A solution for multicollinearity in stochastic frontier production function models

Elkin Castaño and Santiago Gallón

Lecturas de Economía - No. 86. Medellín, enero-junio de 2017

Lecturas de Economía, 86 (enero-junio 2017), pp. 9-23

Elkin Castaño and Santiago Gallón

A solution for multicollinearity in stochastic frontier production function models

Abstract: *This paper considers the problem of collinearity among inputs in a stochastic frontier production model, an issue that has received little attention in the econometric literature. To address this problem, a principal-component-based solution is proposed, which allows carrying out a joint interpretation of technical efficiency and the technology parameters of the model. Applications of the method to simulated and real data show its usability and effective performance.*

Keywords: *stochastic frontier analysis, technical efficiency, productivity, multicollinearity, principal component estimation.*

JEL Classification: *C13, C18, C38, D24, O47.*

Una solución para la multicolinealidad en modelos de función de producción de frontera estocástica

Resumen: *Este artículo considera el problema de colinealidad entre insumos en un modelo de producción de frontera estocástica, un tema que ha recibido poca atención en la literatura econométrica. Para abordar el problema, se propone una solución basada en componentes principales que permite interpretar conjuntamente la eficiencia técnica y los parámetros de tecnología del modelo. Los resultados de la aplicación del método con datos simulados y reales muestran que éste es fácil de usar y presenta un buen desempeño.*

Palabras clave: *análisis de frontera estocástica, eficiencia técnica, productividad, multicolinealidad, estimación de componentes* principales.

JEL Classification: *C13, C18, C38, D24, O47.*

Une solution au probléme de la multicolinéarité dans les modéles de fonction production á frontiére stochastique

Résumé: *Cet article examine le probléme de la colinéarité concernant les inputs dans un modéle de production á frontiére stochastique, une question qui a reçu peu trés d'attention dans la littérature économétrique. Pour résoudre ce probléme, nous proposons une solution basée dans la méthode des composants principaux, laquelle permet d'interpréter á la fois l'efficacité et la technologie des paramétres techniques. Tout en utilisant des données réelles et simulées, les résultats de l'application de la méthode montrent qu'elle est facile á utiliser et elle présente en plus une bonne performance.*

Mots-clés: *analyse de frontière stochastique, efficacité technique, productivité, multicolinéarité, estimation des composants principaux.*

Classification JEL: *C13, C18, C38, D24, O47.*

Lecturas de Economía, 86 (enero-junio), pp. 9-23 © Universidad de Antioquia, 2017

A solution for multicollinearity in stochastic frontier production function models

Elkin Castaño and Santiago Gallón*

doi: 10.17533/udea.le.n86a01

Original manuscript received 8 July 2015; final version accepted 3 May 2016

Introduction

It is well known that the production frontier and technical efficiency analyses on a productive unit assume that deviations of the observed product from its maximum (or potential) attainable output, located on the production frontier, are due exclusively to inefficiencies of the productive unit (see, e.g., Kumbhakar & Lovell, 2000; Coelli, et al., 2005). For instance, if the assumed production function is a Cobb-Douglas technology $y = \boldsymbol{x}^\top \boldsymbol{\beta} + v$, where y and $\boldsymbol{x}$ are the logarithms of the observed output and the input vector respectively, then the production frontier $\boldsymbol{x}^\top \boldsymbol{\beta}$ is deterministic, and $v = y - \boldsymbol{x}^\top \boldsymbol{\beta}$ corresponds to the production inefficiency. The lack of randomness in the production frontier of this kind of models does not correspond to the

* *Elkin Castaño*: Associate Professor. Departamento de Economía, Facultad de Ciencias Económicas, Universidad de Antioquia, and Escuela de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Medellín, Colombia. Postal address: Calle 67 No. 53-108, Oficina 13-116. E-mail: elkincv@gmail.com.

