INFERENCE FOR MULTIVARIATE REGRESSION MODEL BASED ON SYNTHETIC DATA GENERATED UNDER FIXED-POSTERIOR PREDICTIVE SAMPLING: COMPARISON WITH PLUG-IN SAMPLING

Authors: Ricardo Moura
– CMA, Faculty of Sciences and Technology, Nova University of Lisbon
Portugal (rp.moura@campus.fct.unl.pt)

Martin Klein*
– Center for Statistical Research and Methodology, U.S. Census Bureau,
U.S.A. (martin.klein@census.gov)

Carlos A. Coelho
– CMA and Mathematics Department, Faculty of Sciences and Technology,
Nova University of Lisbon
Portugal (cmac@fct.unl.pt)

Bimal Sinha*
– Department of Mathematics and Statistics,
University of Maryland, Baltimore County
and Center for Disclosure Avoidance Research, U.S. Census Bureau
U.S.A. (sinha@umbc.edu)

Abstract:

- The authors derive likelihood-based exact inference methods for the multivariate regression model, for singly imputed synthetic data generated via Posterior Predictive Sampling (PPS) and for multiply imputed synthetic data generated via a newly proposed sampling method, which the authors call Fixed-Posterior Predictive Sampling (FPPS). In the single imputation case, our proposed FPPS method concurs with the usual Posterior Predictive Sampling (PPS) method, thus filling the gap in the existing literature where inferential methods are only available for multiple imputation. Simulation studies compare the results obtained with those for the exact test procedures under the Plug-in Sampling method, obtained by the same authors. Measures of privacy are discussed and compared with the measures derived for the Plug-in Sampling method. An application using U.S. 2000 Current Population Survey data is discussed.

Key-Words:

- *Finite sample inference; Maximum likelihood estimation; Pivotal quantity; Plug-in Sampling; Statistical Disclosure Control; Unbiased estimators.*

AMS Subject Classification:

- 62H10, 62H15, 62H12, 62J05, 62F10, 62E15, 62E10, 62E17, 62D99.

*__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.

1. INTRODUCTION

When releasing microdata to the public, methods of statistical disclosure control (SDC) are used to protect confidential data, that is "data which allow statistical units to be identified, either directly or indirectly, thereby disclosing individual information" [7], while enabling valid statistical inference to be drawn on the relevant population. SDC methods include data swapping, additive and multiplicative noise, top and bottom coding, and also the creation of synthetic data. In this paper, the authors provide inferential tools for the statistical analysis of a singly imputed synthetic dataset when the real dataset cannot be released. The multiple imputation case is also addressed, using a new adapted method of generating synthetic data, which the authors call Fixed-Posterior Predictive Sampling (FPPS).

The use of synthetic data for SDC started with Little [4] and Rubin [10] using multiple imputation [9]. Reiter [8] was the first to present methods for drawing inference based on partially synthetic data. Moura et al. [5] complemented this work with the development of likelihood-based exact inference methods for both single and multiple imputation, that is, inferential procedures developed based on exact distributions, and not on asymptotic results, in the case where synthetic datasets were generated via Plug-in Sampling. The procedures of Reiter [8] are general in that they can be applied to a variety of estimators and statistical models, but these procedures are only applicable in the multiple imputation case, and are based on large sample approximations.

There are two major objectives in the present research. First, to make available likelihood-based exact inference for singly imputed synthetic data via Posterior Predictive Sampling (PPS) where the usual available procedures are not applicable, therefore extending the work of Klein and Sinha [2], under the multivariate linear regression (MLR) model. Second, to propose a different approach for release of multiple synthetic datasets, FPPS, which can use a similar way of gathering information from the synthetic datasets to that used in [5], when these synthetic datasets are generated via the Plug-in Sampling method. This second objective arises from the fact that when using the classical PPS it is too hard to construct an exact joint probability density function (pdf) for the estimators, under the MLR model, since one would face the problem of deriving the distribution of a sum of variables that follow Wishart distributions with different parameter matrices. It is with this problem in mind, that we propose an adapted method that we will call the FPPS method. We show that this method offers a higher level of confidentiality than the Plug-in Sampling method, and it still allows one to draw inference for the unknown parameters using a joint pdf of the proposed estimators.

A brief description of the PPS and FPPS methods follows. Suppose that $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ are the original data which are jointly distributed according to the pdf $f_{\boldsymbol{\theta}}(\mathbf{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar, vector or matrix) parameter. A

prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is assumed and then the posterior distribution of $\boldsymbol{\theta}$ is obtained as $\pi(\boldsymbol{\theta}|Y) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}(x)}$, and used to draw a replication $\boldsymbol{\theta}_f^\bullet$ of $\boldsymbol{\theta}$, when applying the FPPS, or draw $M \geq 1$ independent replications $\boldsymbol{\theta}_1^\bullet, ..., \boldsymbol{\theta}_M^\bullet$ of $\boldsymbol{\theta}$, when applying the PPS. In the case of FPPS, we generate M replicates of $\mathbf{Y}$, namely, $\mathbf{W}_j = (\mathbf{w}_{j1}, ..., \mathbf{w}_{jn})$, $j = 1, ..., M$ drawn all independently from the same $f_{\boldsymbol{\theta}_f^\bullet}$, where $f_{\boldsymbol{\theta}_f^\bullet}$ is the joint pdf of the original $\mathbf{Y}$ with $\boldsymbol{\theta}_f^\bullet$ replacing the unknown $\boldsymbol{\theta}$. In the case of the usual PPS method for each j-th generated synthetic dataset we would use the corresponding j-th posterior draw $\boldsymbol{\theta}_j^\bullet$ and corresponding j-th joint pdf's $f_{\boldsymbol{\theta}_j^\bullet}$, for $j = 1, ..., M$. In either case, these synthetic datasets $\mathbf{W}_1, \ldots, \mathbf{W}_M$ will be the datasets available to the general public. One may observe that, for $M = 1$, the Posterior Predictive Sampling and Fixed-Posterior Predictive Sampling methods concur.

Regarding the MLR model, in our context, we consider the *sensitive* response variables y_j $(j = 1, ..., m)$ forming the vector of response variables $\mathbf{y} = (y_1, ..., y_m)'$, and a set of p non-*sensitive* explanatory variables $\mathbf{x} = (x_1, ..., x_p)'$. It is assumed that $\mathbf{y}|\mathbf{x} \sim N_m(\mathbf{B}'\mathbf{x}, \boldsymbol{\Sigma})$, with $\mathbf{B}$ and $\boldsymbol{\Sigma}$ unknown, and the original data consist of $\mathcal{Y} = \{(y_{1i}, ..., y_{mi}, x_{1i}, ..., x_{pi}), i = 1, ..., n\}$, where n will be the sample size. Let us consider $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ 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 assume $rank(\mathbf{X} : p \times n) = p < n$ and $n \geq m + p$. Therefore the following regression model is considered

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

where $\mathbb{E}_{m\times n}$ is distributed as $N_{mn}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$. Based on the original data,

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

is the Maximum Likelihood Estimator (MLE) and the Uniformly Minimum-Variance Unbiased Estimator (UMVUE) of $\mathbf{B}$, distributed as $N_{pm}(\mathbf{B}, \boldsymbol{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-1})$, independent of $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}(\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X})(\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X})'$ which is the MLE of $\boldsymbol{\Sigma}$, with $n\hat{\boldsymbol{\Sigma}} \sim W_m(\boldsymbol{\Sigma}, n - p)$. Therefore

$$\mathbf{S} = \frac{n\hat{\boldsymbol{\Sigma}}}{n - p} \tag{1.3}$$

will be the UMVUE of $\boldsymbol{\Sigma}$.

The organization of the paper is as follows. In Section 2, based on singly and multiply imputed synthetic datasets generated via Fixed-Posterior Predictive Sampling, two procedures are proposed to draw inference for the matrix of regression coefficients. Under the single imputation case, we recall that the FPPS and the PPS methods coincide. The test statistics proposed will be pivot statistics, different from the classical test statistics for $\mathbf{B}$ under the MLR model (see [1, Secs 8.3 and 8.6]) since it is shown that these classical test statistics are not pivotal in the present context. Section 3 presents some simulations in order to check the accuracy of theoretically derived results. Also in this section, the authors use a measure for the *radius* (distance between the center and the edge)

of the confidence sets for the regression coefficients adapted from [5], computed for the original data and also for the synthetic data generated via FPPS. These *radius* measures are compared with the ones obtained when synthetic datasets are generated via Plug-in Sampling. Section 4 presents data analyses under the proposed methods in the context of public use data from the U.S. Current Population Survey comparing with the same data analysis given by [5] under the Plug-in Sampling method. In Section 5, we compare the level of privacy protection obtained via our FPPS method and via Plug-in Sampling method. Some concluding remarks are added in Section 6. Proofs of the theorems, and other technical derivations are presented in Appendices A and B.

2. ANALYSIS FOR SINGLE AND MULTIPLE IMPUTATION

In this section, we present two new exact likelihood-based procedures for the analysis of synthetic data generated using Fixed-Posterior Predictive Sampling method, under the MLR model in (1.1). For the single imputation case, the two new procedures developed also offer the possibility of drawing inference for a single synthetic dataset generated via Posterior Predictive Sampling.

2.1. A First New Procedure

In this subsection, the synthetic data will consist of M synthetic versions of $\mathbf{Y}$ generated based on the FPPS method.

