Dependence modeling in stochastic frontier analysis

Abstract: This review covers several of the core methodological and empirical developments surrounding stochastic frontier models that incorporate various new forms of dependence. Such models apply naturally to panels where cross-sectional observations on firm productivity correlate over time, but also in situations where various components of the error structure correlate between each other and with input variables. Ignoring such dependence patterns is known to lead to severe biases in the estimates of production functions and to incorrect inference.

Keywords: efficiency, productivity, panel data, endogeneity, determinants of inefficiency, dependence, copulas

MSC 2020: 62J02, 62H05, 91B38

1 Introduction

The roots of stochastic frontier analysis (SFA) can be traced back to the origins of classical growth accounting [82] and, perhaps, production planning [43]. These fields deal with production relationships, which are usually modeled through production functions or, more generally, transformation functions (e.g., Shephard’s distance functions, directional distance functions, cost functions, etc.). In the classical growth accounting approach, all variation in growth, apart from the variation of inputs, is attributed to the so-called Solow’s residual, which under certain restrictions measures what is referred to as the change in total factor productivity (TFP).

Early models assumed that all the decision making units (DMUs) represented in the data as observations (e.g., firms, countries, etc.) were independent of one another and fully efficient. This ignored the numerous inefficiencies that arise in practice, which have arbitrary dependence and can arise due to such factors as simultaneity in DMU decisions, unobserved heterogeneity, common sources of information asymmetry and other market imperfections [84], managerial practices [18], cultural beliefs [34], traditions, expectations, and other unobserved factors inducing unaccounted dependence in models of production [16].

A key feature of several recent developments in SFA is the construction of a statistical model with as few restrictions on its dependence properties as possible. This implicitly recognizes that various forms of...
dependence are empirical questions that can and should be statistically tested against the data. Modern implements of SFA provide a framework where shortfalls from the production potential are decomposed into two terms – statistical noise and inefficiency, both of which are unobserved by a researcher but can be estimated for the sample as a whole (e.g., representing an industry) or for each individual DMU under a variety of possible dependence scenarios.

The extensions of SFA allowing for dependence that has until recently been ignored or overly restrictive are the focus of this review. To a large extent, they are triggered by the fact that restricting the nature of dependence of the composed error can lead to severe biases in estimators and incorrect inference. For example, within the confines of the traditional SFA approach one can test for the presence of either inefficiency or noise \([25,73]\). Thus, the model encompasses the classical approach with a naive assumption of full efficiency (conditional mean) and the deterministic production frontier as special cases. However, such estimators and tests themselves have been derived under the assumption of independence between inefficiency and noise; empirical results suggest that allowing for dependence changes the estimates and tests significantly, potentially distorting the conclusions of the classical models.

Thus, SFA under dependence is a natural relaxation of the extreme assumptions of full efficiency and independence, yet it also encompasses them as special cases, which can still be followed if the data and the statistical tests do not recommend otherwise. If there is evidence in favor of the full efficiency hypothesis after allowing for dependence, one can proceed with regression techniques or growth accounting, but this inference would now be robust to these assumptions.

Accounting for dependence within a production model could be critical for both quantitative and qualitative conclusions and, perhaps more importantly, for the resulting policy implications. For example, El Mehdi and Hafner \([28]\) found that estimated technical efficiency scores across the financing of Moroccan rural districts allowing for dependence tend to be lower than under the assumption of independence but the rankings remained basically the same. Thus, a key difference emerges if one is looking to identify the best versus measure how much improvement can be made.

While some of the methods and models we present here can also be found in previous reviews, e.g., \([67]\) and \([9]\), and it is impossible to give a good review without following them to some degree, here we also summarize many of (what we believe to be) key recent developments as well as (with their help) shed some novel perspectives onto the workhorse methods. We do not claim, however, that this survey comprehensively covers all of the relevant recent developments in modelling dependence in SFA. Many other important references can be found elsewhere.

The rest of the article is structured as follows. Section 2 introduces the classical cross-sectional stochastic frontier models (SFMs) and focuses on dependence between error components in such models. Section 3 considers dependence via sample selection. Section 4 surveys dependence models used in panels. Section 5 discusses dependence that underlies endogeneity in SFM, which is a situation when there is dependence between production inputs and error terms. Section 6 discusses how dependence can help obtain more precise estimates of inefficiency. Section 7 concludes.

## 2 The benchmark SFM and dependence within the composed error

In cross-sectional settings, one of the main approaches to study productivity and efficiency of firms is the SFM, independently proposed by Aigner et al. \([3]\) and Meesuen and van den Broeck \([61]\).\(^1\) Using conventional notation, let \(Y_i\) be the single output for observation (e.g., firm) \(i\) and let \(y_i = \ln(Y_i)\). The cross-sectional SFM can be written for a production frontier as

\[
y_i = m(x_i; \beta) - u_i + \nu_i = m(x_i; \beta) + \epsilon_i. \tag{2.1}
\]

\(^1\) [15] and [62], while appearing in the same year, are applications of the methods.
Here \( m(\mathbf{x}_i; \beta) \) represents the production frontier of a firm (or more generally a DMU), with given input vector \( \mathbf{x}_i \). Observations indexed by \( i = 1, \ldots, N \) are assumed to be independent and identically distributed. Our use of \( \beta \) is to clearly signify that we are parametrically specifying our production function, most commonly as a Cobb-Douglas or translog (see, e.g., [67] or [68] for a detailed treatment on nonparametric estimation of the SFM).

