Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 6 (2024) 2091

Journal of the Nigerian Society of Physical Sciences

Enhanced methods for multicollinearity mitigation in stochastic frontier analysis estimation

Rauf I. Rauf[a,*], Ayinde Kayode[b], Bello A. Hamidu[a], Bodunwa O. Kikelomo[a], Alabi O. Olusegun[a]

[a]*Department of Statistics, Federal University of Technology, Akure-Nigeria*
[b]*Department of Mathematics and Statistics, Northwest Missouri State Unversity, MO, USA.*

Abstract

Efficiency estimation in production technology has been a concern in economics, with methodologies such as Stochastic Frontier Analysis (SFA) playing a key role in this area. SFA has been pivotal in evaluating the efficiency of entities by isolating technical inefficiency from random production errors. However, despite its significance, the application of SFA faces challenges when the multicollinearity assumption underlying the model is violated. Therefore, this study presents a novel estimator, termed "Principal Component Analysis Estimation for Stochastic Frontier Analysis" (PCA-SFA), to address the problem of multicollinearity in the classical SFA model. The PCA-SFA estimator integrates Principal Component Analysis (PCA) to correct assumption violation arising from multicollinearity. Monte Carlo simulation study was conducted to ascertain the PCA-SFA performance, involving no fewer than 2,000 replications based on the Cobb-Douglas production function with varying levels of multicollinearity, represented by correlation coefficients (ρ) ranging from 0.8, 0.9, 0.95, 0.99, and 0.999, and sample sizes (n) of 20, 50, 100, 250, and 1,000. The proposed estimator's performance was compared to the classical SFA model using the Mean Square Error (MSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) as evaluation metrics. The results demonstrate that the PCA-SFA estimator consistently outperforms the classical SFA model. The PCA-SFA model showed significantly lower MSE, AIC, and BIC values, indicating improved precision and reliability in parameter estimation. The study, therefore, recommends that researchers and practitioners in econometrics and related fields consider integrating PCA-SFA into their production efficiency analytical frameworks, particularly when dealing with datasets prone to multicollinearity issues.

DOI:10.46481/jnsps.2024.2091

Keywords: Stochastic frontier analysis, Multicollinearity, Assumption violations, Correction methodologies

Article History :
Received: 21 April 2024
Received in revised form: 21 August 2024
Accepted for publication: 11 September 2024
Published: 27 September 2024

Communicated by: P. Thakur

1. Introduction

Efficiency analysis in production units has long been a focal point in economics, driven especially by methodologies like stochastic frontier analysis (SFA). Based on the foundational works by Aigner, Lovell, and Schmidt [1] and Meeusen and Broeck [2], SFA has been pivotal in evaluating the efficiency of the entity by isolating technical inefficiency from random production errors. However, despite its significance, the application of SFA faces challenges when the assumptions of the underlying model are violated. This study delves into the intricate terrain of the stochastic frontier model, particularly focusing on critical issues like multicollinearity when assumptions are vio-

*Corresponding author: Tel.: +234-813-345-1684.
Email address: rauf.ibrahim@outlook.com (Rauf I. Rauf)

lated. The primary objective of this research is to identify and propose corrective measures and estimators that specifically address multicollinearity within the stochastic frontier analysis model. As highlighted by Wang [3], the model often struggles to capture non-monotonic efficiency effects, necessitating flexible parameterizations. Expanding on the work of Hadri, Guermat, Whittaker [4] and [5], this study extends its scope to include multicollinearity problem and proposes for model estimation while assuming non-violation of collinearity assumptions in the model. Therefore, the objective is to comprehensively evaluate and compare the performance of the proposed measures/estimators with existing models in the literature. The literature on stochastic frontier analysis and related methodologies showcases a diverse range of contributions that have significantly shaped efficiency estimation and production modeling. Wang [3] introduced a model emphasizing flexible parameterizations to account for exogenous influences on inefficiency, highlighting the need to accommodate nonmonotonic efficiency effects. This insight underscores the complexities involved in accurately modeling efficiency. Hadri [6] worked to address heteroscedastic inefficiency, recognizing that multicollinearity assumptions may not hold for the model. Christopoulos, Lolos, and Tsionas [7] explored the cost efficiency of the Greek banking system, employing a stochastic frontier model and uncovering intriguing relationships between bank size, economic performance, and cost efficiency. This empirical application sheds light on the real-world implications of assumption violations in efficiency modeling. Kumbhakar, Denny, and Fuss [8] contributed a stochastic frontier model with random coefficients, acknowledging technological diversity among firms.

This contribution opens avenues for understanding the inherent differences in technological possibilities between firms, challenging the assumption of identical technological capabilities. Karakaplan and Kutlu [9] proposed a maximum likelihood-based framework to address endogeneity in stochastic frontier models, demonstrating superior performance through Monte Carlo experiments. This approach underscores the importance of considering endogeneity in frontier models for robust estimations, aligning with the broader theme of addressing assumptions in SFA.

Furthermore, Obadina *et al.* [10] conducted a comparative study of multiple linear regression estimators under multicollinearity, using the ordinary least squares (OLS), modified ridge regression (MRR), and generalized Liu-Kejian methods (LKM). Through Monte Carlo simulations, the study demonstrated that MRR and LKM significantly outperformed the OLS in terms of reducing the average mean square error (AMSE), particularly in high multicollinearity and larger sample sizes.

