mathbf{B}}_{1A}\right) &= \operatorname{Var}\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\mathbf{X}_1\left(\frac{\sum_{j=1}^{M}\boldsymbol{V}_j'}{M}\right)\right] \\
&= \frac{1}{M}\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} + \boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}\boldsymbol{X}_1\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}\boldsymbol{X}\boldsymbol{X}_1'\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1} \\
&= \frac{1}{M}\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} + \boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} = \left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}
\end{aligned}$$

Expectation of $\mathbf{T_M}$

By arguments similar to that of Case 1 above, we have

$$\begin{aligned}
\operatorname{Var}\left(\boldsymbol{q}_j\right) &= \operatorname{Var}\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\mathbf{X}_1\boldsymbol{V}_j'\right] \\
&= \mathrm{E}\left\{\operatorname{Var}\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\mathbf{X}_1\boldsymbol{V}_j'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right]\right\} + \operatorname{Var}\left\{\mathrm{E}\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\mathbf{X}_1\boldsymbol{V}_j'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right]\right\} \\
&= \mathrm{E}\left[\widehat{\boldsymbol{\Sigma}}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\right] + \operatorname{Var}\left[\begin{pmatrix}\mathbf{I} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{B}}_1\\ \widehat{\mathbf{B}}_2\end{pmatrix}\right] = \boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} + \boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} \\
&= 2\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}
\end{aligned}$$

and for $i \neq j$

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{q}_i\boldsymbol{q}_j'\right) &= \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Ai}\widehat{\mathbf{B}}_{1Aj}'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] = \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Ai}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\cdot\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\
&= \mathrm{E}\left[\left(\widehat{\mathbf{B}}_1+\boldsymbol{L}\widehat{\mathbf{B}}_2\right)\left(\widehat{\mathbf{B}}_1+\boldsymbol{L}\widehat{\mathbf{B}}_2\right)'\right] \\
&= \operatorname{Var}\left(\widehat{\mathbf{B}}_1+\boldsymbol{L}\widehat{\mathbf{B}}_2\right) + \operatorname{vec}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_1+\boldsymbol{L}\widehat{\mathbf{B}}_2\right)\right]\cdot\operatorname{vec}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_1+\boldsymbol{L}\widehat{\mathbf{B}}_2\right)\right]' \\
&= \operatorname{Var}\left[\begin{pmatrix}\mathbf{I} & \mathbf{L}\end{pmatrix}\begin{pmatrix}\widehat{\mathbf{B}}_1\\ \widehat{\mathbf{B}}_2\end{pmatrix}\right] + \operatorname{vec}\left(\mathbf{B}_1\right)\cdot\operatorname{vec}\left(\mathbf{B}_1\right)' = \boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1} + \operatorname{vec}\left(\mathbf{B}_1\right)\cdot\operatorname{vec}\left(\mathbf{B}_1\right)'
\end{aligned}$$

Thus,

$$\mathrm{E}\left[\left(M-1\right)\mathbf{b}_M\right] = \left(M-1\right)\left[\operatorname{Var}\left(\boldsymbol{q}_1\right) - \operatorname{Cov}\left(\boldsymbol{q}_1,\boldsymbol{q}_2\right)\right] = \left(M-1\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$$

In this case $\mathbf{u}_j = \widehat{\mathbf{\Sigma}}_{Aj} \otimes (\mathbf{X}_1\mathbf{X}_1')^{-1}$ where $\widehat{\mathbf{\Sigma}}_{Aj} = \frac{\boldsymbol{V}_j \boldsymbol{R} \boldsymbol{V}_j'}{n-p_1-p_2}$ and it can be shown that $\mathrm{E}\left(\widehat{\mathbf{\Sigma}}_{Aj}\right) = \left(1 + \frac{p_2}{n-p_1}\right)\mathbf{\Sigma}$, so that $\mathrm{E}\left(\overline{\mathbf{u}}_M\right) = \left(1 + \frac{p_2}{n-p_1}\right)\mathbf{\Sigma} \otimes (\mathbf{X}_1\mathbf{X}_1')^{-1}$. Finally, this yields

$$\mathrm{E}\left(\mathbf{T}_M\right) = \mathrm{E}\left(\frac{\mathbf{b}_M}{M} + \overline{\mathbf{u}}_M\right) = \left(1 + \frac{1}{M} + \frac{p_2}{n-p_1}\right)\mathbf{\Sigma} \otimes (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}$$

Case 7: RRF

Variance of $\widehat{\mathbf{B}}_{1A}$

By arguments similar to that of Case 1 above, we have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{j=1}^{M} \boldsymbol{V}_j'\right) &= \mathrm{E}\left[\mathrm{Var}\left(\sum_{j=1}^{M} \boldsymbol{V}_j' \middle| \widehat{\mathbf{B}}_1\,,\, \widehat{\mathbf{\Sigma}}\right)\right] + \mathrm{Var}\left[\mathrm{E}\left(\sum_{j=1}^{M} \boldsymbol{V}_j' \middle| \widehat{\mathbf{B}}_1\,,\, \widehat{\mathbf{\Sigma}}\right)\right] \\
&= M\,\mathrm{E}\left(\widehat{\mathbf{\Sigma}}\right) \otimes \boldsymbol{I}_n + M^2\,\mathrm{Var}\left(\mathbf{X}_1'\widehat{\mathbf{B}}_1\right) = M\,\mathbf{\Sigma} \otimes \boldsymbol{I}_n + M^2\,\mathbf{\Sigma} \otimes \boldsymbol{X}_1'\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}\boldsymbol{X}_1
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\mathbf{B}}_{1A}\right) &= \mathrm{Var}\left\{(\mathbf{X}_1\mathbf{X}_1')^{-1}\mathbf{X}_1\left[\mathbf{I} - \mathbf{X}_{\mathbf{2}}'\left(\mathbf{X}_{\mathbf{2}}\mathbf{R}\mathbf{X}_{\mathbf{2}}'\right)^{-\mathbf{1}}\mathbf{X}_{\mathbf{2}}\mathbf{R}\right]\sum_{j=1}^{M}\boldsymbol{V}_j'/M\right\} \\
&= \frac{1}{M}\mathbf{\Sigma} \otimes \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}'\right] + \mathbf{\Sigma} \otimes (\mathbf{X}_1\mathbf{X}_1')^{-1} \\
&= (1 + \frac{1}{M})\mathbf{\Sigma} \otimes (\mathbf{X}_1\mathbf{X}_1')^{-1} + \frac{1}{M}\mathbf{\Sigma} \otimes \mathbf{L}\mathbf{M}\mathbf{L}'
\end{aligned}$$

Expectation of $\mathbf{T_M}$

By arguments similar to those of Case 1, we have

$$\begin{aligned}
\operatorname{Var}(\boldsymbol{q}_j) = \operatorname{Var}\left(\widehat{\mathbf{B}}_{1Aj}\right) &= \mathrm{E}\left[\operatorname{Var}\left(\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] + \operatorname{Var}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\
&= \mathrm{E}\left(\widehat{\boldsymbol{\Sigma}}\right) \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right] + \operatorname{Var}\left(\widehat{\mathbf{B}}_1\right) \\
&= \boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right] + \boldsymbol{\Sigma} \otimes (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} \\
&= \boldsymbol{\Sigma} \otimes \left[2\,(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]
\end{aligned}$$

and for $i \neq j$

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{q}_i\boldsymbol{q}_j'\right) &= \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Ai}\widehat{\mathbf{B}}_{1Aj}'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] = \mathrm{E}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_{1Ai}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\cdot\mathrm{E}\left(\widehat{\mathbf{B}}_{1Aj}'\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\mathbf{B}}_2\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\
&= \mathrm{E}\left(\widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1'\right) = \operatorname{Var}\left(\widehat{\mathbf{B}}_1\right) + \operatorname{vec}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_1\right)\right]\cdot\operatorname{vec}\left[\mathrm{E}\left(\widehat{\mathbf{B}}_1\right)\right]' \\
&= \boldsymbol{\Sigma} \otimes (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \operatorname{vec}(\mathbf{B}_1)\cdot\operatorname{vec}(\mathbf{B}_1)'
\end{aligned}$$

Thus,

$$\mathrm{E}\left[(M-1)\mathbf{b}_M\right] = (M-1)[\operatorname{Var}(\boldsymbol{q}_1) - \operatorname{Cov}(\boldsymbol{q}_1,\boldsymbol{q}_2)] = (M-1)\boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]$$

In this case, $\mathbf{u}_j = \left[(\mathbf{X}_1\mathbf{X}_1')^{-1} + \mathbf{L}\mathbf{M}\mathbf{L}'\right] \otimes \widehat{\boldsymbol{\Sigma}}_{Aj}$, where $\widehat{\boldsymbol{\Sigma}}_{Aj} = \frac{\boldsymbol{V}_j\left[\mathbf{R}-\mathbf{R}\mathbf{X}_2'(\mathbf{X}_2\mathbf{R}\mathbf{X}_2')^{-1}\mathbf{X}_2\mathbf{R}\right]\boldsymbol{V}_j'}{n-p_1-p_2}$. It can be shown that $\mathrm{E}\left(\widehat{\boldsymbol{\Sigma}}_{Aj}\right) = \boldsymbol{\Sigma}$, for $j = 1,\ldots,M$, so that $\mathrm{E}\left(\overline{\mathbf{u}}_M\right) = \boldsymbol{\Sigma}\otimes\left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]$. Finally, we have

$$\mathrm{E}\left(\mathbf{T}_M\right) = \mathrm{E}\left(\frac{\mathbf{b}_M}{M} + \overline{\mathbf{u}}_M\right) = \left(1 + \frac{1}{M}\right)\boldsymbol{\Sigma} \otimes \left[(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}'\right]$$

Case 8: RRR

Variance of $\widehat{\mathbf{B}}_{1A}$

By arguments similar to those above, we have

$$\begin{aligned}\operatorname{Var}\left(\sum_{j=1}^{M}\widehat{\mathbf{B}}_{1Aj}\right) &= \mathrm{E}\left[\operatorname{Var}\left(\sum_{j=1}^{M}\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] + \operatorname{Var}\left[\mathrm{E}\left(\sum_{j=1}^{M}\widehat{\mathbf{B}}_{1Aj}\middle|\widehat{\mathbf{B}}_1\,,\,\widehat{\boldsymbol{\Sigma}}\right)\right] \\ &= M^2\operatorname{Var}\left(\widehat{\mathbf{B}}_1\right) + M\,\mathrm{E}\left(\widehat{\boldsymbol{\Sigma}}\right)\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} \\ &= M^2\,\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1} + M\,\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}\end{aligned}$$

so that $\operatorname{Var}\left(\widehat{\mathbf{B}}_{1A}\right) = \left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}$.

Expectation of $\mathbf{T_M}$

By arguments similar to those above, it is clear that $\operatorname{Var}(\boldsymbol{q}_j) = 2\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$ and $\mathrm{E}\left(\boldsymbol{q}_i\boldsymbol{q}_j'\right) = \boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$. Then,

$$\mathrm{E}\left[(M-1)\mathbf{b}_M\right] = (M-1)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$$

In this case, $\mathbf{u}_j = \boldsymbol{\Sigma}_{Aj}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$, where $\widehat{\boldsymbol{\Sigma}}_{Aj} = \frac{\boldsymbol{V}_j\boldsymbol{R}\boldsymbol{V}_j'}{n-p_1}$. It can be shown that $\mathrm{E}\left(\widehat{\boldsymbol{\Sigma}}_{Aj}\right) = \boldsymbol{\Sigma}$, for $j=1,\ldots,M$, so that $\mathrm{E}\left(\overline{\mathbf{u}}_M\right) = \boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$. Finally, we have

$$\mathrm{E}\left(\mathbf{T}_M\right) = \mathrm{E}\left(\frac{\mathbf{b}_M}{M}+\overline{\mathbf{u}}_M\right) = \left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$$

Summary

The following summarizes the results derived under the cases 1,5,6,7,8 where the point estimator is unbiased.

