Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue in Memory of Prof. C R Rao
Volume 24, No. 3, 2024 (New Series), pp 119–152
http://www.ssca.org.in/journal

Confidence Ellipsoids of a Multivariate Normal Mean Vector Based on Noise Perturbed and Synthetic Data with Applications

Biswajit Basak[1], Yehenew G. Kifle[2] and Bimal K. Sinha[2,3]
[1]*Department of Statistics, Sister Nivedita University, Kolkata 700156, India*
[2]*Department of Mathematics and Statistics*
University of Maryland Baltimore County, Maryland 21250, USA
[3]*Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Rd, Suitland-Silver Hill, MD 20746, USA*

Received: 19 April 2024; Revised: 09 June 2024; Accepted: 11 June 2024

Abstract

In this paper we address the problem of constructing a confidence ellipsoid of a multivariate normal mean vector based on a random sample from it. The central issue at hand is the sensitivity of the original data and hence the data cannot be directly used/analyzed. We consider a few perturbations of the original data, namely, noise addition and creation of synthetic data based on the plug-in sampling (PIS) method and the posterior predictive sampling (PPS) method. We review some theoretical results under PIS and PPS which are already available based on both frequentist and Bayesian analysis (Klein and Sinha, 2015, 2016; Guin *et al.*, 2023) and derive the necessary results under noise addition. A theoretical comparison of all the methods based on expected volumes of the confidence ellipsoids is provided. A measure of privacy protection (PP) is discussed and its formulas under PIS, PPS and noise addition are derived and the different methods are compared based on PP. Applications include analysis of two multivariate datasets. The first dataset, with $p = 2$, is obtained from the latest Annual Social and Economic Supplement (ASEC) conducted by the US Census Bureau in 2023. The second dataset, with $p = 3$, pertains to renal variables obtained from the book by Harris and Boyd (1995). Using a synthetic version of the original data generated through PIS and PPS methods and also the noise added data, we produce and display the confidence ellipsoids for the unknown mean vector under various scenarios. Finally, the privacy protection measure is evaluated for various methods and different features.

Key words: Bayesian credible Set; Confidence ellipsoid; Noise addition; Plug-in sampling; Posterior predictive sampling; Privacy protection.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Bimal K. Sinha
Email: sinha@umbc.edu

1. Introduction

Statistical data analysis under privacy protection has been the focus of statistical research at many government agencies where the charge is to collect public data or information on many aspects of their lives and then analyze and disseminate the information for public use, policy decisions, and further research by other interested parties. Often the information provided by the public as in the decennial census in the USA contain some sensitive features and it is the responsibility of the data collecting agency to ensure that information related to these features are not compromised, properly hidden, and hard to retrieve from subsequent data analysis and released tables. Statistical research dealing with this task falls into the category of Statistical Disclosure Control (SDC) Methods. Fortunately, many novel methods of SDC have been developed and used over the years, notably noise addition/multiplication, swapping, and synthetic data creation (Drechsler and Reiter (2010); Drechsler (2011); Kinney *et al.* (2014); Kinney *et al.* (2011); Kinney *et al.* (2014); Lin and Wise (2012); Little *et al.* (1993); Meng (1994); Klein *et al.* (2014); Klein and Sinha (2013a); Klein and Sinha (2015); Klein and Sinha (2016); Raghunathan *et al.* (2003); Reiter (2003); Reiter (2004); Reiter (2005a); Reiter (2005b); Reiter (2005c); Reiter and Kinney (2012); Reiter and Mitra (2009); Reiter and Raghunathan (2007); Rubin (1987); Rubin (1993); Rubin (1996); Nayak *et al.* (2011); Sinha *et al.* (2011);Klein and Sinha (2013b)). There are three distinct parts in this process: how to perturb or distort the sensitive parts of the information collected, how to carry out proper statistical analysis based on the perturbed data so as to draw valid inference about some population features (like proportions, means, variances, correlation) and a study of the extent to which privacy has been preserved!

The focus of this paper is on multivariate data analysis in the context of sensitive data collected on p continuous features from a random sample of n units of a population. We assume that data follows a multivariate normal (MVN) model with the mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$, both unknown, and primarily address the problem of constructing confidence sets (CS) for $\boldsymbol{\mu}$ based on suitable perturbations of the original data. Three methods of SDC are considered: noise addition, synthetic data analysis based on Plug-in Sampling (PIS) scheme and synthetic data analysis based on Posterior Predictive Sampling (PPS) scheme. In each case we clearly spell out 1) how to create artificial data, 2) how to analyze it so as to produce a valid CS for $\boldsymbol{\mu}$, and 3) to what extent privacy is protected based on a suitable privacy protection (PP) measure. We should point out that the above methods are widely used in the literature and we have freely used some results which are already available and derived necessary additional results for a complete analysis of the MVN data.

The organization of the paper is as follows. In Section 2 we discuss valid inference based on noise added data, including proper analysis leading to a CS for $\boldsymbol{\mu}$. Section 3 is devoted to valid analysis of synthetic data under PIS and Section 4 to valid analysis under PPS. Both Sections 3 and 4 reside in the frequentist paradigm. We consider Bayesian analysis of PIS and PPS data in Section 5. A comparison of the suggested methods based on the expected volumes is done in Section 6. In Section 7 a measure of privacy protection (PP) suitable for multivariate data is given and explicit expressions of this measure for all the methods are derived. A comparison of the suggested artificial data generation methods based on PP is also given. It should be noted that evaluation of PP depends only on the way the original data are perturbed and not on subsequent data analysis methods. Finally, in Section 8, we apply all the proposed methods in the analysis of two multivariate datasets:

the first, with $p = 2$, is obtained from the US Census Bureau, and the second dataset, encompassing renal variables with $p = 3$, is obtained from the book by Harris and Boyd (1995), providing a comprehensive analysis of both datasets.

Throughout this paper we assume the original data $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ are *iid* as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $n > p$ and $\boldsymbol{\Sigma}$ is a positive definite matrix. Define $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ (sample mean), $\boldsymbol{W}_{\mathbf{x}} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ (sample Wishart matrix) and $\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{W}_{\mathbf{x}}}{n-1}$. Based on the original data, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Define $T^2_{\mathbf{x}} = n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{W}^{-1}_{\mathbf{x}}(\bar{\mathbf{x}} - \boldsymbol{\mu})$, then $\left(\frac{n-p}{p}\right) T^2_{\mathbf{x}} \sim F_{p,n-p}$. A $(1-\gamma)$ level confidence ellipsoid (CE) for $\boldsymbol{\mu}$ based on the original data $\boldsymbol{X}$ will be

$$\Delta(\boldsymbol{\mu}) = \left\{ \boldsymbol{\mu} : T^2_{\mathbf{x}} \leq \left(\frac{p}{n-p}\right) F_{p,n-p;\gamma} \right\}, \tag{1}$$

where $F_{p,n-p;\gamma}$ is the $100(1-\gamma)^{\text{th}}$ percentile of an F-distribution with $(p, n-p)$ degrees of freedoms. The *observed volume* and the *expected volume* of the above CE will be

$$V_{\boldsymbol{\mu}}(\boldsymbol{X}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}} \tag{2}$$

$$E\left[V_{\boldsymbol{\mu}}(\boldsymbol{X})\right] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma}\right)^{p/2} \mathscr{C}_{n,p} |\boldsymbol{\Sigma}|^{\frac{1}{2}}, \tag{3}$$

where $E[|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}] = \mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ and $\mathscr{C}_{n,p} = \prod_{i=1}^p \left[2^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)}\right]$.

2. Inference based on noise added data

In this section our objective is to propose an inferential method of finding a suitable confidence set for the unknown $\boldsymbol{\mu}$ based on the noise added data. The original data $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ are assumed to be independent and identically distributed (*iid*) as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $n > p$. Based on these data, one can define the summary statistics $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ (sample mean) and $\boldsymbol{W}_{\mathbf{x}} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ (sample Wishart matrix). Sometimes the unit level/micro data are available and sometimes they are not. We have encountered these two cases in the following subsections.

2.1. Case 1: Unit level data available

When unit level data are available, they can be perturbed by adding some random noise $\mathbf{e}_i \sim N_p(\mathbf{0}, \boldsymbol{R})$, *iid* for $i = 1, \cdots, n$, to the i^{th} level - resulting in $\mathbf{u}_i = \mathbf{x}_i + \mathbf{e}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{R}), i = 1, 2, \ldots, n$, where $\boldsymbol{R}$ is a known positive definite noise dispersion matrix. Our objective is to propose an inferential method of finding a suitable confidence set for the unknown $\boldsymbol{\mu}$ based on the noise added data $\boldsymbol{U} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$. Define $\bar{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i$ and $\boldsymbol{W}_{\mathbf{u}} = \sum_{i=1}^n (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})'$. It is very easy to verify that, based on the noise added data $\boldsymbol{U}$, $(\bar{\mathbf{u}}, \boldsymbol{W}_{\mathbf{u}})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Clearly $\bar{\mathbf{u}} \sim N_p\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}+\boldsymbol{R}}{n}\right)$, independently of $\boldsymbol{W}_{\mathbf{u}} \sim \text{Wishart}_p(\boldsymbol{\Sigma} + \boldsymbol{R}, n-1)$. We define $T^2_{\mathbf{u}} = n(\bar{\mathbf{u}} - \boldsymbol{\mu})'\boldsymbol{W}^{-1}_{\mathbf{u}}(\bar{\mathbf{u}} - \boldsymbol{\mu})$ which follows $\frac{p}{n-p}$ times an F-distribution with degrees of freedoms $(p, n-p)$. Clearly, $T^2_{\mathbf{u}}$ can be looked upon

as a pivot and can be used to find a $(1-\gamma)$ ellipsoid for $\boldsymbol{\mu}$ as given by

$$\Delta^1_{NA}(\boldsymbol{\mu}) = \left\{ \boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{u}})' \boldsymbol{W}_{\mathbf{u}}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{u}}) \leq \frac{p}{n-p} F_{p,n-p;\gamma} \right\}, \tag{4}$$

where $F_{p,n-p;\gamma}$ is the $100(1-\gamma)^{\text{th}}$ percentile of an $F_{p,n-p}$ distribution. The *volume* of the confidence ellipsoid $\Delta^1_{NA}(\boldsymbol{\mu})$ based on the noise added data $\boldsymbol{U}$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{U}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma} \right)^{p/2} |\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}. \tag{5}$$

Note that $E(|\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}) = \mathscr{C}_{n,p} |\boldsymbol{\Sigma} + \boldsymbol{R}|^{1/2}$ with $\mathscr{C}_{n,p} = \prod_{i=1}^{p} \left[2^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)} \right]$, the *expected volume* is obtained as

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{U})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p} F_{p,n-p;\gamma} \right)^{p/2} \mathscr{C}_{n,p} |\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}. \tag{6}$$

2.2. Case 2: Unit level data not available

If unit level/micro data is not available on $\boldsymbol{X}$, but only summary statistics $\bar{\mathbf{x}}$ and $\boldsymbol{W}_{\mathbf{x}}$ are available, we define $\bar{\mathbf{u}} = \bar{\mathbf{x}} + \bar{\mathbf{e}}$, where $\bar{\mathbf{e}} \sim N_p(\mathbf{0}, \frac{\boldsymbol{R}}{n})$, independent of $\bar{\mathbf{x}}$, and $\boldsymbol{W}_{\mathbf{u}} = \boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r$, where $\boldsymbol{W}_r \sim \text{Wishart}_p(r, \boldsymbol{R})$ with $r \geq p$, independent of $\boldsymbol{W}$. Consequently we have $\bar{\mathbf{u}} \sim N_p(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}+\boldsymbol{R}}{n})$ and $\boldsymbol{W}_{\mathbf{u}}$ follows a distribution which is the sum of two independent Wishart distributions: $\text{Wishart}_p(n-1, \boldsymbol{\Sigma})$ and $\text{Wishart}_p(r, \boldsymbol{R})$. For the sake of simplicity, we write it as $\boldsymbol{W}_{\mathbf{u}} \sim \boldsymbol{W}_p(n-1, \boldsymbol{\Sigma}) + \boldsymbol{W}_p(r, \boldsymbol{R})$. Define $F_{\mathbf{u}} = n(\bar{\mathbf{u}} - \boldsymbol{\mu})' \boldsymbol{W}_{\mathbf{u}}^{-1} (\bar{\mathbf{u}} - \boldsymbol{\mu})$. Here, it should be noted that the distribution of $F_{\mathbf{u}}$ is not independent of the parameter $\boldsymbol{\Sigma}$ and hence can not be used as a pivot. Our goal is to find F^* which is stochastically larger than $F_{\mathbf{u}}$ and which has a distribution free from the parameter.

Consider $\mathbf{v} = \sqrt{n}(\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}(\bar{\mathbf{u}}-\boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$, that is $\sqrt{n}(\bar{\mathbf{u}}-\boldsymbol{\mu}) = (\boldsymbol{\Sigma}+\boldsymbol{R})^{\frac{1}{2}}\mathbf{v}$, and $\boldsymbol{W}^*_{\mathbf{u}} = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{W}_{\mathbf{u}} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}$, we can rewrite $F_{\mathbf{u}} = \mathbf{v}'(\boldsymbol{W}^*_{\mathbf{u}})^{-1}\mathbf{v}$. Note that, $\boldsymbol{W}^*_{\mathbf{u}} \sim \boldsymbol{W}_p(n-1, \boldsymbol{A}_1) + \boldsymbol{W}_p(r, \boldsymbol{A}_2)$, where $\boldsymbol{A}_1 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}$ and $\boldsymbol{A}_2 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{R} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}}$ with $\boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{I}_p$.

Theorem 1: If $\mathbf{w}_1 \sim \text{Wishart}_p(n-1, \boldsymbol{I}_p)$, independently of $\mathbf{w}_2 \sim \text{Wishart}_p(r, \boldsymbol{I}_p)$, the distribution of $F^* = Max\left\{ \mathbf{v}'\mathbf{w}_1^{-1}\mathbf{v}, \mathbf{v}'\mathbf{w}_2^{-1}\mathbf{v} \right\}$ is stochastically larger than $F_{\mathbf{u}}$ and also free from the parameter.