Similarly, Shewa and Ugwuowo [11] explored the Bell Regression Model (BRM) as an alternative to the Poisson regression model for count data with over-dispersion. The authors developed a new estimator that successfully addressed multicollinearity issues within the BRM, as evidenced by both theoretical and simulation results. Their findings emphasize the need for specialized estimators in models where traditional methods like maximum likelihood estimation (MLE) falter due to multicollinearity, a concern also central to the PCA-SFA approach. Further, Jegede *et al.* [12] introduced the Robust Jackknife Kibria-Lukman (RJKL) M-Estimator, designed to combat the dual challenges of multicollinearity and outliers in linear regression models. Through theoretical analysis and Monte Carlo simulations, the study validated the RJKL estimator's superior performance over existing estimators, thereby reinforcing the significance of robust methods in improving estimator efficiency, akin to the objectives of the PCA-SFA estimator. In the context of heteroscedasticity, Rauf *et al.* [13] examined various correction measures within the SFA model, comparing their effectiveness through extensive Monte Carlo simulations. Their research concluded that the HCRTE (heteroscedasticity correction for both random error and technical efficiency error) consistently provided the most accurate parameter estimates under heteroscedasticity.

Despite these studies, a notable gap persists in the literature. Specifically, there is a lack of comprehensive corrected measures or estimators tailored to handle multicollinearity within the Stochastic Frontier Analysis Model. Furthermore, the application of the principal component analysis correction measure for the impact of multicollinearity in the SFA model remains underexplored, necessitating a holistic investigation that addresses the issue of multicollinearity in the SFA model. This study aims to bridge these gaps by proposing novel methodologies and estimators for efficient model estimation in the presence of multicollinearity in data sets to be fitted with SFA.

1.1. Aim and objectives

Aim: The primary aim of this study is to develop and evaluate a new estimator, termed "Principal Component Analysis Estimation for Stochastic Frontier Analysis" (PCA-SFA), that addresses the issue of multicollinearity in the classical stochastic frontier analysis (SFA) model.

Objectives:

1. To introduce the PCA-SFA estimator, combining principal component analysis (PCA) with the traditional SFA model to mitigate the effects of multicollinearity.
2. To formulate a Monte Carlo simulation experiment to test the performance of the PCA-SFA estimator across varying levels of multicollinearity and sample sizes.
3. To compare the effectiveness of the PCA-SFA estimator with the classical SFA model by analyzing their mean square error (MSE), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values under different conditions.
4. To suggest potential extensions and applications of the PCA-SFA estimator for future research and practice in econometrics.

1.2. Scope of the study

This study focuses on the development and evaluation of the PCA-SFA estimator, specifically designed to address multicollinearity in stochastic frontier analysis. The research is confined to the theoretical formulation and empirical testing of this estimator using Monte Carlo simulations. The simulation experiments cover a range of sample sizes (from 20 to 1000)

and varying levels of multicollinearity (with correlation coefficients ranging from 0.8 to 0.999). The study utilizes the Cobb-Douglas production function as the underlying model for generating data, and the results are compared based on the mean square error (MSE), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) criterion.

2. Materials and methods

In this section, the focus is on the methodological framework supporting the empirical exploration of the stochastic frontier analysis (SFA) model. The proposed estimator is introduced to address multicollinearity challenges within the SFA framework. Also, we present details of the simulation procedure designed to rigorously validate the robustness and efficacy of the estimators through meticulous testing on simulated datasets by Monte Carlo simulation study. The goal is to provide empirical evidence supporting the reliability and applicability of these estimators in enhancing the precision of stochastic frontier analysis under challenging empirical conditions.

2.1. Stochastic frontier model (SFA) estimation and properties

Following Aigner *et al.* [1], Meeusen and Broeck [2], Kumbhakar *et al.* [8], Coelli *et al.* [14], considering a cross-sectional data on quantities of N inputs $x_{ni}, n = I, \ldots, N; i = 1, \ldots, I$ are used to produce a single output $y_i, i = 1, \ldots, I$ are available to each of I producers.

The stochastic production frontier for the producers can be written as:

$$y_i = f(x_i;\beta) \cdot \exp(V_i) \cdot TE_i, \tag{1}$$

where the β s are the parameters in the production function. V_i reflects random noise, and TE_i is the output-oriented technical efficiency of producer i. From Eq. (1) we have:

$$TE_i = \frac{y_i}{f(x_i;\beta) \cdot \exp(V_i)}. \tag{2}$$

Assuming the $f(x_i;\beta)$ takes a Cobb-Douglas form, the Eq.(2) becomes:

$$TE_i = \exp\{-U_i\}. \tag{3}$$

Thus, the stochastic production frontier becomes:

$$\ln y_i = \beta_0 + \sum_{n=1}^{N} \beta_n \ln x_{ni} + V_i - U_i. \tag{4}$$

Then the estimate of the technical efficiency can be obtained from: Eq.(2) and (3) below as posited by Battese and Coelli [15]:

$$\widehat{TE_{1i}} = \exp\left\{-E\widehat{(U_i \mid E_i)}\right\}, \tag{5}$$

$$\widehat{TE_{2i}} = E\left(\exp\widehat{\{-U_i\}} \mid E_i\right). \tag{6}$$

The joint density of U and V is then given as follows:

$$f(u,v) = \frac{1}{\sqrt{2\pi}\sigma\theta} \exp\left\{-\frac{v^2}{2\sigma^2}\right\}. \tag{7}$$

Since E + U, the joint density of U and E after the variable transformation is:

$$f_{U,E}(u,\epsilon) = \frac{1}{\sqrt{2\pi}\sigma\theta} \exp\left\{-\frac{(\epsilon+u)^2}{2\sigma^2}\right\}. \tag{8}$$

