A double-copula stochastic frontier model with dependent error components and correction for sample selection

Songsak Sriboonchitta*, Jianxu Liu*,* Aree Wiboonpongse†, and Thierry Denoeux‡

* Faculty of Economics,
Chiang Mai University, Thailand
†Faculty of Economics,
Prince of Songkla University, Thailand
‡ Sorbonne Universités, Université de Technologie de Compiègne,
CNRS, UMR 7253 Heudiasyc, France

August 24, 2016

Abstract

In the standard stochastic frontier model with sample selection, the two components of the error term are assumed to be independent, and the joint distribution of the unobservable in the selection equation and the symmetric error term in the stochastic frontier equation is assumed to be bivariate normal. In this paper, we relax these assumptions by using two copula functions to model the dependences between the symmetric and inefficiency terms on the one hand, and between the errors in the sample selection and stochastic frontier equation on the other hand. Several families of copula functions are investigated, and the best model is selected using the Akaike Information Criterion (AIC). The methodology was applied to a sample of 200 rice farmers from Northern Thailand. The main findings are that (1) the double-copula stochastic frontier model outperforms the standard model in terms of AIC, and (2) the standard model underestimates the technical efficiency scores, potentially resulting in wrong conclusions and

*Corresponding author email: liujianxu1984@163.com

recommendations.

Keywords: stochastic frontier model, copula representation, dependence, families of copula, sample selection, self-selection, technical efficiency.

1 Introduction

The original stochastic frontier model with sample selection was proposed by Greene [6], who provided a general framework for sample selection procedures in stochastic frontier models. This model has been widely used in empirical analyses. For example, Flores et al. [4] examined the impact of *Plataformas de Concertación* (a program aimed at linking small holders to high-value agricultural markets in Ecuador) on productivity growth. Rahman and Rahman [12] evaluated sustainability of maize cultivation in terms of energy use while taking into account factors affecting choice of the growing season and farmers' production environment. Wollni and Brümmer [17] investigated technology choice, productivity and efficiency of coffee farm households in Costa Rica. Rahman et al. [11] evaluated the determinants of switching to Jasmine rice as well as the determinants of Jasmine rice productivity in northern and north-eastern Thailand, etc.

Although the original stochastic frontier model with sample selection has been widely used to analyze technical efficiencies and productivity of crops, it has some limitations. First, the model is usually fitted using a two-stage estimation method, which implies that estimators may not be efficient. Second, the two components of the error term in the stochastic frontier equation are assumed to be independent. This assumption can be relaxed by using copula to fit the joint distribution of the two random error components more appropriately [14], [16]. Third, the original stochastic frontier model with sample selection assumes that the unobservable in the sample selection equation is related to the random error term in the stochastic frontier equation, but these two quantities are further assumed to have a bivariate normal distribution. The restricted form of the bivariate normal distribution may result in strongly biased estimates of parameters and technical efficiencies. To overcome this limitation, Smith [13] proposed a more general copula-based approach to account for data selectivity. Generally speaking, there is no statistical or economic reason to enforce independence between the two error components, or linear correlation between the errors in the stochastic frontier and sample selection equations.

To address these issues, we propose a double-copula stochastic frontier

model with sample selection. In this approach, copula functions are used to model the dependence of the symmetric and asymmetric error components, as well as the dependence between errors of the sample selection and stochastic frontier equations. Several families of copula functions, such as the Gaussian, Frank, Clayton, Gumbel and Joe families and their relevant rotated versions are systematically considered. Each model is fitted globally using the maximum simulated likelihood method [5], and the best model is selected using the Akaike information criterion (AIC). This approach was evaluated using both simulated data and cross-sectional data of rice production in Northern Thailand.

The remainder of this paper is organized as follows. The background on copula functions and sample selection is first recalled in Section 2. Our double-copula stochastic frontier model with sample selection is then introduced in Section 3. The simulation study is then presented in Section 4.1 and the application to rice production efficiency analysis is described in Section 4.2. Finally, Section 5 concludes the paper.

2 Background

In this section, we first recall some basic definitions and results about copula functions in Section 2.1. The sample selection model is then briefly presented in Section 2.2.

2.1 Copula functions

A recent trend in statistics and econometrics is to relax the multivariate Gaussian or Student-t distribution assumptions by using more flexible copula functions [10]. For example, Smith [13] used copula functions to relax the restrictive bivariate normal distributional assumption of the standard Heckman's model; Wu et al. [18] and Sriboonchitta et al. [15] used copula-based generalized autoregressive conditional heteroskedasticity (GARCH) model instead of multivariate GARCH models because the former does not need a multivariate normal or Student-t distribution assumption. A copula function is used to connect the specified marginals of each variable to form a multivariate distribution [10]. In this study, we focus on the presentation of bivariate copula, which will be used later. Given a joint distribution function H of two continuous random variables X and Y, the function $C : [0,1]^2 \to [0,1]$ defined by

$$C(u_1, u_2) = H(F^{-1}(u_1), G^{-1}(u_2)) \tag{1}$$

is a copula; here F and G are the marginal distributions of X, Y, respectively, and F^{-1} and G^{-1} are the corresponding quantile functions. If the random vector (X, Y) has a joint density $h(x, y)$, this density can be expressed as a function of the copula density c by

$$h(x,y) = \frac{\partial^2 H(x,y)}{\partial x \partial y} = c[F(x), G(y)]f(x)g(y), \tag{2}$$

where $f(x)$ and $g(y)$ are the marginal densities.

