Modeling dependence between error components of the stochastic frontier model using copula: Application to Intercrop Coffee Production in Northern Thailand

Aree Wiboonpongse[1], Jianxu Liu[2], Songsak Sriboonchitta[2] and Thierry Denoeux [3]

[1] Faculty of Economics, Prince of Songkla University, Thailand
[2] Faculty of Economics, Chiang Mai University, Thailand
[3] UMR 7253 Heudiasyc, Université de Technologie de Compiègne, CNRS, France

December 19, 2014

Abstract

In the standard stochastic frontier model, the two-sided error term V and the one-sided technical inefficiency error term W are assumed to be independent. In this paper, we relax this assumption by modeling the dependence between V and W using copulas. Nine copula families are considered and their parameters are estimated using maximum simulated likelihood. The best model is then selected using the AIC or BIC criteria. This methodology was applied to coffee production data from Northern Thailand. For these data, the best model was the one based on the Clayton copula. The main finding of this study is that the dependence between V and W is significant and cannot be ignored. In particular, the standard stochastic frontier model with independence assumption grossly overestimated the technical efficiency of coffee production. These results call for a reappraisal of previous production efficiency studies using the SFM with independence assumption, which may occasionally lead to overoptimistic conclusions.

Keywords: Stochastic Frontier, Copula, Technical Efficiency, Econometrics.

1 Introduction

The Stochastic frontier model (SFM) has proved very useful to assess technical efficiency of production units. The stochastic frontier production model for a cross-section of observations was independently proposed by Aigner et al. [1] and Meeusen and Van den Broeck [20]. It is essentially a linear regression model with two independent error components: a two-sided term that captures random variation of the production frontier across firms and a one-sided term that measures inefficiency relative to the frontier. In recent decades, most studies about production, cost or profit efficiency have used the conventional SFM (see, e.g., [1, 9, 10, 15, 19, 22–24, 27–30, 33–36]). In all these studies, it is assumed that the one-sided and two-sided error terms are independent. Based on this assumption, the parameters of the SFM can be estimated using the corrected ordinary least squares or maximum likelihood methods.

The impact of the independence assumption on technical efficiency estimation has long remained an open issue. This assumption can be relaxed by using copula to fit the joint distribution of the two random error components more appropriately. Smith [26] first proposed a SFM allowing for dependence between the two error components using copula functions. Copula functions can be used to capture rank correlation and tail dependence between the two error components, thus making the stochastic frontier analysis much more flexible. However, the log-likelihood function in the copula-based SFM generally does not have a closed form, which makes its maximization numerically intricate.

In this paper, we propose to use the maximum simulated likelihood method, which has numerical and computational advantages over the numerical integration method used by Smith [26]. Furthermore, to explore the dependence structure of the error components in the SFM, we systematically consider several copula families including the Student-t, Clayton, Gumbel and Joe families as well as their relevant rotated versions. The model with the best fit-complexity trade-off is selected using the AIC or BIC criteria. This approach was applied to cross-sectional data about coffee production in Thailand. A comparison between technical efficiencies computed with and without the independence assumption (considering the best copula model) reveals that the standard approach grossly overestimates efficiency, which has important implication for production analysis using the SFM.

The remainder of this paper is organized as follows. Section 2 introduces the necessary background on the SFM and copula. Section 3 presents the copula-based stochastic frontier approach. Empirical results with this model applied to coffee production data are reported in Section 4. Finally, Section 5 concludes the paper.

2 Background and theory

The SFM is a regression-like model with a disturbance term that is asymmetric and distinctly non-normal. This model will first be briefly summarized in Section 2.1. Some background on copula will then be recalled in Section 2.2. These are the two building blocks of the copula-based model introduced in Section 3.

2.1 Stochastic frontier model

Classical models of production [13,19] consider ideal (i.e., highest achievable) production as a function $h(\boldsymbol{x}, \boldsymbol{\beta})$ of a vector $\boldsymbol{x}$ of inputs, where $\boldsymbol{\beta}$ is a vector of parameters. As real production Y can only be less than the ideal one, it can be written as

$$Y = h(\boldsymbol{x}, \boldsymbol{\beta}) \cdot TE, \tag{1}$$

where $TE < 1$, called technical efficiency, is the ratio of actual output y to maximum feasible output $h(\boldsymbol{x}, \boldsymbol{\beta})$. For instance, the Cobb-Douglas production model [37] can be written as

$$\ln Y = \boldsymbol{x}'\boldsymbol{\beta} - W, \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of coefficient and $W = -\ln(TE)$ is a non-negative error term. However, a theoretical problem with this approach is that any measurement error on Y must be embedded in the disturbance W, making the estimation of $\boldsymbol{\beta}$ very sensitive to outliers. To solve this problem, Aigner et al. [1] proposed to add a symmetric random noise V to the right-hand side of (2), resulting in the following model,

$$\ln Y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, \tag{3a}$$

$$\varepsilon = V - W, \tag{3b}$$

where the two error components W and V are assumed to be independent. In this model, the frontier $\exp(\boldsymbol{x}'\boldsymbol{\beta}+V)$ is stochastic, hence the term "stochastic frontier". The disturbances ε are now assumed to arise from two sources: (1) productive inefficiency, resulting in a non-negative error term W, and

(2) firm-specific effects V, which can enter the model with either signs. The technical efficiency TE can then be written as

$$TE = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + V - W)}{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + V)} = \exp(-W). \tag{4}$$

The inefficiency error term W is usually assumed to have a gamma, exponential, or half-normal distribution (defined as the distribution of the absolute value of a normal variable) [11]. In contrast, the symmetric error term V is usually assumed to have a normal or logistic distribution.

