Pak. J. Statist.
2025 Vol. 41(3), 207-222

ENHANCING POPULATION MEAN ESTIMATION THROUGH PCA-BASED ESTIMATORS: ADDRESSING MULTICOLLINEARITY AND DIMENSIONALITY CHALLENGES IN SURVEY SAMPLING

Uzma Nasir and **Zahoor Ahmad**[§]
Department of Statistics, University of Sargodha
Sargodha, Pakistan
[§]Email: zahoor.ahmad@uos.edu.pk

ABSTRACT

Estimating population parameters, such as the finite population mean, becomes challenging when multiple auxiliary variables are involved due to multicollinearity. High correlation among these variables inflates errors and reduces the efficiency and stability of traditional regression-based estimators. To address this, we propose Principal Component Analysis (PCA)-based estimator, which transform correlated auxiliary variables into a smaller set of uncorrelated principal components ranked by their explanatory power. This dimensionality reduction improves estimation precision and mitigates multicollinearity. A synthetic correlated data is generated for fifty auxiliary variables and comparison based on theoretical mean square errors is made using the population data. A simulation study is also presented to evaluate and validate the effectiveness of the PCA-based approach compared to classical methods. Keywords and phrases: Finite population mean; Auxiliary variables; Multi collinearity; Ratio estimators; Dimensionality reduction; Estimator stability; PCA-based estimator.

1. INTRODUCTION

One of the core goals in survey sampling is the estimation of the population mean, a critical summary measure for understanding the distribution of data. Typically, the sample mean is used as an estimator for the population mean, provided that the sampling design is sound. The accuracy of this estimate, however, is influenced by several factors, including sample size, the inherent variability within the population, and the availability of auxiliary information. Traditional methods for estimating the population mean include the expansion estimator, ratio estimator, and regression estimator. Ratio and regression estimators are particularly valued because they utilize auxiliary variables to reduce the mean squared error (MSE) of the estimate (Singh & Vishwakarma, 2008; Bahl & Tuteja, 1991). The integration of auxiliary information refines the estimation process by exploiting the statistical relationship between the study variable and supplementary data (Isaki & Fuller, 1982; Särndal et al., 2003).

Auxiliary variables are additional variables that are not the primary focus of study but are statistically correlated with the study variable. These variables are frequently used in survey sampling to improve the accuracy and efficiency of statistical inference. Their power is primarily a function of the strength of their association with the target variable the 1 stronger the association, the greater their role in improving estimation accuracy

© 2025 Pakistan Journal of Statistics

(Singh and Tailor, 2003). In the case of estimating household income, for example, auxiliary variables like education, employment, and ownership of property may provide useful context, hence resulting in more accurate estimates (Lavrakas, 2008).

If auxiliary variables are closely correlated with the study variable, they can reduce sampling variability significantly and enhance the precision of population mean estimates (Särndal et al., 2003; Singh and Tailor, 2003). Ratio and regression estimators that utilize such auxiliary variables provide more efficient and accurate results compared to the simple sample mean estimator (Bahl and Tuteja, 1991). Additionally, auxiliary variables assist with problems such as nonresponse biasing by making it possible for post-stratification or reweighting of the responses, thereby reducing bias within final estimates (Alam et al., 2021). Where data is missing, they can be utilized to estimate missing values, maintaining the quality of the dataset (Little and Rubin, 2019). Auxiliary variables also facilitate better design of sampling plans by specifying strata or determining suitable sampling weights. Also, auxiliary variables allow for the adjustment of survey weights so that the estimates become more representative of population characteristics due to the adjustment for non-representative samples (Särndal et al., 2003).

The inclusion of several auxiliary variables in survey sampling has the potential to substantially improve the accuracy and precision of population parameter estimates. Being used optimally, they minimize bias and variance while providing a more detailed insight into the relationships between the study variable and its predictors. Unique information from each of the auxiliary variables is combined to increase estimation efficiency. Research here developed from initial contributions of Raj (1965) and Srivastava (1970), which showed that the use of several auxiliary variables helps in minimizing bias and variance. Improvements have come in the form of multi-auxiliary regression estimators, calibration estimators, and generalized multivariate techniques meant for complicated survey designs. Swain (1970) further developed this work by developing generalized and mixture estimators combining different forms of auxiliary data like ranks and product estimates—enhancing further the stability of estimators, particularly with some auxiliary variables having weaker relationships.

Ahmad et al. (2009) developed three types of regression-cum-ratio estimators for two stage sampling covering cases of partial or missing auxiliary data. Hanif et al. (2009) generalized this framework to construct multivariate ratio estimators for estimating more than one variable under two- and multi-phase sampling designs. These works included both theoretical derivations and empirical comparisons to validate the effectiveness of these techniques. Following these developments, Ahmad and co-authors (2010–2014) presented different generalized regression and ratio estimators to deal with nonresponse and measurement error in multi-auxiliary environments. Their simulation findings and theoretical derivations illustrated remarkable improvement in estimation efficiency. Subsequently, Ayaz et al. (2018) came up with a generalized multivariate exponential-type estimator, which worked well even in the case of incomplete population information.