Consider the joint prior distribution $\pi(\mathbf{B}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-\alpha/2}$, leading to the posterior distributions for $\mathbf{\Sigma}$ and $\mathbf{B}$

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

and

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

where we assume that $n + \alpha > p + m + 1$ (see proof in Appendix B.1). Consequently, we draw $\tilde{\mathbf{\Sigma}}$ from (2.1) and $\tilde{\mathbf{B}}$ from (2.2), upon replacing $\mathbf{\Sigma}$ by $\tilde{\mathbf{\Sigma}}$ in this latter expression. We then generate the M synthetic datasets, denoted as $\mathbf{W}_j = (\mathbf{w}_{j1}, ..., \mathbf{w}_{jn})$, for $j = 1, ..., M$, where $\mathbf{w}_{ji} = (w_{1ji}, ..., w_{mji})'$, are independently distributed as

$$\mathbf{w}_{ji}|_{\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}} \sim N_m(\tilde{\mathbf{B}}'\mathbf{x_i}, \tilde{\mathbf{\Sigma}}), \quad i = 1, ..., n, j = 1, ..., M. \tag{2.3}$$

For $i = 1, ..., n$ and $j = 1, ..., M$, let $\mathbf{B}_j^{\bullet} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{W}_j'$ and $\mathbf{S}_j^{\bullet} = \frac{1}{n-p}(\mathbf{W}_j - \mathbf{B}_j^{\bullet\prime}\mathbf{X})(\mathbf{W}_j - \mathbf{B}_j^{\bullet\prime}\mathbf{X})'$ be the estimators of $\mathbf{B}$ and $\mathbf{\Sigma}$, based on the synthetic

data $(w_{1ji}, ..., w_{mji}, x_{1i}, ..., x_{pi})$, which by Lemma 1.1 in [5] are jointly sufficient. Conditional on $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}})$, for every $j = 1, ..., M$, $\mathbf{B}_j^\bullet$ is independent of $\mathbf{S}_j^\bullet$ and $\{(\mathbf{B}_1^\bullet, \mathbf{S}_1^\bullet), ..., (\mathbf{B}_M^\bullet, \mathbf{S}_M^\bullet)\}$ are jointly sufficient estimators for $\mathbf{B}$ and $\mathbf{\Sigma}$. Define then

$$\overline{\mathbf{B}}_M^\bullet = \frac{1}{M}\sum_{j=1}^{M}\mathbf{B}_j^\bullet \quad \text{and} \quad \overline{\mathbf{S}}_M^\bullet = \frac{1}{M}\sum_{j=1}^{M}\mathbf{S}_j^\bullet, \tag{2.4}$$

which are also mutually independent, given $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Sigma}}$. For $p \geq m$ and $n + \alpha > p + 2m + 2$, we derive the following main results.

1. The MLE of $\mathbf{B}$ is $\overline{\mathbf{B}}_M^\bullet$, which is unbiased for $\mathbf{B}$, with $Var(\overline{\mathbf{B}}_M^\bullet) = N_{M,n,m,p,\alpha}\mathbf{\Sigma} \otimes (\mathbf{X}\mathbf{X}')^{-1}$, where $N_{M,n,m,p,\alpha} = \frac{2M(n+\frac{\alpha}{2}-p-m-1)+n-p}{M(n+\alpha-p-2m-2)}$ (see Theorem 2.1 and Appendix B.3).

2. An unbiased estimator (UE) of $\mathbf{\Sigma}$ will be $\hat{\mathbf{S}}_M = \frac{n+\alpha-p-2m-2}{n-p}\overline{\mathbf{S}}_M^\bullet$ (see Theorem 2.1 and Appendix B.3); for $\alpha = 2m + 2$, $\overline{\mathbf{S}}_M^\bullet$ will also be an UE for $\mathbf{\Sigma}$,

3. In Theorem 2.2 (see below), we prove that

$$T_M^\bullet = \frac{|(\overline{\mathbf{B}}_M^\bullet - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}_M^\bullet - \mathbf{B})|}{|M(n-p)\overline{\mathbf{S}}_M^\bullet|}, \tag{2.5}$$

   a statistic somewhat related with the Hotelling T^2, this one built to make inference on a matrix parameter, is a pivotal quantity, and that for $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$, $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ and $F_i \sim F_{p-i+1,M(n-p)-i+1}$ $(i = 1, ..., m)$, all independent random variables,

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

   where $\mathbf{\Omega}$ has the same distribution as $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$ and where $\overset{st}{\sim}$ means 'stochastic equivalent to'.

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

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

   and proceeds by noting that

$$T_{M,\mathbf{C}}^\bullet|\mathbf{W} \overset{st}{\sim} \left\{\prod_{i=1}^{m}\frac{k-i+1}{M(n-p)-i+1}F_{k,i}\right\}\left|\frac{M+1}{M}\mathbf{I}_m + \mathbf{\Omega}\right|, \tag{2.6}$$

   with $F_{k,i} \sim F_{k-i+1,M(n-p)-i+1}$ being independent random variables and $\mathbf{\Omega}$ defined as in the previous item.

(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 $T^{\bullet}_{M,\mathbf{C}_0}$ exceeds $\delta_{M,k,m,p,n;\gamma}$ where $\delta_{M,k,m,p,n;\gamma}$ satisfies $(1-\gamma) = Pr(T^{\bullet}_{M,\mathbf{C}_0} \leq \delta_{M,k,m,p,n;\gamma})$ when H_0 is true. To perform a test for $\mathbf{B} = \mathbf{B}_0$ one has to take $\mathbf{A} = \mathbf{I}_p$.

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

$$\Delta_M(\mathbf{C}) = \{\mathbf{C} : T^{\bullet}_{M,\mathbf{C}} \leq \delta_{M,k,m,n,p;\gamma}\}, \tag{2.7}$$

where the value of $\delta_{M,k,m,n,p;\gamma}$ can be obtained by simulating the distribution in (2.6).

Results in 1-4 are derived based on Theorems 2.1 and 2.2 below.

Theorem 2.1. The joint pdf of $\overline{\mathbf{B}}^{\bullet}_M, \overline{\mathbf{S}}^{\bullet}_M$ and $\tilde{\mathbf{\Sigma}}^{-1}$, for $\overline{\mathbf{B}}^{\bullet}_M$ and $\overline{\mathbf{S}}^{\bullet}_M$ defined in (2.4), is proportional to

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

so that $\overline{\mathbf{B}}^{\bullet}_M$ and $\overline{\mathbf{S}}^{\bullet}_M$, given $\tilde{\mathbf{\Sigma}}$, are independent, with

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

and

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

Proof: See Appendix A. □

Theorem 2.2. The distribution of the statistic $T^{\bullet}_M$ defined in (2.5) can be obtained from the decomposition

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

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

Proof: See Appendix A. □

Remark 2.1. When $m = 1$ and $M = 1$, the statistic in (2.5) reduces to the statistic T^2 used in [2] whose pdf is obtained by noting that

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

Remark 2.2. We remark that the statistic $T_M^\bullet$ in (2.5) degenerates towards zero when $n \to \infty$ or $M \to \infty$, but

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

and

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

where $\xrightarrow{d}$ represents convergence in distribution. Consequently, if instead of using $T_M^\bullet$ one uses $T_{M2}^\bullet = (M(n-p))^m\, T_M^\bullet = \frac{|(\overline{\mathbf{B}}_M^\bullet - \mathbf{B})'(XX')(\overline{\mathbf{B}}_M^\bullet - \mathbf{B})|}{|\overline{\mathbf{S}}_M^\bullet|}$ one would have

$$T_{M2}^\bullet|_{\mathbf{\Omega}} \xrightarrow[n\to\infty]{d} \left\{\prod_{i=1}^m \chi^2_{p-i+1}\right\} \left|\frac{M+1}{M}\mathbf{I}_m + \mathbf{\Omega}\right|$$

and

$$T_{M2}^\bullet|_{\mathbf{\Omega}} \xrightarrow[M\to\infty]{d} \left\{\prod_{i=1}^m \chi^2_{p-i+1}\right\} |\mathbf{I}_m + \mathbf{\Omega}|\,,$$

which corresponds to the use of a simple scale change.

In Table 1, we list the simulated 0.05 cut-off points for $T_M^\bullet$, for $M = 1$ for some values of p, m and n.

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

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

Similar to what was done in [5], one could suggest the following adaptations of the classical test criterion for the multivariate regression model (see [1, Secs 8.3 and 8.6] for the classical criteria):

(a) $T^{\bullet}_{1,M} = |\overline{\mathbf{S}}^{\bullet}_M||\overline{\mathbf{S}}^{\bullet}_M + (\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})'(XX')(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})|^{-1}$ (Wilks' Lambda Criterion),

(b) $T^{\bullet}_{2,M} = tr\left[(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^{\bullet} - \mathbf{B})(\overline{\mathbf{S}}^{\bullet}_M)^{-1}\right]$ (Pillai's Trace Criterion),

(c) $T^{\bullet}_{3,M} = tr\left[(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})[(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B}) + \overline{\mathbf{S}}^{\bullet}_M]^{-1}\right]$ (Hotelling-Lawley Trace Criterion),

(d) $T^{\bullet}_{4,M} = \lambda_1$ where λ_1 denotes the largest eigenvalue of $(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})(\overline{\mathbf{S}}^{\bullet}_M)^{-1}$ (Roy's Largest Root Criterion).

However, these statistics are non-pivotal, since their distributions are function of $\mathbf{\Sigma}$ (see Appendix B.3).