Hence, the marginal density of E can be derived by:

$$f_E(\epsilon) = \int_0^{\theta} \frac{1}{\sqrt{2\pi}\sigma\theta} \exp\left\{-\frac{(\epsilon+u)^2}{2\sigma^2}\right\} \mathrm{d}u \tag{9}$$

$$= \int_{\frac{\epsilon}{\sigma}}^{\frac{\theta+\epsilon}{\sigma}} \frac{1}{\sqrt{2\pi}\theta} \exp\left\{-\frac{z^2}{2}\right\} \mathrm{d}z \tag{10}$$

$$= \frac{1}{\theta}\left[\Phi\left(\frac{\theta+\epsilon}{\sigma}\right) - \Phi\left(\frac{\epsilon}{\sigma}\right)\right], \epsilon \in \Re. \tag{11}$$

Noting that $F_E(\varepsilon)$ is a symmetric density with a mean of:

$$E(\epsilon) = -E(u) = -\frac{\theta}{2}, \tag{12}$$

and variance:

$$\mathrm{Var}(\epsilon) = \mathrm{Var}(v) + \mathrm{Var}(u) = \sigma^2 + \frac{\theta^2}{12}. \tag{13}$$

The $F_E(\varepsilon)$ can be achieved by computing the skewness of the coefficient:

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \tag{14}$$

$$= \frac{E[\epsilon - E[\epsilon]]^3}{\mathrm{Var}(\epsilon)^{3/2}} \tag{15}$$

$$= \frac{E[v - (u - E[u])]^3}{\mathrm{Var}(\epsilon)^{3/2}} \tag{16}$$

$$= \frac{E[-(u - E[u])]^3}{\mathrm{Var}(\epsilon)^{3/2}} \tag{17}$$

$$= 0. \tag{18}$$

The density of E is symmetric around its mean $-0/2$. As presented in Greene [16], Caudill *et al.* [17], we can compute the kurtosis coefficient of $F_E(\varepsilon)$ as follows:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} \tag{19}$$

$$= \frac{E[\epsilon - E[\epsilon]]^4}{\mathrm{Var}(\epsilon)^2} \tag{20}$$

$$= \frac{E[v - (u - E[u])]^4}{\mathrm{Var}(\epsilon)^2} \tag{21}$$

$$= \frac{3\sigma^4 + \frac{\theta^2\sigma^2}{2} + \frac{\theta^4}{80}}{\sigma^4 + \frac{\theta^2\sigma^2}{6} + \frac{\theta^4}{144}} \tag{22}$$

$$= 3 - \frac{\frac{\theta^4}{120}}{\sigma^4 + \frac{\theta^2\sigma^2}{6} + \frac{\theta^4}{144}} \tag{23}$$

$$\leq 3 \text{ for all } \theta \text{ and } \sigma. \tag{24}$$

From the density of Eq. (8), the log-likelihood function is then given by:

$$\ln L = -I \ln\theta + \sum_{i=1}^{I} \ln\left[\Phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \Phi\left(\frac{\epsilon_i}{\sigma}\right)\right], \tag{25}$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{I}{\theta} + \sum_{i=1}^{I} \frac{\frac{1}{\sigma}\phi\left(\frac{\theta+\epsilon_i}{\sigma}\right)}{\Phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \Phi\left(\frac{\epsilon_i}{\sigma}\right)}, \tag{26}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{1}{2\sigma^3} \sum_{i=1}^{I} \frac{-(\theta+\epsilon_i)\,\phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) + \epsilon_i \phi\left(\frac{\epsilon_i}{\sigma}\right)}{\Phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \Phi\left(\frac{\epsilon_i}{\sigma}\right)}. \tag{27}$$

Also, from Greene [16], Aigner and Cain [18], Stevenson [19], the Cobb-Douglas production function is given by:

$$\frac{\partial \ln L}{\partial \beta_0} = \frac{\partial \ln L}{\partial \epsilon_i} * \frac{\partial \epsilon_i}{\partial \beta_0} \tag{28}$$

$$= -\frac{1}{\sigma} \sum_{i=1}^{I} \frac{\phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \phi\left(\frac{\epsilon_i}{\sigma}\right)}{\Phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \Phi\left(\frac{\epsilon_i}{\sigma}\right)}, \tag{29}$$

$$\frac{\partial \ln L}{\partial \beta_n} = -\frac{1}{\sigma} \sum_{i=1}^{I} \ln x_{ni} \cdot \frac{\phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \phi\left(\frac{\epsilon_i}{\sigma}\right)}{\Phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \Phi\left(\frac{\epsilon_i}{\sigma}\right)}. \tag{30}$$

We derive the conditional distribution of Ui/Ei

$$f(u_i \mid \epsilon_i) = \frac{f(u_i, \epsilon_i)}{f(\epsilon_i)}$$

$$= \frac{1}{\sqrt{2\pi}\theta} \cdot \frac{1}{\Phi\left(\frac{\theta+\epsilon_i}{\sigma}\right) - \Phi\left(\frac{\epsilon_i}{\sigma}\right)} \exp\left\{-\frac{(\epsilon_i + u_i)^2}{2\sigma^2}\right\}. \tag{31}$$

