RESEARCH REPORT SERIES
(Statistics #2022-02)

Bayesian Analysis of Multiply Imputed Synthetic Data Under the Multiple Linear Regression Model

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

[1]Department of Mathematics and Statistics, University of Maryland Baltimore County;
[2]Center for Statistical Research and Methodology, U.S. Census Bureau

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

Report Issued: April 4, 2022

Bayesian Analysis of Multiply Imputed Synthetic Data under the Multiple Linear Regression Model

Abhishek Guin[1], Anindya Roy[1,2] and Bimal Sinha[1,2,*]
[1]University of Maryland Baltimore County, [2]U.S. Census Bureau

Abstract

In this paper we consider Bayesian inference of model parameters in a multiple linear regression model when the response variable is sensitive and the covariates are not, analysis being carried out based on multiple synthetic versions of the response variable. Two scenarios of synthetic data generation are considered - plug-in sampling method and posterior predictive sampling method. We also consider the case when part of the response is sensitive and describe how to carry out full Bayesian analysis based on multiply imputed data.
Keywords: Credible sets, Partially sensitive response, Privacy

1 Introduction

For many statistical agencies such as the US Census Bureau it is customary to publish statistical analysis of results collected from surveys as well as the original raw microdata so others can either reproduce the analysis results or can do some further statistical analysis depending on their purpose. When the response data is sensitive or confidential, a redirect release of such data is not possible and statistical agencies often release what is known as a synthetic version of the original microdata. Fortunately, there are several ways to accomplish this goal and valid statistical analysis of such synthetic data is also quite often possible. In fact, there is a rich literature addressing methods to generate synthetic data and also the appropriate statistical methods to analyze such synthetic data, primarily based on a parametric model which is believed to generate the original data [7, 6, 5, 4, 3, 2]. We mention in passing that in this paper we deal with what is known as *partially synthetic data* rather than fully synthetic data [8].

Under the assumption of a classical linear regression model (single or multiple), valid inference about the regression coefficients and residual variance based on synthetic data has been successfully established based on both single and multiple imputations [5, 4]. Broadly, two types of synthetic data generation schemes have been considered - plug-in sampling method and posterior predictive sampling method. Inference about the regression coefficients from a frequentist point of view was extensively studied in [5, 4] and [2], based on both single imputation and multiple imputations. In

*Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, USA, mailto:sinha@umbc.edu

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

this paper we investigate the inference problem from a Bayesian point of view under the multiple linear regression model with multiple synthetic versions of the original data, thus extending the results in [1] who studied the same problem under a single imputation.

Section 2 contains a brief description of synthetic data generation under plug-in sampling and posterior predictive sampling for a very general scenario. In Section 3 we derive Bayesian inferential results for the model parameters (regression coefficients and residual variance) under the plug-in sampling method while Section 4 contains similar inferential results under the posterior predictive sampling method. In Section 5 we consider the case when only a part of the response vector is sensitive (called partially sensitive data) and provide details about how to carry out appropriate inference under both the data generation schemes.

2 Generating Synthetic Data

We consider two ways of generating the $m \geq 1$ synthetic copies of the original data namely, *plug-in sampling* and *posterior predictive sampling*. In the former method, parameter estimates are plugged in the 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 original model to get synthetic data.

Plug-in Sampling. The basic mechanism for generating synthetic data via *plug-in sampling* (PIS) is described as follows: let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ be the original confidential data, which are jointly distributed according to the probability density function (pdf) $f_{\boldsymbol{\theta}}(\boldsymbol{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar or vector) parameter. To generate partially synthetic data, let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ 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 $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$ of the point estimator, is denoted by $f_{\hat{\boldsymbol{\theta}}}$. The singly imputed synthetic data, denoted by $\boldsymbol{Z}$, are then generated by drawing $\boldsymbol{Z}$ from the joint pdf $f_{\hat{\boldsymbol{\theta}}}$. For the multiple imputation case, we draw $m > 1$ samples $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ independently from $f_{\hat{\boldsymbol{\theta}}}$.

Posterior Predictive Sampling. An alternative method to generate partially synthetic data is to use *posterior predictive sampling* (PPS) which proceeds as follows: suppose that $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ are the original data which are jointly distributed according to the pdf $f_{\boldsymbol{\theta}}(\boldsymbol{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar or vector) parameter. Assume a prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, then the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{Y}$ is obtained as $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{Y}) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{Y})$, and used to draw $m \geq 1$ replications $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_m^*$ (known as posterior draws). Next, for each posterior draw of $\boldsymbol{\theta}$, a corresponding replicate of $\boldsymbol{Y}$ is generated, namely $\boldsymbol{Z}_j = (\boldsymbol{z}_{j1}, \ldots, \boldsymbol{z}_{jn})'$ drawn from the pdf $f_{\boldsymbol{\theta}_j^*}(\boldsymbol{X})$ independently for $j = 1, \ldots, m$.

We conclude this chapter with an observation regarding the existence of *sufficient statistics* in the context of synthetic data. The proof is given in [2].

Lemma 2.1. Suppose that when the original data $\boldsymbol{Y}$ are observed, $T(\boldsymbol{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$. Then when the synthetic data $\mathfrak{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M)$ are observed, $(T(\boldsymbol{Z}_1), \ldots, T(\boldsymbol{Z}_M))$ is jointly sufficient for $\boldsymbol{\theta}$. Furthermore, if $M = 1$, the sufficient statistic is simply $T(\boldsymbol{Z}_1)$, and if $M > 1$, then $\sum_{i=1}^{M} T(\boldsymbol{Z}_i)$ is sufficient if $f_{\boldsymbol{\theta}}(\boldsymbol{Y}) = h(\boldsymbol{Y})\psi(\boldsymbol{\theta}) \exp\{\gamma(\boldsymbol{\theta})' T(\boldsymbol{Y})\}$, i.e., if $f_{\boldsymbol{\theta}}(\boldsymbol{Y})$ belongs to the exponential family.

