International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 21(2), 2021, pp 105-144
© 2021 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

Bayesian Analysis of Singly Imputed Partially Synthetic Data generated by Plug-in Sampling and Posterior Predictive Sampling under the Multiple Linear Regression Model

Abhishek Guin[1], **Anindya Roy**[2*] **and Bimal Sinha**[3]

[1]Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, USA, Email:guin1@umbc.edu

[2]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, USA, and Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD, USA, Email:anindya@umbc.edu

[3]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, USA, and Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, USA, Email:sinha@umbc.edu

[*]Correspondence should be addressed to Anindya Roy (Email:anindya@umbc.edu)

[Received October 2, 2021; Accepted November 15, 2021]

Abstract

In this paper we develop Bayesian inference based on singly imputed partially synthetic data, when the original data are derived from a multiple linear regression model. We assume that the synthetic data are generated by using two methods: plug-in sampling, where unknown parameters in the data model are set equal to observed values of their point estimators based on the original data, and synthetic data are drawn from this estimated version of the model; posterior predictive sampling, where an imputed posterior distribution of the unknown parameters is used to generate a posterior draw, which in turn is plugged in the original model to beget synthetic data. Simulation results are presented to demonstrate how the proposed methodology performs compared to the theoretical predictions. We outline some ways to extend the proposed methodology for certain scenarios where the required set of conditions do not hold.

Keywords: Partially synthetic data, Pivotal quantity, Plug-in sampling, Posterior predictive sampling.

AMS Subject Classification: 62F15.

0. Tribute to Professors Bimal and Bikas Sinha

While I have met and interacted with Professor Bikas Sinha several time, and have great respect for his professional contributions, it is Professor Bimal Sinha who I have been fortunate to have as my colleague, my friend and my mentor for more than twenty years. Over the years I have come to truly appreciate and be inspired by his indomitable spirit, his wisdom and his kindness. I feel privileged to contribute to this special issue of Statistics and Application honoring Professors Bimal and Bikas Sinha.

I first met Professor Bimal Sinha in 1999 when I joined the statistics program at the University of Maryland Baltimore County (UMBC), a program that was founded by Professor Sinha or 'Dr. Sinha' as many of us call him (fondly). The statistics world definitely knows about his inspiring work in multivariate statistics, higher order efficiency and in many other topics and recognizes him through several influential books that he has authored/co-authored, particularly the ones on rank set sampling and statistical meta analysis. He has also made substantial contribution to the statistical practices of the U.S federal government. What is often missed is his enormous contribution to the profession in leading the UMBC statistics program to a successful stand alone program in statistics that has nurtured and trained numerous professional statisticians, including myself. I am grateful for his friendship and guidance and look forward to many more years of association. I want to take this opportunity to wish him and Professor Bikas Sinha on their long productive and successful careers and wish them happy and healthy lives.

1. Introduction

Statistical disclosure control (SDC) methodology aims to suitably modify a dataset prior to its release so that the modified dataset does not disclose confidential information about the individual units that contributed their information to the dataset (for example, survey respondents). At the same time, it is also a goal that a dataset that has been modified using SDC methodology would still be useful for drawing inference on the relevant population. SDC methods include data swapping, additive and multiplicative noise, top and bottom coding, and also the creation of synthetic data. The synthetic data approach is a popular form of SDC methodology where (all or part of) the real data deemed confidential are not released, but are instead used to create synthetic data which are released.

Generally, there are two types of synthetic data discussed in the literature: fully synthetic data and partially synthetic data, and methodology for drawing inference based on synthetic data has been developed using concepts of multiple imputation (Rubin, 1987). In fully synthetic data methodology, all units in the population not selected in the sample are treated as missing, and are multiply imputed based on the information from sampled units, to create multiple synthetic populations. A sample is then drawn from each synthetic population, and these samples are released to the public. This approach was suggested by Rubin (1993), and methods for drawing inference based on the synthetic data generated using this approach were developed by Raghunathan et al. (2003). In the partially synthetic data approach, the released data comprise only the originally sampled units, but any responses deemed to be confidential are replaced by multiple imputations. For any particular variable, the responses could be deemed as confidential for some or all respondents. This approach was suggested by Little (1993), and methods for drawing inference based on synthetic data under this approach were developed by Reiter (2003). We refer to the monograph by Drechsler (2011) for a thorough discussion on synthetic data methodology.

In comparison with the standard SDC methods, multiple imputation techniques presents many advantages dealing with many real data problems that other methods cannot. It preserves the joint distribution of the original data offering a better quality analysis; is applicable to both categorical and continuous variables; released fully synthetic datasets gives a very small disclosure risk; with partially synthetic datasets generation one may only synthesize the records at risk, maintaining intact the records that have no need to be protected; it allows the possibility to impute missing values before generating synthetic datasets having no need to give up on some records; preserves linear constraints; allows the analyst to decide if valid results will be given from the synthetic data based on the meta-data information. Some drawbacks exist as well. Since it is a perturbation method there is a question on the utility limit of the data and only the statistical properties gathered by the model are preserved (An and Little, 2007; Drechsler, 2010).

There are several examples where partially synthetic data products have been produced based on major data sources. Some examples in the United States include the Survey of Income and Program Participation (Abowd et al., 2006; Benedetto et al., 2013), the American Community Survey Group Quarters data

(Hawala, 2008), On The Map data on where workers live and where they work (Machanavajjhala et al., 2008), and the Longitudinal Business Database (Kinney et al., 2011; Kinney et al., 2014). To obtain valid inference on population quantities using synthetic data, the current practice requires multiple synthetic datasets to be released, but there are cases where it is prudent to release only a single partially synthetic dataset. For example, the Synthetic Longitudinal Business Database, accessible through the Virtual RDC at Cornell University, is a partially synthetic version of the U.S. Census Bureau's Longitudinal Business Database (LBD). As discussed in Kinney et al. (2011) and Kinney et al. (2014), the decision was made to release only a single version of the LBD in the synthetic file, instead of multiple copies, to avoid the perception of high disclosure risk. Similarly, in the application of partially synthetic data to American Community Survey Group Quarters data presented by Hawala (2008), only a single synthetic dataset is released, because of the concern that releasing multiple synthetic copies may increase disclosure risk.

The primary purpose of this work is to develop Bayesian analyses for drawing inference based on a singly imputed partially synthetic dataset under the *multiple linear regression* (MLR) model. This synthetic data problem fits into the framework of partially synthetic data, and hence the methodology of Reiter (2003) can be used to obtain approximately valid inference if the sample size is sufficiently large and the number of multiply imputed synthetic datasets available is $m \geq 10$, but it breaks down when $m = 1$. However, given the specific structure in this problem, we shall instead exploit the model structure to derive Bayesian inference for the parameters. While the methodology we derive is specific to the problem at hand, it yields exact inference for both large and small samples using the singly imputed synthetic dataset that is available. We essentially extend the work done in Klein and Sinha (2015b) and Klein and Sinha (2015a) that developed exact parametric inferential methods based on singly imputed synthetic data for the MLR model, to the Bayesian domain.

Throughout, we would be dealing with the case of a standard MLR model involving a sensitive response variable y and a $p \times 1$ dimensional vector of *non-sensitive predictors* $\boldsymbol{x}$. We assume that

$$y_1, \ldots, y_n \text{ are independent such that } y_i \sim N(x_i'\beta, \sigma^2) \tag{1}$$

where $x_1,\ldots, x_n$ are fixed p × 1 vectors, and β and σ^2 are both unknown. Thus the original data consist of $\{(y_i; x_i) : i = 1,\ldots, n\}$. We define $y = (y_1,\ldots, y_n)'$ as the n × 1 dimensional vector of response variables, and $\boldsymbol{X} = [x_1 \ \ldots \ x_n]'$ as the n × p dimensional matrix of predictor variables, and we assume that rank $(\boldsymbol{X}) = p < n$. Based on the original data, $\hat{\beta} = \boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the maximum likelihood estimator (MLE) and uniformly minimum variance unbiased estimator (UMVUE) of β, $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$ is the UMVUE of σ^2 where $RSS = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$ with $\boldsymbol{I}_k$ as k-dimensional identity matrix and $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the orthogonal projection matrix to the column space of $\boldsymbol{X}$. Furthermore, $\boldsymbol{b}$ and RSS are independently distributed such that

$$b \sim N_p(\beta;\ \sigma^2(\mathrm{X}'\mathrm{X})^{-1}) \qquad (2)$$

$$RSS \sim \sigma^2\chi^2_{n-p}$$

When the original data are observed, b and RSS are jointly sufficient for β and σ^2.

Since $\boldsymbol{y}$ is sensitive and hence cannot be released, instead it is replaced with a singly imputed synthetic copy which is released. The synthetic data is generated by two methods: *plug-in sampling* and *posterior predictive sampling*. In the former method, parameter estimates are plugged in the MLR model to generate synthetic data. In the latter one, posterior draws of the parameter are generated using an imputed prior, which are then fed into the MLR model to generate synthetic data. The development builds on the exact likelihood based procedures developed in Klein and Sinha (2015b) and Klein and Sinha (2015a).

Plug-in Sampling. The basic mechanism for generating synthetic data via *plug-in sampling* (PIS) is described as follows: let $\boldsymbol{Y} = (y_1,\ldots, y_n)$ be the original confidential data, which are jointly distributed according to the probability density function (pdf) $f_\theta(\boldsymbol{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar or vector) parameter. To generate partially synthetic data, let $\hat{\theta} = \hat{\theta}(\boldsymbol{Y})$ be the observed value of a point estimator of $\boldsymbol{\theta}$, and we plug it into the joint pdf of $\boldsymbol{Y}$. The resulting pdf, with the unknown $\boldsymbol{\theta}$ replaced by the observed value $\hat{\theta}(\boldsymbol{Y})$ of the point estimator, is denoted by $f_{\hat{\theta}}$. The singly imputed synthetic data, denoted by $\boldsymbol{Z}$, are then generated by drawing from the joint pdf $f_{\hat{\theta}}$.

Posterior Predictive Sampling. An alternative method to generate partially synthetic data is to use *posterior predicitve sampling* (PPS) which proceeds as follows: suppose that $\boldsymbol{Y} = (y_1,\ldots, y_n)$ are the original data which are jointly

distributed according to the pdf $f_\theta(\boldsymbol{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar or vector) parameter. Assume a prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, then the imputed posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$ is obtained as $\pi(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto \pi(\theta) f_\theta(\boldsymbol{Y})$, and used to draw $\boldsymbol{\theta}^*$ (known as a posterior draw). Next, for the posterior draw of $\boldsymbol{\theta}$, a corresponding replicate of $\boldsymbol{Y}$ is generated, namely $\boldsymbol{Z} = (z_1,\ldots, z_n)'$ drawn from the pdf $f_{\theta^*}(\boldsymbol{X})$.

The organization of the paper is as follows. In Section 2, we carry out Bayesian inference based on singly imputed synthetic data generated using the plug-in sampling method. In Section 3, we derive Bayesian inference based on singly imputed synthetic data generated using posterior predictive sampling. Here we use a diffuse form of the imputer prior $\pi(\beta, \sigma^2)$, involving a hyper-parameter α. In Section 4 we present results of some simulation studies. In Section 5 we discuss the situation when part of the **y** data is sensitive, referred to as partially sensitive data. We discuss two methods of generating the synthetic data, one based on using only the sensitive part of the data to estimate model parameters and the other based on the entire data. Again, two methods of synthetic data generation are explained, based on plug-in sampling and posterior predictive sampling, and resulting Bayesian analysis are indicated.

We end this section with an observation regarding the existence of sufficient statistics in the context of synthetic data that we will use as our foundation, courtesy of Klein and Sinha (2015b).

Lemma 1.1. Suppose that when the original data $\boldsymbol{Y}$ are observed, $\boldsymbol{T}(\boldsymbol{Y})$ is a sufficient statistic for the unknown parameter $\boldsymbol{\theta}$ in the original model $f_\theta(\boldsymbol{Y})$. Then when the synthetic data $\boldsymbol{Z}$ are generated from $f_{\theta'(\boldsymbol{Y})}(\boldsymbol{Z})$ (where $\boldsymbol{\theta}'(\boldsymbol{Y})$ is a stand-in for $\boldsymbol{\theta}$ derived from the original data $\boldsymbol{Y}$), $\boldsymbol{T}(\boldsymbol{Z})$ is a sufficient statistic for $\boldsymbol{\theta}$.

Proof. Suppose based on the original data $\boldsymbol{Y}$, $\boldsymbol{T(Y)}$ is a sufficient statistic for the unknown parameter $\boldsymbol{\theta}$ in the original model $f_\theta(\boldsymbol{Y})$. Then we can write $f_\theta(\boldsymbol{Y}) = h(\boldsymbol{Y})\, g_\theta[\boldsymbol{T}(\boldsymbol{Y})]$, and the pdf of the synthetic data $\boldsymbol{Z}$ is

$$\int f_{\theta'(\boldsymbol{Y})}(Z) f_\theta(\boldsymbol{Y}) d\boldsymbol{Y} = \int g_{\theta'(\boldsymbol{Y})} T(\boldsymbol{Z}) h(\boldsymbol{Z}) f_\theta(\boldsymbol{Y}) d\boldsymbol{Y} = h(\boldsymbol{Z}) \int g_{\theta'(\boldsymbol{Y})} T(\boldsymbol{Z}) f_\theta(\boldsymbol{Y}) d\boldsymbol{Y} \tag{3}$$

2. Plug-In Sampling method

The singly imputed synthetic data in this case consist of a single synthetic version of $\boldsymbol{y} = (y_1,\ldots, y_n)'$, which is denoted as $\boldsymbol{z} = (z_1,\ldots, z_n)'$, and obtained by drawing

$$z_1,\ldots, z_n \text{ independently such that } z_i \sim N\left(x_i' b,\ \frac{RSS}{n-p}\right) \tag{4}$$

Thus the released data will be of the form $\{(z_i, x_i) : i = 1,\ldots, n\}$, and our goal is to discuss Bayesian inference on β and σ^2 based on this released data.