The conditional distribution of U_i/E_i is truncated. The normal distribution is revealed in the following lemma. Having truncated a_1 and a_2, where $-\infty < a_1 < a_2 < \infty$, is then given by:

$$f(y) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)}, a_1 \le y \le a_2 \tag{32}$$

$$M_Y(t) = E\left[e^{tY} \mid Y \in [a_1, a_2]\right], \tag{33}$$

$$= e^{\mu t + \sigma^2 t^2/2} \frac{\Phi\left(\frac{a_2-\mu}{\sigma} - \sigma t\right) - \Phi\left(\frac{a_1-\mu}{\sigma} - \sigma t\right)}{\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)}, \tag{34}$$

$$E[Y \mid Y \in [a_1, a_2]] = \mu - \sigma \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \tag{35}$$

$$M(Y \mid Y \in [a_1, a_2]) = \begin{cases} a_2 & a_1 \le a_2 \le \mu \\ \mu & a_1 \le \mu \le a_2 \\ a_1 & \mu \le a_1 \le a_2 \end{cases}, \tag{36}$$

where $\alpha_k = \frac{a_k-\mu}{\sigma}$.

Proof. Eq. (32) the probability of Y falling in the interval [a a_1, a_2] is $\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)$. Thus the conditional density of Y is [20]:

$$f(y \mid Y \in [a_1, a_2]) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)}.$$

$$M(t) = E\left[e^{tY} \mid Y \in [a_1, a_2]\right] \tag{37}$$

$$= \frac{\int_{a_1}^{a_2} e^{ty} f(y)\mathrm{d}y}{\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)}. \tag{38}$$

We have the following:

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{a_1}^{a_2} e^{ty} e^{-(y-\mu)^2/2\sigma^2}\,\mathrm{d}y \tag{39}$$

$$= e^{-\frac{1}{2\sigma^2}\left[\mu^2 - \left(\sigma^2 t + \mu\right)^2\right]} \frac{1}{\sigma\sqrt{2\pi}} \int_{a_1}^{a_2} e^{-\frac{\left(y-\sigma^2 t-\mu\right)^2}{2\sigma^2}}\,\mathrm{d}y \tag{40}$$

$$= e^{\mu t + \sigma^2 t^2/2} \int_{a_1}^{a_2} \frac{1}{\sigma}\phi\left(\frac{y-\sigma^2 t-\mu}{\sigma}\right)\mathrm{d}y \tag{41}$$

$$= e^{\mu t + \sigma^2 t^2/2}\left[\Phi\left(\frac{a_2-\sigma^2 t-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\sigma^2 t-\mu}{\sigma}\right)\right]. \tag{42}$$

Then the moment-generating function is given by:

$$M(t) = e^{\mu t + \sigma^2 t^2/2} \frac{\Phi\left(\frac{a_2-\mu}{\sigma} - \sigma t\right) - \Phi\left(\frac{a_1-\mu}{\sigma} - \sigma t\right)}{\Phi\left(\frac{a_2-\mu}{\sigma}\right) - \Phi\left(\frac{a_1-\mu}{\sigma}\right)}. \tag{43}$$

Eq. (33) - (35) from the moment generating function, the expected value is then derived from:

$$E[Y \mid Y \in [a_1, a_2]] = M'(t)\big|_{t=0} \tag{44}$$

$$= \mu - \sigma \frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}, \tag{45}$$

and the variance:

$$\mathrm{Var}[Y \mid Y \in [a_1, a_2]] = M''(t)|_{t=0} \tag{46}$$

$$= \sigma^2\left\{1 - \frac{\alpha_2\phi(\alpha_2) - \alpha_1\phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)} - \left[\frac{\phi(\alpha_2) - \phi(\alpha_1)}{\Phi(\alpha_2) - \Phi(\alpha_1)}\right]^2\right\}, \tag{47}$$

where $\alpha_k = \frac{a_k-\mu}{\sigma}$. The formula for the mode of the distribution easily follows the conditional density [16–18].

2.2. Principal component analysis approach to regression

In the context of an ordinary least squares (OLS) regression model with multicollinearity issues, as discussed by Coxe [21] and Dunteman [24], we begin with the standard OLS regression equation:

$$y = \beta_{01} + X\beta + u, \tag{48}$$

where, y represents the observed output vector (n-dimensional), X is the design matrix of input variables (nxk), β is the coefficient vector (k-dimensional), u is the random error vector, and 1 is an n-dimensional vector of ones.

We compute the sample covariance matrix of the input variables X:

$$S^x = \frac{1}{n} X^\top X. \tag{49}$$

Applying spectral decomposition to S^x yields eigenvalues and eigenvectors:

$$X^\top X = P\Lambda P^\top, \tag{50}$$

where, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is the diagonal matrix of eigenvalues, and $P = (p_1, p_2, \ldots, p_k)$ is the matrix of corresponding orthogonal eigenvectors [25].

The principal components Z are obtained by multiplying X with the eigenvectors:

$$Z = XP = (z_1, z_2, \ldots, z_k), \tag{51}$$

$$z_j = Xp_j. \tag{52}$$

We then select the number of principal components to retain based on the magnitudes of the eigenvalues.

Next, we reparameterize the OLS model using principal components:

$$y = \beta_{01} + Z\theta + u, \tag{53}$$

where, Z represents the matrix of retained principal components, and $\theta = P^\top\beta$ is the transformed coefficient vector [**?**].

