Synthetic Data Generation for Enhanced Covariance Matrix Estimation

Seungkyu Kim, Johan Lim, Donghyeon Yu*

Abstract

Synthetic data generation is an important tool to ensure data confidentiality. Various synthetic data generators have been developed in the literature. The methods in the literature are mostly for general purposes. They aim to generate data whose distributions are the same as the original data set, and the synthesized data are used for every purpose depending on who uses them. However, it could not be good for all purposes. In this paper, we study the synthetic data generation tailored for a specific purpose. We are particularly interested in covariance matrix estimation, which is a key part of many multivariate statistical analyses. To do it, we first see the connection between the sequential regression model and the modified Cholesky decomposition. We then devise a new synthetic data generator, named SynCov, that controls the error variances of the sequential regression model. We show that SynCov results in a shrinkage (synthesized) covariance matrix estimator. We numerically show that our SynCov performs better than other synthetic data generation methods in covariance matrix estimation. Finally, we apply our SynCov to two real data examples, (i) the estimation of the covariance matrix of the (selected) variables of the Los Angeles City Employee Payroll data and (ii) the classification of the Taiwanese Bankruptcy Data.

Keyword: Covariance matrix estimation, data confidentiality, sequential regression multiple imputation, shrinkage estimator, synthetic data generator, disclosure risk of synthetic data

1 Introduction

The concept of generating synthetic data for statistical disclosure control was first introduced by Rubin (1993) based on the multiple imputation method, where the synthetic data is generated from the joint distribution estimated from the original data. The advantage of

*Seungkyu Kim and Johan Lim are at Department of Statistics, Seoul National University, Seoul, Korea. Donghyeon Yu is at Department of Statistics, Inha University, Incheon, Korea. All authors contribute equally and D.Yu is the corresponding author.

synthetic data generation lies in the absence of original samples in the released synthetic dataset. Due to this advantage, synthetic data generation is often considered a risk-free method for data confidentiality (Walonoski et al., 2018). If the original dataset consists solely of categorical variables, the synthetic data set can have identical records to those in the original dataset and thus is not free of disclosure risk (Taub and Elliot, 2019; Taub et al., 2018). However, if the original data contains continuous variables only, theoretically, the synthetic data could not have the same values as the original, although the synthesized dataset can potentially contain data points that are close to those in the original (Smith et al., 2023). In this paper, we are interested in generating synthetic data of continuous variables.

Suppose we have n samples of $p-$dimensional random variables $\mathbf{x} = (X_1, X_2, \ldots, X_p)^\top$. A typical synthetic data generator has two steps. In the first step, we estimate the p-dimensional distribution $P(X_1, X_2, \ldots, X_p)$ from the original data and, then, we generate synthetic samples from the estimated distribution. However, we all know that the estimation of a p-dimensional distribution is a formidable task, even when p is not large. To circumvent the difficulty, we decompose $P(X_1, X_2, \ldots, X_p)$ into

$$P(X_1) \prod_{j=2}^{n} P(X_j \big| X_{j-1}, \ldots, X_1), \tag{1}$$

and sequentially estimate the conditional distributions, and generate synthetic samples from them. In doing these sequential estimations and generations, to make the problem simple, we assume a regression model as, for $j = 2, 3, \ldots, p$,

$$X_j = f_j(X_{j-1}, \ldots, X_1) + \epsilon_j, \tag{2}$$

where ϵ_j are independently distributed as $(0, \sigma_j^2)$ and $f_j(X_{j-1}, \ldots, X_1)$ are regression functions. The conditional mean $f_j(X_{j-1}, \ldots, X_1)$ in (2) is either modelled with a parametric function (e.g. linear regression function) or a non-parametric function (e.g. the classification and regression tree (CART) proposed by Breiman et al. (1984)).

The synthetic data generation methods until now are mostly for a general purpose. To be specific, it aims to generate a data set whose distribution is that of the original data, but does not consider the following use of the synthesized data. The synthesized data are used for various purposes depending on the users including regression, classification, clustering, and the estimation of quantiles depending on the users. However, it could not be good for all purposes.

In this paper, we study the method to generate a synthetic dataset that is tailored for a specific purpose. We are particularly interested in covariance matrix estimation which is

a key part of many multivariate statistical analyses. To do it, we first see the connection between the sequential regression model in (2) and the modified Cholesky decomposition by Pourahmadi (1999). We then devise a new synthetic data generator named SynCov that controls the error variances in (2) and results in a shrinkage synthetized covariance matrix estimator. We propose the optimal control level of the error variance under the Gaussian assumption that theoretically minimizes the mean squared error (MSE, equivalently, the Frobenius risk) of the covariance matrix estimation following the idea of Chen et al. (2010).

The remainder of this paper is organized as follows. In Section 2, we introduce some preliminary results - the sequential regression model and the modified Cholesky decomposition, - to elucidate our new proposal. In Section 3, we introduce a synthetic data generation method that controls error variances to enhance covariance matrix estimation in terms of the MSE. In Section 4, we numerically compare our SynCov to other synthetic data generation methods in covariance matrix estimation. In Section 5, we apply our SynCov to two data examples, (i) the estimation of the covariance matrix of the (selected) variables of the Los Angeles City Employee Payroll data, (ii) the classification of the Taiwanese Bankruptcy Data. Finally, in Section 6, we conclude our study with a summary and some remarks.

2 Preliminaries

2.1 Modified Cholesky Decomposition

In this section, we briefly introduce the modified Cholesky decomposition proposed by Pourahmadi (1999), which bridges the sequential regression models to the covariance estimation. To be specific, let $\mathbf{Y} = \big(Y_1, \ldots, Y_p\big)^\top$ be a random vector from a multivariate distribution with covariance matrix Σ. Without loss of generality, we suppose that the random vector $\mathbf{Y}$ is centered (i.e., $E(\mathbf{Y}) = \mathbf{0}$). We consider the following sequential regression model for a given order of $(1, 2, \ldots, p)$,

$$Y_1 = \epsilon_1, \;\; Y_t = \sum_{j=1}^{t-1} \psi_{tj} Y_j + \epsilon_t, \quad t = 2, \ldots, p, \tag{3}$$

where ϵ_ts are independent random errors with mean 0 and variance σ_t^2. With the matrix and vector notations, the sequential regression model (3) can be represented as

$$\mathbf{\Psi Y} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\psi_{21} & 1 & 0 & \cdots & 0 & 0 \\ -\psi_{31} & -\psi_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\psi_{p1} & -\psi_{p2} & -\psi_{p3} & \cdots & -\psi_{p(p-1)} & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_p \end{pmatrix} = \boldsymbol{\epsilon}. \tag{4}$$

Further, Pourahmadi (1999) considers the following matrix decomposition of the covariance matrix:

$$\mathbf{\Psi}\Sigma\mathbf{\Psi}^\top = \mathbf{D} \quad \text{and} \quad \Sigma = \mathbf{\Psi}^{-1}\mathbf{D}(\mathbf{\Psi}^\top)^{-1} = \mathbf{L}\mathbf{D}\mathbf{L}^\top, \tag{5}$$

where $\mathbf{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\mathbf{L} = \mathbf{\Psi}^{-1}$. Since the matrix $\mathbf{L}$ is a unit lower triangular matrix and $\Sigma = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top$, this matrix decomposition is called the modified Cholesky decomposition (MCD) of the covariance matrix Σ, where $\tilde{\mathbf{L}} = \mathbf{L}\mathbf{D}^{1/2}$ and $\mathbf{D}^{1/2} = \text{diag}(\sigma_1, \ldots, \sigma_p)$. Moreover, the inverse of the covariance matrix can be represented with $\mathbf{\Psi}$ and $\mathbf{D}$ directly as $\Omega = \Sigma^{-1} = \mathbf{\Psi}^\top\mathbf{D}^{-1}\mathbf{\Psi}$.