2.2. A Second New Procedure

We propose yet another likelihood-based approach for exact inference about $\mathbf{B}$ where one may gather more information from the released synthetic data, following a somewhat similar procedure to the one used in [5]. Let us start by recalling that $\mathbf{W_j}$ $(j = 1, ..., M)$ are $m \times n$ matrices formed by the vectors $(\mathbf{w}_{j1}, ..., \mathbf{w}_{jn})$ as columns, generated from $\mathbf{w}_{ji}|_{\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}} \sim N_m(\tilde{\mathbf{B}}'\mathbf{x_i}, \tilde{\mathbf{\Sigma}})$ $(i = 1, ..., n)$. Note that, conditionally on $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Sigma}}$, $(\mathbf{w}_{1i}, ..., \mathbf{w}_{Mi})$ is a random sample from $N_m(\tilde{\mathbf{B}}'\mathbf{x}_i, \tilde{\mathbf{\Sigma}})$, for $i = 1, ..., n$. Consider $\overline{\mathbf{w}}_i = \frac{1}{M}\sum_{j=1}^{M}\mathbf{w}_{ji}$ and $\mathbf{S}_{\mathbf{w}i} = \sum_{j=1}^{M}(\mathbf{w}_{ji} - \overline{\mathbf{w}}_i)(\mathbf{w}_{ji} - \overline{\mathbf{w}}_i)'$ which are sufficient statistics for $\mathbf{\Sigma}$, based on the i-th covariate vector. Defining $\mathbf{S}_{\mathbf{w}} = \sum_{i=1}^{n}\mathbf{S}_{\mathbf{w}i}$, we have $(\overline{\mathbf{w}}_1, ..., \overline{\mathbf{w}}_n, \mathbf{S}_{\mathbf{w}})$ as the joint sufficient statistics for $(\mathbf{B}, \mathbf{\Sigma})$. Conditionally on $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Sigma}}$, we have $\overline{\mathbf{w}}_i \sim N_m(\tilde{\mathbf{B}}'\mathbf{x}_i, \frac{1}{M}\tilde{\mathbf{\Sigma}})$ and $\mathbf{S}_{\mathbf{w}i} \sim W_m(\tilde{\mathbf{\Sigma}}, M-1)$.

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

$$\overline{\mathbf{B}}^{\bullet}_M = (\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\overline{\mathbf{W}}'_M, \tag{2.8}$$

and for $\mathbf{\Sigma}$ its estimator

$$\mathbf{S}^{\bullet}_{comb} = \frac{\mathbf{S}_{\mathbf{w}} + M \times \mathbf{S}^{\bullet}_{mean}}{Mn - p}, \tag{2.9}$$

where we define $\mathbf{S}^{\bullet}_{mean} = (\overline{\mathbf{W}}_M - \overline{\mathbf{B}}^{\bullet'}_M\mathbf{X})(\overline{\mathbf{W}}_M - \overline{\mathbf{B}}^{\bullet'}_M\mathbf{X})'$.

In fact, if the M synthetic datasets are treated as a single synthetic dataset of size nM, the estimators obtained for $\mathbf{B}$ and $\mathbf{\Sigma}$ will be exactly the same as the ones obtained in (2.8) and (2.9). The proof of this fact may be analyzed in Appendix C.

Analogous to what was done in the previous subsection, one can derive the following inferential results, for $p \geq m$ and $n + \alpha > p + 2m + 2$.

1. An UE of $\boldsymbol{\Sigma}$ will be $\hat{\mathbf{S}}_M = \frac{n+\alpha-p-2m-2}{n-p}\mathbf{S}^{\bullet}_{comb}$ (see Corollary 2.3 Appendix B.4), and for $\alpha = 2m+2$, $\mathbf{S}^{\bullet}_{comb}$ will also be an UE for $\boldsymbol{\Sigma}$.

2. In Corollary 2.3 (see below), we prove that

$$T^{\bullet}_{comb} = \frac{|(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})'(XX')(\overline{\mathbf{B}}^{\bullet}_M - \mathbf{B})|}{|(Mn-p)\mathbf{S}^{\bullet}_{comb}|} \tag{2.10}$$

is a pivotal quantity, and that for $\mathbf{A}_1 \sim W_m(\mathbf{I}_m, n+\alpha-p-m-1)$, $\mathbf{A}_2 \sim W_m(\mathbf{I}_m, n-p)$ and $F_i \sim F_{p-i+1,Mn-p-i+1}$ $(i = 1, ..., m)$, all independent random variables,

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

where $\boldsymbol{\Omega}$ has the same distribution as $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$.

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

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

and proceed by noting that

$$T^{\bullet}_{comb,\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{i=1}^{m} \frac{k-i+1}{Mn-p-i+1} F_{k,i} \right\} \left| \frac{M+1}{M}\mathbf{I}_m + \boldsymbol{\Omega} \right|, \tag{2.11}$$

with $F_{k,i} \sim F_{k-i+1,Mn-p-i+1}$ being independent random variables and $\boldsymbol{\Omega}$ defined as in the previous item.

(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 $T^{\bullet}_{comb,\mathbf{C}_0}$ exceeds $\delta_{comb,k,m,p,n;\gamma}$ where $\delta_{comb,k,m,p,n;\gamma}$ satisfies $(1-\gamma) = Pr(T^{\bullet}_{comb,\mathbf{C}_0} \leq \delta_{comb,k,m,p,n;\gamma})$ when H_0 is true. To perform a test for $\mathbf{B} = \mathbf{B}_0$ one has to take $\mathbf{A} = \mathbf{I}_p$.

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

$$\Delta_{comb}(\mathbf{C}) = \{\mathbf{C} : T^{\bullet}_{comb,\mathbf{C}} \leq \delta_{comb,k,m,n,p;\gamma}\}, \tag{2.12}$$

where the value of $\delta_{comb,k,m,n,p;\gamma}$ can be obtained by simulating the distribution in (2.11).

Results in 1-3 are derived based on the following Corollaries 2.3 and 2.4, of Theorems 2.1 and 2.2, respectively.

Corollary 2.3. The joint pdf of $\overline{\mathbf{B}}^{\bullet}_M, \mathbf{S}^{\bullet}_{comb}$ and $\tilde{\mathbf{\Sigma}}^{-1}$, for $\overline{\mathbf{B}}^{\bullet}_M$ and $\mathbf{S}^{\bullet}_{comb}$ defined in (2.8) and (2.9), is proportional to

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

so that $\overline{\mathbf{B}}^{\bullet}_M$ and $\mathbf{S}^{\bullet}_{comb}$, given $\tilde{\mathbf{\Sigma}}$, are independent, with

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

and

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

Proof: See Appendix A. □

Corollary 2.4. The distribution of the statistic $T^{\bullet}_{comb}$ defined in (2.10) can be obtained from the decomposition

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

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

Proof: See Appendix A. □

Remark 2.3. Similar to what happens with the statistic $T^{\bullet}_M$ in (2.5), the statistic $T^{\bullet}_{comb}$ in (2.10) also degenerates towards zero when $n \to \infty$ or $M \to \infty$, and similarly to what happens with $T^{\bullet}_M$,

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

and

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

Using the simple scale change $T^{\bullet}_{comb2} = (Mn-p)^m\, T^{\bullet}_{comb} = \frac{|(\overline{\mathbf{B}}^{\bullet}_M-\mathbf{B})'(XX')(\overline{\mathbf{B}}^{\bullet}_M-\mathbf{B})|}{|\overline{\mathbf{S}}^{\bullet}_{comb}|}$ one would have

$$T^{\bullet}_{comb2}|_{\mathbf{\Omega}} \xrightarrow[n\to\infty]{d} \left\{\prod_{i=1}^{m}\chi^2_{p-i+1}\right\}\left|\frac{M+1}{M}\mathbf{I}_m+\mathbf{\Omega}\right|$$

and

$$T^{\bullet}_{comb2}|\boldsymbol{\Omega} \xrightarrow[M\to\infty]{d} \left\{\prod_{i=1}^{m} \chi^2_{p-i+1}\right\} |\mathbf{I}_m + \boldsymbol{\Omega}|\,,$$

similar to what happens with $T^{\bullet}_{M}$.

3. SIMULATION STUDIES

In order to compare the PPS and the FPPS methods with the Plug-in Sampling method we present the results of some simulations analogous to the ones presented in [5]. The objectives of these simulations are: (i) to show that the inference methods developed in Section 2 perform as predicted, and (ii) to compare the measures (*radius*) obtained from our methods with the ones from the Plug-in method. 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 (1.1), for $m = 2$ and $p = 3$, with matrix of regressor coefficients

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

and covariance matrix

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

We set $\alpha = 6$ in order to have both $\bar{\mathbf{S}}^{\bullet}_{M}$ and $\mathbf{S}^{\bullet}_{comb}$ as the unbiased estimators of $\boldsymbol{\Sigma}$. The regressor variables $x_{1i}, x_{2i}, x_{3i}, i = 1, ..., n$ are generated as i.i.d. $N(1, 1)$ and held fixed for the entire simulation. Based on Monte Carlo simulation with 10^5 iterations, we compute an estimate of the coverage probability of the confidence regions for $\mathbf{B}$ and $\mathbf{C} = \mathbf{AB}$ given by (2.7) and (2.12), defined as percentage of observed values of the statistics smaller than the respective theoretical cut-off points, with $\mathbf{A} = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$, using the methodologies described in Section 2. For $M = 1$, $M = 2$ and $M = 5$, 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 first new procedure in Subsection 2.1, and under the columns $\mathbf{B}(2)$ and $\mathbf{AB}(2)$ for the second new procedure in Subsection 2.2. For $M = 1$, a single column is shown for each confidence region since the two new procedures are the same.

Table 2: Average coverage for $\mathbf{B}$ and $\mathbf{AB}$

n	$M=1$		$M=2$				$M=5$			
	$\mathbf{B}$	$\mathbf{AB}$	1st Approach		2nd Approach		1st Approach		2nd Approach	
			$\mathbf{B}(1)$	$\mathbf{AB}(1)$	$\mathbf{B}(2)$	$\mathbf{AB}(2)$	$\mathbf{B}(1)$	$\mathbf{AB}(1)$	$\mathbf{B}(2)$	$\mathbf{AB}(2)$
10	0.949	0.951	0.949	0.949	0.951	0.949	0.951	0.950	0.949	0.951
50	0.949	0.950	0.951	0.951	0.950	0.951	0.951	0.950	0.949	0.948
100	0.949	0.949	0.951	0.950	0.949	0.951	0.949	0.951	0.951	0.950
200	0.951	0.951	0.949	0.951	0.951	0.949	0.950	0.951	0.950	0.951

The results reported in Table 2 for samples of size $n = 10, 50, 100, 200$, show that, based on singly 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.

