International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 23(2), November, 2023, pp 1-18
DOI: https://doi.org/10.3329/ijss.v23i2.70112

Bayesian Analysis of Singly Imputed Synthetic Data under the Multivariate Normal Model

Abhishek Guin[1], Anindya Roy[1,2] and Bimal Sinha[1,2*]

[1]University of Maryland, Baltimore County, Maryland, USA
[2]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, USA
[*]Correspondence should be addressed to Bimal Sinha
(Email: sinha@umbc.edu)

[Received June 21, 2023; Accepted July 20, 2023]

Abstract

We develop appropriate Bayesian procedures to draw inference about the parameters under a multivariate normal model based on synthetic data. We consider two standard forms of synthetic data, generated under plug in sampling method and posterior predictive sampling method. In addition to point estimates of the mean vector and dispersion matrix, Bayesian credible sets for the mean vector and the generalized variance are also provided under both the scenarios. The analysis in the case when some (partial) features are sensitive and need to be hidden is also briefly indicated.

Keywords and Phrases: Bayesian credible sets, Multivariate normal, Plug-in sampling, Posterior predictive sampling, Synthetic data.

AMS Classification: 62H12, 62H15, 62F15.

1 Introduction

In this paper we present the Bayesian approach for analysis of singly imputed synthetic data generated from a multivariate normal population with both mean vector and covariance matrix unknown. Assume the original confidential data are

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \sim \text{iid} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

where $n > p$, and define $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ (sample mean) and $\boldsymbol{W} = \mathscr{S}_x/(n-1)$ (sample variance) where $\mathscr{S}_x = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$ is the sample Wishart matrix, to be the unbiased estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. We know that $\bar{\boldsymbol{x}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$, $\mathscr{S}_x \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n-1)$, $\bar{\boldsymbol{x}}$ is independent of $\mathscr{S}_x$ and $(\bar{\boldsymbol{x}}, \boldsymbol{W})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when the original data are observed. Our analysis is essentially a sequel to the paper by Klein and Sinha (2016) in which the authors studied the problem of drawing inference about multivariate normal mean vector and dispersion matrix based on the above two types of synthetic data under the *frequentist paradigm.*

The organization of the paper is as follows. In Section 2 we discuss the Plug-in Sampling Method while the Posterior Predictive Sampling Method is studied in Section 3. In both the cases Bayes estimators of μ and Σ as well as Bayesian credible sets for μ and $|\Sigma|$ are derived. Section 4 is devoted to a simulation study under $p = 10$, level of credibility 0.95%, $n = 1000, 10000$ and various choices of δ, the hyper parameter in the prior distribution. Details appear in Tables 1-6. Our recommendation is to use $\delta = 10$ under PIS and $\delta = 20$ and $\alpha = 2$ under PPS, where α is a parameter in the vague prior in this case. The last Section 5 deals with the case when only a part of the data is sensitive and needs to be hidden. Two methods of data analysis are suggested: Method I uses sufficient statistics based only on the sensitive part to impute this part while Method II uses the entire data for imputation purposes. For some standard results on Wishart distribution which will be used throughout this paper, we refer to Anderson (2003) and Muirhead (1982). Some additional relevant references in this context are Klein and Sinha (2013), Klein, Zylstra and Sinha (2019) and Moura et al. (2021).

2 Plug In Sampling method

The singly imputed synthetic data, denoted by $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, are obtained by drawing

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) | \boldsymbol{X} \sim \text{iid} \sim N_p\left(\bar{\boldsymbol{x}}, \boldsymbol{W}\right) \tag{2}$$

Define $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$ (sample mean based on $\boldsymbol{Y}$) and $\mathscr{S}_y = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$ (sample Wishart matrix based on $\boldsymbol{Y}$). Clearly $\bar{\boldsymbol{y}} \sim N_p(\bar{\boldsymbol{x}}, n^{-1}\boldsymbol{W})$, $\mathscr{S}_y \sim \mathcal{W}_p(\boldsymbol{W}, n-1)$. It follows from Lemma 1 in Klein and Sinha (2016) that $(\bar{\boldsymbol{y}}, \mathscr{S}_y)$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also, conditionally given $\boldsymbol{X}$, $\bar{\boldsymbol{y}}$ is independent of $\mathscr{S}_y$, because $\bar{\boldsymbol{y}}$ is independent of $\boldsymbol{y}_i - \bar{\boldsymbol{y}}$ as $\text{Cov}(\bar{\boldsymbol{y}}, \boldsymbol{y}_i - \bar{\boldsymbol{y}}) = 0 \ \forall \ i = 1, \ldots, n$. The following discussion follows from the elucidation in Klein and Sinha (2016).

2.1 Likelihood of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The conditional joint pdf of $(\bar{\boldsymbol{y}}, \mathscr{S}_y)$, given $(\bar{\boldsymbol{x}}, \boldsymbol{W})$, is given by

$$\begin{aligned}
& f(\bar{\boldsymbol{y}}, \mathscr{S}_y \,|\, \bar{\boldsymbol{x}}, \boldsymbol{W}) \\
=& f(\bar{\boldsymbol{y}} \,|\, \bar{\boldsymbol{x}}, \boldsymbol{W}) f(\mathscr{S}_y \,|\, \boldsymbol{W}) \\
\propto & |\boldsymbol{W}|^{-1/2} \exp\left[-\frac{n}{2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}})\right] \times \frac{|\mathscr{S}_y|^{(n-p-2)/2}}{|\boldsymbol{W}|^{(n-1)/2}} \exp\left[-\frac{1}{2}\text{tr}(\mathscr{S}_y \boldsymbol{W}^{-1})\right] \\
=& \frac{|\mathscr{S}_y|^{(n-p-2)/2}}{|\boldsymbol{W}|^{n/2}} \exp\left[-\frac{n}{2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}) - \frac{1}{2}\text{tr}(\mathscr{S}_y \boldsymbol{W}^{-1})\right]
\end{aligned} \tag{3}$$

A similar calculation yields the joint pdf of $(\bar{\boldsymbol{x}}, \boldsymbol{W})$ as

$$f(\bar{\boldsymbol{x}}, \boldsymbol{W} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \frac{|\boldsymbol{W}|^{(n-p-2)/2}}{|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{n}{2}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) - \frac{n-1}{2}\operatorname{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1})\right] \tag{4}$$

We now combine the terms involving $\bar{\boldsymbol{x}}$ from the two exponents as

$$\begin{aligned}
&(\bar{\boldsymbol{y}}-\bar{\boldsymbol{x}})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}}-\bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\\
=&\left\{\bar{\boldsymbol{x}} - [\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]\right\}'\left\{\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}\right\}\\
&\left\{\bar{\boldsymbol{x}} - [\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]\right\}\\
&-\left\{\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\}'\left\{\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}\right\}^{-1}\left\{\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right\} + \bar{\boldsymbol{y}}'\boldsymbol{W}^{-1}\bar{\boldsymbol{y}} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\
=&\left\{\bar{\boldsymbol{x}} - [\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]\right\}'\left\{\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}\right\}\\
&\left\{\bar{\boldsymbol{x}} - [\boldsymbol{W}^{-1}+\boldsymbol{\Sigma}^{-1}]^{-1}[\boldsymbol{W}^{-1}\bar{\boldsymbol{y}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]\right\} + (\bar{\boldsymbol{y}}-\boldsymbol{\mu})'\left(\boldsymbol{\Sigma}+\boldsymbol{W}\right)^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})
\end{aligned}$$

where the simplification is due to Klein and Sinha (2016). Now integrating out $\bar{\boldsymbol{x}}$ from the product of the above two pdfs, we arrive at the following result. S_n^{++} below stands for the set of $p \times p$ positive definite matrices.