3 Plug In Sampling method

Consider a standard multiple linear regression (MLR) model involving a *sensitive response* variable y and a $p \times 1$ dimensional vector of *non-sensitive predictors* $\boldsymbol{x}$. We want to generate synthetic data $\boldsymbol{z}_1 = (z_{11}, \ldots, z_{1n})', \ldots, \boldsymbol{z}_m = (z_{m1}, \ldots, z_{mn})'$ for $m > 1$ under PIS. Consider the point estimates $\boldsymbol{b}$ and $\text{RSS}/(n-p)$, of $\boldsymbol{\beta}$ and σ^2, respectively, where $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$ with $\boldsymbol{I}_k$ as the 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 $\boldsymbol{X}$. The synthetic data are obtained by drawing $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \overset{\text{iid}}{\sim} N_n\left(\boldsymbol{X}\boldsymbol{b}, \frac{\text{RSS}}{n-p}\boldsymbol{I}_n\right)$. Equivalently, the synthetic data are obtained by drawing $z_{ji} \sim N(\boldsymbol{x}_i'\boldsymbol{b}, \frac{\text{RSS}}{n-p})$, independently for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

Let $\bar{z}_i = \frac{1}{m}\sum_{j=1}^m z_{ji}$, $S_{zi}^2 = \sum_{j=1}^m (z_{ji} - \bar{z}_i)^2$, and $S_z^2 = \sum_{i=1}^n S_{zi}^2$. If $m > 1$, then conditional on $\boldsymbol{b}$ and RSS,

$$S_z^2 \sim \frac{\text{RSS}}{(n-p)}\chi^2_{n(m-1)}, \quad \bar{z}_i \sim N\left(\boldsymbol{x}_i'\boldsymbol{b}, \frac{\text{RSS}}{m(n-p)}\right),\ i = 1, \ldots, n,$$

with these terms being (conditionally) independent. If $m = 1$, then $\bar{z}_i = z_{1i}$ and $S_{zi}^2 = 0$ for $i = 1, \ldots, n$, and hence $S_z^2 = 0$.

Let $\bar{\boldsymbol{z}} = (\bar{z}_1, \ldots, \bar{z}_n)'$ and $\boldsymbol{b}_j^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}_j$. We define $\overline{\boldsymbol{b}^*} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{z}} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{b}_j^*$ and $S_{\text{comb}}^2 = S_z^2 + m(\bar{\boldsymbol{z}} - \boldsymbol{X}\overline{\boldsymbol{b}^*})'(\bar{\boldsymbol{z}} - \boldsymbol{X}\overline{\boldsymbol{b}^*})$, and conditional on $\boldsymbol{b}$ and RSS,

$$\overline{\boldsymbol{b}^*} \sim N_p\left(\boldsymbol{b}, \frac{\text{RSS}}{m(n-p)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right), \quad S_{\text{comb}}^2 \sim \frac{\text{RSS}}{(n-p)}\chi^2_{n(m-1)+n-p}$$

which are (conditionally) independent and are jointly sufficient for $(\boldsymbol{\beta}, \sigma^2)$. From [2], we have the following result.

Theorem 3.1. The joint pdf of $(\overline{\boldsymbol{b}^*}, S_{\text{comb}}^2)$ is given by

$$\begin{aligned} f_{\boldsymbol{\beta},\sigma^2}(\overline{\boldsymbol{b}^*}, S_{\text{comb}}^2) \quad \propto \quad & \int_0^\infty e^{-\frac{1}{2}\left[\frac{(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})}{\sigma^2(1+\frac{\psi}{m(n-p)})} + \frac{(n-p)S_{\text{comb}}^2}{\sigma^2\psi} + \psi\right]} \frac{(S_{\text{comb}}^2)^{\frac{nm-p}{2}-1}}{\sigma^{nm}\psi^{\frac{n(m-1)+p+2}{2}}} \\ & \times\left[1 + \frac{m(n-p)}{\psi}\right]^{-p/2} d\psi \end{aligned} \tag{1}$$

3.1 Posterior distributions of $\boldsymbol{\beta}$ and σ^2

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

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta})\pi(\sigma^2)$$

where $\pi(\boldsymbol{\beta}) \propto 1$ 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 $\delta > 0$. For doing posterior computation, we use latent variable augmentation. Consider the latent variable $\psi = (\hat{\sigma}/\sigma)^2$ where $\hat{\sigma}^2 = RSS/(n-p)$. Then following the development in the single imputation case in [1], we use the joint distribution of $(\overline{\boldsymbol{b}^*}, S_{\text{comb}}^2)$ conditional on the latent quantity ψ to

perform posterior computation. Multiplying the prior with the conditional likelihood one obtains the posterior distributions as:

$$\begin{aligned}
\boldsymbol{\beta} \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \sigma^2, \psi &\sim N_p\left(\overline{\boldsymbol{b}^*}, \sigma^2(1+\frac{\psi}{m(n-p)})(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) \\
\sigma^2 \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \psi &\sim \text{Scale-inv-}\chi^2\left(nm-p+\delta-1, \frac{(n-p)S^2_{\text{comb}}}{\psi(nm-p+\delta-1)}\right) \\
\psi &\sim \chi^2_{n-p+\delta-1}
\end{aligned} \tag{2}$$

The posterior distributions are proper as long as $n > \max\{p, p-\delta+1\}$ (this also ensures that $nm-p+\delta-1 > 0$ since $m > 1$, which is necessary for the posterior distribution of σ^2 to be proper). For $m = 1$ in the above formula yields the same results as obtained for the singly imputed plug-in sampling case in [1]. From the joint posterior of $(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \psi)$ the Bayes estimators of $\boldsymbol{\beta}$ and σ^2 follows immediately:

$$\hat{\boldsymbol{\beta}}_{\text{BAYES}} = \text{E}(\boldsymbol{\beta} \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}) = \text{E}_\psi \,\text{E}_{\sigma^2}\, \text{E}(\boldsymbol{\beta} \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \sigma^2, \psi) = \text{E}_\psi \,\text{E}_{\sigma^2}(\overline{\boldsymbol{b}^*}) = \overline{\boldsymbol{b}^*}$$

$$\begin{aligned}
\hat{\sigma}^2_{\text{BAYES}} &= \text{E}(\sigma^2 \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}) = \text{E}_\psi \,\text{E}(\sigma^2 \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \psi) = \text{E}_\psi(\tfrac{(n-p)S^2_{\text{comb}})}{\psi(nm-p+\delta-3)}) \\
&= \tfrac{(n-p)S^2_{\text{comb}}}{(nm-p+\delta-3)}\,\text{E}_\psi(\tfrac{1}{\psi}) = \tfrac{(n-p)S^2_{\text{comb}}}{(nm-p+\delta-3)(n-p+\delta-3)}
\end{aligned}$$

