A trivariate Gaussian copula stochastic frontier model with sample selection

Jianxu Liu, Songsak Sriboonchitta, Aree Wiboonpongse, Thierry Denoeux

A trivariate Gaussian copula stochastic frontier model with sample selection

Jianxu Liu[a,b], Songsak Sriboonchitta[b,*], Aree Wiboonpongse[c], Thierry Denœux[d,e]

[a] *Faculty of Economics, Shandong University of Finance and Economics, Jinan, China*
[b] *Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand*
[c] *Faculty of Agriculture, Chiang Mai University, Chiang Mai, Thailand*
[d] *Université de technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, Compiègne, France*
[e] *Institut universitaire de France, Paris, France*

Abstract

We propose a new stochastic frontier model with sample selection, in which the dependencies between the sample selection mechanism, the inefficiency term and the two-sided error in the production equation are modeled by a trivariate Gaussian copula. This model is compared to Greene's original stochastic frontier model with sample selection, and to an alternative model based on two bivariate copulas. The relative performances of the three models are analyzed using simulated data and cross-sectional data about Jasmine rice production in Thailand. We show that our trivariate Gaussian copula model has the best performance among all models, and that ignoring some correlations may cause estimation bias as well as over or underestimation of technical efficiency scores.

Keywords: Production model, multivariate copula, dependence, sample selection, technical efficiency, rice production.

1. Introduction

Since a selection-corrected stochastic frontier model (SFM) was introduced by Greene [14] in 2006, this model has been widely used. One of the first applications was described by Rahman et al. [34] who analyzed production efficiency of Jasmine rice in Northern and North-Eastern Thailand. Later, Mayen et al. [25], Rahman [33], Bravo-Ureta et al. [4], Wollni and Brummer [46], González-Flores et al. [13], Santos-Montero and Bravo-Ureta [8] and others applied the selection-corrected SFM (hereafter referred to as Greene's model) to estimate the technical efficiency of farm crops. Other applications include assessing the technical efficiency of food retailers [28], labor market [2], fisheries [39], etc.

However, Greene's original model has some limitations. It assumes, without any other justification than technical convenience, the two error components of the production equation to be independent, which may result in over- or underestimation of technical efficiency [45].

*Corresponding author.

Greene [15] also questioned whether it is reasonable to assume that the heterogeneity and the inefficiency in the production model are uncorrelated. Furthermore, the model is usually fitted using a heuristic two-stage estimation method; as a result, the estimators may not be efficient. Finally, the model's distributional assumptions (bivariate normality of the sample selection and symmetric part of the production equation error terms, half-normal distribution of the inefficiency term) can be questioned.

In recent years, some scholars further developed the sample selection and production models, with the aim to overcome some limitations of the original Greene's model. For example, Smith [37] and Kruger et al. [21] proposed copula-based sample selection models to relax the multivariate normality assumption. Smith [38] and Wiboonpongse et al. [45] modeled the dependence between the two error terms of the production model using copulas and demonstrated that accounting for this dependence can improve the estimation of technical efficiency. Mehdi and Hafner [12] also found that the estimated technical efficiencies taking into account dependence through copulas tend to be lower than those under the independence assumption. Huang et al. [20] proposed a simultaneous SFM with correlated composite errors based on copula functions. Greene [16], Beckers and Hammond [3], Stevenson [41], Kumbhakar and Lovell [22], etc., proposed several probability distribution functions for the inefficiency term in SFMs. Sriboonchitta et al. [40] proposed an alternative to Greene's model using two copula functions. The double-copula SFM with sample selection relaxes the assumption of independence between the two error components in the SFM, and also accounts for nonlinear correlation between the error in the selection equation and the composite error in the production equation. However, this double-copula model neglects the correlation between the unobservables in the selection model and the random error in the SFM, in contrast to Greene's model. From this literature review, it appears that: (1) previous studies have laid the foundation for further improvement of Greene's model, and (2) the most advanced extension of Greene's model, the double copula-based model, can be perfected.

To further improve the flexibility of Greene's model, a trivariate Gaussian copula SFM with sample selection is proposed in this paper. This model generalizes Greene's model by modeling the dependence between the unobservables in the selection equation and the two error terms in the production equation using a trivariate Gaussian copula. To assess the feasibility of this approach, we perform a simulation study and compare our model to the double-copula SFM with sample selection and Greene's model. The three models are then applied to cross-sectional data about the technical efficiency of rice production in Thailand.

The remainder of this paper is organized as follows. The previous models considered in this paper are first recalled in Section 2. The new model is then introduced in Section 3, where a simulation study is also presented. Finally, the application to rice production efficiency analysis is described in Section 4, and Section 5 concludes the paper.

2. Previous models

In this section, we briefly review previous SFM's that provide the starting point of this study. The basic SFM is first recalled in Section 2.1. Two SFM's with sample selection are then summarized: the original Greene model in Section 2.2 and the double-copula SFM in Section 2.3.

2.1. Basic SFM

Stochastic frontier analysis [1] is commonly used to fit a production function and to estimate farm-level technical efficiency. The basic SFM is defined by the following equation:

$$Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \tag{1a}$$

$$\varepsilon_i = V_i - W_i, \tag{1b}$$

$i = 1, \ldots, n$, where Y_i represents the output of production unit i, $\mathbf{x}_i$ is a vector of input quantities, $\boldsymbol{\beta}$ is a vector of coefficients, and the random error term ε_i is divided into two parts: a two-sided firm-specific effect V_i (which can be positive or negative) and a positive inefficiency term W_i. The "frontier", or optimal output achievable by production unit i is $\boldsymbol{\beta}^T \mathbf{x}_i + V_i$; it is stochastic, hence the term "stochastic frontier". Typically, it is assumed that V_i and W_i have, respectively, a normal distribution $\mathcal{N}(0, \sigma_v^2)$ and a half-normal distribution with scale parameter σ_w, i.e., $W_i = \sigma_w |U_i|$ with $U_i \sim \mathcal{N}(0, 1)$. The *technical efficiency* (TE) of production unit i is defined as $\exp(-W_i)$. As W_i is not observed, TE is usually measured by its conditional expectation given ε_i, called the *TE score*:

$$TE_i = \mathbb{E}_W[\exp(-W) | \varepsilon = \varepsilon_i]. \tag{2}$$

In the classical SFM, the two error components V_i and W_i are assumed to be independent. Following [38], Wiboonpongse et al. [45] have proposed to relax this assumption and to model the dependence between error terms V and W using a parameterized family of copulas. They proposed a methodology that consists in considering several copula families and selecting the best model according to the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). They advised against the systematic use of the assumption of independence between V and W, which may lead to a gross overestimation of technical efficiency for some datasets. More recently, Wei et al. [44] investigated the use of a skew normal copula to model the asymmetric dependence between V and W.

2.2. SFM with sample selection

To address the problem of selection bias in linear regression, Heckman [19] proposed to model the process of inclusion of an observation in the sample (or "sample selection process") by an equation of the form

$$S_i = \begin{cases} 1 & \text{if } Y_i^* = \boldsymbol{\alpha}^T \mathbf{z}_i + \xi_i \geq 0 \\ 0 & \text{if } Y_i^* = \boldsymbol{\alpha}^T \mathbf{z}_i + \xi_i < 0 \end{cases}, \tag{3}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\alpha}$ is a vector of coefficients, $\mathbf{z}_i$ is a vector of exogenous variables, ξ_i is an error term assumed to have a standard normal distribution $\mathcal{N}(0, 1)$, Y_i^* is a latent variable, and S_i is a dummy variable that indicates whether the response variable is observed

($S_i = 1$) or not ($S_i = 0$). Greene [15] combined the selection equation (3) with the production equation (1) to propose a SFM with sample selection. He assumed that V_i and W_i are independent with, respectively, normal and half-normal distributions, and that the random vector (V_i, ξ_i) has a bivariate normal distribution with zero mean and variance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_v^2 & \rho\sigma_v \\ \rho\sigma_v & 1 \end{pmatrix}.$$

From [15], the conditional probability density function (pdf) for an observation in this model is

$$f(y_i \mid \mathbf{x}_i, |U_i|, \mathbf{z}_i, s_i) = s_i \left[\frac{1}{\sigma_v\sqrt{2\pi}} \exp\left(-\frac{(y_i - \boldsymbol{\beta}^T\mathbf{x}_i + \sigma_w|U_i|)^2}{2\sigma_v^2} \right) \times \right.$$
$$\left. \Phi\left(\frac{\rho(y_i - \boldsymbol{\beta}^T\mathbf{x}_i + \sigma_w|U_i|)/\sigma_\varepsilon + \boldsymbol{\alpha}^T\mathbf{z}_i}{\sqrt{1-\rho^2}} \right) \right] + (1 - s_i)\Phi(-\boldsymbol{\alpha}^T\mathbf{z}_i), \quad (4)$$

where σ_ε is the standard deviation of $\varepsilon = V - W$ and Φ is the standard normal cumulative distribution function (cdf).

To simplify the estimation problem, Greene uses a two-step estimation method. The vector $\boldsymbol{\alpha}$ of coefficients in the selection equation is first estimated by unconstrained maximum likelihood using Eq. (3) only, which defines a Probit model. In the second step, the estimate $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ is plugged in (4), and the log-likelihood is formed by integrating out $|U_i|$ (see [15] for details). This integral is intractable and is approximated by simulation. The simulated log-likelihood is finally given by:

$$\log L_S(\boldsymbol{\beta}, \sigma_w, \sigma_v, \rho) = \sum_{i=1}^{n} \log \frac{1}{M} \sum_{m=1}^{M} \left\{ s_i \left[\frac{1}{\sigma_v\sqrt{2\pi}} \exp\left(-\frac{(y_i - \boldsymbol{\beta}^T\mathbf{x}_i + \sigma_w|U_{im}|)^2}{2\sigma_v^2} \right) \times \right.\right.$$
$$\left.\left. \Phi\left(\frac{\rho(y_i - \boldsymbol{\beta}^T\mathbf{x}_i + \sigma_w|U_{im}|)/\sigma_\varepsilon + \widehat{\boldsymbol{\alpha}}^T\mathbf{z}_i}{\sqrt{1-\rho^2}} \right) \right] + (1 - s_i)\Phi(-\widehat{\boldsymbol{\alpha}}^T\mathbf{z}_i) \right\},$$

where U_{im}, $m = 1, \ldots, M$ is a sequence of M random draws from the standard normal distribution. A gradient-based optimization procedure, such as the BFGS algorithm, can be used to maximize $\log L_S$ and estimate the parameters of the model.