Different families of copula functions have different characteristics and limitations. For example, Gaussian copulas cannot capture tail dependences; Clayton copulas can capture lower tail dependence, while Gumbel copulas can be used to model upper tail dependence. In this study, we used six families of copula functions with relevant rotated versions: the independent, Gaussian, Clayton, Frank, Gumbel, Joe, rotated Clayton, rotated Gumbel, and rotated Joe copulas, to capture potential dependence structure in copula-based stochastic frontier model with sample selection. The main characteristics of copula families used in this study are summarized in Table 1. Kendall's τ coefficient can be computed from the copula function as

$$\tau(X,Y) = 4 \iint_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1. \tag{3}$$

The lower and upper tail dependence coefficients are defined, respectively, as

$$\lambda_L = \lim_{u \to 0^+} P\left[Y \leq G^{-1}(u) | X \leq F^{-1}(u)\right] = \lim_{u \to 0^+} \frac{C(u,u)}{u} \tag{4}$$

and

$$\lambda_U = \lim_{u \to 1^-} P\left[Y > G^{-1}(u) | X > F^{-1}(u)\right] = \lim_{u \to 1^-} \frac{1 - 2u + C(u,u)}{1-u}. \tag{5}$$

Figure 1 displays several copula contour plots under standard normal distribution. The contour plots are generated based on the value of Kendall's tau equals to 0.7. These plots illustrate the fact that different copula functions have different characteristics in terms of tail dependences. These copulas can only capture positive dependence except the Gaussian, T and Frank copulas. However, they can be rotated to capture negative dependence. Thorough reviews of rotated copulas may be found in Cech [3] and Wiboonpongse et al. [16].

Table 1: Main characteristics of copula function families used in this study: distribution function, Kendall's correlation coefficient τ, upper and lower tail dependence coefficients λ_U and λ_L, and parameter range.

Copula	$C(u_1, u_2; \theta)$	τ	λ_U	λ_L	θ range
Gaussian	$\Phi(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta)$	$2 \arcsin \theta / \pi$	—	—	(-1, 1)
Gumbel	$\exp(-w^{1/\theta})$	$1 - 1/\theta$	$2 - 2^{1/\theta}$	—	$[1, +\infty)$
	$w = (-\ln(u_1)^\theta) + (-\ln(u_2)^\theta)$				
Clayton	$(u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$	$2/(2+\theta)$	—	$2^{-1/\theta}$	$(0, +\infty)$
Frank	$-\frac{1}{\theta} \ln(1 + (e^{-\theta} - 1)^{-1} w_1 w_2)$	$1 + \frac{4}{\theta}(D_1(\theta) - 1)$	—	—	$(-\infty, +\infty)/0$
	$w_i = \exp(-\theta u_i) - 1, i = 1, 2$				
Joe	$1 - (w_1 + w_2 - w_1 w_2)^{1/\theta}$	$1 + \frac{4}{\theta^2} g(\theta)$	$2 - 2^{1/\theta}$	—	$[1, +\infty)$
	$w_i = (1 - u_i)^\theta, i = 1, 2$				

Note: $D_1(\theta) = \int_0^\theta \frac{x/\theta}{\exp(x) - 1} dx$ and $g(\theta) = \int_0^1 x \log(x)(1-x)^{2(1-\theta)/\theta} dx$.

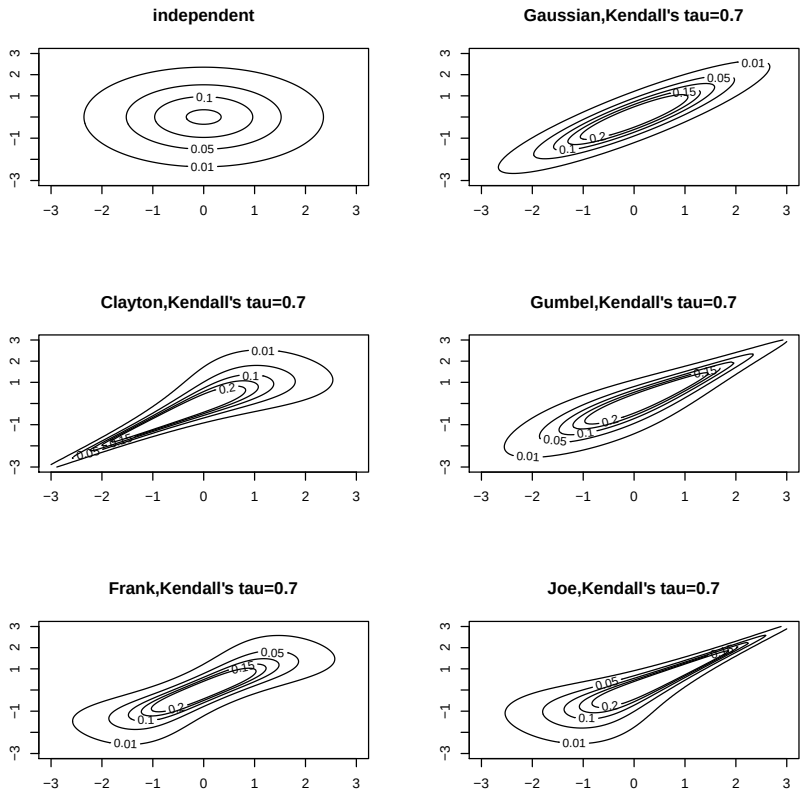

Figure 1: Contour plots of copula functions with normal marginal distributions.

2.2 Sample selection model

Heckman's sample selection model [7] was introduced to deal with the problem of selection bias. Ignoring non-random selection in statistical models can have severe detrimental effects on parameter estimation (see, e.g., Marra and Radice [9]; Bushway et al. [2]). Heckman's sample selection model for linear regression is a two-equation system. The *selection equation* determines if the dependent variable of the regression model is observed or not. It is defined as

$$S_i = \begin{cases} 1 & \text{if } Y_i^* = \alpha^T z_i + \xi_i \geq 0 \\ 0 & \text{if } Y_i^* = \alpha^T z_i + \xi_i < 0 \end{cases}, \quad i = 1, \ldots, n, \tag{6}$$

where α is a vector of coefficients, z_i a vector of exogenous variables, ξ_i an error term assumed to have a standard normal distribution, and Y_i^* a latent variable. The *outcome equation* is the classical linear regression model

$$Y_i = \beta^T x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{7}$$

where Y_i is the dependent variable, β is a vector of coefficients, x_i a vector of covariates, and ε_i an error term assumed to have a normal distribution with mean 0 and variance σ_ε^2. The depend variable Y_i is observed only if $S_i = 1$. The joint distribution of ξ_i and ϵ_i is assumed to be bivariate normal. The full information maximum likelihood estimation and the two-step estimation methods for this model were developed by Heckman [7] (see also [8]). The two-step estimation approach involves estimating a Probit model for selection, and then inserting a correction factor – the inverse Mills ratio, calculated from the Probit model – into the ordinary least squares estimator of the linear model of interest. The initial model was generalized by Smith [13], who proposed to model the dependency between ξ_i and ε_i by a copula, thus relaxing the bivariate normality assumption. The likelihood function for Smith's model is