3.2 Credible Sets for $\boldsymbol{\beta}$ and σ^2

In this section we describe how to get credible sets for the regression parameters. Let $U \coloneqq \dfrac{1}{\sigma^2}$. Then following (2), we have

$$U \,|\, S^2_{\text{comb}}, \psi \sim \Gamma\left(\frac{nm-p+\delta-1}{2}, \frac{(n-p)S^2_{\text{comb}}}{2\psi}\right) \tag{3}$$

Define $K \coloneqq \frac{S^2_{\text{comb}}}{\sigma^2} = U \times S^2_{\text{comb}}$. Then from (3) and the fact that if $X \sim \Gamma(\alpha, \beta)$ then $cX \sim \Gamma(\alpha, \beta/c)$, it follows that $K \,|\, S^2_{\text{comb}}, \psi \sim \Gamma\left(\frac{nm-p+\delta-1}{2}, \frac{(n-p)}{2\psi}\right)$. Since the right hand side is independent of S^2_{comb}, it follows that

$$K \,|\, \psi \sim \Gamma\left(\frac{nm-p+\delta-1}{2}, \frac{(n-p)}{2\psi}\right). \tag{4}$$

Using (4) and the fact that $\psi \sim \chi^2_{n-p+\delta-1}$, an $(1-\gamma)$ level credible set for σ^2 based on K is

$$\left[\frac{S^2_{\text{comb}}}{b_{n,p,\delta;\gamma}}, \frac{S^2_{\text{comb}}}{a_{n,p,\delta;\gamma}}\right]$$

where $a_{n,p,\delta;\gamma}$ and $b_{n,p,\delta;\gamma}$ are any two constants that satisfy $1-\gamma = P(a_{n,p,\delta;\gamma} \le K \le b_{n,p,\delta;\gamma})$. The length of the credible interval is $S^2_{\text{comb}}\left(\dfrac{1}{a_{n,p,\delta;\gamma}} - \dfrac{1}{b_{n,p,\delta;\gamma}}\right)$. To obtain a credible set for $\boldsymbol{\beta}$, we define $V \coloneqq \frac{(\boldsymbol{\beta}-\overline{\boldsymbol{b}^*})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}-\overline{\boldsymbol{b}^*})}{\sigma^2(1+\frac{\psi}{m(n-p)})}$ Then $V \,|\, \overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \sigma^2, \psi \sim \chi^2_p$ and thus unconditionally $V \sim \chi^2_p$.

Also V is independent of $(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \sigma^2, \psi)$ and thus V is independent of U. If $U^* := \frac{U(n-p)S^2_{\text{comb}}}{\psi}$ then $U^* \,|\, S^2_{\text{comb}}, \psi \sim \chi^2_{nm-p+\delta-1}$, and unconditionally $U^* \sim \chi^2_{nm-p+\delta-1}$. As V is independent of U, it is independent of U^*. Finally we define the pivot for $\boldsymbol{\beta}$ as

$$T^2_m := \frac{(\boldsymbol{\beta} - \overline{\boldsymbol{b}^*})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \overline{\boldsymbol{b}^*})}{S^2_{\text{comb}}}$$

. To derive the posterior distribution of T^2_m, note that, conditionally given ψ,

$$\begin{aligned} T^2_m &= \frac{\sigma^2 V\left(1 + \dfrac{\psi}{m(n-p)}\right)}{\dfrac{\psi U^*}{U(n-p)}} \\ &= \frac{V}{U^*}\left[\frac{1 + \frac{\psi}{m(n-p)}}{\psi}\right](n-p) \\ &\sim \frac{\chi^2_p}{\chi^2_{nm-p+\delta-1}}\left(\frac{1}{m} + \frac{n-p}{\psi}\right) \\ &= \left[\frac{p}{nm-p+\delta-1}\right] F_{p,\, n-p+\delta-1}\left(\frac{1}{m} + \frac{n-p}{\psi}\right) \end{aligned}$$

Hence the pivot for $\boldsymbol{\beta}$ is computed from the the distribution of T^2_m which follows from

$$\begin{aligned} \psi &\sim \chi^2_{n-p+\delta-1} \\ T^2_m \,|\, \psi &\sim \left[\frac{p}{nm-p+\delta-1}\right]\left[\frac{1}{m} + \frac{n-p}{\psi}\right] F_{p,\, nm-p+\delta-1}. \end{aligned}$$

A $(1-\gamma)$ level credible ellipsoid for $\boldsymbol{\beta}$ based on T^2_m is given by

$$\{\boldsymbol{\beta} \;:\; T^2_m \le d_{n,p,\delta,m;\gamma}\}$$

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

$$V_{\boldsymbol{\beta}}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m, \boldsymbol{X}) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)}\left(d_{n,p,\delta,m;\gamma} S^2_{\text{comb}}\right)^{p/2}\left|\boldsymbol{X}'\boldsymbol{X}\right|^{-1/2}.$$

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 linear combination of $\boldsymbol{\beta}$, namely, $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\eta}$ where $\boldsymbol{A}$ is a $k \times p$ dimensional matrix with rank$(\boldsymbol{A}) = k < p$, we define $T^2_{m,\boldsymbol{\eta}} = (\boldsymbol{\eta} - \boldsymbol{A}\overline{\boldsymbol{b}^*})'\{\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}'\}^{-1}(\boldsymbol{\eta} - \boldsymbol{A}\overline{\boldsymbol{b}^*})/S^2_{comb}$, and proceed by noting that*

$$T^2_{m,\boldsymbol{\eta}} \,|\, \psi \sim \left[\frac{k}{nm-p+\delta-1}\right]\left[\frac{1}{m} + \frac{n-p}{\psi}\right] F_{k,\, nm-p+\delta-1} \quad \textit{and} \quad \psi \sim \chi^2_{n-p+\delta-1}$$

4 Posterior Predictive Sampling method

In this section we repeat the analysis done under PIS for PPS. We consider the setup described in Section 3. The synthetic data are generated by repeating the following steps below independently for each $j = 1, \ldots, m$.

(a) Draw $(\boldsymbol{\beta}_j^*, \sigma_j^{*2})$ from the posterior distribution (2).

(b) Draw $\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jn})' \sim N_n(\boldsymbol{X}\boldsymbol{\beta}_j^*, \sigma_j^{*2}\boldsymbol{I}_n)$.