2.3. Double-copula SFM with sample selection

In [40], Sriboonchitta et al. proposed a more flexible SFM with sample selection, in which the dependence relations between ξ and ε on the one hand, and between V and W on the other hand, are modeled by two bivariate copulas [27]. Assuming, as before, the distributions of ξ and V to be normal, and the distribution of W to be half-normal, the joint cdf of (ξ, ε) can be written as

$$F_{\xi,\varepsilon}(\xi, \varepsilon) = C_\theta^{(1)}[\Phi(\xi), F_\varepsilon(\varepsilon)],$$

where F_ε is the cdf of ε and $C_\theta^{(1)}$ is a copula function in a family $\mathcal{C}^{(1)} = \{C_\theta^{(1)} : \theta \in \Theta\}$, and the joint cdf of (V, W) can be expressed as

$$F_{V,W}(v, w) = C_\omega^{(2)} \left[\Phi\left(\frac{v}{\sigma_v}\right), F_W(w; \sigma_w) \right],$$

where $F_W(\cdot; \sigma_w)$ is the cdf of the half-normal distribution with scale parameter σ_w and $C_\omega^{(2)}$ is a copula function in a family $\mathcal{C}^{(2)} = \{C_\omega^{(2)} : \omega \in \Omega\}$. Sriboonchitta et al. [40] proposed a methodology that consists in exploring a range of copula families for $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$, fitting the parameters by maximizing the simulated likelihood for each model, and selecting the best model according to AIC or BIC. Using simulated and real data, they showed that improperly assuming independence between the two components of the error term in the SFM may result in biased estimates of technical efficiency scores, hence potentially leading to wrong conclusions and recommendations.

We can remark that, in Greene's model summarized in Section 2.2, V and W are linked by the independence (product) copula, while ξ and V are linked by a Gaussian copula. This corresponds to the following decomposition of the joint density of (V, W, ξ):

$$f(v, w, \xi) = f(v)f(w)f(\xi|v).$$

In contrast, in a double copula model in which V and W are linked by the independence copula, the distribution of ξ depends on the difference $\varepsilon = V - W$, which corresponds to the following decomposition of the joint distribution:

$$f(v, w, \xi) = f(v)f(w)f(\xi|v, w)$$

As a consequence, Greene's model is not a special case of the double-copula model, except in the particular case where we have a fully efficient SFM characterized by the condition $W = 0$. In the following section, we introduce a new model that is, by construction, a direct generalization of Greene's model.

3. A trivariate Gaussian copula SFM with sample selection

Our main purpose in this study is to construct a flexible SFM with sample selection, in which the dependence between the three error terms W, V, and ξ is modeled by a three-dimensional copula that can be learnt from the data. Whereas many parameterized families of bivariate copulas have been proposed, the construction of multivariate copulas with dimension strictly greater than two is still an ongoing research topic [26][48]. In this work, we choose the three-dimensional Gaussian copula family for the following reasons: (1) it can be parameterized by a correlation matrix $\mathbf{R}$ with natural interpretation; (2) it allows for easy calculation of the simulated likelihood, and (3) it makes it possible to recover Greene's model as a special case. This copula family and the unconstrained parameterization of the correlation matrix are first recalled in Section 3.1. Our model is then introduced in Section 3.2, and a simulation study is reported in Section 3.3.

3.1. Trivariate Gaussian copula

A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniformly distributed [27, 42, 43, 7]. Sklar's Theorem [36] states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula that describes the dependence structure between the variables. As noted in [9], the tool of copulas is less universal in the case of m ($m \geq 3$) variables than it is in the case of two. However, an m-dimensional copula function C_H can be constructed from an m-dimensional cdf H with margins $H_1, \ldots, H_m$ as

$$C_H(u_1, \ldots, u_m) = H\left[H_1^{-1}(u_1), \ldots, H_m^{-1}(u_m)\right], \quad (u_1, \ldots, u_m) \in [0,1]^m.$$

In the case $m = 3$, choosing as H the three-dimensional Gaussian cdf $\Phi_{\mathbf{R}}$ with standard normal marginals and covariance matrix equal to the correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix},$$

we get the following trivariate Gaussian copula

$$C_{\mathbf{R}}(u_1, u_2, u_3) = \Phi_{\mathbf{R}}(q_1, q_2, q_3) \tag{5a}$$

$$= \int_{-\infty}^{q_1} \int_{-\infty}^{q_2} \int_{-\infty}^{q_3} \phi_{\mathbf{R}}(x, y, z) dx dy dz, \tag{5b}$$

where

$$\phi_{\mathbf{R}}(x, y, z) = \frac{1}{(2\pi)^{3/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}(x, y, z)\mathbf{R}^{-1}(x, y, z)^T\right) \tag{5c}$$

is the three-dimensional Gaussian pdf with zero mean and covariance matrix $\mathbf{R}$, and $q_k = \Phi^{-1}(u_k)$ for $k \in \{1, 2, 3\}$ are the normal scores. The density of this copula is [47]:

$$c_{\mathbf{R}}(u_1, u_2, u_3) = \frac{\partial^3 C_{\mathbf{R}}(u_1, u_2, u_3)}{\partial u_1 \partial u_2 \partial u_3} \tag{6a}$$

$$= \frac{1}{\phi(q_1)\phi(q_2)\phi(q_3)} \phi_{\mathbf{R}}(q_1, q_2, q_3) \tag{6b}$$

$$= \frac{1}{|\mathbf{R}|^{1/2}} \exp\left(\frac{1}{2}\mathbf{q}^T(\mathbf{I} - \mathbf{R})\mathbf{q}\right), \tag{6c}$$

where $\mathbf{q} = (q_1, q_2, q_3)^T$ is the vector of normal scores, $\mathbf{I}$ is the 3×3 identity matrix, and ϕ is the standard univariate normal pdf.

Unconstrained parameterization of $\mathbf{R}$. When maximizing the likelihood of our model with respect to $\mathbf{R}$, we will need to ensure that $\mathbf{R}$ remains nonnegative. Pinheiro and Bates [29] reviewed different parameterization of covariance matrices that ensure this property. One of those with good properties and easy interpretation is the spherical parameterization, which starts with the Cholesky decomposition:

$$\mathbf{R} = \mathbf{L}\mathbf{L}^T,$$

where $\mathbf{L}$ is a lower triangular matrix with nonnegative diagonal elements. In the three-dimensional case, $\mathbf{L}$ can be parametrized as follows [29][35]:

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ \cos\theta_{12} & \sin\theta_{12} & 0 \\ \cos\theta_{13} & \cos\theta_{23}\sin\theta_{13} & \sin\theta_{23}\sin\theta_{13}, \end{pmatrix}$$

with $(\theta_{12}, \theta_{13}, \theta_{23}) \in \mathbb{R}^3$. The correlation matrix $\mathbf{R}$ can then be expressed as a function of $(\theta_{12}, \theta_{13}, \theta_{23})$ as

$$\mathbf{R} = \begin{pmatrix} 1 & \cos\theta_{12} & \cos\theta_{13} \\ \cos\theta_{12} & 1 & \cos\theta_{12}\cos\theta_{13} + \sin\theta_{12}\cos\theta_{23}\cos\theta_{13} \\ \cos\theta_{13} & \cos\theta_{12}\cos\theta_{13} + \sin\theta_{12}\cos\theta_{23}\cos\theta_{13} & 1 \end{pmatrix}, \tag{7}$$

i.e., the correlation coefficients ρ_{ij} can be recovered as

$$\begin{aligned} \rho_{12} &= \cos\theta_{12}, \\ \rho_{13} &= \cos\theta_{13}, \\ \rho_{23} &= \cos\theta_{12}\cos\theta_{13} + \sin\theta_{12}\cos\theta_{23}\sin\theta_{13}. \end{aligned}$$

3.2. Model description and likelihood

In this section, we describe the proposed generalization of Greene's model, referred to as the Trivariate Gaussian Copula (TGC) model, in which the dependence between the three error terms ξ, V and W is modeled by a Gaussian copula with correlation matrix $\mathbf{R}$. The three parameters in this model, denoted as ρ_{vw}, $\rho_{w\xi}$ and $\rho_{v\xi}$, are correlation coefficients measuring the dependence between, respectively, the pairs (V, W), (W, ξ) and (V, ξ). Greene's model recalled in Section 2.2 is recovered as a special case where $\rho_{vw} = \rho_{w\xi} = 0$.

As shown by Smith [37], the likelihood function of the model described by (1) and (3) is

$$L(\boldsymbol{\psi}) = \prod_{\{i:s_i=0\}} P(Y_i^* \leq 0) \prod_{\{i:s_i=1\}} P(Y_i^* > 0) f(y_i | y_i^* > 0) \tag{8a}$$

$$= \left(\prod_{\{i:s_i=0\}} \Phi(-\boldsymbol{\alpha}^T \mathbf{z}_i) \right) \times \prod_{\{i:s_i=1\}} [1 - \Phi(-\boldsymbol{\alpha}^T \mathbf{z}_i)] f_\varepsilon(\varepsilon_i | S_i = 1), \tag{8b}$$

where $\boldsymbol{\psi}$ is the vector of all parameters in the model (including $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ_v, the three parameters θ_{wv}, $\theta_{w\xi}$ and $\theta_{v\xi}$ defining matrix $\mathbf{R}$ in (7), and the parameters of the distribution of W, which need not be assumed to be half-normal). The difficulty resides in the calculation of the conditional pdf $f_\varepsilon(\varepsilon|s_i=1)$. As $\varepsilon = V - W$, we need to express the joint conditional density of V and W given $S_i = 1$. As shown in [37] and [40], this conditional pdf can be written as

$$f_{V,W}(v,w|S_i=1) = \frac{1}{1-P(Y^*\leq 0)} \frac{\partial^2\left[P(V\leq v, W\leq w) - P(V\leq v, W\leq w, Y^*\leq 0)\right]}{\partial v\partial w} \tag{9a}$$

$$= \frac{1}{1-\Phi(-\boldsymbol{\alpha}^T\mathbf{z})} \frac{\partial^2\left[F_{V,W}(v,w) - H(v,w,-\boldsymbol{\alpha}^T\mathbf{z})\right]}{\partial v\partial w} \tag{9b}$$

$$= \frac{1}{1-\Phi(-\boldsymbol{\alpha}^T\mathbf{z})} \left(f_{V,W}(v,w) - \frac{\partial^2 H(v,w,-\boldsymbol{\alpha}^T\mathbf{z})}{\partial v\partial w}\right), \tag{9c}$$

where H, $f_{V,W}$ and $F_{V,W}$ are, respectively, the joint cdf of (V,W,ξ), and the joint pdf and cdf of V and W. Random variables V and W are linked by a bivariate Gaussian copula $C_{\rho_{vw}}$ with correlation ρ_{vw}. Using a formula similar to (6) for bivariate Gaussian copula, the corresponding copula density is

$$c_{\rho_{vw}}(u_1,u_2) = \frac{1}{1-\rho_{vw}} \exp\left(\frac{2\rho_{vw}q_1q_2 - \rho_{vw}^2(q_1^2+q_2^2)}{2(1-\rho_{vw}^2)}\right),$$

where $q_1 = \Phi^{-1}(u_1)$ and $q_2 = \Phi^{-1}(u_2)$. The pdf $f_{V,W}(v,w)$ can then be written as

$$f_{V,W}(v,w) = c_{\rho_{vw}}(F_V(v), F_W(w)) f_V(v) f_W(w) \tag{10a}$$

$$= c_{\rho_{vw}}[\Phi(v/\sigma_v), F_W(w)] \frac{\phi(v/\sigma_v)}{\sigma_v} f_W(w). \tag{10b}$$