Theorem 1. *The joint pdf of $(\bar{\boldsymbol{y}}, \mathscr{S}_y)$ is given by*

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left(\bar{\boldsymbol{y}}, \mathscr{S}_y\right) \propto \int_{S_n^{++}} \frac{|\mathscr{S}_y|^{\frac{n-p-2}{2}}|\boldsymbol{\Sigma}+\boldsymbol{W}|^{-\frac{1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n-1}{2}}|\boldsymbol{W}|^{\frac{p+1}{2}}} e^{-\frac{1}{2}\left[n(\bar{\boldsymbol{y}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+\boldsymbol{W})^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}) + \operatorname{tr}(\mathscr{S}_y\boldsymbol{W}^{-1}) + (n-1)\operatorname{tr}(\boldsymbol{W}\boldsymbol{\Sigma}^{-1})\right]} d\boldsymbol{W}$$

2.2 Posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

We choose the non-informative joint prior: $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\delta}{2}}$ on the parameters. The posterior distribution can be computed by multiplying the expression inside the above integral in Theorem 1, denoted by $h(\bar{\boldsymbol{y}}, \mathscr{S}_y, \boldsymbol{W} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the prior and the product splits up into exactly three parts corresponding to the three conditional posterior distributions, given W. Recall $W = \mathscr{S}_x/(n-1)$.

$$\begin{aligned}
&h(\bar{\boldsymbol{y}}, \mathscr{S}_y, \boldsymbol{W} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\,\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})\\
\propto\ &\left(|\boldsymbol{\Sigma}+\boldsymbol{W}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\left[n(\bar{\boldsymbol{y}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+\boldsymbol{W})^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})\right]}\right)\left(\frac{|\boldsymbol{W}|^{\frac{n-p+\delta-2}{2}}}{|\boldsymbol{\Sigma}|^{\frac{(n-p+\delta-2)+p+1}{2}}} e^{-\frac{1}{2}\left[\operatorname{tr}((n-1)\boldsymbol{W}\boldsymbol{\Sigma}^{-1})\right]}\right)\\
&\left(\frac{|\mathscr{S}_y|^{\frac{n-p+\delta-2}{2}}}{|\boldsymbol{W}|^{\frac{(n-p+\delta-2)+p+1}{2}}} e^{-\frac{1}{2}\left[\operatorname{tr}(\mathscr{S}_y\boldsymbol{W}^{-1})\right]}\right)
\end{aligned}$$

which concedes that the posterior sampling will be done sequentially in the following manner:

$$\boldsymbol{W} \mid \mathscr{S}_y, \bar{\boldsymbol{y}} \sim \mathcal{W}_p^{-1}\left(\mathscr{S}_y, n-p+\delta-2\right) \tag{5}$$

$$\boldsymbol{\Sigma} \mid \boldsymbol{W}, \bar{\boldsymbol{y}}, \mathscr{S}_y \sim \mathcal{W}_p^{-1}\left((n-1)\boldsymbol{W}, n-p+\delta-2\right) \tag{6}$$

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{W}, \bar{\boldsymbol{y}}, \mathscr{S}_y \sim N_p\left(\bar{\boldsymbol{y}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+\boldsymbol{W}\right)\right) \tag{7}$$

We can reformulate the above posterior distributions as:

$$\mathscr{S}_y^{-1/2}\boldsymbol{W}\mathscr{S}_y^{-1/2} \sim \mathcal{W}_p^{-1}\left(\boldsymbol{I}_p, n-p+\delta-2\right) \tag{8}$$

$$\boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2} \sim \mathcal{W}_p^{-1}\left((n-1)\boldsymbol{I}_p, n-p+\delta-2\right) \tag{9}$$

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \boldsymbol{W}, \bar{\boldsymbol{y}} \sim N_p\left(\bar{\boldsymbol{y}}, \frac{1}{n}(\boldsymbol{\Sigma}+\boldsymbol{W})\right) \tag{10}$$

which has the benefit that $\mathscr{S}_y^{-1/2}\boldsymbol{W}\mathscr{S}_y^{-1/2}$ is independent of $\boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2}$ and their posterior distributions are unconditional. The posterior distributions are proper as long as $n > \max\{p, 2p-\delta+1\}$.

2.3 Bayes Estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\mu}}_{\text{BAYES}} = \mathbb{E}(\boldsymbol{\mu} \mid \bar{\boldsymbol{y}}, \mathscr{S}_y) = \mathbb{E}_{\boldsymbol{W}}\mathbb{E}_{\boldsymbol{\Sigma}}\mathbb{E}(\boldsymbol{\mu} \mid \bar{\boldsymbol{y}}, \boldsymbol{\Sigma}, \boldsymbol{W}) = \mathbb{E}_{\boldsymbol{W}}\mathbb{E}_{\boldsymbol{\Sigma}}\mathbb{E}(\bar{\boldsymbol{y}}) = \bar{\boldsymbol{y}}$$

$$\hat{\boldsymbol{\Sigma}}_{\text{BAYES}} = \mathbb{E}(\boldsymbol{\Sigma} \mid \bar{\boldsymbol{y}}, \mathscr{S}_y) = \mathbb{E}_{\boldsymbol{W}}\mathbb{E}(\boldsymbol{\Sigma} \mid \mathscr{S}_y, \boldsymbol{W}) = \mathbb{E}_{\boldsymbol{W}}\left(\tfrac{(n-1)\boldsymbol{W}}{(n-2p+\delta-3)} \mid \mathscr{S}_y\right) = \frac{(n-1)\mathscr{S}_y}{(n-2p+\delta-3)^2}$$

$$\widehat{|\boldsymbol{\Sigma}|}_{\text{BAYES}} = \mathbb{E}\left(|\boldsymbol{\Sigma}| \mid \bar{\boldsymbol{y}}, \mathscr{S}_y\right) = \mathbb{E}_{\boldsymbol{W}}\mathbb{E}\left(|\boldsymbol{\Sigma}| \mid \mathscr{S}_y, \boldsymbol{W}\right) = \mathbb{E}_{\boldsymbol{W}}\left(|\boldsymbol{W}|\,\mathbb{E}\left(\left|\boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2}\right|\right) \mid \mathscr{S}_y\right)$$
$$= \left(\prod_{j=1}^{p}\frac{n-1}{n-p+\delta-j-3}\right)\mathbb{E}\left(|\boldsymbol{W}| \mid \mathscr{S}_y\right) = \left(\prod_{j=1}^{p}\frac{n-1}{(n-p+\delta-j-3)^2}\right)|\mathscr{S}_y|$$

provided $n > \max\{p, 2p-\delta+5\}$, and we use the results: If $\boldsymbol{S} \sim \mathcal{W}_p^{-1}(\boldsymbol{\Sigma}, \nu)$ then

$$\mathbb{E}(\boldsymbol{S}) = (\nu-p-1)^{-1}\boldsymbol{\Sigma} \quad \text{if } \nu > p+1 \tag{11}$$

$$\mathbb{E}(|\boldsymbol{S}|) = |\boldsymbol{\Sigma}|\prod_{j=1}^{p}(\nu-j-1)^{-1} \quad \text{if } \nu > p+3 \tag{12}$$

2.4 Credible Sets for $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\mu}$

We see that $\boldsymbol{\Sigma}^{-1} \mid \boldsymbol{W} \sim \mathcal{W}_p\left(\boldsymbol{W}^{-1}/(n-1), n-p+\delta-2\right)$, so

$$\frac{\left|\boldsymbol{\Sigma}^{-1}\right|}{\left|\boldsymbol{W}^{-1}/(n-1)\right|} \sim \prod_{i=1}^{p} u_i, \quad \text{where } u_i \sim \chi^2_{n-p+\delta-i-1} \text{ independently for } i = 1, \ldots, p$$

which also shows that the quantity on the left hand side of the above relation is independent of $\boldsymbol{W}$. Here we use the following result: If $\boldsymbol{S} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$ then

$$\frac{|\boldsymbol{S}|}{|\boldsymbol{\Sigma}|} \sim \prod_{i=1}^{p} t_i, \quad \text{where } t_i \sim \chi^2_{n-i+1} \text{ independently for } i = 1, \ldots, p \tag{13}$$

Thus using the above result similarly as before, we can get

$$\frac{|\boldsymbol{W}^{-1}|}{|\mathscr{S}_y^{-1}|} \sim \prod_{j=1}^{p} v_j, \quad \text{where } v_j \sim \chi^2_{n-p+\delta-j-1} \text{ independently for } j = 1, \ldots, p$$

So we can define a pivot for the *generalized variance* $|\boldsymbol{\Sigma}|$ as $N \coloneqq \left|\boldsymbol{\Sigma}\mathscr{S}_y^{-1}\right|$ where

$$N^{-1} \sim \frac{\prod_{i=1}^{p} u_i}{(n-1)^p} \prod_{j=1}^{p} v_j$$

where u_i's and v_j's are as above and they are all pairwise independent. A $(1-\gamma)$ level credible set for $|\boldsymbol{\Sigma}|$ based on N is

$$\left[a_{n,p,\delta;\gamma} \left|\mathscr{S}_y\right|, b_{n,p,\delta;\gamma} \left|\mathscr{S}_y\right|\right]$$

where $a_{n,p,\delta;\gamma}$ and $b_{n,p,\delta;\gamma}$ are any two constants that satisfy $1-\gamma = P(a_{n,p,\delta;\gamma} \le N \le b_{n,p,\delta;\gamma})$. The length of the credible interval is $|\mathscr{S}_y| \, (b_{n,p,\delta;\gamma} - a_{n,p,\delta;\gamma})$.