2.2 Copula

A copula connects a given number of one-dimensional marginal distributions to form a joint multivariate distribution [21]. In the following, we will limit the presentation to bivariate copula, which will be used later. Sklar's theorem [25] states that any cumulative distribution function (cdf) $F(x_1, x_2)$ of a two-dimensional random vector (X_1, X_2) can be expressed as

$$F(x_1, x_2) = C\left(F_1(x_1), F_2(x_2)\right), \tag{5}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the marginal cdfs of X_1 and X_2, and C is a bivariate function, called a *copula*. If X_1 and X_2 are independent, then C is the product. A function $C : [0,1]^2 \to [0,1]$ is a copula if and only if it satisfies the following properties:

1. $C(u_1, 0) = C(0, u_2)$ for all u_1 and u_2 in $[0,1]$;
2. $C(u_1, 1) = u_1$ and $C(1, u_2) = u_2$ for all u_1 and u_2 in $[0,1]$;
3. For all $0 \le u_1 \le u_2 \le 1$ and $0 \le v_1 \le v_2 \le 1$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) - C(u_1, v_1) \ge 0. \tag{6}$$

The Fréchet-Hoeffding theorem states that the following bounds hold any copula C:

$$\max(u_1 + u_2 - 1, 0) \le C(u_1, u_2) \le \min(u_1, u_2), \tag{7}$$

for any $(u_1, u_2) \in [0,1]^2$. The lower and upper Fréchet-Hoeffding bounds correspond to two extreme forms of dependence in which the two variables are, respectively, countermonotonic and comonotonic.

If the random vector (X_1, X_2) has a joint density $f(x_1, x_2)$, it can be expressed as a function of the copula density,

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2} \tag{8}$$

by the following formula,

$$\begin{align} f(x_1, x_2) &= \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} \tag{9a} \\ &= \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2} \frac{\partial F_1(x_1)}{\partial x_1} \frac{\partial F_2(x_2)}{\partial x_2} \tag{9b} \\ &= c(u_1, u_2) f_1(x_1) f_2(x_2), \tag{9c} \end{align}$$

where $f_1(x_1)$ and $f_2(x_2)$ are the marginal densities.

The most common measure of dependence between random variables is Pearson's correlation coefficient. However, it only measures linear dependence and is not very informative for asymmetric distributions [5]. To measure nonlinear dependence, rank correlation coefficients such as Kendall's tau and Spearman's rho are more suitable. They can be expressed in terms of the copula as [21]

$$\tau(X_1, X_2) = 4 \iint_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1 = 4\mathbb{E}[C(U_1, U_2)] - 1, \tag{10}$$

$$\rho(X_1, X_2) = 12 \iint_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 3 = 12\mathbb{E}[C(U_1, U_2)] - 3, \tag{11}$$

where (U_1, U_2) is a two-dimensional random vector with $C(u_1, u_2)$ as a cdf.

Another important concept is that of tail dependence, which describes extreme comovements in the tail of the joint distribution of (X_1, X_2). The lower and upper tail dependence coefficients are defined, respectively, as

$$\lambda_L = \lim_{u \to 0^+} \Pr\left[X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)\right] = \lim_{u \to 0^+} \frac{C(u, u)}{u} \tag{12}$$

and

$$\lambda_U = \lim_{u \to 1^-} \Pr\left[X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)\right] = \lim_{u \to 1^-} \frac{1 - 2u + C(u, u)}{1 - u}, \tag{13}$$

where F_1 and F_2 are the marginal cumulative distribution functions of X_1 and X_2, respectively.

To fit a copula to data, one commonly uses families of copula $\{C_\theta\}$ depending on some parameter θ. The copula families used in this study are summarized in Appendix A (see [21,31,32]). Copula families can be selected based on their ability to capture positive and negative dependence, as well as tail dependence. Some copulas, such as the Gaussian and Frank copulas, possess the characteristic $\lambda_U = \lambda_L = 0$, while most copulas can capture upper or lower tail dependence. For instance, Clayton copulas can measure lower tail dependence, while Gumbel and Joe copulas can measure upper tail dependence. Student-t copula reflects symmetric tail dependence [6].

All of these copulas, except the Gaussian, T and Frank copulas, can only capture positive dependence. However, a copula may be "rotated". There are three rotated forms, with angles 90 degrees, 180 degrees and 270 degrees, defined as follows,

$$C_{90}(u_1, u_2) = u_2 - C(1 - u_1, u_2), \tag{14a}$$

$$C_{180}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, u_2), \tag{14b}$$

$$C_{270}(u_1, u_2) = u_1 - C(u_1, 1 - u_2). \tag{14c}$$

Rotation by 90 and 270 degrees allows for the modeling of negative dependence. A thorough review of rotated copulas may be found in [6, 8, 16].

To summarize, different copulas have different characteristics, such as upper tail dependence, lower dependence, positive and negative dependences, etc. Therefore, the above copula families and relevant rotated copula can potentially capture the appropriate dependence between two random variables. Other popular copula families, such as the Farlie-Gumbel-Morgenstern (FGM) and Ali-Mikhail-Haq (AMH) copulas, although widely used in the econometrics literature, have been discarded because they cannot achieve the Fréchet bounds and they can only accommodate relatively weak dependence between the margins. The ranges of dependence of Kendall's tau and Spearman's rho for the FGM copula are $[-2/9, 2/9]$ and $[1/3, 1/3]$, respectively; the Kendall's tau of the AMH copula is bounded in the interval $[0.1817, 0.3333]$, and the range of Spearman's rho is $[-0.2711, 0.4784]$ (see [18, 26]).