Proof: Suppose $\boldsymbol{S}_1 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{W} (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \sim \text{Wishart}_p(n-1, \boldsymbol{A}_1)$, independently of $\boldsymbol{S}_2 = (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \boldsymbol{W}_r (\boldsymbol{\Sigma}+\boldsymbol{R})^{-\frac{1}{2}} \sim \text{Wishart}_p(r, \boldsymbol{A}_2)$, and $\boldsymbol{S} = \boldsymbol{S}_1 + \boldsymbol{S}_2$, then $F_{\mathbf{u}}$ can be written as $F_{\mathbf{u}} = \mathbf{v}' \boldsymbol{S}^{-1} \mathbf{v}$. We first note that,

$$F_{\mathbf{u}} = \mathbf{v}' \boldsymbol{S}^{-1} \mathbf{v} = \underset{\mathbf{l}:\mathbf{l}'\mathbf{l}=1}{Max} \left\{ \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\boldsymbol{S}\mathbf{l}} \right\}. \tag{*}$$

Let $\mathbf{w}_1 \sim \text{Wishart}_p(n-1, \boldsymbol{I}_p)$, independently of $\mathbf{w}_2 \sim \text{Wishart}_p(r, \boldsymbol{I}_p)$. For any $\mathbf{l}$ such that $\mathbf{l}'\mathbf{l} = 1$, $\mathbf{l}'\mathbf{w}_1\mathbf{l} \sim \chi^2_{n-1}$, and $\mathbf{l}'\mathbf{w}_2\mathbf{l} \sim \chi^2_r$. Again $\frac{\mathbf{l}'\boldsymbol{S}_1\mathbf{l}}{\mathbf{l}'\boldsymbol{A}_1\mathbf{l}} \sim \chi^2_{n-1}$ and $\frac{\mathbf{l}'\boldsymbol{S}_1\mathbf{l}}{\mathbf{l}'\boldsymbol{A}_2\mathbf{l}} \sim \chi^2_r$, which implies

$$\begin{aligned} \mathbf{l}'\boldsymbol{S}_1\mathbf{l} &\overset{d}{=} (\mathbf{l}'\boldsymbol{A}_1\mathbf{l})(\mathbf{l}'\mathbf{w}_1\mathbf{l}), \\ \text{and } \mathbf{l}'\boldsymbol{S}_2\mathbf{l} &\overset{d}{=} (\mathbf{l}'\boldsymbol{A}_2\mathbf{l})(\mathbf{l}'\mathbf{w}_2\mathbf{l}). \end{aligned}$$

Hence

$$\begin{aligned} \mathbf{l}'\boldsymbol{S}\mathbf{l} = \mathbf{l}'(\boldsymbol{S}_1+\boldsymbol{S}_2)\mathbf{l} &\overset{d}{=} (\mathbf{l}'\boldsymbol{A}_1\mathbf{l})(\mathbf{l}'\mathbf{w}_1\mathbf{l}) + (\mathbf{l}'\boldsymbol{A}_2\mathbf{l})(\mathbf{l}'\mathbf{w}_2\mathbf{l}) \\ &\underset{st}{\geq} (\mathbf{l}'\boldsymbol{A}_1\mathbf{l} + \mathbf{l}'\boldsymbol{A}_2\mathbf{l}) Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\} \\ &= (\mathbf{l}'\mathbf{l}) Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\}, \quad [\text{Since, } \boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{I}_p] \\ &= Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\}, \qquad [\text{Since, } \mathbf{l}'\mathbf{l} = 1] \end{aligned}$$

Thus we have

$$\begin{aligned} \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\boldsymbol{S}\mathbf{l}} &\underset{st}{\leq} \frac{(\mathbf{l}'\mathbf{v})^2}{Min\left\{\mathbf{l}'\mathbf{w}_1\mathbf{l}, \mathbf{l}'\mathbf{w}_2\mathbf{l}\right\}} \\ &\overset{d}{=} Max\left\{\frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_1\mathbf{l}}, \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_2\mathbf{l}}\right\}. \end{aligned}$$

From (*),

$$\begin{aligned} F_u &= \underset{\mathbf{l}:\mathbf{l}'\mathbf{l}=1}{Max}\left\{\frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\boldsymbol{S}\mathbf{l}}\right\} \\ &\underset{st}{\leq} \underset{\mathbf{l}:\mathbf{l}'\mathbf{l}=1}{Max}\left\{Max\left\{\frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_1\mathbf{l}}, \frac{(\mathbf{l}'\mathbf{v})^2}{\mathbf{l}'\mathbf{w}_2\mathbf{l}}\right\}\right\} \\ &\overset{d}{=} Max\left\{\mathbf{v}'\mathbf{w}_1^{-1}\mathbf{v}, \mathbf{v}'\mathbf{w}_2^{-1}\mathbf{v}\right\} \\ &= F^* \end{aligned}$$

Clearly the distribution of F^* is free from $\boldsymbol{\Sigma}$, as all of $\mathbf{v}$, $\mathbf{w}_1$ and $\mathbf{w}_2$ are having distributions free from $\boldsymbol{\Sigma}$.

[Note: Here we have used the symbols $\overset{d}{=}$, $\underset{st}{\leq}$ and $\underset{st}{\geq}$, which stands for identically distributed, stochastically smaller and stochastically larger respectively]. □

We determine $F^*_{n,p,r,\gamma}$ such that $P[F^* \leq F^*_{n,p,r,\gamma}] = 1-\gamma$, which implies $P[F_{\mathbf{u}} \leq F^*_{n,p,r,\gamma}] \geq P[F^* \leq F^*_{n,p,r,\gamma}] = 1-\gamma$. Therefore a confidence ellipsoid for $\boldsymbol{\mu}$ with confidence level at least $(1-\gamma)$ is given by

$$\Delta^2_{NA}(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{u}})' \boldsymbol{W}^{-1}_{\mathbf{u}} (\boldsymbol{\mu} - \bar{\mathbf{u}}) \leq F^*_{n,p,r,\gamma}\right\}. \tag{7}$$

The *volume* of the confidence ellipsoid $\Delta^2_{NA}(\boldsymbol{\mu})$ is given by

$$V^*_{\boldsymbol{\mu}} = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(F^*_{n,p,r,\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}. \tag{8}$$

The *expected volume* can be computed by evaluating $E\left[|\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}\right]$. Recall $\boldsymbol{W}_{\mathbf{u}} = \boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r$, we can use the result, $|\boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r|^{\frac{1}{2}} > Max\left\{|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}, |\boldsymbol{W}_r|^{\frac{1}{2}}\right\}$ with probability 1, resulting in $E\left[|\boldsymbol{W}_{\mathbf{x}} + \boldsymbol{W}_r|^{\frac{1}{2}}\right] > E\left[Max\left\{|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}, |\boldsymbol{W}_r|^{\frac{1}{2}}\right\}\right] \geq Max\left\{E\left[|\boldsymbol{W}_{\mathbf{x}}|^{\frac{1}{2}}\right], E\left[|\boldsymbol{W}_r|^{\frac{1}{2}}\right]\right\}$. Therefore a lower bound to the *expected volume* will be

$$\begin{aligned} E[V^*_{\boldsymbol{\mu}}] &\geq \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F^*_{n,p,r,\gamma}\right)^{p/2} Max\left\{\mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}, \mathscr{C}_{r+1,p}|\boldsymbol{R}|^{\frac{1}{2}}\right\} \\ &\approx \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F^*_{n,p,r,\gamma}\right)^{p/2} \mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \qquad (9) \\ &\quad \text{[Assuming } |\boldsymbol{R}| \text{ to be significantly small.]} \end{aligned}$$

Remark 1: We can do a direct comparison of the expected volume in (6) when unit level data are available and the lower bound of the expected volume in (9) when unit level data are not available in situations when R is *small*. This essentially boils down to a comparison of $[p/(n-p)]F_{p,n-p;\gamma}$ and $F^*_{n,p,r,\gamma}$. However, from the definition of F^* it follows that any percentile of F^* is larger than the corresponding percentile of $\mathbf{v}'\mathbf{w}_1^{-1}\mathbf{v}$. Since the latter percentile is $[p/(n-p)]F_{p,n-p;\gamma}$, it readily follows that $F^*_{n,p,r,\gamma}$ is larger than $[p/(n-p)]F_{p,n-p;\gamma}$, regardless of r. In other words, even the lower bound for the expected volume given in (9) is larger than the exact expected volume in (6), whatever be the df r. Table 1 shows a direct comparison of these two cut-off points.

Table 1: The first table presents $F^*_{n,p,r,\gamma}$ cut-off points for various combinations of n, p and r, while the second table displays the $[p/(n-p)]F_{p,n-p;\gamma}$ cut-off points across different values of n and p, with $\gamma = 0.05$. significance level.

n	r=10			r=15			r=20			r=100		
	p=2	p=3	p=4	p=2	p=3	p=4	p=2	p=3	p=4	p=2	p=3	p=4
25	0.921	1.504	2.379	0.534	0.797	1.070	0.394	0.568	0.739	0.307	0.419	0.539
50	0.969	1.518	2.402	0.532	0.795	1.092	0.382	0.523	0.693	0.134	0.179	0.219
100	0.976	1.531	2.437	0.544	0.820	1.097	0.363	0.522	0.679	0.069	0.090	0.109

$[p/(n-p)]F_{p,n-p;\gamma}$ cut-off points			
n	p=2	p=3	p=4
25	0.298	0.416	0.541
50	0.133	0.179	0.224
100	0.063	0.083	0.103

3. Analysis of synthetic data under plug-in sampling

In this section we briefly review the method of analysing synthetic data obtained under plug-in sampling method, which are derived by (Klein and Sinha, 2016). The main objective here is to obtain a confidence ellipsoid for $\boldsymbol{\mu}$, based on the synthetic data, for a given confidence level.

Under plug-in sampling method, singly imputed synthetic data, denoted by $\boldsymbol{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, are obtained by drawing *iid* observations from $N_p(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. Based on these

synthetic data, $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$ and $\boldsymbol{W}_{\mathbf{y}} = \sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (See (Klein and Sinha, 2016)). Clearly, given the original data $\boldsymbol{X}$, $\bar{\mathbf{y}} \sim N_p(\bar{\mathbf{x}}, n^{-1}\hat{\boldsymbol{\Sigma}})$ independently of $\boldsymbol{W}_{\mathbf{y}} \sim \text{Wishart}_p(\hat{\boldsymbol{\Sigma}}, n-1)$. The joint pdf (unconditional) of $(\bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}})$ is given by

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}}) \propto \int_{\hat{\boldsymbol{\Sigma}} \in S_n^{++}} \frac{|\boldsymbol{W}_{\mathbf{y}}|^{\frac{n-p-2}{2}} |\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}}|^{-\frac{1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n-1}{2}} |\hat{\boldsymbol{\Sigma}}|^{\frac{p+1}{2}}} e^{-\frac{1}{2}\left[n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\left(\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}\right)^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}) + Tr(\boldsymbol{W}_{\mathbf{y}}\hat{\boldsymbol{\Sigma}}^{-1}) + (n-1)Tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1})\right]} d\hat{\boldsymbol{\Sigma}},$$

where S_n^{++} stands for the set of $p \times p$ positive definite matrices. For the derivation of the above expression we refer to (Klein and Sinha, 2016).

Based on the synthetic data $\boldsymbol{Y}$, consider $T^2_{\mathbf{y}} = n(\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{W}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$, which has a mixture-type distribution mentioned in the following theorem which is derived by (Klein and Sinha, 2016). The theorem also shows that $T^2_{\mathbf{y}}$ is a pivotal quantity and can be used to find a confidence ellipsoid for $\boldsymbol{\mu}$.

Theorem 2: The distribution of $T^2_{\mathbf{y}} = n(\bar{\mathbf{y}} - \boldsymbol{\mu})' \boldsymbol{W}_{\mathbf{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ has the representation: $T^2_{\mathbf{y}} = T_{\mathbf{y}1} \times T_{\mathbf{y}2}$ where $T_{\mathbf{y}1} \sim \frac{1}{\chi^2_{n-p}}$, independent of $T_{\mathbf{y}2}$, and the conditional distribution of $T_{\mathbf{y}2}$, given a Wishart matrix $\boldsymbol{W}^*$, is $\sum_{i=1}^{p}\lambda_i\chi^2_{1i}$ where χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom and $\lambda_1, \ldots, \lambda_p$ are the roots of $|(n-1)\boldsymbol{I}_p + (1-\lambda)\boldsymbol{W}^*| = 0$ where $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$.

Theorem 2 shows that $T^2_{\mathbf{y}}$ can be used as a pivotal quantity, and hence we can construct a $(1-\gamma)$ ellipsoid for $\boldsymbol{\mu}$ based on $T^2_{\mathbf{y}}$ as given by

$$\Delta_1(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{y}})' \boldsymbol{W}_{\mathbf{y}}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{y}}) \leq a_{n,p,\gamma}\} \tag{10}$$

where $a_{n,p,\gamma}$ is the $(1-\gamma)$ percentile from the distribution of $T^2_{\mathbf{y}}$. The cut-off point $a_{n,p,\gamma}$ can be obtained by simulating from the distribution of $T^2_{\mathbf{y}}$ as given below:

1. Generate $\lambda_1, \lambda_2, \ldots, \lambda_p$, the roots of $|(n-1)\boldsymbol{I}_p + (1-\lambda)\boldsymbol{W}^*| = 0$ where $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$.
2. Generate $T_{\mathbf{y}2} = \sum_{i=1}^{p}\lambda_i\chi^2_{1i}$ where χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom.
3. Generate $T_{\mathbf{y}1} \sim \frac{1}{\chi^2_{n-p}}$, independent of $T_{\mathbf{y}2}$.
4. Finally compute $T^2_{\mathbf{y}} = T_{\mathbf{y}1} \times T_{\mathbf{y}2}$.