Santiago Gallón: Assistant Professor. Departamento de Matemáticas y Estadística, Facultad de Ciencias Económicas, Universidad de Antioquia, Medellín, Colombia. Postal address: Calle 67 No. 53-108, Oficina 13-116. E-mail: santiago.gallon@udea.edu.co.

real economic life, where uncontrollable random production shocks occur commonly.

The stochastic frontier production model (Aigner, Lovell & Schmidt, 1977; Meeusen & van den Broeck, 1977) is specified as

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i - u_i, \qquad i = 1, \ldots, n, \tag{1}$$

where y_i is the observed output and $\boldsymbol{x}_i$ the k-dimensional vector of inputs for the ith firm, $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ and $v_i \overset{\text{iid}}{\sim} (0, \sigma_v^2)$ represent the deterministic and noise components of the frontier respectively, $\boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i$ is the maximum output reached by the firm which constitutes the stochastic frontier, and u_i is the non-negative random technical inefficiency component (i.e., the amount by which the firm fails to achieve its optimum). A symmetric distribution, such as the normal distribution, is usually assumed for v_i. It is also common to assume that v_i and u_i are independent, and that both errors are uncorrelated with $\boldsymbol{x}_i$. Typically, the production function relies on a Cobb-Douglas, translog, or any other logarithmic production model $\log(y_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i - u_i$, where the components of $\boldsymbol{x}_i$ are logarithms of inputs, its squares and cross-products.

Most of the proposed stochastic frontier models in the literature differ mainly on the assumed probability distribution function for the inefficiency component $u \geq 0$ in order to apply the maximum likelihood estimation method. In this regard, Kumbhakar and Lovell (2000), Coelli, et al. (2005), and Greene (2008) present an extensive literature about some distributions. Some instances are the half-normal model $u \sim \mathcal{N}^+(0, \sigma_u^2)$, where $\mathcal{N}^+$ denotes the non-negative half-normal distribution (Aigner, Lovell & Schmidt, 1977); the exponential model $u \sim \text{Exp}(\lambda)$, $\lambda > 0$ (Meeusen & van den Broeck, 1977; Aigner, Lovell & Schmidt, 1977); the gamma model $u \sim \Gamma(\lambda, \theta)$, $\lambda > 0$ and $\theta > 0$ (Stevenson, 1980; Greene, 1980a; Greene, 1980b); and the truncated normal $u \sim \mathcal{N}^+(\mu_u, \sigma_u^2)$ (Stevenson, 1980).

An issue with applications of stochastic frontier analysis emerges when inputs are highly correlated, from which the multicollinearity problem arises, leading to precision loss in estimates. This loss is also given by low input variability. In the presence of collinearity, it is known that: (i) separating

the individual effects of each independent variable could be a difficult task; (ii) the precision loss is expressed in large estimated variances of estimates, and hence the parameters could be non-statistically significant; (iii) the estimated coefficients can have incorrect signs and impossible magnitudes; and (iv) there are instability problems in the sense that small changes in observations, or eliminating an apparently insignificant variable, can produce large changes in estimates (see, e.g., Belsley, Kuh & Welsh, 1980; Fomby, Johnson & Hill, 1984; Groß, 2003). Therefore, it is clear that multicollinearity is a data-driven issue rather than a statistical one (Belsley, Kuh & Welsh, 1980), which can have harmful implications for the estimation of technology coefficients due to their relation with the scale returns generated by the production model.

Despite these drawbacks, a great extent of literature on stochastic frontier analysis considers the multicollinearity problem as unimportant or uses a non-statistical solution. For example, Filippini, et al. (2008) exclude the input whose correlation with other inputs is quite high in order to prevent multicollinearity. Other studies sacrifice the advantages of flexible functional forms for the deterministic component due to the cost of statistically insignificant estimates generated by unreliable parameter estimates resulting from linear dependencies between inputs (Kumbhakar & Lovell, 2000; Puig & Junoy, 2001; Filippini, 2008). Finally, others argue that, when technical inefficiency estimation is the main aim, multicollinearity is not necessarily a serious problem and the interpretation of estimates is secondary (Puig & Junoy, 2001). To the best of our knowledge, no theoretical research has been reported on studying both the stochastic frontier analysis and multicollinearity jointly.