Let us now compute the second derivative in (9c). The multivariate cdf H of V, W and ξ can be expressed using the trivariate Gaussian copula function $C_{\mathbf{R}}$, where correlation matrix $\mathbf{R}$ is composed of ρ_{vw}, $\rho_{w\xi}$, and $\rho_{v\xi}$, as

$$H(v,w,\xi) = C_{\mathbf{R}}[\Phi(v/\sigma_v), F_W(W), \Phi(\xi)].$$

Using the following notation:

$$C''_{\mathbf{R}}(u_1,u_2,u_3) = \frac{\partial^2 C_{\mathbf{R}}(u_1,u_2,u_3)}{\partial u_1 \partial u_2}$$

for the partial derivative of $C_{\mathbf{R}}$ with respect to its first two arguments, we can express the second partial derivatives of H with respect to v and w as

$$\frac{\partial^2 H(v,w,-\boldsymbol{\alpha}^T\mathbf{z})}{\partial v\partial w} = C''_{\mathbf{R}}\left(\Phi(v/\sigma_v), F_W(w), \Phi(-\boldsymbol{\alpha}^T\mathbf{z})\right) \frac{\phi(v/\sigma_v)}{\sigma_v} f_W(w). \tag{11}$$

The expression of function $C''_{\mathbf{R}}$ is derived in Appendix A.

Given that $\varepsilon = V - W$, we can replace v by $\varepsilon + w$ in (10)-(11) and obtain the conditional pdf of (ε, W) as

$$f_{\varepsilon,W}(\varepsilon, w|S_i = 1) = \frac{1}{1-\Phi(-\boldsymbol{\alpha}^T\mathbf{z})}\left\{c_{\rho_{vw}}\left[\Phi\left(\frac{\varepsilon+w}{\sigma_v}\right), F_W(w)\right] - C''_{\mathbf{R}}\left(\Phi\left(\frac{\varepsilon+w}{\sigma_v}\right), F_W(w), \Phi(-\boldsymbol{\alpha}^T\mathbf{z})\right)\right\}\frac{\phi\left(\frac{\varepsilon+w}{\sigma_v}\right)}{\sigma_v}f_W(w).$$

Marginalizing out W, we get the conditional pdf of ε as

$$f_\varepsilon(\varepsilon|S_i = 1) = \int_0^{+\infty} f_{W,\varepsilon}(\varepsilon, w|S_i = 1)dw,$$

which can be expressed as

$$f_\varepsilon(\varepsilon|S_i = 1) = \frac{1}{1-\Phi(-\boldsymbol{\alpha}^T\mathbf{z})}\mathbb{E}_W\left[\left\{c_{\rho_{vw}}\left[\Phi\left(\frac{\varepsilon+W}{\sigma_v}\right), F_W(W)\right] - C''_{\mathbf{R}}\left(\Phi\left(\frac{\varepsilon+W}{\sigma_v}\right), F_W(W), \Phi(-\boldsymbol{\alpha}^T\mathbf{z})\right)\right\}\frac{\phi\left(\frac{\varepsilon+W}{\sigma_v}\right)}{\sigma_v}\right], \quad (12)$$

where $\mathbb{E}_W[\cdot]$ denotes expectation with respect to W. The expectation in (12) can be approximated by Monte Carlo simulation or a quasi-random low-discrepancy sequences such as a Halton sequence [18], which is known to yield better results than a uniform random number generator [17, page 625]. The conditional pdf $f_\varepsilon(\varepsilon|S_i = 1)$ can, thus, be approximated as follows:

$$f_\varepsilon(\varepsilon|S_i = 1) \approx \frac{1}{1-\Phi(-\boldsymbol{\alpha}^T\mathbf{z})}\frac{1}{M}\sum_{m=1}^{M}\left[\left\{c_{\rho_{vw}}\left[\Phi\left(\frac{\varepsilon+F_W^{-1}(q_m)}{\sigma_v}\right), q_m\right] - C''_{\mathbf{R}}\left(\Phi\left(\frac{\varepsilon+F_W^{-1}(q_m)}{\sigma_v}\right), q_m, \Phi(-\boldsymbol{\alpha}^T\mathbf{z})\right)\right\}\frac{\phi\left(\frac{\varepsilon+F_W^{-1}(q_m)}{\sigma_v}\right)}{\sigma_v}\right],$$

where q_m, $m = 1, \ldots, M$ is a Halton sequence of length M. Plugging this approximation into the expression of the likelihood (8), we get the simulated likelihood:

$$L_S(\boldsymbol{\psi}) = \left(\prod_{\{i:s_i=0\}}\Phi(-\boldsymbol{\alpha}^T\mathbf{z}_i)\right) \times \prod_{\{i:s_i=1\}}\frac{1}{M}\sum_{m=1}^{M}\left[\left\{c_{\rho_{vw}}\left[\Phi\left(\frac{\varepsilon_i+F_W^{-1}(q_{i,m})}{\sigma_v}\right), q_{i,m}\right] - C''_{\mathbf{R}}\left(\Phi\left(\frac{\varepsilon_i+F_W^{-1}(q_{i,m})}{\sigma_v}\right), q_{i,m}, \Phi(-\boldsymbol{\alpha}^T\mathbf{z}_i)\right)\right\}\frac{\phi\left(\frac{\varepsilon_i+F_W^{-1}(q_{i,m})}{\sigma_v}\right)}{\sigma_v}\right],$$

where $(q_{i,m})$ for $m = 1, \ldots, M$ and is a Halton sequence for observation i. This function can be maximized using an iterative optimization algorithm.

Estimation of TE scores. After all parameter estimates have been obtained, TE scores can be calculated as well (see [38] and [40]). From (2), we have

$$TE = \frac{1}{f_\varepsilon(\varepsilon)} \int_0^{+\infty} \exp(-w) f_{\varepsilon,W}(\varepsilon, w) \mathrm{d}w \tag{13}$$

From (10), we have

$$f_{\varepsilon,W}(\varepsilon, w) = c_{\rho_{vw}}[\Phi\left((w+\varepsilon)/\sigma_v\right), F_W(w)] \frac{\phi((w+\varepsilon)/\sigma_v)}{\sigma_v} f_W(w).$$

Hence,

$$\begin{aligned} f_\varepsilon(\varepsilon) &= \int_0^{+\infty} c_{\rho_{vw}}[\Phi\left((w+\varepsilon)/\sigma_v\right), F_W(w)] \frac{\phi((w+\varepsilon)/\sigma_v)}{\sigma_v} f_W(w) dw && \text{(14a)} \\ &= \mathbb{E}_W \left(c_{\rho_{vw}}[\Phi\left((W+\varepsilon)/\sigma_v\right), F_W(W)] \frac{\phi((W+\varepsilon)/\sigma_v)}{\sigma_v} \right). && \text{(14b)} \end{aligned}$$

Now, let A denote the integral on the right-hand side of (13); it can be written as

$$\begin{aligned} A &= \int_0^{+\infty} \exp(-w) c_{\rho_{vw}}[\Phi\left((w+\varepsilon)/\sigma_v\right), F_W(w)] \frac{\phi((w+\varepsilon)/\sigma_v)}{\sigma_v} f_W(w) dw && \text{(15a)} \\ &= \mathbb{E}_W \left(\exp(-W) c_{\rho_{vw}}[\Phi\left((W+\varepsilon)/\sigma_v\right), F_W(W)] \frac{\phi((W+\varepsilon)/\sigma_v)}{\sigma_v} \right). && \text{(15b)} \end{aligned}$$

The technical efficients TE_i for each observation i can be estimated by plugging the maximum-likelihood estimates of the parameters in (14b) and (15b), and approximating the expectations using Halton sequences as before.

Comparison with the double-copula model. We can remark that the TGC model introduced in this section and the double-copula model recalled in Section 2.3 rely on different assumptions about the joint distribution of V, W and ξ. The TGC model does not make any independence assumption, so it corresponds to the following general decomposition of the joint pdf of (V, W, ξ):

$$f(v, w, \xi) = f(v) f(w|v) f(\xi|v, w).$$

The double-copula model corresponds to a similar decomposition but it further assumes that $f(\xi|v, w) = f(\xi|v - w)$, i.e., given $V = v$ and $W = w$, the distribution of ξ depends only on the difference $\varepsilon = v - w$. For this reason, the double-copula model with two Gaussian copulas and the TGC model are not nested. We can remark that the former model has two correlation parameters ρ_{vw} and $\rho_{\xi\varepsilon}$, whereas the latter has three: ρ_{vw}, $\rho_{w\xi}$ and $\rho_{v\xi}$. As a consequence, the TGC model is slightly more flexible.

3.3. Simulation study

To demonstrate the feasibility of estimation procedure described in the previous section, and to study the impact of model misspecification, we randomly generated 100 datasets of size $n = 500$ and 100 datasets of size $n = 2000$ from the TGC model described in Section 3.2, with the following parameter values: $\beta = 2$, $\alpha = 1$, $\sigma_v = 0.2$, $\rho_{v,w} = 0.5$, $\rho_{w,\xi} = 0.4$, $\rho_{v,\xi} = 0.2$. The inefficiency W was assumed to have a half-normal distribution with scale parameter $\sigma_w = 0.7$.

We fitted four models to each dataset: the correct TGC model, Greene's model (assuming independence between V and W), and two double-copula models described in Section 2.3: the Double Gaussian copula (DGC) model, and the Gaussian-Clayton copula (GCC) model representing the dependence between V and W by a Gaussian copula and the dependence between ξ and ε by a Clayton copula. To implement the simulated maximum likelihood method, we generated a Halton sequence of size $M = 200$ and we maximized the simulated log-likelihood using the R implementation of the Nelder-Mead algorithm [32]. The starting value of α was obtained by logistic regression using the R function `glm`, and parameters β, σ_v and σ_w were estimated using function `sfa` in the R package `frontier` [6] by neglecting the sample selection process as well as the correlation between V and W.