In order to measure the *radius* (distance between the center and the edge) of the confidence sets, we use the same measure proposed in [5], which is

$$\Upsilon_M = d^*_{M,m,n,p,\gamma} \times |\tilde{\mathbf{S}}^{\bullet}_M|,$$

where $d^*_{M,m,n,p,\gamma}$ is the cut-off point in (2.7) or (2.12). Here we take $M = 0$ for the original data, with $\tilde{\mathbf{S}}^{\bullet}_0 = (n-p)\mathbf{S}$, $M = 1$ for the singly imputed synthetic data and $M = 2, 5$ for the multiply imputed synthetic data, with $\tilde{\mathbf{S}}^{\bullet}_M = M(n-p)\overline{\mathbf{S}}^{\bullet}_M$ for the first new procedure, and $\tilde{\mathbf{S}}^{\bullet}_M = (Mn-p)\mathbf{S}^{\bullet}_{comb}$ 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,

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

for the procedure in Subsection 2.1 and

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

for the procedure in Subsection 2.2, where $\kappa_{n,\alpha,p,m} = n+\alpha-p-m-1$, assuming $n+\alpha > p+2m+2$. For more details about these expected values we refer to Appendix B.5.

We present in Table 3 the average of the simulated values of the *radius* Υ_M and its expected value $E(\Upsilon_M)$ for the confidence sets $\Delta_M(\mathbf{B})$ (first procedure) and $\Delta_{comb}(\mathbf{B})$ (second procedure), and in Table 4 the same values for the confidence sets $\Delta_M(\mathbf{C})$ (first procedure) and $\Delta_{comb}(\mathbf{C})$ (second procedure), for $M = 0, 1, 2, 5$ and $n = 10, 50, 200$. These values may be compared with the values obtained in [5] for the Plug-in Sampling.

Observing Tables 3 and 4 and comparing the entries in these tables with the results in [5] for Plug-in Sampling, we may see that when synthetic data are generated under FPPS, larger *radius* are obtained. In the singly imputed case, one can observe that the PPS synthetic datasets will lead to a *radius* that is approximately two and half times that of the *radius* under Plug-in Sampling. As the number M of released synthetic datasets increases, Υ_M slowly decreases, increasing however the difference of the *radius* between the FPPS and the Plug-in methods. Eventually, one may need very large values of M, in order to have values of Υ_M close to the value of Υ_0. As in [5] we also observe that the values of Υ_M ($M > 1$), for both new FPPS procedures become identical for larger sample sizes.

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

n	Orig	$M=1$		$M=2$			
		avg	exp	1st Procedure		2nd Procedure	
				avg	exp	avg	exp
10	36.97	507.25	512.19	251.55	252.55	237.64	238.68
50	19.11	176.36	176.53	121.23	121.52	121.23	121.48
200	17.52	154.93	156.06	105.81	106.61	105.90	106.72

n	$M=5$			
	1st Procedure		2nd Procedure	
	avg	exp	avg	exp
10	175.34	176.18	163.82	168.92
50	92.25	92.80	92.28	92.84
200	81.89	82.39	81.91	82.40

Table 4: Average values of Υ_M and the values of $E(\Upsilon_M)$ for the confidence set for $\mathbf{C} = \mathbf{AB}$.

n	Orig	$M=1$		$M=2$			
		avg	exp	1st Procedure		2nd Procedure	
				avg	exp	avg	exp
10	13.43	172.64	172.32	92.23	92.44	86.24	86.61
50	7.33	68.93	68.99	47.75	47.86	47.45	47.55
200	7.10	60.65	61.09	41.74	42.05	41.74	42.05

n	$M=5$			
	1st Procedure		2nd Procedure	
	avg	exp	avg	exp
10	63.07	63.38	61.34	61.74
50	35.32	35.52	35.08	35.27
200	32.47	32.51	32.54	32.53

4. AN APPLICATION USING CURRENT POPULATION SURVEY DATA

In this section, we provide an application based on the same real data used in [5] to compare the original data inference with the one obtained via PPS, for the single imputation case, and via FPPS, for the multiple imputation case. The data are from the U.S. 2000 Current Population Survey (CPS) March supplement, available online at http://www.census.gov.cps/. Further details on the data may be found in [5].

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

$$\mathbf{x} = \Big(1, N, L, A, I(E=34), ..., I(E=37), I(E=39), ..., I(E=46),$$
$$I(M=3), ..., I(M=7), I(R=2), I(R=4), I(S=2)\Big)',$$

where N, L, A, are respectively, the number of people in household, the number of people in the household who are less than 18 years old and the age for the head of household, E, M, R and S, are respectively, the education level for the head of

the household (coded to take values 31, 34-37, 39-46), the marital status for the head of the household (coded to take values 1,3-7), the race of the head of the household (coded to take values 1,2,4) and the sex of the head of the household (coded to take values 1,2). I(E = 34) is the indicator variable for $E = 34$, I(E = 35) is the indicator variable for $E = 35$, and so on, and where the indicator variable for the first code present in the sample for each variable is taken out in order to make the model matrix full rank. The vector $\mathbf{y}$ of response variables will be formed by the same three numerical variables used in [5], namely, *total household income*, *household alimony payment* and *household property tax*. After deleting all entries where at least one of these variables are reported as 0, we were left with a sample size of 141, and as such the model matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ has thus $p = 24$ rows, $n = 141$ columns, with rank equal to 24. Throughout this section we will assume $\alpha = 8$ in order to have $\mathbf{S}^{\bullet}_{M}$ and $\mathbf{S}^{\bullet}_{comb}$ as unbiased estimators of $\mathbf{\Sigma}$. Via PPS method we generate a single synthetic dataset and show in expression (4.1) the realizations of the unbiased estimator $\mathbf{S}^{\bullet}$ for $\mathbf{\Sigma}$ and of the estimator $\mathbf{S}$ for the original data, respectively denoted by $\widetilde{\mathbf{S}}^{\bullet}_{1}$ and $\widetilde{\mathbf{S}}$

(4.1)

$$\widetilde{\mathbf{S}}^{\bullet}_{1} = \begin{pmatrix} 1.58572 & -0.20443 & 0.27981 \\ -0.20443 & 1.61395 & 0.16089 \\ 0.27981 & 0.16089 & 0.34648 \end{pmatrix}, \quad \widetilde{\mathbf{S}} = \begin{pmatrix} 1.1980 & -0.0375 & 0.2970 \\ -0.0375 & 1.0699 & 0.1175 \\ 0.2970 & 0.1175 & 0.4045 \end{pmatrix}.$$

In Table 5 we show the realizations of the unbiased estimator $\mathbf{B}^{\bullet}_{1}$ of $\mathbf{B}$ and of the estimator $\hat{\mathbf{B}}$ of the original data, respectively denoted by $\widetilde{\mathbf{B}}^{\bullet}_{1}$ and $\widetilde{\hat{\mathbf{B}}}$.

Table 5: Estimates of the regressor coefficients from the FPPS synthetic data ($\widetilde{\mathbf{B}}^{\bullet}$), Plug-in synthetic data ($\widetilde{\mathbf{B}}^{*}$) and from the original data.

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

At a first glance the estimates originated via Plug-in Sampling (see [5]) seem

to be more in agreement with the original data estimates than the ones drawn from PPS. Nevertheless, this is only one draw and it could be a question of chance to originate 'better' or 'worse' data. Therefore, one must conduct inferences on the regression coefficients based on multiple draws.

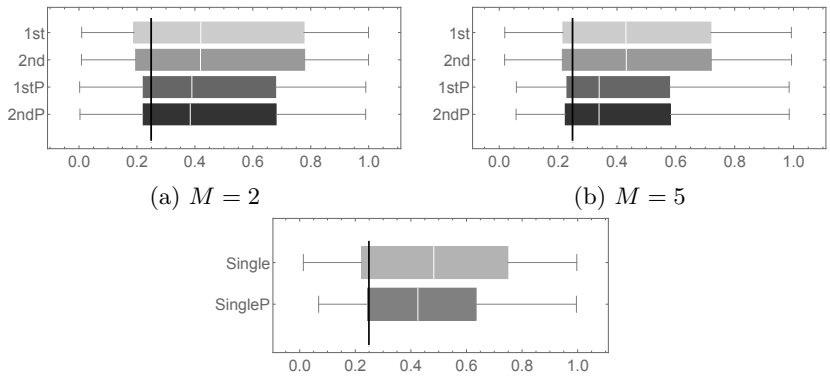
Inferences on regression coefficients are obtained by applying the methodologies in Subsections 2.1 and 2.2, to analyze the singly imputed synthetic dataset and multiply imputed synthetic datasets, considering $M = 1$, $M = 2$ and $M = 5$, using the statistics $T_M^{\bullet}$ and $T_{comb}^{\bullet}$ and their empirical distributions based on simulations with 10^4 iterations, to test the fit of the model and the significance of some regressors for $\gamma = 0.05$. Regarding the test of fit of the model one will find, for all values of M, results equivalent to the ones obtained for the case when synthetic data are generated via Plug-in Sampling, i.e., concluding that the explanatory variables in $\mathbf{x}$ have a significant role in determining the values of the response variables in $\mathbf{y}$ since the obtained p-values, computed as the fraction of values of the empirical distribution of the corresponding statistic that are larger than the computed value of the statistic, were all approximately zero. The cut-off points obtained from the empirical distributions of $T_M^{\bullet}$ and $T_{comb}^{\bullet}$ (respectively associated with the first and second procedures in Subsections 2.1 and 2.2) are approximately equal to 0.50357, for $M = 1$ (where first and second procedures coincide), to 0.03460 and 0.02569, for $M = 2$, and to 0.00149 and 0.00094, for $M = 5$.