In this paper, we propose a principal-component-based solution for multicollinearity in a stochastic frontier model. Basically, we use a re-parameterization of the model in terms of all k principal components and restrict the corresponding coefficient vector to those principal components associated to the $r < k$ nonzero eigenvalues. Finally, estimates of the original model are recovered. The solution permits a joint estimation of the technical efficiency and parameters through this better specified model. Also, through a simulation experiment, the proposed estimator is shown to be consistent and has less mean square error with respect to the traditional stochastic frontier analysis.

The rest of the paper is organized as follows. In Section I., the solution is described, and its performance is studied by a Monte Carlo simulation experiment in Section II. In Section III., an application with real data is carried out. Finally, some conclusions are given.

I. The principal component solution

For the case where there is only near exact multicollinearity (i.e., when one or more nearly exact linear relations exist among the regressors), we consider the matrix representation of the stochastic frontier production model (1),

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} - \boldsymbol{u}, \tag{2}$$

where $\boldsymbol{y}$, $\boldsymbol{v}$, $\boldsymbol{u}$, and $\mathbf{1}$ are n-dimensional vectors of observed outputs, production and inefficiency random errors, and ones respectively; $\boldsymbol{X}$ is the $n \times k$ design matrix of inputs; and $\boldsymbol{\beta}$ the corresponding k-dimensional vector of coefficients. For clarity and notational simplicity, all inputs are assumed to be standardized in the sequel.

Now, based on the spectral decomposition of the $k \times k$ symmetric matrix $\boldsymbol{X}^\top \boldsymbol{X}$,

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top,$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ is the diagonal eigenvalues matrix (with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$), and $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_k)$ the corresponding orthogonal eigenvectors matrix.

By the orthogonality of $\boldsymbol{P}$ (i.e., $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}$), the regression model (2) can be re-parameterized as

$$\begin{aligned} \boldsymbol{y} &= \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{\beta} + \boldsymbol{v} - \boldsymbol{u} \\ &= \beta_0 \mathbf{1} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{v} - \boldsymbol{u}, \end{aligned} \tag{3}$$

where $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{P} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_k)$ is the matrix of principal components $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{p}_j$ with the property $\boldsymbol{z}_j^\top \boldsymbol{z}_j = \lambda_j, \forall j$, and $\boldsymbol{\theta} = \boldsymbol{P}^\top\boldsymbol{\beta}$.

From the theory of principal component analysis –PCA– (see, e.g., Jolliffe, 2002), it is well known that the principal components $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{p}_j$ are

orthogonal, where the first principal component $\boldsymbol{z}_1$ has the maximal variance (i.e., the largest amount of information) of the original variables, the second principal component $\boldsymbol{z}_2$ has the next maximal variance after the first principal component, and so on. Note that if the jth characteristic root λ_j is approximately equal to zero, then $\boldsymbol{z}_j \approx \mathbf{0}$.

Additionally, if all k principal components are used, the same parameter vector $\boldsymbol{\beta}$ is obtained, which is unreliable under collinearity among the exogenous variables as was pointed out in the introduction. In other words, fairly small eigenvalues of the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix generate imprecisions in the OLS estimator $\widehat{\boldsymbol{\beta}}$. Therefore, the strategy consists in preventing that the estimate goes in directions $\lambda_i \boldsymbol{p}_j$ associated to fairly small λ_j (see Fomby, Johnson & Hill, 1984; Groß, 2003).

Thus, to deploy the strategy, we restrict $\boldsymbol{\beta}$ into the subspace spanned by the columns $\lambda_1 \boldsymbol{p}_1, \lambda_2 \boldsymbol{p}_2, \ldots, \lambda_r \boldsymbol{p}_r$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ are the $r < k$ largest eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ and $\lambda_{r+1} \approx \lambda_{r+2} \approx \cdots \approx \lambda_k \approx 0$. This means that range$(\boldsymbol{X}) = r$. Hence, in order to eliminate imprecisions, Massy (1965), Jolliffe (1982), Mason and Gunst (1985), and Hwang and Nettleton (2003) suggest using (i) the first principal components with the largest variance and highly correlated with output y, and (ii) those principal components of low variance but with high output correlation.