Tables 1 and 2 report, respectively, the bias and standard errors of the estimators for the four models, and the mean-square errors (MSE's). Figure 1 displays the histograms of parameter estimates when postulating the correct TGC model, with a normal fit (solid line) together with the 2.5% and 97.5% quantiles shown as dotted vertical lines. As shown in Table 2, the TGC model, which is correctly specified, has the lowest MSE's for all parameters except $\rho_{w,v}$, for which the double-copula models have a lower MSE. Looking at Table 1, we can see that the estimates of $\rho_{w,v}$ in the double-copula models have higher bias, but lower variance as compared to TGC, which is due to the fact that the DGC and GCC models are misspecified, but have fewer parameters that TGC. Somewhat surprisingly, parameters β and, to a lesser extent, α are well estimated by all models, which is not true for the variance and correlation parameters. In particular, Greene's model, which does not represent the dependence between V and W, severely underestimates the scale parameters σ_v and σ_w and gets the correlation coefficient $\rho_{v,\xi}$ completely wrong. As they do not make the wrong assumption of independence between V and W, the two double-copula models do a better job at estimating σ_v and $\rho_{v,w}$, but they overestimate σ_w.

Poor estimation of σ_v, σ_w and $\rho_{v,w}$ by Green's model and, to a lesser extent, by the two double-copula models can be expected to have an impact on the estimation of TE scores. To verify this assumption, we computed, for each dataset, the RMSE's between the true TE scores and their estimates obtained by each of the four models. As shown in Figure 2, Greene's model performs comparatively poorly in terms of TE score estimation, which is due to the wrong assumption of independence between V and W. In contrast, the two double-copula models yield almost as good estimates of TE scores as does the TGC model, which confirms the good performance of these models already reported in [40].

Tables 1-2 and Figure 2 show that the double-copula models fit the TGC-generated data quite well, which suggests that the TGE and double-copula models are actually quite close. To verify this assumption, we fitted the TGC and DGC models on 100 datasets generated

Table 1: Estimated biases and standard errors for the four models. The smallest bias is shown in bold.

$n = 500$	TGC		DGC		GCC		Greene	
	bias	se	bias	se	bias	se	bias	se
α	0.0150	0.1530	0.0401	0.1590	0.0258	0.1520	**-0.0124**	0.1540
β	**5.91e-04**	8.21e-03	6.26e-04	8.24e-03	9.55e-03	8.18e-03	1.39e-03	9.98e-03
σ_w	**0.0115**	0.0318	0.0368	0.0278	0.0366	0.0277	-0.0266	0.0289
σ_v	**3.16e-03**	0.0379	0.0139	0.0370	0.0155	0.0380	-0.0755	0.0264
$\rho_{w,v}$	**-0.0125**	0.0970	0.0174	0.0677	0.0180	0.0684	–	–
$\rho_{w,\xi}$	-0.0969	0.3050	–	–	–	–	–	–
$\rho_{v,\xi}$	**-0.1380**	0.5460	–	–	–	–	-1.01	0.1680
$n = 2000$								
α	**0.0155**	0.0800	0.0320	0.0835	0.0160	0.0842	-0.0251	0.0807
β	2.96e-04	4.7e-03	**1.69e-04**	4.84e-03	4.89e-05	4.82e-03	2.44e-04	5.10e-03
σ_w	**0.0046**	0.0161	0.0332	0.0139	0.0339	0.0159	-0.0312	0.0146
σ_v	**3.19e-03**	0.0158	0.0130	0.0165	0.0136	0.0158	-0.0667	9.22e-03
$\rho_{w,v}$	**-0.0047**	0.0436	0.0153	0.0303	0.0161	0.0329	–	–
$\rho_{w,\xi}$	-0.0285	0.1640	–	–	–	–	–	–
$\rho_{v,\xi}$	**-0.0631**	0.2850	–	–	–	–	-1.03	0.0843

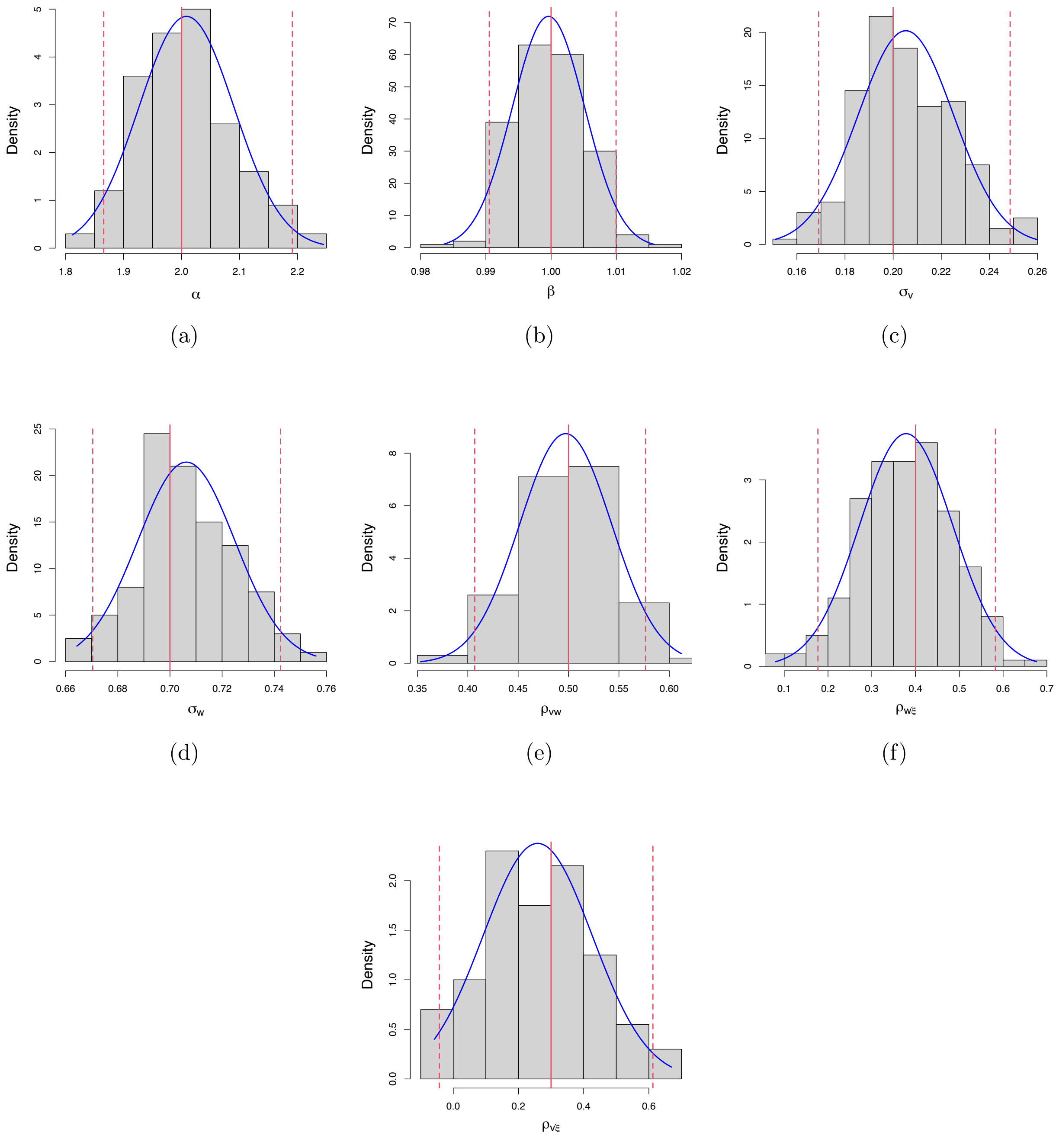
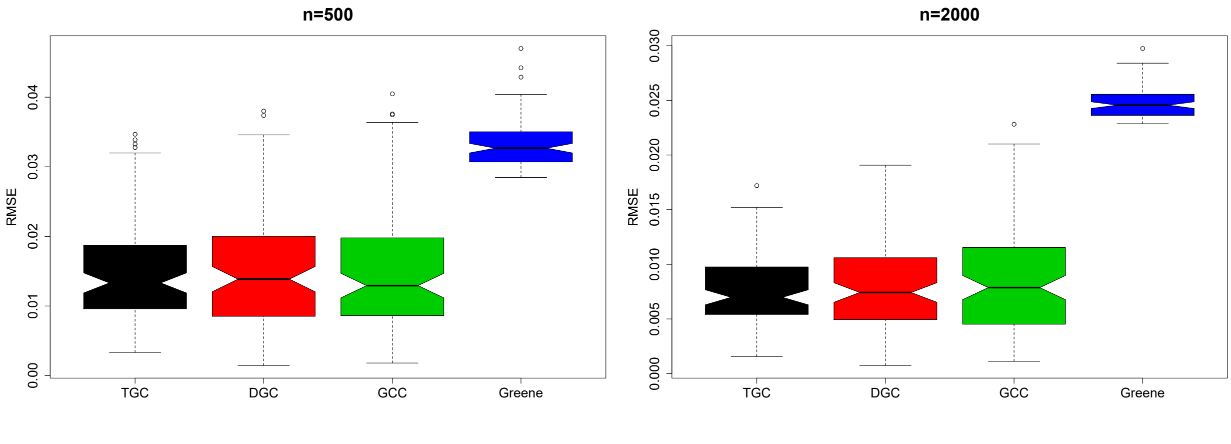


Figure 1: Histograms of parameter estimates for the simulated data of size $n = 2000$, when specifying the true TGC model. The normal fit is represented as a solid blue line. The true value is shown as a solid vertical red line, while the 2.5% and 97.5% quantiles are shown as broken vertical red lines.

Table 2: MSE of the four models. The smallest value is shown in bold.