$$L(\alpha, \beta, \sigma_\varepsilon) = \prod_{\{i:s_i=0\}} P(Y_i^* \leq 0) \prod_{\{i:s_i=1\}} P(Y_i^* > 0) f_c(y_i | y_i^* > 0), \tag{8}$$

where f_c is the conditional probability density function of Y given the event $S = 1$. Following [13], this conditional density can be written as

$$f_c(y|y^* > 0) = \frac{1}{1 - P(Y^* \geq 0)} \frac{\partial \left\{ P(Y \leq y) - P(Y^* \leq 0, Y \leq y) \right\}}{\partial y} \tag{9a}$$

$$= \frac{1}{1 - F_\xi(-\alpha^T z)} \left. \frac{\partial \left\{ F_\varepsilon(\varepsilon) - H(-\alpha^T z, \varepsilon) \right\}}{\partial \varepsilon} \right|_{\varepsilon = y_2 - \beta^T x} \tag{9b}$$

$$= \frac{1}{1 - F_\xi(-\alpha^T z)} \left\{ f_\varepsilon(y - \beta^T x) - \left. \frac{\partial H(-\alpha^T z, \varepsilon)}{\partial \varepsilon} \right|_{\varepsilon = y - \beta^T x} \right\}, \tag{9c}$$

where the joint distribution function H of ε and ξ can be expressed using copula function C_θ, and

$$\frac{\partial H(-\alpha^T z, \varepsilon)}{\partial \varepsilon} = \left. \frac{\partial C_\theta \left(F_\xi(-\alpha^T z), v \right)}{\partial v} \right|_{v = F_\varepsilon(\varepsilon)} f_\varepsilon(\varepsilon). \tag{10}$$

3 Methodology

In this section, we first recall the standard stochastic frontier model, and its version with copula-based modeling of the dependence between the inefficiency and noise terms (Section 3.1). Our double-copula model with sample selection is then introduced in Section 3.2.

3.1 Stochastic frontier model

Aigner et al. [1] proposed a stochastic frontier model that is now commonly used to estimate a production function and to obtain farm-level technical efficiency estimates. The basic form of stochastic frontier model is given by

$$Y_i = \beta^T x_i + \varepsilon_i, \tag{11a}$$

$$\varepsilon_i = V_i - W_i, \tag{11b}$$

$i = 1, \ldots, n$, where Y_i represents the output of individual i, x_i is a vector of input quantities, β is a vector of coefficients, and the random error term is divided into two parts: a firm-specific effect V_i (that can be positive or negative) and a positive inefficiency term W_i. The optimal frontier output pursued by individual i is $\beta' x_i + V_i$; it is stochastic, hence the term "stochastic frontier". In the classical stochastic frontier model, the two components

V_i and W_i of the error term are assumed to be independent. Typically, it is assumed that V_i has a normal distribution $\mathcal{N}(0, \sigma_v^2)$, and W_i has a half normal distribution with scale parameter σ_w, i.e., $W_i = |U_i|$ with $U_i \sim \mathcal{N}(0, \sigma_w^2)$. The technical efficiency of individual i is defined as $\exp(-W_i)$.

Most studies on stochastic frontier analysis assume V and W to be independent. However, this assumption can be relaxed by using a copula function (see [14] and [16]). The likelihood function for the stochastic frontier model is

$$L(\beta, \sigma_v, \sigma_w) = \prod_{i=1}^{n} f_\varepsilon(\varepsilon_i) = \prod_{i=1}^{n} f_\varepsilon(y_i - \beta x_i), \tag{12}$$

where $f_\varepsilon(\varepsilon_i)$ is the density function of the random error term. The error density $f_\varepsilon(\varepsilon)$ can be computed as follows [14]. First, we can express the joint density of (W, ε) from that of (W, V) and use (2) to obtain

$$f(w, \varepsilon) = f(w, w + \varepsilon) = f_W(w) f_V(w + \varepsilon) c_\theta(F_W(w), F_V(w + \varepsilon)). \tag{13}$$

Marginalizing out W then yields

$$f_\varepsilon(\varepsilon) = \int_0^{+\infty} f(w, \varepsilon) dw, \tag{14}$$

or, equivalently,

$$f_\varepsilon(\varepsilon) = \mathbb{E}_W \left[f_V(W + \varepsilon) c_\theta(F_W(W), F_V(W + \varepsilon)) \right], \tag{15}$$

where $\mathbb{E}_W[\cdot]$ denotes the expectation with respect to the W and $c_\theta(\cdot, \cdot)$ is the density of the copula modeling the dependence between V and W. This expectation generally does not have a closed-form expression, but it can be approximated by Monte Carlo simulation. If $W = |U|$ with $U \sim \mathcal{N}(0, \sigma_w^2)$, then $W_0 = W/\sigma_w = |U_0|$ with $U_0 = U/\sigma_w \sim \mathcal{N}(0, 1)$, i.e., W_0 has a standard half-normal distribution. We can then approximate $f_\varepsilon(\varepsilon)$ by

$$f_\varepsilon(\varepsilon) \approx \frac{1}{N} \sum_{r=1}^{N} f_V(\sigma_w w_{0,r} + \varepsilon) c(F_W(\sigma_w w_{0,r}), F_V(\sigma_w w_{0,r} + \varepsilon); \theta), \tag{16}$$

where $w_{0,r}$, $r = 1, \ldots, N$ is a sequence of N random draws from the standard half-normal distribution. The technical efficiency can be measured by the conditional expectation

$$TE = \mathbb{E}_W \left[\exp(-W) | \varepsilon \right] \tag{17a}$$

$$= \frac{1}{f_\varepsilon(\varepsilon)} \int_0^{+\infty} \exp(-w) f(w, \varepsilon) \mathrm{d}w \tag{17b}$$

$$= \frac{\mathbb{E}_W[\exp(-W) f_V(W + \varepsilon) c_\theta(F_W(W), F_V(W + \varepsilon))]}{E_W[f_V(W + \varepsilon) c_\theta(F_W(W), F_V(W + \varepsilon))]}. \tag{17c}$$

The numerator and the denominator in (17) can be approximated by Monte Carlo simulation, in the same way as above.