In order to test the significance of some regressors, we propose to study two different cases, using in each case the same sets of regressors as in [5]. Therefore, we will test the significance of regressor variables R and S, for the first case, and regressor variables A and E, for the second case. As such, in the first case, we will consider a 3×24 matrix

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

and we will be interested in testing the hypothesis $H_0 : \mathbf{AB} = \mathbf{C_0}$, where $\mathbf{C}_0$ is a 3×3 matrix consisting of only zeros. We now generate 100 draws of $M = 1$, $M = 2$ and $M = 5$ synthetic datasets and gather the different p-values obtained when using the statistics in (2.5) and (2.10). In Figure 1, one may analyze the box-plots of the p-values obtained for each procedure together with the ones obtained in [5] for the same sets of variables, where under Single, 1st and 2nd, one has the box-plots associated with the new procedures developed in this paper and under SingleP, 1stP and 2ndP, the box-plots associated to the Plug-in Sampling method. The existing line in the box-plots marks the original data p-value 0.249, obtained using the $T_{O,\mathbf{C}}$ statistic in (3) of [5]. It is important to note that in the case of single imputation ($M = 1$) the FPPS method reduces to the usual PPS method.

In general, from Figure 1, we may note in both new procedures a larger spread of the p-values when compared with the p-values gathered from Plug-in Sampling, presenting a distribution of p-values with larger values than the original, nonetheless with the majority of these p-values leading to similar conclusions



(a) $M = 2$ (b) $M = 5$

(c) $M = 1$

Figure 1: Box-plots of p-values obtained, when testing the joint significance of I(R=2), I(R=4) and I(S=2), from 100 draws of synthetic datasets using procedures in Section 2 and using Plug-in Sampling method from [5], for $M = 1$, $M = 2$ and $M = 5$.

as those obtained from the original data for $\gamma = 0.05$, that is, to not reject the null hypothesis that variables R and S do not have significant influence on the response variables.

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

For the second case, we are interested in testing the hypothesis H_0 : $\mathbf{AB} = \mathbf{C_0}$, where $\mathbf{C_0}$ is a 13×13 matrix consisting of only zeros, with

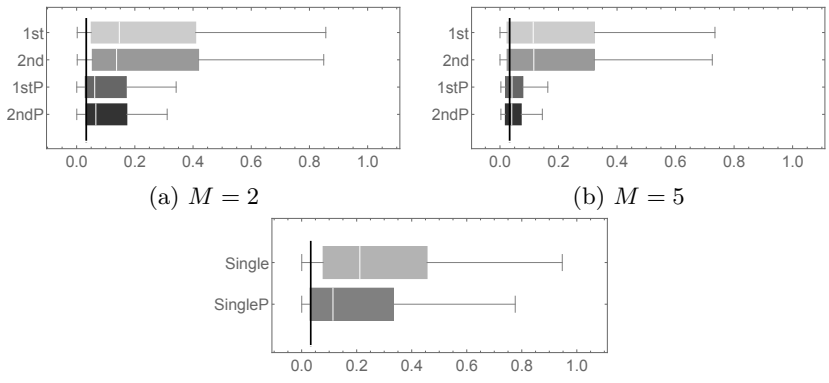
$$\mathbf{A} = \left(\mathbf{0}_{13\times 3} \middle| \mathbf{I}_{13} \middle| \mathbf{0}_{13\times 8} \right),$$

corresponding to the test of joint significance of variables A and E. The p-value obtained for the original data, based on (3) in [5], was 0.033, thus rejecting their non-significance for $\gamma = 0.05$. In Figure 2, we can compare the box-plots obtained for the FPPS and Plug-in Sampling methods obtained by generating 100 draws of synthetic datasets, for $M = 1$, $M = 2$ and $M = 5$. The vertical line represents again the original data's p-value.

From Figure 2, we note that the spread of p-values is again larger for our new procedures based on FPPS than the ones from the Plug-in method, majorly leading to a different conclusion from the inference obtained from the original data.

For the single imputation case, even if the spread of the p-values gathered from the PPS is larger than the ones from the Plug-in Sampling, the distributions of p-values are not that different for the two methods.

For the two cases studied, the two new FPPS multiple imputation procedures presented have very similar p-values. As M increases the spread of the



(a) $M = 2$ (b) $M = 5$

(c) $M = 1$

Figure 2: Box-plots of p-values obtained, when testing the joint significance of A and E, from 100 draws of synthetic datasets using procedures in Section 2 and using Plug-in Sampling method from [5], for $M = 1$, $M = 2$ and $M = 5$.

p-values from FPPS becomes smaller and closer to the original data's p-value but at a smaller rate than the p-values from the Plug-in Sampling.

Nevertheless, this larger spread of the p-values from FPPS will be compensated by an increase of the level of confidentiality, as it can be seen in the next section.

Next, we present the power for the tests

$$\begin{array}{l} H_0 : \mathbf{B} = \mathbf{B_0} (\neq \mathbf{0}) \; vs \;\; H_1 : \mathbf{B} = \mathbf{B_1} \quad \text{and} \\ H_0 : \mathbf{AB} = \mathbf{C_0} (\neq \mathbf{0}) \; vs \;\; H_1 : \mathbf{AB} = \mathbf{C_1} \end{array} \tag{4.2}$$

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

$$\mathbf{A} = \left(\mathbf{0}_{12\times 4} \middle| \mathbf{I}_{12} \middle| \mathbf{0}_{12\times 8} \right),$$

a 12×12 matrix defined appropriately in order to isolate the indicator variables associated with the variable E, and $\mathbf{C}_1 = \mathbf{AB}_1$ where $\mathbf{B}_1$ takes different values, found in Table 6, with $\mathbf{D}$ a $p \times m$ matrix of 1's.

The power for the synthetic data obtained via FPPS was then simulated as well as the power for the case when these synthetic datasets are treated as if they were the original data. We also simulated the power from the original data and refer to [5] for the power values for the synthetic data generated via Plug-in Sampling.

From the power values in Table 6 we may see that tests based on the synthetic data via FPPS show lower values for its power than the ones based in Plug-in generation, as expected, since we are using a method which is supposed to give more confidentiality by generating more perturbed datasets. We may see that these values increase along with the value of M, but with a smaller rate than that for Plug-in Sampling, leading to the conclusion that one will need larger values of M to obtain a closer power value to the one registered when

Table 6: Power for the tests to the hypothesis (4.2), with $\mathbf{B}(1)$, $\mathbf{C}(1)$ and $\mathbf{B}(2)$ and $\mathbf{C}(2)$ denoting the first and second procedures proposed by the authors in Subsections 2.1 and 2.2 for FPPS and in [5] for Plug-in method.

Power for $\mathbf{B}_1 =$	orig data $\mathbf{B}$	Methods	M=1 $\mathbf{B}$	M=2 $\mathbf{B}(1)$	M=2 $\mathbf{B}(2)$	M=5 $\mathbf{B}(1)$	M=5 $\mathbf{B}(2)$	synt as orig $\mathbf{B}$
$\mathbf{B}_0 + 0.005\mathbf{D}$	0.537	FPPS	0.215	0.252	0.253	0.275	0.279	1.000
		Plug-in	0.279	0.382	0.385	0.471	0.472	1.000
$\mathbf{B}_0 * 0.95$	0.945	FPPS	0.535	0.634	0.637	0.700	0.700	1.000
		Plug-in	0.679	0.840	0.841	0.906	0.909	1.000
Power for $\mathbf{C}_1 =$	orig data $\mathbf{C}$	Methods	M=1 $\mathbf{C}$	M=2 $\mathbf{C}(1)$	M=2 $\mathbf{C}(2)$	M=5 $\mathbf{C}(1)$	M=5 $\mathbf{C}(2)$	synt as orig $\mathbf{C}$
$\mathbf{A}(\mathbf{B}_0 + 3\mathbf{D})$	0.465	FPPS	0.185	0.202	0.207	0.245	0.246	0.996
		Plug-in	0.284	0.334	0.343	0.416	0.418	0.975
$\mathbf{A}(\mathbf{B}_0 * 0.5)$	0.393	FPPS	0.136	0.160	0.161	0.179	0.181	0.996
		Plug-in	0.197	0.271	0.279	0.326	0.327	0.959

testing using the original data. If synthetic data is treated as original, we obtain a larger power than the one obtained for the original data, which is obviously misleading, since the estimated coverage probability will be in fact much smaller than the desired 0.95.