3 Copula-based stochastic frontier model

The classical SFM makes the strong assumption that the error components V and W are independent. Smith [26] proposed to relax this assumption and to model the dependence between V and W using a copula. The classical model is recovered as a special case corresponding to the product copula. Following Smith [26], the joint density $f(w,\varepsilon)$ can be obtained from $f(w,v)$ using (9) as

$$f(w,\varepsilon) = f(w,w+\varepsilon) = f_W(w)f_V(w+\varepsilon)c_\theta(F_W(w),F_V(w+\varepsilon)). \tag{15}$$

Marginalizing out W then yields

$$f_\Theta(\varepsilon) = \int_0^{+\infty} f(w,\varepsilon)dw, \tag{16}$$

or, equivalently,

$$f_\Theta(\varepsilon) = \mathbb{E}_W\left[f_V(W+\varepsilon)c_\theta(F_W(W),F_V(W+\varepsilon))\right], \tag{17}$$

where $\mathbb{E}_W[\cdot]$ denotes the expectation with respect to the technical inefficiency W and Θ represents the vector of all parameters of the marginals and the copula function.

Assuming the data to consist of independent cross-sectional observations of n individuals or firms, the log-likelihood function for model (3) is given by

$$L(\boldsymbol{\beta},\sigma_w,\sigma_v,\theta) = \sum_{i=1}^{n} \log f_\Theta(\varepsilon_i) = \sum_{i=1}^{n} \log f_\Theta(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{18}$$

where y_i is the realization of the output from individual or firm i, $\boldsymbol{x}_i$ is explanatory variable vector for individual i, and σ_w and σ_v are scale parameters of the marginal distributions of W and V, respectively. As discussed by Smith [26], there are very few density function of ε for which the maximum likelihood estimate has a closed-form expression. However, each term $f(\varepsilon_i)$ in (18) can be written as an expectation using (17), and can easily be approximated using Monte Carlo simulation. The log-likelihood (18) can then be maximized using the maximum simulated likelihood method [7, 14, 17].

Assuming W and V to have, respectively, half-normal and normal distributions, the density function of ε can be written as

$$f(\varepsilon) = \int_0^\infty f_W(w) f_V(w+\varepsilon) c_\theta\left(F_W(w), F_V(w+\varepsilon)\right) dw \tag{19a}$$

$$= \int_0^\infty \frac{2\exp\left(-\frac{w^2}{2\sigma_w^2}\right)}{\sqrt{2\pi}\sigma_w} f_V(w+\varepsilon) c_\theta(F_W(w), F_V(w+\varepsilon)) dw \tag{19b}$$

$$= \int_0^\infty \frac{2\exp\left(-\frac{(\sigma_w w_0)^2}{2\sigma_w^2}\right)}{\sqrt{2\pi}\sigma_w} f_V(\sigma_w w_0+\varepsilon) c_\theta(F_W(\sigma_w w_0), F_V(\sigma_w w_0+\varepsilon)) d\sigma_w w_0 \tag{19c}$$

$$= \int_0^\infty \frac{2\exp\left(-\frac{w_0^2}{2}\right)}{\sqrt{2\pi}} f_V(\sigma_w w_0+\varepsilon) c_\theta(F_W(\sigma_w w_0), F_V(\sigma_w w_0+\varepsilon)) dw_0. \tag{19d}$$

It can then approximated by

$$\hat{f}(\varepsilon) = \frac{1}{N}\sum_{r=1}^{N} f_V(\sigma_w w_{0,r}+\varepsilon) c_\theta(F_W(\sigma_w w_{0,r}), F_V(\sigma_w w_{0,r}+\varepsilon)), \tag{20}$$

where $w_{0,r}$, $r = 1, \ldots, N$, is a sequence of N random draws from the standard half-normal distribution. The simulated log-likelihood is then

$$L_S(\boldsymbol{\beta}, \sigma_w, \sigma_v, \theta) = \sum_{i=1}^{n} \log\left[\frac{1}{N}\sum_{r=1}^{N} f_V(\sigma_w w_{0,ir}+\varepsilon_i) c_\theta(F_W(\sigma_w w_{0,ir}), F_V(\sigma_w w_{0,ir}+\varepsilon_i))\right]. \tag{21}$$

Following Battese and Coelli [3], the parameters (σ_w, σ_v) can be transformed to (λ, σ) with $\lambda = \sigma_w/\sigma_v$ and $\sigma = \sqrt{\sigma_w^2 + \sigma_v^2}$. The larger λ, the greater the inefficiency component in the model [12]. We can also measure the global inefficiency by $\gamma = \sigma_w^2/(\sigma_w^2 + \sigma_v^2)$. The values of λ and γ reveal whether inefficiency plays an important role in the composite error term [13].

In stochastic frontier analysis, the technical efficiency terms (4) are of primary interest. They are not observed, but we can estimate their conditional expectations given ε,

$$TE_\Theta = \mathbb{E}\left[\exp(-W)|\varepsilon\right] \tag{22a}$$

$$= \frac{1}{f_\Theta(\varepsilon)} \int_0^{+\infty} \exp(-w) f(w, \varepsilon) \mathrm{d}w \tag{22b}$$

$$= \frac{\mathbb{E}_W[\exp(-W) f_V(W+\varepsilon) c_\theta(F_W(W), F_V(W+\varepsilon))]}{E_W[f_V(W+\varepsilon) c_\theta(F_W(W), F_V(W+\varepsilon))]}. \tag{22c}$$

The nominator and denominator on the right-hand side of (22c) can be approximated using Monte Carlo simulation by

$$\widehat{TE}_{\Theta} = \frac{A}{B}, \tag{23a}$$

with

$$A = \frac{1}{N}\sum_{i=1}^{N} \exp(-\sigma_w w_{0,ir}) f_V(\sigma_w w_{0,ir} + \varepsilon_i) c_\theta(F_W(\sigma_w w_{0,ir}), F_V(\sigma_w w_{0,ir} + \varepsilon_i)) \tag{23b}$$

and

$$B = \frac{1}{N}\sum_{i=1}^{N} f_V(\sigma_w w_{0,ir} + \varepsilon_i) c_\theta(F_W(\sigma_w w_{0,ir}), F_V(\sigma_w w_{0,ir} + \varepsilon_i)). \tag{23c}$$

4 Empirical results

The data used in this study were collected by interviewing farmers in the Chiang Mai province of Thailand [2]. The upland areas of the Chiang Mai province, in particular, proved fertile areas for high quality coffee. The area has lower humidity, a shorter monsoon season, and a lower annual temperature, creating a micro-climate among the Chiang Mai mountains. Rainwater from the mountains seeps into the soil, making it particularly rich in mineral content.