Next we define the pivot for $\boldsymbol{\mu}$ as

$$T^2 \coloneqq n(\boldsymbol{\mu} - \bar{\boldsymbol{y}})' \mathscr{S}_y^{-1} (\boldsymbol{\mu} - \bar{\boldsymbol{y}})$$

We will prove that T^2 is a pivot and derive a sampling scheme in what follows. We notice that

$$\sqrt{n}\mathscr{S}_y^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{y}}) \mid \boldsymbol{\Sigma}, \boldsymbol{W} \sim N_p(\boldsymbol{0}, \boldsymbol{A})$$

where $\boldsymbol{A} = \mathscr{S}_y^{-1/2}\boldsymbol{W}^{1/2}(\boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2}+\boldsymbol{I}_p)\boldsymbol{W}^{1/2}\mathscr{S}_y^{-1/2}$, which is obviously defined through the parameters $(\boldsymbol{\Sigma}, \boldsymbol{W})$. If we can prove that the distribution of $\boldsymbol{A}$ is free of $(\boldsymbol{\Sigma}, \boldsymbol{W})$, then by using the fact that if $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{A})$ then $\boldsymbol{Z}'\boldsymbol{Z} \sim \sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{A}$ and χ^2_{1i} are independent χ^2_1 variables, we can conclude that T^2 is a pivot. Taking $\boldsymbol{Z} = \sqrt{n}\mathscr{S}_y^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{y}})$, $\boldsymbol{B} = \mathscr{S}_y^{-1/2}\boldsymbol{W}\mathscr{S}_y^{-1/2}$ it finally follows that:

a) the conditional distribution of $T^2 \mid \boldsymbol{A}$ is $\sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where $\lambda_1, \ldots, \lambda_p$ are the roots of $|\boldsymbol{A} - \lambda \boldsymbol{I}_p| = 0$ such that $\boldsymbol{A} \mid \boldsymbol{B} \overset{d}{=} \mathcal{W}_p^{-1}\left((n-1)\boldsymbol{B}, n-p+\delta-2\right) + \boldsymbol{B}$ by (9) and $\boldsymbol{B} \sim \mathcal{W}_p^{-1}\left(\boldsymbol{I}_p, n-p+\delta-2\right)$ by (8); and

b) the unconditional distribution of T^2 is obtained by averaging over the joint distribution of the roots $\lambda_1, \ldots, \lambda_p$.

We have shown that T^2 is a pivotal quantity, and therefore a $(1-\gamma)$ credible ellipsoid for $\boldsymbol{\mu}$ based on T^2 is given by

$$\left\{\boldsymbol{\mu} \;:\; T^2 \le c_{n,p,\delta;\gamma}\right\}$$

where $c_{n,p,\delta;\gamma}$ satisfies $1-\gamma = P(T^2 \le c_{n,p,\delta;\gamma})$. From the above discussion, it follows that the cut-off point $c_{n,p,\delta;\gamma}$ can be obtained by simulating the distribution of T^2 as follows.

1. Generate $\boldsymbol{B} \sim \mathcal{W}_p^{-1}\left(\boldsymbol{I}_p, n-p+\delta-2\right)$.

2. Generate $\boldsymbol{A} \,|\, \boldsymbol{B} \sim \mathcal{W}_p^{-1}\left((n-1)\boldsymbol{B}, n-p+\delta-2\right) + \boldsymbol{B}$.
3. Generate $\lambda_1, \ldots, \lambda_p$, the roots of $|\boldsymbol{A} - \lambda \boldsymbol{I}_p| = 0$.
4. Generate $T^2 = \sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where χ^2_{1i} are independent χ^2_1 variables.

The volume of the credible ellipsoid is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Y}) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} \left(c_{n,p,\delta;\gamma}/n\right)^{p/2} |\mathscr{S}_y|^{1/2}$$

3 Posterior Predictive Sampling method

We return to the setup of the last section. Under the *posterior predictive sampling* method, starting with a vague prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\alpha/2}$, the joint (imputed) posterior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, given $\boldsymbol{X}$, can be represented as

$$\begin{aligned} \boldsymbol{\Sigma} \,|\, \boldsymbol{X} &\sim \mathcal{W}_p^{-1}\left((n-1)\boldsymbol{W}, n-p+\alpha-2\right) \\ \boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \boldsymbol{X} &\sim N_p\left(\bar{\boldsymbol{x}}, n^{-1}\boldsymbol{\Sigma}\right) \end{aligned} \tag{14}$$

We assume throughout that $n + \alpha > 2p + 1$We now draw $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from the above posterior, resulting in $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, and then draw a random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ as iid from $N_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, which form the singly imputed synthetic data that are released. Define $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$ (sample mean based on $\boldsymbol{Y}$) and $\mathscr{S}_y = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$ (sample Wishart matrix based on $\boldsymbol{Y}$) which are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by Lemma 1 in Klein and Sinha (2016). Proceeding as in Theorem 1, we have the following result. A proof of Theorem 2 appears in Appendix A.

Theorem 2. *The joint pdf of $\bar{\boldsymbol{y}}$ and $\mathscr{S}_y$ is obtained by integrating out $\boldsymbol{\Sigma}^*$ from the joint pdf of $(\bar{\boldsymbol{y}}, \mathscr{S}_y, \boldsymbol{\Sigma}^*)$ given by*

$$\begin{aligned} f(\bar{\boldsymbol{y}}, \mathscr{S}_y, \boldsymbol{\Sigma}^*) \propto\, & e^{-\frac{1}{2}\left[n(\bar{\boldsymbol{y}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*)^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}) + \mathrm{tr}(\mathscr{S}_y \boldsymbol{\Sigma}^{*-1})\right]} |\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^*|^{-\frac{1}{2}} |\boldsymbol{\Sigma}|^{\frac{n-p+\alpha-2}{2}} \\ & |\boldsymbol{\Sigma} + \boldsymbol{\Sigma}^*|^{-\frac{2n-p+\alpha-3}{2}} |\boldsymbol{\Sigma}^*|^{-\left(\frac{p+1}{2}+\alpha\right)} |\mathscr{S}_y|^{\frac{n-p-2}{2}} \end{aligned}$$

3.1 Posterior distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

We choose the same prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\delta}{2}}$ as before and attempt to compute the posterior distribution as before by multiplying the expression inside the above integral in Theorem 2, denoted by $h(\bar{\boldsymbol{y}}, \mathscr{S}_y, \boldsymbol{W} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the prior and the product should split up into exactly three parts corresponding to the three conditional posterior distributions.

$$\begin{aligned} h(\bar{\boldsymbol{y}}, \mathscr{S}_y, \boldsymbol{W} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto & \left(|\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^*|^{-\frac{1}{2}} e^{-\frac{1}{2}\left[n(\bar{\boldsymbol{y}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*)^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})\right]}\right) \\ & \times \left(|\boldsymbol{\Sigma}|^{\frac{n-p+\alpha-\delta-2}{2}} |\boldsymbol{\Sigma} + \boldsymbol{\Sigma}^*|^{-\frac{2n-p+\alpha-3}{2}} |\boldsymbol{\Sigma}^*|^{-\left(\frac{p+1}{2}+\alpha\right)} e^{-\frac{1}{2}\left[\mathrm{tr}(\mathscr{S}_y \boldsymbol{\Sigma}^{*-1})\right]}\right) \end{aligned} \tag{15}$$

We see that the part involving $\boldsymbol{\mu}$ separates out nicely in front and thus its posterior distribution, conditional on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^*$, is obvious. We will now work with the part inside the second parenthesis involving just the determinants below.

$$\begin{aligned}
&|\boldsymbol{\Sigma}|^{\frac{n-p+\alpha-\delta-2}{2}}\,|\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*|^{-\frac{2n-p+\alpha-3}{2}}\,|\boldsymbol{\Sigma}^*|^{-(\frac{p+1}{2}+\alpha)}\\
&=\left|\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}\right|^{\frac{n-p+\alpha-\delta-2}{2}}\left|I_p+\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}\right|^{-\frac{2n-p+\alpha-3}{2}}|\boldsymbol{\Sigma}^*|^{-(\frac{n+p+\delta}{2}+\alpha)} \qquad (16)
\end{aligned}$$