Current advances have also seen integrated calibration approaches under nonresponse 2 conditions. For instance, Alam et al. (2021) and Ozgul (2021) developed methods employing multiple auxiliary variables to improve robustness using calibrated weights for respondent and non-respondent units. These approaches have been found to be effective in

preserving estimator efficiency even in difficult survey situations. However, the use of multiple auxiliary variables presents a serious challenge: multicollinearity. Scholars such as Kmenta (1971) and Gujarati (2004) have underscored the effect multicollinearity has in magnifying standard errors, undermining regression-based estimators and even impairing ability to determine each variable's contribution. Techniques such as ridge regression (Hoerl and Kennard, 1970) have been proposed as a remedy against multicollinearity at the expense of bringing in bias.

The higher the number of auxiliary variables, the more cumbersome is the process of estimation. Multicollinearity can severely reduce the precision and reliability of traditional estimators, making them impractical for real-world applications that involve numerous correlated predictors. To address the limitations introduced by multicollinearity in the use of multiple auxiliary variables, this study proposes a Principal Component Analysis (PCA) based approach for estimating the finite population mean. PCA provides a strong solution by converting very highly correlated auxiliary variables into a new but uncorrelated set of principal components and thus solving the estimation problem with ease while retaining the lion's share of the variance in the data. Not only does this dimensionality reduction method avoid the instability caused by multicollinearity, but it also improves the accuracy and reliability of the estimates of population means. Drawing on early research like Jolliffe (2002), the study elaborates and tests a PCA-based estimator via simulation as well as practical data implementation. By testing the performance of such an estimator in comparison with the standard regression-based estimators, the study endeavors to provide empirical support for the value added from the use of PCA to increase estimation efficiency with more than one auxiliary variable correlated. The overall aim is to provide methodological suggestions for the use of PCA-based approaches in real-world survey sampling situations where efficiency, stability, and accuracy are critical.

2. PRINCIPAL COMPONENT ANALYSIS BASED ESTIMATOR

The ratio estimator is a widely used method in survey sampling, designed to estimate the population mean of a study variable by incorporating auxiliary information. Its strength lies in the assumption that the auxiliary variables are strongly correlated with the variable of interest, thereby improving the efficiency of the estimator. The traditional power ratio estimator that uses multiple auxiliary variables is mathematically expressed as

$$t_x = \bar{y}\prod_{i=1}^{p}\left(\frac{\bar{X}_i}{\bar{x}_i}\right)^{\alpha_i}, \tag{2.1}$$

where t_x denotes the estimated population mean of Y using multiple auxiliary variables, $\bar{y}$ is the mean of the study variable Y in the sample, $\bar{X}_i$ and $\bar{x}_i$ denote the known population mean and sample mean, respectively, of the i-th auxiliary variable X_i, α_i are optimizing constants, and p represents the total number of auxiliary variables. This formulation assumes that the auxiliary variables are available for the entire population and that simple random sampling without replacement has been employed.

Even though the ratio estimator is intuitively easy and efficient, it is significantly limited when it is applied to more than one auxiliary variable. Under these conditions,

multicollinearity with the auxiliary variables being strongly related to one another can critically mislead the process of estimation such that standard errors become inflated while estimates are unstable. In addition, as the auxiliary space increases in dimensionality, weak or redundant variables could weaken the accuracy of the estimator. These disadvantages call for stable alternatives with the capability to handle multicollinearity and decrease the effective dimension of the problem.

The PCA provides a good solution to these problems. The PCA is a statistical method that converts a collection of correlated auxiliary variables into a new collection of linearly uncorrelated variables known as principal components. These components are ranked by the proportion of variance they capture in the original data. By keeping only the most informative components, PCA reduces dimensionality, removes redundancy, and solves multicollinearity, thus resulting in more stable and efficient population mean estimators.

To formalize the estimation framework, the following notation is used. Let Y be the variable of interest, with finite population mean denoted by μ. Let $X=[X_1, X_2, \ldots, X_p]$ represent the matrix of p auxiliary variables for the population. The sample size is denoted by n, and $Z=[Z_1, Z_2, \ldots, Z_k]$ represents the matrix of transformed principal components derived from X, where $k \leq p$ is the number of retained components after dimensionality reduction.

The development of a PCA-based ratio estimator begins by standardizing the auxiliary variables to ensure each has a mean of zero and a variance of one. This step is crucial when auxiliary variables are measured on different scales. The standardized value of the (i, j)-th observation is calculated using

$$X_{ij}^{*}=\frac{X_{ij}-\bar{X}_j}{S_{X_j}}, \quad \text{for } i=1,2,\ldots,N \text{ and } j=1,2,\ldots,p,$$

where $\bar{X}_j=\frac{1}{N}\sum_{i=1}^{N} X_{ij}$ is the sample mean and $S_{X_j}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{ij}-\bar{X}_j)^2}$ is the sample standard deviation of the j-th auxiliary variable.