Variance of point estimator and expected value of Reiter's (2005) variance estimator:

Case	$\mathrm{Var}\left(\widehat{\mathbf{B}}_{1A}\right)$	$\mathrm{E}\left[\widehat{\mathrm{Var}}\left(\widehat{\mathbf{B}}_{1A}\right)\right]$
FFF	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$
RFF	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$
RFR	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$	$\left(1+\frac{1}{M}+\frac{p_2}{n-p_1}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$
RRF	$(1+\frac{1}{M})\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\frac{1}{M}\boldsymbol{\Sigma}\otimes\mathbf{LML}'$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}+\boldsymbol{LML}'\right]$
RRR	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$

Variance of point estimator and expected value of variance estimator from our 1st procedure:

Case	$\mathrm{Var}\left(\widehat{\mathbf{B}}_{1A}\right)$	$\mathrm{E}\left[\widehat{\mathrm{Var}}\left(\widehat{\mathbf{B}}_{1A}\right)\right]$
FFF	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$
RFF	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\mathbf{LML}'\right]$
RFR	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$	$\left(1+\frac{1}{M}\right)\left(1+\frac{p_2}{n-p_1}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$
RRF	$(1+\frac{1}{M})\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}+\frac{1}{M}\boldsymbol{\Sigma}\otimes\mathbf{LML}'$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left[\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}+\boldsymbol{LML}'\right]$
RRR	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\mathbf{X}_1\mathbf{X}_1'\right)^{-1}$	$\left(1+\frac{1}{M}\right)\boldsymbol{\Sigma}\otimes\left(\boldsymbol{X}_1\boldsymbol{X}_1'\right)^{-1}$

We observe that in all the cases where the analyst's estimate of $\mathbf{B}_1$ is unbiased, the covariance matrix estimate of Reiter (2005), and of our approach is expected to over-estimate the covariance. This is consistent with what was demonstrated in Klein and Sinha (2016).

6.3 Data analyst's regression is something other than sensitive variables on non-sensitive variables

Upon observing the released synthetic data, there are a multitude of regression models that may be of interest to the data analyst. For example, the data analyst may be interested in the regression of the nonsensitive variables on the sensitive variables, or in the regression of a combination of sensitive and nonsensitive variables on another combination of sensitive

and nonsensitive variables. To derive the methodology in Sections 2 and 3, we assumed that the data analyst's goal is to use the released synthetic data to draw inference on the regression coefficients in the regression of the sensitive variables, given the nonsensitive variables. We assumed that this model is a multivariate normal model, and it is also the data generation model, and the imputation model. In this section we conduct some simulation studies to evaluate the performance of our methodology under several different cases of the data analyst's choice of a regression model.

To conduct this analysis, let $\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{y}_n \\ \boldsymbol{x}_n \end{pmatrix}$ denote the original data where $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are sensitive while $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are not sensitive. Assume for simplicity that $\boldsymbol{y}_i = (y_{1i}, y_{2i})'$ and $\boldsymbol{x}_i = (x_{1i}, x_{2i})'$ for $i = 1, \ldots, n$. Suppose the data generation model is

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{y}_n \\ \boldsymbol{x}_n \end{pmatrix} \sim iid \sim N_4 \left[\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right], \tag{21}$$

where

$$\boldsymbol{\mu}_y = \begin{pmatrix} \mu_{y_1} \\ \mu_{y_2} \end{pmatrix}, \boldsymbol{\mu}_x = \begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{yy} = \begin{pmatrix} \sigma_{y_1y_1} & \sigma_{y_1y_2} \\ \sigma_{y_1y_2} & \sigma_{y_2y_2} \end{pmatrix}, \boldsymbol{\Sigma}_{yx} = \begin{pmatrix} \sigma_{y_1x_1} & \sigma_{y_1x_2} \\ \sigma_{y_2x_1} & \sigma_{y_2x_2} \end{pmatrix}, \boldsymbol{\Sigma}_{xy} = \begin{pmatrix} \sigma_{y_1x_1} & \sigma_{y_2x_1} \\ \sigma_{y_1x_2} & \sigma_{y_2x_2} \end{pmatrix}, \boldsymbol{\Sigma}_{xx} = \begin{pmatrix} \sigma_{x_1x_1} & \sigma_{x_1x_2} \\ \sigma_{x_1x_2} & \sigma_{x_2x_2} \end{pmatrix}.$$

Under model (21) it follows that

$$\begin{aligned} &(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \text{ are independently distributed such that} \\ &\boldsymbol{y}_i|(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \sim N_2 \left[\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_x) = \boldsymbol{\Omega}_1' \begin{pmatrix} 1 \\ \boldsymbol{x}_i \end{pmatrix}, \ \boldsymbol{\Sigma}_{yy\cdot x} \right] \text{ for } i = 1, \ldots, n, \end{aligned} \tag{22}$$

where

$$\boldsymbol{\Omega}_1' = \left(\boldsymbol{\mu}_y - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\mu}_x, \quad \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1} \right)_{2\times 3}, \ \boldsymbol{\Sigma}_{yy\cdot x} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}.$$

Because $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are sensitive, while $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are nonsensitive, only the y-variables will be synthesized, and the imputation model is the conditional distribution of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ given $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ as displayed in (22). Because (22) is a multivariate linear regression model, we assume that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, the synthetic version of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, are generated via the sampling in equation (5). In case of multiply imputed synthetic data, this sampling is repeated independently $M > 1$ times to get $\boldsymbol{v}_1^{(j)}, \ldots, \boldsymbol{v}_n^{(j)}$ for $j = 1, \ldots, M$. The released data are of the form $\begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{v}_n \\ \boldsymbol{x}_n \end{pmatrix}$ in the case of singly imputed synthetic data, or $\begin{pmatrix} \boldsymbol{v}_1^{(j)} \\ \boldsymbol{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{v}_n^{(j)} \\ \boldsymbol{x}_n \end{pmatrix}$ for $j = 1, \ldots, M$ in case of multiply imputed synthetic data. We consider the following eight cases of the data analysis model.

Case 1. *Regression of (sensitive) on (nonsensitive).* Suppose the analysis model is the conditional distribution of $\{\boldsymbol{y}_i, i = 1, \ldots, n\}$ given $\{\boldsymbol{x}_i, i = 1, \ldots, n\}$. In this case the analysis model is given in equation (22), and is the same model as used for imputation.

Case 2. *Regression of (sensitive) on (sensitive).* Suppose the analysis model is the conditional distribution of $\{y_{1i}, i = 1, \ldots, n\}$ given $\{y_{2i}, i = 1, \ldots, n\}$. In the conditional distribution of $y_{11}, \ldots, y_{1n}$ given $y_{21}, \ldots, y_{2n}$, the random variables $y_{11}, \ldots, y_{1n}$ are independently distributed such that

$$y_{1i} \sim N\left[\mu_{y_1} + \frac{\sigma_{y_1 y_2}}{\sigma_{y_2 y_2}}(y_{2i} - \mu_{y_2}) = \boldsymbol{\Omega}_2' \begin{pmatrix} 1 \\ y_{2i} \end{pmatrix},\ \sigma_{y_1 y_1 \cdot y_2}\right]$$

for $i = 1, \ldots, n$, where

$$\boldsymbol{\Omega}_2' = \left(\mu_{y_1} - \tfrac{\sigma_{y_1 y_2}}{\sigma_{y_2 y_2}}\mu_{y_2},\ \ \tfrac{\sigma_{y_1 y_2}}{\sigma_{y_2 y_2}}\right)_{1\times 2},\ \sigma_{y_1 y_1 \cdot y_2} = \sigma_{y_1 y_1} - \frac{\sigma_{y_1 y_2}^2}{\sigma_{y_2 y_2}}.$$

Case 3. *Regression of (nonsensitive) on (sensitive).* Suppose the analysis model is the conditional distribution of $\{\boldsymbol{x}_i, i = 1, \ldots, n\}$ given $\{\boldsymbol{y}_i, i = 1, \ldots, n\}$. In the conditional distribution of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, the random variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently

distributed such that

$$\boldsymbol{x}_i \sim N_2\left[\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_y) = \boldsymbol{\Omega}_3'\begin{pmatrix}1\\ \boldsymbol{y}_i\end{pmatrix},\ \boldsymbol{\Sigma}_{xx\cdot y}\right]$$

for $i = 1, \ldots, n$, where

$$\boldsymbol{\Omega}_3' = \left(\boldsymbol{\mu}_x - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\mu}_y,\quad \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\right)_{2\times 3},\ \boldsymbol{\Sigma}_{xx\cdot y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}.$$

Case 4. *Regression of (sensitive) on (sensitive and nonsensitive).* Suppose the analysis model is the conditional distribution of $\{y_{1i}, i = 1, \ldots, n\}$ given $\{(y_{2i}, \boldsymbol{x}_i), i = 1, \ldots, n\}$. In the conditional distribution of $y_{11}, \ldots, y_{1n}$ given $\begin{pmatrix}y_{21}\\ \boldsymbol{x}_1\end{pmatrix}, \ldots, \begin{pmatrix}y_{2n}\\ \boldsymbol{x}_n\end{pmatrix}$, the random variables $y_{11}, \ldots, y_{1n}$ are independently distributed such that

$$y_{1i} \sim N\left[\mu_{y_1} + \boldsymbol{\Sigma}_{(y_1)(y_2,x)}\boldsymbol{\Sigma}_{(y_2,x)(y_2,x)}^{-1}\left\{(y_{2i}, \boldsymbol{x}_i')' - (\mu_{y_2}, \boldsymbol{\mu}_x')'\right\} = \boldsymbol{\Omega}_4'\begin{pmatrix}1\\ y_{2i}\\ \boldsymbol{x}_i\end{pmatrix}, \boldsymbol{\Sigma}_{(y_1,y_1)\cdot(y_2,x)}\right]$$

for $i = 1, \ldots, n$, where

$$\boldsymbol{\Sigma}_{(y_1)(y_2,x)} = (\sigma_{y_1y_2}, \sigma_{y_1x_1}, \sigma_{y_1x_2}),\ \boldsymbol{\Sigma}_{(y_2,x)(y_2,x)} = \begin{pmatrix}\sigma_{y_2y_2} & \sigma_{y_2x_1} & \sigma_{y_2x_2}\\ \sigma_{y_2x_1} & \sigma_{x_1x_1} & \sigma_{x_1x_2}\\ \sigma_{y_2x_2} & \sigma_{x_1x_2} & \sigma_{x_2x_2}\end{pmatrix},$$

$$\boldsymbol{\Omega}_4' = \left(\mu_{y_1} - \boldsymbol{\Sigma}_{(y_1)(y_2,x)}\boldsymbol{\Sigma}_{(y_2,x)(y_2,x)}^{-1}(\mu_{y_2}, \boldsymbol{\mu}_x')',\quad \boldsymbol{\Sigma}_{(y_1)(y_2,x)}\boldsymbol{\Sigma}_{(y_2,x)(y_2,x)}^{-1}\right)_{1\times 4}.$$

$$\boldsymbol{\Sigma}_{(y_1,y_1)\cdot(y_2,x)} = \sigma_{y_1y_1} - \boldsymbol{\Sigma}_{(y_1)(y_2,x)}\boldsymbol{\Sigma}_{(y_2,x)(y_2,x)}^{-1}\boldsymbol{\Sigma}_{(y_1)(y_2,x)}'.$$