Next we combine equations (15) and (16) and multiply by the Jacobian of the transformation $\boldsymbol{\Sigma}\mapsto\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}$ which is $|\boldsymbol{\Sigma}^*|^p$ to get

$$\begin{aligned}
&h(\bar{\boldsymbol{y}},\mathscr{S}_y,\boldsymbol{W}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})\,\pi(\boldsymbol{\mu},\boldsymbol{\Sigma})\\
\propto\ &\left(|\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*|^{-\frac{1}{2}}\,e^{-\frac{1}{2}\left[n(\bar{\boldsymbol{y}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*)^{-1}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})\right]}\right)\\
&\left(\left|\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}\right|^{\frac{n-p+\alpha-\delta-2}{2}}\left|I_p+\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}\right|^{-\frac{2n-p+\alpha-3}{2}}\right)\\
&\left(|\boldsymbol{\Sigma}^*|^{-(\frac{n-p+\delta}{2}+\alpha)}\,e^{-\frac{1}{2}\left[\mathrm{tr}(\mathscr{S}_y\boldsymbol{\Sigma}^{*-1})\right]}\right)
\end{aligned}$$

which indicates that the posterior sampling will be done sequentially in the following:

$$\boldsymbol{\Sigma}^*\,|\,\mathscr{S}_y \sim \mathcal{W}_p^{-1}\left(\mathscr{S}_y, n-2p+\delta-1+2\alpha\right) \qquad (17)$$

$$\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} \sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n+\alpha-\delta-1}{2},\frac{n-p+\delta-2}{2}\right) \qquad (18)$$

$$\boldsymbol{\mu}\,|\,\boldsymbol{\Sigma},\boldsymbol{\Sigma}^*,\bar{\boldsymbol{y}} \sim N_p\left(\bar{\boldsymbol{y}},\frac{1}{n}\left(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*\right)\right) \qquad (19)$$

where $\mathrm{B}_{\mathrm{p}}^{\mathrm{II}}(a,b)$ denotes the matrix variate beta type II distribution as described in Muirhead (1982). We can reformulate the above posterior distributions as:

$$\mathscr{S}_y^{-1/2}\boldsymbol{\Sigma}^*\mathscr{S}_y^{-1/2} \sim \mathcal{W}_p^{-1}\left(\boldsymbol{I}_p, n-2p+\delta-1+2\alpha\right) \qquad (20)$$

$$\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} \sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n+\alpha-\delta-1}{2},\frac{n-p+\delta-2}{2}\right) \qquad (21)$$

$$\boldsymbol{\mu}\,|\,\boldsymbol{\Sigma},\boldsymbol{\Sigma}^*,\bar{\boldsymbol{y}} \sim N_p\left(\bar{\boldsymbol{y}},\frac{1}{n}\left(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*\right)\right) \qquad (22)$$

which has the benefit that $\mathscr{S}_y^{-1/2}\boldsymbol{\Sigma}^*\mathscr{S}_y^{-1/2}$ is independent of $\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}$ and its posterior distribution is unconditional. The posterior distributions are proper as long as $n>\max\{p,2p-\alpha+1,3p-\delta,p-\alpha+\delta,2p-\delta+1-2\alpha\}$.

3.2 Bayes Estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$$\hat{\boldsymbol{\mu}}_{\mathrm{BAYES}}=\mathbb{E}(\boldsymbol{\mu}\,|\,\bar{\boldsymbol{y}},\mathscr{S}_y)=\mathbb{E}_{\boldsymbol{\Sigma}^*}\mathbb{E}_{\boldsymbol{\Sigma}}\mathbb{E}(\boldsymbol{\mu}\,|\,\bar{\boldsymbol{y}},\boldsymbol{\Sigma},\boldsymbol{\Sigma}^*)=\mathbb{E}_{\boldsymbol{\Sigma}^*}\mathbb{E}_{\boldsymbol{\Sigma}}\mathbb{E}(\bar{\boldsymbol{y}})=\bar{\boldsymbol{y}}$$

Finding $\hat{\boldsymbol{\Sigma}}_{\text{BAYES}}$ seems very difficult.

$$\begin{aligned}
&\widehat{|\boldsymbol{\Sigma}|}_{\text{BAYES}} = \mathbb{E}\left(|\boldsymbol{\Sigma}| \mid \bar{\boldsymbol{y}}, \mathscr{S}_y\right) = \mathbb{E}_{\boldsymbol{\Sigma}^*}\mathbb{E}\left(|\boldsymbol{\Sigma}| \mid \mathscr{S}_y, \boldsymbol{\Sigma}^*\right) \\
&= \mathbb{E}_{\boldsymbol{\Sigma}^*}\left(|\boldsymbol{\Sigma}^*|\,\mathbb{E}\left(\left|\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}\right|\right) \mid \mathscr{S}_y\right) = \left(\prod_{j=1}^{p}\frac{n+\alpha-\delta-j}{n-p+\delta-j-3}\right)\mathbb{E}\left(|\boldsymbol{\Sigma}^*| \mid \mathscr{S}_y\right) \\
&= \left(\prod_{j=1}^{p}\frac{n+\alpha-\delta-j}{(n-p+\delta-j-3)(n-2p+\delta+2\alpha-j-2)}\right)|\mathscr{S}_y|
\end{aligned}$$

provided that $n > \max(p, p-\alpha+\delta, 2p+3-\delta, 3p+2-\delta+2\alpha)$. We use (12) and the following result for the above derivation: If $\boldsymbol{V} \sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}(a,b)$ then

$$\mathbb{E}\left(|\boldsymbol{V}|\right) = \prod_{j=1}^{p}\frac{a-\frac{1}{2}(j-1)}{b-\frac{1}{2}(j+1)} \quad \text{if } a > \frac{p-1}{2},\, b > \frac{p+1}{2}. \tag{23}$$

3.3 Credible Sets for $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\mu}$

Let $\boldsymbol{C} = \boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2}$. Then by (21), we have $\boldsymbol{C}^{-1} \sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n-p+\delta-3}{2}, \frac{n+\alpha-\delta}{2}\right)$. Also by (13) we can get,

$$\frac{\left|\boldsymbol{\Sigma}^{*-1}\right|}{\left|\mathscr{S}_y^{-1}\right|} \sim \prod_{i=1}^{p} v_j, \quad \text{where } v_j \sim \chi^2_{n-2p+\delta+2\alpha-j} \text{ independently for } j = 1, \ldots, p$$

We can define a pivot for $|\boldsymbol{\Sigma}|$ in the same manner as in the last section to be $N \coloneqq \left|\boldsymbol{\Sigma}\mathscr{S}_y^{-1}\right|$ where

$$N^{-1} \sim |\boldsymbol{M}|\prod_{j=1}^{p} v_j$$

where v_j's are defined as above, $\boldsymbol{M} \sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n-p+\delta-3}{2}, \frac{n+\alpha-\delta}{2}\right)$ and $\boldsymbol{M}$ is independent of v_j, $\forall j$. Since the distribution of N is free of $(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$ we conclude that it is a pivot. A $(1-\gamma)$ level credible set for $|\boldsymbol{\Sigma}|$ is

$$\left[a_{n,p,\alpha,\delta;\gamma}\left|\mathscr{S}_y\right|, b_{n,p,\alpha,\delta;\gamma}\left|\mathscr{S}_y\right|\right]$$

where $a_{n,p,\alpha,\delta;\gamma}$ and $b_{n,p,\alpha,\delta;\gamma}$ are any two constants that satisfy $1-\gamma = P(a_{n,p,\alpha,\delta;\gamma} \le N \le b_{n,p,\alpha,\delta;\gamma})$. The length of the credible interval is $\left|\mathscr{S}_y\right|(b_{n,p,\alpha,\delta;\gamma} - a_{n,p,\alpha,\delta;\gamma})$.

Next we define the pivot similarly as in the last section for $\boldsymbol{\mu}$ as

$$T^2 \coloneqq n(\boldsymbol{\mu}-\bar{\boldsymbol{y}})'\mathscr{S}_y^{-1}(\boldsymbol{\mu}-\bar{\boldsymbol{y}})$$

We will prove that T^2 is a pivot and derive a sampling scheme in what follows. We notice that

$$\sqrt{n}\mathscr{S}_y^{-1/2}(\boldsymbol{\mu}-\bar{\boldsymbol{y}}) \mid \boldsymbol{\Sigma}, \boldsymbol{\Sigma}^* \sim N_p(\boldsymbol{0}, \boldsymbol{A})$$

where $\boldsymbol{A} = \mathscr{S}_y^{-1/2}\boldsymbol{\Sigma}^{*1/2}(\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} + 2\boldsymbol{I}_p)\boldsymbol{\Sigma}^{*1/2}\mathscr{S}_y^{-1/2}$, which is obviously defined through the parameters $(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$. If we can prove that the distribution of $\boldsymbol{A}$ is free of $(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*)$, then by using the fact that if $\boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{A})$ then $\boldsymbol{Z}'\boldsymbol{Z} \sim \sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{A}$ and χ^2_{1i} are independent χ^2_1 variables, we can conclude that T^2 is a pivot. Taking $\boldsymbol{Z} = \sqrt{n}\mathscr{S}_y^{-1/2}(\boldsymbol{\mu} - \bar{\boldsymbol{y}})$, $\boldsymbol{B} = \mathscr{S}_y^{-1/2}\boldsymbol{\Sigma}^*\mathscr{S}_y^{-1/2}$ it finally follows that:

(a) the conditional distribution of $T^2 \mid \boldsymbol{A}$ is $\sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where $\lambda_1, \ldots, \lambda_p$ are the roots of $|\boldsymbol{A} - \lambda \boldsymbol{I}_p| = 0$ such that $\boldsymbol{B} \sim \mathcal{W}_p^{-1}(\boldsymbol{I}_p, n - 2p + \delta + 2\alpha - 1)$ by (20) and

$$\boldsymbol{A} \mid \boldsymbol{B} \overset{d}{=} \mathrm{GB}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n+\alpha-\delta}{2}, \frac{n-p+\delta-3}{2}; \boldsymbol{B}, \boldsymbol{O}\right) + 2\boldsymbol{B}$$

where $\mathrm{GB}_{\mathrm{p}}^{\mathrm{II}}(a, b; \boldsymbol{\Omega}, \boldsymbol{\Psi})$ denotes the generalized matrix variate beta type II distribution as described in [7]. The above derivation follows from (21) and the result: If $\boldsymbol{V} \sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}(a, b)$, $\boldsymbol{A}_{p\times p}$ is a constant, non-singular matrix then $\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}' \sim \mathrm{GB}_{\mathrm{p}}^{\mathrm{II}}(a, b; \boldsymbol{A}\boldsymbol{A}', \boldsymbol{O})$.

(b) the unconditional distribution of T^2 is obtained by averaging over the joint distribution of the roots $\lambda_1, \ldots, \lambda_p$.

We have shown that T^2 is a pivotal quantity, and therefore a $(1-\gamma)$ credible ellipsoid for $\boldsymbol{\mu}$ based on T^2 is given by

$$\left\{\boldsymbol{\mu} \;:\; T^2 \leq c_{n,p,\alpha,\delta;\gamma}\right\}$$

where $c_{n,p,\alpha,\delta;\gamma}$ satisfies $1 - \gamma = P(T^2 \leq c_{n,p,\alpha,\delta;\gamma})$. From the above discussion, it follows that the cut-off point $c_{n,p,\alpha,\delta;\gamma}$ can be obtained by simulating the distribution of T^2 as follows.

1. Generate $\boldsymbol{B} \sim \mathcal{W}_p^{-1}(\boldsymbol{I}_p, n - 2p + \delta + 2\alpha - 1)$.
2. Generate $\boldsymbol{A} \mid \boldsymbol{B} \sim \mathrm{GB}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n+\alpha-\delta}{2}, \frac{n-p+\delta-3}{2}; \boldsymbol{B}, \boldsymbol{O}\right) + 2\boldsymbol{B}$.
3. Generate $\lambda_1, \ldots, \lambda_p$, the roots of $|\boldsymbol{A} - \lambda \boldsymbol{I}_p| = 0$.
4. Generate $T^2 = \sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where χ^2_{1i} are independent χ^2_1 variables.

The volume of the credible ellipsoid is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Y}) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} \left(c_{n,p,\alpha,\delta;\gamma}/n\right)^{p/2} |\mathscr{S}_y|^{1/2}$$

4 Simulation Studies

To conduct the simulation, the population distribution is taken to be the multivariate normal model (1) with

$$p = 10, \;\; \boldsymbol{\mu} = 0.1 \times \begin{pmatrix} 1 & 2 & \ldots & 10 \end{pmatrix}', \;\; \boldsymbol{\Sigma} = 0.25\boldsymbol{I}_p + 0.75\boldsymbol{J}_p, \tag{24}$$

where $\boldsymbol{I}_p$ is the $p \times p$ dimensional identity matrix and $\boldsymbol{J}_p$ is the $p \times p$ matrix of 1's. Based on Monte Carlo simulation with 10^4 iterations, we compute an estimate of the coverage probability, the volume or length (as appropriate) of the respective credible sets and the Bayes estimators of $\boldsymbol{\mu}$ and $|\boldsymbol{\Sigma}|$, where in all cases, the level of credibility is set at 0.95.

In both PIS and PPS cases, increasing δ increases the coverage of $|\mathbf{\Sigma}|$ before it drops off, the effect hastened for small values of n. We thus see the reverse-sigmoid shape of the curve in all situations, it is wider in the PIS case, and the curve seems to shift to the right with increasing α in the PPS case. So the best choice of δ to ensure maximum coverage of $|\mathbf{\Sigma}|$ would increase with increasing α, in the PPS case.
For $\boldsymbol{\mu}$, we see that increasing δ slightly increases the coverage before decreasing steadily, albeit at a much slower pace than that of $|\mathbf{\Sigma}|$.
The size of the credible sets of both quantities shrink with either increasing n or δ. Asymptotically the results conform to our expectations, with the inference worsening for higher δ, quicker for $|\mathbf{\Sigma}|$ than $\boldsymbol{\mu}$. The better inference we get off the PIS method than the PPS method attests to the trade-off between data utility and data privacy.
The recommendation is to use $\delta = 10$ in the PIS case, PPS case with $\alpha = 2$ and $\delta = 20$ in the PPS case with $\alpha = 50$.

Table 1: Inference for $\boldsymbol{\mu}$ and $|\mathbf{\Sigma}|$ for SI PIS MVN data with $n = 1000$

	$\lvert\mathbf{\Sigma}\rvert$		$\boldsymbol{\mu}$	
δ	*avg cvg*	*est len*	*avg cvg*	*est vol*
0.2	0.8093	3.0625e-05	0.9531	1.0384e-09
0.5	0.8125	3.0607e-05	0.9552	1.0399e-09
0.8	0.8319	3.0254e-05	0.9595	1.1277e-09
1	0.8294	3.0172e-05	0.9581	1.0437e-09
2	0.8661	2.9942e-05	0.9572	1.0776e-09
3	0.8755	2.8962e-05	0.9501	1.0287e-09
4	0.8874	2.7952e-05	0.9533	1.0288e-09
5	0.9108	2.7679e-05	0.9542	1.0313e-09
10	0.9517	2.4692e-05	0.9494	9.6189e-10
20	0.8491	2.0096e-05	0.95	9.0293e-10
30	0.5135	1.6400e-05	0.9396	8.0198e-10
50	0.0303	1.1083e-05	0.9297	6.9183e-10
100	0	4.1404e-06	0.8979	4.9770e-10

Table 2: Inference for $\boldsymbol{\mu}$ and $|\boldsymbol{\Sigma}|$ for SI PIS MVN data with $n = 10000$

	$\|\boldsymbol{\Sigma}\|$		$\boldsymbol{\mu}$	
δ	*avg cvg*	*est len*	*avg cvg*	*est vol*
0.2	0.9353	7.5066e-06	0.9475	9.0069e-15
0.5	0.9347	7.4725e-06	0.9494	9.2580e-15
0.8	0.9379	7.4664e-06	0.947	9.1447e-15
1	0.9386	7.4609e-06	0.9491	8.9631e-15
2	0.9408	7.5066e-06	0.9516	9.4023e-15
3	0.9464	7.5633e-06	0.9505	9.4476e-15
4	0.9401	7.3873e-06	0.9492	9.0631e-15
5	0.9483	7.4442e-06	0.9462	8.7338e-15
10	0.9477	7.3689e-06	0.9478	9.3546e-15
20	0.9422	7.3092e-06	0.9442	8.9443e-15
30	0.9136	7.1060e-06	0.9532	9.6100e-15
50	0.7604	6.7308e-06	0.9519	9.2865e-15
100	0.1918	6.1571e-06	0.9465	8.3714e-15

Table 3: Inference for $\boldsymbol{\mu}$ and $|\boldsymbol{\Sigma}|$ for SI PPS MVN data with $\alpha = 2$, $n = 1000$

	$\|\boldsymbol{\Sigma}\|$		$\boldsymbol{\mu}$	
δ	*avg cvg*	*est len*	*avg cvg*	*est vol*
0.2	0.8259	4.0296e-05	0.9552	1.0947e-09
0.5	0.8561	4.0295e-05	0.9523	1.0695e-09
0.8	0.8521	4.0643e-05	0.9571	1.0774e-09
1	0.8745	4.0285e-05	0.9588	1.0814e-09
2	0.8862	3.8019e-05	0.9586	1.0862e-09
3	0.9157	3.7622e-05	0.9591	1.1022e-09
4	0.9444	3.6284e-05	0.9542	1.0461e-09
5	0.9556	3.5436e-05	0.9502	1.0046e-09
10	0.9864	3.0662e-05	0.9536	1.0498e-09
20	0.7985	2.2145e-05	0.9472	9.0976e-10
30	0.2629	1.6346e-05	0.9381	7.8479e-10
50	0.0001	8.8689e-06	0.9298	6.6995e-10
100	0	1.9857e-06	0.8784	3.9465e-10