After standardization, the covariance matrix of the auxiliary variables is computed to capture the linear relationships among them. This matrix is given by

$$\Sigma_X=\frac{1}{N}X^{\top}X,$$

where $X^{\top}$ is the transpose of the standardized matrix X, and N is the number of observation. The next step involves eigen decomposition of the covariance matrix:

$$\Sigma_X=P\Lambda P^{\top},$$

where P is the matrix of eigenvectors, representing the directions of the new orthogonal axes (principal components), and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is the diagonal matrix of eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Each eigenvalue indicates the proportion of the total variance explained by its corresponding principal component.

To achieve dimensionality reduction, only the first k principal components that explain a significant portion of the total variance (e.g., 95%) are retained. These components are computed as

$$Z = X\ P_k,$$

where P_k contains the eigenvectors associated with the k largest eigenvalues, and Z is the resulting matrix of transformed variables (principal components). These components serve as new, uncorrelated auxiliary variables, which can then be incorporated into the ratio estimator to mitigate the effects of multicollinearity and enhance stability.

By replacing the original auxiliary variables with the retained principal components, we construct a PCA-based ratio estimator that retains the advantages of auxiliary variable usage while overcoming the limitations of multicollinearity and high-dimensionality. This transformation improves computational efficiency, enhances interpretability, and results in more robust estimates of the population mean in complex survey designs.

The PCA-based ratio estimator, which utilizes the retained k principal components, is formulated as:

$$t_z = \bar{y} \prod_{i=1}^{k} \left(\frac{\bar{Z}_i}{\bar{z}_i} \right)^{\delta_i}, \tag{2.2}$$

where $\bar{y}$ is the sample mean of the study variable Y, $\bar{Z}_i$ and $\bar{z}_i$ are the population and sample means, respectively, of the i-th principal component, and δ_i are optimizing constants. This estimator exploits the orthogonality and dimensional efficiency of principal components to produce more stable and precise estimates compared to conventional ratio estimators.

2.1 Mean Square Error of PCA-based Estimator

In order to assess the accuracy and performance of the PCA-based ratio estimator, it is crucial to analyze the Mean Squared Error (MSE), measuring the average of the squared difference between the true population mean and the estimated population mean. MSE is an important measure of precision of the estimator and can be used to compare how dimensionality reduction by PCA affects estimation precision while controlling for multicollinearity.

The sampling error for the sample mean of Y, denoted by $\bar{e}_y$, is defined as:

$$\bar{e}_y = \bar{y} - \bar{Y},$$

where $\bar{Y}$ is the population mean of Y.

Similarly, the sampling error for the i-th principal component mean is given by:

$$\bar{e}_{z_i} = \bar{z}_i - \bar{Z}_i.$$

Assuming unbiasedness under simple random sampling without replacement, the expected values of the sampling errors are zero:

$$E(\bar{e}_y) = E(\bar{e}_{z_i}) = 0.$$

The variances of the sampling errors are fundamental components in the computation of MSE. The variance of the sampling error for Y is expressed as

$$E(\bar{e}_y^2) = \theta \bar{Y}^2 C_y^2,$$

where $\theta = \frac{1-f}{n}$, with $f = \frac{n}{N}$ being the sampling fraction, $C_y = \frac{S_y}{\bar{Y}}$ is the coefficient of variation of Y, and S_y is the standard deviation of Y.

Similarly, the variance of the sampling error for the i-th principal component is given by

$$E(\bar{e}_{z_i}^2) = \theta \bar{Z}_i^2 C_{z_i}^2,$$

where $C_{z_i} = \frac{S_{z_i}}{\bar{Z}_i}$ is the coefficient of variation of the i-th principal component. The covariance between the sampling errors of Y and the i-th principal component is

$$E(\bar{e}_{z_i}, \bar{e}_y) = \theta \bar{Z}_i \bar{Y} C_{z_i} C_y \rho_{z_i y},$$

where $\rho_{z_i y}$ is the population correlation coefficient between the i-th principal component and the study variable Y.

In addition, the covariance between two distinct principal components Z_i and Z_j is given by

$$E(\bar{e}_{z_i}, \bar{e}_{z_j}) = \theta \bar{Z}_i \bar{Z}_j C_{z_i} C_{z_j} \rho_{z_i z_j},$$

where $\rho_{z_i z_j}$ is the correlation coefficient between the i-th and j-th principal components.

These expressions constitute the essential building blocks for deriving the full MSE expression of the PCA-based ratio estimator. The simplified correlation structure arising from PCA (with many $\rho_{z_i z_j} \approx 0$) makes it easier to evaluate and minimize the error associated with the estimator. Consequently, the PCA-based approach offers a mathematically elegant and computationally efficient solution for improving estimation in the presence of multiple, potentially collinear, auxiliary variables.