This relationship between the sequential regression model and the covariance matrix via MCD introduces a new way to estimate the covariance matrix Σ and its inverse Ω (a.k.a. the precision matrix). In particular, the covariance estimation by the modified Cholesky factor provides both a positive definite covariance estimator and a regression interpretation of the relationships between covariates. Based on these advantages, Huang et al. (2006) propose the covariance estimator based on the penalized normal likelihood function with the MCD. Additionally, Rothman et al. (2010) consider the banded Cholesky factor of the MCD for the covariance estimation. Rajaratnam and Salzman (2013) study the order of covariates for the banded covariance matrix estimation through the MCD as well. More recently, MCD has been utilized to enhance the estimation of both covariance and precision matrices by permuting covariate orders, as demonstrated in Kang and Deng (2020), Kang et al. (2020a), and Kang et al. (2020b), where the MCD representation with permutation is employed as the ensemble procedure.

2.2 Synthetic Data Generation via Sequential Regression

In this section, we introduce a synthetic data generation approach based on sequential regression, which is the base of our approach. The synthetic data generation through sequential regression originates from the sequential regression multiple imputation (SRMI) approach (Raghunathan et al., 2001) within the context of missing data imputation. To be specific, let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ be an $n \times d$ dimensional matrix of fully observed variables and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p)$ be an $n \times p$ dimensional matrix of partially observed variables, where the columns of $\mathbf{Y}$ are ordered by the amount of missing values from least to most. The SRMI approach considers the following regression models for partially observed variables given the fully observed variables:

$$Y_1 = f_1(\mathbf{X}) + \epsilon_1, \; Y_j = f_j(Y_{j-1}, \ldots, Y_1, \mathbf{X}) + \epsilon_j \text{ for } j = 2, 3, \ldots, p, \tag{6}$$

where $f_j(\cdot)$ is a regression functions of Y_j. The regression functions are determined as the type of the response variable in general. For instance, the usual choice of the regression model for a continuous response is the linear regression model.

Motivated by the first proposal of fully synthetic data (Rubin, 1993), the SRMI is employed to generate synthetic data in Raghunathan et al. (2003). Synthetic data generation via SRMI consists of two steps. In the first step, we fit the regression models in (6) with a given order. Let $\hat{f}_j(\cdot)$ be the estimated regression function of Y_j. Then, the synthetic data is generated by the following equations with a given order $(j_1, j_2, \ldots, j_p)$:

$$Y^*_{j_1} = \hat{f}_{j_1}(\mathbf{X}) + \tilde{\epsilon}_{j_1}, \; Y^*_{j_k} = \hat{f}_{j_k}(Y^*_{j_k-1}, \ldots, Y^*_{j_1}, \mathbf{X}) + \tilde{\epsilon}_{j_k} \text{ for } k = 2, 3, \ldots, p, \tag{7}$$

where $\tilde{\epsilon}_{j_k}$ is randomly drawn from a distribution with mean 0 and variance $\hat{\sigma}^2_{j_k}$ and $\hat{\sigma}^2_{j_k}$ is the estimated variance from the model in (6). Besides fully synthetic data, the partially synthetic data generation (Little, 1993) considers replacing sensitive original observations that pose a high disclosure risk with the synthesized samples. For further details, refer to Reiter (2005a) and Reiter and Raghunathan (2007) for fully synthetic data, and to Reiter (2003), Reiter (2005b), and Drechsler and Reiter (2011) for partially synthetic data.

There are several available software packages for statistical synthetic data generation methods. Among them, the R package `synthpop`, implemented by Nowok et al. (2016), provides various parametric and nonparametric regression functions in the sequential regression models. For example, `synthpop` provides the ordinary linear regression, logistic regression, polynomial regression for parametric functions and classification and regression tree and random forest (Breiman, 2001; Breiman et al., 1984) for nonparametric functions.

Besides statistical methods for synthetic data generation, various techniques based on artificial neural networks have been proposed in recent years. Some examples are variational autoencoder-based models by Tomczak and Welling (2017) and Ma et al. (2020) and generative adversarial network-based models by Park et al. (2018), Xu et al. (2019), and Zhao et al. (2021). However, in this study, we specifically concentrate on the sequential regression model, which serves as our primary generation framework.

3 New synthetic data generator for covariance matrix estimation - SynCov

In this section, we establish a connection between the synthetic generation method using the sequential regression model and the minimum MSE covariance estimation, which is

achieved by controlling error variances in the sequential regression model. Specifically, let $\mathbf{Y} = \big(Y_1, \ldots, Y_p\big)^\top$ be a random vector from a multivariate distribution with a covariance matrix Σ. As described in Section 2.2, we assume that the random vector $\mathbf{Y}$ is centered without loss of generality. In this study, we describe our proposed method for fully synthetic data generation when all the variables in the original dataset are continuous. It is worth noting that our method is versatile and can also be applied to generate partially synthetic data by considering the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, where $\mathbf{X}$ may consist of either categorical or continuous variables.

Recall the sequential regression model in (3):

$$Y_1 = \epsilon_1, \;\; Y_t = \sum_{j=1}^{t-1} \psi_{tj} Y_j + \epsilon_t, \quad t = 2, \ldots, p,$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$ are independent errors with variances $\big(\sigma_t^2\big)_{t=1}^p$. The error variances can be expressed as the following conditional variances for $t = 2, \ldots, p$:

$$\sigma_t^2 = \mathrm{Var}\big(\epsilon_t\big) = \mathrm{Var}\big(Y_t | Y_{[1:(t-1)]}\big),$$

where $Y_{[a:b]} = (Y_a, Y_{a+1}, \ldots, Y_b)$ for $a \le b$. In general, the synthetic data generation via the sequential regression models consists of the following two steps:

(Step 1) Parameter estimation step: for $t = 2, \ldots, p$,

$$\big(\hat{\psi}_{t1}, \ldots, \hat{\psi}_{t(t-1)}\big)^\top = \underset{\psi_{t1}, \ldots, \psi_{t(t-1)}}{\arg\min} \sum_{i=1}^{n} \Big(Y_{it} - \sum_{j=1}^{t-1} \psi_{tj} Y_{ij}\Big)^2, \tag{8}$$

where Y_{ij} is the i-th observation for the j-th variable in the original data.

(Step 2) Synthetic data generation: for $i = 1, \ldots, m$,

$$\begin{aligned} &Y_{i1} \text{ is drawn from the estimated distribution of } Y_1 \\ &Y_{it} = \sum_{j=1}^{t-1} \hat{\psi}_{tj} Y_{ij} + \epsilon_{it}, \text{ for } t = 2, \ldots, p, \end{aligned} \tag{9}$$

where m is a sample size of the synthetic data, Y_{ij} is the i-th synthesized sample for the j-th variable, and ϵ_{it} is the i-th random sample from the estimated error distribution of ϵ_t.

In this study, we consider the modified sequential regression model for the synthetic data generation in Step 2 as follows:

$$\tilde{Y}_1 = \eta\tilde{\epsilon}_1, \;\; \tilde{Y}_t = \sum_{j=1}^{t-1} \psi_{tj} \tilde{Y}_j + \eta\tilde{\epsilon}_t, \quad t = 2, \ldots, p, \tag{10}$$

where $\tilde{Y}_t$ denotes the t-th random variable in the synthetic generation model, and $\tilde{\epsilon}_t$ represents a random error for the t-th regression model having the same distribution with that of ϵ_t. We then find the covariance matrix $\tilde{\Sigma}_\eta$ for $(\tilde{Y}_1, \ldots, \tilde{Y}_p)$ is $\eta^2\Sigma$ as shown in Lemma 1, where Σ is the covariance matrix of the original data $(Y_1, \ldots, Y_p)$.