The *volume* of the confidence ellipsoid $\Delta_1(\boldsymbol{\mu})$ based on the synthetic data $\boldsymbol{Y}$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Y}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}|\boldsymbol{W}_{\mathbf{y}}|^{\frac{1}{2}}. \tag{11}$$

Since $E\left(|\boldsymbol{W}_{\mathbf{y}}|^{\frac{1}{2}}\right) = \frac{\mathscr{C}_{n,p}^2}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ with $\mathscr{C}_{n,p} = \prod_{i=1}^{p}\left[2^{\frac{1}{2}}\frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)}\right]$, the *expected volume* is obtained as

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{Y})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}\frac{\mathscr{C}_{n,p}^2}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \tag{12}$$

4. Analysis of synthetic data under posterior predictive sampling

Likewise in the previous section, the original data $\boldsymbol{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ is assumed to be *iid* as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $n > p$. In this section we briefly discuss a method, illustrated in ((Klein and Sinha, 2015)), to obtain confidence ellipsoid for $\boldsymbol{\mu}$ based on a synthetically generated data under posterior predictive sampling. Consider $\bar{\mathbf{x}}$ and $\boldsymbol{W}_{\mathbf{x}}$, as mentioned in the section (1), which are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Under the posterior predictive sampling method, a vague prior for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is set as $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\alpha}{2}}$, where $n+\alpha > 2p+3$. The joint posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given $\boldsymbol{X}$, can be represented as

$$\begin{aligned} \boldsymbol{\Sigma}^{-1}|\boldsymbol{X} &\sim \text{Wishart}_p(\boldsymbol{W}_{\mathbf{x}}^{-1}, n+\alpha-p-2) \\ \boldsymbol{\mu}|(\boldsymbol{\Sigma}, \boldsymbol{X}) &\sim N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{\Sigma}}{n}\right). \end{aligned} \tag{13}$$

We draw $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ from the above posterior and finally a random sample $\boldsymbol{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$ is drawn from $N_p(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, which constitutes the synthetic data. Based on these synthetic data $\boldsymbol{Z}$, one can easily verify that $\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i$ and $\boldsymbol{W}_{\mathbf{z}} = \sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The joint pdf of $\bar{\mathbf{z}}$ and $\boldsymbol{W}_{\mathbf{z}}$ is obtained by integrating out $\boldsymbol{\Sigma}^*$ from the joint pdf of $(\bar{\mathbf{z}}, \boldsymbol{W}_{\mathbf{z}}, \boldsymbol{\Sigma}^*)$ given by

$$\begin{aligned} f(\bar{\mathbf{z}}, \boldsymbol{W}_{\mathbf{z}}, \boldsymbol{\Sigma}^*) \propto e^{-\frac{1}{2}\left[n(\bar{\mathbf{z}}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*)^{-1}(\bar{\mathbf{z}}-\boldsymbol{\mu})+Tr\left(\boldsymbol{W}_{\mathbf{z}}\boldsymbol{\Sigma}^{*-1}\right)\right]}|\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*|^{-\frac{1}{2}}|\boldsymbol{\Sigma}|^{\frac{n-p+\alpha-2}{2}} \\ |\boldsymbol{\Sigma}+\boldsymbol{\Sigma}^*|^{-\frac{2n-p+\alpha-3}{2}}|\boldsymbol{\Sigma}^*|^{-\left(\frac{p+1}{2}+\alpha\right)}|\boldsymbol{W}_{\mathbf{z}}|^{\frac{n-p-2}{2}}. \end{aligned}$$

Define $T_{\mathbf{z}}^2 = n(\bar{\mathbf{z}}-\boldsymbol{\mu})'\boldsymbol{W}_{\mathbf{z}}^{-1}(\bar{\mathbf{z}}-\boldsymbol{\mu})$, then the distribution of $T_{\mathbf{z}}^2$, as mentioned in (Klein and Sinha, 2015), given in Theorem (3) below.

Theorem 3: $T_{\mathbf{z}}^2$ has the representation: $T_{\mathbf{z}}^2 = T_{\mathbf{z}1} \times T_{\mathbf{z}2}$ with $T_{\mathbf{z}1} \sim \frac{1}{\chi_{n-p}^2}$, independent of $T_{\mathbf{z}2} = \sum_{i=1}^{n}\lambda_i\chi_{1i}^2$ where χ_{1i}^2 are independent χ^2 random variables each with 1 degree of freedom and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the roots of $|\boldsymbol{I}_p + (2-\lambda)\tilde{\boldsymbol{\Sigma}}| = 0$, and the distribution of $\tilde{\boldsymbol{\Sigma}}$ is given by

$$f(\tilde{\boldsymbol{\Sigma}}) \propto |\tilde{\boldsymbol{\Sigma}}|^{\frac{n-p-2}{2}} \times |I+\tilde{\boldsymbol{\Sigma}}|^{-\frac{2n+\alpha-p-3}{2}}.$$

From the above theorem it is clear that $T_{\mathbf{z}}^2$ can be used as a pivot and hence a $(1-\gamma)$ level confidence ellipsoid for $\boldsymbol{\mu}$ based on $T_{\mathbf{z}}^2$ is given by

$$\Delta_2(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : n(\boldsymbol{\mu}-\bar{\mathbf{z}})'\boldsymbol{W}_{\mathbf{z}}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{z}}) \leq b_{n,p,\alpha,\gamma}\}, \tag{14}$$

where $b_{n,p,\alpha,\gamma}$ is the $(1-\gamma)$ level cut-off point from the distribution of $T_{\mathbf{z}}^2$ and it can be obtained by simulating from the distribution of $T_{\mathbf{z}}^2$ as discussed below.

1. To generate $\tilde{\boldsymbol{\Sigma}}$ having the density $f(\tilde{\boldsymbol{\Sigma}})$ as defined in Theorem (3), one can generate $A_1 \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$ independent of $A_2 \sim \text{Wishart}_p(\boldsymbol{I}_p, n+\alpha-p-2)$, and set $\tilde{\boldsymbol{\Sigma}} = A_1^{\frac{1}{2}} A_2^{-1} A_1^{\frac{1}{2}}$. The proof of this representation of $\tilde{\boldsymbol{\Sigma}}$ appears in the proof of Theorem 8.2.8 of (Muirhead, 1982).
2. Obtain the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ as $\delta_1, \delta_2, \ldots, \delta_p$ and take $\lambda_i = 2 + \frac{1}{\delta_i}$, $i = 1, \ldots, p$.
3. Generate $T_{\mathbf{z}2} = \sum_{i=1}^{p} \lambda_i \chi_{1i}^2$ where χ_{1i}^2 are independent χ^2 variables each with 1 degree of freedom.
4. Generate $T_{\mathbf{z}1} \sim \frac{1}{\chi_{n-p}^2}$, independent of $T_{\mathbf{z}2}$.
5. Finally compute $T_{\mathbf{z}}^2 = T_{\mathbf{z}1} \times T_{\mathbf{z}2}$.

The *volume* of the confidence ellipsoid $\Delta_2(\boldsymbol{\mu})$ based on the synthetic data $\boldsymbol{Z}$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Z}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} (b_{n,p,\alpha,\gamma})^{p/2} |\boldsymbol{W}_{\mathbf{z}}|^{\frac{1}{2}}, \tag{15}$$

therefore the *expected volume* is

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{Z})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} (b_{n,p,\alpha,\gamma})^{p/2} \mathscr{D}_{n,p}^2 \mathscr{E}_{n,p,\alpha} |\boldsymbol{\Sigma}|^{\frac{1}{2}}, \tag{16}$$

where $\mathscr{D}_{n,p} = \prod_{i=1}^{p} \left[\frac{\sqrt{2}\Gamma(\frac{n-i+1}{2})}{\Gamma(\frac{n-i}{2})}\right]$ and $\mathscr{E}_{n,p,\alpha} = \prod_{i=1}^{p} \left[\frac{\Gamma(\frac{n+\alpha-p-i-2}{2})}{\sqrt{2}\Gamma(\frac{n+\alpha-p-i-1}{2})}\right]$.

5. Bayesian analysis of PIS and PPS data

In this section, which is essentially based on Guin *et al.* (2023), we discuss the Bayesian credible confidence ellipsoids (BCCE) for the mean vector $\boldsymbol{\mu}$ and their (frequentist) expected volumes under PIS and PPS.

5.1. BCCE under PIS

Referring to the likelihood function of the released data $\mathbf{y}_1, \cdots, \mathbf{y}_n$ under PIS mentioned in Section 3, we now apply a diffuse prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\delta}{2}}$. This results in the posterior joint distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which can be represented in the following manner:

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}} | \boldsymbol{W}_{\mathbf{y}}, \bar{\mathbf{y}} &\sim \text{Wishart}_p^{-1}(\boldsymbol{W}_{\mathbf{y}}, n-p+\delta-2) \\
\boldsymbol{\Sigma} | \hat{\boldsymbol{\Sigma}}, \bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}} &\sim \text{Wishart}_p^{-1}\left((n-1)\hat{\boldsymbol{\Sigma}}, n-p+\delta-2\right) \\
\boldsymbol{\mu} | \boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}, \bar{\mathbf{y}}, \boldsymbol{W}_{\mathbf{y}} &\sim N_p\left(\bar{\mathbf{y}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}\right)\right)
\end{aligned} \tag{17}$$

The above can be further reformulated as:

$$\begin{aligned}
\boldsymbol{W}_{\mathbf{y}}^{-1/2} \hat{\boldsymbol{\Sigma}} \boldsymbol{W}_{\mathbf{y}}^{-1/2} &\sim \text{Wishart}_p^{-1}(\boldsymbol{I}_p, n-p+\delta-2) \\
\hat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1/2} &\sim \text{Wishart}_p^{-1}((n-1)\boldsymbol{I}_p, n-p+\delta-2) \\
\boldsymbol{\mu} \,|\, \boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}, \bar{\mathbf{y}} &\sim N_p\left(\bar{\mathbf{y}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+\hat{\boldsymbol{\Sigma}}\right)\right)
\end{aligned} \tag{18}$$

which has the benefit that $\boldsymbol{W}_{\mathbf{y}}{}^{-1/2}\hat{\boldsymbol{\Sigma}}\,\boldsymbol{W}_{\mathbf{y}}{}^{-1/2}$ is independent of $\hat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1/2}$ and their posterior distributions are unconditional. The posterior distributions are proper as long as $n > \max\{p, 2p-\delta+1\}$. A $(1-\gamma)$ BCCE for $\boldsymbol{\mu}$ can be taken as [(Guin *et al.*, 2023)]

$$\Delta_3(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} \;:\; T^2_{\mathbf{y}} \le c_{n,p,\delta;\gamma}\right\}, \tag{19}$$

where $T^2_{\mathbf{y}} = n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\,\boldsymbol{W}^{-1}_{\mathbf{y}}(\bar{\mathbf{y}}-\boldsymbol{\mu})$ and the cut-off point $c_{n,p,\delta;\gamma}$ is obtained by simulation through the following steps:

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

The observed and expected volumes of the above BCCE under PIS are readily obtained as

$$V^B_{\boldsymbol{\mu}}(\boldsymbol{Y}) \;=\; \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}/n\right)^{p/2}\,|\,\boldsymbol{W}_{\mathbf{y}}|^{1/2} \tag{20}$$

$$E[V^B_{\boldsymbol{\mu}}(\boldsymbol{Y})] \;=\; \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}\right)^{p/2}\frac{\mathscr{C}^2_{n,p}}{(n-1)^{p/2}}\,|\boldsymbol{\Sigma}|^{1/2}, \tag{21}$$

where $\mathscr{C}_{n,p} = \prod_{i=1}^p\left[2^{1/2}\Gamma(\frac{n-i+1}{2})/\Gamma(\frac{n-i}{2})\right]$.

5.2. BCCE under PPS

Referring to the likelihood function of the released data $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ under PPS mentioned in Section 4, we now apply a diffuse prior $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{\delta}{2}}$. This results in the posterior joint distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which can be represented in the following form:

$$\begin{aligned}
\boldsymbol{\Sigma}^*\,|\,\boldsymbol{W}_{\mathbf{z}} &\sim \text{Wishart}_p^{-1}\left(\boldsymbol{W}_{\mathbf{z}}, n-2p+\delta-1+2\alpha\right)\\
\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} &\sim \text{B}_{\text{p}}^{\text{II}}\left(\frac{n+\alpha-\delta-1}{2}, \frac{n-p+\delta-2}{2}\right)\\
\boldsymbol{\mu}\,|\,\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*, \bar{\mathbf{z}} &\sim N_p\left(\bar{\mathbf{z}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*\right)\right)
\end{aligned} \tag{22}$$

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

$$\begin{aligned}
\boldsymbol{W}_{\mathbf{z}}^{-1/2}\boldsymbol{\Sigma}^*\boldsymbol{W}_{\mathbf{z}}^{-1/2} &\sim \text{Wishart}_p^{-1}\left(\boldsymbol{I}_p, n-2p+\delta-1+2\alpha\right)\\
\boldsymbol{\Sigma}^{*-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{*-1/2} &\sim \text{B}_{\text{p}}^{\text{II}}\left(\frac{n+\alpha-\delta-1}{2}, \frac{n-p+\delta-2}{2}\right)\\
\boldsymbol{\mu}\,|\,\boldsymbol{\Sigma}, \boldsymbol{\Sigma}^*, \bar{\mathbf{z}} &\sim N_p\left(\bar{\mathbf{z}}, \frac{1}{n}\left(\boldsymbol{\Sigma}+2\boldsymbol{\Sigma}^*\right)\right)
\end{aligned} \tag{23}$$

which has the benefit that $\boldsymbol{W}_{\mathbf{z}}^{-1/2}\boldsymbol{\Sigma}^{*}\boldsymbol{W}_{\mathbf{z}}^{-1/2}$ 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 $n > \max\{p, 2p-\alpha+1, 3p-\delta, p-\alpha+\delta, 2p-\delta+1-2\alpha\}$.

A BCCE for $\boldsymbol{\mu}$ can be taken as [(Guin *et al.*, 2023)]

$$\Delta_4(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : T_{\mathbf{z}}^2 \leq d_{n,p,\alpha,\delta,\gamma}\}, \tag{24}$$

where $T_{\mathbf{z}}^2 = n(\boldsymbol{\mu}-\bar{\mathbf{z}})'\,\boldsymbol{W}_{\mathbf{z}}^{-1}(\boldsymbol{\mu}-\bar{\mathbf{z}})$ and the cut-off point $d_{n,p,\alpha,\delta,\gamma}$ is obtained by simulation through the following steps:

1. Generate $\boldsymbol{B} \sim \text{Wishart}_p^{-1}(\boldsymbol{I}_p, n-2p+\delta+2\alpha-1)$ and decompose as $\boldsymbol{B} = \boldsymbol{D}\boldsymbol{D}'$.
2. Generate $\boldsymbol{V}_0 \sim \text{Wishart}_p(\boldsymbol{I}_p, n-p+\delta-2)$, $\boldsymbol{V}_1 \sim \text{Wishart}_p(\boldsymbol{I}_p, n+\alpha-\delta-1)$, $\boldsymbol{C} = \boldsymbol{V}_0^{-1}\boldsymbol{V}_1\boldsymbol{V}_0^{-1}$ and $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{C}\boldsymbol{D}' + 2\boldsymbol{B}$ (Gupta and Nagar, 1999).
3. Generate $\lambda_1, \ldots, \lambda_p$, the roots of $|\boldsymbol{A} - \lambda\boldsymbol{I}_p| = 0$.
4. Generate $T_{\mathbf{z}}^2 = \sum_{i=1}^{p} \lambda_i \chi_{1i}^2$ where χ_{1i}^2 are independent χ_1^2 variables.