It is convenient to identify the latent structure of the pseudo randomization involved in the released data. For what follows we would write identities that are sometimes algebraic but also sometimes distributional. The exact case should be clear from the context. Specifically, we could write

$$z \overset{d}{=} \boldsymbol{X}\hat{\beta} + \hat{\sigma}\boldsymbol{W}$$

where $\boldsymbol{W} = (w_1,\ldots, w_n)' \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ with $w_i \overset{iid}{\sim} N(0, 1)$. Then by Lemma 1.1 the sufficient statistics based on the released data are

$$b^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'z \overset{d}{=} \hat{\boldsymbol{\beta}} + \hat{\sigma}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} \overset{d}{=} \hat{\boldsymbol{\beta}} + \hat{\sigma}CU_1$$

$$RSS^* = z'(\boldsymbol{I_n} - \boldsymbol{P_X})z \overset{d}{=} \hat{\sigma}^2\boldsymbol{W}'(\boldsymbol{I_n} - \boldsymbol{P_X})\boldsymbol{W} \overset{d}{=} \hat{\sigma}^2 v \tag{5}$$

where $\boldsymbol{U}_1 \sim N_p(0, \boldsymbol{I}_p)$ and $V \sim \chi^2_{n-p}$, and C is a full rank square root of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ such that $CC' = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. It is easy to check that b^* is independent of RSS^* by using the following result:

If $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{B}_{k\times p}$ and $\boldsymbol{A}_{p\times p}$ are constant matrices, then $\boldsymbol{By}$ and $\boldsymbol{y}'\boldsymbol{Ay}$ are independent if and only if $\boldsymbol{B\Sigma A} = \boldsymbol{O}$. Next, we can write

$$\boldsymbol{b}^* \overset{d}{=} \hat{\boldsymbol{\beta}} + \sigma\left(\frac{\hat{\sigma}}{\sigma}\right)\boldsymbol{CU}_1 \overset{d}{=} \hat{\boldsymbol{\beta}} + \sigma\sqrt{\psi}\boldsymbol{CU}_1$$

$$RSS^* \overset{d}{=} \sigma\left(\frac{\hat{\sigma}}{\sigma}\right)^2 V_1 \overset{d}{=} \sigma^2\psi V \tag{6}$$

where $\psi = (\hat{\sigma}/\sigma)^2$ is a latent quantity. From (2), we have $\hat{\beta} \overset{d}{=} \beta + \sigma\boldsymbol{CU}_2$ where $\boldsymbol{U}_2 \sim N_p(0, \boldsymbol{I}_p)$ independent of $\boldsymbol{U}_1$ and hence from (6) conditional on the parameters, we could write

$$b^* \overset{d}{=} \beta + \sigma\sqrt{1+\psi}\boldsymbol{CU}_3$$

where $\boldsymbol{U}_3 \sim N_p(0, \boldsymbol{I_p})$. Thus the likelihood based on the released data for the parameters $\theta = (\beta, \sigma^2, \psi)$ is given by

$$\mathcal{L}(\beta, \sigma^2, \psi | \boldsymbol{b}^*, RSS^*) = \phi_p(\boldsymbol{b}^*, \beta, \sigma^2(1+\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1})h(RSS^*; n-p, \sigma^2, \psi) \tag{7}$$

where $\phi_p(w; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of $w \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $h(v; d, s)$ is the density of $v \sim s\chi^2_d$. For full Bayesian specification, we need priors on the unknown quantities (β, σ^2, ψ). The prior on ψ is naturally imposed by the original MLR model and the single imputation mechanism. Thus, a priori

$$\psi \sim \pi(\psi) = h(\psi; n-p, (n-p)^{-1}$$

For Bayesian inference on the other unknown parameters we assume non-informative improper priors and assume that all unknown quantities are a priori independent. Specifically, we assume

$$\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2)$$

where $\pi(\beta) \propto 1$ and $\pi(\sigma) \propto \sigma^{-\delta}$ and hence the induced prior on σ^2 is $\pi(\sigma^2) \propto (\sigma^2)^{\frac{\delta+1}{2}}$ for $\delta > 0$. The posterior distribution can be computed in the following manner:

$$\pi(\beta, \sigma^2, \psi | \boldsymbol{b}^*, \boldsymbol{RSS}^*) \propto \mathcal{L}(\beta, \sigma^2, \psi | \boldsymbol{b}^*, \boldsymbol{RSS}^*)\pi(\psi)\pi(\beta, \sigma^2)$$

$$\pi(\beta, \sigma^2, \psi | \boldsymbol{b}^*, \boldsymbol{RSS}^*) = \pi(\beta | \boldsymbol{b}^*, \boldsymbol{RSS}^*, \sigma^2, \psi)\pi(\sigma^2 | \boldsymbol{b}^*, \boldsymbol{RSS}^*, \psi)\pi(\psi | \boldsymbol{b}^*, \boldsymbol{RSS}^*)$$

The conditional posteriors follow from observing that from the above two equations the product of the likelihood of the parameters and their priors break up into three conditional posterior distributions as follows

$$\beta | \boldsymbol{b}^*, \boldsymbol{RSS}^*, \sigma^2, \psi \sim N_p(\boldsymbol{b}^*, \sigma^2(1+\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad (8)$$

$$\sigma^2 | \boldsymbol{b}^*, \boldsymbol{RSS}^*, \psi \sim \text{Scale} - \text{inv} - \chi^2\left(n-p+\delta-1, \frac{\boldsymbol{RSS}^*}{\psi(n-p+\delta-1)}\right) \quad (9)$$

$$\psi \sim (n-p)^{-1}\chi^2_{n-p+\delta-1} \quad (10)$$

The posterior distributions are proper as long as $n > \max\{n-p+\delta-1\}$.

We observe that $\frac{\sigma^2\psi}{\boldsymbol{RSS}^*} | \boldsymbol{RSS}^*, \text{inv} - \chi^2_{n-p+\delta-1}$ so that

$$\frac{\sigma^2\psi}{\boldsymbol{RSS}^*} \sim \text{inv} - \chi^2_{n-p+\delta-1} \quad (11)$$

unconditionally and $\frac{\sigma^2\psi}{\boldsymbol{RSS}^*}$ is independent of the data and ψ. Here we use the fact that if $\boldsymbol{X} \sim$ Scale-inv-$\chi^2(\nu, \tau^2)$ then $\frac{\boldsymbol{X}}{\tau^2\nu} \sim \text{inv} - \chi^2_\nu$.

Marginal Posterior of parameters

$$\beta|\boldsymbol{b}^*, \boldsymbol{RSS}^*, \psi \sim t_{n-p+\delta-1}\left(\boldsymbol{b}^*, \frac{\boldsymbol{RSS}^*(1+\psi)}{\psi(n-p+\delta-1)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

$$\pi(\sigma^2|\boldsymbol{RSS}^*) \propto (\sigma^2)^{-\frac{n-p+\delta-1}{2}} \mathbf{K}_0\left(\sqrt{\frac{(n-p)\boldsymbol{RSS}^*}{\sigma^2}}\right)$$

where $\mathbf{K}_\nu(z)$ is the modified Bessel function of the second kind as defined in Tweedie (1957).

Marginal Distribution of data

$$\pi(\boldsymbol{b}^*, \boldsymbol{RSS}^*) = \int \pi(\boldsymbol{b}^*, \boldsymbol{RSS}^*, \psi|\beta, \sigma^2)\ \pi(\beta, \sigma^2) d\beta d\sigma^2 d\psi \ \propto (\boldsymbol{RSS}^*)^{-\frac{\delta+1}{2}}$$

Posterior Predictive Density

Let D be the original dataset and D_{new} be the new dataset with $(\tilde{b}^*, \widetilde{RSS}^*)$ as the sufficient statistic.

$$\pi(D_{\text{new}}|D) = \int \pi(D_{\text{new}}|\beta, \sigma^2, \psi)\ \pi(\beta, \sigma^2, \psi|D) d\beta d\sigma^2 d\psi$$

$$\propto (\widetilde{\mathbf{RSS}}^*)^{\frac{n-p}{2}-1} \int \left[\frac{(\tilde{\boldsymbol{b}}^* - \boldsymbol{b}^*)'(\boldsymbol{X}'\boldsymbol{X})(\tilde{\boldsymbol{b}}^* - \boldsymbol{b}^*)}{2(1+\psi)} + \frac{\widetilde{\mathbf{RSS}}^* + \mathbf{RSS}^*}{\psi}\right]^{-\frac{2n-p+\delta-1}{2}} \frac{e^{-(n-p)\psi}}{\psi^2(1+\psi)^{\frac{p}{2}}} d\psi$$

Bayes Estimators of β and σ^2

The Bayes estimators for the parameters are calculated as follows:

$$\hat{\beta}_{BAYES} = \mathrm{E}(\beta|\boldsymbol{b}^*, \mathbf{RSS}^*) = \mathrm{E}_\psi \mathrm{E}_{\sigma^2} \mathrm{E}(\beta|\boldsymbol{b}^*, \mathbf{RSS}^*, \sigma^2, \psi) = \mathrm{E}_\psi \mathrm{E}_{\sigma^2}(\boldsymbol{b}^*) = \boldsymbol{b}^*$$

$$\hat{\sigma}^2{}_{BAYES} = E(\sigma^2|\boldsymbol{b}^*, \mathbf{RSS}^*) = \mathrm{E}_\psi \mathrm{E}(\sigma^2|\boldsymbol{b}^*, \mathbf{RSS}^*, \psi) = \mathrm{E}_\psi\left(\frac{\mathbf{RSS}^*}{\psi(n-p+\delta-3)}\right)$$

$$= \frac{\mathbf{RSS}^*}{(n-p+\delta-3)} \mathrm{E}_\psi\left(\frac{\mathbf{1}}{\psi}\right) = \frac{(n-p)\mathbf{RSS}^*}{(n-p+\delta-3)^2}$$

as long as $n > \max\{p, p-\delta+3\}$. Here we use the result that if $X \sim \text{Scale} - \text{inv} - \chi^2(\nu, \tau^2)$ then $\mathrm{E}(X) = \frac{\tau^2\nu}{\nu-2}$ for $\nu > 2$.

Credible Sets for β and σ^2

We will compute pivots (we are misusing the definition a bit, we merely mean a function of data and parameters whose posterior distribution does not depend on parameters) for σ^2 and β. Given that we have closed form posterior expressions in the above equations, we can write down exact posterior intervals in terms of credibility and coverage.

A pivot for σ^2 can be defined as

$$V := \frac{\mathbf{RSS}^*}{\sigma^2}$$

whose distribution is calculated as

$$V \sim V_1 \times V_2$$

where (n - p)V_1, V_2 are independent $\chi^2_{n-p+\delta-1}$ random variables (r.v.'s) due to (11). A (1 - γ) level credible set for σ^2 based on $\mathbf{RSS}^*/\boldsymbol{\sigma^2}$ is

$$\left[\frac{\text{RSS}^*}{b_{n,p,\delta,\gamma}} - \frac{\text{RSS}^*}{a_{n,p,\delta,\gamma}}\right]$$

where $a_{n,p,\delta,\gamma}$ and $b_{n,p,\delta,\gamma}$ are any two constants that satisfy 1 - $\gamma = P\left(a_{n,p,\delta,\gamma} \leq V \leq b_{n,p,\delta,\gamma}\right)$. length of the credible interval is $\text{RSS}^* \left(\frac{1}{a_{n,p,\delta,\gamma}} - \frac{1}{b_{n,p,\delta,\gamma}}\right)$.

Next we define a pivot for β. From (8)

$$\frac{C^{-1/2}(\beta - \boldsymbol{b}^*)}{\sqrt{\mathbf{RSS}^*}} \overset{d}{=} Y_1$$

where $Y_1 \overset{d}{=} \sqrt{\frac{1}{V_2}\left(\frac{1}{V_1} + 1\right)}\, U$ such that V_1, V_2 are defined as before and are independent of $U \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$. Finally we define the pivot for β as

$$T^2 := \frac{(\beta - \boldsymbol{b}^*)'(\boldsymbol{X}'\boldsymbol{X})(\beta - \boldsymbol{b}^*)}{\mathbf{RSS}^*}$$

whose distribution is given by

$$T^2 \sim \frac{p}{n-p+\delta-1}\left(\frac{n-p}{\chi^2_{n-p+\delta-1}}+1\right)F_{p,n-p+\delta-1}$$

where the χ^2 and *F*-distributions above are independent. A $(1-\gamma)$ level credible ellipsoid for β based on T^2 is given by

$$\{\beta : T^2 \leq c_{n,p,\delta,\gamma}\}$$

where $c_{n,p,\delta,\gamma}$ satisfies $(1-\gamma) = P(T^2 \leq c_{n,p,\delta,\gamma})$. The volume of the credible ellipsoid is

$$V_\beta(z,X) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\,(c_{n,p,\delta,\gamma}\mathbf{RSS}^*)^{p/2}\,|\boldsymbol{X}'\boldsymbol{X}|^{-1/2}$$

The above expression follows from the fact that if $\mathcal{A}$ is a $p \times p$ dimensional positive definite (PD) matrix, $a \in \mathbb{R}^p$, and $C > 0$, then the volume of the ellipsoid $\{b \in \mathbb{R}^p : (b-a)'\mathcal{A}(b-a) \leq C\}$ is $\left[\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\right]C^{p/2}|\mathcal{A}|^{-1/2}$.

It is worth noting here that it is easy to show that none of the credible intervals are confidence intervals.

Remark 2.1. *If one is interested in the credible set of a single regression coefficient or moregenerally in the credible set of a vector of linear combination of β, namely* $\mathbf{A}\beta = \boldsymbol{\eta}$ *where* $\mathbf{A}$ *is a* $k \times p$ *dimensional matrix with rank(A) =* $k \leq p$*, we define* $T^2_\eta \sim (\boldsymbol{\eta} - A\boldsymbol{b}^*)'\{A(\boldsymbol{X}'\boldsymbol{X})^{-1}A'\}^{-1}(\boldsymbol{\eta} - A\boldsymbol{b}^*)/RSS^*$*, and proceed by noting that*

$$T^2_\eta \sim \frac{k(n-p)}{n-p+\delta-1}\left(\frac{1}{\chi^2_{n-p+\delta-1}}+1\right)F_{k,n-p+\delta-1}$$

where the χ^2 *and F-distributions above are independent.*

3. Posterior Predictive Sampling method

We now proceed as follows to generate the singly imputed synthetic data $\boldsymbol{z} = (z_1,\ldots, z_n)$ under posterior predictive sampling. We start from a joint prior distribution $\pi(\beta,\sigma^2) \propto (\sigma^2)^{-\frac{\alpha+1}{2}}$ for $\beta \in \mathbb{R}^p, \sigma^2 > 0$ and $\alpha > 0$ resulting in the posterior

$$\sigma^2 \mid b, \text{RSS} \sim \text{Scale} - \text{inv} - \chi^2 \left(n - p + \alpha - 1, \frac{\text{RSS}}{n-p+\alpha-1}\right) \tag{12}$$

$$\beta \mid b, \text{RSS}, \sigma^2 \sim N_p(b, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \tag{13}$$

We assume throughout that n + α > p + 1. We first draw (β^*, σ^*) from the above posterior, and then independently $z_i \sim N(x_i'\beta^*), (\sigma^*)^2), i = 1, \ldots, n$. As before, $\boldsymbol{b}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'z$ and $\text{RSS}^* = (z - X\boldsymbol{b}^*)'(z - X\boldsymbol{b}^*)$, which are jointly sufficient for (β, σ^2) by Lemma 1.1.

Similarly as in the last section we can write

$$z \overset{d}{=} \boldsymbol{X}\beta^* + \sigma^*\boldsymbol{W}$$

where $\boldsymbol{W} \sim N_n(0, \boldsymbol{I}_n)$. Then the sufficient statistics based on the released data can be written as

$$b^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'z \overset{d}{=} \beta^* + \sigma^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W} \overset{d}{=} \beta^* + \sigma^* C U_1$$

$$\text{RSS}^* = z'(\boldsymbol{I}_n - \boldsymbol{P}_X)z \overset{d}{=} \sigma^{*2}\boldsymbol{W}'(\boldsymbol{I}_n - \boldsymbol{P}_X)\boldsymbol{W} \overset{d}{=} \sigma^{*2}V \tag{14}$$

where $\boldsymbol{U}_1 \sim N_p(0, \boldsymbol{I}_p)$, $V \sim \chi^2_{n-p}$, $\boldsymbol{C}$ is such that $CC' = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, $\boldsymbol{b}^*$ and RSS^* are independent.