Therefore, the model (3) can be re-expressed using the subdivision of the eigenvalues into groups $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and $\lambda_{r+1} \approx \lambda_{r+2} \approx \cdots \approx \lambda_k \approx 0$ and defining the corresponding partition $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2) = (\boldsymbol{X}\boldsymbol{P}_1, \boldsymbol{X}\boldsymbol{P}_2)$, where $\boldsymbol{Z}_1$ is the $n \times r$ matrix with principal components associated to the nonzero eigenvalues and $\boldsymbol{Z}_2$ the $n \times (k - r)$ matrix with the rest of the principal components associated to the eigenvalues approximately equal to zero. Then, assuming that the first r principal components are highly correlated with y in order to simplify the notation, and using $\boldsymbol{Z}_2 \approx \mathbf{0}$, the reparameterized model (3) can be expressed as

$$\begin{aligned} \boldsymbol{y} &= \beta_0 \mathbf{1} + \boldsymbol{Z}_1 \boldsymbol{\theta}_1 + \boldsymbol{Z}_2 \boldsymbol{\theta}_2 + \boldsymbol{v} - \boldsymbol{u} \\ &= \beta_0 \mathbf{1} + \boldsymbol{Z}_1 \boldsymbol{\theta}_1 + \boldsymbol{v} - \boldsymbol{u}, \end{aligned}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top\right)^\top$, with $\boldsymbol{\theta}_1 = \boldsymbol{P}_1^\top \boldsymbol{\beta}_1$ and $\boldsymbol{\theta}_2 = \boldsymbol{P}_2^\top \boldsymbol{\beta}_2$. The constraint $\boldsymbol{Z}_2 \approx \mathbf{0}$ is equivalent to $\boldsymbol{\theta}_2 \approx \mathbf{0}$.

Finally, the least squares estimator of $\boldsymbol{\theta}_1$ is $\widehat{\boldsymbol{\theta}}_1 = \left(\boldsymbol{Z}_1^\top \boldsymbol{Z}_1\right)^{-1} \boldsymbol{Z}_1^\top \boldsymbol{y}$. Thus, the principal component estimator of $\boldsymbol{\beta}$ in (2) is given by

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{P}_1 \widehat{\boldsymbol{\theta}}_1, \tag{4}$$

with covariance matrix $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{P}_1 \mathrm{Cov}(\widehat{\boldsymbol{\theta}}_1) \boldsymbol{P}_1^\top$.

II. Simulation study

To evaluate the performance of the proposed principal-component-based method, we carried out a Monte Carlo simulation experiment with 20,000 replications on the stochastic frontier model

$$\log(y_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) + v_i - u_i, \quad i = 1, \ldots, n(= 100), \tag{5}$$

with a half-normal/normal specification, $u_i \overset{\text{iid}}{\sim} \mathcal{N}^+(0, \sigma_u^2)$ and $v_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2)$, where $\sigma_u = 3$, $\sigma_v = 2.5$, $\sigma^2 = \sigma_u^2 + \sigma_v^2 = 15.25$, $\gamma = \sigma_u^2/\sigma^2 = 0.59$, $(\beta_0, \beta_1, \beta_2) = (1, 0.8, 0.7)$; and $(x_1, x_2) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (20, 25)$ and $\boldsymbol{\Sigma} = \boldsymbol{DRD}$, where $\boldsymbol{D} = \mathrm{diag}(\sigma_{x_1}, \sigma_{x_2}) = \mathrm{diag}(1, 2)$; and $\boldsymbol{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ with $\rho = \mathrm{Corr}(x_1, x_2) = 0.7,\ 0.8,\ 0.9$. For the most severe multicollinearity problem, where $\rho = 0.9$, we performed the simulations with $n = 1000$ to study the large sample properties of the estimator. We used the frontier: Stochastic Frontier Analysis `R` package version 1.1-0 by Coelli and Henningsen (2013).