5. PRIVACY PROTECTION OF SINGLY VERSUS MULTIPLY IMPUTED SYNTHETIC DATA

In order to evaluate the level of protection and at the same time compare it with the level obtained from synthetic data generated via Plug-in Sampling, we perform, in this section, a similar evaluation as in [5] using CPS data. Let us consider $\mathbf{W}_l = (\mathbf{w}_{1l}, ..., \mathbf{w}_{nl})$, $l = 1, ..., M$, M synthetic datasets generated via FPPS, where $\mathbf{w}_{il} = (w_{1il}, ..., w_{mil})', i = 1, ..., n$. The estimate of the original values $\mathbf{y}_i = (y_{1i}, ..., y_{mi})'$ will be $\hat{\mathbf{y}}_i = \frac{1}{M}\sum_{l=1}^{M} \mathbf{w}_{il}$. Let us recall the three criteria used in [5] as measures of the level of privacy protection:

$$\Gamma_{1,\epsilon} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n} Pr\left[\left|\frac{\hat{y_{ji}} - y_{ji}}{y_{ji}}\right| < \epsilon \,\middle|\, \mathbf{Y}\right]; \tag{5.1}$$

$$\Gamma_{2,\epsilon} = \frac{1}{n}\sum_{i=1}^{n} Pr\left[\sqrt{\frac{1}{m}\sum_{j=1}^{m}\frac{(\hat{y_{ji}} - y_{ji})^2}{y_{ji}^2}} < \epsilon \,\middle|\, \mathbf{Y}\right]; \tag{5.2}$$

$$\Gamma_{3,\epsilon} = Pr\left[\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\left|\frac{\hat{y_{ji}} - y_{ji}}{y_{ji}}\right| < \epsilon \,\middle|\, \mathbf{Y}\right]. \tag{5.3}$$

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

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

and, from $\Gamma_{3,\epsilon}$,

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

We use a Monte Carlo simulation with 10^4 iterations to estimate all three measures in (5.1)–(5.3) based on the $n = 141$ households in the CPS data. In Table 7, we show the values of $\Gamma_{1,0.01}$, $\Gamma_{2,0.01}$ and the minimum, 1st quartile (Q_1), median, 3rd quartile (Q_3) and maximum of $D_{1,\epsilon}$, displaying also the values gathered when using Plug-in Sampling. In Table 8, we show the values of $\Gamma_{3,0.1}$ and the minimum, Q_1, median, Q_3 and maximum of D_3 also displaying the values gathered when using Plug-in Sampling.

Table 7: Values of $\Gamma_{1,0.01}$, $\Gamma_{2,0.01}$ and a summary of the distribution of $D_{1,0.01}$.

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

Table 8: Values of $\Gamma_{3,0.1}$ and a summary of the distribution of D_3.

M	Method	$\Gamma_{3,0.1}$	Min	Q_1	Median	Q_3	Max
$M = 1$	FPPS	0.0000	0.1091	0.1248	0.1287	0.1325	0.1544
	Plug-in	0.0000	0.1050	0.1202	0.1233	0.1264	0.1379
$M = 2$	FPPS	0.0021	0.0960	0.1088	0.1116	0.1145	0.1324
	Plug-in	0.0694	0.0948	0.1026	0.1051	0.1072	0.1159
$M = 5$	FPPS	0.5008	0.0896	0.0980	0.1000	0.1020	0.1131
	Plug-in	1.0000	0.0846	0.0905	0.0920	0.0936	0.0992

Looking at Tables 7 and 8, we observe that the values of the privacy measures in (5.1)–(5.3) increase for increasing values of M for both procedures developed in Subsections 2.1 and 2.2, showing that the disclosure risk increases with the increase in the number of released synthetic datasets. Compared with the measures obtained under Plug-in Sampling, we may observe a smaller disclosure risk in all cases, leading to the conclusion that the proposed FPPS procedures have an overall higher level of confidentiality. Regarding measures $\Gamma_{2,\epsilon}$ and $\Gamma_{3,\epsilon}$ this increase reaches in some cases an increase of 50% or more in confidentiality. In the single imputation case, under the PPS we also register an increase of confidentiality when comparing the same measure under Plug-in Sampling, nevertheless this increase is relatively small.

6. CONCLUDING REMARKS

In this paper the authors derive likelihood-based exact inference for single and multiple imputation cases where synthetic datasets are generated via Fixed-Posterior Predictive Sampling (FPPS). If only one synthetic dataset is released, then FPPS is equivalent to the usual Posterior Predictive Sampling (PPS) method. Thus the proposed methodology can be used to analyze a singly imputed synthetic data set generated via PPS under the multivariate linear regression (MLR) model. Therefore this work fills a gap in the literature because the state of the art methods apply only to multiply imputed synthetic data. Under the MLR model, the authors derived two different exact inference procedures for the matrix of regression coefficients, when multiply imputed synthetic datasets are released. It is shown that the methodologies proposed lead to confidence sets matching the expected level of confidence, for all sample sizes. Furthermore, while the second proposed procedure displays a better precision for smaller samples and/or smaller values of M by yielding smaller confidence sets, the two procedures concur for larger sample sizes and larger values of M, as it is corroborated in theory by remarks 2.2 and 2.3. When compared with inference procedures for Plug-in Sampling, the procedures proposed based on FPPS lead to synthetic datasets that give respondents a higher level of confidentiality, that is, a reduced disclosure risk, nevertheless at the expense of accuracy, since the confidence sets are larger, as illustrated in the application with the CPS data. Once likelihood-based exact inferential methods are now made available both for FPPS/PPS and Plug-in Sampling, it is therefore the responsibility of those in charge of releasing the data to decide which method to use in order to better respect the demands and objectives of their institution.

ACKNOWLEDGMENTS

Ricardo Moura's research is supported by a Fulbright Research Grant, and he sincerely thanks the faculty of Mathematics and Statistics at UMBC for their support and encouragement. Ricardo Moura and Carlos A. Coelho also thank FCT (Portuguese Foundation for Science and Technology) project UID/MAT/00297/2013 awarded through CMA/UNL. Martin Klein and Bimal Sinha thank Laura McKenna, Eric Slud, William Winkler, and Tommy Wright at the U.S. Census Bureau for their support. The authors would also like to thank the referees for the helpful comments and suggestions leading to the improvement of the paper.

A. Proof of Theorems 2.1 and 2.2 and Corollaries 2.3 and 2.4

Proof of Theorem 2.1: Given $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}})$, from (2.3) we have that, for every $j = 1, ..., M$,

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

and

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

Therefore, we have for $\overline{\mathbf{B}}_M^{\bullet}$ and $\overline{\mathbf{S}}_M^{\bullet}$ in (2.4),

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

and

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

Since $\overline{\mathbf{B}}_M^{\bullet}$ and $\overline{\mathbf{S}}_M^{\bullet}$ are independent, the conditional joint pdf of $(\overline{\mathbf{B}}_M^{\bullet}, \overline{\mathbf{S}}_M^{\bullet})$, given $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Sigma}}$, is

$$f(\overline{\mathbf{B}}_M^{\bullet}, \overline{\mathbf{S}}_M^{\bullet}|\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}) \propto e^{-\frac{1}{2}tr\{M\tilde{\mathbf{\Sigma}}^{-1}\left[(\overline{\mathbf{B}}_M^{\bullet}-\tilde{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\overline{\mathbf{B}}_M^{\bullet}-\tilde{\mathbf{B}})+M(n-p)\overline{\mathbf{S}}_M^{\bullet}\right]\}} \times \frac{|\overline{\mathbf{S}}_M^{\bullet}|^{\frac{M(n-p)-m-1}{2}}}{|\tilde{\mathbf{\Sigma}}|^{\frac{M(n-p)+p}{2}}}, \tag{A.1}$$

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

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

On the other hand, given the independence of $\hat{\mathbf{B}}$ and $\mathbf{S}$, defined in (1.2) and (1.3), the joint pdf of $(\hat{\mathbf{B}}, \mathbf{S})$ is given by

$$f(\hat{\mathbf{B}}, \mathbf{S}) \propto e^{-\frac{1}{2}tr\{\mathbf{\Sigma}^{-1}\left[(\hat{\mathbf{B}}-\mathbf{B})'\mathbf{X}\mathbf{X}'(\hat{\mathbf{B}}-\mathbf{B})+(n-p)\mathbf{S}\right]\}} \frac{|\mathbf{S}|^{\frac{n-p-m-1}{2}}}{|\mathbf{\Sigma}|^{\frac{n}{2}}}. \tag{A.3}$$

Thus, by multiplying the three pdf's in (A.1), (A.2) and (A.3), we obtain the joint pdf of $(\overline{\mathbf{B}}_M^{\bullet}, \overline{\mathbf{S}}_M^{\bullet}, \tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}^{-1}, \hat{\mathbf{B}}, \mathbf{S})$.

Since

$$tr\{M(\overline{\mathbf{B}}_M^{\bullet} - \tilde{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\overline{\mathbf{B}}_M^{\bullet} - \tilde{\mathbf{B}})\} = tr\{M(\tilde{\mathbf{B}} - \overline{\mathbf{B}}_M^{\bullet})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}} - \overline{\mathbf{B}}_M^{\bullet})\},$$

and since from Appendix B.2 we may write

$$M(\tilde{\mathbf{B}} - \overline{\mathbf{B}}_M^{\bullet})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}} - \overline{\mathbf{B}}_M^{\bullet}) + (\tilde{\mathbf{B}} - \hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}} - \hat{\mathbf{B}}) =$$

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

by integrating out $\tilde{\mathbf{B}}$, we obtain the joint pdf of $(\overline{\mathbf{B}}^{\bullet}_M, \overline{\mathbf{S}}^{\bullet}_M, \tilde{\mathbf{\Sigma}}^{-1}, \hat{\mathbf{B}}, \mathbf{S})$ proportional to

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

Since

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

and since from the identities in 1.-3. in Appendix B1 in [5] we may write

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

integrating out $\hat{\mathbf{B}}$ we will have the joint pdf of $(\overline{\mathbf{B}}^{\bullet}_M, \overline{\mathbf{S}}^{\bullet}_M, \tilde{\mathbf{\Sigma}}^{-1}, \mathbf{S})$ proportional to

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

Consequently, if we integrate out $\mathbf{S}$ we will end up with the joint pdf of $(\overline{\mathbf{B}}^{\bullet}_M, \overline{\mathbf{S}}^{\bullet}_M, \tilde{\mathbf{\Sigma}}^{-1})$ proportional to

(A.4)

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

as we wanted to prove. It is easy to see that in (A.4), $\overline{\mathbf{S}}^{\bullet}_M$ and $\overline{\mathbf{B}}^{\bullet}_M$, given $\tilde{\mathbf{\Sigma}}^{-1}$, are separable, with the distributions in the body of the Theorem. □