To account for variability arising from sampling, we express both $\bar{y}$ and $\bar{z}_i$ in terms of their population means and sampling errors. Specifically, $\bar{y} = \bar{Y} + \bar{e}_y$ *and* $\bar{z}_i = \bar{Z}_i + \bar{e}_{z_i}$. Substituting these into the estimator (2.1) yields

$$t_z = (\bar{Y} + \bar{e}_y)\prod_{i=1}^{k}\left(\frac{\bar{Z}_i}{\bar{Z}_i + \bar{e}_{z_i}}\right)^{\delta_i} = (\bar{Y} + \bar{e}_y)\prod_{i=1}^{k}\left(1 + \frac{\bar{e}_{z_i}}{\bar{Z}_i}\right)^{-\delta_i}.$$

Applying the binomial expansion to each term, where $(1+x)^{-n} \approx 1 - nx + \frac{n(n-1)}{2}x^2 + \cdots$, and retaining only the first-order terms (i.e., neglecting squares and higher-order terms), we approximate

$$t_z \approx (\bar{Y} + \bar{e}_y)\prod_{i=1}^{k}\left(1 - \delta_i \frac{\bar{e}_{z_i}}{\bar{Z}_i}\right).$$

Expanding the product and again retaining only first-order terms

$$t_z \approx (\bar{Y} + \bar{e}_y)\left(1 - \sum_{i=1}^{k}\delta_i \frac{\bar{e}_{z_i}}{\bar{Z}_i}\right).$$

Distributing the terms and ignoring second-order interactions results in

$$t_z \approx \bar{Y} + \bar{e}_y - \sum_{i=1}^{k}\delta_i \frac{\bar{Y}}{\bar{Z}_i}\bar{e}_{z_i}.$$

To obtain the Mean Square Error of the estimator t_z, by definition we can write

$$E\left(t_z - \bar{Y}\right)^2 = E\left(\bar{e}_y - \sum_{i=1}^{k}\delta_i \frac{\bar{Y}}{\bar{Z}_i}\bar{e}_{z_i}\right)^2.$$

This leads to the compact form of the MSE of the PCA-based ratio estimator as

$$\mathrm{MSE}(t_z) = E\left(\bar{e}_y - \sum_{i=1}^{k}\delta_i \frac{\bar{Y}}{\bar{Z}_i}\bar{e}_{z_i}\right)^2. \qquad (2.3)$$

This expression forms the basis for further theoretical derivations and optimization. In particular, it enables one to determine the optimal values of the constants δ_i that minimize the MSE, and to analytically compare the PCA-based estimator's performance with that of traditional ratio and regression estimators under conditions of multicollinearity and auxiliary variable redundancy.

To obtain the optimal values of the weights δ_i that minimize the Mean Square Error (MSE) of the PCA-based ratio estimator, we differentiate the MSE expression with respect to each δ_i and equate the derivatives to zero. Recall from the previous result:

$$\mathrm{MSE}(t_z) = E\left(\bar{e}_y - \sum_{i=1}^{k} \delta_i \frac{\bar{Y}}{\bar{Z}_i} \bar{e}_{z_i} \right)^2 .$$

Taking the derivative of the MSE with respect to δ_i and setting it equal to zero, we obtain:

$$\frac{\partial \mathrm{MSE}(t_z)}{\partial \delta_i} = 2E\left(\bar{e}_y - \sum_{j=1}^{k} \delta_j \frac{\bar{Y}}{\bar{Z}_j} \bar{e}_{z_j} \right)\left(-\frac{\bar{Y}}{\bar{Z}_i} \bar{e}_{z_i} \right) = 0, \quad i = 1,2,\ldots,k.$$

This simplifies to

$$E\left(\bar{e}_y \cdot \frac{\bar{Y}}{\bar{Z}_i} \bar{e}_{z_i} \right) = E\left(\sum_{j=1}^{k} \delta_j \frac{\bar{Y}}{\bar{Z}_j} \bar{e}_{z_j} \cdot \frac{\bar{Y}}{\bar{Z}_i} \bar{e}_{z_i} \right), \quad i = 1,2,\ldots,k.$$

This leads to

$$\frac{\bar{Y}}{\bar{Z}_i} E(\bar{e}_y \bar{e}_{z_i}) = \sum_{j=1}^{k} \delta_j \frac{\bar{Y}^2}{\bar{Z}_i \bar{Z}_j} E(\bar{e}_{z_i} \bar{e}_{z_j}).$$

Substituting the expectations using their variance and covariance forms

$$E(\bar{e}_y \bar{e}_{z_i}) = \theta \bar{Y}\bar{Z}_i C_y C_{z_i} \rho_{z_i y}, \quad E(\bar{e}_{z_i} \bar{e}_{z_j}) = \theta \bar{Z}_i \bar{Z}_j C_{z_i} C_{z_j} \rho_{z_i z_j}.$$

Cancelling out common terms $\theta \bar{Y}^2$ from both sides, we obtain the system of equations

$$C_y \rho_{z_i y} = \sum_{j=1}^{k} \delta_j C_{z_i} C_{z_j} \rho_{z_i z_j}, \quad i = 1,2,\ldots,k.$$

Explicitly writing the system for each i, we get

For $i = 1$,

$$C_y \rho_{z_1 y} = \delta_1 C_{z_1}^2 + \delta_2 C_{z_1} C_{z_2} \rho_{z_1 z_2} + \cdots + \delta_k C_{z_1} C_{z_k} \rho_{z_1 z_k}.$$