The observed and expected volumes of the above BCCE under PPS are readily obtained as

$$V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{z}}|^{1/2}, \tag{25}$$

$$E[V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta,\gamma}\right)^{p/2} \mathscr{D}_{n,p}^{2}\mathscr{E}_{n,p,\alpha} \times |\boldsymbol{\Sigma}|^{1/2}, \tag{26}$$

where $\mathscr{D}_{n,p} = \prod_{i=1}^{p}\left[\frac{\sqrt{2}\Gamma(\frac{n-i+1}{2})}{\Gamma(\frac{n-i}{2})}\right]$ and $\mathscr{E}_{n,p,\alpha} = \prod_{i=1}^{p}\left[\frac{\Gamma(\frac{n+\alpha-p-i-2}{2})}{\sqrt{2}\Gamma(\frac{n+\alpha-p-i-1}{2})}\right]$.

6. Comparison of the suggested methods based on the expected volumes

6.1. Expressions of observed and expected volumes

In this subsection, we provide a brief overview of various expressions for *observed* and *expected* volumes for $\boldsymbol{\mu}$ within the noise-added data context and also both frequentist and Bayesian frameworks under PIS and PPS methods.

The observed and expected volumes of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 4), derived from noise added data $\boldsymbol{U}$, when unit level data are available, are given below.

$$V_{\boldsymbol{\mu}}(\boldsymbol{U}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p}F_{p,n-p;\gamma}\right)^{p/2} |\,\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}},$$

$$E[V_{\boldsymbol{\mu}}(\boldsymbol{U})] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(\frac{p}{n-p}F_{p,n-p;\gamma}\right)^{p/2} \mathscr{C}_{n,p}|\boldsymbol{\Sigma}+\boldsymbol{R}|^{\frac{1}{2}}. \tag{27}$$

where $\mathscr{C}_{n,p} = \prod_{i=1}^{p}\left[2^{\frac{1}{2}}\frac{\Gamma\left(\frac{n-i+1}{2}\right)}{\Gamma\left(\frac{n-i}{2}\right)}\right]$.

If unit level data are not available, the observed volume and a lower bound to the expected volume of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 7) are given by,

$$\begin{aligned} V_{\boldsymbol{\mu}}^{*} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F_{n,p,r,\gamma}^{*}\right)^{p/2}|\,\boldsymbol{W}_{\mathbf{u}}|^{\frac{1}{2}}, \\ E[V_{\boldsymbol{\mu}}^{*}] &\geq \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F_{n,p,r,\gamma}^{*}\right)^{p/2} Max\left\{\mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}, \mathscr{C}_{r+1,p}|\boldsymbol{R}|^{\frac{1}{2}}\right\} \\ &\geq \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(F_{n,p,r,\gamma}^{*}\right)^{p/2}\mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \qquad (28) \\ &\quad [\text{Assuming } |\boldsymbol{R}| \text{ to be significantly small}] \end{aligned}$$

Below are the observed and expected volumes of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 10), derived from synthetic data $\boldsymbol{Y}$ using the PIS method.

$$\begin{aligned} V_{\boldsymbol{\mu}}(\boldsymbol{Y})_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}|\,\boldsymbol{W}_{\mathbf{y}}|^{\frac{1}{2}}, \\ E[V_{\boldsymbol{\mu}}(\boldsymbol{Y})]_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(a_{n,p,\gamma})^{p/2}\frac{\mathscr{C}_{n,p}^{2}}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \qquad (29) \end{aligned}$$

Likewise, the observed and expected volumes of the confidence ellipsoid for $\boldsymbol{\mu}$ (see 14), utilizing synthetic data $\boldsymbol{Z}$ under the PPS method, are presented below.

$$\begin{aligned} V_{\boldsymbol{\mu}}(\boldsymbol{Z})_{PPS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(b_{n,p,\alpha,\gamma})^{p/2}|\,\boldsymbol{W}_{\mathbf{z}}|^{\frac{1}{2}}, \\ E[V_{\boldsymbol{\mu}}(\boldsymbol{Z})]_{PPS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(b_{n,p,\alpha,\gamma})^{p/2}\mathscr{D}_{n,p}^{2}\mathscr{E}_{n,p,\alpha}|\boldsymbol{\Sigma}|^{\frac{1}{2}}, \qquad (30) \end{aligned}$$

where $\mathscr{D}_{n,p} = \prod_{i=1}^{p}\left[\frac{\sqrt{2}\Gamma(\frac{n-i+1}{2})}{\Gamma(\frac{n-i}{2})}\right]$ and $\mathscr{E}_{n,p,\alpha} = \prod_{i=1}^{p}\left[\frac{\Gamma(\frac{n+\alpha-p-i-2}{2})}{\sqrt{2}\Gamma(\frac{n+\alpha-p-i-1}{2})}\right]$.

In the Bayesian framework, we provide below the observed and expected volumes of credible confidence ellipsoids for $\boldsymbol{\mu}$ within the context of synthetic data generated using the PIS method (see 19),

$$\begin{aligned} V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Y})_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}\right)^{p/2}|\boldsymbol{W}_{\mathbf{y}}|^{1/2}, \\ E[V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Y})]_{PIS} &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}\left(c_{n,p,\delta;\gamma}\right)^{p/2}\frac{\mathscr{C}_{n,p}^{2}}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{1/2}, \qquad (31) \end{aligned}$$

Table 2: Coefficients of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression under various perturbation schemes ($\gamma = 0.05$).

DIFFERENT SCHEMES	n	p		
		2	3	4
NA DATA ($r = 100$)	25	0.8619	0.9690	1.1219
	50	0.4061	0.2999	0.2231
	100	0.2390	0.1279	0.0683
PIS	25	1.7928	2.8911	5.0991
	50	0.8181	0.8554	0.9147
	100	0.3949	0.2770	0.2017
PPS ($\alpha = 4$)	25	2.2241	5.9769	14.3850
	50	1.2394	1.6682	2.3476
	100	0.5877	0.5294	0.4796
PIS BAYES ($\delta = 10$)	25	1.1445	1.6016	2.3651
	50	0.6672	0.6517	0.6477
	100	0.3572	0.2468	0.1717
PPS BAYES ($\alpha = 1$, $\delta = 10$)	25	1.457	2.4228	5.1696
	50	0.7489	0.7758	0.9171
	100	0.3773	0.2768	0.2032

and under PPS method (see 24),

$$\begin{aligned} V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z}) &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta;\gamma}\right)^{p/2} |\boldsymbol{W}_{\mathbf{z}}|^{1/2}, \\ E[V_{\boldsymbol{\mu}}^{B}(\boldsymbol{Z})] &= \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)} \left(d_{n,p,\alpha,\delta,\gamma}\right)^{p/2} \mathscr{D}_{n,p}^{2} \mathscr{E}_{n,p,\alpha} \times |\boldsymbol{\Sigma}|^{1/2}. \end{aligned} \tag{32}$$

6.2. Comparison of expected volumes - all are proportional to $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$

Note that the expected volume expressions presented in equations 29, 30, 31 and 32 for various methods are directly proportional to $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$. The coefficient of $|\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}$ in the equation 27 is the same as that of the expected volume under the original data, hence it is immaterial to consider it for the comparison. Rather we compare the expected volume under noise added data when unit level data are not available. We assume $|\boldsymbol{R}|$ to be small enough and calculate the coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in (28). Consequently, a straightforward comparison of these methods can be made by examining the coefficients of the expected volume expressions, without considering the population parameter $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$. In Table (2), we present the coefficients obtained from various perturbation schemes in different combinations of n and p values. Specifically, we used n values of 25, 50, and 100, and p values of 2, 3, and 4. The parameters $\alpha = 4$ and $\delta = 10$ remain fixed in the frequentist approach, while in the Bayesian framework we used $\alpha = 1$ and $\delta = 10$. Additionally, for data with added noise, we set $r = 100$. Throughout the analysis, we maintain a consistent value of $\gamma = 0.05$.

From Table (2), it is clear that the expected volume decreases as the sample size (n) increases under any schemes, which is quite natural. Also, in all the choices of the pair (n, p), we can

see that the expected volumes under the noise added data (taking $r = 100$) are quite smaller than the other schemes. As anticipated, in both the frequentist and Bayesian frameworks, the expected volumes under PPS exceed those under PIS.

Remark 2: Referring to Remark 1, it is obvious that in case unit level data are available, the expected volume then will be the least among all reported above. Therefore, if one were to make practical recommendations based on the expected volumes only, gathering unit level data and subsequent noise addition will certainly pay off, followed by the same noise addition mechanism based on summary data.

7. Measure of privacy protection

Disclosure risk evaluation

When the original (unit level) microdata is considered to be sensitive and thus hidden through the use of a masked version, it is natural to examine the extent to which sensitivity of a data point has been protected. A slight variation of a popular privacy measure to study the disclosure risk of a single scalar value x_i, given in Klein and Sinha (2016), can be taken as

$$P[|\hat{x}_i - x_i| < \epsilon | X] = \theta_i \tag{33}$$

where X is the entire original data, and $\hat{x}_i$ is an intruder's prediction of x_i based upon seeing the released (artificial/synthetic) data, ϵ be any small positive quantity. Naturally, a high value of the above probability indicates a low level of protection and vice versa. This privacy measure (PM) is computed based on the random mechanism producing the masked data, given the original data X.

In the multivariate case, a generalization of (33) can be taken as

$$\theta_i = P[(\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_i)^t A (\hat{\mathbf{x}}_\mathbf{i} - \mathbf{x}_i) \leq \epsilon | X] \tag{34}$$

where A is a positive definite symmetric matrix.

Returning to our specific problem, based on the synthetic multivariate data released by the data producer, a naive intruder's best guess about $\mathbf{x}_i$, the original value for the ith unit, can be discussed under two circumstances: (a) the identities of the perturbed data are released by the data producer and $\mathbf{u}_i, \mathbf{y}_i$ or $\mathbf{z}_i$, the perturbed value of $\mathbf{x}_i$ based on NA/PIS/PPS, corresponding to the identifiable ith unit, is taken as intruder's choice, and (b) the identities of the perturbed data are lost/retained by the data producer in which case $\bar{\mathbf{u}} = [\sum_{i=1}^n \mathbf{u}_i]/n$, $\bar{\mathbf{y}} = [\sum_{i=1}^n \mathbf{y}_i]/n$ or $\bar{\mathbf{z}} = [\sum_{i=1}^n \mathbf{z}_i]/n$ is taken as intruder's choice.

There is also a 3rd case in the multivariate data context in which an intruder may be interested in a particular component, say component 1, of the p vector multivariate data. If original data $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are available, intruder's obvious choice is $\bar{x}_1 = (x_{11} + ... + x_{n1})/n$ where we write $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, \cdots, n$. In the absence of the original data, we can take $\bar{u}_1$, $\bar{y}_1$ and $\bar{z}_1$ as intruder's choice under NA/PIS/PPS, respectively.

In subsection 7.1 we discuss PP under noise added data, in subsection 7.2 we discuss PP under PIS and PPS is taken up in subsection 7.3.

All the above methods discussed in Sections (7.1), (7.2) and (7.3), are from a naive intruder's

perspective. However, a smart intruder with an excellent training in statistics can think in a different way. We have added a remark to this effect at the end of this section.

7.1. Three cases under noise added (NA) data

7.1.1. Case (a)

Here we assume that the identities of the released perturbed data are known and hence the intruder's best choice of $\mathbf{x}_i$ will be $\mathbf{u}_i$. Recall that $\mathbf{u}_i = \mathbf{x}_i + \mathbf{e}_i$ where $\mathbf{e}_i \sim N_p(\mathbf{0}, \boldsymbol{R})$ is independent of the original data $\boldsymbol{X}$. Note that $\mathbf{e}_i^* = \boldsymbol{R}^{-\frac{1}{2}}\mathbf{e}_i \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$. Define $\boldsymbol{B} = \boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}$, which is a symmetric positive definite matrix, there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{B} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}$, where $\boldsymbol{\Lambda} = Diag(\lambda_1, \cdots, \lambda_p)$ be a diagonal matrix with diagonal elements λ_i's $(i = 1, \cdots, p)$, which are the solutions to the equation $|\boldsymbol{B} - \lambda \boldsymbol{I}_p|$. Considering $\mathbf{m}_i = \boldsymbol{\Gamma}\mathbf{e}_i^* \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$ we can deduce the privacy measure (θ_i) corresponding to the i^{th} unit as given by

$$\begin{aligned}
\theta_i &= P\left[(\mathbf{u}_i - \mathbf{x}_i)'\boldsymbol{A}(\mathbf{u}_i - \mathbf{x}_i) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\mathbf{e}_i'\boldsymbol{A}\mathbf{e}_i \leq \epsilon\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}(\mathbf{e}_i^*) \leq \epsilon\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{B}(\mathbf{e}_i^*) \leq \epsilon\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}(\mathbf{e}_i^*) \leq \epsilon\right] \\
&= P\left[\mathbf{m}_i'\boldsymbol{\Lambda}\mathbf{m}_i \leq \epsilon\right] \\
&= P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \leq \epsilon\right] \qquad (35)
\end{aligned}$$

In the above expression χ_{1j}^2, $j = 1, \cdots, p$ are independent central chi square variables each with 1 d.f. Note that the quantity $\theta_i = P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \leq \epsilon\right] = \theta^*$ is independent of any specific unit i and hence it can be taken as a measure of overall privacy protection. The following are two special cases based on the choice of matrix $\boldsymbol{A}$.

Case 1: $\boldsymbol{A} = \boldsymbol{I}_p \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\boldsymbol{R} - \lambda \boldsymbol{I}_p| = 0$.

Case 2: $\boldsymbol{A} = Diag(a_{11}, \cdots, a_{pp}) \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\boldsymbol{R} - \lambda Diag(\frac{1}{a_{11}}, \cdots, \frac{1}{a_{pp}})| = 0$.