Tables 1-3 show the means, biases, and mean squared errors −MSE− of estimators of β_1 and β_2 approximated by the principal-component-based ($\widehat{\beta}_1^{\text{pc}}$ and $\widehat{\beta}_2^{\text{pc}}$) and the usual stochastic frontier analysis ($\widehat{\beta}_1^{\text{sfa}}$ and $\widehat{\beta}_2^{\text{sfa}}$) methods for the assumed values of ρ. Results indicate that, in general, the coefficient estimators obtained with the principal-component-based method are biased, as these biases do not decrease asymptotically. However, the estimators have less MSE with respect to the ones obtained by the traditional method, even in large samples. The usual estimators are biased for finite samples with greater

biases than for the proposed method, although these decrease asymptotically. The estimations for γ and σ_2 remain unaffected if the principal components are chosen correctly. Finally, when keeping fixed the number of principal components, the biases increase as the linear relationship among variables decreases.

Table 1. $\rho = 0.7$

n	$\widehat{\beta}_1^{\text{pc}}$	$\widehat{\beta}_1^{\text{sfa}}$	$\widehat{\beta}_2^{\text{pc}}$	$\widehat{\beta}_2^{\text{sfa}}$	$\widehat{\sigma}_{\text{pc}}^2$	$\widehat{\sigma}_{\text{sfa}}^2$	$\widehat{\gamma}_{\text{pc}}$	$\widehat{\gamma}_{\text{sfa}}$
				Mean				
100	0.991	0.851	0.618	0.717	15.341	15.320	0.525	0.531
				Bias				
100	0.191	0.051	−0.082	0.017	0.091	0.070	−0.065	−0.059
				MSE				
100	3.400	8.829	2.113	5.502	5.053	5.127	0.315	0.315

Source: author's elaboration.

Table 2. $\rho = 0.8$

n	$\widehat{\beta}_1^{\text{pc}}$	$\widehat{\beta}_1^{\text{sfa}}$	$\widehat{\beta}_2^{\text{pc}}$	$\widehat{\beta}_2^{\text{sfa}}$	$\widehat{\sigma}_{\text{pc}}^2$	$\widehat{\sigma}_{\text{sfa}}^2$	$\widehat{\gamma}_{\text{pc}}$	$\widehat{\gamma}_{\text{sfa}}$
				Mean				
100	0.983	0.901	0.611	0.677	15.502	15.492	0.537	0.544
				Bias				
100	0.183	0.101	−0.089	−0.023	0.252	0.242	−0.053	−0.046
				MSE				
100	3.312	10.542	2.062	6.523	5.014	5.079	0.306	0.305

Source: author's elaboration.

Table 3. $\rho = 0.9$

n	$\widehat{\beta}_1^{\text{pc}}$	$\widehat{\beta}_1^{\text{sfa}}$	$\widehat{\beta}_2^{\text{pc}}$	$\widehat{\beta}_2^{\text{sfa}}$	$\widehat{\sigma}^2_{\text{pc}}$	$\widehat{\sigma}^2_{\text{sfa}}$	$\widehat{\gamma}_{\text{pc}}$	$\widehat{\gamma}_{\text{sfa}}$
				Mean				
100	0.974	0.600	0.606	0.853	15.574	15.563	0.542	0.549
1000	0.953	0.791	0.593	0.695	15.153	15.143	0.574	0.574
				Bias				
100	0.174	−0.200	−0.094	0.153	0.324	0.313	−0.049	−0.041
1000	0.153	−0.009	−0.107	−0.005	−0.097	−0.107	−0.016	−0.016
				MSE				
100	3.211	14.525	1.997	9.048	5.037	5.094	0.302	0.300
1000	1.005	4.449	0.627	2.763	1.791	1.793	0.111	0.112

Source: author's elaboration.