Table 4: Inference for $\boldsymbol{\mu}$ and $|\boldsymbol{\Sigma}|$ for SI PPS MVN data with $\alpha = 2$, $n = 10000$

	$\|\boldsymbol{\Sigma}\|$		$\boldsymbol{\mu}$	
δ	*avg cvg*	*est len*	*avg cvg*	*est vol*
0.2	0.974	9.4585e-06	0.9481	9.0549e-15
0.5	0.9722	9.2176e-06	0.955	9.4496e-15
0.8	0.975	9.2989e-06	0.949	9.3798e-15
1	0.9729	9.3212e-06	0.9552	9.8925e-15
2	0.9728	9.0660e-06	0.9522	9.6263e-15
3	0.9799	9.1529e-06	0.9477	8.8715e-15
4	0.9789	9.2158e-06	0.9569	9.8428e-15
5	0.9813	9.2459e-06	0.9541	1.0064e-14
10	0.9835	8.9542e-06	0.9494	8.8645e-15
20	0.9672	8.6689e-06	0.9489	8.7975e-15
30	0.924	8.4805e-06	0.9502	9.0701e-15
50	0.7123	8.1242e-06	0.9449	8.5905e-15
100	0.0291	6.8814e-06	0.9431	8.3882e-15

Table 5: Inference for $\boldsymbol{\mu}$ and $|\boldsymbol{\Sigma}|$ for SI PPS MVN data with $\alpha = 50$, $n = 1000$

	$\|\boldsymbol{\Sigma}\|$		$\boldsymbol{\mu}$	
δ	*avg cvg*	*est len*	*avg cvg*	*est vol*
0.2	0.0742	6.4298e-05	0.961	1.1587e-09
0.5	0.0841	6.3811e-05	0.9641	1.3207e-09
0.8	0.0826	6.4185e-05	0.9616	1.2371e-09
1	0.1073	6.3435e-05	0.9596	1.2032e-09
2	0.1248	6.1302e-05	0.9654	1.2527e-09
3	0.1657	6.0413e-05	0.9601	1.2380e-09
4	0.1925	5.8477e-05	0.9541	1.1280e-09
5	0.2418	5.7401e-05	0.9602	1.1985e-09
10	0.5246	4.7836e-05	0.9598	1.1788e-09
20	0.9323	3.5425e-05	0.953	1.0537e-09
30	0.963	2.6544e-05	0.948	9.4866e-10
50	0.1188	1.4721e-05	0.9335	7.2119e-10
100	0	3.2798e-06	0.9007	4.6731e-10

Table 6: Inference for $\boldsymbol{\mu}$ and $|\boldsymbol{\Sigma}|$ for SI PPS MVN data with $\alpha = 50$, $n = 10000$

	$\lvert\boldsymbol{\Sigma}\rvert$		$\boldsymbol{\mu}$	
δ	*avg cvg*	*est len*	*avg cvg*	*est vol*
0.2	0.8845	9.7714e-06	0.9542	9.4786e-15
0.5	0.8776	9.7090e-06	0.9532	9.3734e-15
0.8	0.8921	9.7755e-06	0.9502	9.3442e-15
1	0.8949	9.6187e-06	0.9463	9.0667e-15
2	0.9002	9.7131e-06	0.9549	1.0231e-14
3	0.9110	9.7131e-06	0.9534	9.7385e-15
4	0.9151	9.6263e-06	0.9501	9.2920e-15
5	0.9147	9.5988e-06	0.9495	9.5697e-15
10	0.9489	9.4674e-06	0.9532	9.2661e-15
20	0.981	9.3402e-06	0.9496	9.1140e-15
30	0.9822	8.9777e-06	0.951	9.4796e-15
50	0.897	8.4004e-06	0.9466	8.8161e-15
100	0.1373	7.2256e-06	0.9427	8.1619e-15

5 Partially Sensitive Data

In this section we deal with the case when only a part of the data is sensitive and hence needs to be hidden. Assume without any loss of generality that the first r components of $\mathbf{x}$ are sensitive. Two methods of data analysis are suggested: Method I uses sufficient statistics based only on the sensitive part to impute this part while Method II uses the entire data for imputation purposes.

Method I: Using only estimates of sensitive part to impute synthetic data

Plug-In Sampling

Let us now assume, referring to (1), that $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r)$ is sensitive and the rest $(\boldsymbol{x}_{r+1}, \ldots, \boldsymbol{x}_n)$ is not. The sufficient statistics for the sensitive part, assuming $r > p$, is given by

$$\bar{\boldsymbol{x}}_r = \frac{1}{r}\sum_{i=1}^{r} \boldsymbol{x}_i \sim N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{r}\right)$$

$$\boldsymbol{W}^{(r)} = \frac{1}{r-1}\sum_{i=1}^{r}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_r)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_r)' \sim \frac{\mathcal{W}_p(\boldsymbol{\Sigma}, r-1)}{r-1}$$

and the sufficient statistics for the non-sensitive part, assuming $n-r>p$, is given by

$$\bar{\boldsymbol{x}}_{n-r} = \frac{1}{n-r}\sum_{i=r+1}^{n}\boldsymbol{x}_i \sim N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{n-r}\right)$$

$$\boldsymbol{W}^{(n-r)} = \frac{1}{n-r-1}\sum_{i=r+1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{n-r})(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{n-r})' \sim \frac{\mathcal{W}_p(\boldsymbol{\Sigma}, n-r-1)}{n-r-1}$$

We will impute the synthetic counterparts to the sensitive data using only the sufficient statistics for the sensitive part so as to ensure the imputed data is independent of the non-sensitive data. Thus we generate

$$\boldsymbol{y}_1,\ldots,\boldsymbol{y}_r \sim \text{iid} \sim N_p\left(\bar{\boldsymbol{x}}_r, \boldsymbol{W}^{(r)}\right)$$

so that the released data is $(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_r,\boldsymbol{x}_{r+1},\ldots,\boldsymbol{x}_n)$. Since $(\bar{\boldsymbol{x}}_r, \boldsymbol{W}^{(r)}, \bar{\boldsymbol{x}}_{n-r}, \boldsymbol{W}^{(n-r)})$ is the sufficient statistics of the original data, so by Lemma 1 of Klein and Sinha (2016), $(\bar{\boldsymbol{y}}_r, \boldsymbol{W}_y^{(r)}, \bar{\boldsymbol{x}}_{n-r}, \boldsymbol{W}^{(n-r)})$ is the sufficient statistics for the released data where

$$\bar{\boldsymbol{y}}_r = \frac{1}{r}\sum_{i=1}^{r}\boldsymbol{y}_i, \quad \bar{\boldsymbol{y}}_r \mid \bar{\boldsymbol{x}}_r, \boldsymbol{W}^{(r)} \sim N_p\left(\bar{\boldsymbol{x}}_r, \frac{\boldsymbol{W}^{(r)}}{r}\right)$$

$$\boldsymbol{W}_y^{(r)} = \frac{1}{r-1}\sum_{i=1}^{r}(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_r)(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_r)', \quad \boldsymbol{W}_y^{(r)} \mid \boldsymbol{W}^{(r)} \sim \frac{\mathcal{W}_p(\boldsymbol{W}^{(r)}, r-1)}{r-1}$$

We can then compute the likelihood of the released data and multiply it with our regular prior $\pi(\boldsymbol{\mu},\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{\delta}{2}}$ to find the following posterior distributions

$$\begin{aligned}
&\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \bar{\boldsymbol{y}}_r, \boldsymbol{W}^{(r)}, \bar{\boldsymbol{x}}_{n-r} \\
&\sim N_p\left[\left(r\left(\boldsymbol{\Sigma}+\boldsymbol{W}^{(r)}\right)^{-1} + (n-r)\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(r(\boldsymbol{\Sigma}+\boldsymbol{W}^{(r)})^{-1}\bar{\boldsymbol{y}}_r + (n-r)\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}_{n-r}\right),\right. \\
&\qquad \left.\left(r\left(\boldsymbol{\Sigma}+\boldsymbol{W}^{(r)}\right)^{-1} + (n-r)\boldsymbol{\Sigma}^{-1}\right)^{-1}\right]
\end{aligned} \tag{25}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}, \boldsymbol{W}^{(r)} \mid \bar{\boldsymbol{y}}_r, \boldsymbol{W}_y^{(r)}, \bar{\boldsymbol{x}}_{n-r}, \boldsymbol{W}^{(n-r)} \\
&\propto \exp\left[-\frac{r(n-r)}{2}(\bar{\boldsymbol{y}}_r-\bar{\boldsymbol{x}}_{n-r})'(n\boldsymbol{\Sigma}+(n-r)\boldsymbol{W}^{(r)})^{-1}(\bar{\boldsymbol{y}}_r-\bar{\boldsymbol{x}}_{n-r})\right] \\
&\quad \left|n\boldsymbol{\Sigma}+(n-r)\boldsymbol{W}^{(r)}\right|^{-\frac{1}{2}}\left|\boldsymbol{W}^{(r)}\right|^{-\frac{p+1}{2}}|\boldsymbol{\Sigma}|^{-\frac{n+\delta-2}{2}} \\
&\quad \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}((n-r-1)\boldsymbol{W}^{(n-r)}+(r-1)\boldsymbol{W}^{(r)}))\right] \\
&\quad \exp\left[-\frac{r-1}{2}\operatorname{tr}({\boldsymbol{W}^{(r)}}^{-1}\boldsymbol{W}_y^{(r)})\right]
\end{aligned} \tag{26}$$

We point out that the distributions of $\boldsymbol{\Sigma}$ and the latent matrix $\boldsymbol{W}^{(r)}$ are inextricably entangled, which requires further investigation!