3.2 Double-Copula stochastic frontier model with sample selection

The idea of the double-copula stochastic frontier model with sample selection proposed in this paper is to combine the copula-based sample selection model described in Section 2.2, with the copula-based stochastic frontier model recalled in Section 3.1. We thus use two copula functions: one to model the dependence between the errors in the sample selection and production equation, and another one to model the dependence between the two components of the error in the stochastic frontier model. This general model will allow us to fit a wide range of production data, without relying on restricted (and often unsupported) assumptions about the dependence between various random variables in the model. Our approach will be based on maximum simulated likelihood estimation of all model parameters, including those of the two copula functions, for different copula families. The best model will then be selected using the AIC.

Let us first describe the model precisely, and give the expression of the likelihood function. The model is defined by the selection equation (6) and the stochastic frontier equation (11). The joint distribution of the error terms ξ and ε is

$$F_{\xi,\varepsilon}(\xi,\varepsilon) = C_\theta[\Phi(\xi), F_\varepsilon(\varepsilon)], \tag{18}$$

where Φ is the normal cdf, F_ε is the cdf of ε and C_θ is a copula function in a family $\{C_\theta : \theta \in \Theta\}$. The joint distribution of V and W in (11) is

$$F_{V,W}(v,w) = C'_{\theta'}\left\{\Phi\left(\frac{v}{\sigma_v}\right), F_W(w;\sigma_w)\right\}, \tag{19}$$

where $F_W(\cdot;\sigma_w)$ is the cdf of the half-normal distribution with scale parameter σ_w and $C'_{\theta'}$ is a copula function in a family $\{C'_{\theta'} : \theta' \in \Theta'\}$. The vector of parameters in the model is thus $\boldsymbol{\omega} = (\alpha, \beta, \sigma_v, \sigma_w, \theta, \theta')$.

From (9) and (10), the likelihood function is

$$L(\boldsymbol{\omega}) = \left(\prod_{\{i:s_i=0\}} F_\xi(-\alpha^T z_i)\right) \times \prod_{\{i:s_i=1\}} \frac{1}{1-F_\xi(-\alpha^T z_i)} f_\varepsilon(\varepsilon_i)\left[1 - \left.\frac{\partial C_\theta(F_\xi(-\alpha^T z_i), v)}{\partial v}\right|_{v=F_\varepsilon(\varepsilon)}\right], \tag{20}$$

where the density f_ε given by

$$f(\varepsilon) = \mathbb{E}_W \left[f_V(W+\varepsilon) c'_{\theta'}(F_W(W), F_V(W+\varepsilon)) \right] \tag{21}$$

can be approximated by Monte Carlo simulation using (16). The log-likelihood function is

$$\log L(\boldsymbol{\omega}) = \sum_{i=1}^{n} (1-s_i) \log F_\xi(-\alpha^T z_i) - s_i \log(1 - F_\xi(-\alpha^T z_i)) + s_i \left(\log(f_\varepsilon(\varepsilon_i)) + \log \left[1 - \left. \frac{\partial C_\theta(F_\xi(-\alpha^T z_i), v)}{\partial v} \right|_{v=F_\varepsilon(\varepsilon)} \right] \right). \tag{22}$$

4 Experimental results

4.1 Simulation study

To check the feasibility of estimating the parameters in the copula-based model introduced in the Section 3.2, we first performed a simulation experiment. The dependences between the error pairs (V, W) and (ε, ξ) were successively assumed to be positive-positive, positive-negative, negative-positive, and negative-negative. Thirty datasets of size $n = 200$ were generated from the model described in Section 3.2, with the following parameters: $\sigma_v = 1$, $\sigma_w = 4$, $\beta = (10, 0.7)^T$, $\alpha = (0.2, 0.5)^T$. The copulas C_θ and $C'_{\theta'}$ were chosen, respectively, in the Clayton and Frank families. In each case, the Kendall's tau coefficient was fixed at ± 0.7, which corresponds to $\theta = \pm 4.67$ and $\theta' = \pm 11.41$. To obtain a negative correlation between ε and ξ we used a Clayton copula rotated by 90 degrees.

For each of the four dependence patterns, we generated 30 datasets and we estimated the model parameters; we then computed the means and standard deviations of the estimates. To implement the simulated maximum likelihood method, we fixed $N = 500$ and we maximized the simulated log-likelihood using the BFGS algorithm in the R statistical software, using starting values obtained from the `glm` and `sfa` functions in the R packages `statistics` and `frontier`, respectively. The initial values for the copula parameters were $\theta'_0 = \pm 11.41$ and $\theta_0 = \pm 4.67$.

Figure 2 shows 2-D plots of V vs W and ξ vs. ε for the four different patterns of positive or negative dependence. The first column (Figures 2(a), 2(c) and 2(e)) corresponds to a positive dependence between V and W, while the second column (Figures 2(b), 2(d) and 2(f)) corresponds to negative dependence. The second and third rows correspond, respectively, to positive

and negative dependence between ξ and ε. For instance, Figure 2(d) shows the graph of ξ vs. ϵ in the case of negative correlation between V and W, and positive correlation between ξ and ε. The Frank copula is characterized by weak and symmetric tail dependences, and a strong correlation in the center of the distribution. In contrast, the Clayton copula models asymmetric dependence, with higher correlation in the lower tail.

Tables 2 and 3 show point estimates and approximate 95% confidence intervals (mean plus or minus two standard deviations) for the model parameters, in the case of positive and negative dependence between ε and ξ, respectively. In each case, we present the results for two models: the correct one, based on Frank and Clayton copulas, and a model that wrongly assumes independence between V and W. Somewhat unexpectedly, the impact of model misspecification is rather limited in terms of estimation error: the parameter estimates obtained from the two models are similar.