Lemma 1. *Let Σ be the covariance matrix of the original data $(Y_1, \ldots, Y_p)$ and $\tilde{\Sigma}_\eta$ be the covariance matrix of $(\tilde{Y}_1, \ldots, \tilde{Y}_p)$ in the modified sequential model* (10)*. Then, $\tilde{\Sigma}_\eta$ can be represented as $\eta^2\Sigma$.*

Proof. From the modified sequential regression model (10), we can represent the model with the vector and matrix from as follows:

$$\boldsymbol{\Psi}\tilde{\mathbf{Y}} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\psi_{21} & 1 & 0 & \cdots & 0 & 0 \\ -\psi_{31} & -\psi_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\psi_{p1} & -\psi_{p2} & -\psi_{p3} & \cdots & -\psi_{p(p-1)} & 1 \end{pmatrix} \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \tilde{Y}_3 \\ \vdots \\ \tilde{Y}_p \end{pmatrix} = \begin{pmatrix} \eta\tilde{\epsilon}_1 \\ \eta\tilde{\epsilon}_2 \\ \eta\tilde{\epsilon}_3 \\ \vdots \\ \eta\tilde{\epsilon}_p \end{pmatrix} = \eta\tilde{\boldsymbol{\epsilon}}. \tag{11}$$

Hence, $\tilde{\Sigma}_\eta = \mathrm{Var}(\tilde{\mathbf{Y}}) = \eta^2\boldsymbol{\Psi}^{-1}\mathrm{Var}(\tilde{\boldsymbol{\epsilon}})\big(\boldsymbol{\Psi}^T\big)^{-1} = \eta^2\Sigma$ because $\mathrm{Var}(\tilde{\boldsymbol{\epsilon}}) = \mathrm{Var}(\boldsymbol{\epsilon}) = \mathbf{D}$ and $\Sigma = \boldsymbol{\Psi}^{-1}\mathbf{D}(\boldsymbol{\Psi}^T)^{-1}$ in (5). □

From Lemma 1, the sample covariance matrix of the synthetic data generated from the modified sequential regression model can be regarded as a re-scaled version of the sample covariance matrix of the original data. We denote η as the parameter to control the error variance. Additionally, when $\eta = 1$, the modified sequential regression model reduces to the original sequential regression model. To select the optimal control parameter η, we consider the following minimization problem motivated by the minimum MSE covariance estimation in Chen et al. (2010):

$$\begin{aligned} \min_{\eta} \quad & E\Big(\big\|\widehat{\tilde{\Sigma}}_\eta - \Sigma\big\|_{\mathrm{F}}^2\Big) \\ \text{s.t.} \quad & \widehat{\tilde{\Sigma}}_\eta = \eta^2 S_n, \end{aligned} \tag{12}$$

where $\|A\|_{\mathrm{F}}^2 = \mathrm{tr}(AA^\top)$ and $S_n = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}$ is the sample covariance matrix of the original data.

For η^*, the minimizer of the problem (12), $\widehat{\tilde{\Sigma}}_{\eta^*}$ is the covariance matrix estimator to minimize the Frobenius risk in (12) in the class $\{\tilde{\Sigma}_\eta : \tilde{\Sigma}_\eta = \eta^2 S_n\}$. The optimal shrinkage level η^* is given in Theorem 1 below.

Theorem 1. *Let $S_n = \frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top$ be the sample covariance matrix of the p-dimensional i.i.d. random vectors $\{\mathbf{y}_i\}_{i=1}^n$ with $E(\mathbf{y}_i) = \mathbf{0}$ and $\mathrm{Var}(\mathbf{y}_i) = \Sigma$ for all $i = 1, 2, \ldots, n$. If $\{\mathbf{y}_i\}_{i=1}^n$ follow*

Gaussian distribution, then the solution to (12) *is*

$$(\eta^*)^2 = \frac{E\{\mathrm{tr}(S_n\Sigma)\}}{E\{\mathrm{tr}(S_n^2)\}} = \frac{n\mathrm{tr}(\Sigma^2)}{(n+1)\mathrm{tr}(\Sigma^2) + (\mathrm{tr}(\Sigma))^2}. \tag{13}$$

Moreover, the estimator $(\eta^*)^2 S_n$ *for* Σ *satisfies*

$$E\{\|(\eta^*)^2 S_n - \Sigma\|_{\mathrm{F}}^2\} \le E\{\|S_n - \Sigma\|_{\mathrm{F}}^2\}. \tag{14}$$

Proof. By plugging in the constraint form $\eta^2 S_n$ to $\widehat{\widetilde{\Sigma}}_\eta$, we easily get the following equation

$$\begin{aligned}(\eta^*)^2 &= \arg\min_{\eta^2} E\left\{\|\widehat{\widetilde{\Sigma}}_\eta - \Sigma\|_{\mathrm{F}}^2\right\} \text{ subject to } \widehat{\widetilde{\Sigma}}_\eta = \eta^2 S_n \\ &= \arg\min_{\eta^2} E\left\{\mathrm{tr}(\eta^2 S_n - \Sigma)(\eta^2 S_n - \Sigma)^\top\right\} \\ &= \frac{E\{\mathrm{tr}(S_n\boldsymbol{\Sigma})\}}{E\{\mathrm{tr}(S_n^2)\}}.\end{aligned} \tag{15}$$

From the Gaussian assumption and results in Chen et al. (2010) and Letac and Massam (2004), we can express the numerator and denominator terms as

$$E\{\mathrm{tr}(S_n\Sigma)\} = \mathrm{tr}(\Sigma^2) \quad \text{and} \quad E\{\mathrm{tr}(S_n^2)\} = \frac{n+1}{n}\mathrm{tr}(\Sigma^2) + \frac{1}{n}(\mathrm{tr}(\Sigma))^2 \tag{16}$$

and so

$$(\eta^*)^2 = \frac{n\mathrm{tr}(\Sigma^2)}{(n+1)\mathrm{tr}(\Sigma^2) + \{\mathrm{tr}(\Sigma)\}^2}.$$

In addition, the inequality (14) is trivial since $(\eta^*)^2$ is the minimizer of the function $f(\eta^2) = E\{\|\eta^2 S_n - \Sigma\|_{\mathrm{F}}^2\}$, where S_n can be considered as $\eta^2 S_n$ with $\eta = 1$. □

As described in Theorem 1, the theoretical minimizer $(\eta^*)^2$ is defined with the unknown parameter Σ. To resolve this difficulty, we can refer (16) and

$$E\{(\mathrm{tr}(S_n))^2\} = (\mathrm{tr}(\Sigma))^2 + \frac{2}{n}\mathrm{tr}(\Sigma^2)$$

so that $(\eta^*)^2$ has another representation as

$$(\eta^*)^2 = \frac{nE\{\mathrm{tr}(S_n^2)\} - E[\{\mathrm{tr}(S_n)\}^2]}{(n+1-2n^{-1})E\{\mathrm{tr}(S_n^2)\}}. \tag{17}$$

Then, we propose the estimate $\hat{\eta}^2$ as

$$\hat{\eta}^2 = E\big((\eta^*)^2 | S_n\big) = \frac{n\mathrm{tr}(S_n^2) - \{\mathrm{tr}(S_n)\}^2}{(n+1-2n^{-1})\mathrm{tr}(S_n^2)}. \tag{18}$$

In summary, the proposed synthetic data generation method for enhanced minimum MSE covariance estimation considers the following two steps after Step 1 instead of Step 2 in (9).

(Step 2-M) Estimation of the optimal control parameter $\hat{\eta}^2$ using the equation (18).

(Step 3-M) Synthetic data generation: for $i = 1, \ldots, m$,

$$\begin{aligned} &Y_{i1} \text{ is drawn from the estimated distribution of } Y_1 \\ &Y_{it} = \sum_{j=1}^{t-1} \hat{\psi}_{tj} Y_{ij} + \hat{\eta}\epsilon_{it}, \text{ for } t = 2, \ldots, p, \end{aligned} \tag{19}$$

where m represents the sample size of the synthetic data, $\tilde{Y}_{ij}$ denotes the i-th synthesized sample for the j-th variable, $\tilde{\epsilon}_{it}$ is the i-th random sample from the estimated error distribution of ϵ_t, and $\hat{\eta}$ can be either $+\sqrt{\hat{\eta}^2}$ or $-\sqrt{\hat{\eta}^2}$.

4 Numerical study

In this section, we numerically compare the performance of the proposed SynCov to other existing sequential regression methods in estimating the covariance matrix.