$n = 500$	TGC	DGC	GCC	Greene
α	**2.36e-02**	2.69e-02	2.38e-02	2.40e-02
β	**6.78e-05**	6.84e-05	6.79e-05	1.02e-04
σ_w	**1.14e-03**	2.12e-03	2.11e-03	1.54e-03
σ_v	**1.45e-03**	1.57e-03	1.69e-03	6.41e-03
$\rho_{w,v}$	9.57e-03	**4.88e-03**	5.00e-03	–
$\rho_{w,\xi}$	1.02e-01	–	—	–
$\rho_{v,\xi}$	**3.17e-01**	–	—	1.06
$n = 2000$				
α	**6.64e-03**	8.00e-03	7.35e-03	7.14e-03
β	**2.22e-05**	2.34e-05	2.32e-05	2.60e-05
σ_w	**2.81e-04**	1.30e-03	1.40e-03	1.18e-03
σ_v	**2.59e-04**	4.42e-04	4.35e-04	4.53e-03
$\rho_{w,v}$	1.92e-03	**1.15e-03**	1.34e-03	—
$\rho_{w,\xi}$	2.77e-02	—	—	—
$\rho_{v,\xi}$	**8.52e-02**	—	—	1.07



(a) (b)

Figure 2: Box plots of RMSEs on TE scores estimated using the four models, for 100 randomly generated datasets of size $n = 500$ (a) and $n = 2000$ (b). (The scales of the two figures on the vertical axis are different).

from the TGC model with the previous parameter values, and 100 datasets generated from the DGC model with the following parameter values: $\alpha = 5$, $\beta = 0.7$, $\sigma_w = 1$, $\sigma_v = 1$, $\rho_{wv} = 0.7$ and $\rho_{\xi\varepsilon} = 0.5$. We repeated this experiment with two sample sizes: $n = 500$ and $n = 2000$. Figure 3 shows that the AIC values of both models are quite close, under both data distributions. With TGC-generated data, the TGC model achieves a lower AIC than the DGC model for 62% of the datasets of size $n = 500$ and 93% of the datasets of size $n = 2000$. For DGC-generated data, the DGC model achieves a lower AIC for 88% of the datasets of size $n = 500$, but only 11% of the datasets of size $n = 2000$. The TGC model would, thus, be selected more often when fitted to the larger datasets according to AIC, even when the true distribution is that of the DGC model. However, using the BIC for model selection would lead to different conclusions: the TGC model would be selected only for 15% and 47% of the TGC-generated data of size, respectively, 500 and 2000, while the DGC model would be selected for, respectively, 100% and 97% of the DGC-generated data of size, respectively, 500 and 2000. The conclusion of this simulation experiment is that it would be very difficult to select the true model for any of the two data distributions. However, the analysis of a real dataset presented in the next section shows that the TGC model may indeed fit the data better than double-copula models in some cases, and yield significantly different TE scores.

4. Application to Jasmine rice data

In this section, we compare our TGC model to the Greene and double-copula models using a real dataset about Jasmine rice production in Thailand. The data and model specification will first be described in Section 4.1, and the results will be reported in Section 4.2.

4.1. Data and model specification

The dataset used in this study was collected in the crop year 1999-2000 by interviewing farmers in three provinces of Thailand: Chiang Mai, Phitsanulok and Tung Gula Rong Hai (TGR). A total of 348 farmers were interviewed, of which 141 were purely Jasmine rice producers, while the remaining 207 farmers were mainly non-Jasmine rice producers.

The selection equation of the three models was specified as

$$Y_i^* = \alpha_0 + \alpha_1 \mathsf{return}_i + \alpha_2 \mathsf{edu}_i + \alpha_3 \mathsf{temp}_i + \alpha_4 \mathsf{rain}_i + \alpha_5 \mathsf{rice\_ratio}_i + \alpha_6 \mathsf{attitude}_i + \alpha_7 \mathsf{irrigation\_ratio}_i + \alpha_8 \mathsf{Phitsanulok}_i + \alpha_9 \mathsf{TGR}_i + \xi_i,$$

where the explanatory variables for the selection of Jasmine rice are the gross return from growing rice (return), the highest level of education in the household (edu), the mean annual temperature (temp), the total annual rainfall (rain), a dummy variable to account for farmers who transplanted rice (rice_ratio), the farmers' attitude towards commercialisation (attitude), a measure of access to irrigation (irrigation_ratio), and dummy variables for the Phitsanulok (Phitsanulok) and TGR (TGR) provinces. It is assumed that farmer i chooses to produce Jasmine rice if $Y_i^* > 0$.

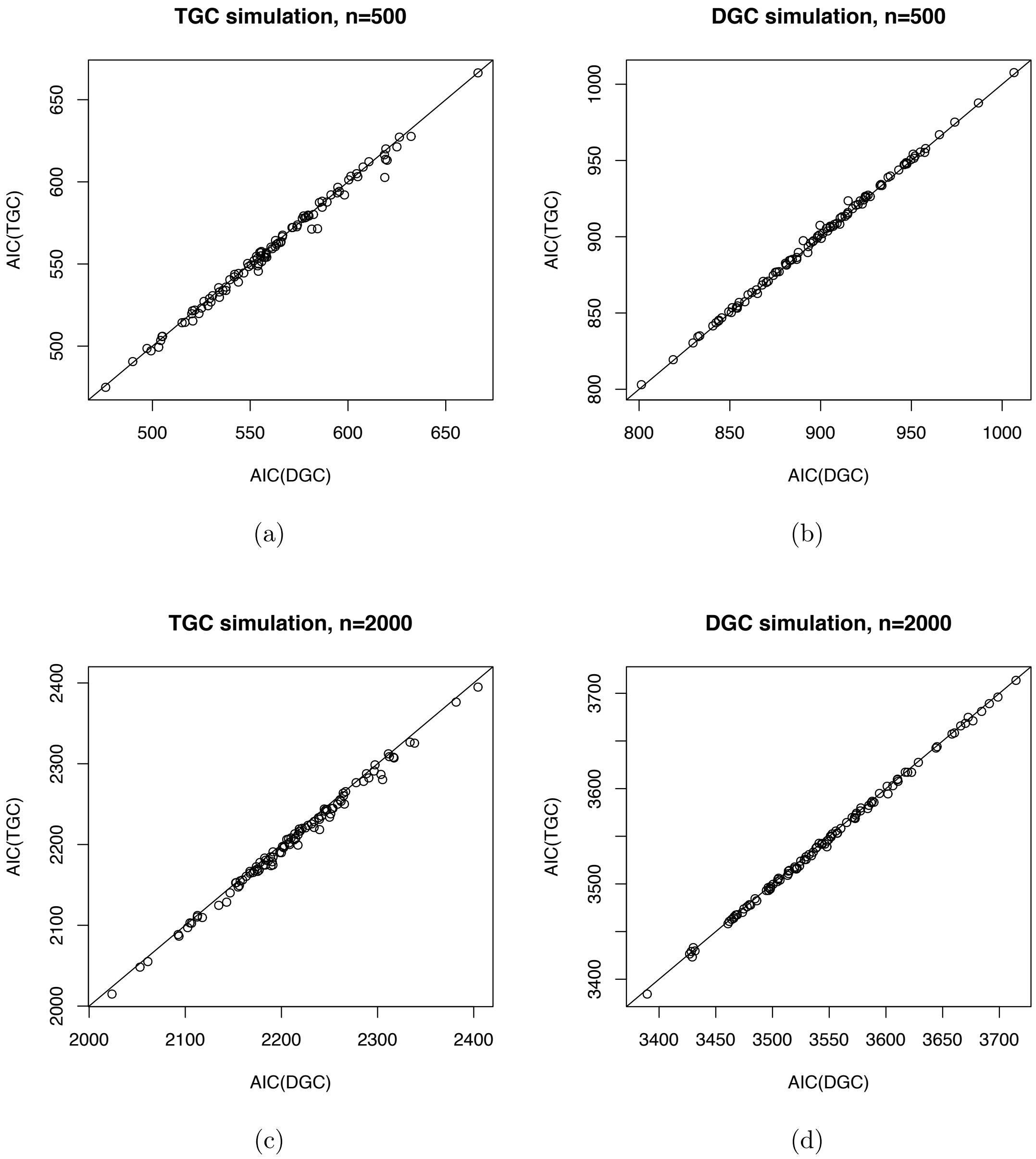


Figure 3: Biplots of AIC values for the TGC model (y-axis) vs. the DGC model (x-axis) fitted on 100 datasets generated from the TGC model (a,c) and from the DGC model (b,d). Size of datasets: $n = 500$ (a,b) and $n = 2000$ (c,d).

The stochastic frontier equation for Jasmine rice production is

$$\log \mathsf{output}_i = \beta_0 + \beta_1 \log \mathsf{labor}_i + \beta_2 \log \mathsf{fertilizer}_i + \beta_3 \log \mathsf{irrigation}_i + \beta_4 \log \mathsf{land}_i + \beta_5 \mathsf{Phitsanulok}_i + \beta_6 \mathsf{TGR}_i + \varepsilon_i,$$

$$\varepsilon_i = V_i - W_i,$$

where the four input variables are labour, chemical fertilizers, irrigation and land, all of which are expected to have a positive influence on rice output. Moreover, the same two regional dummy variables used in the selection equation were included to account for differences with respect to bio-physical and environmental factors.

Table 3 shows the correlation coefficients between the quantitative covariates in the selection and stochastic frontier equations. The highest correlations are observed between temp and rain in the selection equation, and between log fertilizer and log land in the frontier equation. In least-squares linear regression, multicollinearity is known to cause a high variance of coefficient estimates. Although this issue has not received as much attention in stochastic frontier modeling as it has in least-squares regression [5], it is likely that multicollinearity may cause similar problems in SFM's too, the main possible effect being a high standard error of some coefficient estimates making them statistically nonsignificant. As will be seen in the next section (Table 5), this does not seem to be the case with the dataset under study. Furthermore, multicollinearity is not likely to have an important effect on the estimation of technical efficiencies, which is often the main objective of stochastic frontier analysis [31].

4.2. Results and discussion

For parameter estimation, we used Halton sequences of length $M = 200$ for each observation. Table 4 shows the values of the log-likelihood as well as three information criteria: AIC, BIC, and the Hannan-Quinn Information Criterion (HQIC) [24] for the three models. Every model was evaluated with four different distributions of the inefficiency W: half-normal, exponential, gamma and truncated normal. The results of the double-copula model is for the best fitted model among several copula families including Gaussian, Clayton, Rotated Clayton, Gumbel, Rotated Gumbel, and Frank copulas. The double-copula best model has a Clayton copula rotated by 90 degrees for the dependence between V and W, and a Gaussian copula for the dependence between ε and ξ.

Overall, we can see that the TGC model with a gamma-distributed inefficiency has the best explanatory ability according to log-likelihood and the three information criteria. As the Greene and TGC models are nested, the likelihood ratio (LR) test can be used to compare them. According to this test, the correlations coefficients ρ_{vw} and $\rho_{w\xi}$ are significantly different from zero with a p-value less than 10^{-4}, whatever the inefficiency distribution. The double-copula model with a gamma-distributed inefficiency is a better fit than the Greene model, which can be explained by the fact that it accounts for the dependence between V and W; however, it is not as good as the TGC model.

Table 5 shows the parameter estimates and their standard errors for the three models with gamma-distributed inefficiency. We observe that the standard errors of all parameter

Table 3: Correlations between quantitative covariates in the selection and stochastic frontier equations. We separate in the table the five covariates of the selection equation and the four covariates of the stochastic frontier equation. The symbols *, ** and *** represent significance at levels 10%, 5%, and 1%, respectively.