We do, however, observe large discrepancies between the estimates of technical efficiency. This is shown in Figure 3, which displays the distributions of technical efficiency scores computed using (17), for the correct and misspecified models, using one arbitrary dataset. We also show in this figure the technical efficiencies computed using the true parameter values and true errors ε_i. We can see that the misspecified model overestimates (respectively, underestimates) the technical efficiency scores in the case of positive (respectively, negative) dependence between V and W. This observation is consistent with previous findings by Smith [14] and Wiboonpongse et al. [16]. It confirms the importance of accounting for possible correlation in the error terms of the stochastic frontier model.

4.2 Application to rice production data

The double-copula model described in Section 3.2 was applied to data collected from 200 rice farmers from Kamphaeng Phet province, Thailand, during the year 2012. A random sampling procedure was employed. Output and input data were collected from face to face interviews performed by graduate students from Chiang Mai University, Chiang Mai, Thailand. The Kamphaeng Phet province is one of the most important rice production areas in the lower part of Northern Thailand. It contains an area of 1,436,934 rai (2299 km^2) of rice plantation, which constitutes 3% of the rice production area in Thailand. Also, the Ping river with the Bhumibol dam provides convenient conditions for planting rice.

We divided the farmers into two groups depending on the farmers' planting techniques. Transplanting is the traditional method, while the other

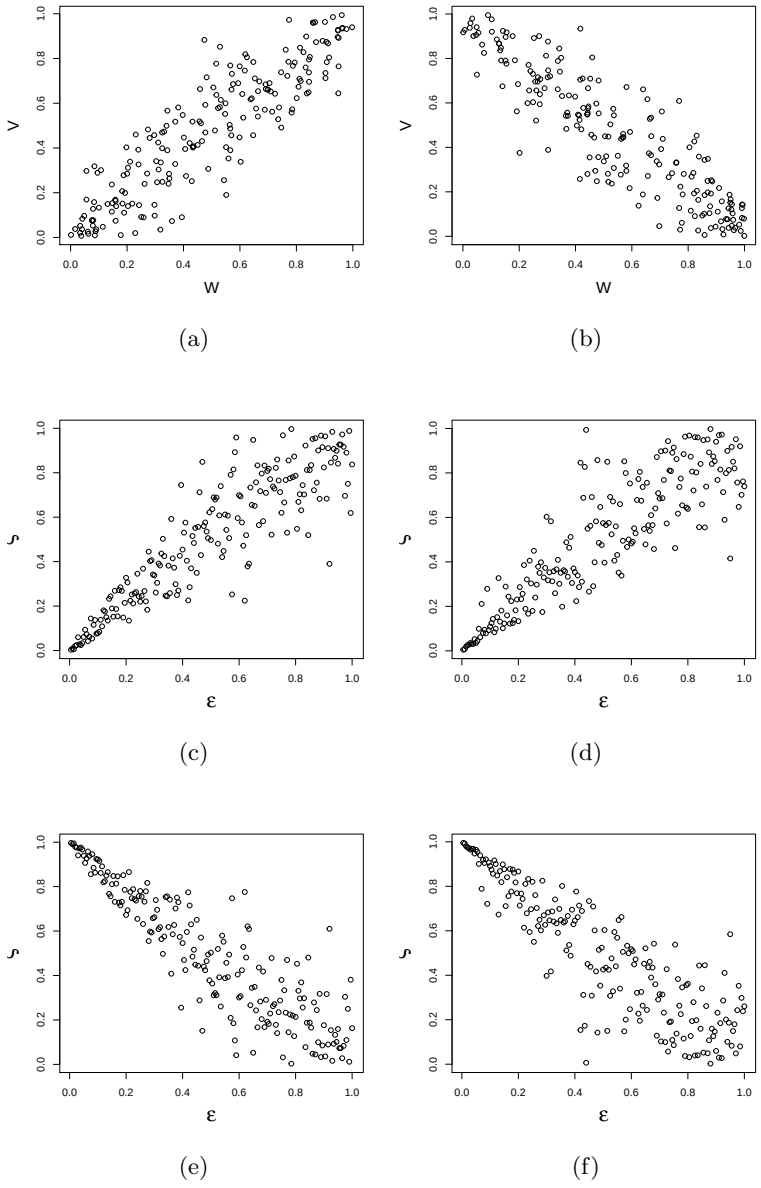

Figure 2: Simulated data based on Frank and (rotated) Clayton copulas. The first and second columns correspond to positive and negative dependence between V and W, respectively; the second and third rows correspond to positive and negative dependence between ξ and ε, respectively.

Table 2: Parameter estimates and approximate 95% confidence intervals (means plus or minus two standard deviations) for the simulated data, using the two models (Independent-Clayton and Frank-Clayton) in the case of positive correlation between ξ and ε ($\tau(\xi, \varepsilon) = 0.7$). The last two rows are the mean and standard deviation of AIC.

		$\tau(V,W) = 0.7$		$\tau(V,W) = -0.7$	
	True	Ind-Clt	Frk-Clt	Ind-Clt	Frk-Clt
β_0	10	8.34	9.19	10.9	10.17
		[7.92, 8.76]	[7.53,10.85]	[10.58, 11.54]	[9.57, 10.77]
β_1	0.7	0.67	0.7	0.69	0.68
		[0.45, 0.89]	[0.48, 0.92]	[0.21, 1.17]	[0.22, 1.14]
α_0	0.2	0.53	0.53	0.23	0.24
		[0.33, 0.73]	[0.33, 0.73]	[0.11,0.35]	[0.12,0.36]
α_1	0.5	0.58	0.58	0.49	0.49
σ_w	4	3.07	4.04	2.96	2.07
		[2.11, 4.03]	[1.94, 6.14]	[2.04, 3.88]	[1.23, 3.3]
σ_v	1	0.42	0.91	1.06	1.04
		[0.06, 0.78]	[0, 1.85]	[0.5, 1.62]	[0.62, 1.46]
		[0.52, 0.64]	[0.50,0.66]	[0.37, 0.61]	[0.37, 0.61]
θ'	$\pm$11.4		9.34		-7.55
			[-1.78, 20.46]		[-15.49, 0.39]
θ	4.67	4.63	4.51	4.89	5.28
		[1.61, 7.65]	[1.61, 7.41]	[2.17, 7.61]	[0.36, 10.2]
$\overline{\log L}$		-288.01	-286.1	-334.21	-333.12
$\overline{AIC}$		590.03	588.2	682.42	682.24
s.d.		27.45	26.87	30.55	29.98

Table 3: Parameter estimates and approximate 95% confidence intervals (means plus or minus two standard deviations) for the simulated data, using the two models (Independent-Clayton and Frank-Clayton) in the case of negative correlation between ξ and ε. ($\tau(\xi, \varepsilon) = -0.7$). The last two rows are the mean and standard deviation of AIC.