Posterior Predictive Sampling

Assuming $r > \max\{p, 2p - \alpha + 1\}$, we generate a posterior draw $(\boldsymbol{\mu}_r^*, \boldsymbol{\Sigma}_r^*)$ from

$$\boldsymbol{\Sigma} \,|\, \boldsymbol{W}^{(r)} \sim \mathcal{W}_p^{-1}\left((r-1)\boldsymbol{W}^{(r)}, r - p + \alpha - 2\right)$$
$$\boldsymbol{\mu} \,|\, \bar{\boldsymbol{x}}_r, \boldsymbol{\Sigma} \sim N_p\left(\bar{\boldsymbol{x}}_r, \frac{\boldsymbol{\Sigma}}{r}\right)$$

so that the released data is $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_r, \boldsymbol{x}_{r+1}, \ldots, \boldsymbol{x}_n)$ where $\boldsymbol{y}_i \sim \text{iid} \sim N_p(\boldsymbol{\mu}_r^*, \boldsymbol{\Sigma}_r^*)$ for $i = 1, \ldots, r$. Thus the sufficient statistics for released data is $(\bar{\boldsymbol{y}}_r, \boldsymbol{W}_y^{(r)}, \bar{\boldsymbol{x}}_{n-r}, \boldsymbol{W}^{(n-r)})$, the quantities are defined as in the preceding page, whose distributions are as follows

$$\bar{\boldsymbol{y}}_r \,|\, \boldsymbol{\mu}_r^*, \boldsymbol{\Sigma}_r^* \sim N_p\left(\boldsymbol{\mu}_r^*, \frac{\boldsymbol{\Sigma}_r^*}{r}\right); \quad \boldsymbol{W}_y^{(r)} \,|\, \boldsymbol{\Sigma}_r^* \sim \frac{\mathcal{W}_p(\boldsymbol{\Sigma}_r^*, r-1)}{r-1}$$
$$\bar{\boldsymbol{x}}_{n-r} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{r}\right); \quad \boldsymbol{W}^{(n-r)} \,|\, \boldsymbol{\Sigma} \sim \frac{\mathcal{W}_p(\boldsymbol{\Sigma}, n-r-1)}{n-r-1}$$

where the last two quantities need the assumption $n - r > p$ for their distributions to be defined. Then the likelihood of the released data is computed as ($h(\cdot)$ below stands for a generic notation for the pdf)

$$\int h(\bar{\boldsymbol{y}}_r \,|\, \boldsymbol{\mu}_r^*, \boldsymbol{\Sigma}_r^*)\, h(\boldsymbol{W}_y^{(r)} \,|\, \boldsymbol{\Sigma}_r^*)\, h(\boldsymbol{\mu}_r^* \,|\, \bar{\boldsymbol{x}}_r, \boldsymbol{\Sigma})\, h(\boldsymbol{\Sigma}_r^* \,|\, \boldsymbol{W}^{(r)})\, h(\bar{\boldsymbol{x}}_r \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, h(\boldsymbol{W}^{(r)} \,|\, \boldsymbol{\Sigma})$$
$$h(\bar{\boldsymbol{x}}_{n-r} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, h(\boldsymbol{W}^{(n-r)} \,|\, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}_r^*\, d\boldsymbol{\Sigma}_r^*\, d\bar{\boldsymbol{x}}_r\, d\boldsymbol{W}^{(r)}$$

We integrate out $\boldsymbol{\mu}_r^*$, $\bar{\boldsymbol{x}}_r$, $\boldsymbol{W}^{(r)}$ one by one from the above likelihood and then multiply with our usual prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{\delta}{2}}$ to obtain the following posterior distributions

$$\begin{aligned}
&\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}_r^*, \bar{\boldsymbol{y}}_r, \bar{\boldsymbol{x}}_{n-r} \\
&\sim N_p\left[\left(r\,(\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_r^*)^{-1} + (n-r)\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(r\,(\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_r^*)^{-1}\,\bar{\boldsymbol{y}}_r + (n-r)\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}_{n-r}\right),\right. \\
&\qquad \left.\left(r\,(\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}_r^*)^{-1} + (n-r)\boldsymbol{\Sigma}^{-1}\right)^{-1}\right]
\end{aligned} \tag{27}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_r^* \,|\, \bar{\boldsymbol{y}}_r, \boldsymbol{W}_y^{(r)}, \bar{\boldsymbol{x}}_{n-r}, \boldsymbol{W}^{(n-r)} \\
&\propto \exp\left[-\frac{r(n-r)}{2}(\bar{\boldsymbol{y}}_r - \bar{\boldsymbol{x}}_{n-r})'(n\boldsymbol{\Sigma} + 2(n-r)\boldsymbol{\Sigma}_r^*)^{-1}(\bar{\boldsymbol{y}}_r - \bar{\boldsymbol{x}}_{n-r})\right] \\
&\quad |n\boldsymbol{\Sigma} + 2(n-r)\boldsymbol{\Sigma}_r^*|^{-\frac{1}{2}}\, |\boldsymbol{\Sigma}_r^*|^{\frac{2r-2p-4}{2}}\, |\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_r^*|^{-\frac{2r-p+\alpha-3}{2}}\, |\boldsymbol{\Sigma}|^{-\frac{n-2r+p-\alpha+\delta+1}{2}} \\
&\quad \exp\left[-\frac{n-r-1}{2}\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}^{(n-r)})\right]\exp\left[-\frac{r-1}{2}\,\mathrm{tr}({\boldsymbol{\Sigma}_r^*}^{-1}\boldsymbol{W}_y^{(r)})\right]
\end{aligned} \tag{28}$$

We can check that the results of this section match the case when all responses are sensitive, those obtained in Sections 2 and 3, by suppressing the quantities $\bar{\boldsymbol{x}}_{n-r}, \boldsymbol{W}^{(n-r)}$; replacing $\boldsymbol{W}_y^{(r)} = \dfrac{\mathscr{S}_y}{n-1}$, $\bar{\boldsymbol{y}}_r = \bar{\boldsymbol{y}}$, $r = n$ and $\boldsymbol{W}_y^{(r)} = \boldsymbol{W}$ (in the PIS case), $\boldsymbol{\Sigma}_r^* = \boldsymbol{\Sigma}^*$ (in the PPS case).

Method II: Using whole data estimates to impute synthetic data

Plug-In Sampling

In this method, the released data is $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_r, \boldsymbol{x}_{r+1}, \ldots, \boldsymbol{x}_n)$ where $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_r \sim \text{iid} \sim N_p(\bar{\boldsymbol{x}}, \boldsymbol{W})$. Then the portion of the likelihood of the released data required to calculate the posterior distributions is computed as

$$\int \left(\prod_{i=1}^{r} h(\boldsymbol{y}_i \,|\, \bar{\boldsymbol{x}}, \boldsymbol{W}) \right) h(\bar{\boldsymbol{x}} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, h(\boldsymbol{W} \,|\, \boldsymbol{\Sigma})\, d\bar{\boldsymbol{x}}\, d\boldsymbol{W}$$

We integrate out $\bar{\boldsymbol{x}}$ from the above likelihood and then multiply with our usual prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{\delta}{2}}$ to obtain the following posterior distributions

$$\begin{aligned}
\boldsymbol{W}_y^{(r)^{-1/2}} \boldsymbol{W} \boldsymbol{W}_y^{(r)^{-1/2}} &\sim \mathcal{W}_p^{-1}\left((r-1)\boldsymbol{I}_p, r-p+\delta-2\right) && (29)\\
\boldsymbol{W}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{W}^{-1/2} &\sim \mathcal{W}_p^{-1}\left((n-1)\boldsymbol{I}_p, n-p+\delta-2\right) && (30)\\
\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \boldsymbol{W}, \bar{\boldsymbol{y}}_r &\sim N_p\left(\bar{\boldsymbol{y}}_r, \frac{\boldsymbol{\Sigma}}{n} + \frac{\boldsymbol{W}}{r}\right) && (31)
\end{aligned}$$

where $\boldsymbol{W}_y^{(r)} = \sum_{i=1}^{r} (\boldsymbol{y}_i - \bar{\boldsymbol{y}}_r)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_r)'$, which is equivalent to $\mathscr{S}_y$ when $r = n$. The distributions are proper as long as $n > p$, $r > 2p - \delta + 1$.