Variables	return	edu	temp	rain	attitude	log irrig.	log labor	log fertil.	log land
return	1.000								
edu	0.055	1.000							
temp	−0.320***	−0.030	1.000						
rain	−0.454***	−0.025	0.881***	1.000					
attitude	0.243***	−0.062	0.059	−0.088	1.000				
log irrig.	0.168**	0.014	−0.215***	−0.184***	0.077	1.000			
log labor	−0.025	−0.104*	0.303***	0.237***	0.054	−0.076	1.000		
log fertil.	−0.438***	−0.096*	0.440***	0.556***	−0.110**	−0.232***	0.337***	1.000	
log land	−0.422***	−0.105*	0.610***	0.620***	0.031	−0.216***	0.510***	0.831***	1.000

Table 4: Information criteria of the TGC, Greene and double-copula models for the Jasmine rice data. HN, EX, GA, and TN stand for half-normal, exponential, gamma and truncated normal distributions, respectively. For each criterion, the best value for each model is underlined, and the overall best value is printed in bold.

	HN	EX	GA	TN
TGC				
Log-likelihood	−252.92	−247.97	**−235.3**	−248.97
AIC	549.85	539.94	**516.6**	543.93
BIC	634.6	624.68	**605.2**	632.53
HQIC	544.72	534.81	**511.24**	538.57
Greene				
Log-likelihood	−273.06	−275.03	−273.16	271.25
AIC	586.13	590.05	588.31	584.49
BIC	663.17	667.1	669.21	665.39
HQIC	581.46	585.39	583.41	579.6
χ^2 stat.	40.28	54.12	75.72	44.56
p-value	$<.0001$	$<.0001$	$<.0001$	$<.0001$
Double copula				
Log-likelihood	−285.50	−268.71	−266.58	−285.49
AIC	613.01	579.42	577.15	614.98
BIC	693.90	660.32	661.90	699.73
HQIC	608.11	574.52	572.02	609.85

estimates for the TGC model are smaller than those of the two other models, which suggests a better fit to the data. The double-copula and TGC models agree on finding a high negative correlation between V and W, which shows the necessity of relaxing the independence assumption. Greene's model and the TGC model both find a high positive correlation between V and ξ, which confirms that a serious selection bias exists, i.e., estimation using observations from only Jasmine or non-Jasmine rice producer data would provide biased estimates of productivity. This finding confirms the importance of accounting for sample selection in the estimation. The estimates of the parameters related to the error distributions (shape and scale of the gamma distribution, and σ_v) are quite different for the three models, which can be expected to impact the influence of technical efficiencies. This assumption will be confirmed later.

Except for α_7 and α_9, the estimates of the coefficients in the selection equation are similar across the three models. The estimates of coefficients $\beta_1, \ldots, \beta_4$ are of particular interest because they are elasticities, i.e., β_j is interpreted as the percentage change in output per one percent change in input x_j. We can see that the TGC and Greene models do not have much difference between elasticities. According to the result of the TGC model, the production elasticity with respect to changes in land area has the highest value of 0.67, implying that a 1% increase in land area allocated to Jasmine rice increases production by 0.67%. The production elasticities with respect to irrigation, fertilizer and labor are estimated, respectively, at 0.17, 0.13, and 0.09. The elasticity estimates of the double-copula model depart from those of the two other models. In particular, the negative estimate of the production elasticity with respect to labor is not realistic from an economic point of view. This observation shows that caution should be exercised when interpreting results obtained with an ill-specified model.

Summary statistics of technical efficiency scores for the three models are reported in Table 6. We observe large differences in the distributions of technical efficiency scores for the three models, which suggests that the correlations between W and V, and between W and ξ have a big impact on the estimates of technical efficiency, as was already observed in other studies [40]. Both Greene's model and, to an even larger extent, the double-copula model appear to overestimate technical efficiency. According to the TGC model, farmers also exhibit a wider range of production technical efficiency in Jasmine rice farming, which is consistent with previous findings reported by Ebers et al. [11] and Piya et al. [30].

Figures 4 and 5 show scatter plots of the TE scores estimated, respectively, using the Greene model and the double-copula model, vs. the TGC estimates, with different inefficiency distributions. Regardless of the distribution postulated for W, both the Greene model and the double-copula model overestimate TE as compared to the TGC model. Figure 6 shows kernel density estimates of the TE distributions for the three models. The TE distribution appears to be more robust with respect to the choice of the positive error distribution for the trivariate-copula model than it is for the other two models, which can be regarded as additional evidence for the superiority of the TGC model.

Table 5: Parameter estimates and standard errors for the three models with gamma-distributed inefficiency applied to the Jasmine rice data. For the coefficients α_j and β_j, one, two and three stars correspond, respectively, to significance at the 5%, 1% and 0.1% levels.

	TGC		Greene		Double copula	
	estimate	se	estimate	se	estimate	se
α_0	296.14***	2.41	296.23***	6.09	294.11***	4.30
α_1	−0.001***	< 0.001	−0.001***	< 0.001	−0.001***	< 0.001
α_2	0.06*	0.03	0.06*	0.03	0.07**	0.03
α_3	−91.65***	0.67	−91.65***	1.58	−91.45***	1.73
α_4	0.47	0.35	0.47	0.57	0.45	0.69
α_5	0.02	0.11	0.09	0.17	−0.02	0.19
α_6	0.05**	0.02	0.05	0.03	0.06	0.03
α_7	0.46***	0.14	1.18***	0.23	1.07***	0.23
α_8	2.40***	0.35	2.15***	0.46	2.88***	0.56
α_9	−0.35	0.26	−0.71*	0.34	0.24	0.39
β_0	6.46***	0.01	5.90***	0.34	6.39***	0.03
β_1	0.08***	< 0.001	0.12**	0.05	−0.04***	< 0.001
β_2	0.13***	< 0.001	0.11*	0.05	0.02***	0.005
β_3	0.17***	0.002	0.41***	0.11	0.12***	0.006
β_4	0.67***	<0.001	0.65***	0.08	0.97***	0.006
β_5	0.49***	0.001	−0.31*	0.13	−0.42***	0.004
β_6	0.52***	0.002	−0.57***	0.10	−0.58***	0.006
Shape	2.09	0.04	2.27	0.83	0.33	0.04
Scale	0.60	0.003	0.22	0.04	0.55	0.01
σ_v	0.11	0.005	0.39	0.06	0.30	0.03
ρ_{wv}	−0.99				−0.93	
$\rho_{w\xi}$	−0.96					
$\rho_{v\xi}$	0.96		0.98			
$\rho_{\xi\varepsilon}$					0.15	

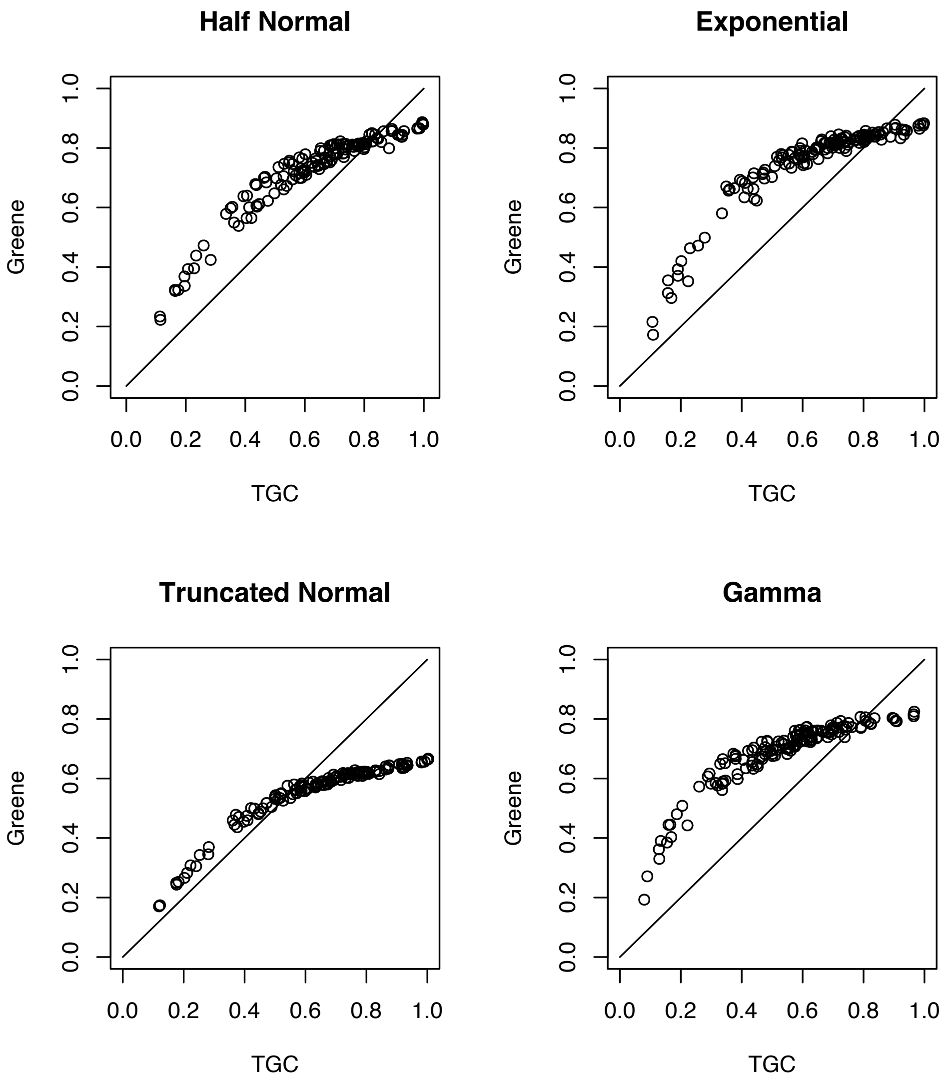


Figure 4: TE scores estimated using Greene's model (y-axis) versus those estimated using the TGC model (x-axis) for the three different inefficiency distributions.