		$\tau(V,W) = 0.7$		$\tau(V,W) = -0.7$	
	True	Ind-Clt	Frk-Clt	Ind-Clt	Frk-Clt
β_0	10	8.12	9.2	7.95	7.04
		[7.74, 8.5]	[7.58, 10.82]	[6.43, 9.47]	[4.04,10.04]
β_1	0.7	0.74	0.74	0.64	0.7
		[0.46, 1.02]	[0.5, 0.98]	[0, 1.28]	[0.04, 1.36]
α_0	0.2	0.17	0.17	0.19	0.19
		[-0.01, 0.35]	[-0.01, 0.35]	[0.07, 0.31]	[0.07, 0.31]
α_1	0.5	0.5	0.5	0.46	0.46
		[0.28, 0.72]	[0.28, 0.72]	[0.36, 0.56]	[0.36, 0.56]
σ_w	4	2.8	4.05	3.69	2.6
		[2.02, 3.58]	[1.91, 6.19]	[1.93, 5.45]	[0, 5.82]
σ_v	1	0.54	1.11	1.9	2.17
		[0.24, 0.84]	[0, 2.37]	[0.92, 2.88]	[0.27, 4.07]
θ'	$\pm$11.4		7.94		-11.31
			[-1.2, 17.08]		[-22.91, 0.29]
θ	-4.67	-4.24	-4.39	-5.34	-5.48
		[-5.88, -2.6]	[-5.99, -2.79]	[-7.36, -3.32]	[-7.22, -3.74]
$\overline{\log L}$		-305.66	-303.52	-375.45	-374.31
$\overline{AIC}$		625.32	623.04	764.91	764.62
s.d.		33.82	33.11	26.51	26.21

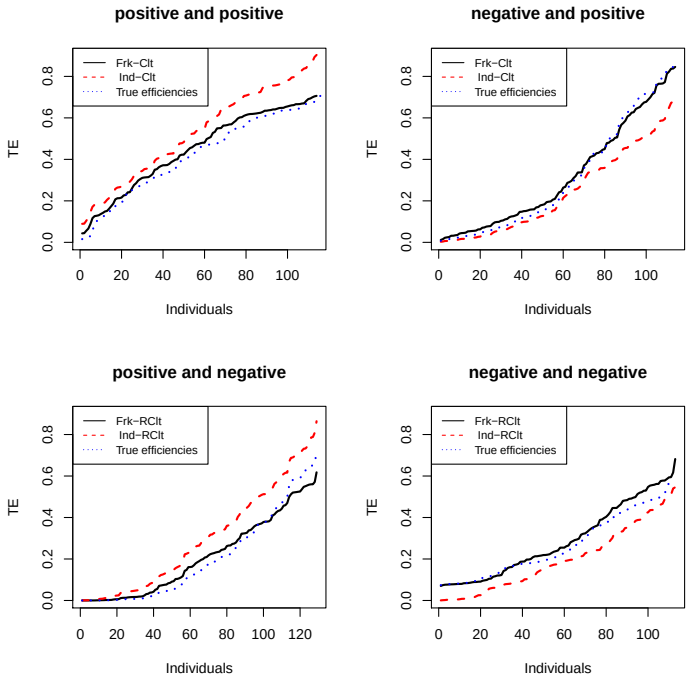

Figure 3: Estimated technical efficiencies for the correct model (solid lines), the misspecified model (interrupted line) and the corrected model with the true parameters (dotted line). The first and second columns correspond to positive and negative correlation between V and W, respectively. The first and second rows correspond, respectively, to positive and negative correlation between ξ and ε.

planting techniques, such as manual or mechanical seeding, are more recent. The choice of a planting technique is likely to influence the farmer's production and inputs. If we do not consider self-selection in the estimation of separate production frontiers, coefficient estimates will be biased. Following Heckman [7], we estimated the probability for the farmers to choose the transplanting technique by means of a bivariate Probit model. Also, a stochastic production frontier model was used to analyze the productive efficiency of rice farmers using the transplanting technique. The copula-based stochastic frontier model with sample selection was thus specified as

$$Y_i^* = \alpha_0 + \alpha_1 \times \text{member}_i + \alpha_2 \times \text{education}_i + \alpha_3 \times \text{hiring}_i + \xi_i \qquad (23)$$

$$\log(\text{output}_i) = \beta_0 + \beta_1 \times \log(\text{labor})_i + \beta_2 \times \log(\text{land})_i + \beta_3 \times \log(\text{input})_i + \beta_4 \times \text{machine}_i + V_i - W_i. \qquad (24)$$

In (23), Y_i^* is the unobserved propensity to choose transplanting relatively to the other planting techniques, which is a function of the number of family members, the years of education, and the share of hired labor. In (24), the production inputs are: labor (person days), land area (rai), material inputs (which includes inorganic fertilizers, pesticides, and seeds) (baht), and mechanical power (baht). Figure 4 shows scatter plots of the output and the independent variables. There clearly exist positive correlations between the output and independent variables.

To select the best model, i.e., the best pair $(C_\theta, C'_{\theta'})$ of copulas in (18) and (19), we first estimated the sign of the correlation between the two pairs (V, W) and (ε, ξ). For that purpose, we used the Gaussian copula family. We found a negative correlation between V and W, and a positive correlation between ξ and ε. We thus considered Gaussian, Clayton, Gumbel, Frank, Joe, 180-degree rotated Clayton, 180-degree rotated Gumbel, and 180-degree rotated Joe copulas to identify the dependence between ε and ξ. The dependence between V and W was modeled by independent, Gaussian, Frank, 90 and 270-degree rotated Clayton, 90 and 270-degree rotated Gumbel, and 90 and 270-degree rotated Joe copulas. Each model was fitted to the data, and we computed the AIC to select the best model.