Case 5. *Regression of (nonsensitive) on (sensitive and nonsensitive).* Suppose the analysis model is the conditional distribution of $\{x_{1i}, i = 1, \ldots, n\}$ given $\{(x_{2i}, \boldsymbol{y}_i), i = 1, \ldots, n\}$. In the conditional distribution of $x_{11}, \ldots, x_{1n}$ given $\begin{pmatrix}x_{21}\\ \boldsymbol{y}_1\end{pmatrix}, \ldots, \begin{pmatrix}x_{2n}\\ \boldsymbol{y}_n\end{pmatrix}$, the random variables

$x_{11}, \ldots, x_{1n}$ are independently distributed such that

$$x_{1i} \sim N\left[\mu_{x_1} + \boldsymbol{\Sigma}_{(x_1)(x_2,y)}\boldsymbol{\Sigma}^{-1}_{(x_2,y)(x_2,y)}\left\{(x_{2i}, \boldsymbol{y}_i')' - (\mu_{x_2}, \boldsymbol{\mu}_y')'\right\} = \boldsymbol{\Omega}_5' \begin{pmatrix} 1 \\ x_{2i} \\ \boldsymbol{y}_i \end{pmatrix}, \boldsymbol{\Sigma}_{(x_1,x_1)\cdot(x_2,y)}\right]$$

for $i = 1, \ldots, n$, where

$$\boldsymbol{\Sigma}_{(x_1)(x_2,y)} = (\sigma_{x_1x_2}, \sigma_{y_1x_1}, \sigma_{y_2x_1}),\ \boldsymbol{\Sigma}_{(x_2,y)(x_2,y)} = \begin{pmatrix} \sigma_{x_2x_2} & \sigma_{y_1x_2} & \sigma_{y_2x_2} \\ \sigma_{y_1x_2} & \sigma_{y_1y_1} & \sigma_{y_1y_2} \\ \sigma_{y_2x_2} & \sigma_{y_1y_2} & \sigma_{y_2y_2} \end{pmatrix},$$

$$\boldsymbol{\Omega}_5' = \left(\mu_{x_1} - \boldsymbol{\Sigma}_{(x_1)(x_2,y)}\boldsymbol{\Sigma}^{-1}_{(x_2,y)(x_2,y)}(\mu_{x_2}, \boldsymbol{\mu}_y')',\quad \boldsymbol{\Sigma}_{(x_1)(x_2,y)}\boldsymbol{\Sigma}^{-1}_{(x_2,y)(x_2,y)}\right)_{1\times 4}.$$

$$\boldsymbol{\Sigma}_{(x_1,x_1)\cdot(x_2,y)} = \sigma_{x_1x_1} - \boldsymbol{\Sigma}_{(x_1)(x_2,y)}\boldsymbol{\Sigma}^{-1}_{(x_2,y)(x_2,y)}\boldsymbol{\Sigma}'_{(x_1)(x_2,y)}.$$

Case 6. *Regression of (sensitive and nonsensitive) on (sensitive).* Suppose the analysis model is the conditional distribution of $\{(y_{1i}, x_{1i}), i = 1, \ldots, n\}$ given $\{y_{2i}, i = 1, \ldots, n\}$. In the conditional distribution of $\begin{pmatrix} y_{11} \\ x_{11} \end{pmatrix}, \ldots, \begin{pmatrix} y_{1n} \\ x_{1n} \end{pmatrix}$ given $y_{21}, \ldots, y_{2n}$, the random vectors $(y_{11}, x_{11})', \ldots, (y_{1n}, x_{1n})'$ are independently distributed such that

$$\begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix} \sim N_2\left\{\begin{pmatrix} \mu_{y_1} \\ \mu_{x_1} \end{pmatrix} + \boldsymbol{\Sigma}_{(y_1,x_1)(y_2)}\sigma^{-1}_{y_2y_2}\left[y_{2i} - \mu_{y_2}\right] = \boldsymbol{\Omega}_6' \begin{pmatrix} 1 \\ y_{2i} \end{pmatrix}, \boldsymbol{\Sigma}_{(y_1,x_1)(y_1,x_1)\cdot(y_2)}\right\}$$

for $i = 1, \ldots, n$, where

$$\boldsymbol{\Sigma}_{(y_1,x_1)(y_2)} = (\sigma_{y_1y_2}, \sigma_{y_2x_1})',$$

$$\boldsymbol{\Omega}_6' = \left((\mu_{y_1}, \mu_{x_1})' - \boldsymbol{\Sigma}_{(y_1x_1)(y_2)}\sigma^{-1}_{y_2y_2}\mu_{y_2},\quad \boldsymbol{\Sigma}_{(y_1x_1)(y_2)}\sigma^{-1}_{y_2y_2}\right)_{2\times 2}.$$

$$\boldsymbol{\Sigma}_{(y_1,x_1)(y_1,x_1)\cdot y_2} = \begin{pmatrix} \sigma_{y_1y_1} & \sigma_{y_1x_1} \\ \sigma_{y_1x_1} & \sigma_{x_1x_1} \end{pmatrix} - \boldsymbol{\Sigma}_{(y_1,x_1)(y_2)}\sigma^{-1}_{y_2y_2}\boldsymbol{\Sigma}'_{(y_1,x_1)(y_2)}.$$

Case 7. *Regression of (sensitive and nonsensitive) on (nonsensitive).* Suppose the analysis model is the conditional distribution of $\{(y_{1i}, x_{1i})i = 1, \ldots, n\}$ given $\{x_{2i}, i = 1, \ldots, n\}$. In the conditional distribution of $\begin{pmatrix} y_{11} \\ x_{11} \end{pmatrix}, \ldots, \begin{pmatrix} y_{1n} \\ x_{1n} \end{pmatrix}$ given $x_{21}, \ldots, x_{2n}$, the random vectors $(y_{11}, x_{11})', \ldots, (y_{1n}, x_{1n})'$ are independently distributed such that

$$\begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} \mu_{y_1} \\ \mu_{x_1} \end{pmatrix} + \mathbf{\Sigma}_{(y_1,x_1)(x_2)} \sigma^{-1}_{x_2x_2} \left[x_{2i} - \mu_{x_2} \right] = \mathbf{\Omega}'_7 \begin{pmatrix} 1 \\ x_{2i} \end{pmatrix}, \mathbf{\Sigma}_{(y_1,x_1)(y_1,x_1)\cdot x_2} \right\}$$

for $i = 1, \ldots, n$, where

$$\begin{aligned} \mathbf{\Sigma}_{(y_1,x_1)(x_2)} &= (\sigma_{y_1x_2}, \sigma_{x_1x_2})', \\ \mathbf{\Omega}'_7 &= \left((\mu_{y_1}, \mu_{x_1})' - \mathbf{\Sigma}_{(y_1,x_1)(x_2)} \sigma^{-1}_{x_2x_2} \mu_{x_2}, \quad \mathbf{\Sigma}_{(y_1,x_1)(x_2)} \sigma^{-1}_{x_2x_2} \right)_{2\times 2}. \\ \mathbf{\Sigma}_{(y_1,x_1)(y_1,x_1)\cdot x_2} &= \begin{pmatrix} \sigma_{y_1y_1} & \sigma_{y_1x_1} \\ \sigma_{y_1x_1} & \sigma_{x_1x_1} \end{pmatrix} - \mathbf{\Sigma}_{(y_1,x_1)(x_2)} \sigma^{-1}_{x_2x_2} \mathbf{\Sigma}'_{(y_1x_1)(x_2)}. \end{aligned}$$

Case 8. *Regression of (nonsensitive and sensitive) on (nonsensitive and sensitive).* Suppose the analysis model is the conditional distribution of $\{(y_{1i}, x_{1i}), i = 1, \ldots, n\}$ given $\{(y_{2i}, x_{2i}), i = 1, \ldots, n\}$. In the conditional distribution of $(y_{11}, x_{11})' \ldots, (y_{1n}, x_{1n})'$ given $(y_{21}, x_{21})', \ldots, (y_{2n}, x_{2n})'$, the random vectors $(y_{11}, x_{11})', \ldots, (y_{1n}, x_{1n})'$ are independently distributed such that

$$\begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix}$$

$$\sim N_2 \left\{ \begin{pmatrix} \mu_{y_1} \\ \mu_{x_1} \end{pmatrix} + \mathbf{\Sigma}_{(y_1,x_1)(y_2,x_2)} \mathbf{\Sigma}^{-1}_{(y_2,x_2)(y_2,x_2)} \begin{pmatrix} y_{2i} - \mu_{y_2} \\ x_{2i} - \mu_{x_2} \end{pmatrix} = \mathbf{\Omega}'_8 \begin{pmatrix} 1 \\ y_{2i} \\ x_{2i} \end{pmatrix}, \mathbf{\Sigma}_{(y_1,x_1)(y_1,x_1)\cdot(y_2,x_2)} \right\}$$

for $i = 1, \ldots, n$, where

$$\boldsymbol{\Sigma}_{(y_1,x_1)(y_2,x_2)} = \begin{pmatrix} \sigma_{y_1y_2} & \sigma_{y_1x_2} \\ \sigma_{y_2x_1} & \sigma_{x_1x_2} \end{pmatrix}, \ \boldsymbol{\Sigma}_{(y_2,x_2)(y_2,x_2)} = \begin{pmatrix} \sigma_{y_2y_2} & \sigma_{y_2x_2} \\ \sigma_{y_2x_2} & \sigma_{x_2x_2} \end{pmatrix}$$

$$\boldsymbol{\Omega}_8' = \Big((\mu_{y_1}, \mu_{x_1})' - \boldsymbol{\Sigma}_{(y_1,x_1)(y_2,x_2)} \boldsymbol{\Sigma}^{-1}_{(y_2,x_2)(y_2,x_2)} (\mu_{y_2}, \mu_{x_2})', \quad \boldsymbol{\Sigma}_{(y_1,x_1)(y_2,x_2)} \boldsymbol{\Sigma}^{-1}_{(y_2x_2)(y_2x_2)} \Big)_{2\times 3}.$$

$$\boldsymbol{\Sigma}_{(y_1,x_1)(y_1,x_1)\cdot(y_2x_2)} = \begin{pmatrix} \sigma_{y_1y_1} & \sigma_{y_1x_1} \\ \sigma_{y_1x_1} & \sigma_{x_1x_1} \end{pmatrix} - \boldsymbol{\Sigma}_{(y_1,x_1)(y_2,x_2)} \boldsymbol{\Sigma}^{-1}_{(y_2,x_2)(y_2,x_2)} \boldsymbol{\Sigma}'_{(y_1,x_1)(y_2,x_2)}.$$

In the following simulation results, we show the coverage of the data analyst's confidence region for the regression parameters $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_8$ in each of the Cases 1-8. Although the methodology of Sections 2 and 3 was not designed for the range of scenarios in Cases 1-8, our goal is to evaluate the performance of these methods in these cases. Thus we assume that the data analyst constructs a point estimate and confidence region for some desired regression parameters using the methodology in Sections 2 and 3, with the definitions of $\boldsymbol{X}$ and $\boldsymbol{V}$ appropriately modified to reflect the analyst's choice of the regression model. For example, in Case 8 the data analyst is interested in the regression model where $\{(y_{1i}, x_{1i}), i = 1, \ldots, n\}$ are the response variables and $\{(y_{2i}, x_{2i}), i = 1, \ldots, n\}$ are the regressor variables. Thus in Case 8 the target parameter is $\boldsymbol{\Omega}_8$, and upon observing the released data, we assume the analyst constructs a confidence region for $\boldsymbol{\Omega}_8$ using the methodology of Sections 2 and 3 treating $\begin{pmatrix} v_{21} & v_{22} & \cdots & v_{2n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix}$ as the $\boldsymbol{X}$ matrix, and $\begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ x_{11} & x_{12} & \cdots & x_{1n} \end{pmatrix}$ as the $\boldsymbol{V}$ matrix. For the sake of comparison, we also show the coverage of the confidence region obtained using the methodology of Reiter (2005) in each case. Table 7 provides Monte Carlo estimates of $||E(\hat{\boldsymbol{\Omega}}_j) - \boldsymbol{\Omega}_j||_1$ for each of the target parameters $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_8$ in Cases 1-8. Table 8 provides Monte Carlo estimates of confidence region coverage probability for each of the target parameters in $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_8$ in Cases 1-8 .