The released synthetic data are $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ along with the matrix of predictor variables $\boldsymbol{X}$. The sufficient statistics for the synthetic data are: $\boldsymbol{b}_j^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}_j$ and $\text{RSS}_j^* = (\boldsymbol{z}_j - \boldsymbol{X}\boldsymbol{b}_j^*)'(\boldsymbol{z}_j - \boldsymbol{X}\boldsymbol{b}_j^*)$, for $j = 1, \ldots, m$. It can be shown that $(\boldsymbol{b}_1^*, \text{RSS}_1^*), \ldots, (\boldsymbol{b}_m^*, \text{RSS}_m^*)$ are jointly sufficient for $(\boldsymbol{\beta}, \sigma^2)$. In view of the sampling mechanism above, the joint distribution of $\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_m^*$, $\text{RSS}_1^*, \ldots, \text{RSS}_m^*$, $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*$, $\sigma_1^{*2}, \ldots, \sigma_m^{*2}$, $\boldsymbol{b}$ and RSS has the following hierarchical structure:

$$
\begin{aligned}
\boldsymbol{b}_j^* \,|\, \text{RSS}_1^*, \ldots, \text{RSS}_m^*, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*, \sigma_1^{*2}, \ldots, \sigma_m^{*2}, \boldsymbol{b}, \text{RSS} &\sim N_p(\boldsymbol{\beta}_j^*,\, \sigma_j^{*2}(\boldsymbol{X}'\boldsymbol{X})^{-1}), \\
\text{RSS}_j^* \,|\, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*, \sigma_1^{*2}, \ldots, \sigma_m^{*2}, \boldsymbol{b}, \text{RSS} &\sim \sigma_j^{*2}\chi^2_{n-p}, \\
\boldsymbol{\beta}_j^* \,|\, \sigma_1^{*2}, \ldots, \sigma_m^{*2}, \boldsymbol{b}, \text{RSS} &\sim N_p(\boldsymbol{b},\, \sigma_j^{*2}(\boldsymbol{X}'\boldsymbol{X})^{-1}), \\
\sigma_j^{*2} \,|\, \boldsymbol{b}, \text{RSS} &\sim \frac{\text{RSS}}{\chi^2_{n-p+\alpha-1}}, \\
\boldsymbol{b} &\sim N_p(\boldsymbol{\beta},\, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}), \\
\text{RSS} &\sim \sigma^2\chi^2_{n-p}.
\end{aligned}
$$

which are generated independently for $j = 1, \ldots, m$, whenever applicable. Hence,

$$
\begin{aligned}
&f(\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_m^*, \text{RSS}_1^*, \ldots, \text{RSS}_m^*, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*, \sigma_1^{*2}, \ldots, \sigma_m^{*2}, \boldsymbol{b}, \text{RSS}) = \\
&\quad \prod_{j=1}^{m} (2\pi\sigma_j^{*2})^{-p/2}\ |\boldsymbol{X}'\boldsymbol{X}|^{1/2}\ \exp\left[-\frac{1}{2\sigma_j^{*2}}(\boldsymbol{b}_j^* - \boldsymbol{\beta}_j^*)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b}_j^* - \boldsymbol{\beta}_j^*)\right] \\
&\quad \times \prod_{j=1}^{m} \frac{(\text{RSS}_j^*)^{\frac{n-p}{2}-1}}{2^{\frac{n-p}{2}}\,\Gamma(\frac{n-p}{2})}(\sigma_j^{*2})^{-(n-p)/2} \exp\left[-\frac{\text{RSS}_j^*}{2\sigma_j^{*2}}\right] \\
&\quad \times \prod_{j=1}^{m} (2\pi\sigma_j^{*2})^{-p/2}\ |\boldsymbol{X}'\boldsymbol{X}|^{1/2}\ \exp\left[-\frac{1}{2\sigma_j^{*2}}(\boldsymbol{\beta}_j^* - \boldsymbol{b})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta}_j^* - \boldsymbol{b})\right] \\
&\quad \times \prod_{j=1}^{m} \frac{(\text{RSS})^{(n-p+\alpha-1)/2}}{2^{(n-p+\alpha-1)/2}\,\Gamma\left(\frac{n-p+\alpha-1}{2}\right)}(\sigma_j^{*2})^{-(n-p+\alpha-1)/2-1} \exp\left[-\frac{\text{RSS}}{2\sigma_j^{*2}}\right] \\
&\quad \times (2\pi\sigma^2)^{-p/2}\ |\boldsymbol{X}'\boldsymbol{X}|^{1/2}\ \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{b} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b} - \boldsymbol{\beta})\right] \\
&\quad \times \frac{(\text{RSS})^{\frac{n-p}{2}-1}}{2^{\frac{n-p}{2}}\,\Gamma(\frac{n-p}{2})}(\sigma^2)^{-(n-p)/2} \exp\left[-\frac{\text{RSS}}{2\sigma^2}\right]
\end{aligned}
$$

Patterned after Theorem 3.1, integrating out $\boldsymbol{\beta}_j^*$'s, $\boldsymbol{b}$, RSS, we get the relevant likelihood as