Thus, we get

$$\boldsymbol{b}^* \overset{d}{=} \beta^* + \sigma(\sigma^*/\sigma)CU_1 \overset{d}{=} \beta^* + \sigma\sqrt{\psi}CU_1$$

$$\text{RSS}^* \overset{d}{=} \sigma^2\left(\frac{\sigma^*}{\sigma}\right)^2 V \overset{d}{=} \sigma^2\psi V \tag{15}$$

where $\psi = \left(\frac{\sigma^*}{\sigma}\right)^2$ is a latent quantity. From (13) and (2), we have

$$\beta^* \overset{d}{=} b + \sigma^* C U_0 \overset{d}{=} \beta + \sigma C U^0 + \sigma^* C U_0 \overset{d}{=} \beta + \sigma\sqrt{1+\psi}CU_2$$

where $U_0, U^0, U_2 \sim N_p(0, \boldsymbol{I}_p)$ are all independent of each other and of $\boldsymbol{U}_1$ and hence from (15) conditional on the parameters, we could write

$$\beta^* \overset{d}{=} \beta + \sigma\sqrt{1+2\psi}CU_3$$

where $U_3 \sim N_p(0, \boldsymbol{I}_p)$. Thus the likelihood based on the released data for the parameters $\theta = (\beta, \sigma^2, \psi)$ is given by

$$\mathcal{L}(\beta, \sigma^2, \psi | \boldsymbol{b}^*, \mathbf{RSS}^*) = \phi_{\boldsymbol{p}}(\boldsymbol{b}^*; \beta, \sigma^2(1+2\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1} h(\mathrm{RSS}^*; n-p, \sigma^2\,\psi) \quad (16)$$

The prior on ψ is naturally imposed by the original MLR model and the single imputation method. From (12), $\mathrm{RSS}/\sigma^{*2} \mid \mathrm{RSS} \sim \chi^2_{n-p+\delta-1}$ and thus unconditionally $\mathrm{RSS}/\sigma^{*2} \sim \chi^2_{n-p+\delta-1}$ which also implies RSS/σ^{*2} is independent of RSS. Hence

$$\psi = \frac{\sigma^{*2}}{\sigma^2} = \frac{\mathrm{RSS}/\sigma^2}{\mathrm{RSS}/\sigma^{*2}} \overset{d}{=} \frac{n-p}{n-p+\alpha-1} F_{n-p,n-p+\alpha-1} \overset{d}{=} \beta'\left(\frac{n-p}{2}, \frac{n-p+\alpha-1}{2}\right)$$

For Bayesian inference on the other unknown parameters we assume the same independent non-informative improper priors as before. Thus for $\delta > 0$ we assume

$$\pi(\beta, \sigma^2) = \pi(\beta)\pi(\sigma^2) \propto (\sigma^2)^{-\frac{\delta+1}{2}}$$

The conditional posteriors can be determined similarly as in the last section as follows:

$$\beta \mid \boldsymbol{b}^*, \mathbf{RSS}^*, \sigma^2, \psi \sim N_p(\boldsymbol{b}^*, \sigma^2(1+2\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1}) \quad (17)$$

$$\sigma^2 \mid \boldsymbol{b}^*, \mathbf{RSS}^*, \psi \sim \mathrm{Scale-inv}-\chi^2\left(n-p+\delta-1, \frac{\mathbf{RSS}^*}{\psi(n-p+\delta-1)}\right) \quad (18)$$

$$\psi \sim \beta'\left(\frac{n-p+\delta-1}{2}, \frac{n-p+\alpha-\delta}{2}\right) \quad (19)$$

The posterior distributions are proper as long as $n > \max\{p,\ p-\delta+1,\ p-\alpha+1, p-\alpha+\delta\}$.

Here again as before we can see that $\frac{\mathbf{RSS}^*}{\sigma^2\psi} \sim \chi^2_{n-p+\delta-1}$ and thus $\frac{\mathbf{RSS}^*}{\sigma^2\psi}$ is independent of the data and ψ.

Marginal Posterior of parameters

$$\beta \mid \boldsymbol{b}^*, \mathbf{RSS}^*, \psi \sim t_{n-p+\delta-1}\left(\boldsymbol{b}^*, \frac{\mathbf{RSS}^*(1+2\psi)}{\psi(n-p+\delta-1)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

$$\pi(\sigma^2 \mid \mathbf{RSS}^*) \propto (\sigma^2)^{-\frac{n-p+\delta+1}{2}} \mathrm{U}\left(\frac{2n-2p+\alpha-1}{2}, 1, \frac{\mathbf{RSS}^*}{2\sigma^2}\right)$$

where U(*a*,*b*,*x*) is the confluent hypergeometric function of the second kind.

Marginal Distribution of data

$$\pi(\boldsymbol{b}^*, \mathbf{RSS}^*) = \int \pi(\boldsymbol{b}^*, \mathbf{RSS}^*, \psi|\beta, \sigma^2)\pi(\beta, \sigma^2)\ d\beta d\sigma^2 d\psi \ \propto (\mathbf{RSS}^*)^{-\frac{\delta+1}{2}}$$

Posterior Predictive Density

$$\pi(D_{\text{new}}|D) = \int \pi(D_{\text{new}}|\beta, \sigma^2, \psi)\ \pi(\beta, \sigma^2, \psi|D) d\beta d\sigma^2 d\psi$$

$$\propto (\widetilde{\mathbf{RSS}}^*)^{\frac{n-p}{2}-1} \int \left[\frac{(\widetilde{\boldsymbol{b}}^* - \boldsymbol{b}^*)'(\boldsymbol{X}'\boldsymbol{X})(\widetilde{\boldsymbol{b}}^* - \boldsymbol{b}^*)}{2(1+2\psi)} + \frac{\widetilde{\mathbf{RSS}}^* + \mathbf{RSS}^*}{\psi} \right]^{-\frac{2n-p+\delta-1}{2}} \frac{(1+\psi)^{-2n-2p+\alpha-1}}{\psi^2(1+\psi)^{\frac{p}{2}}} d\psi$$

Bayes Estimators of β and σ^2

The Bayes estimators for the parameters are calculated as follows:

$$\hat{\beta}_{BAYES} = \mathrm{E}(\beta|\boldsymbol{b}^*, \mathbf{RSS}^*) = \mathrm{E}_{\psi}\mathrm{E}_{\sigma^2}\mathrm{E}(\beta|\boldsymbol{b}^*, \mathbf{RSS}^*, \sigma^2, \psi) = \mathrm{E}_{\psi}\mathrm{E}_{\sigma^2}(\boldsymbol{b}^*) = \boldsymbol{b}^*$$

$$\hat{\sigma}^2{}_{BAYES} = E(\sigma^2|\boldsymbol{b}^*, \mathbf{RSS}^*) = \mathrm{E}_{\psi}\mathrm{E}(\sigma^2|\boldsymbol{b}^*, \mathbf{RSS}^*, \psi) = \mathrm{E}_{\psi}\left(\frac{\mathbf{RSS}^*}{\psi(n-p+\delta-3)}\right)$$

$$= \frac{\mathbf{RSS}^*}{(n-p+\delta-3)}\mathrm{E}_{\psi}\left(\frac{\mathbf{1}}{\psi}\right) = \frac{(n-p+\alpha-\delta)\mathbf{RSS}^*}{(n-p+\delta-3)^2}$$

as long as $n > \max\{p, p-\delta+3, p-\alpha+1, p-\alpha+\delta\}$. Here use the following facts: if $X \sim \beta'(\alpha, \beta)$ then $X^{-1} \sim \beta'(\beta, \alpha)$ and $\mathrm{E}(X) = \frac{\alpha}{\beta-1}$ for $\beta > 1$.

Credible Sets for β and σ^2

As $\frac{\mathbf{RSS}^*}{\sigma^2\psi}$ is independent of ψ so a pivot for σ^2 can be defined as

$$N := \frac{\mathbf{RSS}^*}{\sigma^2} = \left(\frac{\mathbf{RSS}^*}{\sigma^2\psi}\right)\psi = N_1 \times N_2$$

where $N_1 \sim \chi^2{}_{2\zeta}$, $N_2 \sim \beta'(\zeta, \eta)$ and N_1 is independent of N_2 where $\eta = \frac{n-p+\alpha-\delta}{2}, \zeta = \frac{n-p+\delta-1}{2}$. A $(1-\gamma)$ level credible set for σ^2 based on $N = \frac{\mathbf{RSS}^*}{\sigma^2}$ is

$$\left[\frac{\mathbf{RSS}^*}{b_{n,p,\alpha,\delta,\gamma}} - \frac{\mathbf{RSS}^*}{a_{n,p,\alpha,\delta,\gamma}}\right]$$

where $a_{n,p,\alpha,\delta,\gamma}$ and $b_{n,p,\alpha,\delta,\gamma}$ are any two constants that satisfy $(1-\gamma) = P\left(a_{n,p,\alpha,\delta,\gamma} \le N \le b_{n,p,\alpha,\delta,\gamma}\right)$.

The length of the credible interval is $\mathbf{RSS}^* \left(\frac{1}{a_{n,p,\alpha,\delta,\gamma}} - \frac{\mathbf{1}}{b_{n,p,\alpha,\delta,\gamma}}\right)$.

Let us now consider

$$T^2 := \frac{(\beta - \boldsymbol{b}^*)'(\boldsymbol{X}'\boldsymbol{X})(\beta - \boldsymbol{b}^*)}{\mathbf{RSS}^*}$$

We will compute the posterior distribution of $T^2 \,|\boldsymbol{b}^*, \mathbf{RSS}^*$. Observe that we can write

$$T^2 = \left[\frac{(\beta - \boldsymbol{b}^*)'(\boldsymbol{X}'\boldsymbol{X})(\beta - \boldsymbol{b}^*)}{\sigma^2(1+2\psi)}\right]\left[\frac{\sigma^2\psi}{\mathbf{RSS}^*}\right]\left[\frac{(1+2\psi)}{\psi}\right] = T_1 \times T_2 \times T_3$$

Now

(a) $T_1 \,|\boldsymbol{b}^*, \mathbf{RSS}^*, \sigma^2, \psi \sim {\chi^2}_p$ and hence $T_1 \sim {\chi^2}_p$ unconditionally. This also shows that T_1 is independent of $(\boldsymbol{b}^*, \mathbf{RSS}^*, \sigma^2, \psi)$ and thus T_1 is independent of T_2 and T_3.

$(b) T_2 \sim {\chi^2}_{2\zeta}$ and is independent of T_3.

$(c)\ T_3 - 2 \sim \beta'(\eta,\ \zeta)$ or alternatively $T_3 \overset{d}{=} 1 + \frac{1}{\boldsymbol{M}}$ where $\boldsymbol{M} \sim \beta(\zeta, \eta)$. This is because if $X \sim \beta'(\alpha, \beta)$ then $\frac{1}{X} \sim \beta'(\beta,\ \alpha)$ and $\frac{1}{1+X} \sim \beta(\beta,\ \alpha)$.

Hence finally we see that T^2 is a pivot for β and

$$\boldsymbol{T^2} \sim \frac{p}{n-p+\delta-1} F_{p,\ n-p+\delta-1}\left(1 + \frac{1}{\boldsymbol{M}}\right) \quad \text{where } \boldsymbol{M} \sim \beta(\zeta, \eta)$$

A $(1-\gamma)$ level credible ellipsoid for β based on T^2 is given by

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

where $c_{n,p,\alpha,\delta,\gamma}$ satisfies $(1-\gamma) = P\left(T^2 \le c_{n,p,\alpha,\delta,\gamma}\right)$. The volume of the credible ellipsoid is

$$\boldsymbol{V}_{\beta}(z,\boldsymbol{X}) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} \left(c_{n,p,\alpha,\delta,\gamma}\mathbf{RSS}^{*}\right)^{p/2} |\boldsymbol{X}'\boldsymbol{X}|^{-1/2}$$

Remark 3.1. If one is interested in the credible set of a single regression coefficient or more generally in the credible set of a vector of linear combination of β, namely, $\boldsymbol{A}\beta = \boldsymbol{\eta}$ where $\boldsymbol{A}$ is a $k \times p$ *dimensional matrix with rank(*$\boldsymbol{A}$*)* = $k \leq p$, *we define* $T_{\eta}^{2} \sim (\boldsymbol{\eta} - \boldsymbol{A}\boldsymbol{b}^{*})'\{\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'\}^{-1}(\boldsymbol{\eta} - \boldsymbol{A}\boldsymbol{b}^{*})/RSS^{*}$, *and proceed by noting that*

$$T^{2} \sim \frac{k}{n-p+\delta-1} F_{k,\ n-p+\delta-1}\left(1+\frac{1}{\boldsymbol{M}}\right) \qquad \text{where } \boldsymbol{M} \sim \beta(\zeta,\eta)$$

4. Simulation studies

In order to conduct the simulation, the population distribution is taken to be the linear regression model (1) with

$$p = 10, \quad x_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \\ x_{3i} \\ x_{4i} \\ \boldsymbol{I}(x_{5i}=2) \\ \boldsymbol{I}(x_{5i}=3) \\ \boldsymbol{I}(x_{5i}=4) \\ \boldsymbol{I}(x_{5i}=5) \\ \boldsymbol{I}(x_{5i}=6) \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \end{pmatrix} = \begin{pmatrix} 10 \\ 2 \\ 2 \\ -3 \\ -1 \\ -2 \\ 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \quad \sigma^{2} = 1. \qquad (20)$$

The regressor variables in x_i are generated one time at the beginning of the simulation, and then held fixed from one iteration to the next. We generate the regressor variables (all independently) as follows:

$$x_{1i} \sim N(1,1), \qquad \text{Log}(x_{2i}) \sim N(0,1) \qquad x_{3i} \sim \text{Exponential(mean} = 1),$$

$$x_{4i} \sim \text{Poisson}(1), \qquad x_{5i} = \begin{cases} 1 \text{ with probability } 0.2 \\ 2 \text{ with probability } 0.1 \\ 3 \text{ with probability } 0.2 \\ 4 \text{ with probability } 0.2 \\ 5 \text{ with probability } 0.2 \\ 6 \text{ with probability } 0.1 \end{cases}$$

Based on Monte Carlo simulation with 10^4 iterations, we compute an estimate of the coverage probability, the volume or length (as appropriate) of the respective credible sets and the Bayes estimators of the parameters, where in all cases, the level of credibility is set at 0.95.

Plug-In Sampling Tables 1, 2, 3 includes the simulation results for a plug-in sampling data where the sample size n equals 500, 1000 and 10000 respectively for different values of the tuning parameter δ. Some interesting observations are in order. The coverage for σ^2 gets slightly better initially as we increase δ, starts worsening beyond $\delta \geq 10$, and at large values of δ it is significantly worse. This effect is more prominent when n is small, in which case the coverage is not the best anyway as is to be expected. The same effect is observed for the coverage of β though not as severe. The coverage of β decreases at a much slower rate compared to that of σ^2 with increasing δ. The size of the credible sets shrink for both the parameters as n or δ increases. With decreasing n or increasing δ there seems to be no effect on the Bayes estimator of β, while the Bayes estimator of σ^2 becomes slightly worse, which is what we expect since $\hat{\beta}_{\text{BAYES}}$ does not involve δ while σ^2_{BAYES} has δ in the denominator. All of this suggests that there is a sweet spot for the choice of δ to ensure maximum coverage along with the smallest possible size of the credible sets of the parameters. For both σ^2 and β, from Table 3 asymptotically the results imply that the Bernstein-von Mises theorem holds, with the caveat that inference worsens with increasing δ, quicker for σ^2 than for β. In the asymptotic case, the credible sets are tighter and the Bayes estimators perform admirably for both the parameters, as expected. The behavior of the coverage of σ^2 and β with respect to different values of δ in the case n = 500 (depicted by alternating dashes and dots), n = 1000 (depicted by solid lines), asymptotic case n = 10000 (depicted by dashed lines) are represented in Figure 1(a) and Figure 1(b) respectively.