A questionnaire was constructed to ask for details about the irrigated rice production at the farms. In particular, there was interest in the area grown, the yields obtained, as well as the use of fertilizer and labor. Information was also obtained on social characteristics of the sample farmers. Data on a sample of 111 farmers were obtained in the survey. In this study, we considered the relation between production output, fertilizer and labor. These data are displayed in Figure 1.

4.1 Estimation of the stochastic frontier model

We considered the following linear translog production model

$$\begin{aligned}\ln Y_i = \beta_0 + \beta_1 \ln(labor) + \beta_2 \ln(fertilizer) +& \\ \frac{\beta_3}{2}(\ln(labor))^2 + \frac{\beta_4}{2}(\ln(fertilizer))^2 +& \\ \beta_5 \ln(labor)\ln(fertilizer) + V_i - W_i,&\end{aligned} \tag{24}$$

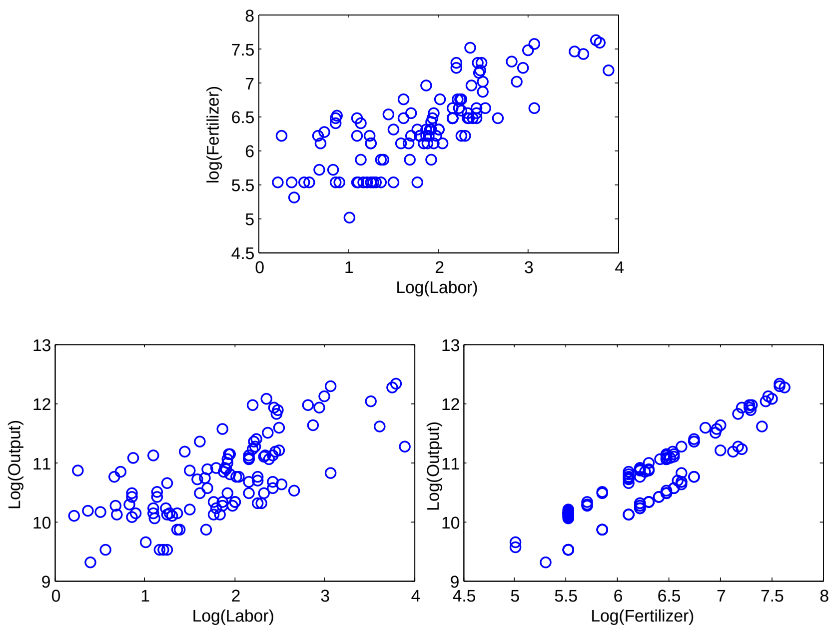

Figure 1: Matrix plot of coffee production data.

where Y_i represents intercrop coffee output of farmer i. Fertilizer and labor are expressed, respectively, in kg/ha and in man-days per hectare. We made the usual assumptions of normal and half-normal distributions for noise terms V_i and inefficiency terms W_i, respectively. In addition to the independence copula, we considered nine different copula families (see Appendix A): Gaussian copula, T copula, Frank copula, Clayton copula, Gumbel copula, Joe copula, rotated Clayton copula (180°), rotated Gumbel copula (180°) and rotated Joe copula (180°). The independence and Gaussian copulas were estimated first, to obtain initial values and determine the sign of correlation between V and W. If it is negative, the rotated copulas at 90° and 270° should be used in this model. The simulated log-likelihood function (21) was computed using $N = 500$ and maximized using the Nelder-Mead algorithm in the R statistical software, using starting values obtained from the `sfa` function in the package `frontier` in R.

Figure 2 displays the AIC and BIC for each copula-based SFM. According to both criteria, the best model is the one based on the Clayton copula, whereas the independence copula performs the worst. This result confirms the interest of relaxing the assumption of independence between the two error components of the SFM. We also performed the likelihood ratio (LR)

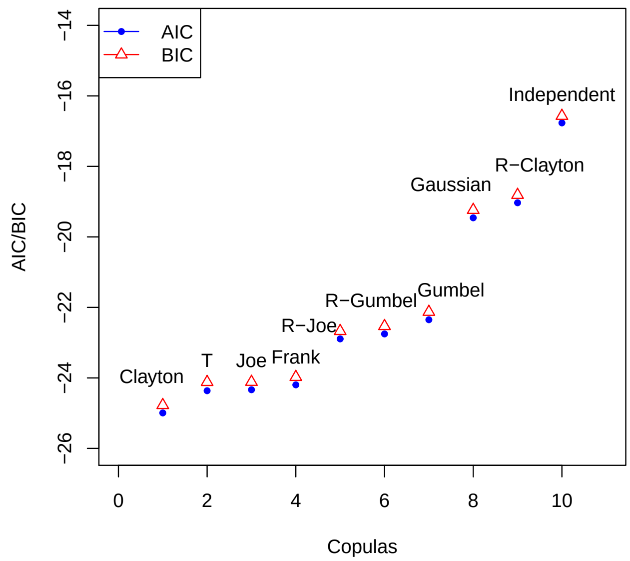

Figure 2: AIC and BIC for each copula-based stochastic frontier model.

test to test the null hypothesis of independence between U and V, vs. the alternative hypothesis of a dependence structure characterized by each of the copula families. The LR test statistic has approximately a chi-squared distribution with degrees of freedom equal to the difference of the number of parameters of the two models. The results are reported in Table 1. We can be see that, except in one case, the p-values are less than 10%, which implies that the null hypothesis of independence is rejected with a confidence level at least equal to 10%.