$$
\begin{aligned}
&L(\boldsymbol{\beta},\sigma^2,\sigma_j^{*2}, j=1,...,m \,|\, \boldsymbol{b}_j^*, \mathrm{RSS}_j^*, j=1,\ldots,m) = \\
&\quad (2\pi)^{-p/2} \frac{\left(\frac{1}{\sigma^2}\right)^{p/2} \left(\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}\right)^{p/2}}{\left(\frac{1}{\sigma^2} + \sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}\right)^{p/2}} \, |\boldsymbol{X}'\boldsymbol{X}|^{1/2} \\
&\quad \times \exp\left[-\frac{1}{2} \frac{\left(\frac{1}{\sigma^2}\right)\left(\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}\right)}{\left(\frac{1}{\sigma^2} + \sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}\right)} \left(\boldsymbol{\beta} - \frac{\sum_{j=1}^m \frac{\boldsymbol{b}_j^*}{2\sigma_j^{*2}}}{\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\boldsymbol{\beta} - \frac{\sum_{j=1}^m \frac{\boldsymbol{b}_j^*}{2\sigma_j^{*2}}}{\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}}\right) \right] \\
&\quad \times \prod_{j=1}^m (2\pi)^{-p/2} \left(2\sigma_j^{*2}\right)^{-p/2} |\boldsymbol{X}'\boldsymbol{X}|^{1/2} \exp\left[-\frac{1}{2} \cdot \frac{1}{2\sigma_j^{*2}} \left(\boldsymbol{b}_j^* - \frac{\sum_{j=1}^m \frac{\boldsymbol{b}_j^*}{2\sigma_j^{*2}}}{\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}}\right)' (\boldsymbol{X}'\boldsymbol{X}) \left(\boldsymbol{b}_j^* - \frac{\sum_{j=1}^m \frac{\boldsymbol{b}_j^*}{2\sigma_j^{*2}}}{\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}}\right) \right] \\
&\quad \times \frac{\left(\sum_{j=1}^m \frac{1}{2\sigma_j^{*2}}\right)^{-\frac{p}{2}} \left(\frac{1}{\sigma^2} + \sum_{j=1}^m \frac{1}{\sigma_j^{*2}}\right)^{-\frac{(m+1)(n-p)+m(\alpha-1)}{2}}}{(\sigma^2)^{\frac{n-p}{2}} \left(\prod_{j=1}^m (\sigma_j^{*2})^{\frac{n-p+\alpha-1}{2}} + \frac{n-p}{2} + 1\right)} \exp\left[-\sum_{j=1}^m \frac{\mathrm{RSS}_j^*}{2\sigma_j^{*2}} \right] \\
&\quad \times \frac{\Gamma\left(\frac{m(n-p+\alpha-1)+(n-p)}{2}\right)}{\left(\Gamma\left(\frac{n-p+\alpha-1}{2}\right)\right)^m \left(\Gamma\left(\frac{n-p}{2}\right)\right)^{m+1}} (2\pi)^{p/2} \, |\boldsymbol{X}'\boldsymbol{X}|^{-1/2} \left(\prod_{j=1}^m \frac{(\mathrm{RSS}_j^*)^{\frac{n-p}{2}-1}}{2^{\frac{n-p}{2}}} \right). \qquad (5)
\end{aligned}
$$

Observe that the quantity inside the exponential in the second line vanishes for $m = 1$. Multiplying the likelihood in (5) by the prior $\pi(\beta, \boldsymbol{\sigma^2}) \propto (\boldsymbol{\sigma^2})^{-\frac{\delta+1}{2}}$, and separating the parameters we get the posteriors as:

$$
\begin{aligned}
\boldsymbol{\beta} \,|\, \sigma^2, \sum_{j=1}^m \frac{\boldsymbol{b}_j^*}{\sigma_j^{*2}}, \sum_{j=1}^m \frac{1}{\sigma_j^{*2}} \quad &\sim \quad N_p \left(\frac{\sum_{j=1}^m \frac{\boldsymbol{b}_j^*}{\sigma_j^{*2}}}{\sum_{j=1}^m \frac{1}{\sigma_j^{*2}}}, \left(1 + \sigma^2 \left(\sum_{j=1}^m \frac{1}{\sigma_j^{*2}}\right)\right) (\boldsymbol{X}'\boldsymbol{X})^{-1} \right), \\
\sigma^2 \left(\sum_{j=1}^m \frac{1}{\sigma_j^{*2}}\right) \;|\; \left(\sum_{j=1}^m \frac{1}{\sigma_j^{*2}}\right) \quad &\sim \quad \beta' \left(\frac{m(n-p+\alpha-1)-\delta+1}{2}, \frac{n-p+\delta-1}{2} \right) \qquad (6)
\end{aligned}
$$

Thus, $\sigma^2 \left(\sum_{j=1}^m 1/\sigma_j^{*2}\right)$ is independent of both latent variables and data. The posterior distributions are proper as long as $n > \max\left\{p, p-\delta+1, p-\alpha+1, p-\alpha+1+\frac{\delta-1}{m}\right\}$. The latent variables

have the following distribution:

$$g(\sigma_1^{*2},\ldots,\sigma_m^{*2}\,|\,\boldsymbol{b}_1^*,\ldots,\boldsymbol{b}_m^*,\mathrm{RSS}_1^*,\ldots,\mathrm{RSS}_m^*)\propto \frac{\left(\prod_{j=1}^m(\sigma_j^{*2})\right)^{-\frac{2n-p+\alpha+1}{2}}}{\left(\sum_{j=1}^m\frac{1}{\sigma_j^{*2}}\right)^{\frac{m(n-p+\alpha-1)+p-\delta}{2}}} \exp\left[-\frac{1}{4}\sum_{j=1}^m\frac{1}{\sigma_j^{*2}}\left\{\left(\boldsymbol{b}_j^*-\frac{\sum_{j=1}^m\frac{\boldsymbol{b}_j^*}{\sigma_j^{*2}}}{\sum_{j=1}^m\frac{1}{\sigma_j^{*2}}}\right)'(\boldsymbol{X}\boldsymbol{X}')\left(\boldsymbol{b}_j^*-\frac{\sum_{j=1}^m\frac{\boldsymbol{b}_j^*}{\sigma_j^{*2}}}{\sum_{j=1}^m\frac{1}{\sigma_j^{*2}}}\right)+2\mathrm{RSS}_j^*\right\}\right].$$

Defining $\sigma_j^2=1/\sigma_j^{*2}$, we have the joint posterior of the transformed latent variables as

$$h(\sigma_1^2,\ldots,\sigma_m^2\,|\,\boldsymbol{b}_1^*,\ldots,\boldsymbol{b}_m^*,\mathrm{RSS}_1^*,\ldots,\mathrm{RSS}_m^*)\propto \frac{\left(\prod_{j=1}^m(\sigma_j^2)\right)^{\frac{2n-p+\alpha-1}{2}}}{\left(\sum_{j=1}^m\sigma_j^2\right)^{\frac{m(n-p+\alpha-1)+p-\delta}{2}}} \exp\left[-\frac{1}{4}\sum_{j=1}^m\sigma_j^2\left\{\left(\boldsymbol{b}_j^*-\frac{\sum_{j=1}^m\sigma_j^2\boldsymbol{b}_j^*}{\sum_{j=1}^m\sigma_j^2}\right)'(\boldsymbol{X}\boldsymbol{X}')\left(\boldsymbol{b}_j^*-\frac{\sum_{j=1}^m\sigma_j^2\boldsymbol{b}_j^*}{\sum_{j=1}^m\sigma_j^2}\right)+2\mathrm{RSS}_j^*\right\}\right]. \tag{7}$$

Let us denote the quantity inside the exponential of (7) as Q. Our goal here is to sample from (7).