For the comparison, we consider four covariance structures following Kang et al. (2020b), which are:

(M1) Independent and unequal variances: $\mathbf{\Sigma}_1 = \text{diag}(1, 2^{-1}, \ldots, p^{-1})$

(M2) Autoregressive structure with homogeneous variance Σ_2:

$$(\Sigma_2)_{ij} = 0.5^{|i-j|} \text{ for } 1 \le i, j \le p.$$

(M3) Linearly decreasing correalation and possibly banded covariance matrix Σ_3:

$$(\Sigma_3)_{ij} = \max\{1 - 2|i-j|/p, 0\}$$

(M4) Block diagonal matrix structure with the compound symmetry and identity matrix structure Σ_4:

$$\Sigma_4 = \begin{pmatrix} \text{CS}(0.5) & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \text{ and } \big(\text{CS}(0.5)\big)_{ij} = \begin{cases} 1 & \text{for } i = j = 1, 2, \ldots, \lceil p/2 \rceil \\ 0.5 & \text{for } i \ne j, 1 \le i, j \le \lceil p/2 \rceil \end{cases}.$$

For each model, the original data is simulated from the multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with the sample size $n = 250, 500$ and the number of variables $p = 25, 50, 100$. For each original data set, we generate a synthetic data set with the same size using four synthetic methods, which are

(1) Sequential regression (SR): the synthetic data is generated by Step 1 and Step 2 described in Section 3.

(2) SynCov: the synthetic data is generated by Step 1, Step 2-M, and Step 3-M described in Section 3.

(3) Sequential regression preserving the marginal distribution (SR-PMD): It is a default option of the R package `synthpop`. In each stage of sequential regression, this method transforms the response variables $\{Y_{ij}\}_{i=1}^n$ into normal quantile values $\{Q_{ij}\}_{i=1}^n$ using the rank of Y_{ij} among $\{Y_{ij}\}_{i=1}^n$ in the original data. It regresses Q_{ij} against Y_{ik}, $k = 1, 2, \ldots, j-1$, and generates the synthesized response value $\hat{Q}_{ij}$ from the estimated regression model. Finally, it transforms $\hat{Q}_{ij}$ to $\hat{Y}_{ij}$ with Y_{ik} whose normal quantile value is the closest to $\hat{Q}_{ij}$.

(4) Classification and regression tree (CART): the synthetic data is generated by the CART model. This method fits the regression tree model by binary recursive partitioning in each sequential regression model. In the terminal node, it randomly draws a sample Y from the raw observations in the node and takes it as the synthetic value.

To apply the synthetic data generation models mentioned above, we use R package `synthpop` with `norm`, `normrank`, and `cart` in the `method` argument for SR, SR-PMD, and CART, respectively. We consider two performance measures of the covariance matrix estimation, the Frobenius loss (the Frobebius norm) and the matrix ℓ_2 loss (the spectral norm) of the sample covariance matrix evaluated from the synthesized data. We replicate the simulation of the original data and the generation of the synthetic data 1000 times, and report the averages and standard errors of the two performance measures in Table 1.

We observe several interesting features from Table 1. First, the proposed method SynCov outperforms SR for all cases of Models 2–4 and demonstrates similar performance for cases of Model 1, which supports the theoretical result in Theorem 1 in Section 3. Second, our SynCov also outperforms SR-PMD for all cases of Models 2–4. All methods exhibit similar performance in Model 1, where the variables are independent with unequal variances. Third, SR-PMD exhibits significantly worse estimation performance than the other methods for cases of Model 3, where the dependency of variables linearly decreases in the variable order. This might be due to the information loss resulting from the rank transformation. Finally, SynCov exhibits either similar or slightly worse performance than CART; CART is a nonparametric method, while SynCov is a parametric method. We conjecture that this

order between SynCov and CART is the advantage of (i) the variable selection and (ii) the resampling of the original data. In CART, important variables are chosen to construct a tree structure, whereas SynCov includes all variables at each step in the sequential regression model. In addition, CART (also SR-PMD) generates synthetic data by resampling the observations in the original data and thus, inevitably, all values in the synthetic data are the values that appeared in the original data. This increases the disclosure risk of synthesized data and contradicts the purpose of the use of synthetic data.

5 Real Data examples

In this section, we apply our SynCov and the CART method in R package `synthpop` to two real data examples, LA City Employee Payroll Data and Taiwanese Bankruptcy Data. The two methods show better performance than others in the numerical study in Section 4. In both data sets, we aim to evaluate or compare the performances of the synthetic data not only in the estimation of the covariance matrix but also in the subsequent multivariate analysis. In the first example, we evaluate the performance in estimating the first few eigenvectors and their principal component scores which are the basics of the principal component analysis. In the second example, we investigate the performance of the synthetic data in linear discriminant analysis (LDA).

5.1 Application to LA City Employee Payroll Data

In this section, we apply the methods in Section 4 including our SynCov to a subset of Los Angeles City Employee Payroll data provided by the LA City Controller's Office. The full data are available at `https://controllerdata.lacity.org/Payroll/City-Employee-Payroll-Current-/g9h8-fvhu/about_data`, which consist of $371,455$ observations of 35 variables including 21 continuous variables. We build a subset of the full data set with the inclusion/exclusion criteria that are: (i) We consider only continuous variables and 21 variables are left. (ii) We only consider the samples with no missing values. $141,993$ samples are left. (iii) We exclude variables whose proportions of zero values are more than 40% and, in consequence, 6 variables among 21 are excluded. (iv) We additionally remove the observations that have negative and zero annual salaries. (v) We only consider two sub-populations with jobs 'Police Officer (II, III)' and 'Firefighter III', which are the two most popular public jobs in LA. In the end, we have two sub-datasets of 'Police Officer (II, III)' and 'Firefighter III', which have $19,684$ observations and $5,728$ observations, respectively. The number of variables is 14 and they

are listed below.

Y_1: Hourly or Event Rate (HER)	Y_2: Projected Annual Salary (PAS)
Y_3: Q_1 Payments (Q1P)	Y_4: Q_2 Payments (Q2P)
Y_5: Q_3 Payments (Q3P)	Y_6: Q_4 Payments (Q4P)
Y_7: Payments Over Base Pay (POBP)	Y_9: Total Payments (TP)
Y_{10}: Base Pay (BP)	Y_{17}: Other Pay Payroll Explorer (OPPE)
Y_{18}: Average Health Cost (AHC)	Y_{19}: Average Dental Cost (ADC)
Y_{20}: Average Basic Life (ABL)	Y_{21}: Average Benefit Cost (ABC)

In each sub-dataset, we randomly select $1,000$ samples and treat them as testing data. We use the remaining samples as training data to estimate the sequential regression models. To understand better the performance in estimating the covariance matrix, we consider the detailed measures as:

- the Frobenius risk and the matrix ℓ_2 risk in estimating the covariance matrix as in the numerical study (Table 2),
- the vector ℓ_2 error in estimating the first three principal eigenvectors (Table 3),
- the vector ℓ_2 error in estimating the first three predictive principal component (PC) scores (Table 4).

In the above, the first three predictive (PC) scores are defined as, for test samples $\mathbf{Y}^{\text{test}}$,

$$PC_i^{(\text{test})} = \mathbf{Y}^{(\text{test})}\mathbf{v}_i, \quad i = 1, 2, 3,$$

where $\mathbf{v}_i$ is the i-th eigenvector (of the i-th largest eigenvalue) of the sample covariance matrix S_n of the original data (i.e. training data). Similarly, for the synthesized data and its sample covariance matrix, we can define the estimates of the PC scores as

$$\hat{PC}_i^{(\text{test})} = \mathbf{Y}^{(\text{test})}\hat{\mathbf{v}}_i, \quad i = 1, 2, 3,$$

where $\hat{\mathbf{v}}_i$ is the i-th eigenvector of the sample covariance matrix of the synthesized data.

With these two datasets, we compare the estimation performance of the proposed SynCov and CART in `synthpop`. It's worth noting that we only consider SynCov and CART here because they are two methods that outperformed others (SR and SR-PMD) in the numerical study. In generating synthetic data, we consider $m = n/5, n/2, n, 2n, 5n$, where n is the sample size of the original data and m is that of the synthetic data. We repeat these steps 100 times and report their summaries in Tables 2 – 4.

We have made a few findings from Tables 2–4. First, SynCov outperforms CART in all cases considered in not only the estimation error of the covariance matrix itself but also in the estimation of the first few principal eigenvectors and their PC scores. Second, the performances of the synthetic data are improved as the sample size of the synthetic data m increases in all cases for both SynCov and CART. However, the rate of the improvement per sample decreases as m increases. Finally, interestingly, SynCov with $m = n/5$ (a small size synthetic data) outperforms CART with $m = 5n$ (a large-size synthetic data) in all cases except one case, the matrix ℓ_2 error of the covariance matrix estimation of the Firefighter (III) case. Further, in this single case, the difference in the estimation error is nearly negligible. In summary, the proposed synthetic data generation method provides better covariance matrix estimation in terms of the Frobenius and matrix ℓ_2 errors. This also results in improved estimation performance in estimating the first few eigenstructures of the covariance matrix and principal component analysis.