7.1.2. Case (b)

When the identities of the released perturbed data are not known, the intruder's best choice of $\mathbf{x}_i$ $(i = 1, \cdots, n)$ will be $\bar{\mathbf{u}}$. Note that $\bar{\mathbf{u}} - \mathbf{x}_i = \bar{\mathbf{e}} - (\mathbf{x}_i - \bar{\mathbf{x}})$, and for conditionally given $\boldsymbol{X}$, it follows $N_p\left(\mathbf{x}_i - \bar{\mathbf{x}}, \frac{\boldsymbol{R}}{n}\right)$. Define $\mathbf{e}_i^* = \sqrt{n}\boldsymbol{R}^{-\frac{1}{2}}(\bar{\mathbf{u}} - \mathbf{x}_i)$, which implies $\mathbf{e}_i^* | \boldsymbol{X} \sim N_p(\boldsymbol{\delta}_i, \boldsymbol{I}_p)$, where $\boldsymbol{\delta}_i = \sqrt{n}\boldsymbol{R}^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}})$. Here we take $\boldsymbol{B} = \frac{\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}}{n}$, which is a symmetric and positive definite matrix, there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{B} = \boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}$, where $\boldsymbol{\Lambda} = Diag(\lambda_1, \cdots, \lambda_p)$ be a diagonal matrix with diagonal elements λ_j's $(j = 1, \cdots, p)$, which are the solutions to the equation $|\boldsymbol{B} - \lambda \boldsymbol{I}_p|$. Likewise **Case (a)**, we define $\mathbf{m}_i = \boldsymbol{\Gamma}\mathbf{e}_i^*$, which conditionally for given $\boldsymbol{X}$, follows $N_p(\boldsymbol{\eta}_i, \boldsymbol{I}_p)$, where $\boldsymbol{\eta}_i = \boldsymbol{\Gamma}\boldsymbol{\delta}_i$. We

proceed in a similar fashion as mentioned in **Case (a)** and deduce the privacy measure (θ_i) corresponding to the i^{th} unit as

$$\begin{aligned}
\theta_i &= P\left[(\bar{\mathbf{u}} - \mathbf{x}_i)'\boldsymbol{A}(\bar{\mathbf{u}} - \mathbf{x}_i) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[(\mathbf{e}_i^*)'\frac{\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{R}^{\frac{1}{2}}}{n}(\mathbf{e}_i^*) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{B}(\mathbf{e}_i^*) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[(\mathbf{e}_i^*)'\boldsymbol{\Gamma}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}(\mathbf{e}_i^*) \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\mathbf{m}_i'\boldsymbol{\Lambda}\mathbf{m}_i \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2(\eta_{ij}^2) \leq \epsilon\right], \qquad (36)
\end{aligned}$$

where $\chi_{1j}^2(\eta_{ij}^2)$, $j = 1, \cdots, p$ are independent noncentral chi-squared variables each with 1 d.f. and noncentrality parameters η_{ij}^2, which is the squared j^{th} component ($j = 1, \cdots, p$) of $\boldsymbol{\eta}_i$.

Unlike **Case (a)**, here θ_i depends on the specific unit i through the noncentrality parameters η_{ij}'s. We can write,

$$\theta_i \leq P[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \leq \epsilon] = \theta^* \text{ (say)}.$$

The quantity θ^* is independent of i and can be taken as a measure of overall privacy measure. Two special choices of $\boldsymbol{A}$ as similar to **Case (a)** are given below.

Case 1: $\boldsymbol{A} = \boldsymbol{I}_p \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{R}}{n} - \lambda \boldsymbol{I}_p| = 0$.

Case 2: $\boldsymbol{A} = Diag(a_{11}, \cdots, a_{pp}) \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{R}}{n} - \lambda Diag(\frac{1}{a_{11}}, \cdots, \frac{1}{a_{pp}})| = 0$.

7.1.3. Case (c)

When an intruder is interested in a particular component, say component 1, of the p-component vector multivariate data, based on the original data $\mathbf{x}_1, \cdots, \mathbf{x}_n$, intruder's obvious choice is $\bar{x}_1 = (x_{11} + ... + x_{n1})/n$ where we write $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, \cdots, n$. In the absence of the original data, the best choice would be $\bar{u}_1 = \frac{1}{n}\sum_{i=1}^{n} u_{i1}$. Clearly, $\bar{u}_1 - \bar{x}_1 = \bar{e}_1$, independently of the original data $\boldsymbol{X}$, follows $N(0, \frac{r_{11}}{n})$, where r_{11} be the $(1,1)^{\text{th}}$ element of $\boldsymbol{R}$. Therefore the privacy measure (θ) is given by

$$\begin{aligned}
\theta &= P\left[(\bar{u}_1 - \bar{x}_1)^2 \leq \epsilon | \boldsymbol{X}\right] \\
&= P\left[\bar{e}_1^2 \leq \epsilon\right] \\
&= P\left[\chi_1^2 \leq \frac{n\epsilon}{r_{11}}\right] \qquad (37)
\end{aligned}$$

From the above, it readily follows that, more the variability in a particular noise component, more the privacy protection for the same component.

7.2. Three cases under PIS

7.2.1. Case (a)

Since the identities of the released masked data are known, the intruder's choice of $\mathbf{x}_i$ can be taken as $\mathbf{y_i}$, which conditionally given $\boldsymbol{X}$, is $N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n-1}\right)$ according to the PIS scheme. It is interesting to observe that $\mathbf{y}_i$ has no bearing with the index i as far as the PIS scheme is concerned.

Before we compute the PM θ in Case (a), let us look at Case (b).

7.2.2. Case (b)

Since in the absence of the identity of the i^{th} unit $\bar{\mathbf{y}}$ seems to be the intruder's obvious choice of $\mathbf{x}_i$, to compute the PM θ, we proceed as follows. Recall that

$$\theta = P\left[(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})^t \boldsymbol{A} (\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i}) \leq \epsilon | \boldsymbol{X}\right]. \tag{38}$$

Note that under PIS, $\bar{\mathbf{y}}_\mathbf{1}, \cdots, \bar{\mathbf{y}}_\mathbf{n}$ are iid following $N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n-1}\right)$, implying $\bar{\mathbf{y}}|\boldsymbol{X} \sim N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n(n-1)}\right)$. Define $\boldsymbol{D} = \frac{\boldsymbol{W_x}}{n(n-1)}$, we have $(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})|\boldsymbol{X} \sim N_p\left((\bar{\mathbf{x}} - \bar{\mathbf{x}}_\mathbf{i}), \boldsymbol{D}\right)$, which implies $\boldsymbol{D}^{-1/2}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})|\boldsymbol{X} \sim N_p\left(\boldsymbol{D}^{-1/2}(\bar{\mathbf{x}} - \mathbf{x_i}), \boldsymbol{I}_p\right)$. Write $\boldsymbol{Z} = \boldsymbol{D}^{-1/2}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_\mathbf{i})$, then $\theta_i = P[\boldsymbol{Z}^t \boldsymbol{D}^{1/2} \boldsymbol{A} \boldsymbol{D}^{1/2} \boldsymbol{Z} \leq \epsilon | \boldsymbol{X}] = P[\mathbf{Z}^t \boldsymbol{B} \boldsymbol{Z} \leq \epsilon | \boldsymbol{X}]$, where $\boldsymbol{D}^{1/2} \boldsymbol{A} \boldsymbol{D}^{1/2} = \boldsymbol{B}$: $p \times p$ symmetric pd and $\boldsymbol{Z}|\boldsymbol{X} \sim N_p(\boldsymbol{\delta}_i, \boldsymbol{I}_p)$ with $\boldsymbol{\delta}_i = \boldsymbol{D}^{-1/2}(\bar{\mathbf{x}} - \mathbf{x_i})$.

Since $\boldsymbol{B}$ is symmetric pd, there exists an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Gamma}^t \boldsymbol{\Lambda} \boldsymbol{\Gamma} = \boldsymbol{B}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with elements $\lambda_1, \cdots, \lambda_p$ as the characteristic roots of $\boldsymbol{B}$. Let $\boldsymbol{U} = \boldsymbol{\Gamma} \boldsymbol{Z} \sim N_p[\boldsymbol{\eta}_i, \boldsymbol{I}_p]$, where $\boldsymbol{\eta}_i = \Gamma \boldsymbol{\delta}_i$. Then

$$\begin{aligned} \theta_i &= P[\boldsymbol{Z}^t \boldsymbol{\Gamma}^t \boldsymbol{\Lambda} \boldsymbol{Z} \boldsymbol{\Gamma} \leq \epsilon] \\ &= P[\boldsymbol{U}^t \boldsymbol{\Lambda} \boldsymbol{U} \leq \epsilon] \\ &= P[\sum_{j=1}^{p} \lambda_j \chi^2_{1j}(\eta^2_{ij}) \leq \epsilon]. \end{aligned} \tag{39}$$

Note that the roots of $\boldsymbol{B}$ are the solutions of $|\boldsymbol{B} - \lambda \boldsymbol{I}_p| = 0 \iff |\boldsymbol{D}^{1/2} \boldsymbol{A} \boldsymbol{D}^{1/2} - \lambda \boldsymbol{I}_p| = 0 \iff |\boldsymbol{A} - \lambda \boldsymbol{D}^{-1} \boldsymbol{I}_p| = 0 \iff |\boldsymbol{A} - \lambda n(n-1) \mathbf{S}^{-\mathbf{1}}_\mathbf{x}| = \mathbf{0}$. Moreover, $\chi^2_{1j}(\eta^2_{ij})$, $j = 1, \cdots, p$ are independent noncentral chisquare variables each with 1 d.f. and noncentrality parameters as appear above.

For any specific unit i, θ_i above can be taken as a privacy measure. Obviously, for any i

$$\theta_i \leq P[\sum_{j=1}^{p} \lambda_j \chi^2_{1j} \leq \epsilon] = \theta^* \text{ (say)}. \tag{40}$$

We can take the absolute quantity θ^*, which is independent of any specific unit i, as a measure of overall privacy protection. In the above, $\chi^2_{11}, \cdots, \chi^2_{1p}$ are iid central chi-square each with 1 d.f. Here are two special cases:

Case 1: $\boldsymbol{A} = \boldsymbol{I}_p \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{W_x}}{n(n-1)} - \lambda \boldsymbol{I}_p| = 0$.

Case 2: $\boldsymbol{A} = Diag(a_{11}, \cdots, a_{pp}) \Rightarrow \lambda_1, \cdots, \lambda_p$ are the solutions of $|\frac{\boldsymbol{W_x}}{n(n-1)} - \lambda Diag(\frac{1}{a_{11}}, \cdots, \frac{1}{a_{pp}})| = 0$. Note that $a_{11}, \cdots, a_{pp}$ can be interpreted as quantities representing relative importance of the p components of the vector $\mathbf{x}$.

Returning now to Case (a), we proceed as in Case (b) and it is easy to check that the PM θ_i simplifies to

$$\theta_i = P[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2(\eta_{ij}^2) \leq \epsilon]. \tag{41}$$

$$\leq P[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 \leq \epsilon]. \tag{42}$$

where $\lambda_1, \cdots, \lambda_p$ are now the roots of the equation $|\boldsymbol{A} - \lambda(n-1)\boldsymbol{W}_{\mathbf{x}}^{-1}| = 0$ and $\chi_{11}^2, \cdots, \chi_{1p}^2$ are independent central chi-square variables each with 1 d.f. The two special cases of choice of $\boldsymbol{A}$ can be similarly dealt here.

7.2.3. Case (c)

From the conditional multivariate normal distribution of $\bar{\mathbf{y}}|\boldsymbol{X} \sim N_p\left(\bar{\mathbf{x}}, \frac{\boldsymbol{W_x}}{n(n-1)}\right)$, it readily follows that the conditional univariate distribution of $\bar{y}_1$, given X, is normal with mean $\bar{x}_1$ and variance $\frac{\boldsymbol{W}_{\mathbf{x}11}}{n(n-1)} = d$ (say). Therefore the privacy measure (PM) θ, which is $P[(\bar{y}_1 - \bar{x}_1)^2 \leq \epsilon|\boldsymbol{X}]$, can be simplified as

$$\theta = P\left[(\bar{y}_1 - \bar{x}_1)^2 \leq \epsilon|\boldsymbol{X}\right] = P\left[\chi_1^2 \leq \frac{\epsilon}{d}\right]. \tag{43}$$

The implication of the PM in this case is obvious - the component having the maximum sampling variation will offer maximum privacy protection.

7.3. Three cases under PPS

7.3.1. Case (a)

Since the identities of the released masked data are known in this case, the intruder's obvious choice of $\mathbf{x_i}$ is $\mathbf{z_i}$, which (under the PPS scheme) conditionally given $\boldsymbol{X}$ and Σ^*, is $N_p\left(\bar{\mathbf{x}}, (1 + \frac{1}{n})\Sigma^*\right)$ with Σ^* having an Inverted Wishart distribution (see (44) below). Again, as under PIS, here also the unit i has no direct relevance.

Before we compute the PM θ in Case (a), let us look at Case (b).

7.3.2. Case (b)

Recall that $\bar{\mathbf{z}}$ is the intruder's choice of $\mathbf{x_i}$ in this case. To compute the PM θ_i, we proceed as follows.