For $i = 2$,

$$C_y \rho_{z_2 y} = \delta_1 C_{z_1} C_{z_2} \rho_{z_1 z_2} + \delta_2 C_{z_2}^2 + \cdots + \delta_k C_{z_2} C_{z_k} \rho_{z_2 z_k}.$$

For $i = k$,

$$C_y \rho_{z_k y} = \delta_1 C_{z_1} C_{z_k} \rho_{z_1 z_k} + \delta_2 C_{z_2} C_{z_k} \rho_{z_2 z_k} + \cdots + \delta_k C_{z_k}^2.$$

The system of normal equations derived for the optimal constants δ_i can be compactly expressed using matrix notation. Let us define the following

Let $c \in \mathbb{R}^{k \times 1}$ be the correlation vector between the study variable Y and the principal components

$$c = \begin{bmatrix} \rho_{yz_1} \\ \rho_{yz_2} \\ \vdots \\ \rho_{yz_k} \end{bmatrix},$$

and let $R \in \mathbb{R}^{k \times k}$ be the symmetric correlation matrix among the k retained principal - components

$$R = \begin{bmatrix} 1 & \rho_{z_1 z_2} & \cdots & \rho_{z_1 z_k} \\ \rho_{z_1 z_2} & 1 & \cdots & \rho_{z_2 z_k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{z_1 z_k} & \rho_{z_2 z_k} & \cdots & 1 \end{bmatrix}.$$

Let $a \in \mathbb{R}^{k \times 1}$ denote the scaled vector of optimal weights

$$a = \begin{bmatrix} \delta_1 C_{z_1} \\ \delta_2 C_{z_2} \\ \vdots \\ \delta_k C_{z_k} \end{bmatrix}.$$

Then, the set of linear equations for optimal δ_i can be written in matrix form as $C_y c = Ra.$

Solving this equation yields the optimal vector a, provided that R is invertible:

$$a = C_y R^{-1} c.$$

Alternatively, using the Adjoint matrix and determinant, the solution can be expressed as:

$$a = \frac{C_y \cdot Adj(R) \cdot c}{|R|}.$$

Each element of a corresponds to $\delta_i C_{z_i}$; so solving for δ_i gives:

$$\delta_i = \frac{1}{C_{z_i}} a_i = (-1)^{i+1} \frac{C_y}{C_{z_i}} \cdot \frac{\left|R_{yz_i}\right|}{|R|}, \quad i = 1, 2, \ldots, k,$$

where $\left|R_{yz_i}\right|$ is the determinant of the matrix obtained by replacing the i-th column of R with the vector c.

Substituting these optimal weights into the original MSE expression of the PCA-based estimator, we recall:

$$\mathrm{MSE}(t_z) = \theta \bar{Y}^2 \left[C_y^2 - C_y \sum_{i=1}^{k} \delta_i C_{z_i} \rho_{yz_i} \right].$$

Substituting the value of δ_i into the expression:

$$\mathrm{MSE}(t_z) = \theta \bar{Y}^2 C_y^2 \left[1 - \sum_{i=1}^{k} (-1)^{i+1} \rho_{yz_i} \cdot \frac{\left| \mathrm{R}_{yz_i} \right|}{|\mathrm{R}|} \right].$$

After simplification, we can write,

$$\mathrm{MSE}(t_z) = \frac{\theta \bar{Y}^2 C_y^2}{|\mathrm{R}|} \left[|\mathrm{R}| - \sum_{i=1}^{k} (-1)^{i+1} \rho_{yz_i} \left| \mathrm{R}_{yz_i} \right| \right] = \frac{\theta \bar{Y}^2 C_y^2}{|\mathrm{R}|} \left[|\mathrm{R}| - \left| \mathrm{R}_{y\tilde{z}} \right| \right].$$

According to Ahmad, in Result 2 from Section 4.2.2 of Ahmad (2008) we can write

$$\frac{\left| \mathrm{R}_{y\tilde{z}} \right|}{|\mathrm{R}|} = (1 - \rho^2_{y.\tilde{z}_k}),$$

then $\mathrm{MSE}(t_z)$ expression simplifies to

$$\mathrm{MSE}(t_z) = \theta \bar{Y}^2 C_y^2 \left(1 - \rho^2_{y.\tilde{z}_k} \right),$$

where $\rho^2_{y.\tilde{z}_k}$ denotes the squared multiple correlation coefficient between Y and the retained principal components $\tilde{z}_k$. This final result clearly shows that the efficiency of the PCA-based ratio estimator depends entirely on how well the retained components collectively explain the variability in the study variable.

3. NUMERICAL STUDY

The objective of this numerical study is to conduct an in-depth comparison of the PCA-based regression estimator and the classical regression estimator in estimating the finite population mean. By simulating various data scenarios, we aim to assess how both estimators perform under a range of conditions, focusing on accuracy, efficiency, and stability.

This empirical study is designed to evaluate and compare the performance of the PCA-based ratio estimator and the classical regression estimator for estimating the finite population mean. A synthetic finite population of size $N = 10{,}000$, consisting of 50 auxiliary variables per unit, is generated for this purpose.