Proof of Theorem 2.2: From the distributions of $\overline{\mathbf{S}}_M^{\bullet}$ and $\overline{\mathbf{B}}_M^{\bullet}$ in Theorem 2.1, and by Theorem 2.4.1 in [3] we have that, for $p \geq m$,

$$(\overline{\mathbf{B}}_M^{\bullet} - \mathbf{B})'(XX')(\overline{\mathbf{B}}_M^{\bullet} - \mathbf{B})|_{\tilde{\mathbf{\Sigma}}^{-\mathbf{1}}} \sim W_m\left(\frac{M+1}{M}\tilde{\mathbf{\Sigma}} + \mathbf{\Sigma}, p\right).$$

From Theorem 2.4.2 in [3] and Subsection 7.3.3 in [1] we have

(A.5)
$$\mathbf{H} = \left(\frac{M+1}{M}\tilde{\mathbf{\Sigma}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}}(\overline{\mathbf{B}}_M^{\bullet}-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}_M^{\bullet}-\mathbf{B})\left(\frac{M+1}{M}\tilde{\mathbf{\Sigma}}+\mathbf{\Sigma}\right)^{\prime-\frac{1}{2}} \sim W_m(\mathbf{I}, p)$$

and

(A.6)
$$\mathbf{G} = M(n-p)\tilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\overline{\mathbf{S}}_M^{\bullet}\tilde{\mathbf{\Sigma}}^{\prime-\frac{1}{2}} \sim W_m(\mathbf{I}, M(n-p)).$$

We may thus write $T_M^{\bullet}$ in (2.5) as

$$T_M^{\bullet} = \frac{|(\overline{\mathbf{B}}_M^{\bullet} - \mathbf{B})'(XX')(\overline{\mathbf{B}}_M^{\bullet} - \mathbf{B})|}{|M(n-p)\overline{\mathbf{S}}_M^{\bullet}|} = \frac{\left|\frac{M+1}{M}\tilde{\mathbf{\Sigma}} + \mathbf{\Sigma}\right|}{|\tilde{\mathbf{\Sigma}}|} \times \frac{|\mathbf{H}|}{|\mathbf{G}|},$$

where, $|\mathbf{G}| \sim \prod_{i=1}^{m} \chi^2_{n-p-i+1}$ and $|\mathbf{H}| \sim \prod_{i=1}^{m} \chi^2_{p-i+1}$, with independent chi-square random variables in each product, we end up with a product of independent F-distributions, due to the independence of $\mathbf{H}$ and $\mathbf{G}$, inherited from the independence of $\overline{\mathbf{B}}_M^{\bullet}$ and $\overline{\mathbf{S}}_M^{\bullet}$. So, conditionally on $\tilde{\mathbf{\Sigma}}^{-\mathbf{1}}$, we have

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

where

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

As such, from (A.4), integrating out $\overline{\mathbf{B}}_M^{\bullet}$ and $\overline{\mathbf{S}}_M^{\bullet}$, we end up with the pdf of $\tilde{\mathbf{\Sigma}}^{-1}$ proportional to

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

Making the transformation $\mathbf{\Omega}=\mathbf{\Sigma}^{\frac{1}{2}}\tilde{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}^{\frac{1}{2}}$, which implies $\tilde{\mathbf{\Sigma}}^{-1}=\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Omega}\mathbf{\Sigma}^{-\frac{1}{2}}$, with the Jacobian of the transformation from $\tilde{\mathbf{\Sigma}}^{-1}$ to $\mathbf{\Omega}$ being $|\mathbf{\Sigma}|^{-\frac{m+1}{2}}$, we have the pdf of $\mathbf{\Omega}$ proportional to

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

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

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

independent of $\mathbf{\Sigma}$. Therefore, we may conclude that

$$T_M^{\bullet}|_{\mathbf{\Omega}}\sim\left\{\prod_{i=1}^{m}\frac{p-i+1}{n-p-i+1}F_{p-i+1,M(n-p)-i+1}\right\}\left|\frac{M+1}{M}\mathbf{I_m}+\mathbf{\Omega}\right|$$

where from [6, Theorem 8.2.8.] $\mathbf{\Omega}$ has the same distribution as $\mathbf{A}_1^{\frac{1}{2}}\mathbf{A}_2^{-1}\mathbf{A}_1^{\frac{1}{2}}$ with $\mathbf{A}_1\sim W_m(\mathbf{I}_m,n+\alpha-p-m-1)$ and $\mathbf{A}_2\sim W_m(\mathbf{I}_m,n-p)$, two independent random variables. □

Proof of Corollary 2.3: The proof is identical to the proof of Theorem 2.1 replacing the joint pdf of $(\overline{\mathbf{B}}_M^{\bullet},\overline{\mathbf{S}}_M^{\bullet})$ by the joint pdf of $(\overline{\mathbf{B}}_M^{\bullet},\mathbf{S}_{comb}^{\bullet})$, noting that we have

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

Proof of Corollary 2.4: The proof is identical to that of Theorem 2.2 replacing $\overline{\mathbf{S}}_M^{\bullet}$ by $\mathbf{S}_{comb}^{\bullet}$, noting that from Corollary 2.3, conditional on $\tilde{\mathbf{\Sigma}}$, $\overline{\mathbf{B}}_M^{\bullet}$ is $N_{pm}(\mathbf{B},(\mathbf{\Sigma}+\frac{\mathbf{1}}{\mathbf{M}}\tilde{\mathbf{\Sigma}})\otimes(\mathbf{XX}')^{-\mathbf{1}})$ and $(Mn-p)\mathbf{S}_{comb}^{\bullet}$ is $W_m(\tilde{\mathbf{\Sigma}},Mn-p)$, independent of $\overline{\mathbf{B}}_M^{\bullet}$. □

B. Details on several results

B.1. The posterior distributions for Σ and B

Let us start by observing that $\mathbf{Y}|_{\mathbf{B},\mathbf{\Sigma}}\sim N_{mn}(\mathbf{B}'\mathbf{X},\mathbf{I_n}\otimes\mathbf{\Sigma})$ and that the likelihood function for $\mathbf{Y}$ will be

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

We may then get the joint posterior distribution of $(\mathbf{B},\mathbf{\Sigma})$ from the product of the prior and likelihood functions as

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

The exponent in (B.1) may be written as

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

where, using $\hat{\mathbf{B}}' = \left[(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Y}'\right]' = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}$,

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

Therefore, the joint posterior distribution of $(\mathbf{B},\boldsymbol{\Sigma})$ is proportional to

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

In conclusion, by Corollary 2.4.6.2. in [3], the posterior distribution for $\boldsymbol{\Sigma}$ is

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

and the posterior distribution for $\mathbf{B}$ is

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

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

B.2. Matrix calculations required in the proof of Theorem 2.1

For $\tilde{\mathbf{B}}$, $\mathbf{B}$ and $\mathbf{X}$ defined as in Section 2 we have

$$\begin{aligned}
&M(\tilde{\mathbf{B}}-\overline{\mathbf{B}}^{\bullet}_{M})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}}-\overline{\mathbf{B}}^{\bullet}_{M})+(\tilde{\mathbf{B}}-\hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}}-\hat{\mathbf{B}})= \\
&\quad=(M+1)\tilde{\mathbf{B}}'\mathbf{X}\mathbf{X}'\tilde{\mathbf{B}}-M\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\tilde{\mathbf{B}}-M\tilde{\mathbf{B}}'\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M}+M\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} \\
&\qquad -\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\tilde{\mathbf{B}}-\tilde{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}}+\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} \\
&\quad=(M+1)\tilde{\mathbf{B}}'\mathbf{X}\mathbf{X}'\tilde{\mathbf{B}}-\tilde{\mathbf{B}}'\mathbf{X}\mathbf{X}'(M\overline{\mathbf{B}}^{\bullet}_{M}+\hat{\mathbf{B}})-(M\overline{\mathbf{B}}^{\bullet}_{M}+\hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'\tilde{\mathbf{B}} \\
&\qquad +M\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M}+\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}}
\end{aligned}$$

$$= (M+1)\left[\tilde{\mathbf{B}} - \frac{1}{M+1}(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}})\right]' \mathbf{X}\mathbf{X}' \left[\tilde{\mathbf{B}} - \frac{1}{M+1}(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}})\right]$$
$$+ M\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} - \frac{1}{M+1}(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}}).$$

Since,

$$\begin{aligned}
&M\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} - \frac{1}{M+1}(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}}) \\
&= M\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} \\
&\qquad - \frac{M^2}{M+1}\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} - \frac{1}{M+1}\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} \\
&\qquad - \frac{M}{M+1}\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} - \frac{M}{M+1}\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} \\
&= \frac{M}{M+1}\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} + \frac{M}{M+1}\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} - \frac{M}{M+1}\overline{\mathbf{B}}^{\bullet'}_{M}\mathbf{X}\mathbf{X}'\hat{\mathbf{B}} \\
&\qquad - \frac{M}{M+1}\hat{\mathbf{B}}'\mathbf{X}\mathbf{X}'\overline{\mathbf{B}}^{\bullet}_{M} \\
&= \frac{M}{M+1}(\overline{\mathbf{B}}^{\bullet}_{M} - \hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\overline{\mathbf{B}}^{\bullet}_{M} - \hat{\mathbf{B}})
\end{aligned}$$

we may write

$$\begin{aligned}
&M(\tilde{\mathbf{B}} - \overline{\mathbf{B}}^{\bullet}_{M})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}} - \overline{\mathbf{B}}^{\bullet}_{M}) + (\tilde{\mathbf{B}} - \hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\tilde{\mathbf{B}} - \hat{\mathbf{B}}) = \\
&= (M+1)\left[\tilde{\mathbf{B}} - \frac{1}{M+1}(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}})\right]' \mathbf{X}\mathbf{X}' \left[\tilde{\mathbf{B}} - \frac{1}{M+1}(M\overline{\mathbf{B}}^{\bullet}_{M} + \hat{\mathbf{B}})\right] \\
&\qquad + \frac{M}{M+1}(\overline{\mathbf{B}}^{\bullet}_{M} - \hat{\mathbf{B}})'\mathbf{X}\mathbf{X}'(\overline{\mathbf{B}}^{\bullet}_{M} - \hat{\mathbf{B}})
\end{aligned}$$