4.1 Approach I:

To suitably transform $\sigma_1^2,\ldots,\sigma_m^2$ to draw a sample, our first method is to use

$$v_1=\sigma_1^2, v_2=\frac{\sigma_2^2}{\sigma_1^2},\ldots,v_m=\frac{\sigma_m^2}{\sigma_1^2} \tag{8}$$

so that $(v_1,v_2,\ldots,v_m)$ can be sampled as

$$v_1\,|\,v_2,\ldots,v_m,\mathrm{data}\sim\frac{2\,\chi^2_{nm-p+\delta-1}}{Q},$$
$$\pi(v_2,\ldots,v_m\,|\,\mathrm{data})\le\frac{\mathbf{B}\left(\frac{2n-p+\alpha-1}{2},\ldots,\frac{2n-p+\alpha-1}{2}\right)}{(\mathrm{RSS}^*_{\min})^{\frac{nm-p+\delta-1}{2}}}\,g_{\mathrm{InvDir}}(v_2,\ldots,v_m), \tag{9}$$

where $g_{\mathrm{InvDir}}(v_2,\ldots,v_m)$ is the pdf of an m^{th} order Inverse-Dirichlet $\left(\frac{2n-p+\alpha-1}{2},\ldots,\frac{2n-p+\alpha-1}{2}\right)$ distribution.

4.2 Approach II:

One could also use the following transformation:

$$v_1=\sigma_1^2, v_2=\frac{\sigma_1^2}{\sigma_2^2}, v_3=\frac{\sigma_1^2+\sigma_2^2}{\sigma_3^2},\ldots,v_m=\frac{\sigma_1^2+\cdots+\sigma_{m-1}^2}{\sigma_m^2}, \tag{10}$$

so that $(v_1, v_2, \ldots, v_m)$ can be sampled as

$$v_1 \,|\, v_2, \ldots, v_m, \text{data} \sim \frac{2\,\chi^2_{nm-p+\delta-1}}{Q},$$

$$\pi(v_2, \ldots, v_m \,|\, \text{data}) \leq \frac{\prod_{j=2}^{m} \mathrm{B}\left(\frac{(2n-p+\alpha-1)(j-1)}{2}, \frac{2n-p+\alpha-1}{2}\right)}{(\mathrm{RSS}^*_{\min})^{\frac{nm-p+\delta-1}{2}}} \left(\prod_{j=2}^{m} g'_j(v_j)\right), \tag{11}$$

where $g'_j(v_j)$ is the pdf of a Beta-Prime $\left(\frac{(2n-p+\alpha-1)(j-1)}{2}, \frac{2n-p+\alpha-1}{2}\right)$ distribution, independently for $j = 2, \ldots, m$. Since if $X \sim \beta'(a, b)$ then $X^{-1} \sim \beta'(b, a)$, the reciprocal transformation would also work in (11).

5 Partially Sensitive Data

In this section we discuss the situation when only a part of the response vector $\mathbf{y}$, say $(y_1, \ldots.., y_r)$ is sensitive and hence needs to be protected. We provide two methods, depending on the nature of imputation of the synthetic data.

Method I: Using whole data estimates to impute synthetic data

Plug-In Sampling

The original data now has the same setup as in Section 3 with both $r > p$ and $n - r > p$ assumed to hold. We synthesize m copies of the original data $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ given by $\{\boldsymbol{y}^{*j} = (\boldsymbol{y}_1^{*j}, \boldsymbol{y}_2) : j = 1, \ldots, m\}$ whose sufficient statistics are given by $(\boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m}, \boldsymbol{b}_2, \mathrm{RSS}_2)$. We denote $\overline{\boldsymbol{b}_1^*} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{b}_1^{*j}$. Then we can derive the following posterior distributions in a similar manner as before

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

$$\sigma^2 \,|\, \psi, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m}, \boldsymbol{b}_2, \mathrm{RSS}_2 \sim \text{Scale-inv-}\chi^2\left(\nu, \tau_1^2\right),$$

$$\pi(\psi \,|\, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m}, \boldsymbol{b}_2, \mathrm{RSS}_2),$$

$$\propto \left|\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \frac{\boldsymbol{X}_2'\boldsymbol{X}_2}{m}\right|^{-\frac{1}{2}} \psi^{-\frac{(m-1)(r-p)}{2}-1}(1+\psi)^{-\frac{mp}{2}} e^{\frac{(r-p)\psi}{2}} \left\{\nu\tau_1^2\right\}^{-\frac{\nu}{2}}$$

where $\nu = n + (m-1)r - p + \delta - 1$ and

$$\nu\tau_1^2 = \sum_{j=1}^{m}\left(\boldsymbol{b}_1^{*j} - \overline{\boldsymbol{b}_1^*}\right)' \frac{\boldsymbol{X}_1{}'\boldsymbol{X}_1}{1+\psi}\left(\boldsymbol{b}_1^{*j} - \overline{\boldsymbol{b}_1^*}\right)$$

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

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

Let $\boldsymbol{X}'\boldsymbol{X}_{\psi,m} = \dfrac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+\psi} + \dfrac{\boldsymbol{X}_2'\boldsymbol{X}_2}{m}$. Note that

$$\boldsymbol{X}'\boldsymbol{X}_{\psi,m} \geq \frac{\boldsymbol{X}'\boldsymbol{X}}{m(1+\psi)} \implies \left|\boldsymbol{X}'\boldsymbol{X}_{\psi,m}\right|^{-\frac{1}{2}} \leq m^{\frac{p}{2}}(1+\psi)^{\frac{p}{2}}\left|\boldsymbol{X}'\boldsymbol{X}\right|^{-\frac{1}{2}}\nu\tau_1^2 \geq \sum_{j=1}^{m}\frac{\mathrm{RSS}_1^{*j}}{\psi}$$