The auxiliary variables X are drawn from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where the mean vector $\mu = (10, 10, \ldots, 10)$. The correlation structure of the auxiliary variables is embedded in the covariance matrix Σ, and this structure is varied across

scenarios to reflect low, moderate, and strong multicollinearity conditions. The study variable Y is generated using a linear model $Y = X\beta + \epsilon$, where the regression coefficients are fixed as $\beta = [1.5, 1.5, \ldots, 1.5]^\top$, and the error term ϵ is sampled independently from a normal distribution with zero mean and variance $\sigma^2 = 2500$ (*i.e.*, $\epsilon \sim \mathcal{N}(0, 50^2)$).

In the low multicollinearity setting, the auxiliary variables are only weakly correlated with each other. This is achieved by constructing the covariance matrix Σ such that all off-diagonal elements are set to 0.2, while the diagonal elements remain 1. This corresponds to a uniform pairwise correlation coefficient of $\rho = 0.2$, reflecting near-orthogonal relationships among the auxiliary variables.

In the moderate multicollinearity scenario, the correlation between auxiliary variables is increased to represent a more realistic case where predictors exhibit moderate linear dependencies. Here, the covariance matrix Σ is specified with off-diagonal entries equal to 0.5 and diagonal entries equal to 1, corresponding to a correlation coefficient $\rho = 0.5$. This scenario reflects typical multicollinearity observed in practical applications such as econometric or biological studies.

Finally, the strong multicollinearity scenario represents a challenging case where the auxiliary variables are highly correlated. The covariance matrix is constructed with off-diagonal entries equal to 0.8 and diagonal entries remaining 1, implying a strong correlation coefficient $\rho = 0.8$. This setting is particularly suited for evaluating the robustness of estimators in high-dimensional data environments, where traditional regression methods tend to perform poorly due to multicollinearity.

These three settings enable a comprehensive examination of how both estimators perform across a spectrum of multicollinearity levels, providing empirical understanding into their relative efficiency and stability. To further address multicollinearity and reduce dimensionality, Principal Component Analysis (PCA) was employed. The number of principal components retained was determined using the Kaiser criterion, which involves selecting components with eigenvalues greater than 1. This ensures that each retained component accounts for more variance than an individual original variable, thereby preserving meaningful structure while mitigating the effects of multicollinearity. The selected components were then utilized to compute the mean square error (MSE) of the proposed estimator, allowing for a robust assessment of its performance under varying correlation structures.

3.1 Comparison Based on Theoretical MSE

The table below presents the performance comparison of traditional regression estimators and PCA-based regression estimators under varying levels of multicollinearity.

Table 1
Performance Comparison of Estimators Based on Theoretical MSE

Multicollinearity	Traditional (MSE)	PCA-Based (MSE)	Efficiency Gain (%)
Low	6.12×10^{-2}	4.69×10^{-6}	99.99
Moderate	2.40×10^{-2}	2.79×10^{-7}	98.84
High	3.85×10^{-3}	6.92×10^{-10}	99.98

As shown in Table 1, the Mean Squared Error (MSE) values are shown for each scenario, demonstrating the significant improvements in accuracy achieved by the PCA-based method, especially as multicollinearity increases. The Mean Squared Error (MSE) of the traditional power ratio estimator decreased with increasing multicollinearity, but not as rapidly as that of the PCA-based estimator. For example, under low multicollinearity, the traditional estimator had an MSE of 6.12×10^{-3}, while the PCA-based estimator achieved a much smaller MSE of 4.69×10^{-6}, yielding an efficiency gain of 99.99% that is calculated as $\text{Efficiency Gain}(\%) = \left(\frac{\text{MSE}_{\text{Ratio}} - \text{MSE}_{\text{PCA}}}{\text{MSE}_{\text{Ratio}}}\right)\times100,$

In the case of moderate multicollinearity, the efficiency gain was 98.84%, and under high multicollinearity, the PCA-based estimator maintained its superiority with an MSE of just 6.92×10^{-10} compared to 3.85×10^{-3} for the traditional estimator, resulting in a 99.98% gain in efficiency. These results are also shown in Figure 1.