To address collinearity, we choose the first $r < k$ principal components highly correlated with y and partition Z into Z_1 and Z_2:

$$Z = (Z_1, Z_2) = (XP_1, XP_2). \tag{54}$$

By assuming $Z_2 \approx 0$ for simplification, the re-parameterized model becomes:

$$y = \beta_{01} + Z_1\theta_1 + v - u, \tag{55}$$

where, $\theta_1 = P_1^\top\beta_1$ includes coefficients associated with significant principal components.

We estimate the parameters using the re-parameterized model:

$$\hat{\theta}_1 = \left(Z_1^\top Z_1\right)^{-1} Z_1^\top y, \tag{56}$$

$$\hat{\beta} = P_1\hat{\theta}_1. \tag{57}$$

Finally, we compute the covariance matrix of the estimated coefficients as suggested by Jolliffe [27], Johnson and Wichern [28], Meredith and Millsap [29] and Rao [30] as follows:

$$\text{Cov}(\hat{\beta}) = P_1 \text{Cov}(\hat{\theta})P_1^\top. \tag{58}$$

2.3. The principal component solution to multicollinearity in SFA

Considering a scenario where multicollinearity assumption violations occur in the stochastic frontier analysis (SFA) model, a modified OLS model, we incorporate principal component analysis (PCA) into OLS regression as discussed in the previous Section. The matrix representation of the stochastic frontier production model is given by Eq. (1):

$$y = \beta_{01} + X\beta + v - u, \tag{59}$$

where $-y, v, u$, and 1 are n-dimensional vectors of n dimensional observed outputs, random errors of production and inefficiency, and ones, respectively. - X is the nxk design matrix of the inputs. - β is the corresponding k-dimensional vector of coefficients. - All inputs are assumed to be standardized.

Applying spectral decomposition to the symmetric matrix $X^\top X$ yields:

$$X^\top X = P\Lambda P^\top, \tag{60}$$

where: $-\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is the diagonal eigenvalues matrix. - $P = (p_1, p_2, \ldots, p_k)$ is the corresponding orthogonal eigenvector matrix.

The re-parameterized SFA model using principal components is given by:

$$y = \beta_{01} + Z\theta + v - u, \tag{61}$$

where, $Z = XP = (z_1, z_2, \ldots, z_k)$ is the matrix of principal components, and $\theta = P^\top\beta$.

We partition Z into Z_1 and Z_2, where Z_1 contains components with significant eigenvalues and Z_2 contains negligible components. Assuming $Z_2 \approx 0$ for simplification, the re-parameterized SFA model becomes:

$$y = \beta_{01} + Z_1\theta_1 + v - u, \tag{62}$$

where $\theta = \left(\theta_1^\top, \theta_2^\top\right)^\top, \theta_1 = P_1^\top\beta_1$, and $\theta_2 = P_2^\top\beta_2$. The constraint $Z_2 \approx 0$ is equivalent to $\theta_2 \approx 0$.

The SFA estimator as a method of least squares (MOLS) of θ_1 is given by:

$$\hat{\theta}_1 = \left(Z_1^\top Z_1\right)^{-1} Z_1^\top y. \tag{63}$$

Finally, the principal component estimator of β is:

$$\hat{\beta} = P_1\hat{\theta}_1. \tag{64}$$

With the covariance matrix of $\hat{\beta}$ calculated as:

$$\text{Cov}(\hat{\beta}) = P_1 \text{Cov}(\hat{\theta})P_1^\top. \tag{65}$$

2.4. Proposed estimator to address multicollinearity in the classical stochastic frontier analysis (SFA) model

From the principles outlined in the above, this study introduces a new estimator, termed "Principal Component Analysis Estimation for Stochastic Frontier Analysis" (PCA-SFA). This estimator leverages principal component analysis (PCA) to rectify violations of assumptions in the classical stochastic frontier analysis (SFA) model.

The proposed estimator ("PCA-SFA") is a mathematical combination of equations Eq. (63) and (64), given by:

$$\hat{\beta} = P_1\left(\left(Z_1^\top Z_1\right)^{-1} Z_1^\top y\right), \tag{66}$$

where - Z_1 is the nx matrix containing principal components associated with non-zero eigenvalues. - P_1 is the corresponding orthogonal eigenvector matrix in the PCA estimators without collinear variables.

3. Monte Carlo simulation study

The procedures to generate the input variables and error terms is conducted using the Monte Carlo simulation technique.

Table 1: Comparison of MSE between SFA and PCA-SFA for different sample sizes and multicollinearity levels.

Sample Size (n)	Multicollinearity (ρ)	MSE_SFA	MSE_PCA_SFA
20	0.8	0.0799572	0.3198883
	0.9	0.945719767	0.3513819
	0.95	1.058591	0.25155777
	0.99	1.204909433	0.29942287
	0.999	0.4724872	0.3316227
50	0.8	0.9559107	0.23743835
	0.9	0.369043	0.24205851
	0.95	0.198039567	0.27531365
	0.99	0.343385533	0.23598764
	0.999	0.8569634	0.25339741
100	0.8	0.186406367	0.25832282
	0.9	0.220669033	0.21779611
	0.95	0.195958267	0.23261347
	0.99	1.1087664	0.28035442
	0.999	0.986483533	0.24524453
250	0.8	0.2270343	0.2058508
	0.9	1.0082384	0.22229198
	0.95	0.122285133	0.23206008
	0.99	1.034549567	0.21420244
	0.999	0.9809092	0.203331998
1000	0.8	0.823431667	0.200252164
	0.9	0.206892367	0.21108376
	0.95	0.824043667	0.208946244
	0.99	1.022383567	0.200680016
	0.999	1.0331049	0.209547976

3.1. Model formulation

To evaluate the performance of the proposed estimator ('PCA-SFA'), we perform a Monte Carlo simulation experiment of not fewer than 2,000 replications on the stochastic frontier model following the Cobb-Douglass production function in (1), where:

$$y = \beta_0 1 + X\beta + vu, \tag{67}$$

$$\ln(y) = \beta_1 + \beta_2 \ln(x_1) + \ldots + \beta_k \ln(x_k) + v - u, \tag{68}$$

y, is the observed output (dependent variable), v and u are the random errors and the technical inefficiency component, respectively, x_{1-k}, is the of production inputs (independent variables); β_{1-k}, is the corresponding k$^{\text{th}}$ coefficients.