We now turn to the frontier estimates of the best model based in the

Table 1: LR texts between independence copula and other copula models. R-C, R-G and R-Joe represent rotated Clayton, rotated Gumbel and rotated Joe copulas by 180 degrees, respectively.

	Gau	T	Cla	Fra	Gum	Joe	R-Cla	R-Gum	R-Joe
LR stat.	2.71	7.63	8.25	7.45	5.60	7.59	2.28	6.01	6.15
p-value	0.10	0.02	< 0.01	0.01	0.02	0.01	0.13	0.01	0.01

Clayton copula. The results are shown in Table 2. Most parameters are significantly non-null at the 10% level. The estimated parameter of the Clayton copula is 4.65, which is significantly different from zero at the 10% level. This result confirms that significant dependence between error components exists, thereby justifying the use of copula-based SFM. Parameter γ, which measures the relative importance of the technical inefficient term, is equal to 0.99. The estimated Kendall's tau and Spearman's rho are, respectively, 0.70 and 0.87. Additionally, the lower tail dependence coefficient equals 0.86.

4.2 Technical efficiencies

Figure 3 displays the technical efficiencies computed using the best model based on the Clayton copula (solid line) and the independent copula-based model (interrupted line) in term of technical efficiency. The efficiency range is 0.44-0.99 (average 0.73) for the classical model but only 0.35-0.95 (average 0.66) for the best copula-based model. Obviously, the usual assumption of independence between the random error and inefficiency terms result in a severe overestimation of technical efficiencies in this study.

Figure 4 shows the distribution of technical efficiency scores for all sample farmers using the independence and Clayton copula models. According to the classical SFM with independence assumption, more than half of the farmers have high technical efficiencies (i.e., greater than 0.7). In contrast, according to the best model, a large proportion (39%) of the sample farmers have low technical efficiency scores (< 0.50). This finding suggests that a considerable amount of productivity is lost due to inefficiency.

More generally, we may wonder whether positive or negative dependence between the error components of the SFM systematically result in, respectively, overestimation and underestimation of technical efficiencies. To try to answer this question, we performed a Monte Carlo experiment. Model (24) was used to generate new samples of the same size as the dataset used in this study, but with different degrees of dependence between V and W. The cofactor values (labor and fertilizer) were the same as in the original data, and the coefficients β_j as well as the parameters of the marginal distributions of V and W were assigned their maximum likelihood estimates computed from the real data. Five datasets were then generated by simulating the error terms V_i and W_i from a joint distribution based on Clayton copulas with parameter θ such that the Kendall's correlation correlation coefficient between V and W equals -0.9, 0.4, 0, 0.4 and 0.9. (The Clayton copula was rotated by 90° to obtain negative dependence.) The conventional

Table 2: Estimated parameters and standard errors for the Clayton-copula-based model. Significance codes: ***=0.001,**=0.01,*=0.05,$\cdot$=0.1.

Parameters	β_0	β_1	β_2	β_3	β_4	β_5
Estimates	12.96***	1.09 $\cdot$	-1.80***	0.09	0.49***	-0.19
s.e	1.66	0.66	0.57	0.09	0.11	0.12
Parameters	σ_w	σ_u	θ	$\log L$	LR stats	τ
Estimates	0.59***	0.07***	4.65 $\cdot$	12.58	8.25**	0.70***
s.e	0.05	0.01	2.40			0.06
Parameters	ρ	λ	σ^2	γ	λ_U	λ_L
Estimates	0.87***	8.35***	0.35***	0.99***	0	0.86
s.e	0.05	0.70	0.03	0.02		

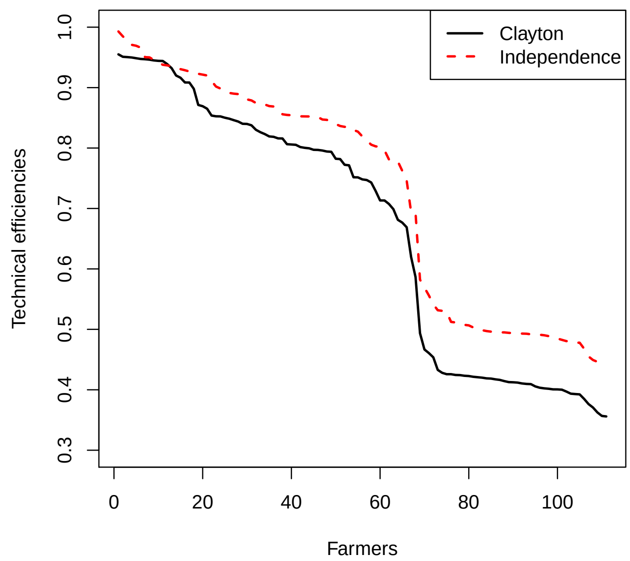

Figure 3: Technical efficiencies for the independence and Clayton copula-based models.