Posterior Predictive Sampling The general trend of Bayesian inference for model parameters observed under PIS is also mirrored when data is generated by posterior predictive sampling, as illustrated in Tables 4, 5, 6, 7, 8 and 9. Overall for σ^2, compared to PIS, the coverage is lower, the credible interval is wider, but the Bayes estimator performs similarly well. For β, compared to PIS, the coverage is similar, the Bayes estimator performs similarly well, but the volume of the credible ellipsoid is one order of magnitude bigger. The interaction of the hyperparameter α and tuning parameter δ is also pretty interesting to observe.

Increasing α seems to have no effect on the coverage of the parameters but the size of the credible sets narrow down marginally, although asymptotically there seems to be no significant difference (as seen by comparing Tables 6 and 9).

We should be able to find a combination of the two that yields the best inference. The inference for β seems to be unaffected by the increase in α, except again for the fact that the credible set for β contracts a bit. The behavior of the coverage of σ^2 and β with respect to different values of δ in the case n = 500 (depicted by alternating dashes and dots), n = 1000 (depicted by solid lines), asymptotic case n = 10000 (depicted by dashed lines) are represented in Figures 1(c), 1(e) and Figures 1(d), 1(f) respectively.

After assessing the results, the recommendation would be to use $2 \leq \delta \leq 4$.

The PIS method offers smaller radius of the confidence sets than the PPS method and also gives estimates of the parameters closer to the ones obtained from the original data, despite giving slightly higher levels of disclosure risk (Moura (2016)). So we have a trade off between data utility and data privacy.

In general, the Bayesian posterior intervals, credible intervals and HPD intervals need not have valid frequentist coverage. This is because the Bayesian intervals are not derived using a repeated sampling paradigm; their objective is to characterize reasonable parameter values that conform with the specific model and prior for a given situation. However, some researchers have advocated a more principled approach to the practice where the Bayes intervals are calibrated to frequentist calculations so that Bayesian statements can be rejected based on empirical tests. Such calibrated Bayes approach (Rubin, 1984; Little, 2006) looks for reconciliation between the two paradigms. Another approach for reconciliation (asymptotically) is to choose priors that provided credible intervals with accurate frequentist coverage. Such priors are called Probability Matching Priors (Datta and Ghosh, 1995).

Usually, Bayesian credible intervals have good frequentist properties provided the problem admits some type of Bernstein-von Mises theorem. In the present case however, in the presence of latent structure, such Bernstein-von Mises results may not be readily available. From the limited empirical investigation it seems that the coverage of the credible intervals depends on the δ in the prior even asymptotically. It will be interesting to determine the limits of coverage as $\delta > 0$ varies. We will pursue such investigation in the future.

Table 1: Inference for β and σ^2 for PIS data with n = 500

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	Bayes est
0.2	0.953	0.360	1.011	0.953	1.06e-03	(10.002, 2.000, 2.000, -2.999, -1.000, -2.000, 0.997, 1.998, 1.998, 3.998)’
0.5	0.948	0.359	1.010	0.950	8.81e-04	(10.001, 1.999, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.000, 4.000)’
0.8	0.952	0.359	1.009	0.950	8.25e-04	(9.998, 2.000, 2.000, -2.999, -1.001, -2.000, 1.006, 2.001, 2.002, 4.005)’
1	0.951	0.359	1.010	0.949	9.71e-04	(10.002, 2.001, 2.000, -3.000, -0.999, -2.004, 0.997, 1.995, 1.998, 3.998)’
2	0.951	0.357	1.005	0.949	8.05e-04	(10.000, 2.001, 2.000, -2.999, -0.999, -2.003, 0.998, 1.996, 1.998, 3.997)’
3	0.948	0.355	1.000	0.947	3.99e-04	(9.998, 2.000, 2.000, -3.000, -1.000, -2.001, 1.002, 1.998, 2.000, 3.997)’
4	0.945	0.353	0.996	0.946	6.73e-04	(9.997, 2.001, 2.001, -3.000, -0.999, -2.000, 1.002, 1.999, 2.001, 4.000)’
10	0.933	0.342	0.972	0.944	4.31e-04	(10.000, 2.000, 2.000, -3.001, -1.000, -2.002, 1.000, 2.002, 2.001, 4.001)’
20	0.863	0.326	0.934	0.931	4.78e-04	(10.001, 2.000, 2.001, -3.001, -1.001, -2.002, 1.000, 2.002, 2.003, 4.000)’
30	0.745	0.310	0.899	0.919	4.96e-04	(9.997, 2.000, 2.000, -2.999, -1.000, -2.000, 1.002, 2.001, 2.003, 4.000)’
50	0.426	0.282	0.834	0.898	3.86e-04	(10.000, 2.000, 2.001, -2.999, -1.001, -2.003, 1.002, 1.998, 1.998, 3.998)’
100	0.010	0.226	0.697	0.825	1.94e-04	(9.998, 2.000, 2.000, -3.000, -1.000, -2.000, 1.001, 2.000, 2.000, 4.000)’

Table 2: Inference for β and σ^2 for PIS data with n = 1000

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.951	0.251	1.006	0.949	2.48e-05	(10.002, 1.999, 2.000, -3.000, -1.000, -2.001, 1.000, 1.999, 1.999, 4.001)’
0.5	0.953	0.251	1.004	0.951	2.13e-05	(10.002, 1.999, 2.000, -3.000, -1.000, -2.000, 0.999, 1.997, 1.999, 4.000)’
0.8	0.951	0.250	1.003	0.953	2.19e-05	(10.000, 2.000, 2.000, -3.000, -1.000, -1.998, 1.000, 2.000, 2.000, 4.000)’
1	0.950	0.251	1.004	0.951	2.22e-05	(10.002, 1.999, 2.000, -3.000, -1.000, -2.001, 0.998, 1.998, 1.997, 3.997)’
2	0.949	0.250	1.004	0.946	2.36e-05	(10.000, 2.000, 2.000, -3.001, -1.001, -2.000, 1.002, 2.000, 2.000, 3.999)’
3	0.949	0.250	1.001	0.947	2.47e-05	(10.000, 2.000, 2.000, -3.000, -1.000, -1.998, 1.001, 2.001, 1.990, 3.997)’
4	0.948	0.248	0.997	0.949	2.17e-05	(10.000, 2.000, 2.000, -3.000, -1.000, -1.998, 0.999, 2.000, 2.000, 4.000)’
10	0.939	0.245	0.986	0.943	1.56e-05	(10.000, 2.000, 2.000, -3.000, -1.000, -2.001, 1.000, 2.002, 2.001, 3.999)’
20	0.906	0.239	0.965	0.943	2.20e-05	(10.001, 2.000, 2.000, -3.000, -1.000, -1.999, 1.001, 1.998, 1.998, 3.996)’
30	0.843	0.233	0.947	0.935	2.16e-05	(10.001, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 1.999, 2.000, 4.000)’
50	0.661	0.222	0.912	0.926	1.53e-05	(9.996, 2.000, 2.001, -3.001, -1.000, -1.997, 1.003, 2.005, 2.004, 4.004)’
100	0.133	0.198	0.830	0.899	1.12e-05	(10.000, 1.999, 2.000, -3.001, -1.000, -1.998, 1.001, 2.001, 1.999, 4.002)’

Table 3: Inference for β and σ^2 for PIS data with n = 10000

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.947	0.078	1.000	0.945	1.81e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.001, 1.000, 1.999, 2.000, 4.000)’
0.5	0.949	0.078	1.000	0.950	1.94e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.001, 1.000, 2.000, 2.000, 4.001)’
0.8	0.95	0.078	1.001	0.951	2.03e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.001, 2.001, 2.000, 4.001)’
1	0.948	0.078	1.000	0.951	2.10e-10	(10.000, 2.000, 2.000, -3.000, -2.000, -2.000, 1.000, 2.000, 2.000, 3.999)’
2	0.95	0.078	1.000	0.950	2.00e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.001, 4.000)’
3	0.949	0.078	1.000	0.947	2.09e-10	(9.999, 2.000, 2.000, -3.000, -1.000, -1.999, 1.001, 2.001, 2.000, 4.001)’
4	0.951	0.078	1.000	0.949	2.05e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.001, 2.000, 4.000)’
10	0.947	0.078	0.999	0.952	1.93e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.001, 2.000, 2.000, 4.001)’
20	0.946	0.078	0.997	0.951	1.90e-10	(10.000, 2.000, 3.000, -3.000, -1.000, -1.999, 1.001, 2.001, 2.000, 4.001)’
30	0.944	0.078	0.995	0.948	1.91e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.000, 4.000)’
50	0.927	0.078	0.991	0.944	1.83e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.001, 4.000)’
100	0.823	0.077	0.981	0.946	1.61e-10	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.001, 4.001)’

Table 4: Inference for β and σ^2 for PPS data with α = 2, n = 500

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.955	0.443	1.017	0.951	7.19e-04	(10.003, 2.001, 2.000, -3.001, -1.000, -2.004, 0.995, 1.999, 1.997, 3.997)’
0.5	0.955	0.441	1.015	0.949	5.62e-03	(9.996, 2.001, 2.000, -3.000, -1.000, -1.997, 1.005, 2.004, 2.004, 4.005)’
0.8	0.949	0.441	1.013	0.946	5.36e-03	(10.000, 1.999, 2.000, -2.998, -1.001, -1.996, 0.999, 2.001, 1.999, 3.999)’
1	0.947	0.441	1.013	0.949	5.07e-03	(9.999, 2.000, 2.000, -3.000, -0.999, -1.998, 1.000, 2.001, 2.000, 4.000)’
2	0.944	0.438	1.006	0.95	4.86e-03	(10.001, 2.000, 2.000, -3.000, -1.000, -1.996, 0.998, 1.999, 2.002, 3.998)’
3	0.951	0.435	1.000	0.946	6.30e-03	(9.998, 2.001, 2.000, -3.000, -0.999, -2.000, 1.000, 1.999, 1.999, 4.002)’
4	0.948	0.432	0.994	0.952	8.09e-03	(10.001, 1.999, 2.000, -2.999, -1.001, -1.998, 1.000, 2.002, 2.002, 3.999)’
10	0.927	0.415	0.958	0.94	5.26e-03	(10.002, 2.000, 2.000, -3.001, -0.998, -2.005, 0.998, 1.996, 1.995, 3.997)’
20	0.818	0.389	0.901	0.929	5.40e-03	(9.998, 2.000, 2.000, -2.999, -1.000, -2.002, 1.000, 2.002, 2.003, 4.003)’
30	0.638	0.366	0.848	0.919	3.99e-03	(10.001, 2.000, 2.000, -3.000, -1.000, -1.998, 0.999, 1.998, 1.997, 3.996)’
50	0.232	0.323	0.752	0.891	3.00e-03	(10.001, 2.001, 2.001, -3.000, -1.000, -2.002, 0.999, 1.996, 1.998, 3.998)’
100	4.0e-04	0.239	0.559	0.805	1.25e-03	(9.998, 2.001, 2.000, -3.000, -1.000, -1.997, 1.003, 2.002, 2.003, 4.003)’

Table 5: Inference for β and σ^2 for PPS data with $\alpha = 2$, $n = 1000$

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.950	0.308	1.009	0.947	2.38e-04	(10.000, 2.001, 2.000, -3.000, -1.000, -1.999, 1.003, 1.999, 2.002, 4.004)’
0.5	0.947	0.308	1.008	0.951	1.56e-04	(10.000, 2.000, 2.000, -3.000, -1.000, -2.003, 1.001, 2.001, 2.000, 4.000)’
0.8	0.949	0.308	1.007	0.953	1.78e-04	(10.005, 1.999, 1.999, -3.000, -1.001, -2.002, 0.995, 1.997, 1.997, 3.996)’
1	0.945	0.308	1.007	0.949	1.60e-04	(10.000, 2.000, 2.000, -3.000, -1.000, -2.002, 0.999, 2.001, 2.002, 4.000)’
2	0.950	0.307	1.004	0.949	1.48e-04	(9.999, 2.000, 2.000, -3.000, -1.000, -2.000, 0.999, 1.999, 1.999, 4.003)’
3	0.951	0.306	1.000	0.950	2.09e-04	(9.999, 2.001, 2.000, -3.000, -1.000, -2.003, 0.998, 2.001, 2.001, 3.998)’
4	0.949	0.304	0.996	0.952	1.67e-04	(9.999, 2.000, 2.000, -3.000, -1.000, -1.999, 1.001, 2.001, 2.003, 4.006)’
10	0.937	0.299	0.979	0.944	1.42e-04	(9.999, 2.000, 2.000, -2.999, -1.000, -2.001, 1.000, 2.001, 2.001, 3.999)’
20	0.885	0.289	0.949	0.935	1.09e-04	(10.000, 2.000, 2.000, -3.000, -1.000, -2.002, 0.999, 2.000, 1.998, 3.998)’
30	0.796	0.280	0.921	0.937	1.36e-04	(10.002, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.001, 2.001, 4.001)’
50	0.521	0.263	0.867	0.922	1.20e-04	(10.000, 2.000, 2.000, -3.000, -1.001, -2.000, 1.000, 2.000, 2.000, 4.000)’
100	0.030	0.226	0.749	0.893	8.03e-05	(10.001, 2.000, 2.000, -2.999, -1.001, -2.004, 1.000, 1.998, 2.000, 3.999)’

Table 6: Inference for β and σ^2 for PPS data with $\alpha = 2$, $n = 10000$

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.951	0.096	1.001	0.948	1.54e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.000, 4.001)’
0.5	0.954	0.096	1.001	0.953	1.52e-09	(9.999, 2.000, 2.000, -3.000, -1.000, -1.999, 1.001, 2.000, 2.000, 4.000)’
0.8	0.952	0.096	1.001	0.950	1.52e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.002, 0.999, 2.000, 1.999, 4.000)’
1	0.949	0.096	1.000	0.952	1.54e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.000, 4.001)’
2	0.951	0.096	1.001	0.945	1.55e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.000, 4.000)’
3	0.952	0.096	1.000	0.951	1.52e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.001, 4.000)’
4	0.949	0.096	1.000	0.950	1.59e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.001, 2.001, 4.001)’
10	0.948	0.096	0.998	0.950	1.59e-09	(10.001, 2.000, 2.000, -3.000, -1.000, -2.001, 1.000, 2.000, 2.000, 3.999)’
20	0.945	0.096	0.995	0.946	1.50e-09	(9.999, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.000, 4.000)’
30	0.935	0.095	0.992	0.946	1.56e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.001, 1.000, 2.000, 2.000, 3.999)’
50	0.91	0.095	0.986	0.946	1.47e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.000, 4.001)’
100	0.773	0.093	0.972	0.942	1.51e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.001, 0.999, 2.000, 2.000, 4.000)’