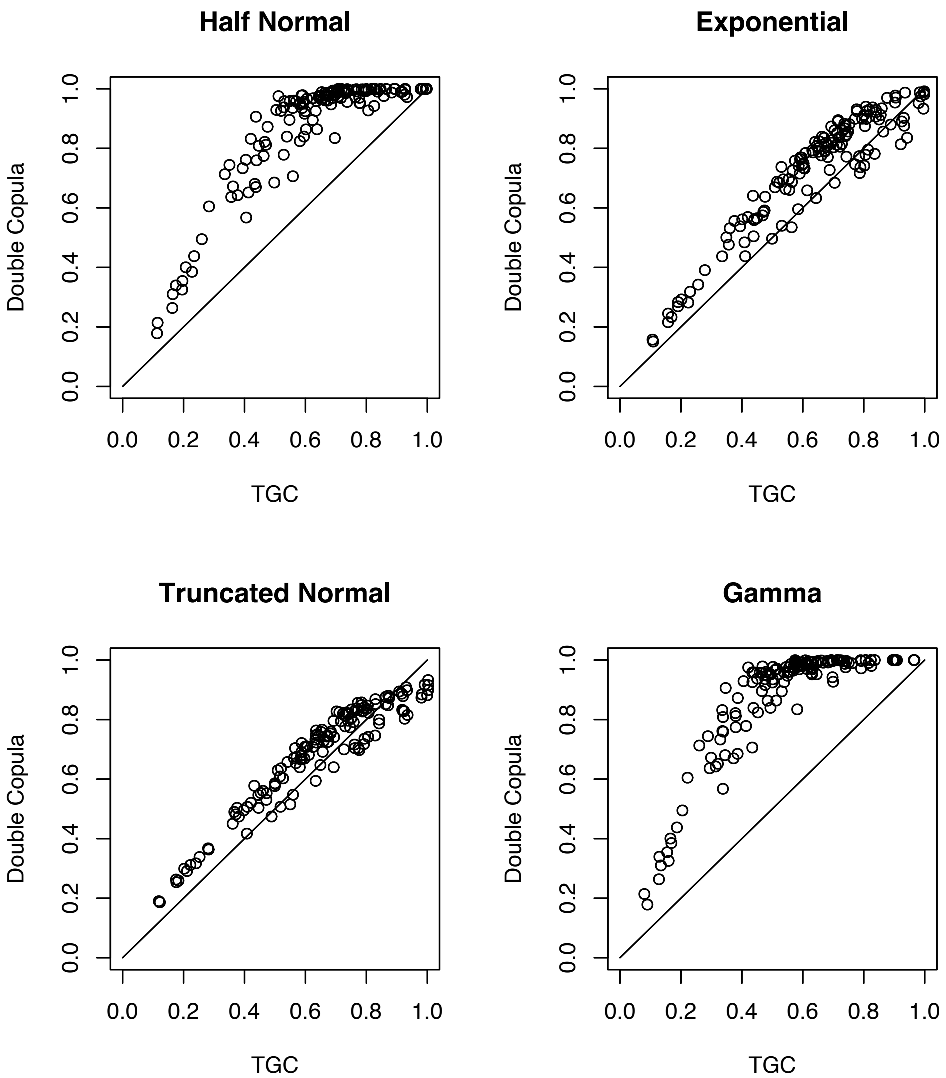
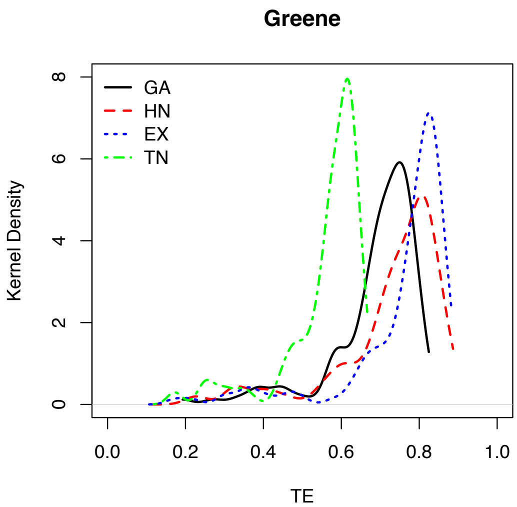
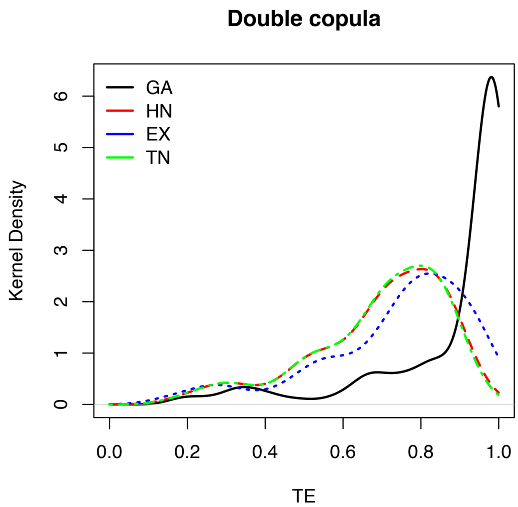


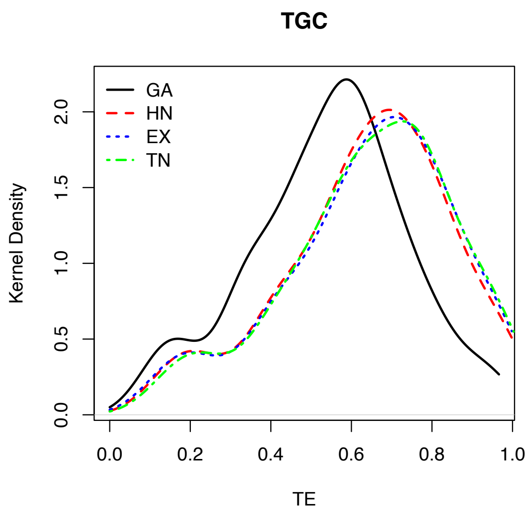
Figure 5: TE scores estimated using the double-copula model (y-axis) versus those estimated using the TGC model (x-axis) for the three different inefficiency distributions.



(a)



(b)



(c)

Figure 6: Kernel density estimates of the technical efficiency distributions from the Greene (a), double-copula (b) and TGC (c) models, with different inefficiency distributions.

Table 6: Range and frequency of TE scores.

	TGC		Greene		Double	
Range	# Farmers	%	# Farmers	%	# Farmers	%
$(0, 0.25]$	12	0.08	0	0.00	2	0.01
$(0.25, 0.5]$	41	0.29	9	0.06	9	0.06
$(0.5, 0.6]$	32	0.23	8	0.06	1	0.01
$(0.6, 0.7]$	29	0.21	49	0.35	8	0.06
$(0.7, 0.8]$	14	0.10	75	0.53	8	0.06
$(0.8, 1]$	13	0.09	0	0.00	113	0.80
Mean		0.54		0.68		0.88
sd		0.20		0.09		0.19
Min		0.08		0.27		0.18
Max		0.97		0.80		0.99

5. Conclusions

In recent years, it has been realized that adequately representing the dependencies between error terms is a key issue when designing SFMs, and that wrong assumptions on these dependencies can result in large errors in the estimation of technical efficiency. Copulas have proved to be a useful device for building more flexible SFMs [38, 45, 20]. For instance, in [45], we showed that wrongly assuming independence between the two-sided error term and the inefficiency term in the production equation may result in gross overestimation of technical efficiency, and that modeling this dependency using Gaussian copulas allows for a better fit to some datasets.

In this paper, we have applied a similar approach to stochastic frontier analysis with sample selection. We have relaxed the assumption of independence between two-sided random error and inefficiency in Greene's original model [15], by representing the dependencies between these two terms and the random error in the selection equation using a trivariate Gaussian copula parameterized by a correlation matrix. Our model is, thus, a proper generalization of Greene's model. We have compared the new model to Greene's model and to an alternative solution based on two bivariate copulas introduced in [40], using both simulated data and real data about Jasmine rice production. Our model has been shown to fit the real data better than the other two models, which tend to overestimate technical efficiency, confirming the trend already reported in [45].

In the future, it will be interesting to investigate alternative multidimensional copula families such as proposed by Durante et al. [10], Liebscher [23], Mazo et al. [26] or Zhu et al. [48].

Acknowledgements

This work has been supported by the Faculty of Economics and the Centre of Excellence in Econometrics at Chiang Mai University, as well as Faculty of Economics at Shandong University of Finance and Economics under research grant 19BJCJ46 and the education plan of the youth and creative talents in Shandong higher education institutions.

References

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

[2] S. Bazen and K. Waziri. The assimilation of young workers into the labour market in France: A stochastic earnings frontier approach. Technical Report IZA DP No. 10841, IZA Institute of Labor Economics, Bonn, Germany, 2017.

[3] D. E. Beckers and C. J. Hammond. A tractable likelihood function for the normal-gamma stochastic frontier model. *Economics Letters*, 24(1):33–38, 1987.

[4] B. E. Bravo-Ureta, W. Greene, and D. Solís. Technical efficiency analysis correcting for biases from observed and unobserved variables: an application to a natural resource management project. *Empirical Economics*, 43(1):55–72, 2012.

[5] E. Castaño and S. Gallón. A solution for multicollinearity in stochastic frontier production function models. *Lecturas de Economía*, pages 9–23, January 2017.

[6] T. Coelli and A. Henningsen. *frontier: Stochastic Frontier Analysis*, 2020. R package version 1.1-8.

[7] B. De Baets and H. De Meyer. Cutting levels of the winning probability relation of random variables pairwisely coupled by a same Frank copula. *International Journal of Approximate Reasoning*, 112:22–36, 2019.

[8] L. A. De los Santos-Montero and B. E. Bravo-Ureta. Productivity effects and natural resource management: econometric evidence from POSAF-II in Nicaragua. *Natural Resources Forum*, 41(4):220–233, 2017.

[9] D. Drouet Mari and S. Kotz. *Correlation and dependence*. Imperial College Press, London, UK, 2001.

[10] F. Durante, J. Quesada-Molina, and M. Úbeda-Flores. A method for constructing multivariate copulas. In *New Dimensions in Fuzzy Logic and Related Technologies - Proceedings of the 5th EUSFLAT Conference*, volume 1, pages 191–195, Ostrava, Czech Republic, September 2007.

[11] A. Ebers, T. T. Nguyen, and U. Grote. Production efficiency of rice farms in Thailand and Cambodia: a comparative analysis of ubon ratchathani and stung treng provinces. *Paddy and Water Environment*, 15(1):79–92, 2017.

[12] R. El Mehdi and C. M. Hafner. Inference in stochastic frontier analysis with dependent error terms. *Mathematics and Computers in Simulation*, 102:104–116, 2014.

[13] M. González-Flores, B. E. Bravo-Ureta, D. Solís, and 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*, 44:237–247, 2014.

[14] W. Greene. A general approach to incorporating selectivity in a model. Working Papers 06-10, New York University, Leonard N. Stern School of Business, Department of Economics, 2006.

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

[16] W. H. Greene. A gamma-distributed stochastic frontier model. *Journal of Econometrics*, 46(1):141–163, 1990.

[17] W. H. Greene. *Econometric analysis*. Prentice Hall, Upper Saddle River, NJ, USA, 7th edition, 2012.

[18] J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, 1964.