Recall that under PPS:

$$\bar{\mathbf{z}} - \mathbf{x}_i|\Sigma^*, X \sim N_p\left(\bar{\mathbf{x}} - \mathbf{x}_i, \frac{2}{n}\Sigma^*\right) \text{ and } \Sigma^*|X \sim \text{Wishart}_p^{-1}\left(\boldsymbol{W}_{\mathbf{x}}^{-1}, n + \alpha - p - 2\right) \tag{44}$$

with (Anderson (2003))

$$h(\Sigma^*) \sim e^{-\frac{1}{2}tr\Sigma^{*-1}S_x}|\Sigma^*|^{-(\frac{n+\alpha-1}{2})}|\boldsymbol{W}_{\mathbf{x}}|^{(\frac{n+\alpha-p-2}{2})} \tag{45}$$

Combining (44) and (45), the marginal density of $\bar{\mathbf{z}}$, given X, is readily obtained as:

$$\begin{aligned} f(\bar{\mathbf{z}}|X) &\sim \int_{\Sigma^*} \frac{e^{-\frac{n}{4}(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t\Sigma^{*-1}(\bar{\mathbf{z}}-\bar{\mathbf{x}})}}{|\Sigma^*|^{p/2}} e^{-\frac{1}{2}tr\Sigma^{*-1}\boldsymbol{W}_{\mathbf{x}}}|\Sigma^*|^{-(\frac{n+\alpha-1}{2})}|\boldsymbol{W}_{\mathbf{x}}|^{(\frac{n+\alpha-p-2}{2})}d\Sigma^* \\ &\sim \int_{\Sigma^*} e^{-\frac{1}{2}tr\Sigma^{*-1}[\boldsymbol{W}_{\mathbf{x}}+\frac{n}{2}(\bar{\mathbf{z}}-\bar{\mathbf{x}})(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t]}|\Sigma^*|^{-(\frac{n+\alpha-1+p}{2})}|\boldsymbol{W}_{\mathbf{x}}|^{(\frac{n+\alpha-p-2}{2})}d\Sigma^* \\ &\sim \frac{|\boldsymbol{W}_{\mathbf{x}}|^{\frac{n+\alpha-p-2}{2}}}{|\boldsymbol{W}_{\mathbf{x}}+\frac{n}{2}(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t(\bar{\mathbf{z}}-\bar{\mathbf{x}})|^{\frac{n+\alpha-2}{2}}} \\ &\sim \frac{|\boldsymbol{W}_{\mathbf{x}}|^{-\frac{p}{2}}}{|1+\frac{n}{2}(\bar{\mathbf{z}}-\bar{\mathbf{x}})^t\boldsymbol{W}_{\mathbf{x}}^{-1}(\bar{\mathbf{z}}-\bar{\mathbf{x}})|^{\frac{n+\alpha-2}{2}}} \end{aligned} \tag{46}$$

which is a multivariate t-distribution. The privacy measure (PM) θ_i can then be written as

$$\begin{aligned} \theta_i &= P\Big[(\bar{\mathbf{z}}-\mathbf{x}_i)^t\boldsymbol{A}(\bar{\mathbf{z}}-\mathbf{x}_i) \le \epsilon|\boldsymbol{X}\Big] \\ &= P\Big[\Big\{(\bar{\mathbf{z}}-\bar{\mathbf{x}})+(\mathbf{x}_i-\bar{\mathbf{x}})\Big\}^t\boldsymbol{A}\Big\{(\bar{\mathbf{z}}-\bar{\mathbf{x}})+(\mathbf{x}_i-\bar{\mathbf{x}})\Big\} \le \epsilon|\boldsymbol{X}\Big] \\ &= P\Big[(\mathbf{y}-\boldsymbol{\zeta}_i)^t\boldsymbol{A}(\mathbf{y}-\boldsymbol{\zeta}_i) \le \epsilon|\boldsymbol{X}\Big] \end{aligned} \tag{47}$$

where $\mathbf{y} = \bar{\mathbf{z}} - \bar{\mathbf{x}}$ and $\boldsymbol{\zeta}_i = \mathbf{x}_i - \bar{\mathbf{x}}$. Note from (46) that the pdf of $\mathbf{y}$ can be written as

$$h(\mathbf{y}) \sim |\mathbf{B}|^{p/2}[1+\mathbf{y}^t\mathbf{B}\mathbf{y}]^{-\frac{n+\alpha-2}{2}} \tag{48}$$

where $\mathbf{B} = \frac{n}{2}\boldsymbol{W}_{\mathbf{x}}^{-1}$. It is well known that a multivariate t-distribution is a scale-mixture of normal and gamma. This follows because (48) can be written as

$$\sim \int_0^\infty \Big[e^{-\frac{\mathbf{y}^t\mathbf{B}\mathbf{y}}{2}u}|\mathbf{B}|^{p/2}u^{p/2}\Big]\Big[e^{-\frac{u}{2}}u^{\frac{\nu-p}{2}}\Big]du \tag{49}$$

$$\sim |\mathbf{B}|^{p/2}(1+\mathbf{y}^t\mathbf{B}\mathbf{y})^{-(\frac{\nu}{2}+1)} \text{ where } \nu = n+\alpha-4 \tag{50}$$

$$\mathbf{y}|u \sim N_p\left(\mathbf{0}, \frac{\mathbf{B}^{-1}}{u}\right), \; u \sim e^{-\frac{u}{2}}u^{\frac{\nu-p}{2}}, 0 < u < \infty. \tag{51}$$

Let $\boldsymbol{\Gamma} : p \times p$ be a nonsingular matrix such that $\boldsymbol{\Gamma}\mathbf{B}^{-1}\boldsymbol{\Gamma}^t = \boldsymbol{I}_p \Leftrightarrow \mathbf{B}^{-1} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Gamma}^t)^{-1} = (\boldsymbol{\Gamma}^t\boldsymbol{\Gamma})^{-1}$. Then $\mathbf{V}_i \stackrel{\text{def}}{=} \boldsymbol{\Gamma}(\mathbf{y}-\boldsymbol{\zeta})_i|u \sim N_p\left(-\boldsymbol{\Gamma}\boldsymbol{\zeta}_i = \boldsymbol{\delta}_i, \frac{I_p}{u}\right)$.

The privacy measure θ_i from 47) can be expressed as

$$\begin{aligned} \theta_i &= P\Big[(\mathbf{y}-\boldsymbol{\zeta}_i)^t\boldsymbol{A}(\mathbf{y}-\boldsymbol{\zeta}_i) \le \epsilon|\boldsymbol{X}\Big] \\ &= P\Big[\mathbf{V}_i^t\left((\boldsymbol{\Gamma}^{-1})^t\boldsymbol{A}\boldsymbol{\Gamma}^{-1}\right)\mathbf{V}_i \le \epsilon|\boldsymbol{X}\Big]. \end{aligned}$$

Finally, let us write $\mathbf{C} = (\mathbf{\Gamma}^{-1})^t \boldsymbol{A} \mathbf{\Gamma}^{-1}$ and choose an orthogonal matrix $\mathbf{\Lambda}$ satisfying $\mathbf{C} = \mathbf{\Lambda}^t D(\boldsymbol{\lambda}) \mathbf{\Lambda}$, where $D(\boldsymbol{\lambda})$ is a diagonal matrix with the diagonal elements as the roots of $\mathbf{C}$. Then

$$\begin{aligned}
\theta_i &= P\left[\mathbf{V}_i^t \mathbf{\Lambda}^t D(\boldsymbol{\lambda}) \mathbf{\Lambda} \mathbf{V}_i \le \epsilon | \boldsymbol{X}\right] \\
&= P\Big[\mathbf{V}_i^{*t} D(\boldsymbol{\lambda}) \mathbf{V}_i^* \le \epsilon | \boldsymbol{X}\Big], \ \mathbf{V}_i^* = \mathbf{\Lambda} \mathbf{V}_i \sim N_p\left(\boldsymbol{\eta}_i, \frac{1}{u} \boldsymbol{I}_p\right), \text{ where } \boldsymbol{\eta}_i = -\mathbf{\Lambda} \mathbf{\Gamma} \boldsymbol{\zeta}_i \\
&= E_u \left\{ P\left[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 (u \eta_{ij}^2) \le u\epsilon | u \right] \right\}, \text{ where } \eta_{ij} \text{ be the } j^{th} \text{ component of } \boldsymbol{\eta}_i. \quad (52) \\
&\le P\Big[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 (\text{central}) \le \epsilon \chi_{\nu-p+2}^2 \Big]. \quad (53)
\end{aligned}$$

Recall that $\lambda_1, \cdots, \lambda_p$ are the roots of $\mathbf{C}$, which are the same as the roots of $\boldsymbol{A}(\mathbf{\Gamma}^t \mathbf{\Gamma})^{-1} = \boldsymbol{A}\mathbf{B}^{-1} = \frac{2}{n}(\boldsymbol{A}\boldsymbol{W}_{\mathbf{x}})$, and $\chi_{11}^2, \cdots, \chi_{1p}^2$ are independent central χ^2 with 1 degree of freedom. The universal upper bound in (53) can be used as a privacy measure for any unit.

Three special cases follow.

Case 1: $\boldsymbol{A} = \mathbf{I}_p \implies \theta \le P\Big[\sum_{i=1}^{p} \lambda_i \chi_{1i}^2 (\text{central}) \le \epsilon \chi_{\nu-p+2}^2 \Big]$, where $\lambda_1, \cdots, \lambda_p$ are the roots of $\frac{2}{n} \boldsymbol{W}_{\mathbf{x}}$.

Case 2: $\boldsymbol{A} = \boldsymbol{W}_{\mathbf{x}}^{-1} \implies \lambda_1 = \cdots = \lambda_p = \frac{2}{n}$, which implies $\theta \le P\Big[\chi_p^2 \le \frac{n}{2} \epsilon \chi_{\nu-p+2}^2\Big]$.

Case 3: $\boldsymbol{A} = Diag(a_1, \cdots, a_p) \implies \lambda_1, \cdots, \lambda_p$ are the roots of $\frac{2}{n} \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_p \end{bmatrix} \boldsymbol{W}_{\mathbf{x}}$.

Returning now to Case (a), it is easy to verify from the distributional property of $\mathbf{z_i}$ and the derivation under Case (b) that here

$$\theta \le P\Big[\sum_{j=1}^{p} \lambda_j \chi_{1j}^2 (\text{central}) \le \epsilon \chi_{\nu-p+2}^2 \Big] \quad (54)$$

where $\lambda_1, \cdots, \lambda_p$ are now the roots of $(1 + \frac{1}{n}) \boldsymbol{A} \boldsymbol{W}_{\mathbf{x}}$. Three special cases as in Case b can be easily dealt here.

7.3.3. Case (c)

From the derivation under case (a), referring to equation (51) which displays the conditional multivariate normal distribution of $\bar{\mathbf{z}}$, given $\boldsymbol{X}$ and u, it readily follows that the conditional univariate distribution of $\bar{z}_1$, given $\boldsymbol{X}$ and u, is $N\left(\bar{x}_1, (\frac{2}{n}) \boldsymbol{W}_{\mathbf{x}11}\right)$ with the marginal pdf of u as $\sim e^{-u/2} u^{\frac{\nu-p}{2}}$, $0 < u < \infty$. Hence the privacy measure (PM) $P[(\bar{z}_1 - \bar{x}_1)^2 \le \epsilon]$ can be computed as

$$P[(\bar{z}_1 - \bar{x}_1)^2 \le \epsilon] = P\Big[\chi_1^2 \le \frac{n\epsilon}{2\boldsymbol{W}_{\mathbf{x}11}} \chi_{\nu-p+2}^2\Big] = P\left[F_{1,n+\alpha-p-2} \le \frac{n\epsilon(n+\alpha-p-2)}{2\boldsymbol{W}_{\mathbf{x}11}}\right] \quad (55)$$

since $\nu = n + \alpha - 4$.

Remark 3: Smart intruder's case

A smart intruder with sufficient training in statistics is likely to think in a completely different manner than a naive intruder. A general result to predict unobserved $\mathbf{X}$ from an observed $\mathbf{Y}$ is to use the conditional mean formula: $E(\mathbf{X}|\mathbf{Y})$. In our case upon observing the released data ($\mathbf{u}$, $\mathbf{y}$, $\mathbf{z}$) under the three data generation or perturbation schemes, it is possible to compute the conditional means $E(\mathbf{X}|\mathbf{u}$ or $\mathbf{y}$ or $\mathbf{z}]$ although the expressions will be quite complicated in some cases. We do not pursue this aspect here.

8. Applications

In this section, we consider one publicly accessible multivariate dataset obtained from the US Census Bureau website and another multivariate dataset on renal variables from the book by Harris and Boyd (1995). Subsequently, we employ the various data masking procedures described in the prevision sections. The goal is to construct a credible ellipsoid for the unknown mean vector based on the original data and its perturbed versions, and display and compare them. We also study which component of the multivariate data vector is expected to provide least to most privacy protection based on the criterion used in Section 7.

Subsection 8.1 provides a description and summary of the Census Bureau data for $p = 2$, while subsection 8.2 focuses on the renal dataset for $p = 3$, presenting its description and analysis. Privacy protection measures for both datasets are presented in subsection 8.3.

8.1. Description and summary of census bureau data

This subsection provides an overview of the 2023 Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) data, conducted by the Bureau of the Census for the Bureau of Labor Statistics. The ASEC Supplement includes crucial monthly demographic and labor force data, supplemented by additional details on work experience, income, noncash benefits, health insurance coverage, and migration. Our data analysis focused on the District of Columbia (D.C.) for $p = 2$, we have examined two variables, Total Household Earnings (THHE), which includes Wages and Salary income, and Other Household Earnings (OHHE), encompassing retirement, interest, dividend, and social security income, chosen from a diverse range of available data. The "2023 Annual Social and Economic Supplements" can be accessed at https://www.census.gov/data/datasets/2023/demo/cps/cps-asec-2023.html.

In our analysis for the District of Columbia data from the Census Bureau, utilizing two variables ($p = 2$: THHE and OHHE), we have examined a sample of 171 households with Total Household Earnings (THHE) more than 200,000 USD. The resulting mean vector and dispersion matrix (in thousands) are: $\bar{\mathbf{X}} = \begin{bmatrix} \text{THHE} & \text{OHHE} \\ 347.51113 & 26.44435 \end{bmatrix}$, and $\boldsymbol{S} = \boldsymbol{W}/(n-1) = \begin{bmatrix} 19649.7273 & 548.1169 \\ 548.1169 & 1241.4463 \end{bmatrix}$. Based on the original data, the observed volume (2) of the $(1-\gamma)$

level confidence ellipsoid (1) is 553.25 and the coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation (3) is 0.11205.

8.1.1. CE Under NA Data: Census bureau data

Case 1: **Unit level data available**

We have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 1000 & 10 \\ 10 & 100 \end{bmatrix}$ and $r = 100$. If unit level data are available, then for $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (5) under NA data is 577.6583. The coefficient of $|\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}$ in the expected volume expression of Equation 6 is the same as for the original data, that is 0.11205. Figure 2 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from noise added data when unit level data are available.

Case 2: **Unit level data not available**

Likewise the previous case, here we also have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 1000 & 10 \\ 10 & 100 \end{bmatrix}$ and $r = 100$. If unit level data are not available, then for $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (8) under NA data is 1026.853. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 9 is 0.2012185. Figure 3 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from noise added data when unit level data are available.

8.1.2. CE Under PIS: Census bureau data

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, the observed volume (11) under PIS is 1140.265. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 12 is 0.2258409. Figure 6 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PIS.

8.1.3. CE Under PPS: Census bureau sata

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 4$, the observed volume (15) under PPS is 1639.902. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 16 is 0.3382968. Figure 7 displays the confidence ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PPS.

8.1.4. BCCE Under PIS: Census bureau data

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (20) under PIS is 1062.07. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 21 is 0.2104. Figure 9 displays the credible ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PIS.

8.1.5. BCCE Under PPS: Census bureau data

For $p = 2$, $n = 171$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 1$, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (25) under PPS is 1019.4310. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 26 is 0.22239. Figure 10 displays the credible ellipsoid for the unknown mean vector $\boldsymbol{\mu}$ derived from synthetic data using PPS.