III. Application

To see how the proposed solution behaves with real data, we use the production data of the agricultural and livestock sector with a sample of $n = 23$ livestock farms. The output variable is the total income, and inputs are labor, capital and other inputs; all have been measured in nominal Colombian −COL− pesos.

Then, a stochastic frontier production model was fitted assuming a Cobb-Douglas functional form with normal-exponential specification, $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $u_i \overset{iid}{\sim} Exp(\lambda)$, $\lambda > 0$. Estimations were carried out using the LIMited DEPendent −LIMDEP− econometric software (version 10). As can be seen in Table 4 the only statistically significant parameter is the input corresponding to $\log(\text{Other inputs}_2)$. Although the variable $\log(Capital)$ is insignificant, its estimated coefficient has an unexpected opposite sign, indicating a signal of possible multicollinearity.

Table 4. *Estimated Stochastic Frontier Production Function*

Variable	$\widehat{\beta}_j$	$\widehat{\text{s.e.}(\widehat{\beta}_j)}$	$\widehat{\beta}_j/\widehat{\text{s.e.}(\widehat{\beta}_j)}$	$\mathbb{P}(\lvert Z\rvert \geq z)$	$\overline{X}_j$
Constant	4.30	2.42	1.77	0.076	
log(Labor)	0.25	0.36	0.68	0.498	17.92
log(Other inputs$_1$)	0.11	0.17	0.62	0.534	14.89
log(Other inputs$_2$)	0.53	0.22	2.38	0.018	17.83
log(Capital)	−0.10	0.15	−0.62	0.53	5.60
Variance parameters for compound error					
γ	3.33	2.07	1.60	0.11	
σ_u	0.26	0.13	2.05	0.04	

Source: author's elaboration.

To detect multicollinearity, we computed the scaled condition indexes. Table 5 shows there are two harmful condition indexes (with values greater than 30), indicating two possible near-linear dependencies among inputs. Thus, under the multicollinearity problem, we applied the proposed principal-component-based solution. The proportion of variance explained by the first principal component was 88.6%. Therefore, we applied the solution using this principal component. Table 6 displays the corresponding results. Based on these results, the estimates of the principal-component-based stochastic frontier using the equation (4) are in Table 7. Results show that all inputs are statistically significant with correct signs in accordance to production theory.

Table 5. *Condition Indexes*

Condition Index
1.000
12.829
42.981
101.730

Source: author's elaboration.

Table 6. *Estimated Principal Component Model*

Variable	$\widehat{\beta}_j$	$\widehat{\text{s.e.}(\widehat{\beta}_j)}$	$\widehat{\beta}_j/\widehat{\text{s.e.}(\widehat{\beta}_j)}$	$\mathbb{P}(\lvert Z\rvert \geq z)$	$\overline{X}_j$
Constant	19.33	0.18	107.2	0.000	
PC_1	0.39	0.05	7.23	0.000	0.15e-12
Variance parameters for compound error					
γ	2.75	0.98	2.82	0.005	
σ_u	0.29	0.07	3.90	0.00	

Source: author's elaboration.

Table 7. *Estimated Principal-Component-Based Stochastic Frontier Model*

Variable	$\widehat{\beta}_j$	$\widehat{\text{s.e.}(\widehat{\beta}_j)}$	$\widehat{\beta}_j/\widehat{\text{s.e.}(\widehat{\beta}_j)}$
log(Labor)	0.1658	0.0229	7.2335
log(Capital)	0.1969	0.0272	7.2321
log(Other inputs$_1$)	0.1914	0.0264	7.2336
log(Other inputs$_2$)	0.2182	0.0301	7.2331
Scale returns	0.7723		

Source: author's elaboration.