Posterior Predictive Sampling

We draw $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ from (14) so that the released data is $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_r, \boldsymbol{x}_{r+1}, \ldots, \boldsymbol{x}_n)$ where $\boldsymbol{y}_i \sim \text{iid} \sim N_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ for $i = 1, \ldots, r$. Then the portion of the likelihood of the released data required to calculate the posterior distributions is computed as

$$\int \left(\prod_{i=1}^{r} h(\boldsymbol{y}_i \,|\, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \right) h(\boldsymbol{\mu}^* \,|\, \bar{\boldsymbol{x}}, \boldsymbol{\Sigma})\, h(\boldsymbol{\Sigma}^* \,|\, \boldsymbol{W})\, h(\bar{\boldsymbol{x}} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})\, h(\boldsymbol{W} \,|\, \boldsymbol{\Sigma})\, d\boldsymbol{\mu}^*\, d\boldsymbol{\Sigma}^*\, d\bar{\boldsymbol{x}}\, d\boldsymbol{W}$$

We integrate out $\boldsymbol{\mu}^*$, $\bar{\boldsymbol{x}}$, $\boldsymbol{W}$ one by one from the above likelihood and then multiply with our usual prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{\delta}{2}}$ to obtain the following posterior distributions

$$\begin{aligned}
\mathscr{S}_{y,r}^{-1/2} \boldsymbol{\Sigma}^* \mathscr{S}_{y,r}^{-1/2} &\sim \mathcal{W}_p^{-1}\left(\boldsymbol{I}_p, r-2p+\delta-1\right) && (32)\\
\boldsymbol{\Sigma}^{*-1/2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{*-1/2} &\sim \mathrm{B}_{\mathrm{p}}^{\mathrm{II}}\left(\frac{n+\alpha-\delta-1}{2}, \frac{n-p+\delta-2}{2}\right) && (33)\\
\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*, \bar{\boldsymbol{y}}_r &\sim N_p\left(\bar{\boldsymbol{y}}_r, \frac{1}{n}\left(\boldsymbol{\Sigma} + \left(1 + \frac{n}{r}\right)\boldsymbol{\Sigma}^*\right)\right) && (34)
\end{aligned}$$

where $\mathscr{S}_{y,r} = \sum_{i=1}^{r}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_r)(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_r)'$, which is equivalent to $\mathscr{S}_y$ when $r = n$. The conditions for existence are $r > 3p - \delta$, $n > \max\{p, 2p - \alpha + 1, p - \alpha + \delta, 2p - \delta + 1\}$. All the conditions for existence throughout this work can also be expressed as inequalities for δ, since once we have the data at hand, that would enable us to choose a proper value of δ to get the best inference.

Appendix A

Proof of Theorem 2

Under the setup of Section 2, Klein and Sinha (2015a) proved the following result (Theorem 2.1) in their paper.

Theorem 2.1

The joint pdf of $\bar{z}$ and $\mathscr{S}_z$ is obtained by integrating out Σ^* from the joint pdf of $(\bar{z}, \mathscr{S}_z, \Sigma^*)$ given by

$$f(\bar{z}, \mathscr{S}_z, \Sigma^*) \propto \exp\left[-\frac{1}{2}(n(\bar{z}-\mu)'(\Sigma + 2\Sigma^*)^{-1}(\bar{z}-\mu) + \operatorname{tr}\mathscr{S}_z(\Sigma^*)^{-1})\right]$$
$$\times |\Sigma^{-1} + \frac{1}{2}(\Sigma^*)^{-1}|^{-\frac{1}{2}} \times |\Sigma|^{-\frac{n}{2}} \times |\Sigma^*|^{-\frac{2n+\alpha-1}{2}} \times |\Sigma^{-1} + (\Sigma^*)^{-1}|^{-\frac{2n+\alpha-p-3}{2}} \times |\mathscr{S}_z|^{\frac{n-p-2}{2}}.$$

The equivalence of our Theorem 2 with their Theorem 2.1 follows from the following observations. Write $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{D_\delta}\boldsymbol{\Gamma'}$ and $\boldsymbol{\Sigma}^* = \boldsymbol{\Gamma}\boldsymbol{D_\lambda}\boldsymbol{\Gamma'}$ where $\boldsymbol{\Gamma}$ is non-singular and D_δ and D_λ are diagonal matrices. We show that

$$|\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^*|^{-\frac{1}{2}}\,|\boldsymbol{\Sigma}|^{\frac{n-p+\alpha-2}{2}}\,|\boldsymbol{\Sigma} + \boldsymbol{\Sigma}^*|^{-\frac{2n-p+\alpha-3}{2}}\,|\boldsymbol{\Sigma}^*|^{-(\frac{p+1}{2}+\alpha)} \approx \tag{35}$$
$$\left|\boldsymbol{\Sigma}^{-1} + \frac{1}{2}(\boldsymbol{\Sigma}^*)^{-1}\right|^{-\frac{1}{2}}\,|\boldsymbol{\Sigma}|^{-\frac{n}{2}}\,|\boldsymbol{\Sigma}^*|^{-(\frac{2n+\alpha-1}{2})}\,\left|\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^*)^{-1}\right|^{-(\frac{2n+\alpha-p}{2})}$$

A direct computation leads to LHS of equation (1) as

$$\left[|\boldsymbol{\Gamma\Gamma'}|^{-(\frac{n+p+1}{2})}\right]\Big\{\prod_{i=1}^{p}(\delta_i + 2\lambda_i)\Big\}^{-\frac{1}{2}}\Big\{\prod_{i=1}^{p}\delta_i\Big\}^{-\frac{n-p+\alpha-2}{2}}\Big\{\prod_{i=1}^{p}\lambda_i\Big\}^{-(\frac{p+1}{2}+\alpha)}\Big\{\prod_{i=1}^{p}\delta_i + \lambda_i\Big\}^{-(\frac{2n-p+\alpha-3}{2})} \tag{36}$$

Likewise, the RHS of equation (35) simplifies to

$$\left[|\boldsymbol{\Gamma\Gamma'}|^{-(\frac{n+p+1}{2})}\right]\Big\{\prod_{i=1}^{p}(\frac{2\lambda_i + \delta_i}{\delta_i\lambda_i})\Big\}^{-\frac{1}{2}}\Big\{\prod_{i=1}^{p}\delta_i\Big\}^{-\frac{n}{2}}\Big\{\prod_{i=1}^{p}\lambda_i\Big\}^{-(\frac{2n+\alpha-1}{2})}\Big\{\prod_{i=1}^{p}\frac{\lambda_i + \delta_i}{\lambda_i\delta_i}\Big\}^{-(\frac{2n+\alpha-p-3}{2})}2^{p/2} \tag{37}$$

The proof follows upon simplifying equation (37) and comparing to (36).

Acknowledgment

Abhishek Guin is thankful to the Department of Mathematics and Statistics, UMBC, for providing financial support for his graduate studies. Anindya Roy and Bimal Sinha are thankful to Dr. Tommy Wright, Chief of CSRM/Census Bureau for his support and encouragement. Our sincere thanks are due to Professor Gaurisankar Datta of the University of Georgia for his critical reading of the manuscript and making some excellent suggestions which led to improved presentation.

References

[1] Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York.

[2] Klein, M., and Sinha, B. (2013). Statistical analysis of noise-multiplied data using multiple imputation. Journal of Official Statistics, 29(3), 425-465.

[3] Klein, M. and Sinha, B. (2015a). Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models. Sankhya B, 77(2), 293-311.

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

[5] Klein, M. D., Zylstra, J., and Sinha, B. K. (2019). Finite sample inference for multiply imputed synthetic data under a multiple linear regression model. Calcutta Statistical Association Bulletin, 71(2), 63-82.

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

[7] Muirhead, R. J. (1982). Aspects of Multivariate Statistical Theory. Wiley, New York.