5.2 Application to Taiwanese Bankruptcy Data

In this section, we apply synthetic data generation methods to the Taiwanese Bankruptcy dataset available at `https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction`. The Taiwanese bankruptcy data consists of 96 variables for 6819 companies in Taiwan, collected from 1999 to 2009. Among the 96 variables, the variable `Bankrupt` denotes the bankruptcy status of the company, where 1 indicates that the corresponding company was bankrupt and 0 indicates non-bankruptcy. The other 95 variables are related to the financial status of companies, such as the cash flow rate, the operating profit growth rate, and the net value growth rate. Among the 95 variables, there are two categorical variables: the liability assets flag and the net income flag. The liability assets flag indicates whether the company has liability assets, while the net income flag indicates whether the company has positive net income. In this application, we exclude these two variables because all companies have positive net income values and only 8 out of 6819 companies have liability assets. Therefore, we have one binary variable indicating bankruptcy and 93 continuous financial status-related variables reported in Table 5. In Table 5, we denote 24 variables having large variances with a symbol (*) and apply the logarithmic transformation $f(x) = \log(1 + x)$ to these variables to stabilize the variance for further analysis.

In this application, we consider the LDA to illustrate the advantage of the synthetic data generation method in enhancing the covariance estimation. Specifically, let B_i be the bankruptcy indicator variable of the i-th company and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$ be a vector of

the financial-status related variables of the i-th company. Suppose that $\mathbf{Z}|B = 0 \sim N(\boldsymbol{\mu}_0, \Sigma)$ and $\mathbf{Z}|B = 1 \sim N(\boldsymbol{\mu}_1, \Sigma)$. Then, the LDA decision rule for the company's bankruptcy with the observed vector $\mathbf{z}$ is represented as

$$\mathbf{z}^\top \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) > \frac{1}{2}\boldsymbol{\mu}_1^\top \Sigma^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu}_0 + \log(n_0/n) - \log(n_1/n), \tag{20}$$

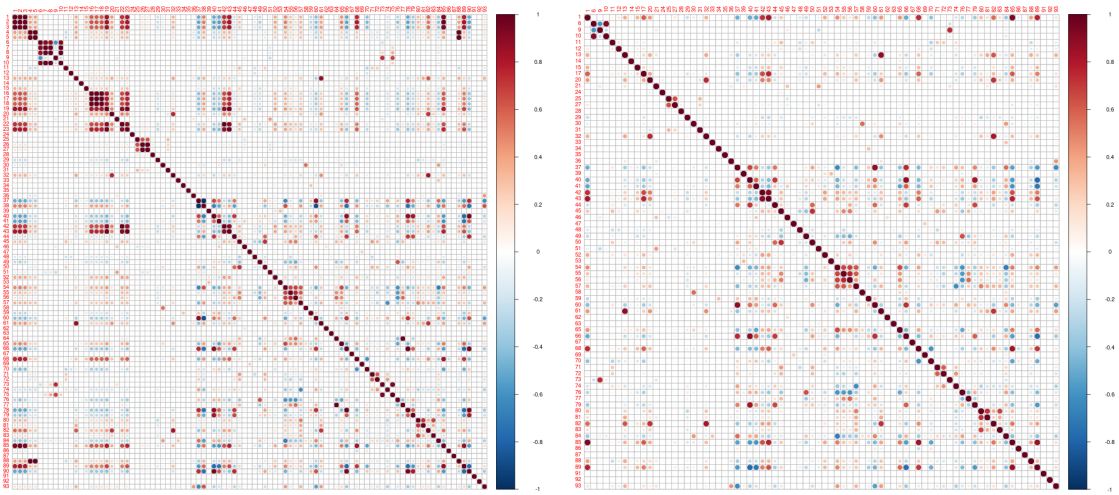
where n is the total number of companies, n_0 is the number of non-bankrupt companies, and n_1 is the number of bankrupt companies. To apply the LDA, we first examine the correlation structure of the Taiwanese bankruptcy data, where the financial status-related variables are highly correlated in general, and it could make the sample covariance matrix nearly singular. We illustrate the correlation map of the 93 variables in Figure 1(a). As depicted in Figure 1 (a), there are several groups of highly correlated variables. For instance, Z_1, Z_2, Z_4, and Z_{85} exhibit high correlations, with values ranging between 0.9327 and 0.9917. To identify these groups, we construct a network of 93 nodes (i.e., variables) with an edge that connects two nodes when the absolute correlation of two nodes is greater than 0.95. We depict the networks of all 93 nodes and highly correlated variables in Figure 2. From the network representation, we found 11 groups of highly correlated variables and excluded 17 variables (Z_2, Z_3, Z_4, Z_5, Z_7,Z_8,Z_{16}, Z_{18},Z_{19},Z_{22}, Z_{23},Z_{26},Z_{38},Z_{64},Z_{75},Z_{78},Z_{90}) that have redundant information and make the sample covariance matrix singular. We illustrate the correlation map of the remaining 76 variables in Figure 1(b).

To evaluate the performance of LDA, we divide the 6819 companies into training and testing datasets. Given that the number of bankrupt companies (220) is considerably smaller than the number of non-bankrupt companies (6659), we randomly select 25 companies from each class, where the class indicates the bankruptcy status of the companies. Then, we consider the following four methods:

M1: (Original) The decision rule of the LDA is estimated using the original data, comprising 6634 non-bankrupt and 195 bankrupt companies.

M2: (Original-Up) The decision rule of the LDA is estimated using the original data with upsampling on the bankrupt company class, resulting in 6634 non-bankrupt and 6634 bankrupt companies. Note that upsampling is done by randomly resampling the bankrupt class observations.

M3: (CART-Up) The decision rule of the LDA is estimated using synthetic data obtained by CART with upsampling on the bankrupt company class, resulting in 6634 non-bankrupt and 6634 bankrupt companies.



(a) All 93 variables (b) Chosen 76 variables

Figure 1: Correlation maps of all variables and chosen variables used in the linear discriminant analysis in Taiwanese bankruptcy data.

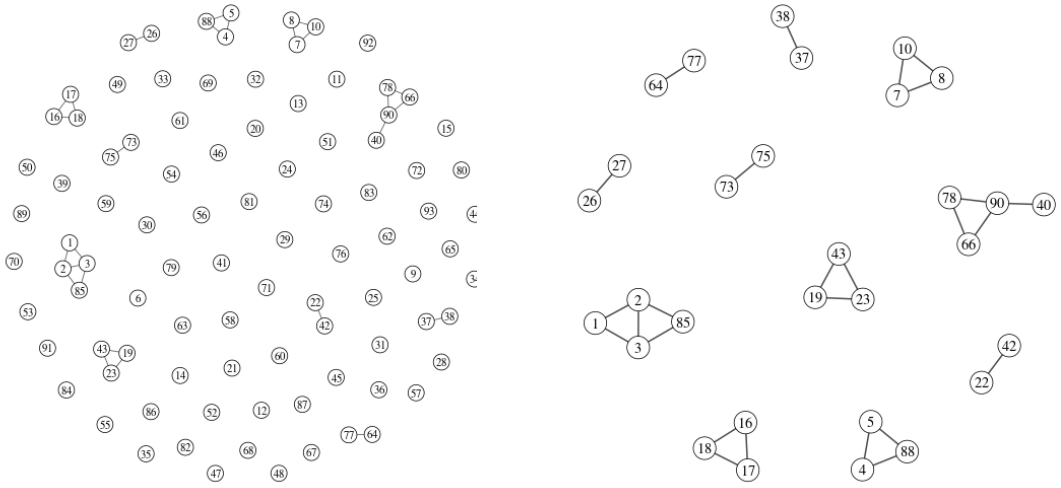
M4: (SynCov-Up) The decision rule of the LDA is estimated using synthetic data obtained by SynCov with upsampling on the bankrupt company class, resulting in 6634 non-bankrupt and 6634 bankrupt companies.