[19] J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[20] T.-H. Huang, C.-N. Hu, and B.-G. Chang. Competition, efficiency, and innovation in taiwan's banking industry – an application of copula methods. *The Quarterly Review of Economics and Finance*, 67:362–375, 2018.
[21] S. Krüger, T. Oehme, D. Rösch, and H. Scheule. A copula sample selection model for predicting multi-year lgds and lifetime expected losses. *Journal of Empirical Finance*, 47:246–262, 2018.
[22] S. C. Kumbhakar and C. A. Knox Lovell. *Stochastic Frontier Analysis.* Cambridge University Press, 2003.
[23] E. Liebscher. Construction of asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 99(10):2234–2250, 2008.
[24] R. Q. Liew and Y. Wu. Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19(1):12–30, 2013.
[25] C. D. Mayen, J. V. Balagtas, and C. E. Alexander. Technology adoption and technical efficiency: Organic and conventional dairy farms in the United States. *American Journal of Agricultural Economics*, 92(1):181–195, 2011.
[26] G. Mazo, S. Girard, and F. Forbes. A class of multivariate copulas based on products of bivariate copulas. *Journal of Multivariate Analysis*, 140:363–376, 2015.
[27] R. B. Nelsen. *An introduction to copulas.* Springer, London, UK, 2nd edition, 2010.
[28] T. A. Park. Assessing performance impacts in food retail distribution systems: A stochastic frontier model correcting for sample selection. *Agricultural & Resource Economics Review*, 43(3):373–389, 2014.
[29] J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996.
[30] S. Piya, A. Kiminami, and H. Yagi. Comparing the technical efficiency of rice farms in urban and rural areas: A case study from Nepal. *Trends in Agricultural Economics*, 5(2):2793–2798, 2012.
[31] J. Puig-Junoy. Technical inefficiency and public capital in U.S. states: A stochastic frontier approach. *Journal of Regional Science*, 41(1):75–96, 2001.
[32] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020.
[33] S. Rahman. Resource use efficiency under self-selectivity: the case of Bangladeshi rice producers. *Australian Journal of Agricultural and Resource Economics*, 55(2):273–290, 2011.
[34] S. Rahman, A. Wiboonpongse, S. Sriboonchitta, and Y. Chaovanapoonphol. Production efficiency of jasmine rice producers in Northern and North-Eastern Thailand. *Journal of Agricultural Economics*, 60(2):419–435, 2010.
[35] F. Rapisarda, D. Brigo, and F. Mercurio. Parameterizing correlations: a geometric interpretation. *IMA Journal of Management Mathematics*, 18(1):55–73, 2007.
[36] A. Sklar. Fonctions de répartition a n dimensions et leurs marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8:229–231, 1959.
[37] M. D. Smith. Modelling sample selection using archimedean copulas. *Econometrics Journal*, 6(1):99–123, 2003.
[38] M. D. Smith. Stochastic frontier models with dependent error components. *The Econometrics Journal*, 11(1):172–192, 2008.
[39] D. Solís, J. del Corral, L. Perruso, and J. J. Agar. Evaluating the impact of individual fishing quotas (IFQs) on the technical efficiency and composition of the US Gulf of Mexico red snapper commercial fishing fleet. *Food Policy*, 46:74–83, 2014.
[40] S. Sriboonchitta, J. Liu, A. Wiboonpongse, and T. Denœux. A double-copula stochastic frontier model with dependent error components and correction for sample selection. *International Journal of Approximate Reasoning*, 80:174–184, 2017.
[41] R. E. Stevenson. Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics*, 13(1):57–66, 1980.
[42] S. Tasena. Polynomial copula transformations. *International Journal of Approximate Reasoning*, 107:65–78, 2019.
[43] S. Tasena. On a distribution form of subcopulas. *International Journal of Approximate Reasoning*,

128:1–19, 2021.

[44] Z. Wei, E. M. Conlon, and T. Wang. Asymmetric dependence in the stochastic frontier model using skew normal copula. *International Journal of Approximate Reasoning*, 128:56–68, 2021.

[45] A. Wiboonpongse, J. Liu, S. Sriboonchitta, and T. Denœux. 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*, 65:34–44, 2015.

[46] M. Wollni and B. Brümmer. Productive efficiency of specialty and conventional coffee farmers in costa rica: Accounting for technological heterogeneity and self-selection. *Food Policy*, 37(1):67–76, 2012.

[47] P. Xue-Kun Song. Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

[48] X. Zhu, T. Wang, and V. Pipitpojanakarn. Constructions of multivariate copulas. In V. Kreinovich, S. Sriboonchitta, and V.-N. Huynh, editors, *Robustness in Econometrics*, pages 249–265. Springer International Publishing, Cham, 2017.

Appendix A. Second derivative of the trivariate Gaussian copula

From (5), the first derivative of trivariate Gaussian copula $C_{\mathbf{R}}$ w.r.t u_1 can be expressed as

$$\frac{\partial C_{\mathbf{R}}(u_1,u_2,u_3)}{\partial u_1} = \underbrace{\frac{d\Phi^{-1}(u_1)}{du_1}}_{1/\phi(q_1)} \int_{-\infty}^{q_2}\int_{-\infty}^{q_3} \phi_{\mathbf{R}}(q_1,y,z)dydz,$$

where $q_k = \Phi^{-1}(u_k)$, $k \in \{1,2,3\}$, and its second derivative w.r.t. u_1 and u_2 is

$$C''_{\mathbf{R}}(u_1,u_2,u_3) = \frac{1}{\phi(q_1)\phi(q_2)}\int_{-\infty}^{q_3}\phi_{\mathbf{R}}(q_1,q_2,z)dz \quad \text{(A.1)}$$

$$= \frac{1}{\phi(q_1)\phi(q_2)(2\pi)^{3/2}|\mathbf{R}|^{1/2}} \times I, \quad \text{(A.2)}$$

$$= (2\pi|\mathbf{R}|)^{-1/2}\exp\left(\frac{q_1^2+q_2^2}{2}\right) \times I \quad \text{(A.3)}$$

where I is the integral

$$I = \int_{-\infty}^{q_3} \exp\left(-\frac{1}{2}(q_1,q_2,z)\mathbf{R}^{-1}(q_1,q_2,z)^T\right)dz \quad \text{(A.4)}$$

with

$$(q_1,q_2,z)\mathbf{R}^{-1}(q_1,q_2,z)^T = [(1-\rho_{23}^2)q_1^2 + (1-\rho_{13}^2)q_2^2 + (1-\rho_{12}^2)z^2 - 2\rho_{12}q_1q_2 - 2\rho_{13}q_1z - 2\rho_{23}q_2z + 2\rho_{13}\rho_{23}q_1q_2 + 2\rho_{12}\rho_{23}q_1z + 2\rho_{12}\rho_{13}q_2z] \,/\, |\mathbf{R}|\,. \quad \text{(A.5)}$$

From (A.4) and (A.5), we get

$$I = \exp\left\{-\frac{1}{2\,|\mathbf{R}|}[(1-\rho_{23}^2)q_1^2 + (1-\rho_{13}^2)q_2^2 - 2(\rho_{12}-\rho_{13}\rho_{23})q_1q_2]\right\} \times J, \quad \text{(A.6)}$$

with

$$J=\int_{-\infty}^{q_3}\exp\left\{-\frac{1}{2\,|\mathbf{R}|}[(1-\rho_{12}^2)z^2-2(\rho_{13}q_1+\rho_{23}q_2-\rho_{12}\rho_{23}q_1-\rho_{12}\rho_{13}q_2)z]\right\}dz.$$

Let $\sqrt{\frac{1-\rho_{12}^2}{|\mathbf{R}|}}z=t$, then $z=t\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}$ and $dz=dt\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}$. With these notations, J can be written as

$$J=\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}\times$$
$$\int_{-\infty}^{q_3\sqrt{1-\rho_{12}^2}/\sqrt{|\mathbf{R}|}}\exp\left\{-\frac{1}{2}\left[t^2-2[(\rho_{13}-\rho_{12}\rho_{23})q_1+(\rho_{23}-\rho_{12}\rho_{13})q_2]\frac{t}{\sqrt{(1-\rho_{12}^2)\,|\mathbf{R}|}}\right]\right\}dt.$$

458 Let

$$D=\frac{2[(\rho_{13}-\rho_{12}\rho_{23})q_1+(\rho_{23}-\rho_{12}\rho_{13})q_2]}{\sqrt{(1-\rho_{12}^2)\,|\mathbf{R}|}},$$

then

$$\begin{aligned}J&=\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}\times\int_{-\infty}^{q_3\sqrt{1-\rho_{12}^2}/\sqrt{|\mathbf{R}|}}\exp\left\{-\frac{1}{2}\left(t-\frac{D}{2}\right)^2-\frac{D^2}{4}\right\}dt\\&=\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}\exp\left(\frac{D^2}{8}\right)\int_{-\infty}^{q_3\sqrt{1-\rho_{12}^2}/\sqrt{|\mathbf{R}|}}\exp\left\{-\frac{1}{2}(t-\frac{D}{2})^2\right\}dt\\&=\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}\exp\left(\frac{D^2}{8}\right)(2\pi)^{1/2}\Phi\left(q_3\sqrt{\frac{1-\rho_{12}^2}{|\mathbf{R}|}}-\frac{D}{2}\right)\end{aligned}\quad\text{(A.7)}$$

From (A.1), (A.6) and (A.7), we get

$$\begin{aligned}C''_{\mathbf{R}}(u_1,u_2,u_3)=(2\pi|\mathbf{R}|)^{-1/2}\exp\left(\frac{q_1^2+q_2^2}{2}\right)\times\\\exp\left\{-\frac{1}{2\,|\mathbf{R}|}[(1-\rho_{23}^2)q_1^2+(1-\rho_{13}^2)q_2^2-2(\rho_{12}-\rho_{13}\rho_{23})q_1q_2]\right\}\times\\\sqrt{\frac{|\mathbf{R}|}{1-\rho_{12}^2}}\exp\left(\frac{D^2}{8}\right)(2\pi)^{1/2}\Phi\left(q_3\sqrt{\frac{1-\rho_{12}^2}{|\mathbf{R}|}}-\frac{D}{2}\right)\end{aligned}\quad\text{(A.8)}$$

To further simplify the notation, let

$$B=\exp\left\{-\frac{1}{2\,|\mathbf{R}|}[(1-\rho_{23}^2)q_1^2+(1-\rho_{13}^2)q_2^2-2(\rho_{12}-\rho_{13}\rho_{23})q_1q_2]\right\}.$$

We have finally:

$$C''_{\mathbf{R}}(u_1, u_2, u_3) = \frac{B}{\sqrt{1-\rho_{12}^2}} \exp\left(\frac{D^2}{8}\right) \exp\left(\frac{q_1^2+q_2^2}{2}\right) \Phi\left(q_3\sqrt{\frac{1-\rho_{12}^2}{|\mathbf{R}|}} - \frac{D}{2}\right). \tag{A.9}$$