Setting: sample size (n) to initial $(20, 50, 100, 250, 1000)$, $\beta_1 = 0.7; \beta_2 = 0.8; \beta_3 = 0.9; \beta_4 = 1.0; \beta_5 = 1.1; \beta_6 = 1.2;$ k = 6.

3.2. Procedure for generating the input variables with varying levels of collinearity.

The simulation procedure used by Gibbons [31], Wichern and Churchill [32], McDonald and Galarneau [33], Kibria [34], Lukman and Ayinde [35], Fayose and Ayinde [36] is also used to generate the exposure variables in this study. This is given as follows:

$$X_{ti} = \left(1 - \rho^2\right)^{\frac{1}{2}} Z_{ti} + \rho Z_{tp}, \tag{69}$$

$$t = 1, 2, 3 \ldots, n, \tag{70}$$

$$i = 1, 2 \ldots p, \tag{71}$$

where

- Z_{ti} is independent standard normal distribution with mean zero and unit variance;
- ρ is the correlation between any two exposure variables and p is the number of exposure variables. The values of ρ is taken as 0.8, 0.9, 0.95, 0.99, and 0.999 respectively. Also, the number of input variables (p) is set to six (6).

3.3. Criteria for evaluating estimators

The performance of the estimators is compared using the Mean Square Error (MSE) criterion. For any fitted $\widehat{y}$, MSE is defined as follows:

$$\text{MSE}(\hat{y}) = \frac{1}{2000} \sum_{i=1}^{n} \sum_{j=1}^{2000} \left(\hat{y}_{ij} - y_i\right)^2, \tag{72}$$

$$\text{Log-Likelihood} = \sum_{i=1}^{n} \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\right), \tag{73}$$

$$\text{AIC} = 2k - 2\ln(\text{Log-Likelihood}), \tag{74}$$

$$\text{BIC} = k\ln(n) - 2\ln(\text{Log-Likelihood}), \tag{75}$$

Table 2: Comparison of mean AIC and mean BIC between SFA and PCA-SFA for different sample sizes and multicollinearity levels.

Sample Size (n)	Multicollinearity (ρ)	Mean AIC		Mean BIC	
		SFA	PCA-SFA	SFA	PCA-SFA
20	0.8	159.8985641	119.5663982	165.7834924	126.1483142
	0.9	297.2111758	278.3136749	302.9043291	284.6072158
	0.95	321.0591032	254.1588451	326.9601932	260.5537689
	0.99	383.1029814	298.3828374	389.3241508	305.2779558
	0.999	219.2365098	209.5326583	224.4457895	216.2268895
50	0.8	371.5519675	240.3985592	377.1239547	246.7843789
	0.9	277.0912937	259.0466128	282.4783211	265.2345043
	0.95	225.0835647	275.3712325	231.0734572	221.6614896
	0.99	283.8459965	258.4263815	289.2496873	264.7198941
	0.999	359.4869274	289.1586543	364.8798429	295.7523046
100	0.8	224.3678328	218.2108456	229.7753895	214.0012381
	0.9	227.1063819	241.7904983	233.1998628	217.8853681
	0.95	229.4365823	247.3218465	235.5307482	223.7189862
	0.99	369.6789574	283.9842365	375.6902342	290.4823941
	0.999	339.5078485	287.2346357	344.8972341	293.7278917
250	0.8	239.6745128	214.3405963	244.6875398	220.4318924
	0.9	344.0238463	233.4539471	349.1219782	239.9472386
	0.95	209.6728345	205.0819378	215.8741928	201.5753287
	0.99	349.8523094	237.9212938	354.9752438	244.5183124
	0.999	323.6123431	229.4892143	329.4789231	236.0781241
1000	0.8	302.3124578	199.8764324	307.1253481	205.7639283
	0.9	218.3471256	210.0783917	224.5628379	215.4738293
	0.95	293.1324563	213.4789214	299.4328546	219.3758916
	0.99	343.8492317	212.3875639	349.8731235	218.6847529
	0.999	353.6891421	218.1379345	359.8701346	224.5273481

where $\hat{y}_{ij}$ is t^{th} element of the model in the j-th replication which gives the estimate of $y_1 \cdot y_n$ are the true value of " y " previously mentioned. The estimator with the minimum MSE is considered the best.

4. Results

This section presents the results from the simulation study with varying the level of multicollinearity and sample sizes in SFA model.

5. Discussion

Table 1 presents a comparative analysis of the MSE values for the Classical SFA model and the PCA-SFA model at varying sample sizes (n) ranging from 20 to 1000, along with different levels of multicollinearity represented by ρ values ranging from 0.8 to 0.999.