B.3. Details about the derivations of results 1, 2 and 5 in Section 2.1

Details on Result 1

From (A.4) we may immediately conclude that the MLE of $\mathbf{B}$ based on the synthetic data will be $\overline{\mathbf{B}}^{\bullet}_{M}$ with

$$E(\overline{\mathbf{B}}^{\bullet}_{M}) = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\frac{1}{M}\sum_{j=1}^{M} E(\mathbf{W}'_j) = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{X}'E(\tilde{\mathbf{B}}) = E(\hat{\mathbf{B}}) = \mathbf{B}$$

and

$$Var(\overline{\mathbf{B}}^{\bullet}_{M}) = Var[E(\overline{\mathbf{B}}^{\bullet}_{M}|\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}})] + E[Var(\overline{\mathbf{B}}^{\bullet}_{M}|\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}})]. \tag{B.2}$$

For the first term in (B.2), we have

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

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

and for the second term, we have

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

so that

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

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

Details on Result 2

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

Details on Result 5

Let us consider $\mathbf{H}$ and $\mathbf{G}$ given by (A.5) and (A.6). We will begin by rewriting all four classical statistics $T_{1,M}^{\bullet}$, $T_{2,M}^{\bullet}$, $T_{3,M}^{\bullet}$ and $T_{4,M}^{\bullet}$ in Subsection 2.1, in order to make them assume the same kind of form and then we will prove why all of them are non-pivotal, without loss of generality considering $M=1$. The first statistic, $T_{1,M}^{\bullet}$ may be rewritten as

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

while $T_{2,M}^{\bullet}$ and $T_{3,M}^{\bullet}$ may be rewritten as

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

$$T_{3,1}^{\bullet} = tr\{\mathbf{H} \times [\mathbf{H}+(\mathbf{2}\tilde{\mathbf{\Sigma}}+\mathbf{\Sigma})^{-\mathbf{1/2}}\tilde{\mathbf{\Sigma}}^{\mathbf{1/2}} \times (n-p)\mathbf{G} \times \tilde{\mathbf{\Sigma}}^{\mathbf{1/2}}(\mathbf{2}\tilde{\mathbf{\Sigma}}+\mathbf{\Sigma})^{-\mathbf{1/2}}]^{-1}\}.$$

Concerning $T_{4,1}^{\bullet}$, we have $T_{4,1}^{\bullet} = \lambda_1$ where λ_1 denotes the largest eigenvalue of

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

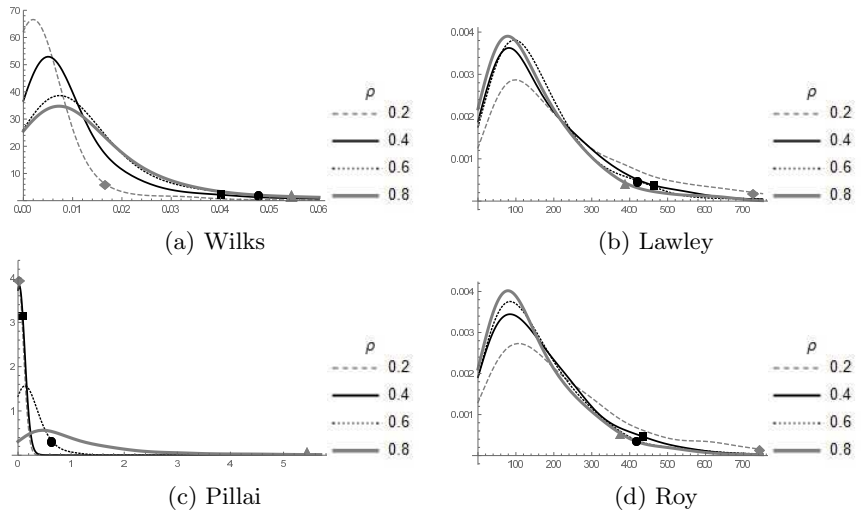
We can observe that a term in the denominator of the expression $T_{1,1}^{\bullet}$ is

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

while in the expressions for the other statistics there are similar terms. These terms involve a product similar to $\tilde{\mathbf{\Sigma}}^{-1/2}(\mathbf{2}\tilde{\mathbf{\Sigma}}+\mathbf{\Sigma})^{1/2}$ that cannot be simplified

to an expression which is not a function of $\mathbf{\Sigma}$, therefore making these statistics non-pivotal.

Thus, in order to illustrate how these statistics are dependent on $\mathbf{\Sigma}$, we can analyze in Figure 3 the empirical distributions of $T^{\bullet}_{1,1}$, $T^{\bullet}_{2,1}$, $T^{\bullet}_{3,1}$ and $T^{\bullet}_{4,1}$ when we consider a simple case where $m = 2$, $p = 3$, $\alpha = 4$, $n = 100$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with $\rho = \{0.2, 0.4, 0.6, 0.8\}$ for a simulation size of 1000.



(a) Wilks (b) Lawley

(c) Pillai (d) Roy

Figure 3: Smoothed empirical distributions and cut-off points (γ=0.05) of $T^{\bullet}_{1,1}$, $T^{\bullet}_{2,1}$, $T^{\bullet}_{3,1}$ and $T^{\bullet}_{4,1}$ for $\rho = \{0.2,0.4,0.6,0.8\}$.

B.4. Details about the derivation of result 1 in Subsection 2.2

Recalling that $(Mn - p)\mathbf{S}^{\bullet}_{comb}|_{\tilde{\mathbf{\Sigma}}} \sim W_m(\tilde{\mathbf{\Sigma}}, Mn - p)$ and that $\tilde{\mathbf{\Sigma}}^{-1}|_{\mathbf{S}} \sim W_m(\frac{1}{n-p}\mathbf{S}^{-1}, n + \alpha - p - m - 1)$ we immediately obtain

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

B.5. Details about the derivations of the results in Section 3

Details on the Expected Values in Section 3

Recall that $(n - p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n - p)$, thus implying that

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

and recall that

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

thus implying that, making $\kappa_{n,\alpha,p,m} = n+\alpha-p-m-1$, given $\mathbf{S}$,

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

since $\prod_{i=1}^m \chi^2_{\kappa_{n,\alpha,p,m}-i+1}$ is a product of independent χ^2 variables. Also recalling that, given $\tilde{\mathbf{\Sigma}}$, we have $M(n-p)\overline{\mathbf{S}}^{\bullet}_M \sim W_m(\tilde{\mathbf{\Sigma}}, M(n-p))$ and $(Mn-p)\mathbf{S}^{\bullet}_{comb} \sim W_m(\tilde{\mathbf{\Sigma}}, Mn-p)$, we may conclude that, given $\tilde{\mathbf{\Sigma}}$,

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

and

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

Combining the results for $E(|(n-p)\mathbf{S}|)$ and $E(|\tilde{\mathbf{\Sigma}}|)|_{\mathbf{S}}$ with each of the expected values for $|M(n-p)\mathbf{S}^{\bullet}_M|$ and $|(Mn-p)\mathbf{S}^{\bullet}_{comb}|$, we end up with the expression for $E(\Upsilon_M)$ found in Section 3.

C. Joining multiple datasets into a single dataset

Let us consider the M synthetic datasets as one only dataset of size nM

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

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

Let

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

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

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

which is same estimator for $\mathbf{B}$ as in (2.8).

Now let

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

be the estimator for $\mathbf{\Sigma}$, based on the dataset of size nM, obtained by joining the M synthetic datasets in one only dataset.

Observe that $\overline{\mathbf{W}}_M = \frac{1}{M}\sum_{j=1}^{M}\mathbf{W}_j$, defined before expression (2.8), can be written as

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

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

Now let us consider the estimator $\mathbf{S}_{\mathbf{w}}$ of $\mathbf{\Sigma}$, defined in the text, before expression (2.8). This estimator may be written as

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

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

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

and the estimator $\mathbf{S}_{mean}$ of $\mathbf{\Sigma}$, defined right after expression (2.9) as

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

We may therefore write the combination estimator $\mathbf{S}_{comb}$ defined in (2.9) as

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

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

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

$$= \mathbf{W}_a\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' + \frac{1}{M^2}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{R}\mathbf{R}'\mathbf{W}_a' + \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\mathbf{R}'\mathbf{W}_a' - \frac{1}{M}\mathbf{W}_a\mathbf{R}\mathbf{R}'\mathbf{X}_a'\mathbf{B}_a + \frac{1}{M}\mathbf{B}_a'\mathbf{X}_a\mathbf{R}\mathbf{R}'\mathbf{X}_a'\mathbf{B}_a\,,$$

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

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

REFERENCES

[1] Anderson, T.W. (2003). *An Introduction To Multivariate Statistical Analysis*, 3rd ed., Wiley, New Jersey.

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

[3] Kollo, T. and Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*, Springer, New York

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

[5] Moura, R., Klein, M., Coelho, C. A. and Sinha, B. (2016). Inference for multivariate regression model based on synthetic data generated using plug-in sampling. *Technical Report - Centro de Matemática e Aplicações (CMA)* 2/2016, (submited for publication).

[6] Muirhead, R.J. (1985). *Aspects of Multivariate Statistical Theory*, 2nd ed., John Wiley & Sons, Inc., New Jersey.

[7] Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009. *Official Journal of the European Union*, **L 87**, 164–173.

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

[9] Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New Jersey.

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