Table 7: Inference for β and σ^2 for PPS data with $\alpha = 50$, $n = 500$

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.951	0.436	1.017	0.949	4.02e-03	(10.002, 2.000, 2.000, -3.000, -1.000, -1.998, 1.000, 1.999, 1.998, 3.999)’
0.5	0.952	0.435	1.014	0.953	4.33e-03	(10.000, 2.001, 2.000, -3.000, -0.999, -2.004, 1.000, 1.997, 1.996, 3.994)’
0.8	0.947	0.433	1.011	0.945	5.19e-03	(10.002, 1.999, 1.999, -3.000, -0.999, -2.003, 1.000, 1.999, 1.998, 4.001)’
1	0.953	0.433	1.009	0.953	4.02e-03	(9.998, 2.000, 2.001, -2.999, -1.001, -2.002, 1.001, 1.997, 2.000, 4.004)’
2	0.949	0.431	1.006	0.947	6.53e-03	(10.002, 2.000, 2.000, -3.000, -1.001, -2.001, 1.000, 1.996, 2.000, 4.000)’
3	0.949	0.428	0.999	0.949	5.48e-03	(10.001, 1.999, 2.000, -2.999, -0.999, -2.003, 0.998, 1.999, 1.998, 4.001)’
4	0.944	0.425	0.994	0.949	4.12e-03	(9.998, 2.001, 2.000, -3.000, -1.000, -1.999, 0.999, 1.997, 2.000, 4.003)’
10	0.92	0.409	0.959	0.938	4.19e-03	(10.002, 2.000, 2.000, -3.000, -1.002, -2.001, 1.000, 2.001, 1.999, 3.999)’
20	0.822	0.385	0.904	0.931	4.66e-03	(9.998, 2.000, 2.000, -3.000, -1.000, -1.997, 1.004, 2.001, 2.002, 4.001)’
30	0.64	0.361	0.852	0.919	3.83e-03	(10.001, 1.999, 2.000, -3.001, -1.000, -2.000, 0.999, 1.998, 2.000, 4.007)’
50	0.254	0.321	0.761	0.891	2.08e-03	(9.996, 2.001, 2.000, -3.000, -1.000, -1.995, 1.006, 2.005, 2.002, 4.003)’
100	4.0E-04	0.239	0.571	0.809	9.47e-04	(10.001, 2.000, 2.000, -3.001, -1.000, -1.999, 0.998, 1.999, 2.001, 3.996)’

Table 8: Inference for β and σ^2 for PPS data with $\alpha = 50$, $n = 1000$

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.947	0.306	1.010	0.951	1.41e-04	(10.000, 2.000, 2.000, -3.000, -1.001, -2.004, 1.001, 2.000, 1.998, 3.999)’
0.5	0.948	0.305	1.007	0.950	1.73e-04	(10.002, 2.000, 2.000, -3.001, -1.000, -2.002, 0.997, 1.999, 1.999, 3.999)’
0.8	0.949	0.305	1.007	0.948	1.42e-04	(9.998, 2.000, 2.000, -2.999, -1.000, -1.999, 1.001, 1.999, 1.999, 3.999)’
1	0.953	0.305	1.005	0.945	1.54e-04	(10.000, 1.999, 2.000, -3.000, -1.000, -1.999, 1.001, 2.003, 2.001, 4.000)’
2	0.949	0.304	1.003	0.951	1.68e-04	(9.998, 2.000, 2.000, -3.000, -1.000, -2.001, 1.001, 2.002, 2.002, 4.003)’
3	0.947	0.303	1.000	0.948	1.14e-04	(10.002, 2.000, 2.000, -3.001, -1.001, -1.999, 1.001, 2.000, 2.003, 3.990)’
4	0.949	0.302	0.998	0.949	1.70e-04	(10.000, 2.000, 2.000, -3.000, -1.000, -1.997, 1.002, 2.001, 2.001, 4.000)’
10	0.936	0.297	0.980	0.947	1.36e-04	(10.002, 1.999, 2.000, -3.000, -1.001, -1.999, 0.999, 1.999, 2.000, 4.001)’
20	0.882	0.287	0.951	0.937	1.19e-04	(9.998, 2.001, 2.000, -3.000, -1.001, -1.996, 1.005, 2.002, 2.005, 4.002)’
30	0.791	0.278	0.922	0.933	1.42e-04	(10.000, 2.000, 2.000, -2.999, -1.000, -2.000, 1.000, 1.997, 2.000, 3.998)’
50	0.535	0.262	0.871	0.924	8.79e-05	(10.001, 2.000, 2.000, -3.000, -1.000, -2.003, 1.000, 1.999, 1.999, 4.003)’
100	0.034	0.225	0.752	0.888	4.55e-05	(9.999, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.002, 3.997)’

Table 9: Inference for β and σ^2 for PPS data with $\alpha = 50$, $n = 10000$

δ	σ^2			β		
	avg cvg	*est len*	*Bayes est*	*avg cvg*	*est vol*	*Bayes est*
0.2	0.953	0.096	1.001	0.950	1.48e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.001, 1.000, 1.999, 1.999, 3.999)’
0.5	0.947	0.096	1.001	0.950	1.62e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.001, 2.000, 4.000)’
0.8	0.950	0.096	1.001	0.950	1.46e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.001, 2.000, 2.000, 4.001)’
1	0.947	0.096	1.001	0.952	1.60e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 0.999, 2.000, 2.000, 4.000)’
2	0.948	0.096	1.000	0.950	1.40e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 1.998, 1.999, 3.999)’
3	0.952	0.096	1.000	0.951	1.42e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.001, 2.000, 2.001, 4.000)’
4	0.946	0.096	1.000	0.949	1.38e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.001, 3.000, 4.000)’
10	0.949	0.096	0.998	0.951	1.32e-09	(10.001, 2.000, 2.000, -3.000, -1.000, -2.000, 0.999, 2.000, 1.999, 3.999)’
20	0.940	0.095	0.994	0.948	1.45e-09	(9.999, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.001, 2.000, 4.000)’
30	0.935	0.095	0.992	0.946	1.52e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -1.999, 1.000, 2.000, 2.001, 4.000)’
50	0.909	0.095	0.986	0.950	1.44e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.001, 4.000)’
100	0.773	0.093	0.971	0.944	1.38e-09	(10.000, 2.000, 2.000, -3.000, -1.000, -2.000, 1.000, 2.000, 2.000, 4.001)’

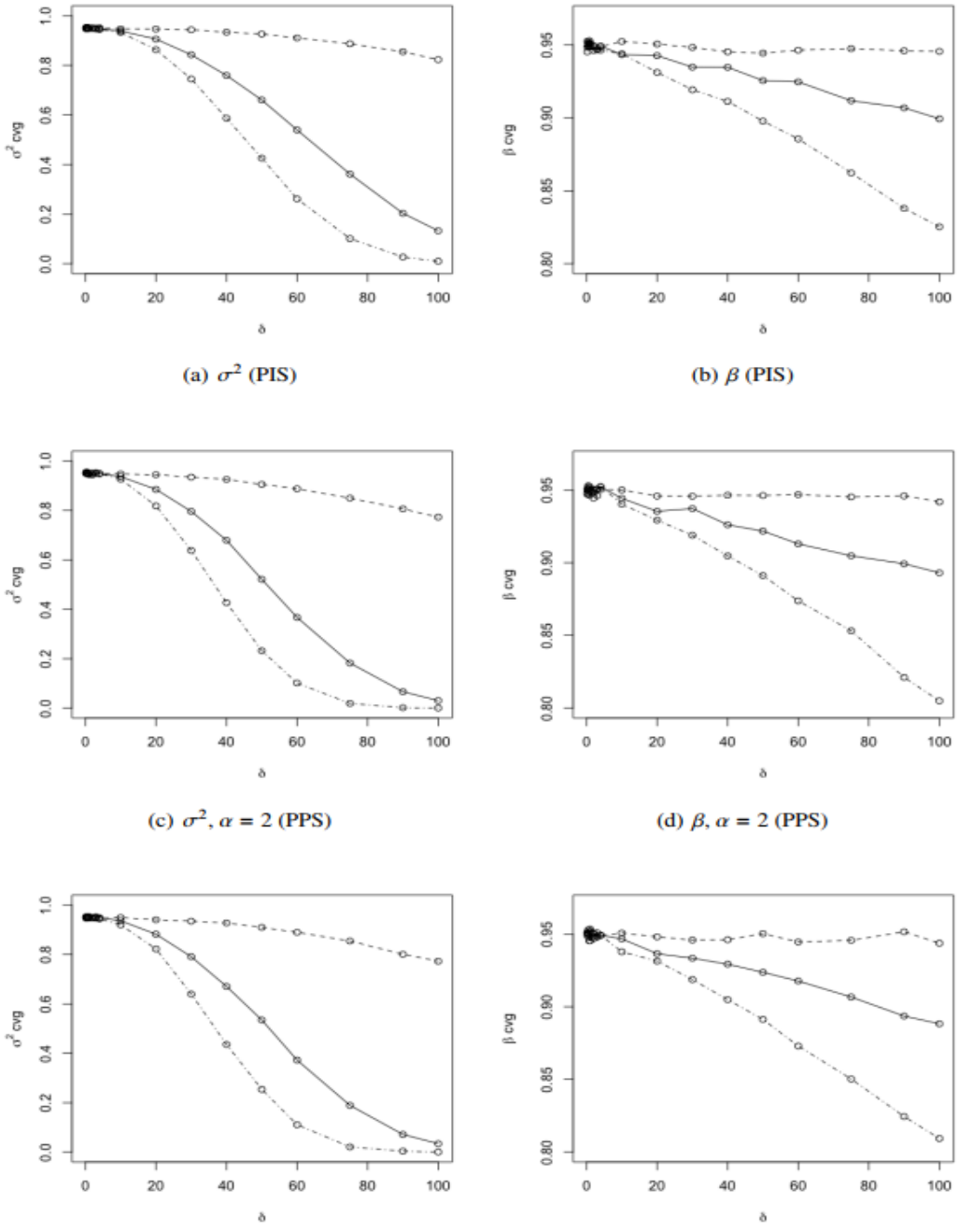

Figure 1: Variation in coverage of β and σ^2 with respect to δ for SI MLR data (--- n = 500, ---n = 1000, --- n = 10000)

5. Partially Sensitive Data

We have assumed so far that all the n observations $\boldsymbol{y} = (y_1, \ldots, y_n)'$ in the multiple linear regression model are sensitive. Of course, this need not be the case, and quite generally we can partition $\boldsymbol{y}$ into two parts: $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ of dimensions r and $(n - r)$, respectively, and assume that the first r observations $\boldsymbol{y}_1$ are sensitive, thus requiring privacy protection, and the remaining $(n - r)$ observations $\boldsymbol{y}_2$ are non-sensitive, and can remain unprotected. Let $\boldsymbol{X} = [\boldsymbol{X}_1', \boldsymbol{X}_2']'$ be the corresponding partitioning of the matrix $\boldsymbol{X}$, so that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are of dimensions $r \times p$ and $(n - r) \times p$, respectively. The reasons for some of the y-values being sensitive can vary depending on the context. For example, for income data, large incomes (extreme values) may be sensitive. The sensitive nature of y may also depend on the (extreme) values of the corresponding covariates $\boldsymbol{x}$. We outline below two data analysis procedures when the latter situation holds, namely, the sensitivity of the first r values of $\boldsymbol{y}$ is due to the nature of the covariates, which makes r a *non-random integer*.

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

Plug-In Sampling

We propose to synthesize the r sensitive y-values $\boldsymbol{y}_1$ by applying the plug-in sampling method based on these r y-values, as discussed in Section 2. The reason for using only the sensitive part of the data for imputing synthetic data is to ensure that in the released data the synthetic part and the non-sensitive part are independent. The synthetic version of $\boldsymbol{y}_1$ is $\boldsymbol{y}_1^* = (y_1^*, \ldots, y_r^*)'$ such that $y_i^* \sim N(x_i' b_1, \hat{\sigma}_1)$ generated independently for $i = 1, \ldots, r$, where $\boldsymbol{b}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1$ and $\text{RSS}_1 = \boldsymbol{y}_1'(\boldsymbol{I}_r - P_{\boldsymbol{X}_1})\boldsymbol{y}_1$ are the sufficient statistics of $\boldsymbol{y}_1$, and $\hat{\sigma}_1^2 = \text{RSS}_1/(r - p)$. We assume that $r > p$ and $n - r > p$ so that we can draw valid inference about the p regression coefficients β separately for each data set. Thus similarly, $\boldsymbol{b}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}_2$ and $\text{RSS}_2 = \boldsymbol{y}_2'(\boldsymbol{I}_{n-r} - P_{\boldsymbol{X}_2})\boldsymbol{y}_2$ are the sufficient statistics of $\boldsymbol{y}_2$. The released data is $\boldsymbol{y}^* = (\boldsymbol{y}_1^{*\prime}, \boldsymbol{y}_2')'$. Then by Lemma 1.1 the sufficient statistics for the imputed data are

$$\boldsymbol{b}_1^* = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1^* \overset{d}{=} b_1 + \hat{\sigma}_1 C_1 U_0 \overset{d}{=} \beta + \sigma\sqrt{1 + \psi}\, C_1 U_1$$

$$\text{RSS}_1^* = \boldsymbol{y}_1^{*\prime}(\boldsymbol{I}_r - P_{\boldsymbol{X}_1})\boldsymbol{y}_1^* \overset{d}{=} \hat{\sigma}_1^2 \boldsymbol{W}_1'(\boldsymbol{I}_r - P_{\boldsymbol{X}_1})\boldsymbol{W}_1 \overset{d}{=} \sigma^2 \psi V_1$$

where $U_0, U_1 \sim N_p(0, \boldsymbol{I_p})$ independently, $C_1\boldsymbol{C}_1' = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$, $\psi = (\hat{\sigma}_1/\sigma)^2$ is a latent quantity, $\boldsymbol{W}_1 \sim N_r(0, \boldsymbol{I_r})$ and $V_1 \sim \chi^2_{r-p}$. Now suppose we represent $\boldsymbol{b}_1^* = B\boldsymbol{y}_1^*$, $\text{RSS}_1^* = \boldsymbol{y}_1^{*\prime} A \boldsymbol{y}_1^*$ and $\boldsymbol{y}_1^* \sim N_r(\boldsymbol{X}_1\hat{\beta}_1, \Sigma)$, then $\boldsymbol{b}_1^*$ is independent of RSS_1^* since $\boldsymbol{B\Sigma A} = \hat{\sigma}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{I}_r - P_{\boldsymbol{X_1}}\right) = \boldsymbol{O}$. Thus the likelihood based on the released data for the parameters $\theta = (\beta, \sigma^2, \psi)$ is given by

$$\mathcal{L}(\beta, \sigma^2, \psi | \boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{y}_2) = \phi_{\boldsymbol{p}}(\boldsymbol{b}_1^*; \beta, \sigma^2(1+\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} h(\text{RSS}_1^*; r - p, \sigma^2\,\psi)\phi_{n-r}(\boldsymbol{y}_2;\, \boldsymbol{X}_2\beta, \sigma^2\boldsymbol{I}_{n-r})$$