Table 4 shows the values of AIC for all the considered models. We can see that the best model is the one based on a Gaussian copula to model the dependence between ξ and ε, and a Clayton copula rotated by 270 degrees to represent the dependence between V and W. The AIC for this model (-83.84) is much smaller that that for the Gaussian-Independent model

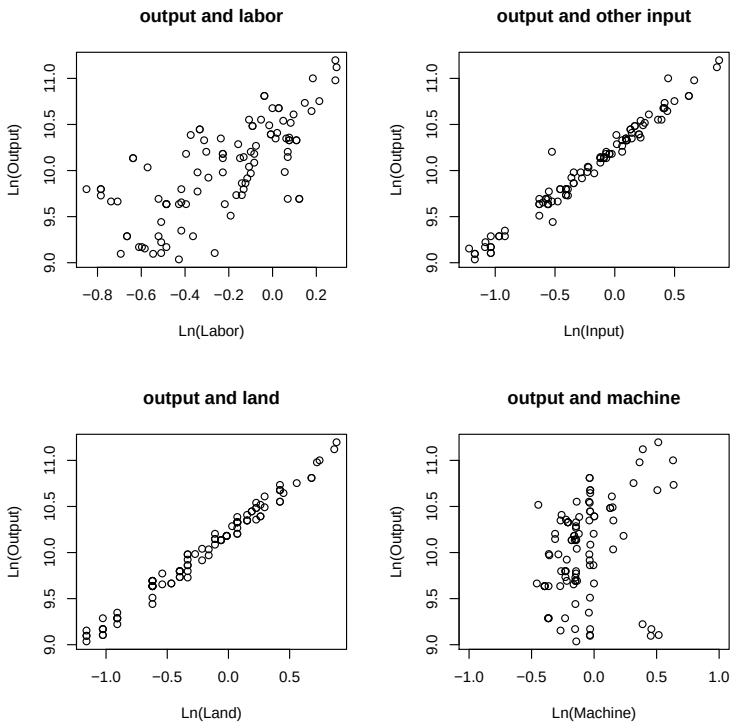

Figure 4: Matrix plot of the rice production data.

(−53.80). This result confirms the relevance of relaxing the assumption of independence between the inefficiency and random error components of the stochastic frontier model.

Table 5 shows the parameter estimates for the Gaussian/Independent and Gaussian/270-degree rotated Clayton models. The standard errors tend to be smaller for the model that considers the correlation between V and W; as a result, the significance level of coefficients in the sample selection and stochastic frontier equations tend to be higher. Overall, however, the estimated parameters in the two models are quite similar. In the sample selection equation, the parameter estimates for the number of family members and the years of education are significant at the 10% level: the probability of choosing the transplanting technique is higher when the family is larger and the education level is smaller. The estimate of σ_w for the best model is smaller than that of the alternative model: the assumption of independence between the two component errors tends to result in overestimation of the variability of technical efficiencies. We can see that the estimated tau coefficient between V and W (-0.975) indicates a very high negative correlation between the two components of the stochastic frontier equation. Assuming independence between V and W thus seems to be very inaccurate here. There is also a highly significant positive correlation between ξ and ε.

Figure 5 shows the distribution of technical efficiency scores for the best model, and the model assuming independence between V and W. We can see that the latter model underestimates the technical efficiency, which confirms the phenomenon observed with simulated data in Section 4.1. According to the best model, all the farmers are operating at an efficiency level of 96%, whereas only 46% of the farmers have an efficiency equal to or above that level according to the alternative model. The uniformly high efficiency observed with our model may be explained by the fact that transplanting, being the traditional technique, is well mastered by the farmers. It is clear that mistakenly assuming independence between the inefficiency and noise terms in the stochastic frontier equation would result in wrong conclusions and prescription for this case study.

5 Conclusions

In this paper, we have proposed a very general stochastic frontier model with self-selection based on two copulas. This model is very flexible, in that it does not assume any specific form for the dependence between the error terms of the stochastic frontier equation on the one hand, and between the

Table 4: Values of AIC for all copula-based models. The rows correspond to the copula between ξ and ε. The columns correspond to copulas between V and W. The best value is marked in bold.

$(\xi, \epsilon) \setminus (V, W)$	Ind	Gau	Frank	R-Clt 90°	R-Clt 270°	R-Gum 90°	R-Gum 270°	R-Joe 90°	R-Joe 270°
Gaussian	-53.80	-54.04	-60.49	-54.38	**-83.84**	-60.15	-69.58	-58.29	-54.49
Frank	-52.25	-57.33	-57.33	-57.39	-70.91	-78.30	-76.94	-57.04	-57.38
Gumbel	-52.85	-52.76	-53.69	-47.58	-50.52	-63.55	-55.43	-54.45	-60.33
Clayton	-54.41	-58.16	-55.71	-56.78	-54.37	-54.46	-80.11	-60.68	-55.73
Joe	-59.63	-60.61	-55.27	-54.23	-59.74	-59.64	-57.81	-55.92	-60.07
R-Clayton 180°	-49.59	-57.92	-51.75	-49.59	-56.96	-75.42	-58.63	-56.23	-57.74
R-Gumbel 180°	-57.37	-58.38	-56.72	-56.19	-56.32	-55.99	-63.81	-55.39	-57.10
R-Joe 180°	-64.02	-62.39	-64.99	-64.37	-63.26	-64.69	-63.01	-63.95	-64.98

Table 5: Parameter estimates of the Gaussian/Independent and Gaussian/270°-rotated Clayton models. The standard errors are shown in parentheses. Significance codes: *=0.1, **=0.05, ***=0.01.