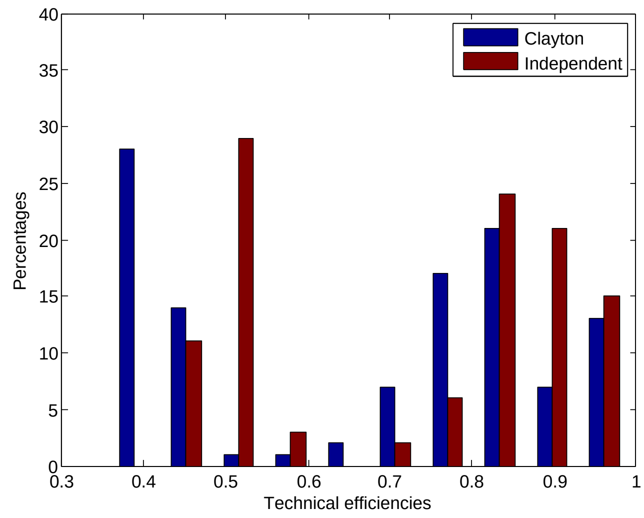

Figure 4: Distribution of technical efficiencies for the independence and Clayton copula models

and Cayton-copula-based SFMs were then estimated using each of the five datasets. The conventional model thus wrongly assumed independence, except for the dataset corresponding to $\tau = 0$. The distribution of differences between technical efficiencies computed from the two models are shown in Figure 5. A positive difference corresponds to overestimation of technical efficiency for the conventional model. We can verify that large positive and negative dependence between the error terms result, respectively, in severe over- and underestimation of technical efficiencies by the conventional model based on the independence assumption.

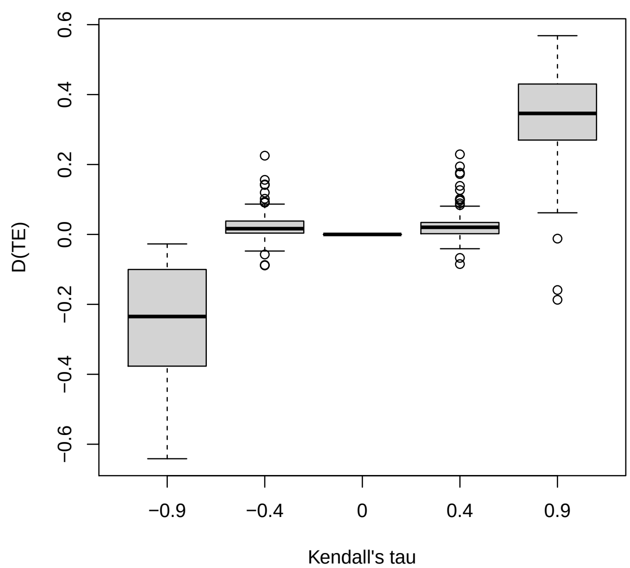

Figure 5: Distribution of differences between technical efficiencies estimated using the conventional and copula-based models, for five datasets with randomly generated error terms V_i and W_i with $\tau \in \{-0.9, -0.4, 0, 0.4, 0.9\}$.

5 Conclusions

The SFM is a linear regression model in which the error component is assumed to consist in the sum of a random error term V with a symmetric distribution and a non-negative inefficiency term W. In the classical model, these two terms are assumed to be independent. Following Smith [26], we have relaxed this assumption and modeled the dependency between the two error components using copula. The methodology proposed in this paper is based on (1) the consideration of a large number of copula families to capture a wide range of dependence patterns between V and W, (2) the use of the maximum simulated likelihood method to estimate the parameters of both the copula and the marginal distribution, and (3) the selection of the best model using the AIC and BIC criteria. In this study, the marginal distributions of the errors were assumed to be normal and half-normal, but other distributions could be considered as well, and tested jointly with the copula families.

This methodology was applied to intercrop coffee production data from Northern Thailand. For these data, all investigated copula families outperformed the independence copula according to the AIC and BIC criteria, the best results being obtained for the Clayton copula model. The main finding of this study is that the independence assumption leads to a gross overestimation of production efficiency. These results raise the question of the reliability of stochastic frontier analyses based on the classical model. They suggest that the dependence between random error and the inefficiency term cannot be ignored and call for a systematic use of copula-based models in stochastic frontier analyses. The economic interpretation of this observed dependence is an interesting and important open issue, which is left for further research.

Acknowledgements

The support of the Thailand Research Fund (TRF) for the Faculty of Economics and the Faculty of Agriculture in Chiang Mai University is here gratefully acknowledged. We also gratefully thank Profs Murray D. Smith, Hung T. Nguyen and Vladik Kreinovich for their valuable comments and suggestions.

References

[1] D.J. Aigner, C.A.K. Lovell, P.S. Schmidt, Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics* 6 (1977) 21-37.

[2] K.N. Ayuttaya. Planting Season Choice and Technical Efficiency in Thailand's Arabica Coffee and Macadamia Nut Production: Case Study Northern of Thailand. Diploma thesis, Chiang Mai University, 2011.

[3] G.E. Battese, T.J. Coelli, Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India, *The journal of productivity analysis* 3 (1992) 153-169.

[4] C.R. Bhat, N. Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling, Transportation Research Part B: Methodological 43 (7) (2009) 749-765.

[5] B.H. Boyer, M.S. Gibson, M. Loretan, Pitfalls in tests for changes in correlations. Federal Reserves Board, *IFS Discussion Paper* No 597R (1999).

[6] E.C. Brechmann, U. Schepsmeier, Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine, *Journal of Statistical Software* 52 (3) (2013) 1-29.

[7] R. Burns, The Simulated Maximum Likelihood Estimation of Stochastic Frontier Models with Correlated Error Components. Sydney: The University of Sydney. (2004).

[8] C. Cech, Copula-based top-down approaches in financial risk aggregation, 2006. Available from: http://ssrn.com/abstract=953888 or http://dx.doi.org/10.2139/ssrn.953888.