The prior distribution on the parameters is given by for $\delta > 0$

$$\pi(\beta, \sigma^2, \psi) = \pi(\beta)\pi(\sigma^2)\pi(\,\psi) \;\propto\; (\sigma^2)^{-\frac{\delta+1}{2}} \psi^{\frac{r-p}{2}-1} e^{\frac{-(r-p)\psi}{2}}$$

The posterior distribution can be computed in the following manner:

$$\pi(\beta, \sigma^2, \psi | \boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{y}_2) \;\propto\; \mathcal{L}(\beta, \sigma^2, \psi | \boldsymbol{b}_1^*, \text{RSS}_1^*)\, \pi(\,\psi)\pi(\beta, \sigma^2)$$

$$\pi(\beta, \sigma^2, \psi | \boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{y}_2) = \pi(\beta\,|\boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{y}_2, \sigma^2, \psi)\pi(\sigma^2\,|\boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{y}_2, \psi)\,\pi(\psi\,|\boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{y}_2)$$

The conditional posteriors are as follows

$$\beta|\sigma^2, \psi, \boldsymbol{b}_1^*, \boldsymbol{b}_2 \sim N_p\left[\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi}\boldsymbol{b}_1^* + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\right), \sigma^2\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\right]$$

$$\sigma^2|\,\psi, \boldsymbol{b}_1^*, \mathbf{RSS}_1^*, \boldsymbol{b}_2, \mathbf{RSS}_2 \sim \text{Scale} - \text{inv} - \chi^2\left[n - p + \delta - 1, \frac{1}{n-p+\delta-1}\left(\frac{\text{RSS}_1^*}{\psi} + \text{RSS}_2 + (\boldsymbol{b}_1^* - \boldsymbol{b}_2)'\left((1+\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^* - \boldsymbol{b}_2)\right)\right]$$

$$\pi(\psi|\, \boldsymbol{b}_1^*, \mathrm{RSS}_1^*, \boldsymbol{b}_2, \mathrm{RSS}_2)$$
$$\propto \left|\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right|^{-\frac{1}{2}} (1+\psi)^{-\frac{p}{2}} \psi^{-1} e^{\frac{-(r-p)\psi}{2}}$$
$$\times \left\{(\boldsymbol{b}_1^* - \boldsymbol{b}_2)'\left((1+\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^* - \boldsymbol{b}_2)\right.$$
$$\left. + \frac{\mathrm{RSS}_1^*}{\psi} + \mathrm{RSS}_2\right\}^{-\frac{n-p+\delta-1}{2}}$$

We see that the expressions match the case when all of $\boldsymbol{y}$ is sensitive as in Section 2 by deleting all quantities involving $\boldsymbol{y}_2$, $\boldsymbol{X}_2$; replacing $\boldsymbol{X}_1$ by $\boldsymbol{X}$, $\boldsymbol{b}_1^*$ by $\boldsymbol{b}^*$ and r by n. The posterior distributions are proper as long as $r > p$, $n > \max\{r + p, p - \delta + 1\}$.

Now as $\pi(\psi)$ we use this shorthand from here on) is a non-standard pdf, we devise a sampling scheme below using the Accept-Reject method. Let us denote

$$Q(\psi) = (\boldsymbol{b}_1^* - \boldsymbol{b}_2)'\left((1+\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^* - \boldsymbol{b}_2) + \frac{\mathrm{RSS}_1^*}{\psi} + \mathrm{RSS}_2$$

$$\boldsymbol{X}'\boldsymbol{X}_\psi = \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2$$

We notice that, if we had started with only the sole assumption $r > p$, $\boldsymbol{X}'\boldsymbol{X}_\psi > 0\ \forall\ \psi > 0$ (as it is a covariance matrix), then letting $\psi \to \infty$ would yield $\boldsymbol{X}_2'\boldsymbol{X}_2 > 0$ and thus $n - r > p$, necessitating both of those assumptions in the first place. Now turning our attention to $Q(\psi)$, we see that $Q(\psi) > 0\ \forall\ \psi > 0$ by definition and also by design. Since the r.v.'s $(\boldsymbol{b}_1^*, \boldsymbol{b}_2, \mathbf{RSS}_1^*, \mathbf{RSS}_2)$ embroiled in the expression of $Q(\psi)$ are mutually independent, $Q(\psi) > 0$ even when $\mathbf{RSS}_1^*$ is arbitrarily small, hence $(\psi) \geq \frac{\mathbf{RSS}_1^*}{\psi}$. This coupled with the fact that $\boldsymbol{X}'\boldsymbol{X}_\psi = \frac{\boldsymbol{X}'\boldsymbol{X}}{1+\psi} \Rightarrow \left|\boldsymbol{X}'\boldsymbol{X}_\psi\right| > (1+\psi)^{-\frac{p}{2}} |\boldsymbol{X}'\boldsymbol{X}|$ (as $A > B \Rightarrow \lambda_i(B)\ \forall\ i = 1,\ldots,$ n where $\{\lambda_i(A) : i = 1, \ldots, n\}$ and $\{\lambda_i(B) : i = 1, \ldots, n\}$ are the ordered eigenvalues of $n \times n$ PD matrices A and B respectively) produces $\pi(\psi) \leq Lg(\psi)$ where

$$L = \frac{|\boldsymbol{X}'\boldsymbol{X}|^{-\frac{1}{2}}\, 2^{n-p+\delta-1}\, \Gamma\left(\frac{n-p+\delta-1}{2}\right)}{\left((r-p)\mathbf{RSS}_1^*\right)^{\frac{n-p+\delta-1}{2}}}$$

and $g(\psi)$ is the pdf of a $\frac{\chi^2_{n-p+\delta-1}}{r-p} \equiv \Gamma\left(\frac{n-p+\delta-1}{2}, \frac{r-p}{2}\right)$ r.v.

Algorithm for sampling from $\pi(\psi)$:

1. We have the *i*-th sample $\psi^{(i)}$.
2. Draw a sample $\psi' \sim g(\psi)$ where $g(\psi) \sim \frac{\chi^2_{n-p+\delta-1}}{r-p}$ and also draw $u \sim \boldsymbol{U}[0; 1]$.
3. If $u \sim \frac{\pi(\psi)}{Lg(\psi)}$ then $\psi^{(i+1)} = \psi'$, else discard ψ' and go back to step 2.

Theorem 5.1. The joint pdf of $(\boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{b}_2, \text{RSS}_2)$ is given by

$$f_{\beta,\sigma^2}(\boldsymbol{b}_1^*, \text{RSS}_1^*, \boldsymbol{b}_2, \text{RSS}_2)$$

$$\propto \int_0^\infty \phi_{\boldsymbol{p}}[\beta; \left[\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi}\boldsymbol{b}_1^* + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\right), \sigma^2\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\right]$$

$$\times \frac{(\text{RSS}_1^*)^{\frac{r-p}{2}-1}(\text{RSS}_2)^{\frac{n-r-p}{2}-1}}{(\sigma^2)^{\frac{n-p}{2}}} e^{-\frac{1}{2\sigma^2}\left|(\boldsymbol{b}_1^*-\boldsymbol{b}_2)'\left((1+\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}+(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^*-\boldsymbol{b}_2)+\frac{\text{RSS}_1^*}{\psi}+\text{RSS}_2\right|}$$

$$\times \left|\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right|^{-\frac{1}{2}} (1+\psi)^{-\frac{p}{2}} \psi^{-1} e^{\frac{-(r-p)\psi}{2}} d\psi$$

Posterior Predictive Sampling

We similarly synthesize $\boldsymbol{r}$ sensitive y-values $\boldsymbol{y}_1$ by applying the posterior predictive sampling method based on these $\boldsymbol{r}$ *y*-values, as discussed in Section 3. The synthetic version of $\boldsymbol{y}_1$ is $\boldsymbol{y}_1^* = (y_1^*, \ldots, y_r^*)'$ such that $y_i^* \sim N\left(\boldsymbol{x}_i'\beta_1^*, {\sigma_1^*}^2\right)$ generated independently for i = 1,…, r, where following from equations (12) and (13), $\left(\beta_1^*, {\sigma_1^*}^2\right)$ are drawn from the imputed posterior

$$\sigma_1^2 \mid \boldsymbol{b}_1, \mathbf{RSS}_1 \sim \text{Scale} - \text{inv} - \chi^2\left(r - p + \alpha - 1, \frac{\mathbf{RSS}_1}{(r - p + \alpha - 1)}\right)$$

$$\beta_1 \mid \boldsymbol{b}_1, \mathbf{RSS}_1, \sigma_1^2 \sim N_p(\boldsymbol{b}_1, \sigma_1^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})$$

where we assume throughout that $r + \alpha > p + 1$.

Then the sufficient statistics for the imputed data are

$$\boldsymbol{b}_1^* \overset{d}{=} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1^* = \beta_1^* + \sigma_1^* C_1 U_0 = \beta + \sigma\sqrt{1+\psi} C_1 U_1$$

$$\mathrm{RSS}_1^* \overset{d}{=} \boldsymbol{y}_1^{*\prime}\left(\boldsymbol{I}_r - P_{\boldsymbol{X_1}}\right)\boldsymbol{y}_1^* = \hat{\sigma}_1^{*2}\boldsymbol{W}_1'\left(\boldsymbol{I}_r - \boldsymbol{P}_{\boldsymbol{X_1}}\right)\boldsymbol{W}_1 = \sigma^2\,\psi V_1$$

where $U_0, U_1 \sim N_p(0, \boldsymbol{I_p})$ independently, $C_1\boldsymbol{C}_1' = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$, $\psi = (\sigma_1^*/\sigma)^2$ is a latent quantity, $\boldsymbol{W}_1 \sim N_r(0, \boldsymbol{I_r})$ and $V_1 \sim \chi^2_{r-p}$. Next we can basically adapt the same procedure as before to obtain the conditional posteriors as follows

$$\beta|\sigma^2, \psi, \boldsymbol{b}_1^*, \boldsymbol{b}_2 \sim N_p\left[\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi}\boldsymbol{b}_1^* + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\right), \sigma^2\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\right]$$

$$\sigma^2|\,\psi, \boldsymbol{b}_1^*, \mathbf{RSS}_1^*, \boldsymbol{b}_2, \mathbf{RSS}_2 \sim \mathrm{Scale-inv}$$
$$-\chi^2\left[n-p+\delta\right.$$
$$\left.-1, \frac{1}{n-p+\delta-1}\left(\frac{\mathrm{RSS}_1^*}{\psi} + \mathrm{RSS}_2\right.\right.$$
$$\left.\left.+ (\boldsymbol{b}_1^* - \boldsymbol{b}_2)'\left((1+\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^* - \boldsymbol{b}_2)\right)\right]$$

$$\pi(\psi|\,\boldsymbol{b}_1^*, \mathrm{RSS}_1^*, \boldsymbol{b}_2, \mathrm{RSS}_2)$$
$$\propto \left|\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right|^{-\frac{1}{2}}(1+2\psi)^{-\frac{p}{2}}(1+\psi)^{-\frac{2r-2p+\alpha-1}{2}}\psi^{-1}$$
$$\times\left\{(\boldsymbol{b}_1^* - \boldsymbol{b}_2)'\left((1+2\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^* - \boldsymbol{b}_2)\right.$$
$$\left.+ \frac{\mathrm{RSS}_1^*}{\psi} + \mathrm{RSS}_2\right\}^{-\frac{n-p+\delta-1}{2}}$$

The posterior distributions are proper as long as $r > \max\left\{p, p-\alpha+1, \frac{n+p-\alpha+\delta}{2}\right\}$, $n > \max\{r+p, p-\delta+1\}$ and expressions align as well with our results in Section 3 when *r* = *n*.

Algorithm for sampling from $\pi(\psi)$:

1. We have the *i*-th sample $\psi^{(i)}$.
2. Draw a sample $\psi' \sim g(\psi)$ where $g(\psi) \sim \beta'\left(\frac{n-p+\delta-1}{2}, \frac{2r-n-p+\alpha-\delta}{2}\right)$ and also draw $u \sim \boldsymbol{U}[0; 1]$. This necessitates the assumption $2r + \alpha > n + p + \delta$.
3. If $u \leq \frac{\pi(\psi)}{Lg(\psi)}$ then $\psi^{(i+1)} = \psi'$, else discard ψ' and go back to step 2. Here

$$L = \frac{|\boldsymbol{X}'\boldsymbol{X}|^{-\frac{1}{2}}\,\mathrm{B}\left(\frac{n-p+\delta-1}{2}, \frac{2r-n-p+\alpha-\delta}{2}\right)}{(\mathrm{RSS}_1^*)^{\frac{n-p+\delta-1}{2}}}$$

where B(a, b) is the Beta function.

Theorem 5.2. The joint pdf of $(\boldsymbol{b}_1^*, \mathrm{RSS}_1^*, \boldsymbol{b}_2, \mathrm{RSS}_2)$ is given by

$$f_{\beta,\sigma^2}(\boldsymbol{b}_1^*, \mathrm{RSS}_1^*, \boldsymbol{b}_2, \mathrm{RSS}_2)$$
$$\propto \int_0^\infty \phi_{\boldsymbol{p}}[\beta;\ \left[\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi}\boldsymbol{b}_1^* + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\right), \sigma^2\left(\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\right]$$
$$\times \frac{(\mathrm{RSS}_1^*)^{\frac{r-p}{2}-1}(\mathrm{RSS}_2)^{\frac{n-r-p}{2}-1}}{(\sigma^2)^{\frac{n-p}{2}}}\, e^{-\frac{1}{2\sigma^2}\left|(\boldsymbol{b}_1^*-\boldsymbol{b}_2)'\left((1+2\psi)(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}+(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\right)^{-1}(\boldsymbol{b}_1^*-\boldsymbol{b}_2)+\frac{\mathrm{RSS}_1^*}{\psi}+\mathrm{RSS}_2\right|}$$
$$\times \left|\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi} + \boldsymbol{X}_2'\boldsymbol{X}_2\right|^{-\frac{1}{2}} (1+2\psi)^{-\frac{p}{2}}(1+\psi)^{-\frac{2r-2p+\alpha-1}{2}}\psi^{-1}\,d\psi$$