When examining the performance of the two models at smaller sample sizes, notably n = 20 and n = 50, a trend emerges where the Classical SFA model tends to outperform the PCA-SFA model in terms of MSE. This trend suggests that with limited data points, the added complexity introduced by the PCA approach may not yield significant improvements in predictive accuracy. At these smaller sample sizes, the Classical SFA model may provide a more stable and reliable estimation of the stochastic frontier. However, it is observed that as the sample size increases, such as at $n = 100, n = 250$, and $n = 1000$, a shift occurs in the model performance. The PCA-SFA model starts to demonstrate superior performance with lower MSE values compared to the Classical SFA model. This shift is particularly pronounced as the sample size grows, indicating that the benefits of incorporating PCA to mitigate multicollinearity become more evident and impactful with larger datasets. The PCA-SFA model's ability to handle multicollinearity and capture underlying patterns in the data improves its predictive accuracy and efficiency compared to the Classical SFA model under these conditions. It is further observed from Table 1, that higher levels of multicollinearity, reflected with higher ρ, generally lead to higher MSE values in both the Classical SFA and PCA-SFA models. Additionally, the analysis of AIC and BIC values supports these findings. As shown in Table 2, the PCA-SFA model often achieves lower AIC and BIC values compared to the Classical SFA model, especially as the sample size increases. These lower AIC and BIC values indicate that the PCA-SFA model not only provides a better fit to the data but also robust in fitting large-sample and high-multicollinearity scenarios. This aligns with the expected behaviour, as multicollinearity can introduce instability and reduce the accuracy of regression based models.

6. Conclusion

Based on the findings discussed above, it is evident that the Principal Component Analysis-based Stochastic Frontier Analysis (PCA-SFA) model offers substantial advantages over the Classical Stochastic Frontier Analysis (SFA) model, especially in scenarios with higher levels of multicollinearity and large sample size. The consistently lower mean square error (MSE), AIC and BIC of PCA-SFA across different sample sizes and levels of multicollinearity underscores its robustness and effectiveness in addressing multicollinearity challenges in econometric modeling. In conclusion, researchers and practitioners in econometrics and related fields are encouraged to consider integrating PCA-SFA into their analytical frameworks, particularly when dealing with datasets prone to multicollinearity issues. Using the power of principal component analysis to reduce the impact of multicollinearity, PCA-SFA not only improves the accuracy of the estimates, but also improves the interpretability of the results by minimizing the noise and potential biases introduced by multicollinearity. Future studies could extend the PCA-SFA estimator by exploring its application across different functional forms, such as Translog production functions, and testing it on real-world datasets in various industries. Research could also compare PCA-SFA with other dimensionality reduction techniques like partial least squares and investigate its performance with larger datasets. Incorporating heteroscedasticity, exploring the estimator's theoretical properties, and developing user-friendly software tools would further enhance its practical utility and adoption in stochastic frontier analysis.

Acknowledgment

We would like to acknowledge the Department of Statistics, Federal University of Akure, and the entire staff, both academic and non-academic, for their valuable contributions that have contributed to the success of this work.

References

[1] D. Aigner, C. A. K. Lovell & P. Schmidt, "Formulation and estimation of stochastic frontier production function models", J. Econom. **6** (1977) 21. https://doi.org/10.1016/0304-4076(77)90052-5.

[2] W. Meeusen & J. van Den Broeck, "Efficiency estimation from Cobb-Douglas production functions with composed error", Int. Econ. Rev. **18** (1977) 435. https://doi.org/10.2307/2525757.

[3] H.-J. Wang, "Heteroscedasticity and non-monotonic efficiency effects of a stochastic frontier model", J. Product. Anal. **18** (2002) 241. https://ink.springer.com/article/10.1023/A:1020638827640.

[4] K. Hadri, C. Guermat & J. Whittaker, "Estimating farm efficiency in the presence of double heteroscedasticity using panel data", J. Appl. Econ. **6** (2003) 255. https://doi.org/10.1080/15140326.2003.12040594.

[5] K. Hadri, C. Guermat & J. Whittaker, "Estimation of technical inefficiency effects using panel data and doubly heteroscedastic stochastic production frontiers", Empir. Econ. **28** (2003) 203. https://link.springer.com/article/10.1007/s001810100127.

[6] K. Hadri, "Estimation of a doubly heteroscedastic stochastic frontier cost function", J. Bus. Econ. Stat. **17** (1999) 359. https://doi.org/10.1080/07350015.1999.10524824.

[7] D. K. Christopoulos, S. E. G. Lolos & E. G. Tsionas, "Efficiency of the Greek banking system in view of the EMU: a heteroscedastic stochastic frontier approach", J. Policy Model. **24** (2002) 813. https://doi.org/10.1016/S0161-8938(02)00174-6.

[8] S. C. Kumbhakar, M. Denny & M. Fuss, "Estimation and decomposition of productivity change when production is not efficient: A panel-data approach", Econom. Rev. **19** (2000) 312. https://doi.org/10.1080/07474930008800481.

[9] M. U. Karakaplan & L. Kutlu, "Handling endogeneity in stochastic frontier analysis", Economics Bulletin, AccessEcon **37** (2017) 889. https://ideas.repec.org/a/ebl/ecbull/eb-16-00551.html.

[10] O. G. Obadina, A. F. Adedotuun & O. A. Odusanya, "Ridge estimation's effectiveness for multiple linear regression with multicollinearity: An investigation using Monte-Carlo simulations", Journal of the Nigerian Society of Physical Sciences **3** (2021) 278. https://doi.org/10.46481/jnsps.2021.304.