To conduct the simulation study, we set the model parameters as follows:

$$\boldsymbol{\mu}_x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \ \boldsymbol{\mu}_y = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} \sigma_{y_1y_1} & \sigma_{y_1y_2} & \sigma_{y_1x_1} & \sigma_{y_1x_2} \\ \sigma_{y_1y_2} & \sigma_{y_2y_2} & \sigma_{y_2x_1} & \sigma_{y_2x_2} \\ \sigma_{y_1x_1} & \sigma_{y_2x_1} & \sigma_{x_1x_1} & \sigma_{x_1x_2} \\ \sigma_{y_1x_2} & \sigma_{y_2x_2} & \sigma_{x_1x_2} & \sigma_{x_2x_2} \end{pmatrix} = \begin{pmatrix} 1 & 1.5 & 0.1 & 0.2 \\ 1.5 & 3 & 0.1 & 0.3 \\ 0.1 & 0.1 & 3 & 1.6 \\ 0.2 & 0.3 & 1.6 & 1 \end{pmatrix}.$$

For the simulation study we take $n = 500$ which should be sufficiently large so that the procedures of Reiter (2005) can also be applied in each case. The results of Table 7 indicate that the point estimator of the regression parameter matrix (which for multiple imputation is the same under our 1st procedure, 2nd procedure, and the procedure of Reiter (2005)) is approximately unbiased in this scenario. In Case 1 the data analysis model is the regression of the sensitive variables on the non-sensitive variables, and this data analysis model is the one that was assumed to derive the procedures in Sections 2 and 3. However, the assumptions of Sections 2 and 3 do not quite hold in Case 1, because in Case 1 the regressor variables are random, while in Sections 2 and 3 the regressor variables are fixed. Nonetheless, we observe in Table 8 that the our proposed procedures, as well as the procedures of Reiter (2005) for multiple imputation, give confidence regions whose coverage is approximately equal to the nominal value of 0.95. In Cases 2-8 the data analysis regression is something other than the regression of the sensitive variables on the non-sensitive variables, and here we observe in Table 8 that the coverage of our proposed confidence regions tends to be greater than the nominal value. An exception is Case 5 where our 2nd proposed procedure for multiple imputation has coverage less than the nominal value. On the other hand, in each of the Cases 1-8, we observe in Table 8 that the confidence regions of Reiter (2005) have coverage probability approximately equal to the nominal value of 0.95.

7 Concluding Remarks

The data analysis methodology of Reiter (2003), Reiter (2005) and Raghunathan et al. (2003) are asymptotic in nature and can only be used when multiply imputed synthetic datasets are released, and, as such, are not meant for analysis of singly imputed synthetic data.

In this paper the authors derived likelihood-based exact inference for the single imputation case and also two exact likelihood-based solutions are offered for the case when multiple synthetic datasets are released. Inference procedures were obtained for the matrix regression coefficients matrix under a Multivariate Linear Regression Model when synthetic data are generated via Plug-in Sampling. The simulation studies showed that the methodologies developed lead to confidence sets with the expected level of confidence, even for small sample sizes, both for single and multiple imputation. Our simulations also reveal that as the number of synthetic datasets releases increase, the inference derived from synthetic dataset comes closer to the one based on the original data, but of course at the expense of compromising privacy, namely by increasing the disclosure risk. It turns out that the second procedure proposed for multiple synthetic data has a better performance than the first one for small sample sizes, and their performances are nearly the same for larger sample sizes.

It will be interesting to compare the performance of our two exact test procedures with the asymptotic procedure of Reiter (2005) for a real large dataset. It is expected that all three methods will provide similar performances. Moreover, the adapted procedure of Reiter (2005) is the only method to be used in the multiple imputation setting when the number of tested regressors is smaller than the number of response variables.

References

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Wiley.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*(2), 171–178.

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

Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics 13*(4), 271–275.

Klein, M., T. Mathew, and B. Sinha (2013). A comparison of statistical disclosure control methods: Multiple imputation versus noise multiplication. *U.S. Census Bureau Research Report Series*.

Klein, M. and B. Sinha (2013). Statistical analysis of noise multiplied data using multiple imputation. *Journal of Official Statistics 29*, 425–465.

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

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

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

Klein, M. and B. Sinha (2016). Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models. *Journal of Privacy and Confidentiality 7*(1), 43–98.

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

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

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

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

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

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

Reiter, J. and S. Kinney (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics 28*, 583–590.

Reiter, J. and T. Raghunathan (2007). The multiple adaptations of multiple imputation. *Journal of American Statistical Association 102*, 1462–1471.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

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

Seber, G. and A. Lee (2003). *Linear Regression Analysis* (2nd ed.). Wiley.

Table 5: Average coverage for **B** and **AB** when error distribution is multivariate t

n	ν	$M=1$	$M=2$			$M=5$			$M=10$			$M=20$		
			vec(AB)	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)
10	3	0.944	0.822	0.944	0.945	0.742	0.946	0.944	0.735	0.944	0.944	0.739	0.943	0.943
	5	0.946	0.828	0.944	0.947	0.743	0.948	0.946	0.735	0.945	0.946	0.742	0.945	0.946
	10	0.951	0.830	0.951	0.949	0.753	0.953	0.950	0.751	0.950	0.950	0.750	0.951	0.951
	15	0.950	0.828	0.951	0.950	0.751	0.950	0.949	0.744	0.949	0.950	0.747	0.950	0.948
	20	0.950	0.827	0.947	0.948	0.752	0.948	0.948	0.743	0.948	0.947	0.747	0.948	0.948
20	3	0.952	0.925	0.953	0.951	0.881	0.951	0.951	0.877	0.951	0.954	0.879	0.952	0.952
	5	0.948	0.922	0.948	0.947	0.881	0.947	0.948	0.872	0.950	0.951	0.874	0.951	0.950
	10	0.951	0.921	0.952	0.950	0.876	0.948	0.948	0.867	0.948	0.948	0.873	0.948	0.947
	15	0.952	0.926	0.952	0.950	0.880	0.949	0.949	0.879	0.951	0.952	0.879	0.953	0.953
	20	0.951	0.925	0.951	0.951	0.876	0.951	0.950	0.874	0.950	0.952	0.875	0.952	0.951
50	3	0.948	0.960	0.951	0.952	0.931	0.952	0.951	0.930	0.954	0.953	0.932	0.953	0.952
	5	0.948	0.953	0.947	0.948	0.927	0.951	0.950	0.926	0.950	0.949	0.929	0.951	0.950
	10	0.951	0.952	0.949	0.950	0.925	0.951	0.950	0.921	0.949	0.948	0.924	0.950	0.948
	15	0.950	0.950	0.949	0.950	0.928	0.950	0.949	0.924	0.952	0.951	0.924	0.950	0.949
	20	0.949	0.956	0.952	0.953	0.929	0.955	0.955	0.924	0.952	0.950	0.929	0.952	0.951
100	3	0.950	0.963	0.950	0.951	0.940	0.950	0.952	0.938	0.953	0.953	0.940	0.952	0.951
	5	0.952	0.962	0.950	0.950	0.938	0.949	0.951	0.938	0.950	0.950	0.937	0.949	0.947
	10	0.949	0.961	0.950	0.951	0.939	0.951	0.952	0.933	0.952	0.953	0.938	0.950	0.948
	15	0.951	0.963	0.949	0.949	0.944	0.951	0.952	0.940	0.949	0.950	0.944	0.951	0.950
	20	0.948	0.965	0.947	0.947	0.944	0.954	0.955	0.942	0.951	0.952	0.942	0.952	0.951
200	3	0.950	0.966	0.948	0.947	0.943	0.951	0.951	0.940	0.949	0.949	0.944	0.948	0.949
	5	0.954	0.965	0.950	0.948	0.943	0.946	0.946	0.939	0.948	0.948	0.943	0.946	0.947
	10	0.952	0.963	0.950	0.948	0.947	0.948	0.948	0.942	0.947	0.947	0.942	0.947	0.948
	15	0.950	0.963	0.950	0.949	0.947	0.951	0.951	0.942	0.950	0.950	0.944	0.947	0.948
	20	0.948	0.964	0.951	0.950	0.944	0.951	0.952	0.939	0.950	0.950	0.942	0.949	0.950