[9] T.J. Coelli, D.S.P. Rao and G.E. Battese, An Introduction to Efficiency and Productivity Analysis. Boston: Kluwer. (1998).

[10] T.J. Coelli, S. Rahman, C. Thirtle, Technical, allocative, cost and scale efficiencies in Bangladesh rice cultivation: A non-parametric approach, *Journal of Agricultural Economics* 53 (2002) 607-626.

[11] K. Cullinane, D.W. Song. Estimating the Relative Efficiency of European Container Ports: A Stochastic Frontier Analysis. Research in Transportation Economics.Volume 16 (2006) 85-115.

[12] W. Greene, Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems, Health Econ. 13 (2004) 959-980.

[13] W. H. Greene. Econometric analysis, 6th edition, Prentice Hall, Upper Saddle River, NJ, USA, 2008

[14] W. Greene, A stochastic frontier model with correction for sample selection, *Journal of Productivity Analysis* Springer 34(1) (2010) 15-24.

[15] J. Jondrow, C.A.K. Lovell, I.S. Materov, P. Schmidt, On the estimation of technical inefficiency in the stochastic frontier production function model, *Journal of Econometrics* 19 (1982) 233-38.

[16] J. Luo, Stepwise estimation of D-Vines with arbitrary specified copula pairs and EDA Tools, Diploma thesis, Technische Universität München, 2010.

[17] C. Kao, L. Lee, M.M. Pitt. Simulated Maximum Likelihood Estimation of The Linear Expenditure System with Binding Non-negativity Constraints, *Annals of Economics and Finance* 2 (2001) 203-223.

[18] P. Kumar. Probability Distributions and Estimation of Ali-Mikhail-Haq Copula, Applied Mathematical Sciences 4 (14) (2010) 657-666.

[19] S.C. Kumbhakar, C.A.K. Lovell, Stochastic Frontier Analysis. Cambridge UK: Cambridge University Press, 2000.

[20] W. Meeusen, J. van den Broeck, Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review* 18 (1977) 435-444.

[21] R.B. Nelsen. An Introduction to Copulas (2nd ed.), Springer-Verlag, New York (2006).

[22] S. Rahman, A. Wiboonpongse, S. Sriboonchitta, Y. Chaovanapoonphol, Production Efficiency of Jasmine Rice Producers in Northern and North-eastern Thailand, *Journal of Agricultural Economics* 60 (2) (2009) 419-435.

[23] S. Rahman, A. Wiboonpongse, S. Sriboonchitta, K. Kanmanee, Total factor productivity growth and convergence in Northern Thai agriculture, *African Journal of Agricultural Research* 17(7) (2012) 2689-2700.

[24] S. Rahman, M.K. Hasan, Impact of environmental production conditions on productivity and efficiency: A case study of wheat farmers in Bangladesh, *Journal of environmental management* 88(4) (2008) 1495-1504.

[25] M. Sklar, Fonctions de repartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8 (1959) 229-231.

[26] M.D. Smith, Stochastic frontier models with dependent error components, *Econometrics Journal* 11 (2008) 172-192.

[27] S. Sriboonchitta, Evaluation of Cost Efficiency of Thai Public Universities, *International Journal of Intelligent Technologies & Applied Statistics* 5 (2012) 361-374.

[28] S. Sriboonchitta, A. Wiboonpongse, On the Estimation of Stochastic Production Frontiers with Self-Selectivity: Jasmine and Non-Jasmine Rice in Thailand, CMU. Journal 4 (2005) 105-124.

[29] S. Sriboonchitta, A. Wiboonpongse, Technical Efficiency of Rural Micro and Community Enterprises in the Upper North of Thailand. Asia-Pacific Productivity Conference, Seoul National University, Seoul, Korea (2006) 17-19.

[30] S. Sriboonchitta, A. Wiboonpongse, The Effects of Production Inputs, Technical Efficiency and Other Factors on Jasmine and Non-Jasmine Rice Yields in Thailand. Asia-Pacific Productivity conference, Brisbane, Australia (2004) 14-16.

[31] S. Sriboonchitta, H.T. Nguyen, A. Wiboonpongse, J. Liu, Modeling volatility and dependency of agricultural price and production indices of Thailand: Static versus time-varying copulas, *International Journal of Approximate Reasoning* 54 (2013) 793-808.

[32] P. Trivedi, D. Zimmer. Copula Modeling: An Introduction for Practitioners', *Foundations and Trends in Econometrics* 1 (1) (2005) 1-111.

[33] A. Wiboonpongse, S. Sriboonchitta, Agricultural Production Growth Assessment for Agroeconomic Zones in Northern Thailand Using Statistical Data. Asia Pacific Productivity Conference 2008, Academic Sinica, Taipei, Taiwan, July 17-19, 2008.

[34] A. Wiboonpongse, S. Sriboonchitta, Impact of Technical Skill on Potato Production Efficiency in Thailand. Asia Pacific Productivity Conference 2008, Academic Sinica, Taipei, Taiwan, July 17-19, 2008.

[35] A. Wiboonpongse, S. Sriboonchitta, E.G. Battese, The Effects of Production Input, Technical Inefficiency and Biological Risk on Jasmine

and Non-Jasmine Rice Yields in Thailand, The Empirical Economics Letters 4 (3) (2005).

[36] A. Wiboonpongse, S. Sriboonchitta, S. Rahman, P. Calkins, Thanes Sriwichailumphun, Joint determination of the choice of planting season and technical efficiency of potato in Northern Thailand: A comparison of Greene's versus Heckman's sample selection approach, *African Journal of Business Management* 6 (12) (2012) 4504-4513.

[37] A. Zellner and J. Kmenta and F. Drèze. Specification and estimation of Cobb-Douglas production functions. *Econometrica* 34 (1966) 784-795.