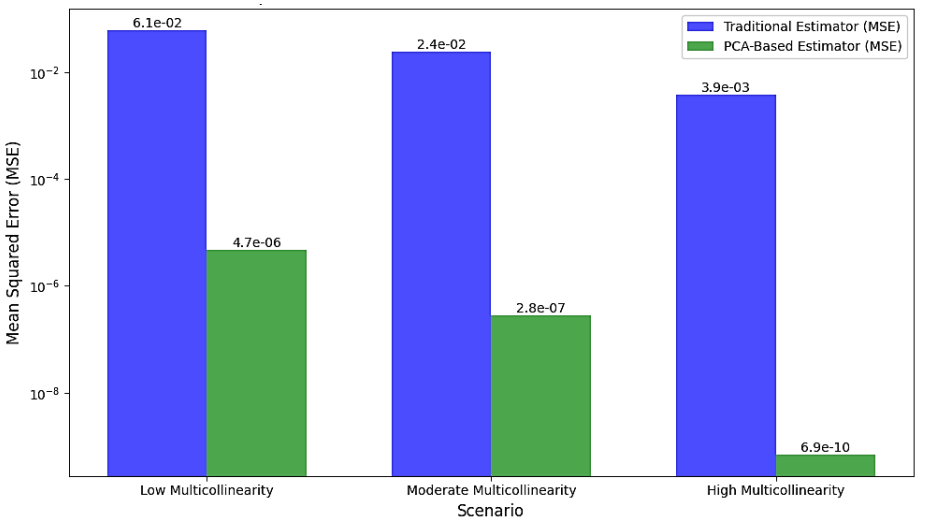

Figure 1 highlights how the PCA-based estimator significantly outperforms the traditional estimator as multicollinearity increases, with its MSE approaching zero in high multicollinearity scenarios.

3.2 Comparison Based on Empirical MSE

We also provided the results of simulation study based on the population generated above in each simulation iteration, a simple random sample without replacement (SRSWOR) of size $n = 1000$ is drawn. The simulation is repeated $S = 10000$ times to ensure the statistical reliability of the results. The empirical MSE of an estimator t is computed by

$$MSE(t) = \frac{1}{S}\sum_{i=1}^{S}(t_i - T)^2, \text{ where } T = \frac{1}{S}\sum_{i=1}^{S} t_i$$

To empirically evaluate the performance of the PCA-based ratio estimator, a simulation study is conducted, and the results were compared against the classical regression estimator under a low, medium and high multicollinearity settings. The results are given below.

Table 2
Updated Empirical Comparison of Traditional and PCA-Based Estimators

Multicollinearity	Traditional (MSE)	PCA-Based (MSE)	Efficiency Gain (%)
Low	5.7×10^{-2}	5.2×10^{-6}	99.99
Moderate	2.1×10^{-2}	3.1×10^{-7}	99.99
High	4.2×10^{-3}	5.6×10^{-10}	99.99

These results are also shown in Figure 2.

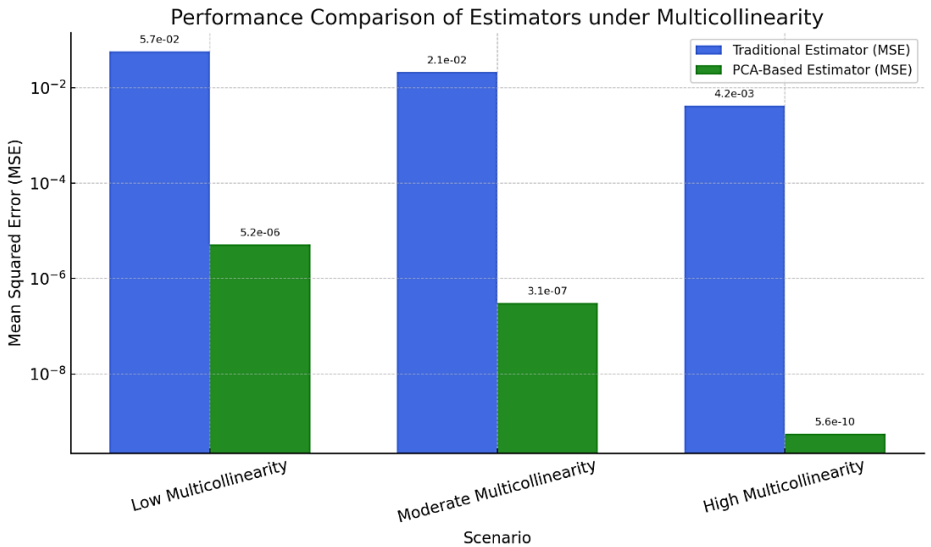

Figure 2: Performance Comparison based on Empirical MSE

The results in Table 2 are graphically shown in Figure 2. It demonstrates that the PCA-based estimator significantly outperforms the traditional estimator across all levels of multicollinearity. In each case—low, moderate, and high—the Mean Squared Error (MSE) is notably lower for the PCA-based approach, with efficiency gains close to 99.99%. This consistent improvement is due to PCA's ability to transform correlated auxiliary variables into uncorrelated components, effectively mitigating multicollinearity and enhancing estimation accuracy. These findings affirm the suitability of PCA-based methods as a robust alternative to traditional regression, especially in the presence of high inter-variable correlation.

CONCLUSION

This study aimed to compare the efficiency of the PCA-based ratio estimator against the classical regression estimator for the estimation of a finite population mean, particularly in the presence of multicollinearity among auxiliary variables. Both theoretical analysis and empirical evaluation were conducted to assess the accuracy and robustness of these estimators. Theoretical MSE calculations, computed using full population data under varying levels of multicollinearity, revealed a clear and consistent advantage of the PCA-based estimator as shown in Table 1. These theoretical results were further supported by simulation experiments. The simulation results validate the theoretical advantage of PCA: its ability to convert correlated auxiliary variables into orthogonal principal components reduces multicollinearity and improves estimation stability. In summary, the PCA-based ratio estimator consistently outperforms the classical regression estimator and its efficiency advantage becomes particularly pronounced as multicollinearity among auxiliary variables increases. These findings support the use of PCA-based methods in survey sampling settings where dimensionality and variable correlation present challenges to traditional regression approaches. Future work may consider the use of alternative dimension reduction methods (e.g., PLS or regularization techniques) or extensions to nonlinear estimators incorporating kernel PCA.