(a) Average coverage for **B**

n	ν	$M=1$	$M=2$			$M=5$			$M=10$			$M=20$		
			vec(AB)	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)
10	3	0.943	0.956	0.944	0.944	0.777	0.945	0.944	0.776	0.944	0.944	0.772	0.943	0.943
	5	0.944	0.958	0.946	0.946	0.778	0.945	0.945	0.781	0.944	0.944	0.776	0.942	0.943
	10	0.949	0.965	0.948	0.949	0.790	0.951	0.952	0.797	0.950	0.949	0.787	0.949	0.950
	15	0.952	0.967	0.951	0.951	0.798	0.953	0.953	0.799	0.950	0.950	0.789	0.951	0.951
	20	0.949	0.962	0.951	0.949	0.791	0.949	0.948	0.794	0.948	0.948	0.787	0.948	0.949
20	3	0.956	0.996	0.949	0.950	0.898	0.954	0.952	0.896	0.949	0.950	0.894	0.952	0.952
	5	0.949	0.995	0.950	0.950	0.896	0.953	0.952	0.896	0.950	0.951	0.891	0.949	0.949
	10	0.950	0.994	0.950	0.950	0.892	0.951	0.950	0.889	0.945	0.947	0.886	0.948	0.948
	15	0.952	0.996	0.950	0.953	0.896	0.952	0.950	0.899	0.948	0.949	0.896	0.949	0.949
	20	0.950	0.996	0.952	0.954	0.892	0.949	0.949	0.894	0.950	0.950	0.891	0.951	0.952
50	3	0.948	0.999	0.950	0.950	0.933	0.952	0.952	0.935	0.953	0.952	0.933	0.954	0.952
	5	0.949	0.999	0.950	0.951	0.934	0.952	0.951	0.932	0.952	0.950	0.933	0.953	0.952
	10	0.952	0.999	0.952	0.952	0.930	0.953	0.951	0.928	0.951	0.950	0.927	0.949	0.947
	15	0.955	0.999	0.955	0.955	0.935	0.953	0.952	0.935	0.954	0.952	0.932	0.951	0.951
	20	0.948	0.999	0.948	0.948	0.935	0.952	0.951	0.934	0.950	0.949	0.934	0.951	0.950
100	3	0.948	1.000	0.950	0.952	0.941	0.950	0.951	0.942	0.951	0.952	0.939	0.951	0.952
	5	0.950	1.000	0.947	0.948	0.944	0.952	0.952	0.943	0.950	0.952	0.940	0.949	0.950
	10	0.952	1.000	0.948	0.948	0.937	0.947	0.947	0.937	0.948	0.950	0.938	0.947	0.948
	15	0.949	1.000	0.950	0.950	0.940	0.950	0.950	0.939	0.950	0.952	0.941	0.951	0.951
	20	0.951	1.000	0.950	0.951	0.945	0.950	0.950	0.944	0.951	0.952	0.943	0.951	0.952
200	3	0.950	1.000	0.953	0.952	0.943	0.951	0.952	0.944	0.954	0.951	0.943	0.952	0.951
	5	0.957	1.000	0.953	0.952	0.949	0.950	0.951	0.947	0.953	0.951	0.946	0.952	0.952
	10	0.952	1.000	0.955	0.954	0.944	0.948	0.948	0.942	0.949	0.948	0.942	0.948	0.947
	15	0.952	1.000	0.952	0.952	0.945	0.952	0.953	0.946	0.952	0.949	0.944	0.951	0.950
	20	0.948	1.000	0.953	0.952	0.947	0.952	0.953	0.947	0.954	0.951	0.946	0.953	0.952

(b) Average coverage for **AB**

Table 6: Average coverage for **B** and **AB** when error distribution is skew-normal

n	λ	$M=1$	$M=2$			$M=5$			$M=10$			$M=20$		
			vec(AB)	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)
10	10	0.944	0.817	0.942	0.941	0.738	0.944	0.942	0.731	0.942	0.941	0.734	0.942	0.94
	-10	0.949	0.819	0.942	0.942	0.736	0.944	0.942	0.728	0.945	0.944	0.729	0.944	0.943
	100	0.944	0.815	0.942	0.941	0.739	0.943	0.942	0.73	0.941	0.943	0.731	0.942	0.942
	-100	0.95	0.816	0.944	0.944	0.735	0.944	0.942	0.726	0.943	0.942	0.73	0.944	0.942
50	10	0.942	0.947	0.941	0.944	0.917	0.945	0.945	0.913	0.941	0.944	0.917	0.946	0.945
	-10	0.955	0.954	0.95	0.951	0.922	0.946	0.945	0.913	0.942	0.945	0.918	0.944	0.943
	100	0.943	0.948	0.942	0.944	0.918	0.946	0.945	0.914	0.944	0.946	0.916	0.944	0.944
	-100	0.953	0.953	0.951	0.952	0.923	0.946	0.945	0.913	0.943	0.945	0.917	0.944	0.943
200	10	0.952	0.971	0.953	0.952	0.95	0.957	0.956	0.943	0.953	0.953	0.943	0.949	0.95
	-10	0.948	0.964	0.947	0.946	0.947	0.949	0.949	0.943	0.953	0.952	0.948	0.953	0.954
	100	0.953	0.969	0.953	0.952	0.947	0.956	0.956	0.945	0.953	0.953	0.941	0.95	0.952
	-100	0.95	0.963	0.95	0.948	0.948	0.949	0.949	0.946	0.95	0.95	0.947	0.953	0.954

(a) Average coverage for **B**

n	λ	$M=1$	$M=2$			$M=5$			$M=10$			$M=20$		
			vec(AB)	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)	**vec(AB)**	**AB** (1)	**AB** (2)
10	10	0.95	0.964	0.95	0.95	0.798	0.95	0.95	0.796	0.949	0.948	0.793	0.95	0.95
	-10	0.953	0.965	0.952	0.95	0.792	0.95	0.95	0.798	0.951	0.95	0.791	0.95	0.95
	100	0.948	0.967	0.95	0.95	0.795	0.95	0.95	0.799	0.952	0.95	0.791	0.95	0.951
	-100	0.953	0.965	0.952	0.951	0.791	0.951	0.95	0.797	0.95	0.95	0.789	0.948	0.949
50	10	0.946	0.999	0.944	0.945	0.927	0.946	0.944	0.928	0.949	0.947	0.927	0.95	0.948
	-10	0.95	0.999	0.95	0.95	0.928	0.948	0.948	0.926	0.95	0.948	0.924	0.948	0.947
	100	0.945	0.999	0.946	0.946	0.926	0.947	0.946	0.928	0.95	0.948	0.926	0.951	0.949
	-100	0.951	1.	0.953	0.953	0.928	0.947	0.946	0.926	0.948	0.947	0.924	0.948	0.947
200	10	0.95	1.	0.955	0.954	0.949	0.953	0.954	0.947	0.951	0.95	0.946	0.956	0.955
	-10	0.951	1.	0.951	0.951	0.944	0.946	0.947	0.947	0.951	0.949	0.945	0.95	0.95
	100	0.953	1.	0.955	0.954	0.949	0.951	0.953	0.947	0.953	0.951	0.949	0.954	0.952
	-100	0.951	1.	0.952	0.95	0.943	0.948	0.948	0.944	0.952	0.95	0.946	0.951	0.95

(b) Average coverage for **AB**

Table 7: Monte Carlo estimates of $||E(\hat{\mathbf{\Omega}}_j) - \mathbf{\Omega}_j||_1$ in each Case 1-8 for $n = 500$ and $\gamma = 0.05$.

	Target Parameter	$M = 1$	$M = 10$
Case 1	$\mathbf{\Omega}_1$	0.00832404	0.00725665
Case 2	$\mathbf{\Omega}_2$	0.00327056	0.00196958
Case 3	$\mathbf{\Omega}_3$	0.0055076	0.00545225
Case 4	$\mathbf{\Omega}_4$	0.00493499	0.00343167
Case 5	$\mathbf{\Omega}_5$	0.00233433	0.00183536
Case 6	$\mathbf{\Omega}_6$	0.00714904	0.00616492
Case 7	$\mathbf{\Omega}_7$	0.000624239	0.000564305
Case 8	$\mathbf{\Omega}_8$	0.00256119	0.00201542

Table 8: Monte Carlo estimate of confidence region coverage probabilities in each Case 1-8 for $n = 500$ and $\gamma = 0.05$.

	Target Parameter	$M = 1$	$M = 10$		
			1st Approach	2nd Approach	Reiter (2005)
Case 1	$\mathbf{\Omega}_1$	0.951	0.952	0.952	0.947
Case 2	$\mathbf{\Omega}_2$	0.953	0.985	1.000	0.946
Case 3	$\mathbf{\Omega}_3$	0.980	1.000	0.995	0.946
Case 4	$\mathbf{\Omega}_4$	0.951	0.976	1.000	0.949
Case 5	$\mathbf{\Omega}_5$	0.984	0.993	0.886	0.950
Case 6	$\mathbf{\Omega}_6$	0.973	0.998	1.000	0.947
Case 7	$\mathbf{\Omega}_7$	0.987	0.956	0.957	0.950
Case 8	$\mathbf{\Omega}_8$	0.977	0.990	0.997	0.947

Appendices to "Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling"

Appendix A

Proof of Theorem 2.1. : From (5), given $(\hat{\mathbf{B}}, \mathbf{S})$, we have for $\mathbf{B}^*$ and $\mathbf{S}^*$ in (6),

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

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

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

while, given the independence of $\hat{\mathbf{B}}$ and $\mathbf{S}$, defined in the Introduction after (1), the joint pdf of $(\hat{\mathbf{B}}, \mathbf{S})$ is proportional to

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

Therefore, we obtain the joint pdf of $(\mathbf{B}^*, \mathbf{S}^*, \hat{\mathbf{B}}, \mathbf{S})$ by multiplying the two joint pdf's (A.1) and (A.2).

Since

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

where, from the identities in Appendix B.1,

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

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

$$\exp\left\{-\frac{1}{2}tr\left[(\mathbf{\Sigma} + \mathbf{S})^{-1}(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B}) + (n-p)\mathbf{S}^{-\mathbf{1}}\mathbf{S}^* + (n-p)\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}\right]\right\}$$
$$\times |\mathbf{S}^*|^{\frac{n-p-m-1}{2}} \times \frac{|\mathbf{S}|^{-\frac{p+m+1}{2}}}{|\mathbf{\Sigma}|^{\frac{n}{2}}} \times |\mathbf{\Sigma}^{-\mathbf{1}} + \mathbf{S}^{-1}|^{-p/2}. \quad \text{(A.3)}$$

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

Proof of Theorem 2.2. : In (A.3), $\mathbf{S}^*$ and $\mathbf{B}^*$, conditional on $\mathbf{S}$, are separable, with $\mathbf{B}^* \sim N_{pm}(\mathbf{B}, (\mathbf{\Sigma} + \mathbf{S}) \otimes (\mathbf{X}\mathbf{X}')^{-\mathbf{1}})$ and $(n-p)\mathbf{S}^* \sim W_m(\mathbf{S}, n-p)$, independent of $\mathbf{B}^*$.

Then, $(\mathbf{B}^* - \mathbf{B})'|_{\mathbf{S}} \sim N(\mathbf{0}, (\mathbf{X}\mathbf{X}')^{-\mathbf{1}} \otimes (\mathbf{\Sigma} + \mathbf{S}))$, and by Theorem 2.4.1 in Kollo and Rosen (2005) we have that, for $p \geq m$,

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

Therefore, from Theorem 2.4.2 in Kollo and Rosen (2005) and subsection 7.3.3 in Anderson (2003) we have

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

where $\mathbf{H}$ and $\mathbf{G}$ are two independent random variables.