A Copula families used in this study

In the following, we briefly summarize the main properties of copula families used in this study.

Gaussian copulas

Gaussian copulas take the form

$$C_{Ga}(u_1, u_2; \rho) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{x_1^2 + x_2^2 - 2r x_1 x_2}{2(1-r^2)}\right] \mathrm{d}x_1 \mathrm{d}x_2, \quad (25)$$

where $-1 < r < 1$ is Pearson's correlation coefficient and Φ is the cdf the standard normal distribution function. This is the copula pertaining to a bivariate normal distribution with standard normal marginals and Pearson's linear correlation coefficient r. Parameter ρ is related to Kendall's τ and Spearman's ρ coefficients by the relations $\tau = (2/\pi)\sin^{-1}(r)$ and $\rho = (6/\pi)\sin^{-1}(r/2)$.

T copula

As Gaussian copulas, T copulas belong to the class of elliptical copula (i.e., they are the copulas of elliptically contoured distributions). However, T copulas can capture tail dependence. They are defined as

$$C(u_1, u_2 | r, \nu) = \int_{-\infty}^{T_\nu^{-1}(u_1)} dx \int_{-\infty}^{T_\nu^{-1}(u_2)} dy \left[1 + \frac{x^2 - 2rxy + y^2}{\nu(1-r^2)}\right]^{-\frac{\nu+2}{2}}, \quad (26)$$

where r is the Pearson's correlation coefficient and T_ν is the cdf of a Student distribution with ν degrees of freedom. When ν tends to infinity, the T copula converges to the Gaussian copula. The symmetric tail dependence coefficient can be calculated as a function of parameters r and ν as

$$\lambda_U = \lambda_L = 2T_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-r}{1+r}}\right). \tag{27}$$

Clayton copula

A copula of the form $C(u_1, u_2) = \varphi^{-1}(\varphi(u_1)+\varphi(u_2))$ is said to be Archimedean and function φ is said to be its generator [21]. There are different Archimedean copulas based on different generators, such as the Clayton, Gumbel, Frank, Joe and AMH copulas, among others. The Clayton family can reflect lower tail dependence for $\theta > 0$. It is characterized by the following formula,

$$C_{Cl}(u_1, u_2|\theta) = (u_1^\theta + u_2^\theta - 1)^{-1/\theta}. \tag{28}$$

This copula can only capture a strong lower tail and positive dependence, but it can be rotated and used in capture negative dependence or reflect strong upper tail dependence (see [8,16]). The corresponding Kendall's tau and lower tail dependence coefficients are, respectively, $\tau_{CL} = \theta/(\theta+2)$ and $\lambda_L = 2^{-1/\theta}$. The expression of Spearman's rho as a function of θ is more complicated.

Frank copula

Copulas in the Frank family are the only Archimedean copulas that attain the lower and upper Fréchet bounds, thus allowing for positive and negative dependence. The corresponding copula function is given by

$$C_{Fr}(u_1, u_2) = -\frac{1}{\theta}\ln\left(1+\frac{(\exp(-\theta u_1)-1)(\exp(-\theta u_2)-1)}{\exp(-\theta)-1}\right), \tag{29}$$

where $\theta \in (-\infty, +\infty)\setminus\{0\}$. Positive (resp., negative) values of θ correspond to positive (resp., negative) dependence. The independence copula is recovered in the limit when $\theta \to 0$. The rank correlation coefficients are given by

$$\tau = 1 - \frac{4}{\theta} + 4\frac{D_1(\theta)}{\theta} \tag{30a}$$

and

$$\rho = 1 - \frac{12}{\theta}[D_1(\theta) - D_2(\theta)], \tag{30b}$$

where $D_k(\theta)$ is Debye function

$$D_k(\theta) = \frac{k}{\theta^k} \int_0^\theta \frac{t^k}{\exp(t) - 1} dt, \quad k = 1, 2. \tag{30c}$$

Gumbel copula

The bivariate Gumbel copula is given by

$$C_{Gum}(u_1, u_2|\theta) = \exp\left[-\left((-\ln u_1)^{1/\theta} + (-\ln u_2)^{1/\theta} \right)^\theta \right], \tag{31}$$

where $\theta \in (1, +\infty)$. It is an asymmetric Archimedean copula that allows for strong upper tail dependence. The rotated Gumbel copula can be applied to capture negative dependence as well. The Kendall's tau of Gumbel copula is given by $\tau_{Gum} = 1 - 1/\theta$, but the Spearman's rho does not have a closed form. The upper tail dependence coefficient is $\lambda_U = 2 - 2^{1/\theta}$.

Joe copula

The Joe copula is defined as follows:

$$C_{Joe}(u_1, u_2|\theta) = 1 - \left[(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta (1-u_2)^\theta \right]^{1/\theta}, \tag{32}$$

where $\theta \geq 1$. This copula can capture upper tail dependence, as does Gumbel copula. But it can capture a stronger upper tail dependence than does Gumbel copula (see [4]); the coefficient of upper tail dependence is $\lambda_U = 2 - 2^{1/\theta}$. The Kendall's tau is related to parameter θ by

$$\tau_{Joe} = 1 + \frac{4}{\theta^2} \int_0^1 t \ln(t)(1-t)^{2(1-\theta)/\theta} dt. \tag{33}$$

The relationship between Spearman's rho and parameter θ does not have a closed form expression. The rotated Joe copula can describe negative dependence as well.

Highlights

- A copula-based stochastic frontier model is investigated.
- The use of copula allows us to capture dependency between the two error components.

- The method was applied to coffee production in Northern Thailand.
- The conventional stochastic frontier model severely overestimates the technical efficiencies.