8.2. Description and summary of renal data

In this section, we used a renal data set from the book by Harris and Boyd (1995), Appendix 4.2 on page 137. Serum creatinine (SCR), urea nitrogen (BUN), and uric acid (UA) levels were assessed from a single blood specimen collected from a group of male medical students at the University of Virginia between 1987 and 1991 (Harris and Boyd, 1995). To demonstrate the methodologies introduced in this paper, we applied them to a subset of renal data with $p = 3$ (SCR, BUN, and UA) and a sample size of $n = 150$. The resulting mean vector and dispersion matrix are: $\bar{\mathbf{X}} = \begin{bmatrix} \text{BUN} & \text{SCR} & \text{UA} \\ 15.3600 & 1.0967 & 6.4680 \end{bmatrix}$, and $\boldsymbol{S} = \boldsymbol{W}/(n-1) = \begin{bmatrix} 12.9970 & 0.0495 & 0.3478 \\ 0.0495 & 0.0183 & 0.0574 \\ 0.3478 & 0.0574 & 1.5086 \end{bmatrix}$. Based on the original data, the observed volume (2) of the $(1-\gamma)$ level confidence ellipsoid (1) is 0.0294 and the coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation (3) is 0.0518.

8.2.1. CE Under NA Data: Renal data

Case 1: **Unit level data available**

We have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 0.7 & -0.3 & -0.3 \\ -0.3 & 0.7 & -0.3 \\ -0.3 & -0.3 & 0.7 \end{bmatrix}$ and $r = 100$. If unit level data are available, then for $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (5) under NA data is 0.24303. The coefficient of $|\boldsymbol{\Sigma} + \boldsymbol{R}|^{\frac{1}{2}}$ in the expected volume expression of Equation 6 is the same as for the original data, that is 0.0518.

Case 2: **Unit level data not available**

Likewise the previous case, here we also have taken the noise dispersion matrix as $\boldsymbol{R} = \begin{bmatrix} 0.7 & -0.3 & -0.3 \\ -0.3 & 0.7 & -0.3 \\ -0.3 & -0.3 & 0.7 \end{bmatrix}$ and $r = 100$. If unit level data are not available, then for $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error the observed volume (8) under NA data is 0.35897. The coefficient of $|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 9 is 0.10454.

8.2.2. CE Under PIS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, the observed volume (11) under PIS is 0.09941. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 12 is 0.14711.

8.2.3. CE Under PPS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 4$, the observed volume (15) under PPS is 0.15295. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 16 is 0.27928.

8.2.4. BCCE Under PIS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (20) under PIS is 0.0904. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 21 is 0.1338.

8.2.5. BCCE Under PPS: Renal data

For $p = 3$, $n = 150$, a significance level of $\gamma = 5\%$ for type-I error, $\alpha = 1$, and a hyperparameter $\delta = 10$ in the prior distribution, the observed volume (25) under PPS is 0.0605. The coefficient of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression of Equation 26 is 0.1482.

The outcomes of observed and expected volumes for both datasets under various perturbation schemes have been summarized into a single, as shown in Table (3).

Table 3: Observed volumes and the coefficients of $|\mathbf{\Sigma}|^{\frac{1}{2}}$ in the expected volume expression (denoted as *Expected) for various perturbation schemes and two data sets ($\gamma = 0.05$).**

DIFFERENT SCHEMES	Volumes	CB Data Set ($n = 171, p = 2$)	Renal Data Set ($n = 150, p = 3$)
NA DATA (Microdata Available)	Observed	577.6583	0.24303
	Expected	0.11205	0.05180
NA DATA (Microdata NOT Available)	Observed	1026.853	0.35897
	Expected	0.20122	0.10454
PIS	Observed	1140.265	0.09941
	Expected	0.22584	0.14711
PPS ($\alpha = 4$)	Observed	1639.902	0.15295
	Expected	0.33830	0.27928
PIS BAYES ($\delta = 10$)	Observed	1062.070	0.09040
	Expected	0.21040	0.13380
PPS BAYES ($\alpha = 1$, $\delta = 10$)	Observed	1019.4310	0.0605
	Expected	0.2239	0.1482

8.3. Privacy protection measures

Here we have obtained privacy protection measures for selective units from both Census Bureau data set ($p = 2$) and Renal data set ($p = 3$). Under noise added data, as it is immaterial to consider the second scenario, that is when the original microdata are not available, we have only considered the scenario when all the units of the original data are available. However we have considered the situation when only the summary statistics corresponding to the perturbed data are available. We have used privacy measures as given in the equations (36), (39) and (52) under NA data, PIS and PPS respectively. Privacy measures for two different data sets have been obtained in the following subsections.

8.3.1. Census bureau dataset

For CB data set, as mentioned in section (8.1), with two variables ($p = 2$: THHE and OHHE), we have examined a sample of 171 households with Total Household Earnings (THHE) more than 200,000 USD. We choose three responses with the values (in thousands) in the two categories as $(214.735, 113.943)$, $(305, 134.217)$ and $(500, 155)$. Under any perturbation scheme, privacy measures for each unit are obtained taking $\epsilon = 0.6(0.05)1$ and for the i^{th} selected unit $\mathbf{x}_i = (x_{i1}, x_{i2})$, the matrix $\boldsymbol{A}$ is chosen as $\boldsymbol{A} = \begin{bmatrix} \frac{1}{x_{i1}^2} & 0 \\ 0 & \frac{1}{x_{i2}^2} \end{bmatrix}$. For noise added data, the noise dispersion matrix is taken as $\boldsymbol{R} = \begin{bmatrix} 10000 & 100 \\ 100 & 1000 \end{bmatrix}$ and $\alpha = 4$ under PPS.

Table 4: Privacy Measures under different schemes of perturbation and for three different units from CB data set.

Units	Schemes	ϵ								
		0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
Unit 1	NA Data	0	0	0	0	0.0004	0.0085	0.0828	0.3353	0.6920
	PIS	0	0	0	0.0001	0.0030	0.0274	0.1342	0.3662	0.6521
	PPS	0	0.0001	0.0006	0.0051	0.0252	0.0858	0.2117	0.3988	0.6045
Unit 2	NA Data	0.0122	0.3084	0.8799	0.9971	1	1	1	1	1
	PIS	0.0206	0.3183	0.8516	0.9939	1	1	1	1	1
	PPS	0.0691	0.3584	0.7628	0.9597	0.9968	0.9999	1	1	1
Unit 3	NA Data	0	0	0.0012	0.1300	0.7500	0.9919	1	1	1
	PIS	0	0	0.0058	0.1680	0.7126	0.9793	1	1	1
	PPS	0	0.0016	0.0356	0.2449	0.6527	0.9245	0.9926	0.9996	1

8.3.2. Renal dataset

For Renal data set, as mentioned in section (8.2), with three variables ($p = 3$: SCR, BUN and UA), we have examined a sample of 150 male medical students at the University of Virginia between 1987 and 1991 (Harris and Boyd, 1995). We choose three responses with the values in the three categories as $(12, 0.9, 6.1)$, $(15, 1.1, 6.9)$ and $(25, 1.1, 6.6)$. Under any perturbation scheme, privacy measures for each unit are obtained for the choices of $\epsilon \in \{0.005, 0.01, 0.05, 0.1, 0.12, 0.14, 0.145, 0.15, 0.16\}$ and for the i^{th} selected unit $\mathbf{x}_i =$

(x_{i1}, x_{i2}, x_{i3}), the matrix $\boldsymbol{A}$ is chosen as $\boldsymbol{A} = \begin{bmatrix} \frac{1}{x_{i1}^2} & 0 & 0 \\ 0 & \frac{1}{x_{i2}^2} & 0 \\ 0 & 0 & \frac{1}{x_{i3}^2} \end{bmatrix}$. For noise added data, the noise dispersion matrix is taken as $\boldsymbol{R} = \begin{bmatrix} 0.7 & -0.3 & -0.3 \\ -0.3 & 0.7 & -0.3 \\ -0.3 & -0.3 & 0.7 \end{bmatrix}$ and $\alpha = 4$ under PPS.

Table 5: Privacy Measures under different schemes of perturbation and for three different units from Renal data set

Units	Schemes	ϵ								
		0.005	0.01	0.05	0.1	0.12	0.14	0.145	0.15	0.16
Unit 1	NA Data	0	0	0	0.1079	0.3657	0.6169	0.6697	0.7165	0.7961
	PIS	0	0	0	0.0195	0.2549	0.7275	0.8177	0.8850	0.9607
	PPS	0	0	0	0.0684	0.3109	0.6588	0.7331	0.7971	0.8916
Unit 2	NA Data	0.2478	0.7521	0.9995	1	1	1	1	1	1
	PIS	0.5071	0.9755	1	1	1	1	1	1	1
	PPS	0.4415	0.8959	1	1	1	1	1	1	1
Unit 3	NA Data	0	0	0	0	0	0	0.0009	0.3152	0.8991
	PIS	0	0	0	0	0.0003	0.1461	0.3094	0.5236	0.8710
	PPS	0	0	0	0	0.0080	0.2221	0.3560	0.5078	0.7815

9. Conclusion

Referring to Table 2 in Section 6.2, it is evident that the expected volume decreases with increasing sample size (n). Conversely, regardless of the scheme used, the expected volume increases with an increase in the number of components (p). In particular, among all perturbation schemes, the smallest expected volumes are consistently observed with the noise-added data. Moreover, in both frequentist and Bayesian frameworks, PIS resulted in a smaller expected volumes compared to PPS.

We have performed some data analyses in section (8) for Census Bureau data set ($p = 2$) and for the Renal data set ($p = 3$). The observed and expected volumes for both data sets under any scheme are summarized in Table (3). The volumes under noise added data are the smallest among all the schemes and for both the data sets, whereas under two schemes of noise added data (units available and units not available), we can see smaller volumes when units are available. For both data sets, under frequentist setup, PPS is showing larger volumes than PIS. On the other hand, under the Bayes framework, the observed volumes under PPS scheme are marginally smaller than those under the PIS scheme. The diagrams (2, 3, 6, 7, 9, 10) of the ellipsoids obtained for the CB data set under different schemes are given in the appendix. Also some diagrams are obtained overlapping the ellipsoids obtained under two different schemes as, 1. NA 1 and NA 2 (fig : 4), 2. PIS and PPS under the frequentist framework (fig : 8) and 3. PIS and PPS under Bayesian framework (fig : 11). From the diagrams it is clear that one should expect a smaller volume under NA 1 scheme than that under NA 2 scheme as the ellipsoid under NA 2 scheme is containing the ellipsoid under NA 2 scheme. Under frequentist setup, ellipsoid obtained under PPS contains the ellipsoid obtained under PIS. However, in the Bayesian framework, the scenario is not the same, where none of the ellipsoids, under PIS or PPS, contain another.

For privacy protection analysis, as carried out in section (8.3), we have selected three units from both the data sets. Units are so chosen that, one is very close to the sample mean which is happen to be the second unit in both the cases, the third units are a bit distant from the mean and the first units are taken to be extreme. Privacy measures for CB data set are shown in Table (4) and those for Renal data set are shown in Table (5). As expected, the privacy measures for the second units for each data set and under any scheme are very high, which means lower privacy protection. For both data sets, we can see a higher privacy protection for Unit 3 compared to Unit 1. Comparing the perturbation schemes in terms of the privacy measure, we can say that, no scheme can be chosen over others through out for any choices of ϵ. It depends on the choice of specific units and also upon the choices of ϵ.

Acknowledgements

Our sincere thanks are due to Professor Thomas Mathew (UMBC) for his help with Theorem 1 in Section 2.2 as well as for some other critical comments. Bimal Sinha is thankful to Dr. Tommy Wright (CSRM/Census Bureau) for his encouragement and support.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics, 3rd edition.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, volume 201. Springer Science & Business Media.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, **105**, 1347–1357.

Guin, A., Roy, A., and Sinha, B. (2023). Bayesian analysis of singly imputed synthetic data under the multivariate normal model. *International Journal of Statistical Sciences*, **23**, 1–18.

Gupta, A. and Nagar, D. (1999). *Matrix Variate Distributions*. Chapman and Hall/CRC.

Harris, E. K. and Boyd, J. C. (1995). *Statistical Bases of Reference Values in Laboratory Medicine*. CRC Press.

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

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

Klein, M., Mathew, T., and Sinha, B. (2014). Noise multiplication for statistical disclosure control of extreme values in log-normal regression samples. *Journal of Privacy and Confidentiality*, **6**, 77–125.

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

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

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**, 293–311.

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*, **7**, 43–98.

Lin, Y.-X. and Wise, P. (2012). Estimation of regression parameters from noise multiplied data. *Journal of Privacy and Confidentiality*, **4**, 61 – 94.

Little, R. J. et al. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, **9**, 407–407.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical science*, , 538–558.

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

Nayak, T., Sinha, B., and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection. *Journal of Official Statistics*, **27**, 527–544.

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

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

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, **30**, 235–242.

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

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

Reiter, J. P. (2005c). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, **21**, 441–462.

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

Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality*, **1**, 99–110.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, **91**, 473–489.

Sinha, B., Nayak, T., and Zayatz, L. (2011). Privacy protection and quantile estimation from noise multiplied data. *Sankhya B*, **73**, 297–315.

ANNEXURE

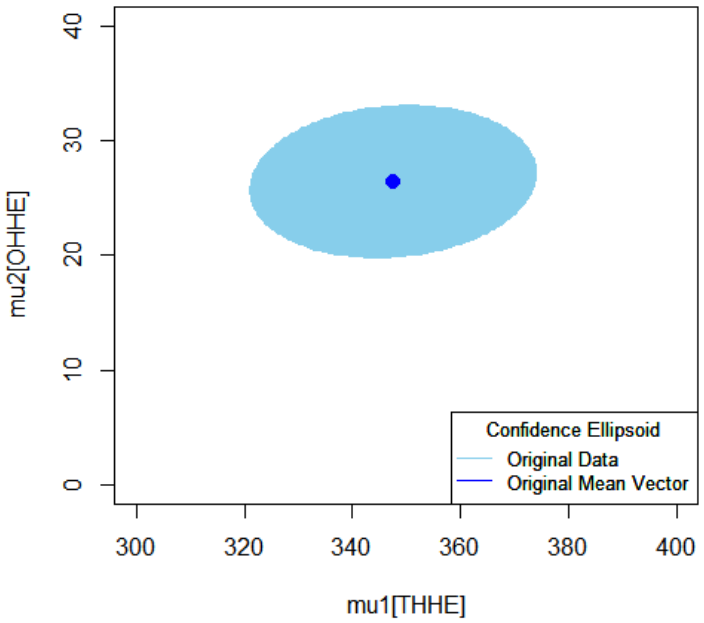

Figure 1: Confidence Ellipsoid for the unknown mean vector using original Data.