Since we may write

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

where, given $\mathbf{S}$, $|\mathbf{G}| \sim \prod_{i=1}^m \chi^2_{n-p-i+1}$ and $|\mathbf{H}| \sim \prod_{i=1}^m \chi^2_{p-i+1}$, with the chi-square random variables in each product independent, we end up with a product of independent F-distributions. So, conditionally on $\mathbf{S}$, we have

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

Note that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$, thus implying $\frac{1}{n-p}\mathbf{S}^{-1} \sim W_m^{-1}((n-p)\mathbf{\Sigma^{-1}}, n-p)$, or $\frac{1}{n-p}\mathbf{\Sigma}^{1/2}\mathbf{S}^{-1}\mathbf{\Sigma}^{1/2} \sim W_m^{-1}(\mathbf{I}, n-p)$, which shows that the distribution of $|\mathbf{S^{-1}}(\mathbf{\Sigma}+\mathbf{S})| = |\mathbf{\Sigma^{1/2}S^{-1}\Sigma^{1/2}} + \mathbf{I}|$ is independent of $\mathbf{\Sigma}$, concluding the proof. □

Proof of Corollary 3.1. : The proof is identical to the proof of Theorem 2.1 replacing, conditional on $\hat{\mathbf{B}}$ and $\mathbf{S}$, the joint pdf of $(\mathbf{B}^*, \mathbf{S}^*)$ by the joint pdf of $(\overline{\mathbf{B}}_M^*, \overline{\mathbf{S}}_M^*)$ and observing that

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

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

Proof of Corollary 3.2. : The proof is identical to the proof of Theorem 2.2 replacing, conditional on $\mathbf{S}$, $\mathbf{B}^*$ and $\mathbf{S}^*$ by $\overline{\mathbf{B}}_M^*$ and $\overline{\mathbf{S}}_M^*$, noting that from the distribution in Corollary 3.1 we have that $\overline{\mathbf{B}}_M^*$ has $N_{pm}(\mathbf{B}, (\mathbf{\Sigma} + \frac{\mathbf{1}}{\mathbf{M}}\mathbf{S})\otimes(\mathbf{XX}')^{-\mathbf{1}})$ distribution and that $M(n-p)\overline{\mathbf{S}}_M^*$ is $W_m(\mathbf{S}, M(n-p))$, independent of $\overline{\mathbf{B}}_M^*$. □

Proof of Corollary 3.3. The proof is identical to the proof of Theorem 2.1 replacing, conditional on $\hat{\mathbf{B}}$ and $\mathbf{S}$, the joint pdf of $(\mathbf{B}^*, \mathbf{S}^*)$ by the joint pdf of $(\overline{\mathbf{B}}_M^*, \mathbf{S}_{comb})$, noting that we have

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

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

Proof of Corollary 3.4. The proof is identical to the proof of Theorem 2.2 replacing, conditional on $\mathbf{S}$, $\mathbf{B}^*$ and $\mathbf{S}^*$ by $\overline{\mathbf{B}}_M^*$ and $\mathbf{S}_{comb}$, noting that from the distribution of Corollary 3.3 we have that $\overline{\mathbf{B}}_M^*$ has $N_{pm}(\mathbf{B}, (\mathbf{\Sigma} + \frac{\mathbf{1}}{\mathbf{M}}\mathbf{S})\otimes(\mathbf{XX}')^{-\mathbf{1}})$ distribution and that $(Mn-p)\mathbf{S}_{comb}$ is $W_m(\mathbf{S}, Mn-p)$, independent of $\overline{\mathbf{B}}_M^*$. □

Appendix B

B.1

Some matrix identities and matrix calculations required in the proof of Theorem 2.1.

1. If the matrices A and B are p.d. then
   (*i*) $\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}$ and
   (*ii*) $(\mathbf{A^{-1}}+\mathbf{B^{-1}})^{-\mathbf{1}} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{-\mathbf{1}}\mathbf{B} = \mathbf{B}(\mathbf{A}+\mathbf{B})^{-\mathbf{1}}\mathbf{A}$.

2. Let $\mathbf{S}$ and $\mathbf{\Sigma}$ be symmetric:

$$(\mathbf{C}-\mathbf{X})\mathbf{S}^{-1}(\mathbf{C}-\mathbf{X})' + (\mathbf{X}-\mathbf{D})\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{D})' =$$
$$\left[\mathbf{X}-(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\mathbf{\Sigma}^{-1})(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\right](\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})\left[\mathbf{X}-(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\mathbf{\Sigma}^{-1})(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\right]'$$
$$+\mathbf{C}\mathbf{S}^{-1}\mathbf{C}'+\mathbf{D}\mathbf{\Sigma}^{-1}\mathbf{D}'-(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\mathbf{\Sigma}^{-1})(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}(\mathbf{C}\mathbf{S}^{-1}+\mathbf{D}\mathbf{\Sigma}^{-1})'.$$

3. Taking the last three terms of the previous sum, using identities from item 1, we have,

$$\mathbf{C}\mathbf{S}^{-1}\mathbf{C}'-\mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{S}^{-1}\mathbf{C}'+\mathbf{D}\mathbf{\Sigma}^{-1}\mathbf{D}'-\mathbf{D}\mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1}\mathbf{D}'$$
$$-\mathbf{C}\mathbf{S}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1}\mathbf{D}'-\mathbf{D}\mathbf{\Sigma}^{-1}(\mathbf{S}^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{S}^{-1}\mathbf{C}'=$$
$$\mathbf{C}(\mathbf{S}+\mathbf{\Sigma})^{-1}(\mathbf{C}'-\mathbf{D}')+\mathbf{D}(\mathbf{S}+\mathbf{\Sigma})^{-1}(\mathbf{D}'-\mathbf{C}')=(\mathbf{C}-\mathbf{D})(\mathbf{S}+\mathbf{\Sigma})^{-1}(\mathbf{C}-\mathbf{D})'.$$

B.2

We provide some details about the derivations of the results in Section 2.
Details of Result 1: From the joint pdf in (A.3) we see immediately that the MLE of $\mathbf{B}$ is $\hat{\mathbf{B}}_{MLE}(\mathbf{V})=\mathbf{B}^*$ with $E(\mathbf{B}^*)=(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}E(\mathbf{X}'\hat{\mathbf{B}})=\mathbf{B}$ and that

$$Var(\mathbf{B}^*)=E(Var_{\mathbf{B}^*|\hat{\mathbf{B}},\mathbf{S}}(\mathbf{B}^*|\hat{\mathbf{B}},\mathbf{S}))+Var(E_{\mathbf{B}^*|\hat{\mathbf{B}},\mathbf{S}}(\mathbf{B}^*|\hat{\mathbf{B}},\mathbf{S}))=2\mathbf{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-1}.$$

Details of Result 2: Noting that $(n-p)\mathbf{S}^*|_{\mathbf{S}}\sim W_m(\mathbf{S},n-p)$ and that $(n-p)\mathbf{S}\sim W_m(\mathbf{\Sigma},n-p)$ then immediately $E(\mathbf{S}^*)=E(\mathbf{S})=\mathbf{\Sigma}$.
Details of Result 5: Let us star by writing $T_{\mathbf{\Delta}}$ as

$$T_{\Delta}=T_{\mathbf{\Delta}}^{(1)}\times T_{\mathbf{\Delta}}^{(2)}=\frac{|(\mathbf{\Delta}^*-\mathbf{\Delta})'[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}^*-\mathbf{\Delta})|}{|\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D}|}\times\frac{|\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D}|}{|(n-p)\mathbf{D}'\mathbf{S}^*\mathbf{D}|}.$$

Recalling that we had

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

in the proof of Theorem 2.2, we obtain

$$\mathbf{\Delta}^*|\mathbf{S}\sim N_{kr}[\mathbf{\Delta},\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D}\otimes\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}'].$$

Analogous to what was done in Theorem 2.2, we may conclude that

$$(\mathbf{\Delta}^*-\mathbf{\Delta})'[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}^*-\mathbf{\Delta})|\mathbf{S}\sim W_r[\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D},k]$$

and hence

$$T_{\mathbf{\Delta}}^{(1)}=\frac{|(\mathbf{\Delta}^*-\mathbf{\Delta})'[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}^*-\mathbf{\Delta})|}{|\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D}|}|\mathbf{S}\sim\prod_{i=1}^{r}\chi^2_{k-i+1},$$

which will be independent of $\mathbf{S}$. Let us write $T_{\mathbf{\Delta}}^{(2)}$ as

$$T_{\mathbf{\Delta}}^{(2)} = \frac{|\mathbf{D}'\mathbf{S}\mathbf{D}|}{|(n-p)\mathbf{D}'\mathbf{S}^*\mathbf{D}|} \times \frac{|\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D}|}{|\mathbf{D}'\mathbf{S}\mathbf{D}|}$$

Recalling that $(n-p)\mathbf{S}^*|\mathbf{S} \sim W_m(\mathbf{S}, n-p)$ from Theorem 2.2, we analogously obtain

$$|(n-p)\mathbf{D}'\mathbf{S}^*\mathbf{D}|/|\mathbf{D}'\mathbf{S}\mathbf{D}||\mathbf{S} \sim \prod_{i=1}^{r} \chi^2_{n-p-i+1},$$

which will also be independent of $\mathbf{S}$. Lastly, it is easy to show by standard arguments that

$$\frac{|\mathbf{D}'(\mathbf{\Sigma}+\mathbf{S})\mathbf{D}|}{|\mathbf{D}'\mathbf{S}\mathbf{D}|} \sim \frac{|\mathbf{W}^* + \mathbf{I}_r(n-p)|}{|\mathbf{W}^*|} \sim |\mathbf{I}_r + (n-p)\mathbf{W}^{*-1}|,$$

where $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, which is independent of $\mathbf{S}$. Combining the above terms, we conclude that

$$T_{\mathbf{\Delta}} \overset{st}{\sim} \left\{ \prod_{i=1}^{r} \frac{k-i+1}{n-p-i+1} F_i \right\} \frac{|\mathbf{W}^* + (n-p)\mathbf{I}_r|}{|\mathbf{W}^*|}$$

where $F_i \sim F_{k-i+1,n-p-i+1}$, completing the proof.

Adaptations of classical test criterion that are not pivotal: Let us consider $\mathbf{H}$ and $\mathbf{G}$ as we did in Appendix A. We will begin to decompose all the four statistics in order to assume the same kind of form and then prove why all of them are non-pivotal. The first statistic is

$$T_1 = \frac{|\mathbf{S}^*|}{|\mathbf{S}^* + (\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|}$$

that we can decompose as

$$\begin{aligned} T_1 &= \frac{|\mathbf{S}||(n-p)\mathbf{S}^{-1/2}\mathbf{S}^*\mathbf{S}^{-1/2}|}{(n-p)^m|\mathbf{S}^* + (\mathbf{\Sigma}+\mathbf{S})^{1/2}(\mathbf{\Sigma}+\mathbf{S})^{-1/2}(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})(\mathbf{\Sigma}+\mathbf{S})^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}|} \\ &= \frac{|\mathbf{G}|}{|\mathbf{G} + (n-p)\mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{H}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2}|}. \end{aligned}$$

Now let us consider the following statistics

$$T_2 = (n-p)tr\left[\mathbf{H} \times (\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2} \times \mathbf{G}^{-1} \times \mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\right],$$

$$T_3 = tr\{\mathbf{H} \times [\mathbf{H} + (\mathbf{\Sigma}+\mathbf{S})^{-1/2}\mathbf{S}^{1/2} \times (n-p)\mathbf{G} \times \mathbf{S}^{1/2}(\mathbf{\Sigma}+\mathbf{S})^{-1/2}]^{-1}\}$$

and $T_4 = \lambda_1$ where λ_1 denotes the largest eigenvalue of

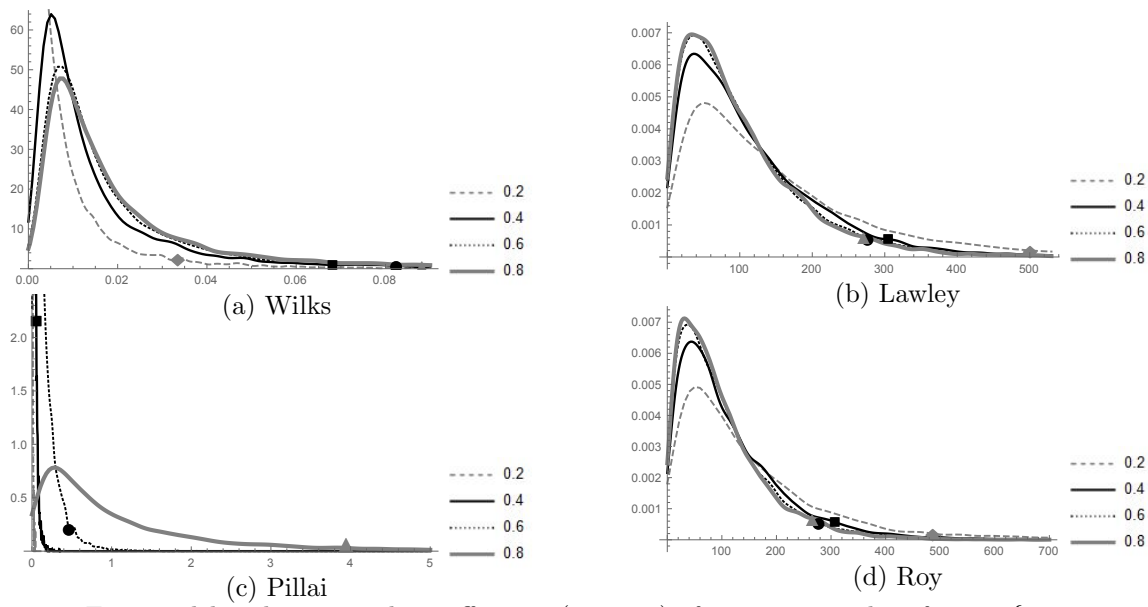
$$(n-p)\mathbf{H} \times (\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2} \times \mathbf{G}^{-1} \times \mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}.$$