To generate synthetic data for the Taiwanese bankruptcy dataset, we apply the synthetic data generation method to each class dataset and then combine the two datasets. For **M3**, we use the R package `synthpop` with the CART method. For **M4**, we apply our own R functions for the SynCov For **M3** and **M4**, we generate additional observations for the bankrupt class from the synthetic data generation methods to match the number of non-bankrupt class observations; we refer to this approach as *upsampling*.

For the performance measures of the LDA, we consider four metrics: accuracy, sensitivity (true positive rate), specificity (true negative rate), and F_1 score, defined as follows:

$$\text{ACC} = \frac{TP+TN}{P+N},\ \text{SEN} = \frac{TP}{P},\ \text{SPC} = \frac{TN}{N},\ \text{F}_1 = \frac{2TP}{2TP+FP+FN},$$

where P is the number of positives ($B = 1$), N is the number of negatives ($B = 0$), TP (TN) is the number of true positives (negatives), and FP (FN) is the number of false positives (negatives). We evaluate these four metrics using the prediction results for the testing data. Since the results may vary depending on the selection of the testing data, we repeat the described evaluation procedure 100 times.



(a) All 93 variables (b) Highly correlated groups

Figure 2: Network graphs of all variables and identified groups of highly correlated variables in Taiwanese bankruptcy data.

We report the summary of the four performance measures in Table 6. First, by comparing the results of the original data (**M1**) and the original data with upsampling (**M2**), we observe that the class imbalance significantly affects the prediction performance of the LDA, as shown in Table 6. Moreover, by comparing the results of **M2**, **M3**, **M4**, we observe that our SynCov with upsampling (**M4**) slightly enhances prediction performance across all four metrics compared to the original data with upsampling (**M2**), while the CART with upsampling (**M3**) performs significantly worse than **M2** and **M4**.

Table 1: Performance of covariance matrix estimation by SR, SynCov, SR-PMD, and CART for Models 1–4. In the table, F and ℓ_2 denote the Frobenius and the matrix ℓ_2 loss.

n	p	Method	Model 1		Model 2		Model 3		Model 4	
			F	ℓ_2	F	ℓ_2	F	ℓ_2	F	ℓ_2
250	25	SynCov	0.35	0.22	2.26	1.17	2.65	2.55	2.44	1.84
			(0.04)	(0.05)	(0.17)	(0.20)	(1.46)	(1.51)	(0.48)	(0.62)
		SR	0.36	0.23	2.43	1.35	2.53	2.42	2.46	1.83
			(0.04)	(0.05)	(0.21)	(0.25)	(1.37)	(1.43)	(0.48)	(0.64)
		SR-PMD	0.36	0.23	2.38	1.28	4.30	4.12	2.46	1.83
			(0.04)	(0.06)	(0.20)	(0.23)	(2.86)	(2.94)	(0.43)	(0.58)
		CART	0.37	0.23	2.29	1.21	2.64	2.53	2.61	2.00
			(0.04)	(0.06)	(0.18)	(0.21)	(1.37)	(1.43)	(0.48)	(0.64)
	50	SynCov	0.41	0.24	4.43	1.87	5.36	4.82	4.69	2.52
			(0.03)	(0.05)	(0.19)	(0.24)	(2.34)	(2.46)	(0.30)	(0.52)
		SR	0.43	0.25	5.10	2.38	5.52	5.00	5.04	2.65
			(0.04)	(0.05)	(0.27)	(0.30)	(2.44)	(2.56)	(0.35)	(0.62)
		SR-PMD	0.43	0.25	4.96	2.26	11.47	10.80	4.93	2.55
			(0.04)	(0.06)	(0.25)	(0.28)	(7.81)	(8.05)	(0.28)	(0.49)
		CART	0.42	0.25	4.27	1.85	5.27	4.69	4.58	2.66
			(0.04)	(0.06)	(0.19)	(0.25)	(2.09)	(2.22)	(0.31)	(0.51)
	100	SynCov	0.48	0.25	8.80	3.14	13.38	11.56	9.14	3.57
			(0.03)	(0.05)	(0.26)	(0.31)	(4.23)	(4.54)	(0.22)	(0.50)
		SR	0.51	0.27	11.40	4.42	14.17	12.32	11.15	4.13
			(0.04)	(0.06)	(0.38)	(0.38)	(4.41)	(4.70)	(0.31)	(0.61)
		SR-PMD	0.50	0.27	10.95	4.16	33.15	30.40	10.84	3.92
			(0.04)	(0.06)	(0.35)	(0.36)	(16.89)	(17.28)	(0.28)	(0.53)
		CART	0.46	0.26	7.95	2.76	11.34	8.98	8.31	3.65
			(0.04)	(0.06)	(0.21)	(0.26)	(3.05)	(3.22)	(0.25)	(0.52)
500	25	SynCov	0.25	0.16	1.61	0.81	1.81	1.73	1.69	1.27
			(0.03)	(0.04)	(0.11)	(0.13)	(0.97)	(1.01)	(0.33)	(0.43)
		SR	0.25	0.16	1.68	0.89	1.86	1.78	1.74	1.31
			(0.03)	(0.04)	(0.13)	(0.15)	(1.00)	(1.04)	(0.34)	(0.45)
		SR-PMD	0.25	0.16	1.65	0.86	3.21	3.10	1.70	1.28
			(0.03)	(0.04)	(0.13)	(0.15)	(2.25)	(2.30)	(0.29)	(0.40)
		CART	0.26	0.16	1.67	0.88	1.87	1.79	1.91	1.48
			(0.03)	(0.04)	(0.12)	(0.15)	(1.02)	(1.06)	(0.38)	(0.50)
	50	SynCov	0.29	0.17	3.17	1.27	3.69	3.35	3.29	1.80
			(0.02)	(0.04)	(0.13)	(0.15)	(1.53)	(1.60)	(0.22)	(0.40)
		SR	0.30	0.17	3.40	1.48	3.78	3.42	3.40	1.82
			(0.03)	(0.04)	(0.15)	(0.18)	(1.59)	(1.69)	(0.22)	(0.41)
		SR-PMD	0.30	0.17	3.35	1.44	9.17	8.79	3.36	1.79
			(0.02)	(0.04)	(0.14)	(0.17)	(6.92)	(7.06)	(0.22)	(0.38)
		CART	0.30	0.17	3.12	1.32	3.70	3.29	3.37	2.01
			(0.03)	(0.04)	(0.13)	(0.17)	(1.43)	(1.52)	(0.25)	(0.43)
	100	SynCov	0.34	0.18	6.23	2.01	8.13	6.81	6.38	2.48
			(0.02)	(0.04)	(0.14)	(0.17)	(2.41)	(2.57)	(0.15)	(0.36)
		SR	0.35	0.18	7.14	2.52	8.50	7.18	7.07	2.63
			(0.02)	(0.04)	(0.19)	(0.19)	(2.57)	(2.73)	(0.18)	(0.42)
		SR-PMD	0.34	0.18	7.05	2.46	24.79	23.25	7.00	2.57
			(0.02)	(0.04)	(0.19)	(0.19)	(13.77)	(14.05)	(0.17)	(0.39)
		CART	0.33	0.18	5.77	1.91	7.93	6.19	6.04	2.65
			(0.02)	(0.04)	(0.13)	(0.17)	(2.12)	(2.21)	(0.17)	(0.35)

Table 2: The error in covariance matrix estimation error by SynCov and the CART using `synthpop` for Police Officer (II, III) and Firefighter (III) in the LA Employee Payroll data. F and ℓ_2 denote the Frobenius and matrix ℓ_2 distance between the sample covariance matrix of the synthetic data and that of the original data, respectively.