Method II: Using whole data estimates to impute synthetic data

Plug-In Sampling

We can relax the assumption $n - r > p$ needed before if we use estimates of the entire data to impute the r sensitive y-values $\boldsymbol{y}_1$. In the case, the synthetic version of $\boldsymbol{y}_1$ is $\boldsymbol{y}_1^* = (y_1^*, \ldots, y_r^*)'$ such that $y_i^* \sim N(\boldsymbol{x}_i'b, \mathrm{RSS})$ generated independently for i = 1,…, r and $\boldsymbol{y}_2$ is defined as before. The likelihood of the released data is proportional to what follows below, since we only retain quantities containing parameters (β, σ^2) necessary for posterior distribution calculation, also using the fact that $\boldsymbol{y}_2|b$, RSS is independent of (β, σ^2) by the definition of sufficient statistic

$$\pi(\boldsymbol{y}_1^*, \boldsymbol{y}_2 \mid \beta, \sigma^2) =$$
$$\int \pi(\boldsymbol{y}_1^*, \boldsymbol{y}_2 \mid \boldsymbol{b}, \mathrm{RSS})\, \pi(\boldsymbol{b}, \mathrm{RSS} \mid \beta, \sigma^2) db\, d\mathrm{RSS} =$$
$$\int \pi(\boldsymbol{y}_1^*, \mid \boldsymbol{y}_2, \boldsymbol{b}, \mathrm{RSS})\, \pi(\boldsymbol{y}_2 \mid \boldsymbol{b}, \mathrm{RSS})\, \pi(b, \mathrm{RSS} \mid \beta, \sigma^2) db\, d\mathrm{RSS} \propto$$

$$\int \pi(\boldsymbol{y}_1^*, | \, \boldsymbol{b}, \text{RSS}) \; \pi(b | \, \beta, \sigma^2) \; \pi(\text{RSS} \mid \sigma^2) db \; d\text{RSS} \propto \int \frac{1}{(\sigma^2\psi)^{r/2}} \exp\left[-\frac{1}{2\sigma^2\psi}(\boldsymbol{y}_1^* - \boldsymbol{X}_1\boldsymbol{b})'(\boldsymbol{y}_1^* - \boldsymbol{X}_1\boldsymbol{b})\right] \times \frac{1}{(\sigma^2)^{p/2}} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{b}-\beta)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b}-\beta)\right] \times \psi^{\frac{n-p}{2}-1} e^{\frac{-(n-p)\psi}{2}} db \; d\psi \tag{21}$$

The last line is due to a change in variable $\text{RSS}/(\text{n}-\text{p})\sigma^2 = \psi$. Next we collect terms for $\boldsymbol{b}$ as

$$\frac{1}{\psi}(\boldsymbol{y}_1^* - \boldsymbol{X}_1\boldsymbol{b})'(\boldsymbol{y}_1^* - \boldsymbol{X}_1\boldsymbol{b}) + (\boldsymbol{b}-\beta)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b}-\beta) = \boldsymbol{b}'\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)\boldsymbol{b} - 2\boldsymbol{b}\left(\boldsymbol{X}'\boldsymbol{X}\beta + \frac{\boldsymbol{X}_1'\boldsymbol{y}_1^*}{\psi}\right) + \frac{\boldsymbol{y}_1^{*\prime}\boldsymbol{y}_1^*}{\psi} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta = \left(\boldsymbol{b} - \left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\beta + \frac{\boldsymbol{X}_1'\boldsymbol{y}_1^*}{\psi}\right)\right)' \left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)\left(\boldsymbol{b} - \left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\beta + \frac{\boldsymbol{X}_1'\boldsymbol{y}_1^*}{\psi}\right)\right) + \frac{\boldsymbol{y}_1^{*\prime}\boldsymbol{y}_1^*}{\psi} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta - \left(\boldsymbol{X}'\boldsymbol{X}\beta + \frac{\boldsymbol{X}_1'\boldsymbol{y}_1^*}{\psi}\right)'\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\beta + \frac{\boldsymbol{X}_1'\boldsymbol{y}_1^*}{\psi}\right) \tag{22}$$

where we know $\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}$ is invertible because $\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi} > 0$ due to $\boldsymbol{X}'\boldsymbol{X} > 0$, $\boldsymbol{X}_1'\boldsymbol{X}_1 > 0, \psi > 0$. The last three quantities above simplify to

$$\beta'\left[\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right] - 2\beta'\left[\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\frac{\boldsymbol{X}_1'\boldsymbol{y}_1^*}{\psi}\right] + \boldsymbol{y}_1^{*\prime}\left[\frac{\boldsymbol{I}_r}{\psi} - \frac{\boldsymbol{X}_1}{\psi}\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\frac{\boldsymbol{X}_1'}{\psi}\right]\boldsymbol{y}_1^* \tag{23}$$

from which it is clear that the (conditional) posterior variance of β is

$$\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} > 0$$

$$\Leftrightarrow (\boldsymbol{X}'\boldsymbol{X})^{1/2}\left[\boldsymbol{I}_p - (\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}}\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}(\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}}\right](\boldsymbol{X}'\boldsymbol{X})^{1/2} > 0$$

$$\Leftrightarrow \boldsymbol{I}_p > (\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}}\left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1}(\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}}$$

$$\Leftrightarrow (\boldsymbol{X}'\boldsymbol{X})^{-1} > \left(\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right)^{-1} \Leftrightarrow \boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi} > \boldsymbol{X}'\boldsymbol{X}$$

$$\Leftrightarrow \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi} > 0 \quad \forall\psi \quad \Leftrightarrow \boldsymbol{X}_1'\boldsymbol{X}_1 > 0 \qquad \Leftrightarrow r > p$$

so that we still have to respect the condition $r > p$ while employing this method. Thus (23) further simplifies to

$$\beta'[(\boldsymbol{X}'\boldsymbol{X})^{-1} - \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]^{-1}\beta \; - 2\beta'[(\boldsymbol{X}'\boldsymbol{X})^{-1} - \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]^{-1}\boldsymbol{b}_1^*$$

$$+\boldsymbol{b}_1^{*\prime}[(\boldsymbol{X}'\boldsymbol{X})^{-1} - \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]^{-1}\boldsymbol{b}_1^* + \frac{\boldsymbol{y}_1^{*\prime}\boldsymbol{y}_1^*}{\psi} - \boldsymbol{b}_1^{*\prime}\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\boldsymbol{b}_1^*$$

$$= (\beta - \boldsymbol{b}_1^*)'[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]^{-1}(\beta - \boldsymbol{b}_1^*) + \frac{\boldsymbol{y}_1^{*\prime}\boldsymbol{y}_1^*}{\psi} - \boldsymbol{y}_1^{*\prime}\frac{\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{X}_1}}{\psi}\boldsymbol{y}_1^*$$

$$= (\beta - \boldsymbol{b}_1^*)'[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]^{-1}(\beta - \boldsymbol{b}_1^*) + \boldsymbol{y}_1^{*\prime}\frac{\left(\boldsymbol{I}_r - \boldsymbol{P}_{\boldsymbol{X}_1}\right)}{\psi}\boldsymbol{y}_1^*$$

$$= (\beta - \boldsymbol{b}_1^*)'[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]^{-1}(\beta - \boldsymbol{b}_1^*) + \frac{\mathrm{RSS}_1^*}{\psi}\boldsymbol{y}_1^* \qquad (24)$$

where $\boldsymbol{b}_1^* = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1^*$, $\mathrm{RSS}_1^* = \boldsymbol{y}_1'\left(\boldsymbol{I}_r - \boldsymbol{P}_{\boldsymbol{X}_1}\right)\boldsymbol{y}_1$ are sufficient statistics for $\boldsymbol{y}_1^*$.

So integrating out $\boldsymbol{b}$ from (21) using (22) and (24) and multiplying by our usual prior $\pi(\beta, \sigma^2) \propto (\sigma^2)^{-\frac{\delta+1}{2}}$ we get the joint posterior distribution to be

$$\begin{aligned}\pi(\beta, \sigma^2, \psi \,|\boldsymbol{y}_1^*, \boldsymbol{y}_2) &\propto \phi_{\boldsymbol{p}}(\beta; \boldsymbol{b}_1^*, \sigma^2[(\boldsymbol{X}'\boldsymbol{X})^{-1} + \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}]) \\ &\times \frac{(\mathrm{RSS}_1^*/\psi)^{\frac{r-p+\delta-1}{2}}}{(\sigma^2)^{\frac{r-p+\delta-1}{2}+1}} \exp\left[\frac{RSS_1^*}{2\sigma^2\psi}\right] \\ &\times |(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &+ \psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}|^{\frac{1}{2}} \left|\boldsymbol{X}'\boldsymbol{X} + \frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right|^{-\frac{1}{2}} \psi^{\frac{n-r-p}{2}-1} e^{-\frac{(n-p)\psi}{2}} \psi^{\frac{r-p+\delta-1}{2}}\end{aligned}$$

where after observing

$$|(\boldsymbol{X}'\boldsymbol{X})^{-1}+\psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}|^{\frac{1}{2}}\left|\boldsymbol{X}'\boldsymbol{X}+\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\psi}\right|^{-\frac{1}{2}}=\psi^{\frac{p}{2}}$$

it leads us to the hierarchical (conditional) posterior distributions as follows

$$\beta|\,\sigma^2,\psi,\boldsymbol{b}_1^*\sim N_p\big(\boldsymbol{b}_1^*,\sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}+\psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})\big)$$

$$\sigma^2|\psi,\mathrm{RSS}_1^*\sim \mathrm{Scale}-\mathrm{inv}-\chi^2\left(r-p+\delta-1,\frac{\mathbf{RSS}_1^*}{\psi(r-p+\delta-1)}\right)$$

$$\psi\ \sim\frac{\chi^2_{n-p+\delta-1}}{n-p}\equiv\Gamma\left(\frac{n-p+\delta-1}{2},\frac{n-p}{2}\right)$$

The posterior distributions are proper as long as $r>\max\{p,p-\delta+1\}$ and the expressions align as well with our results from Section 2 when $r = n$. We notice that an advantage of this method is that the sampling of ψ is straightforward.

Posterior Predictive Sampling

We synthesize $\boldsymbol{y}_1^*=(y_1^*,\ldots,y_r^*)'$ such that $y_i^*\sim N(\boldsymbol{x}_i'\beta^*,\sigma^{*2})$ generated independently for i = 1,…, r, where (β^*,σ^{*2}) are drawn from the imputed posterior given by equations (12) and (13). The likelihood of the released data is given by

$$\pi(\boldsymbol{y}_1^*,\boldsymbol{y}_2\mid\beta,\sigma^2)$$

$$=\int\pi(\boldsymbol{y}_1^*,\boldsymbol{y}_2\mid\beta^*,\sigma^{*2},\boldsymbol{b},\mathrm{RSS})\,\pi(\beta^*,\sigma^{*2}|\boldsymbol{b},\mathrm{RSS})\pi(\boldsymbol{b},\mathrm{RSS}|\beta,\sigma^2)d\beta^*\,d\sigma^{*2}db\,d\mathrm{RSS}$$

$$\propto\int\pi(\boldsymbol{y}_1^*,|\beta^*,\sigma^{*2})\,\pi(\beta^*\,|\sigma^{*2},\boldsymbol{b})\ \pi(\sigma^{*2}|\,\mathrm{RSS})\pi(\boldsymbol{b}|\beta,\sigma^2)\pi(\,\mathrm{RSS}|\sigma^2)\,d\beta^*\,d\sigma^{*2}db\,d\mathrm{RSS}$$

$$\propto\int\frac{1}{(\sigma^{*2})^{r/2}}\exp\left[-\frac{1}{2\sigma^{*2}}(\boldsymbol{y}_1^*-\boldsymbol{X}_1\beta^*)'(\boldsymbol{y}_1^*-\boldsymbol{X}_1\beta^*)\right]$$

$$\times\ \frac{1}{(\sigma^{*2})^{p/2}}\exp\left[-\frac{1}{2\sigma^{*2}}(\beta^*-b)'(\boldsymbol{X}'\boldsymbol{X})(\beta^*-b)\right]\times\frac{\exp\left[-\frac{RSS}{\sigma^{*2}}\right](\mathrm{RSS})^{\frac{n-p+\alpha-1}{2}}}{(\sigma^{*2})^{\frac{n-p+\alpha+1}{2}}}$$

$$\times\frac{1}{(\sigma^{*2})^{p/2}}\exp\left[-\frac{1}{2\sigma^2}(b-\beta)'(\boldsymbol{X}'\boldsymbol{X})(b-\beta)\right]$$

$$\times\ \frac{\exp\left[-\frac{RSS}{2\sigma^2}\right](\mathrm{RSS})^{\frac{n-p}{2}-1}}{(\sigma^2)^{\frac{n-p}{2}}}\,d\beta^*\,d\sigma^{*2}db\,d\mathrm{RSS}$$

We begin by collecting terms for β^* as

$$\begin{aligned}(\boldsymbol{y}_1^* - \boldsymbol{X}_1\beta^*)'(\boldsymbol{y}_1^* - \boldsymbol{X}_1\beta^*) + (\beta^* - b)'(\boldsymbol{X}'\boldsymbol{X})(\beta^* - b)\\ = \beta^{*'}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)\beta^* - 2\beta^*(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}_1'\boldsymbol{y}_1^*) + \boldsymbol{y}_1^{*'}\boldsymbol{y}_1^* + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}\\ = \left(\beta^* - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}_1'\boldsymbol{y}_1^*)\right)'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)\left(\beta^*\right.\\ \left. - (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}_1'\boldsymbol{y}_1^*)\right) + \boldsymbol{y}_1^{*'}\boldsymbol{y}_1^* + b'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}\\ - (\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}_1'\boldsymbol{y}_1^*)'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} + \boldsymbol{X}_1'\boldsymbol{y}_1^*)\end{aligned}$$

After integrating out β^* the likelihood stands at

$$\int \frac{1}{(\sigma^{*2})^{r/2}} \exp\left[-\frac{1}{2\sigma^{*2}}\left(\boldsymbol{y}_1^{*'}\boldsymbol{y}_1^* + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} - \right.\right.$$

$$\left.\left.(\boldsymbol{X}_1'\boldsymbol{y}_1^* + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1'\boldsymbol{y}_1^* + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b})\right)\right] \times \frac{\exp\left[-\frac{RSS}{\sigma^{*2}}\right]}{(\sigma^{*2})^{\frac{n-p+\alpha+1}{2}}} \times \frac{\exp\left[-\frac{RSS}{2\sigma^2}\right]}{(\sigma^2)^{\frac{n-p}{2}}} \times$$

$$(\mathrm{RSS})^{\frac{2n-2p+\alpha-1}{2}-1} \times \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{b} - \beta)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b} - \beta)\, d\sigma^{*2} db\, d\mathrm{RSS}$$

Next we collect terms for $\boldsymbol{b}$ as follows

$$\frac{1}{\sigma^{*2}}\left(\boldsymbol{y}_1^{*'}\boldsymbol{y}_1^* + \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} - (\boldsymbol{X}_1'\boldsymbol{y}_1^* + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}_1'\boldsymbol{y}_1^* + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{b})\right) +$$

$$\frac{1}{\sigma^2}(\boldsymbol{b} - \beta)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b} - \beta) =$$

$$\boldsymbol{b}'\left[(\boldsymbol{X}'\boldsymbol{X})\left(\frac{1}{\sigma^2} + \frac{1}{\sigma^{*2}}\right) - \frac{(\boldsymbol{X}'\boldsymbol{X})}{\sigma^{*2}}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}'\boldsymbol{X})\right]\boldsymbol{b} - 2\boldsymbol{b}'\left[\frac{(\boldsymbol{X}'\boldsymbol{X})}{\sigma^{*2}}(\boldsymbol{X}'\boldsymbol{X} + \right.$$

$$\left.\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1^* + \frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}\beta\right] + \frac{\boldsymbol{y}_1^{*'}\boldsymbol{y}_1^*}{\sigma^{*2}} - \frac{\boldsymbol{y}_1^{*'}}{\sigma^{*2}}\boldsymbol{X}_1(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1^* + \beta'\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}\beta \qquad (25)$$