From T_1 we can observe that a term of the denominator is

$$\mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{H}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2}|_{\mathbf{S}} \sim W_m(\mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})\mathbf{S}^{-1/2}, p) \equiv W_m((\mathbf{S}^{-1/2}\mathbf{\Sigma}\mathbf{S}^{-1/2}+\mathbf{I}), p),$$

and in the other statistics there are similar terms. We can also observe that all of the terms involve a product similar to $\mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}$ that cannot be simplified the same way we could do when using the determinant as in the statistic T used in this paper.

Thus, in order to prove that these statistics are dependent on $\mathbf{\Sigma}$, we can see the empirical distributions of T_1, T_2, T_3 and T_4 when we consider a simple case where $m = 2$, $p = 3$, $n = 100$ and $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ with $\rho = \{0.2, 0.4, 0.6, 0.8\}$ for a simulation size of 10^4, in Figure 1. After making the above simulations we can observe from its distributions and cut-off



(a) Wilks (b) Lawley (c) Pillai (d) Roy

Figure 1: Empirical distributions and cut-off points (γ=0.05) of T_1, T_2, T_3 and T_4 for $\rho = \{0.2,0.4,0.6,0.8\}$.

points ($\gamma = 0.05$) that these four statistics are non-pivotal. We also see that the statistic based on the ratio between $(\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})$ and $\mathbf{S}^*$ was the best try to find a pivotal statistic.

B.3

We provide some details about the derivations of the results in Section 3.1.

Details of Result 1: $E(\overline{\mathbf{B}}^*_M) = (\mathbf{XX'})^{-\mathbf{1}}\mathbf{X}\frac{1}{M}\sum_{i=1}^{M} E(\mathbf{V'}_i) = (\mathbf{XX'})^{-\mathbf{1}}\mathbf{X}\frac{1}{M}\sum_{i=1}^{M} E(\mathbf{X'}\hat{\mathbf{B}}) = \mathbf{B}$ and that

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

Details of Result 2: Noting that $M(n-p)\overline{\mathbf{S}}^*_M|_{\mathbf{S}} \sim W_m(\mathbf{S}, M(n-p))$ and that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$ then, immediately, $E(\overline{\mathbf{S}}^*_M) = E(\mathbf{S}) = \mathbf{\Sigma}$.

B.4

We provide some details about the derivations of the results in subsection 3.2.
Details of Result 1: Noting that $(Mn-p)\mathbf{S}_{comb}|_{\mathbf{S}} \sim W_m(\mathbf{S}, Mn-p)$ and that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$ then, immediately, $E(\mathbf{S}_{comb}) = E(\mathbf{S}) = \mathbf{\Sigma}$.

B.5

Lastly, we provide some details about the derivations of the results in section 4.
Details of Expect Values in Section 4: Recall that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$, thus implying that $E(|(n-p)\mathbf{S}|) = |\mathbf{\Sigma}|E(\prod_{i=1}^{m}\chi^2_{n-p-i+1}) = \frac{(n-p)!}{(n-p-m)!}|\mathbf{\Sigma}|$, since $\prod_{i=1}^{m}\chi^2_{n-p-i+1}$ is a product of independent χ^2 variables. Also recalling that, conditionally on $\mathbf{S}$, we have $(n-p)\mathbf{S}^* \sim W_m(\mathbf{S}, n-p)$, $M(n-p){\mathbf{S}^*}_M \sim W_m(\mathbf{S}, M(n-p))$ and $(Mn-p)\mathbf{S}_{comb} \sim W_m(\mathbf{S}, Mn-p)$, thus concluding that, conditionally on $\mathbf{S}$,

$$E(|(n-p)\mathbf{S}^*|) = \frac{1}{(n-p)^m} \times \frac{(n-p)!}{(n-p-m)!} \times |(n-p)\mathbf{S}|,$$

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

and

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

Combining the result of $E(|(n-p)\mathbf{S}|)$ with each of the synthetic expected values, conditionally on $\mathbf{S}$, we end up with the expression for $E(\Upsilon_M)$ found in Section 4.

B.6 Individual regressor coefficients confidence interval for the original data

$$
\begin{pmatrix}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
3.0276 & 3.45314 & 7.22404 & 7.81773 \\
0.101763 & 0.128775 & 0.50271 & 0.540397 \\
-0.100943 & -0.0676293 & -0.55975 & -0.513273 \\
0.000592196 & 0.00205233 & -0.0160693 & -0.0140322 \\
-0.140275 & 0.307819 & -0.392277 & 0.232881 \\
-0.181448 & 0.249481 & -0.339044 & 0.262165 \\
-0.0719615 & 0.345724 & -0.146577 & 0.436157 \\
-0.124093 & 0.302649 & -0.0634389 & 0.53193 \\
-0.065135 & 0.356647 & 0.0415413 & 0.629991 \\
-0.0445564 & 0.377282 & 0.00218244 & 0.59071 \\
-0.0126063 & 0.434573 & 0.0703166 & 0.694198 \\
0.155254 & 0.563845 & 0.512783 & 1.08283 \\
0.244623 & 0.654163 & 0.828421 & 1.39979 \\
0.237687 & 0.653551 & 0.812801 & 1.39299 \\
0.365105 & 0.782933 & 1.04715 & 1.63009 \\
0.400888 & 0.810689 & 1.39139 & 1.96312 \\
0.680126 & 1.09312 & 1.7045 & 2.28069 \\
0.686246 & 1.1133 & 2.21108 & 2.80688 \\
0.696918 & 1.12775 & 2.16138 & 2.76245 \\
-0.432626 & 0.0225628 & -0.420386 & 0.214671 \\
-0.152141 & 0.020387 & -0.560916 & -0.320213 \\
-0.0876056 & -0.0161402 & -0.513375 & -0.41367 \\
-0.118923 & -0.057579 & -0.397948 & -0.312365 \\
-0.143437 & 0.0075315 & -0.766689 & -0.556065 \\
-0.0741748 & -0.00465169 & -0.619728 & -0.522733 \\
-0.113329 & -0.0443148 & -0.174004 & -0.0777186 \\
-0.341987 & -0.162744 & -0.514953 & -0.264881 \\
-0.08713 & 0.0295535 & -0.0262251 & 0.136566 \\
-0.00459359 & 0.0347371 & -0.135873 & -0.0810009
\end{pmatrix},
\begin{pmatrix}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.24037 & 7.52088 \\
0.115269 & 0.521554 \\
-0.0842859 & -0.536512 \\
0.00132226 & -0.0150507 \\
0.0837721 & -0.0796979 \\
0.0340165 & -0.0384396 \\
0.136881 & 0.14479 \\
0.0892781 & 0.234245 \\
0.145756 & 0.335766 \\
0.166363 & 0.296446 \\
0.210983 & 0.382257 \\
0.35955 & 0.797806 \\
0.449393 & 1.11411 \\
0.445619 & 1.1029 \\
0.574019 & 1.33862 \\
0.605789 & 1.67726 \\
0.886624 & 1.9926 \\
0.899771 & 2.50898 \\
0.912332 & 2.46191 \\
-0.205032 & -0.102857 \\
-0.0658772 & -0.440565 \\
-0.0518729 & -0.463522 \\
-0.0882509 & -0.355157 \\
-0.0679526 & -0.661377 \\
-0.0394132 & -0.571231 \\
-0.0788219 & -0.125861 \\
-0.252366 & -0.389917 \\
-0.0287882 & 0.0551704 \\
0.0150718 & -0.108437
\end{pmatrix}
$$

B.7 Individual regressor coefficients confidence interval for the single synthetic data

$$
\begin{pmatrix}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
3.02817 & 3.6323 & 7.13692 & 7.98322 \\
0.0972511 & 0.1356 & 0.491859 & 0.545581 \\
-0.106496 & -0.059202 & -0.567982 & -0.50173 \\
-0.000170873 & 0.00190206 & -0.0158549 & -0.012951 \\
-0.386709 & 0.249444 & -0.562915 & 0.328241 \\
-0.339504 & 0.272278 & -0.58278 & 0.274238 \\
-0.242414 & 0.350569 & -0.362125 & 0.468556 \\
-0.249879 & 0.355959 & -0.312773 & 0.535917 \\
-0.179851 & 0.418947 & -0.16971 & 0.669118 \\
-0.229205 & 0.369672 & -0.189196 & 0.649742 \\
-0.109485 & 0.525368 & -0.0587802 & 0.830556 \\
-0.0070404 & 0.57303 & 0.321972 & 1.13457 \\
0.0776444 & 0.659063 & 0.627344 & 1.44183 \\
0.0440082 & 0.634403 & 0.650395 & 1.47745 \\
0.198631 & 0.791815 & 0.923888 & 1.75485 \\
0.231117 & 0.812905 & 1.18828 & 2.00328 \\
0.471254 & 1.05758 & 1.46725 & 2.28861 \\
0.491991 & 1.09827 & 1.94475 & 2.79406 \\
0.507036 & 1.11868 & 2.00075 & 2.85757 \\
-0.614778 & 0.0314465 & -0.642398 & 0.262868 \\
-0.192984 & 0.0519521 & -0.586147 & -0.243027 \\
-0.090293 & 0.0111654 & -0.54331 & -0.401182 \\
-0.114903 & -0.027814 & -0.389401 & -0.267403 \\
-0.141936 & 0.0723915 & -0.788623 & -0.488382 \\
-0.0592764 & 0.0394245 & -0.654749 & -0.516484 \\
-0.139882 & -0.041904 & -0.208622 & -0.0713688 \\
-0.417745 & -0.163276 & -0.574755 & -0.218281 \\
-0.154133 & 0.0115208 & -0.0584952 & 0.173561 \\
-0.0061631 & 0.0496741 & -0.144825 & -0.066605
\end{pmatrix},
\begin{pmatrix}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.33024 & 7.56007 \\
0.116426 & 0.51872 \\
-0.0828492 & -0.534856 \\
0.000865592 & -0.0144029 \\
-0.0686325 & -0.117337 \\
-0.0336128 & -0.154271 \\
0.0540774 & 0.0532158 \\
0.0530399 & 0.111572 \\
0.119548 & 0.249704 \\
0.0702339 & 0.230273 \\
0.207941 & 0.385888 \\
0.282995 & 0.728269 \\
0.368354 & 1.03459 \\
0.339206 & 1.06392 \\
0.495223 & 1.33937 \\
0.522011 & 1.59578 \\
0.764416 & 1.87793 \\
0.795129 & 2.3694 \\
0.812856 & 2.42916 \\
-0.291666 & -0.189765 \\
-0.0705159 & -0.414587 \\
-0.0395638 & -0.472246 \\
-0.0713583 & -0.328402 \\
-0.0347721 & -0.638503 \\
-0.00992597 & -0.585616 \\
-0.0908932 & -0.139996 \\
-0.290511 & -0.396518 \\
-0.0713061 & 0.0575331 \\
0.0217555 & -0.105715
\end{pmatrix}
$$