Job	Norm	Method	Synthetic data sample size m				
			$m = n/5$	$n/2$	n	$2n$	$5n$
Police officer (II, III)	F	SynCov	0.1574 (0.0066)	0.1023 (0.0038)	0.0730 (0.0028)	0.0533 (0.0021)	0.0336 (0.0013)
		CART	0.7130 (0.0363)	0.4693 (0.0226)	0.3509 (0.0148)	0.2913 (0.0128)	0.2062 (0.0070)
	ℓ_2	SynCov	0.1416 (0.0070)	0.0916 (0.0042)	0.0660 (0.0030)	0.0482 (0.0022)	0.0303 (0.0014)
		CART	0.6514 (0.0384)	0.4232 (0.0243)	0.3109 (0.0159)	0.2621 (0.0137)	0.1752 (0.0076)
Firefighter (III)	F	SynCov	0.3359 (0.0100)	0.2079 (0.0069)	0.1357 (0.0047)	0.1045 (0.0033)	0.0706 (0.0024)
		CART	1.3738 (0.0487)	0.8467 (0.0297)	0.6489 (0.0248)	0.4650 (0.0145)	0.3412 (0.0107)
	ℓ_2	SynCov	0.2913 (0.0109)	0.1777 (0.0074)	0.1142 (0.0049)	0.0887 (0.0034)	0.0616 (0.0026)
		CART	1.2199 (0.0512)	0.7385 (0.0302)	0.5598 (0.0259)	0.3979 (0.0149)	0.2876 (0.0114)

Table 3: The vector ℓ_2 error between the first three principal component vectors $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v_3})$ of the original data and those of the synthetic data for the Police Officer (II, III) and Firefighter (III) in the LA Employee Payroll data.

Job	Eigenvector	Method	Synthetic data sample size m				
			$n/5$	$n/2$	n	$2n$	$5n$
Police officer (II, III)	$\mathbf{v}_1$	SynCov	0.0126 (0.0005)	0.0082 (0.0003)	0.0055 (0.0002)	0.0040 (0.0002)	0.0026 (0.0001)
		CART	0.0530 (0.0030)	0.0334 (0.0013)	0.0273 (0.0012)	0.0213 (0.0007)	0.0172 (0.0006)
	$\mathbf{v}_2$	SynCov	0.0298 (0.0014)	0.0199 (0.0008)	0.0121 (0.0005)	0.0102 (0.0005)	0.0060 (0.0002)
		CART	0.1179 (0.0047)	0.0885 (0.0034)	0.0787 (0.0026)	0.0754 (0.0021)	0.0720 (0.0019)
	$\mathbf{v}_3$	SynCov	0.0381 (0.0015)	0.0252 (0.0008)	0.0164 (0.0007)	0.0132 (0.0005)	0.0079 (0.0003)
		CART	0.1529 (0.0055)	0.1136 (0.0040)	0.0923 (0.0028)	0.0826 (0.0020)	0.0760 (0.0019)
Firefighter (III)	$\mathbf{v}_1$	SynCov	0.0372 (0.0018)	0.0232 (0.0012)	0.0159 (0.0008)	0.0120 (0.0006)	0.0080 (0.0004)
		CART	0.1589 (0.0081)	0.1095 (0.0061)	0.0817 (0.0042)	0.0563 (0.0027)	0.0429 (0.0021)
	$\mathbf{v}_2$	SynCov	0.0492 (0.0018)	0.0288 (0.0012)	0.0215 (0.0008)	0.0154 (0.0005)	0.0102 (0.0004)
		CART	0.1962 (0.0091)	0.1246 (0.0058)	0.0954 (0.0040)	0.0690 (0.0026)	0.0546 (0.0019)
	$\mathbf{v}_3$	SynCov	0.0770 (0.0032)	0.0473 (0.0017)	0.0354 (0.0014)	0.0245 (0.0010)	0.0145 (0.0005)
		CART	0.2897 (0.0139)	0.1997 (0.0096)	0.1415 (0.0068)	0.1107 (0.0052)	0.0894 (0.0030)

Table 4: The vector L_2 norm of the difference between the first three principal component scores $\left(PC_1^{(\text{test})} = \mathbf{Y}^{(\text{test})}\mathbf{v}_1, PC_2^{(\text{test})} = \mathbf{Y}^{(\text{test})}\mathbf{v}_2, PC_3^{(\text{test})} = \mathbf{Y}^{(\text{test})}\mathbf{v}_3\right)$ of the test data using the principal components of the original data and those of the synthetic data for the Police Officer (II, III) and Firefighter (III) in the LA Employee Payroll data.

Job	PC score	Method	Synthetic data sample size m				
			$n/5$	$n/2$	n	$2n$	$5n$
Police officer (II, III)	$\mathbf{Y}^{(\text{test})}\mathbf{v}_1$	SynCov	0.3712	0.2398	0.1536	0.1149	0.0758
			(0.0212)	(0.0128)	(0.0088)	(0.0064)	(0.0042)
		CART	1.4964	0.9180	0.7475	0.5320	0.4492
			(0.1169)	(0.0503)	(0.0480)	(0.0253)	(0.0207)
	$\mathbf{Y}^{(\text{test})}\mathbf{v}_2$	SynCov	0.9802	0.6466	0.3771	0.3151	0.1908
			(0.0479)	(0.0284)	(0.0198)	(0.0157)	(0.0087)
		CART	3.8425	2.7118	2.3798	2.1181	2.0007
			(0.2209)	(0.1077)	(0.0929)	(0.0668)	(0.0531)
	$\mathbf{Y}^{(\text{test})}\mathbf{v}_3$	SynCov	1.0458	0.7121	0.4282	0.3836	0.2142
			(0.0511)	(0.0312)	(0.0204)	(0.0196)	(0.0095)
		CART	4.2420	3.2104	2.8753	2.7586	2.6298
			(0.1913)	(0.1380)	(0.1021)	(0.0819)	(0.0713)
Firefighter (II, III)	$\mathbf{Y}^{(\text{test})}\mathbf{v}_1$	SynCov	1.5598	0.9401	0.6456	0.4863	0.3407
			(0.0987)	(0.0657)	(0.0404)	(0.0296)	(0.0224)
		CART	6.6879	4.6253	3.2639	2.3336	1.7397
			(0.4555)	(0.3191)	(0.2186)	(0.1542)	(0.1159)
	$\mathbf{Y}^{(test)}\mathbf{v}_2$	SynCov	2.3925	1.3909	0.9833	0.7364	0.5182
			(0.1374)	(0.0951)	(0.0592)	(0.0428)	(0.0317)
		CART	9.6011	6.6258	4.7722	3.4706	2.5950
			(0.6233)	(0.4730)	(0.3175)	(0.2114)	(0.1649)
	$\mathbf{Y}^{(\text{test})}\mathbf{v}_3$	SynCov	2.1368	1.2988	0.9588	0.6889	0.3991
			(0.0906)	(0.0412)	(0.0332)	(0.0259)	(0.0151)
		CART	8.1840	5.3122	3.9970	2.9909	2.4760
			(0.3798)	(0.2233)	(0.1832)	(0.1216)	(0.0848)

Table 5: Summary of variables in Taiwanese bankruptcy data.