	Gauss/indep		Gaussian/270°-Clt	
Stochastic frontier equation				
intercept	10.30***	(0.028)	10.17***	(0.0011)
log(labor)	0.0911**	(0.028)	0.138***	(0.0007)
log(land)	0.781***	(0.069)	0.783***	(0.0011)
log(input)	0.187***	(0.073)	0.173***	(0.0012)
log(machine)	0.023*	(0.017)	0.0162***	(0.0003)
Sample selection equation				
intercept	0.419	(0.40)	0.207	(0.359)
members	0.0422	(0.034)	0.051*	(0.0304)
education	−0.226*	(0.099)	−0.188*	(0.0943)
hiring	0.131	(0.426)	0.293	(0.388)
SD and dependence parameters				
σ_w	0.0747	(0.016)	0.0513	(0.0003)
σ_v	0.0369	(0.0092)	0.0616	(0.0004)
θ'			−78.26	(2.25)
θ	0.808	(0.069)	0.892	(0.039)
$\tau(V, W)$			−0.975***	(0.0008)
$\tau(\varepsilon, \xi)$	0.599***	(0.067)	0.701***	(0.047)

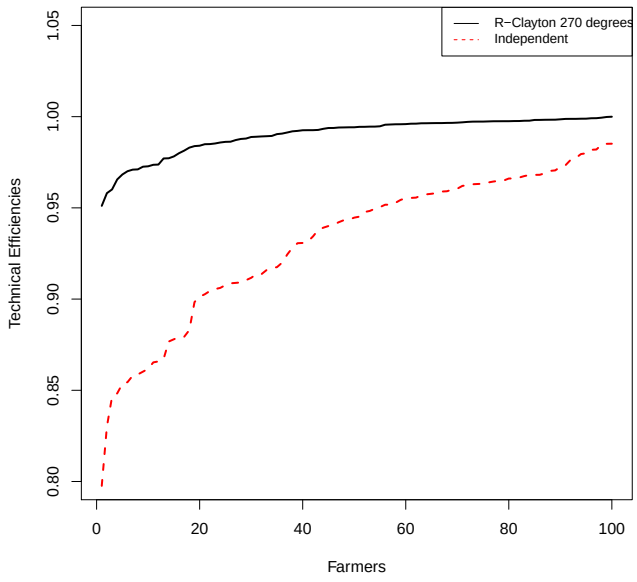

Figure 5: Technical efficiencies for the Gaussian/Independent (interrupted line) and Gaussian/270°-rotated Clayton (solid line) models.

errors of the sample selection and stochastic function equations on the other hand. Rather, our approach considers several families of copula functions to model these dependences. The maximum simulated likelihood method is used to fit each model, and the best model is selected using the AIC.

Using both simulated data and cross-sectional rice production data, we have shown that improperly assuming independence between the two components of the error term in the stochastic frontier model may result in biased estimates of technical efficiency scores, hence potentially leading to wrong conclusions and recommendations. In particular, for the rice production data, we found a strong negative correlation between the noise and inefficiency terms in the stochastic frontier equation, resulting in severe underestimation of the technical efficiencies if this correlation is ignored. This result confirms previous findings by Smith [14] and Wiboonpongse [16], among others; however, our approach goes beyond that of other studies, because we also accounts for sample self-selection in the model. In view of these findings, a lesson to be learnt from this study is that great caution should be exercised when interpreting the technical efficiencies in stochastic frontier analysis, because of the sensitivity of the approach to model misspecification. Our methodology could be extended by also considering different models for the marginal error distributions, and by using multivariate copulas.

Acknowledgements

This work has been supported by the Thailand Research Fund (TRF) for the Faculties of Economics and Agriculture at Chiang Mai University, as well as the Puey Ungphakorn Centre of Excellence in Econometrics at Chiang Mai University.

References

[1] D. Aigner, K. Lovell, and P. Schmidt, Formulation and Estimation of Stochastic Frontier Production Function Models, *Journal of Econometrics* 6 (1) (1977) 21-37.

[2] S. Bushway, B.D. Johnson, and L.A. Slocum, Is the Magic Still There? The Relevance of the Heckman Two-Step Correction for Selection Bias in Criminology, *Journal of Quantitative Criminology* 23 (2) (2007)151-178.

[3] 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.

[4] M.G. Flores, B.B. Ureta, D. Solís, P. Winters, The impact of high value markets on smallholder productivity in the Ecuadorean Sierra: A Stochastic Production Frontier approach correcting for selectivity bias. *Food Policy* 44 (2014) 237-247.

[5] W.H. Greene, Simulated Likelihood Estimation of the Normal-Gamma Stochastic Frontier Function, *Journal of Productivity Analysis* 19 (2003) 179-190.

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

[7] J. Heckman, Sample selection bias as a specification error, *Econometrica* 47(1979) 153-161.

[8] G.S. Maddala, Limited-Dependent and Qualitative Variables in Economics, New York: Cambridge University Press, (1983) 257-291.

[9] G. Marra, R. Radice, Estimation of a regression spline sample selection model, *Computational Statistics and Data Analysis* 61 (2013) 158-173.

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

[11] 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.

[12] S. Rahman, M. S. Rahman. Energy productivity and efficiency of maize accounting for the choice of growing season and environmental factors: An empirical analysis from Bangladesh, *Energy* 49 (2013) 329-336.

[13] M.D. Smith, Modelling sample selection using Archimedean copulas, *Econometrics Journal* 6 (2003) 99-123.

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

[15] 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.

[16] A. Wiboonpongse, J. Liu, S. Sriboonchitta, T. Denoeux, Modeling dependence between error components of the stochastic frontier model using copula: Application to intercrop coffee production in Northern Thailand, *International Journal of Approximate Reasoning* 65(2015) 34-44.

[17] M. Wollni, B. Brümmer, Productive efficiency of specialty and conventional coffee farmers in Costa Rica: Accounting for technological heterogeneity and self-selection, *Food Policy* 37 (2012) 67-76.

[18] C.C.Wu, S.S. Liang, The economic value of range-based covariance between stock and bond returns with dynamic copulas, *Journal of Empirical Finance* 18 (4) (2011) 711-727.

Highlights

- We propose a stochastic frontier model with self-selection, based on two copulas
- The model is estimated using maximum simulated likelihood
- Several copula families are considered; the best model is selected using AIC
- The model was applied to cross-sectional rice production data
- Our model provides more reliable estimates of technical efficiency scores