B.8 Individual regressor coefficients confidence interval for the multiple M=5 synthetic data (1st procedure)

$$
\begin{pmatrix}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
2.93058 & 3.39889 & 7.17972 & 7.83333 \\
0.0972081 & 0.126935 & 0.498923 & 0.540413 \\
-0.0968526 & -0.0601912 & -0.559855 & -0.508687 \\
0.000700045 & 0.00230693 & -0.0159042 & -0.0136615 \\
-0.108209 & 0.38492 & -0.423626 & 0.264626 \\
-0.147025 & 0.327213 & -0.381087 & 0.2808 \\
-0.0115419 & 0.448123 & -0.160998 & 0.480549 \\
-0.0618431 & 0.407787 & -0.0978739 & 0.557582 \\
-0.0053142 & 0.458858 & 0.00350566 & 0.651344 \\
-0.0116139 & 0.45262 & -0.0418757 & 0.606048 \\
0.0441611 & 0.536283 & 0.0454834 & 0.73233 \\
0.204335 & 0.653991 & 0.48795 & 1.11553 \\
0.29911 & 0.749811 & 0.803562 & 1.4326 \\
0.286226 & 0.743884 & 0.800312 & 1.43906 \\
0.410814 & 0.870635 & 1.02998 & 1.67174 \\
0.451483 & 0.90247 & 1.3587 & 1.98813 \\
0.724857 & 1.17936 & 1.67296 & 2.3073 \\
0.73176 & 1.20173 & 2.18023 & 2.83616 \\
0.727331 & 1.20146 & 2.13127 & 2.79301 \\
-0.489139 & 0.0117984 & -0.561753 & 0.137396 \\
-0.168175 & 0.0216926 & -0.529872 & -0.264877 \\
-0.0989116 & -0.0202637 & -0.548395 & -0.438628 \\
-0.121477 & -0.0539678 & -0.401783 & -0.307562 \\
-0.145168 & 0.0209728 & -0.824159 & -0.592279 \\
-0.066873 & 0.00963738 & -0.614174 & -0.507389 \\
-0.115578 & -0.0396281 & -0.163004 & -0.0570012 \\
-0.344065 & -0.146807 & -0.568873 & -0.293563 \\
-0.116207 & 0.0122035 & -0.0401409 & 0.13908 \\
-0.00421745 & 0.0390661 & -0.1346 & -0.0741894
\end{pmatrix},
\begin{pmatrix}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.16474 & 7.50652 \\
0.112072 & 0.519668 \\
-0.0785219 & -0.534271 \\
0.00150349 & -0.0147829 \\
0.138355 & -0.0794999 \\
0.0900943 & -0.0501434 \\
0.21829 & 0.159775 \\
0.172972 & 0.229854 \\
0.226772 & 0.327425 \\
0.220503 & 0.282086 \\
0.290222 & 0.388906 \\
0.429163 & 0.801739 \\
0.524461 & 1.11808 \\
0.515055 & 1.11969 \\
0.640725 & 1.35086 \\
0.676976 & 1.67342 \\
0.952109 & 1.99013 \\
0.966745 & 2.5082 \\
0.964395 & 2.46214 \\
-0.23867 & -0.212178 \\
-0.0732414 & -0.397375 \\
-0.0595876 & -0.493511 \\
-0.0877223 & -0.354672 \\
-0.0620976 & -0.708219 \\
-0.0286178 & -0.560782 \\
-0.0776033 & -0.110003 \\
-0.245436 & -0.431218 \\
-0.0520018 & 0.0494693 \\
0.0174243 & -0.104394
\end{pmatrix}
$$

B.9 Individual confidence intervals for the multiple $M = 5$ synthetic data (2nd procedure))

$$
\begin{pmatrix}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
2.93001 & 3.39946 & 7.17893 & 7.83412 \\
0.0971719 & 0.126972 & 0.498873 & 0.540463 \\
-0.0968972 & -0.0601466 & -0.559917 & -0.508625 \\
0.000698089 & 0.00230888 & -0.0159069 & -0.0136588 \\
-0.108809 & 0.38552 & -0.424461 & 0.265461 \\
-0.147602 & 0.327791 & -0.38189 & 0.281603 \\
-0.0121014 & 0.448682 & -0.161776 & 0.481327 \\
-0.0624147 & 0.408359 & -0.0986693 & 0.558377 \\
-0.00587915 & 0.459423 & 0.00271958 & 0.65213 \\
-0.0121789 & 0.453185 & -0.0426619 & 0.606834 \\
0.0435621 & 0.536882 & 0.0446499 & 0.733163 \\
0.203788 & 0.654538 & 0.487188 & 1.11629 \\
0.298562 & 0.75036 & 0.802799 & 1.43336 \\
0.285669 & 0.744441 & 0.799537 & 1.43983 \\
0.410254 & 0.871195 & 1.0292 & 1.67252 \\
0.450934 & 0.903019 & 1.35793 & 1.9889 \\
0.724303 & 1.17991 & 1.67219 & 2.30807 \\
0.731188 & 1.2023 & 2.17944 & 2.83696 \\
0.726754 & 1.20204 & 2.13047 & 2.79381 \\
-0.489748 & 0.0124081 & -0.562601 & 0.138245 \\
-0.168407 & 0.0219237 & -0.530194 & -0.264555 \\
-0.0990073 & -0.0201679 & -0.548528 & -0.438494 \\
-0.121559 & -0.0538856 & -0.401897 & -0.307447 \\
-0.14537 & 0.021175 & -0.82444 & -0.591997 \\
-0.0669662 & 0.00973051 & -0.614303 & -0.50726 \\
-0.115671 & -0.0395357 & -0.163132 & -0.0568725 \\
-0.344305 & -0.146567 & -0.569207 & -0.293229 \\
-0.116363 & 0.0123598 & -0.0403584 & 0.139297 \\
-0.00427013 & 0.0391188 & -0.134673 & -0.0741161
\end{pmatrix},
\begin{pmatrix}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.16474 & 7.50652 \\
0.112072 & 0.519668 \\
-0.0785219 & -0.534271 \\
0.00150349 & -0.0147829 \\
0.138355 & -0.0794999 \\
0.0900943 & -0.0501434 \\
0.21829 & 0.159775 \\
0.172972 & 0.229854 \\
0.226772 & 0.327425 \\
0.220503 & 0.282086 \\
0.290222 & 0.388906 \\
0.429163 & 0.801739 \\
0.524461 & 1.11808 \\
0.515055 & 1.11969 \\
0.640725 & 1.35086 \\
0.676976 & 1.67342 \\
0.952109 & 1.99013 \\
0.966745 & 2.5082 \\
0.964395 & 2.46214 \\
-0.23867 & -0.212178 \\
-0.0732414 & -0.397375 \\
-0.0595876 & -0.493511 \\
-0.0877223 & -0.354672 \\
-0.0620976 & -0.708219 \\
-0.0286178 & -0.560782 \\
-0.0776033 & -0.110003 \\
-0.245436 & -0.431218 \\
-0.0520018 & 0.0494693 \\
0.0174243 & -0.104394
\end{pmatrix}
$$

B.10 Individual confidence intervals for the multiple $M = 5$ synthetic data using procedure of Reiter (2005)

$$
\begin{pmatrix}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
2.96796 & 3.36151 & 7.23729 & 7.77576 \\
0.0995434 & 0.1246 & 0.502667 & 0.536668 \\
-0.0939313 & -0.0631125 & -0.55526 & -0.513283 \\
0.000821743 & 0.00218523 & -0.0156967 & -0.0138691 \\
-0.0718543 & 0.348565 & -0.365043 & 0.206044 \\
-0.108755 & 0.288943 & -0.322426 & 0.222139 \\
0.023707 & 0.412874 & -0.105777 & 0.425328 \\
-0.0205464 & 0.366491 & -0.0427216 & 0.502429 \\
0.0339183 & 0.419626 & 0.0605533 & 0.594296 \\
0.0259288 & 0.415078 & 0.0167955 & 0.547377 \\
0.0872027 & 0.493241 & 0.104035 & 0.673778 \\
0.241437 & 0.616889 & 0.542842 & 1.06064 \\
0.335716 & 0.713205 & 0.857966 & 1.37819 \\
0.320877 & 0.709233 & 0.858478 & 1.38089 \\
0.448814 & 0.832636 & 1.08679 & 1.61493 \\
0.488064 & 0.865889 & 1.41258 & 1.93425 \\
0.757956 & 1.14626 & 1.72799 & 2.25227 \\
0.769124 & 1.16437 & 2.23578 & 2.78061 \\
0.763264 & 1.16553 & 2.19283 & 2.73145 \\
-0.442428 & -0.0349121 & -0.507455 & 0.0830979 \\
-0.15406 & 0.00757737 & -0.50452 & -0.29023 \\
-0.0960601 & -0.0231151 & -0.539017 & -0.448006 \\
-0.117733 & -0.0577119 & -0.396741 & -0.312603 \\
-0.136677 & 0.0124814 & -0.808989 & -0.607448 \\
-0.0605667 & 0.00333109 & -0.606661 & -0.514902 \\
-0.110188 & -0.0450188 & -0.157626 & -0.0623794 \\
-0.326546 & -0.164325 & -0.546601 & -0.315834 \\
-0.106702 & 0.00269888 & -0.0241126 & 0.123051 \\
-0.000640254 & 0.0354889 & -0.1289 & -0.0798891
\end{pmatrix},
\begin{pmatrix}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.16474 & 7.50652 \\
0.112072 & 0.519668 \\
-0.0785219 & -0.534271 \\
0.00150349 & -0.0147829 \\
0.138355 & -0.0794999 \\
0.0900943 & -0.0501434 \\
0.21829 & 0.159775 \\
0.172972 & 0.229854 \\
0.226772 & 0.327425 \\
0.220503 & 0.282086 \\
0.290222 & 0.388906 \\
0.429163 & 0.801739 \\
0.524461 & 1.11808 \\
0.515055 & 1.11969 \\
0.640725 & 1.35086 \\
0.676976 & 1.67342 \\
0.952109 & 1.99013 \\
0.966745 & 2.5082 \\
0.964395 & 2.46214 \\
-0.23867 & -0.212178 \\
-0.0732414 & -0.397375 \\
-0.0595876 & -0.493511 \\
-0.0877223 & -0.354672 \\
-0.0620976 & -0.708219 \\
-0.0286178 & -0.560782 \\
-0.0776033 & -0.110003 \\
-0.245436 & -0.431218 \\
-0.0520018 & 0.0494693 \\
0.0174243 & -0.104394
\end{pmatrix}
$$