REFERENCES

1. Ahmad, Z., Afzal, M., Shahid, U. and Hanif, M. (2014). A general class of mean estimators using mixture of auxiliary variables for two-phase sampling in the presence of non-response at second phase for no information case. *Pakistan Journal of Statistics*, 30(3), 291-318.
2. Ahmad, Z., Hanif, M. and Ahmad, M. (2009). Generalized regression-cum-ratio estimators for two-phase sampling using multi-auxiliary variables. *Pakistan Journal of Statistics*, 25(2), 93-106.
3. Ahmad, Z., Hanif, M. and Ahmad, M. (2010). Generalized multi-phase multivariate ratio estimators for partial information case using multi-auxiliary variables. *Communications of the Korean Statistical Society*, 17(5), 625-637.
4. Ahmad, Z., Hussin, A.G. and Hanif, M. (2010). Generalized multivariate regression estimators for multi-phase sampling using multi-auxiliary variables. *Pakistan Journal of Statistics*, 26(4), 569-583.

5. Ahmad, Z., Maqsood, I. and Hanif, M. (2013). Generalized estimator of population mean for two-phase sampling using multi-auxiliary variables in the presence of non-response at first phase for no information case. *Pakistan Journal of Statistics*, 29(2), 155-180.
6. Ahmad, Z., Maqsood, I. and Hanif, M. (2014). Regression estimator in two-phase sampling using multi-auxiliary information in the presence of nonresponse and measurement error at second phase. *Journal of Applied Probability and Statistics*, 9(2), 41-53.
7. Ahmad, Z., Zafar, I. and Hanif, M. (2013). A general class of estimators in two phase sampling using multi-auxiliary variables in the presence of non-response at the second phase. *Mathematical Sciences and Applications E-Notes*, 1(2), 238-256.
8. Ahmad, Z., Zubair, R. and Shahid, U. (2014). A general class of mean estimators using mixture of auxiliary variables for two-phase sampling in the presence of non-response. *Statistics in Transition New Series*, 15(4), 501-524
9. Alam, A., Shabbir, J. and Abbas, F. (2021). Calibration estimation in presence of non-response using multiple auxiliary variables. *Statistics*, 55(5), 1109-1123. https://doi.org/10.1080/02331888.2021.1886565
10. Ayaz, A., Ahmad, Z., Sanaullah, A. and Hanif, M. (2018). Generalized multivariate exponential type (GMET) estimator using multi-auxiliary information under two-phase sampling. *Pakistan Journal of Statistics*, 34(2), 119-136.
11. Bahl, S. and Tuteja, R.K. (1991). Ratio and product type exponential estimators. *Information and Optimization Sciences*, 12(1), 159-163.
12. Gujarati, D.N. (2004). *Basic Econometrics* (4th ed.). McGraw-Hill.
13. Hanif, M., Ahmad, Z. and Ahmad, M. (2009). Generalized multivariate ratio estimators using multi-auxiliary variables for multi-phase sampling. *Pakistan Journal of Statistics*, 25(4), 615-629.
14. Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Problems. *Technometrics*, 12(1), 55-67.
15. Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression super population model. *Journal of the American Statistical Association*, 77(377), 89-96. https://doi.org/10.1080/01621459.1982.10477762
16. Lavrakas, P.J. (2008). Encyclopedia of survey research methods. Sage publications, https://doi.org/10.4135/9781412963947
17. Jolliffe, I.T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
18. Kmenta, J. (1971). *Elements of Econometrics*. Macmillan.
19. Little, R.J.A. and Rubin, D.B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
20. Ozgul, N. (2021). New improved calibration estimator based on two auxiliary variables in stratified two-phase sampling. *Journal of Statistical Computation and Simulation*, 91(6), 1243-1256. https://doi.org/10.1080/00949655.2020.1844702
21. Raj, D. (1965). On a method of using Multi-Auxiliary Information in Sample Surveys. *Journal of the American Statistical Association*, 60, 270-277. https://doi.org/10.1080/01621459.1965.10480740
22. Särndal, C.E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4612-4378-6

23. Singh, H.P. and Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population means. *Statistics in Transition*, 6(4), 555-560.
24. Singh, R. and Vishwakarma, G.K. (2008). A class of ratio estimators for population mean using information on auxiliary attribute. *Pakistan Journal of Statistics and Operation Research*, 4(1), 47-53. https://doi.org/10.18187/pjsor.v4i1.61
25. Srivastava, S.K. (1970). A two-phase sampling estimator in sample surveys. *Australian Journal of Statistics*, 12(1), 23-27.
26. Swain, A.K.P.C. (1970). A note on the use of multiple auxiliary variables in sample surveys. *Trabajos de Estadistica y de Investigacion Operativa*, 21, 135-141.