We can figure out what the variance-covariance matrix will be when we would integrate out $\boldsymbol{b}$, and thus by definition after a change of variable $\sigma^{*2}/\sigma^2 = \psi$ we have

$$\begin{aligned}(1+\psi)\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}'\boldsymbol{X} > 0 \;&\Leftrightarrow (1+\psi)\boldsymbol{I}_p\\ > (\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}(\boldsymbol{X}'\boldsymbol{X})^{\frac{1}{2}} \;&\Leftrightarrow (1+\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1}\\ > (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} \;&\Leftrightarrow \frac{\boldsymbol{X}'\boldsymbol{X}}{(1+\psi)} < \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1\\ &\Leftrightarrow \boldsymbol{X}_1'\boldsymbol{X}_1 + \frac{\psi}{1+\psi}\boldsymbol{X}'\boldsymbol{X} > 0\end{aligned}$$

which is true for all values of $\psi > 0$. We let $\psi \to 0$ to get $\boldsymbol{X}_1'\boldsymbol{X}_1 > 0$, so $r > p$ and

$$(1+\psi)\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}'\boldsymbol{X} = \psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}$$

Here we use the following fact: for any two PD matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$A^{-1} - A^{-1}(A^{-1} + B^{-1})^{-1}A^{-1} = A^{-1}(A^{-1} + B^{-1})^{-1}B^{-1} = (A+B)^{-1} \qquad (26)$$

Next we follow up from (25) to get, after taking out the common factor $\frac{1}{\sigma^2\psi}$

$$\begin{aligned}(b - [\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}\,[\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\,\boldsymbol{X}_1'\boldsymbol{y}_1^* \\ + \psi\boldsymbol{X}'\boldsymbol{X}\beta])'\,[\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]\end{aligned}$$

$$\begin{aligned}(\boldsymbol{b} - [\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}\,[\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\,\boldsymbol{X}_1'\boldsymbol{y}_1^* \\ + \psi\boldsymbol{X}'\boldsymbol{X}\beta]) + \boldsymbol{y}_1^{*\prime}\boldsymbol{y}_1^* - \boldsymbol{y}_1^{*'}\boldsymbol{X}_1(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1^* + \psi\beta'\boldsymbol{X}'\boldsymbol{X}\beta \\ - [\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\,\boldsymbol{X}_1'\boldsymbol{y}_1^* + \psi\boldsymbol{X}'\boldsymbol{X}\beta]'\,[\psi\boldsymbol{X}'\boldsymbol{X} \\ + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}\,[\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\,\boldsymbol{X}_1'\boldsymbol{y}_1^* \\ + \psi\boldsymbol{X}'\boldsymbol{X}\beta]\end{aligned}$$

The last three lines give us by repeated application of (26)

$$\begin{aligned}&\beta'(\psi\boldsymbol{X}'\boldsymbol{X} - \psi\boldsymbol{X}'\boldsymbol{X}[\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}\,\psi\boldsymbol{X}'\boldsymbol{X})\beta - \\ &2\beta'(\psi\boldsymbol{X}'\boldsymbol{X}[\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \\ &\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1)\boldsymbol{b}_1^* + \boldsymbol{y}_1^{*'}(\boldsymbol{I}_r - \boldsymbol{P}_{X_1})\boldsymbol{y}_1^* + \\ &\boldsymbol{b}_1^{*'}(\boldsymbol{X}_1'\boldsymbol{X}_1 - \boldsymbol{X}_1'\boldsymbol{X}_1(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_1)\boldsymbol{b}_1^{*'} - \\ &\boldsymbol{b}_1^{*'}\boldsymbol{X}_1'\boldsymbol{X}_1(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\,\boldsymbol{X}'\boldsymbol{X}[\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}\,\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \\ &\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\,\boldsymbol{X}_1'\boldsymbol{X}_1\boldsymbol{b}_1^* = \beta'\left(\frac{1}{\psi}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)^{-1}\beta\ - \\ &2\beta'\left(\frac{1}{\psi}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)^{-1}\boldsymbol{b}_1^* + \mathrm{RSS}_1^* + \boldsymbol{b}_1^{*'}((\boldsymbol{X}'\boldsymbol{X})^{-1} + \\ &(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}\,\boldsymbol{b}_1^* - \boldsymbol{b}_1^{*'}((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}\,[\psi\boldsymbol{X}'\boldsymbol{X} + ((\boldsymbol{X}'\boldsymbol{X})^{-1} + \\ &(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}]^{-1}((\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}\boldsymbol{b}_1^* = (\beta - \boldsymbol{b}_1^*)'\left(\frac{1}{\psi}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \right. \\ &\left.(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)^{-1}(\beta - \boldsymbol{b}_1^*) + \mathrm{RSS}_1^* \qquad (27)\end{aligned}$$

We integrate out $\boldsymbol{b}$ to find the likelihood to be

$$\int \frac{1}{(\sigma^2)^{p/2}} \exp\left[-\frac{1}{2\sigma^2}\,(\beta-\boldsymbol{b}_1^*)'\left((1+\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1}+\psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)^{-1}(\beta-\boldsymbol{b}_1^*)\right]$$
$$\times \frac{\exp\left[-\frac{\mathrm{RSS}_1^*}{2\sigma^2\psi}\right]}{(\psi)^{\frac{n+r-2p+\alpha+1}{2}}} \times \frac{\exp\left[-\frac{\mathrm{RSS}_1^*}{2\sigma^2}\left(1+\frac{1}{\psi}\right)\right]}{(\sigma^2)^{\frac{2n+r-2p+\alpha+1}{2}-1}} \times (\mathrm{RSS})^{\frac{2n-2p+\alpha-1}{2}-1}$$
$$\times (\sigma^2)^{\frac{p}{2}}|\psi\boldsymbol{X}'\boldsymbol{X}+((\boldsymbol{X}'\boldsymbol{X})^{-1}+(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}|^{-\frac{1}{2}}\,d\psi\,d\mathrm{RSS}$$

Next integrating out RSS we have

$$\int \frac{1}{(\sigma^2)^{p/2}} \exp\left[-\frac{1}{2\sigma^2}\,(\beta-\boldsymbol{b}_1^*)'\left((1+\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1}+\psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)^{-1}(\beta-\boldsymbol{b}_1^*)\right]$$
$$\times \frac{\exp\left[-\frac{\mathrm{RSS}_1^*}{2\sigma^2\psi}\right]}{(\psi)^{\frac{n+r-2p+\alpha+1}{2}}} \times \frac{\left(\frac{\sigma^2\psi}{1+\psi}\right)^{\frac{2n-2p+\alpha-1}{2}}}{(\sigma^2)^{\frac{2n+r-2p+\alpha+1}{2}-1}}$$
$$\times\ (\sigma^2)^{\frac{p}{2}}|\psi\boldsymbol{X}'\boldsymbol{X}+((\boldsymbol{X}'\boldsymbol{X})^{-1}+(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1})^{-1}|^{-\frac{1}{2}}\,d\psi$$

Finally multiplying the integrand by our regular prior $\pi(\beta,\sigma^2) \propto (\sigma^2)^{-\frac{\delta+1}{2}}$, we find that the product breaks up into exactly three parts corresponding to the following posterior distributions

$$\beta|\,\sigma^2,\psi,\boldsymbol{b}_1^* \sim N_p\left(\boldsymbol{b}_1^*,\sigma^2\left((1+\psi)(\boldsymbol{X}'\boldsymbol{X})^{-1}-\psi(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)\right) \tag{28}$$

$$\sigma^2|\psi,\mathrm{RSS}_1^* \sim \mathrm{Scale}-\mathrm{inv}-\chi^2\left(r-p+\delta-1,\frac{\mathbf{RSS}_1^*}{\psi(r-p+\delta-1)}\right) \tag{29}$$

$$\psi\ \sim \beta'\left(\frac{n-p+\delta-1}{2},\frac{n-p+\alpha-\delta}{2}\right) \tag{30}$$

The posterior distributions are proper as long as $r > \max\{p, p-\delta+1, p-\alpha+1\}, n > p-\alpha+\delta$ and they match our results in Section 3 when $r = n$.

All the conditions for existence throughout this work can also be expressed as inequalities for δ, since once we have the data at hand, that would enable us to choose a proper value of δ to get the best inference.

Remark 5.1. *We can think of an r based decision rule to analyze synthetic MLR data as follows. If r < p we ignore the part of the data that is sensitive and base our analysis only on the non-sensitive part. This makes sense in the light of our simulation data where n >> p. If r > p, then we use Method II (use whole data estimates to impute synthetic data). If r > p, n - r > p then we use Method I (use sensitive part estimates to impute synthetic data). If r = n then we use our regular methods of analyses outlined in Sections 2 and3.*

6. Discussion

In this paper, we have developed model based Bayesian inference based on a singly imputed partially synthetic dataset, generated via plug-in sampling, or posterior predictive sampling, under the multiple linear regression model. The methods developed here have the desirable property that they are exact, and based on sufficient statistics. Furthermore, these methods allow a data user to draw valid inference when (perhaps due to privacy concerns or limitations in resources) a statistical agency can only release a single synthetic dataset instead of multiple synthetic copies. The simulation studies presented in Section 4 illustrate that these methods perform just as our theory predicts. It should be noted that the methodology developed here is model based, and thus it does not immediately generalize to cases that do not fall under the multiple linear regression model. In other cases, such as when there are a mixture of continuous and categorical variables, it may very well be possible to derive analogous methods for analyzing singly imputed partially synthetic data, and we hope to pursue this problem in future work.

In what follows, we outline some directions for future research. We have used a non-informative diffuse prior here, and it would be interesting to examine how the inference is affected by the choice of other (non-informative) priors, probability matching priors, and also prions which are conjugate in nature with a suitable choice of the hyperparameters so as not to affect data influence. The simulation results in Section 4 confirm our theoretical results. It is clear from the simulation results in the preceding chapters that the coverage is a decreasing function of δ. It is desirable to express the nature of this dependence exactly, or even within bounds. We would also like to apply this methodology to real-life data to verify our results. We can also look into construction of highest posterior density (HPD) sets of the parameters discussed in various chapters of this dissertation. This will necessitate a judicial choice of the cut-off points of the proposed credible sets.

In deriving the methodology of Sections 2 and 3, we have made assumptions about the process that generated the original data, and about the mechanism used to create synthetic data. Indeed, these assumptions are used to derive the Bayesian inference for singly imputed synthetic data. We leave it as future work to explore the performance of our methodology when some of the conditions do not hold (i.e., scenarios where the imputer and/or data analyst overfit or underfit the regression model; and a scenario where the imputer's model is the regression of y on x, but the data analyst's model is the regression of x on y). Another future research topic would be to consider extensions of our methodology to non-ideal situations that frequently mar real life data (for e.g., non-normal errors; y's have unequal variances and/or are correlated; the original data are from a census, not a

sample; only part of y is sensitive; response and covariates are all sensitive; original data contain missing observations and so on). We would like to point out that the case of partially sensitive data has been addressed in Section 5.

Since one of our prime objectives is to provide valid inference while protecting privacy, we would like to devise methods to quantify privacy in the synthetic data (for e.g., Disclosure Risk Analysis as discussed in Klein and Sinha (2015b)) and observe the trade-off between quality of inference and privacy of survey respondents. It is worth mentioning here that since the data generating methods are still the same as in the frequentist case, the disclosure risk is the same for the cases considered here as in Klein and Sinha (2015a), Klein and Sinha (2015b), and Klein, Zylstra and Sinha (2019).

An excellent new direction of research would be to go beyond the MLR model, and to develop both frequentist and Bayesian analysis of singly and multiply imputed data under a GLM framework. Bayesian analysis of Noise Multiplied data (Klein, Mathew and Sinha, 2014) will also be quite relevant.

Acknowledgments: The authors thank Dr. Gauri Sankar Dutta for reviewing a previous version of the manuscript and Dr. Tommy Wright for helpful comments and encouragement.

References

[1] Abowd, J., Stinson, M., and Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical Report.

[2] Abramowitz, M. and Stegun, I. A. (1972). Handbook of Mathematical Functions, Dover.

[3] An, D. and Little, R. J. A. (2007). Multiple imputation: an alternative to top coding for statistical disclosure control, Journal of the Royal Statistical Society: Series A, 170(4), 923-940.

[4] Banerjee, S., Roy, A. (2014). Linear Algebra and Matrix Analysis for Statistics, CRC Press.

[5] Benedetto, G., Stinson, M. H., and Abowd, J. M. (2013). The Creation and Use of the SIPP Synthetic Beta, Technical Report.

[6] Datta, G., and Ghosh, J. K. (1995). On Priors Providing Frequentist Validity for Bayesian Inference, Biometrika, 82(1), 37-45.

[7] Drechsler, J. (2010). Generating Multiply Imputed Synthetic Datasets: Theory and Implementation, Ph.D. Thesis, Otto-Friedrich-University Bamberg.

[8] Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control, Springer, New York.

[9] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). Bayesian Data Analysis, CRC Press.

[10] Hawala, S. (2008). Producing Partially Synthetic Data to Avoid Disclosure, Proceedings of the Joint Statistical Meetings, American Statistical Association.

[11] Harville, D. A. (1997). Matrix Algebra From a Statistician's Perspective , Springer.

[12] 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.

[13] Kinney, S. K., Reiter, J. P., and Miranda, J. (2014). SynLBD 2.0: Improving the Synthetic Longitudinal Business Database, Statistical Journal of the International Association for Official Statistics, 30, 129-135.

[14] Klein M. D., Mathew, T., and Sinha, B. K. (2014). Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-normal Regression Samples, Journal of Privacy and Confidentiality, 6(1).

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

[16] Klein, M. D., and Sinha, B. K. (2015). Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models, Journal of Privacy and Confidentiality, 7(1).

[17] Klein, M. Zylstra J. and Sinha, B. K. (2019). Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model, Calcutta Statistical Association Bulletin, 71(2), 63-82.

[18] Little, R. J. A. (1993). Statistical Analysis of Masked Data, Journal of Official Statistics, 9, 407-426.

[19] Little, R. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap, The American Statistician, 60(3), 213-223.

[20] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory Meets Practice on the Map. IEEE 24th International Conference on Data Engineering, 277-286.

[21] Martino, L., Garc´ıa, D. L., Arenas, J.M. (2018). Independent Random Sampling Methods, Springer.

[22] Moura, R. (2016). Likelihood-based Inference for Multivariate Regression Models using Synthetic Data, Ph.D. thesis, NOVA University of Lisbon.

[23] Mukhopadhyay, N. (2000). Probability and Statistical Inference, Marcel Dekker Inc.

[24] Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution.

[25] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple Imputation for Statistical Disclosure Limitation, Journal of Official Statistics, 19, 1-16.

[26] Reiter, J. P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets, Survey Methodology, 29, 181-188.

[27] Reiter, J. P. (2005). Significance Tests for Multi-Component Estimands From Multiply Imputed, Synthetic Microdata, Journal of Statistical Planning and Inference, 131, 365-377.

[28] 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.

[29] Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician, The Annals of Statistics, 12(4), 1151-1172.

[30] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys, Wiley.

[31] Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation, Journal of O_cial Statistics, 9, 461-468.

[32] Tweedie, M. C. K. (1957). Statistical Properties of Inverse Gaussian Distributions I, Annals of Mathematical Statistics, 28(2), 362-377.