[11] G. A. Shewa & F. I. Ugwuowo, "Combating the multicollinearity in Bell regression model: Simulation and application", Journal of the Nigerian Society of Physical Sciences **4** (2022) 713. https://doi.org/10.46481/jnsps.2022.713.

[12] S. L. Jegede, A. F. Lukman, K. Ayinde & K. A. Odeniyi, "Jackknife Kibria-Lukman M-Estimator: Simulation and application", Journal of the Nigerian Society of Physical Sciences **3** (2022) 251. https://doi.org/10.46481/jnsps.2022.664.

[13] R. I. Rauf, A. H. Bello, B. O. Kikelomo, K. Ayinde & O. O. Alabi, "Heteroscedasticity correction measures in stochastic frontier analysis", The Annals of the University of Oradea. Economic Sciences - TOM XXXIII **1** (2024) 155. https://anale.steconomiceuoradea.ro/en/wp-content/uploads/2024/08/AUOES.July_.2024.pdf.

[14] T. J. Coelli, D. S. P. Rao, C. J. O'Donnell & G. E. Battese, *An introduction to efficiency and productivity analysis*, Springer Science & Business Media, New York, 2005. https://doi.org/10.1007/b136381.

[15] G. E. Battese & T. J. Coelli, "Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data", J. Econom. **38** (1988) 387. https://doi.org/10.1016/0304-4076(88)90053-X.

[16] W. H. Greene, "A gamma-distributed stochastic frontier model", J. Econom. 46 (1990) 141. https://doi.org/10.1016/0304-4076(90)90052-U

[17] S. B. Caudill, J. M. Ford & D. M. Gropper, "Frontier estimation and firm-specific inefficiency measures in the presence of heteroscedasticity", J. Bus. Econ. Stat. **13** (1995) 105. https://doi.org/10.1080/07350015.1995.10524583.

[18] D. J. Aigner & G. G. Cain, "Statistical theories of discrimination in labor markets", Ilr Rev. **30** (1977) 175. https://doi.org/10.1177/001979397703000204.

[19] R. Stevenson, "Measuring technological bias", Am. Econ. Rev. **70** (1980) 162. http://www.jstor.org/stable/1814745.

[20] A. K. Bera & S. C. Sharma, "Estimating production uncertainty in stochastic frontier production function models", J. Product. Anal. **12** (1999) 187. https://doi.org/10.1023/A:1007828521773.

[21] K. L. Coxe, "Principal components regression analysis", in *Encycl. Stat. Sci.*, S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic and N. L. Johnson (Ed.), Wiley, 2006. https://doi.org/10.1002/0471667196.ess2056.pub2.

[22] J. N. Darroch, "An optimal property of principal components", Ann. Math. Stat. **36** (1965) 1579. https://www.jstor.org/stable/2238448.

[23] S. Daultrey, *Principal components analysis*, Geo Abstracts Ltd., Norwich, 1976. https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://quantile.info/wp-content/uploads/2014/09/8-principle-components-analysis.pdf.

[24] G. H. Dunteman, "Principal components analysis", in *Quantitative Applications in the Social Sciences*, Sage, 1989. https://us.sagepub.com/en-us/nam/book/principal-components-analysis.

[25] J. D. Jobson, "Principal components, factors and correspondence analysis", in *Appl. Multivar. Data Anal*, Springer Texts in Statistics, Springer, New York, NY, 1992, pp. 345. https://doi.org/10.1007/978-1-4612-0921-8_4.

[26] B. Flury, *Common principal components & related multivariate models*, John Wiley & Sons, Inc., 1988. https://doi.org/10.1007/s00180-020-01041-8.

[27] I. T. Jolliffe, "Principal component analysis", Technometrics **45** (2003) 276. https://www.tandfonline.com/doi/abs/10.1198/tech.2003.s783.

[28] R. A. Johnson & D. W. Wichern, "Applied multivariate statistical analysis", Biometrics **54** (1998) 1203. https://doi.org/10.2307/2533879.

[29] W. Meredith & R. E. Millsap, "On component analyses", Psychometrika **50** (1985) 495. https://doi.org/10.1007/BF02296266.

[30] C. R. Rao, "The use and interpretation of principal component analysis in applied research", Sankhyā Indian J. Stat. Ser. A **26** (1964) 329. https://www.jstor.org/stable/25049339.

[31] D. G. Gibbons, "A simulation study of some ridge estimators", J. Am. Stat. Assoc. **76** (1981) 131. https://doi.org/10.1080/01621459.1981.10477619.

[32] D. W. Wichern & G. A. Churchill, "A comparison of ridge estimators", Technometrics **20** (1978) 301. https://doi.org/10.1080/00401706.1978.10489675.

[33] G. C. McDonald & D. I. Galarneau, "A Monte Carlo evaluation of some ridge-type estimators", J. Am. Stat. Assoc. **70** (1975) 407. https://doi.org/10.1080/01621459.1975.10479882.

[34] B. M. G. Kibria, "Performance of some new ridge regression estimators", Commun. Stat. Comput. **32** (2003) 419. https://doi.org/10.1081/SAC-120017499.

[35] A. F. Lukman & K. Ayinde, "Review and classications of the ridge parameter estimation techniques", Hacettepe J. Math. Stat. **46** (2017) 953. https://dergipark.org.tr/en/pub/hujms/issue/38493/446611.

[36] T. S. Fayose & K. Ayinde, "Different forms biasing parameter for generalized ridge regression estimator", Int. J. Comput. Appl. **181** (2019) 2. https://doi.org/10.214/ssrn.2607276.