Conclusions

Based on simulation results, the estimators for inputs obtained under the proposed principal-component-based solution are biased, and such biases do not decrease asymptotically. Besides, the estimators have less MSE with respect to the usual ones even in large samples. For finite samples, the estimators are biased, and seem to have greater biases than the principal-component-based estimators. Also, the bias diminishes when the sample size increases. If the principal components are correct, the estimation of $\gamma = \sigma_u^2/\sigma^2$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$ remains unaffected with the proposed method. Furthermore, when keeping fixed the number of principal components, the biases of the proposed estimator increase as the linear relation between covariates decreases. The choice of the number of principal components is critical to the estimation of $\boldsymbol{\beta}$, γ and σ^2, as well as for the efficiency component. After applying the proposed method on real data from the agricultural and livestock sectors to evaluate its technical inefficiency, our method seems to provide better estimation results for the coefficients, as well as for the scale returns, in comparison with the traditional method.

References

Aigner, Dennis; Lovell, Knox & Schmidt, Peater (1977). "Formulation and estimation of stochastic frontier production function models", *Journal of Econometrics*, Vol. 6, Issue 1, pp. 21-37.

Belsley, David; Kuh, Edwin & Welsh, Roy (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Inc.

Coelli, Timothy & Henningsen, Arne (2013). *Frontier: Stochastic Frontier Analysis. Retrieved from*: http://CRAN.R-Project.org/package=frontier. R package version 1.1-0. (Accessed on July 2014).

Coelli, Timothy; Rao, Prasada D.S.; O'Donnell, Christopher J. & Battese, George E. (2005). *An Introduction to Efficiency and Productivity Analysis* (2nd. Ed.). New York: Springer.

Filippini, Massimo; Hrovatin, Nevenka & Zoric, Jelena (2008). "Cost efficiency of slovenian water distribution utilities: an application of stochastic frontier methods", *Journal of Productivity Analysis*, Vol. 29. Issue 2, pp. 169-182.

Fomby, Thomas B.; Johnson, Stanley R. & Hill, Carter (1984). *Advanced Econometric Methods*. New York: Springer.

Greene, William (1980a). "Maximum likelihood estimation of econometric frontier functions", *Journal of Econometrics*, Vol. 13, Issue 1, pp. 27-56.

Greene, William (1980b). "On the estimation of a flexible frontier production model", *Journal of Econometrics*, Vol. 13, Issue 1, pp. 101-115.

Greene, William (2008). "The econometric approach to efficiency analysis". In: Fried, Harold; Lovell, Knox & Schmidt, Shelton (Eds.), *The Measurement of Productive Efficiency and Productivity Growth* (pp. 92-150). New York, Oxford University Press.

Groß, Jürgen (2003). "Linear Regression", *Lecture Notes in Statistics*, Vol. 175. Springer.

Hwang, Gene J. T. & Nettleton, Dan (2003). "Principal components regression with data chosen components and related methods", *Technometrics*, Vol. 45, No. 1, pp. 70-79.

Jolliffe, Ian T. (1982). "A note on the use of principal components in regression", *Journal of the Royal Statistical Society*. Series C (Applied Statistics), Vol. 31, No. 3, pp. 300-303.

Jolliffe, Ian T. (2002). *Principal Component Analysis* (2nd Ed.). New York: Springer.

Kumbhakar, Subal C. & Lovell, C. Knox (2000). *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press.

Mason, Robert & Gunst, Richard (1985). "Selecting principal components in regression", *Statistics and Probability Letters*, Vol. 3, Issue 6, pp. 299-301.

Massy, William F. (1965). "Principal components regression in exploratory statistical research", *Journal of the American Statistical Association*, Vol. 60, Issue 309, pp. 234-256.

Meeusen, Wim & van Den Broeck, Julien (1977). "Efficiency estimation from Cobb-Douglas production functions with composed error", *International Economic Review*, Vol. 18, No. 2, pp. 435-444.

Puig-Junoy, Jaume (2001). "Technical inefficiency and public capital in U.S. states: A stochastic frontier approach", *Journal of Regional Science*, Vol. 41, Issue 1, pp. 75-96.

Stevenson, Rodney (1980). "Likelihood functions for generalized stochastic frontier estimation", *Journal of Econometrics*, Vol. 13, Issue 1, pp. 58-66.