. Here $A \geq B$ is used to mean $(A-B)$ is positive semidefinite. Now using the fact that $(1+\psi)^{-\frac{(m-1)p}{2}} \leq 1$ (as $\psi > 0$), we can sample from the distribution of latent variables using the Accept-Reject algorithm as follows:

$$\pi(\psi \,|\, \text{data}) \leq \frac{m^{\frac{p}{2}}\left|\boldsymbol{X}'\boldsymbol{X}\right|^{-\frac{1}{2}}2^{\frac{n+(m-2)p+\delta-1}{2}}\Gamma\left(\frac{n+(m-2)p+\delta-1}{2}\right)}{(r-p)^{\frac{n+(m-2)p+\delta-1}{2}}\left(\sum_{j=1}^{m}\mathrm{RSS}_1^{*j}\right)^{\frac{n+(m-1)r-p+\delta-1}{2}}}\tilde{f}_{\text{ScaledChi}}(\psi)$$

where $\tilde{f}_{\text{ScaledChi}}(\psi)$ is the pdf of a Gamma$\left(\frac{n+(m-2)p+\delta-1}{2}, \frac{r-p}{2}\right)$ distribution. We also assume $n+(m-2)p+\delta-1 > 0$. All expressions in the partially sensitive case coincide with the results in the fully sensitive case when all of $\boldsymbol{y}$ is sensitive.

Posterior Predictive Sampling

We follow the same process as in Section 5 for the PPS case to derive the following posterior distributions

$$\begin{aligned}
&\boldsymbol{\beta} \,|\, \sigma^2, \psi_1, \ldots, \psi_m, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \boldsymbol{b}_2 \\
&\sim N_p\left[\left(\sum_{j=1}^{m}\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi_j} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\sum_{j=1}^{m}\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi_j}\boldsymbol{b}_1^{*j} + \boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{b}_2\right), \sigma^2\left(\sum_{j=1}^{m}\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi_j} + \boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\right], \\
&\sigma^2 \,|\, \psi_1, \ldots, \psi_m, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m}, \boldsymbol{b}_2, \mathrm{RSS}_2 \sim \text{Scale-inv-}\chi^2\left(\nu, \tau_1^2\right), \\
&\pi(\psi_1, \ldots, \psi_m \,|\, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m}, \boldsymbol{b}_2, \mathrm{RSS}_2) \\
&\propto \left|\sum_{j=1}^{m}\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{1+2\psi_j} + \boldsymbol{X}_2'\boldsymbol{X}_2\right|^{-\frac{1}{2}}\left(\prod_{j=1}^{m}\psi_j^{-1}(1+2\psi_j)^{-\frac{p}{2}}(1+\psi_j)^{-\frac{2r-2p+\alpha-1}{2}}\right)\left\{\nu\tau_2^2\right\}^{-\frac{\nu}{2}}
\end{aligned}$$

where $\nu = n+(m-1)r-p+\delta-1$ and

$$\begin{aligned}
&\nu\tau_2^2 \\
&= \sum_{j=1}^{m}\left(\boldsymbol{b}_1^{*j} - \left(\sum_{j=1}^{m}\frac{1}{1+2\psi_j}\right)^{-1}\left(\sum_{j=1}^{m}\frac{\boldsymbol{b}_1^{*j}}{1+2\psi_j}\right)\right)'\frac{\boldsymbol{X}_1{}'\boldsymbol{X}_1}{1+2\psi_j}\left(\boldsymbol{b}_1^{*j} - \left(\sum_{j=1}^{m}\frac{1}{1+2\psi_j}\right)^{-1}\left(\sum_{j=1}^{m}\frac{\boldsymbol{b}_1^{*j}}{1+2\psi_j}\right)\right) \\
&+ \left(\left(\sum_{j=1}^{m}\frac{1}{1+2\psi_j}\right)^{-1}\left(\sum_{j=1}^{m}\frac{\boldsymbol{b}_1^{*j}}{1+2\psi_j}\right) - \boldsymbol{b}_2\right)'\left(\left(\sum_{j=1}^{m}\frac{1}{1+2\psi_j}\right)^{-1}\left(\boldsymbol{X}_1{}'\boldsymbol{X}_1\right)^{-1} + \left(\boldsymbol{X}_2{}'\boldsymbol{X}_2\right)^{-1}\right)^{-1} \\
&\left(\left(\sum_{j=1}^{m}\frac{1}{1+2\psi_j}\right)^{-1}\left(\sum_{j=1}^{m}\frac{\boldsymbol{b}_1^{*j}}{1+2\psi_j}\right) - \boldsymbol{b}_2\right) + \sum_{j=1}^{m}\frac{\mathrm{RSS}_1^{*j}}{\psi_j} + \mathrm{RSS}_2.
\end{aligned}$$

The posterior distributions are proper as long as $r > \max\left\{p, p-\alpha+1, \frac{n+(2m-1)p-\alpha m+\delta+m-1}{m+1}\right\}$, $n-r > \max\{p, p-rm-\delta+1\}$. Note that

$$\left(\sum_{j=1}^{m}\frac{1}{1+2\psi_j}\right)\boldsymbol{X}_1'\boldsymbol{X}_1+\boldsymbol{X}_2'\boldsymbol{X}_2 \geq \left(\prod_{j=1}^{m}\frac{1}{1+2\psi_j}\right)\boldsymbol{X}_1'\boldsymbol{X}_1+\boldsymbol{X}_2'\boldsymbol{X}_2 \geq \left(\prod_{j=1}^{m}\frac{1}{1+2\psi_j}\right)\boldsymbol{X}'\boldsymbol{X}$$

and hence using rejection sampling one can sample the latent variables as:

$$\pi(\psi_1,\ldots,\psi_m \,|\, \text{data}) \leq \frac{\left(\prod_{j=1}^{m}\mathrm{B}\left(\frac{n+(m-1)r-p+\delta-1}{2m}, \frac{(m+1)r-n-(2m-1)p+m\alpha-\delta-m+1}{2m}\right)\right)}{|\boldsymbol{X}'\boldsymbol{X}|^{\frac{1}{2}}\left(m\prod_{j=1}^{m}\mathrm{RSS}_j^{*\frac{1}{m}}\right)^{\frac{n+(m-1)r-p+\delta-1}{2}}}\left(\prod_{j=1}^{m}\overline{g_j}'(\psi_j)\right),$$

where $\overline{g_j}'(\psi_j)$ is the pdf of a Beta-Prime $\left(\frac{n+(m-1)r-p+\delta-1}{2m}, \frac{(m+1)r-n-(2m-1)p+m\alpha-\delta-m+1}{2m}\right)$ distribution, independently for $j=1,\ldots,m$.