B: Bankruptcy	Z_1: ROA C before interest and depreciation before interest
Z_2: ROA A before interest and after tax	Z_3: ROA B before interest and depreciation after tax
Z_4: Operating Gross Margin	Z_5: Realized Sales Gross Margin
Z_6: Operating Profit Rate	Z_7: Pre tax net Interest Rate
Z_8: After tax net Interest Rate	Z_9: Non industry income and expenditure revenue
Z_{10}: Continuous interest rate after tax	Z^*_{11}: Operating Expense Rate
Z^*_{12}: Research and development expense rate	Z_{13}: Cash flow rate
Z^*_{14}: Interest bearing debt interest rate	Z_{15}: Tax rate A
Z_{16}: Net Value Per Share B	Z_{17}: Net Value Per Share A
Z_{18}: Net Value Per Share C	Z_{19}: Persistent EPS in the Last Four Seasons
Z_{20}: Cash Flow Per Share	Z^*_{21}: Revenue Per Share Yuan
Z_{22}: Operating Profit Per Share Yuan	Z_{23}: Per Share Net profit before tax Yuan
Z_{24}: Realized Sales Gross Profit Growth Rate	Z_{25}: Operating Profit Growth Rate
Z_{26}: After tax Net Profit Growth Rate	Z_{27}: Regular Net Profit Growth Rate
Z_{28}: Continuous Net Profit Growth Rate	Z^*_{29}: Total Asset Growth Rate
Z^*_{30}: Net Value Growth Rate	Z_{31}: Total Asset Return Growth Rate Ratio
Z_{32}: Cash Reinvestment	Z^*_{33}: Current Ratio
Z^*_{34}: Quick Ratio	Z_{35}: Interest Expense Ratio
Z^*_{36}: Total debt Total net worth	Z_{37}: Debt ratio
Z_{38}: Net worth Assets	Z_{39}: Long term fund suitability ratio A
Z_{40}: Borrowing dependency	Z_{41}: Contingent liabilities Net worth
Z_{42}: Operating profit Paid in capital	Z_{43}: Net profit before tax Paid in capital
Z_{44}: Inventory and accounts receivable Net value	Z_{45}: Total Asset Turnover
Z^*_{46}: Accounts Receivable Turnover	Z^*_{47}: Average Collection Days
Z^*_{48}: Inventory Turnover Rate times	Z^*_{49}: Fixed Assets Turnover Frequency
Z_{50}: Net Worth Turnover Rate times	Z^*_{51}: Revenue per person
Z_{52}: Operating profit per person	Z^*_{53}: Allocation rate per person
Z_{54}: Working Capital to Total Assets	Z_{55}: Quick Assets Total Assets
Z_{56}: Current Assets Total Assets	Z_{57}: Cash Total Assets
Z^*_{58}: Quick Assets Current Liability	Z^*_{59}: Cash Current Liability
Z_{60}: Current Liability to Assets	Z_{61}: Operating Funds to Liability
Z_{62}: Inventory Working Capital	Z^*_{63}: Inventory Current Liability
Z_{64}: Current Liabilities Liability	Z_{65}: Working Capital Equity
Z_{66}: Current Liabilities Equity	Z^*_{67}: Long term Liability to Current Assets
Z_{68}: Retained Earnings to Total Assets	Z_{69}: Total income Total expense
Z_{70}: Total expense Assets	Z^*_{71}: Current Asset Turnover Rate
Z^*_{72}: Quick Asset Turnover Rate	Z_{73}: Working capitcal Turnover Rate
Z^*_{74}: Cash Turnover Rate	Z_{75}: Cash Flow to Sales
Z^*_{76}: Fixed Assets to Assets	Z_{77}: Current Liability to Liability
Z_{78}: Current Liability to Equity	Z_{79}: Equity to Long term Liability
Z_{80}: Cash Flow to Total Assets	Z_{81}: Cash Flow to Liability
Z_{82}: CFO to Assets	Z_{83}: Cash Flow to Equity
Z_{84}: Current Liability to Current Assets	Z_{85}: Net Income to Total Assets
Z^*_{86}: Total assets to GNP price	Z_{87}: No credit Interval
Z_{88}: Gross Profit to Sales	Z_{89}: Net Income to Stockholder s Equity
Z_{90}: Liability to Equity	Z_{91}: Degree of Financial Leverage DFL
Z_{92}: Interest Coverage Ratio Interest expense to EBIT	Z_{93}: Equity to Liability

Table 6: Summary of the four performance metrics for Taiwanese bankruptcy data. Numbers in parentheses denote the standard errors.

Method	ACC	SEN	SPC	F_1 score
Original (**M1**)	0.6480	0.3120	0.9840	0.4644
	(0.0040)	(0.0079)	(0.0023)	(0.0091)
Original-Up (**M2**)	0.8342	0.8124	0.8560	0.8297
	(0.0056)	(0.0082)	(0.0078)	(0.0059)
CART-Up (**M3**)	0.8240	0.7964	0.8516	0.8181
	(0.0051)	(0.0079)	(0.0073)	(0.0055)
SynCov-Up (**M4**)	0.8380	0.8184	0.8576	0.8342
	(0.0053)	(0.0075)	(0.0075)	(0.0055)

6 Conclusion

In this paper, we propose to shrink the error variances in generating synthetic data based on the sequential regression model. We prove that our new synthetic data generator, named SynCov, provides a better covariance matrix estimator than the original without shrinking. We numerically show that the improved covariance matrix further (i) provides better estimates of the first few eigenstructures of the covariance matrix, which are important in the principal component analysis, and also (ii) introduces an enhanced classifier performing better than that with the existing synthetic data generators. We illustrate our SynCov with two real data examples, the LA Employee Payroll Data and the Taiwanese Bankruptcy Data.

We remark that our SynCov is a synthetic data generator aiming for a specific tailored purpose, the estimation of the covariance matrix, while all existing synthetic data generators are for general and universal purposes. The covariance matrix plays an important role in many multivariate procedures and, we expect that the synthetic data by our SynCop also perform better than that by the original (unshrinking) synthetic data in subsequent multivariate procedures including the testing of the mean vectors or covariance matrices of two or more populations and the classification of observations. We numerically investigate this in this paper. However, it should be further investigated with pencils.

Two additional remarks on our SynCov are as follows. First, our SynCov can straightforwardly be applied to synthetic data generation of mixed data containing both categorical variables and continuous variables, if the number of categorical variables is not many. It can be done by first generating categorical variables and next generating continuous variables for each case of categorical variables. Second, our SynCov in this paper only considers the cases with $n > p$ to apply sequential regression models. However, for the cases $p \geq n$, we adopt sequential regularized regressions which are closely connected to the modified Cholesky decomposition-based estimation of the covariance matrix in the literature. We leave this as a part of future work.

Acknowledgements

J. Lim's research is supported by the National Research Foundation of Korea (NRF-2021R1A2C1010786), and D. Yu's research is supported by the National Research Foundation of Korea (NRF-2022R1A5A7033499, NRF-2020S1A5C2A02093223) and Inha University Research Grant.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Chapman and Hall, New York.

Chen, Y., Wiesel, A., Eldar, Y. C., and Hero, A. O. (2010). Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing*, 58:5016–5029.

Drechsler, J. and Reiter, J. (2011). An empirical evaluation of easily implemented, non-parametric methods for generating synthetic datasets. *Computational Statistics and Data Analysis*, 55:3232–3243.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Kang, X. and Deng, X. (2020). An improved modified cholesky decomposition approach for precision matrix estimation. *Journal of Statistical Computation and Simulation*, 90(3):443–464.

Kang, X., Deng, X., Tsui, K.-W., and Pourahmadi, M. (2020a). On variable ordination of modified cholesky decomposition for estimating time-varying covariance matrices. *International Statistical Review*, 88(3):616–641.

Kang, X., Xie, C., and Wang, M. (2020b). A cholesky-based estimation for large-dimensional covariance matrices. *Journal of Applied Statistics*, 47(6):1017–1030.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Letac, G. and Massam, H. (2004). All invariant moments of the wishart distribution. *Scandinavian Journal of Statistics*, 31:285–318.

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

Ma, C., Tschiatschek, S., Hernández-Lobato, J., Turner, R., and Zhang, C. (2020). Vaem: a deep generative model for heterogeneous mixed type data. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 11237–11247.

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74:1–26.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB*, 11:1071–1083.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86:677–690.

Raghunathan, T., Lepkowski, J., Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiple imputation missing values using a sequence of regression models. *Survery Methodology*, 27:85–95.

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

Rajaratnam, B. and Salzman, J. (2013). Best permutation analysis. *Journal of Multivariate Analysis*, 121:193–223.

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–189.

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

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

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

Rothman, A., Levina, E., and Zhu, J. (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97:539–550.

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

Smith, D., Elliot, M., and Sakshaug, J. W. (2023). To link or synthesize? an approach to data quality comparison. *Journal of Data and Information Quality*, 15(2):1–20.

Taub, J. and Elliot, M. J. (2019). The synthetic data challenge. *UNECE: Conference of European Statisticians*.

Taub, J., Elliot, M. J., Pampaka, M., and Smith, D. (2018). Differential correct attribution probability for synthetic data: An exploration. In *Privacy in Statistical Databases*, pages 122–137.

Tomczak, J. and Welling, M. (2017). Vae with a vamp prior. *International Conference of Artificial Intelligence and Statistics*, pages 1214–1223.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., and McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *NIPS'19: Proceedings of the 33th International Conference on Neural Information Processing Systems*, pages 7335–7345.

Zhao, Z., Kunar, A., Birke, R., and Chen, L. (2021). Ctab-gan: effective table data synthesizing. *Asian Conference on Machine Learning*, pages 97–112.