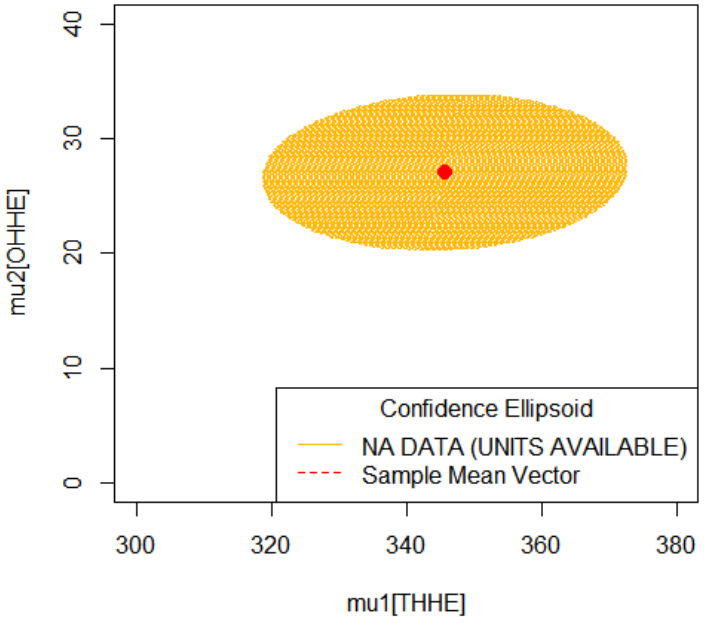

Figure 2: Confidence Ellipsoid for the unknown mean vector using Noise Added Data (Microdata Available).

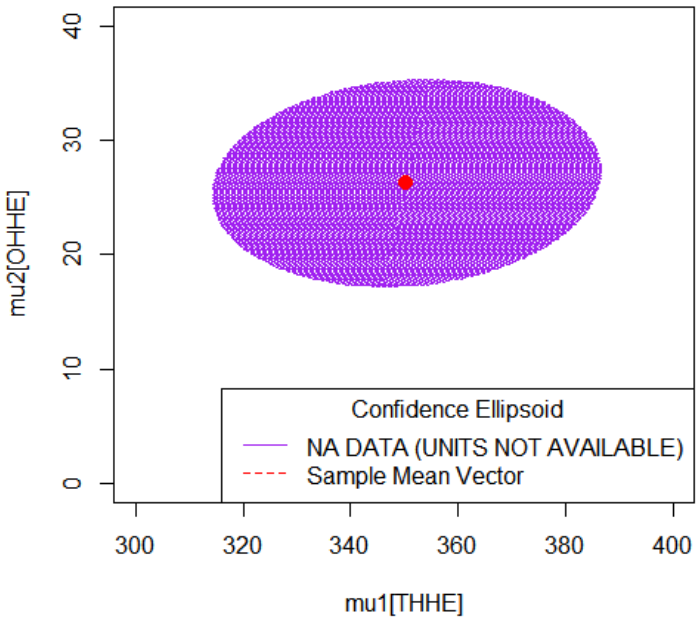

Figure 3: Confidence Ellipsoid for the unknown mean vector using Noise Added Data (Microdata Not Available).

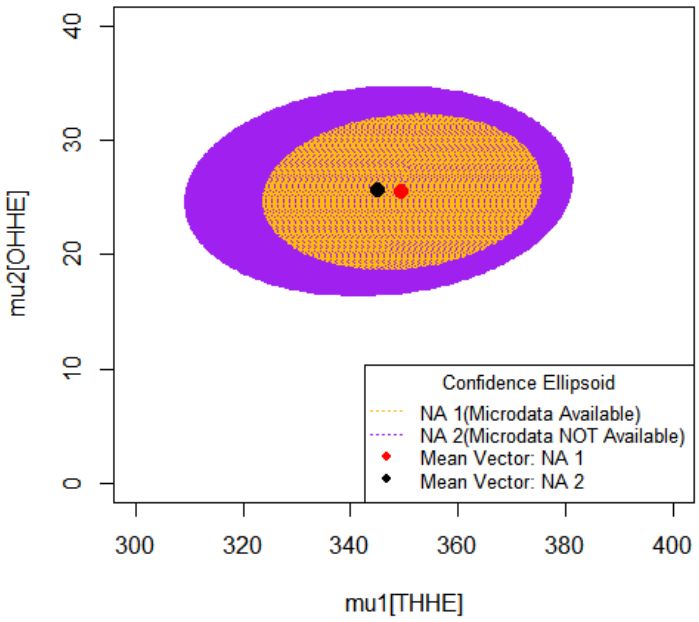

Figure 4: Confidence ellipsoids for the unknown mean vector under NA 1 (Microdata Available) and NA 2 (Microdata Not Available).

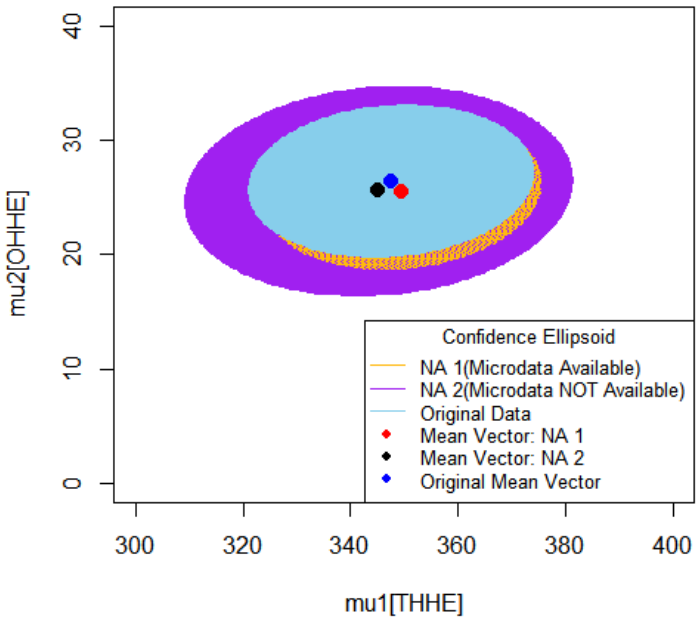

Figure 5: Confidence ellipsoids for the unknown mean vector under Original, NA 1 (Microdata Available) and NA 2 (Microdata Not Available).

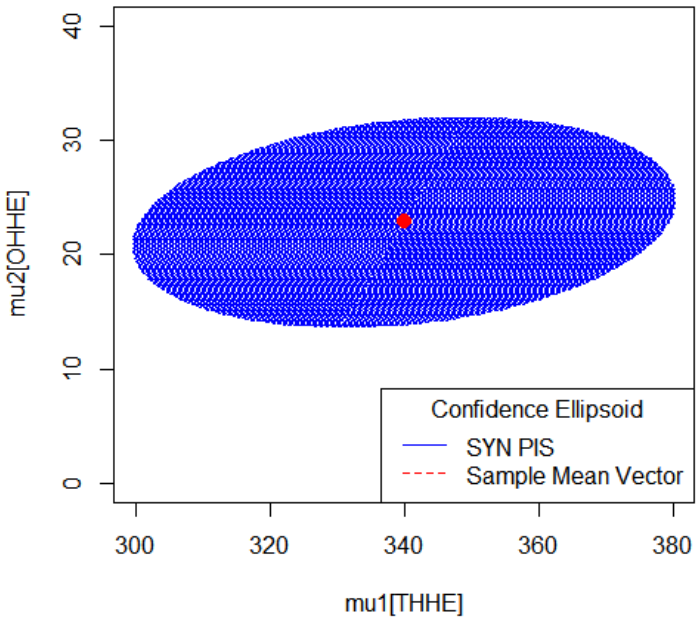

Figure 6: Confidence Ellipsoid for the unknown mean vector using synthetic data under PIS.

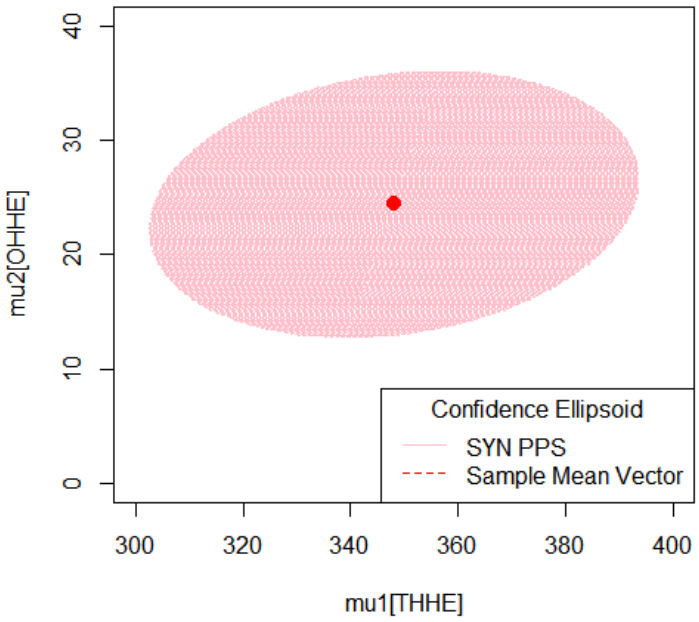

Figure 7: Confidence Ellipsoid for the unknown mean vector using synthetic data under PPS.

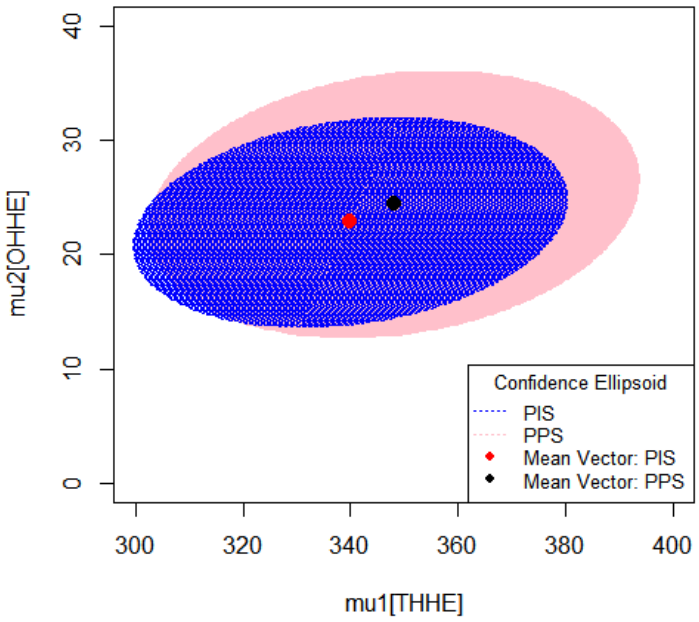

Figure 8: Confidence ellipsoids for the unknown mean vector using synthetic data under PIS and PPS.

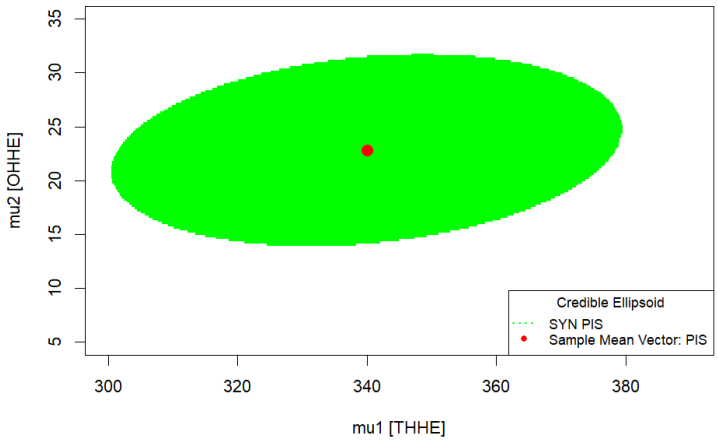

Figure 9: Credible ellipsoid for the unknown mean vector using synthetic data under PIS.

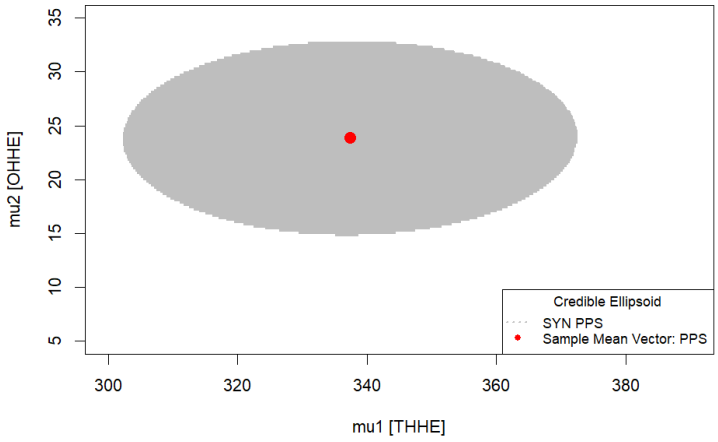

Figure 10: Credible ellipsoid for the unknown mean vector using synthetic data under PPS.

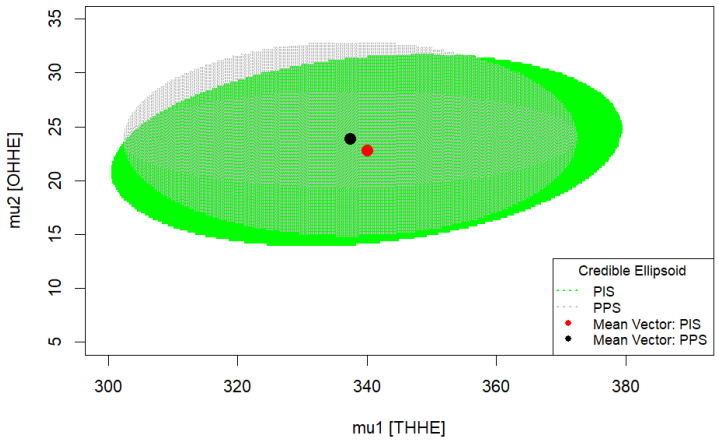

Figure 11: Credible ellipsoids for the unknown mean vector using synthetic data under PIS and PPS.