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

Plug-In Sampling

As in Section 5, our analysis in this case will be based solely on the synthetic part. We require only $r > p$. The posterior distributions are given by

$$\begin{aligned}
\boldsymbol{\beta} \,|\, \sigma^2, \psi, \overline{\boldsymbol{b}_1^*} &\sim N_p\left(\overline{\boldsymbol{b}_1^*}, \sigma^2\left((\boldsymbol{X}'\boldsymbol{X})^{-1}+\frac{\psi}{m}(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\right)\right), \\
\sigma^2 \,|\, \psi, \boldsymbol{b}_1^{*1},\ldots,\boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1},\ldots,\mathrm{RSS}_1^{*m} &\sim \text{Scale-inv-}\chi^2\left(\tilde{\nu}, \tilde{\tau}_1^2\right), \\
\psi &\sim \frac{\chi^2_{n-p+\delta-1}}{n-p} \equiv \Gamma\left(\frac{n-p+\delta-1}{2}, \frac{n-p}{2}\right)
\end{aligned}$$

where $\tilde{\nu} = rm-p+\delta-1$, $\tilde{\nu}\tilde{\tau_1}^2 = \frac{1}{\psi}\left(\sum_{j=1}^{m}\mathrm{RSS}_1^{*j} + \sum_{j=1}^{m}\left(\boldsymbol{b}_1^{*j}-\overline{\boldsymbol{b}_1^*}\right)'(\boldsymbol{X}_1{}'\boldsymbol{X}_1)\left(\boldsymbol{b}_1^{*j}-\overline{\boldsymbol{b}_1^*}\right)\right)$. The posterior distributions are proper as long as $r > \max\left\{p, \frac{p-\delta+1}{m}\right\}$, $n > p-\delta+1$ and the results match our expressions from Section 3 when $r=n$.

Posterior Predictive Sampling

Following the derivations in the plug-in sampling case for partially sensitive response, one can derive analogous expressions for the posterior in the posterior predictive sampling scheme. The posteriors

are

$$\begin{aligned}
&\boldsymbol{\beta} \,|\, \sigma^2, \psi_1, \ldots, \psi_m, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m} \\
&\sim N_p \left[\left(\sum_{j=1}^{m} \psi_j^{-1} \right)^{-1} \left(\sum_{j=1}^{m} \frac{\boldsymbol{b}_1^{*j}}{\psi_j} \right), \sigma^2 \left(\left(\sum_{j=1}^{m} \psi_j^{-1} \right)^{-1} \left(\left(\boldsymbol{X}_1' \boldsymbol{X}_1 \right)^{-1} + \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \right) + \left(\boldsymbol{X}' \boldsymbol{X} \right)^{-1} \right) \right], \\
&\sigma^2 \,|\, \psi_1, \ldots, \psi_m, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m} \sim \text{Scale-inv-}\chi^2 \left(\nu, \tilde{\tau_2}^2 \right), \\
&\pi(\psi_1, \ldots, \psi_m \,|\, \boldsymbol{b}_1^{*1}, \ldots, \boldsymbol{b}_1^{*m}, \mathrm{RSS}_1^{*1}, \ldots, \mathrm{RSS}_1^{*m}) \\
&\propto \left(\prod_{j=1}^{m} \psi_j \right)^{-\frac{n-p+r+\alpha+1}{2}} \left(\sum_{j=1}^{m} \psi_j^{-1} \right)^{-\frac{p}{2}} \left(1 + \left(\sum_{j=1}^{m} \psi_j^{-1} \right) \right)^{-\frac{2n-2p+\alpha-1}{2}} \left\{ \nu \tilde{\tau_2}^2 \right\}^{-\frac{\nu}{2}}
\end{aligned}$$

where $\nu = m(r-2) - p + \delta + 1$ as before and

$$\begin{aligned}
\nu \tilde{\tau_2}^2 = \sum_{j=1}^{m} &\left(\boldsymbol{b}_1^{*j} - \left(\sum_{j=1}^{m} \psi_j^{-1} \right)^{-1} \left(\sum_{j=1}^{m} \frac{\boldsymbol{b}_1^{*j}}{\psi_j} \right)^{-1} \right)' \frac{\left((\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} + (\boldsymbol{X}' \boldsymbol{X})^{-1} \right)^{-1}}{\psi_j} \\
&\left(\boldsymbol{b}_1^{*j} - \left(\sum_{j=1}^{m} \psi_j^{-1} \right)^{-1} \left(\sum_{j=1}^{m} \frac{\boldsymbol{b}_1^{*j}}{\psi_j} \right)^{-1} \right) + \sum_{j=1}^{m} \frac{\mathrm{RSS}_1^{*j}}{\psi_j}
\end{aligned}$$

The latent variables can be sampled using Accept-Reject algorithm similarly as before, using an Inverse-Dirichlet distribution as proposal distribution.

6 Discussion

We have described how to perform full Bayesian analysis in the multiple regression model based on multiply imputed synthetic data. The priors used are non-informative priors. The frequentist coverage of the credible set (based on a limited simulation study that is not reported here) is sensitive to the choice of the hyperparameter and depending on the value the coverage may be significantly lower than the nominal level. This is expected due to the latent variable structure of the problem. However, more investigation is needed to understand the impact of the hyperparameters. The analysis can be extended to related cases such as where all or some of the independent variables are also sensitive or when the response is multivariate. We will consider such an investigation in the future.

Acknowledgemnt

The authors thank Andrew Raim for reviewing the paper and providing many helpful comments and suggestions. The authors also thank Tommy Wright for his support and encouragement.

References

[1] A. Guin, A. Roy, and B.K. Sinha. Bayesian analysis of singly imputed partially synthetic data generated by plug-in sampling and posterior predictive sampling under the multiple linear regression model. *International Journal of Statistical Sciences*, to appear, 2021.

[2] Martin Klein, John Zylstra, and Bimal Sinha. Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model. *Calcutta Statistical Association Bulletin*, 71(2):63–82, 2019.

[3] M.D. Klein, T. Mathew, and B. Sinha. Noise multiplication for statistical disclosure control of extreme values in log-normal regression samples. *Journal of Privacy and Confidentiality*, 6(1), Jun. 2014.

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

[5] M.D. Klein and B. Sinha. 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), Dec. 2015.

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

[7] J.P. Reiter. Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology*, 29(2):181–188, 2003.

[8] D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.