2.1 Canonical independence framework

The main difference between a standard production function setup and the SFM is the presence of two distinct error terms in the model. The \( u_i \) term captures inefficiency, shortfall from maximal output dictated by the production technology, while the \( v_i \) term captures stochastic shocks. The standard neoclassical production function model assumes full efficiency – so SFA embraces it as a special case, when \( u_i = 0, \forall i \), and allows the researcher to test this statistically.² It is commonly assumed that inputs are exogenous, in the sense that \( \mathbf{x} \) is independent of \( u \) and \( v \), \( \mathbf{x} \perp (u, v) \), and the two components of the error term are independent, \( u \perp v \).

Many estimation methods require distributional assumptions for both \( u \) and \( v \) (beyond the assumption of independence). For an assumed distributional pair, one can obtain the implied distribution for \( e_i \) and then estimate all of the parameters of the SFM with the maximum likelihood estimator (MLE). The most common assumption is that \( v_i \sim N(0, \sigma_v^2) \) and \( u_i \) is from a Half Normal distribution, \( N_i(0, \sigma_u^2) \), or \( u_i \) is from an Exponential distribution with parameter \( \sigma_u \).

The most popular case for the density of the composed error \( e \) is obtained for the Normal Half Normal specification under independence \( u \perp v \). According to Aigner et al. [3], the distribution function of a sum of a normal and Truncated Normal was first derived by Weinstein [94]. Let \( f_v \) and \( f_u \) denote the density of \( v \) and \( u \), respectively. For the Normal Half Normal case, \( f_v(v) = \phi \left( \frac{v}{\sigma} \right) \) and \( f_u(u) = \frac{2}{\sigma_u} \phi \left( \frac{u}{\sigma_u} \right) \), where \( \phi(\cdot) \) is the standard Normal probability density function (pdf). The closed form expression for the pdf can be obtained by convolution as follows:

\[
 f(e) = \int_0^\infty f_v(e + u)f_u(u)du = \frac{2}{\sigma} \phi \left( \frac{e}{\sigma} \right) \Phi \left( -\frac{e\lambda}{\sigma} \right), \tag{2.2}
\]

where \( \Phi(\cdot) \) is the standard normal cumulative distribution function (cdf), with the parameterization \( \sigma = \sqrt{\sigma_u^2 + \sigma_v^2} \) and \( \lambda = \sigma_u / \sigma_v \). \( \lambda \) is commonly interpreted as the proportion of variation in \( e \) due to inefficiency. The density of \( e \) in (2.2) can be characterized as that of a Skew Normal random variable with location parameter 0, scale parameter \( \sigma \), and skew parameter \( -\lambda \).³ This connection has only recently appeared in the efficiency and productivity literature [24].

It is worth noting that the closed form expression in (2.2) is equivalent to

\[
 f(e) = \mathbb{E}_u f_v(e + u) = \mathbb{E}_v f_u(v - e),
\]

where expectations are taken with respect to the relevant distribution. This suggests an alternative, simulation-based, way to construct the density by sampling from the distribution of \( u \) or \( v \) and evaluating the corresponding sample averages. Among the two sampling options (from the distribution of \( u \) or from the

² Prior to the development of the SFM, approaches which intended to model inefficiency typically ignored \( v_i \) leading to estimators of the SFM with less desirable statistical properties: see [1,2,27,71,72,86].

³ The pdf of a Skew Normal random variable \( x \) is \( f(x) = 2\phi(x)\Phi(\alpha x) \). The distribution is right skewed if \( \alpha > 0 \) and is left skewed if \( \alpha < 0 \). We can also place the Normal, Truncated Normal pair of distributional assumptions in this class. The pdf of \( x \) with location \( \xi \), scale \( \omega \), and skew parameter \( a \) is \( f(x) = \frac{2}{\omega} \phi \left( \frac{x - \xi}{\omega} \right) \Phi \left( a \left( \frac{x - \xi}{\omega} \right) \right) \). See [10] and [65] for more details.
distribution of $v$), sampling the $u$’s is more practical as it avoids to need to ensure that $v - \varepsilon > 0$. Sampling $u$ can be easily done by sampling from the standard normal distribution and taking the absolute values $\sigma_d N(0, 1)$ (in the case of the Half Normal distribution).\(^4\)

Our mathematical formulation will focus on a production frontier as it is the most popular object of study. The framework for dual characterizations (e.g., cost, revenue, profit) or other frontiers is similar and follows with only minor changes in notation. For example, the cost function formulation is obtained by changing the sign in front of $u$ to a “+,” which will represent excess, rather than shortfall, of cost above the minimum level.

### 2.2 Modeling dependence

Smith [81] relaxed the assumption of independence between $u$ and $v$ by introducing a copula function to model their joint distribution. This is one of the first relaxations of the independence assumptions available for SFA and it allowed testing the adequacy of this assumption. If the marginal distributions of $u$ and $v$ are linked by a copula density $c(\cdot, \cdot)$, then their joint density can be expressed as follows:

$$f(v, u) = f_d(v) f_u(u) c(F_v(v), F_u(u)),$$

where $F_v$ and $F_u$ denote the respective cdfs. It then follows by a similar construction to (2.2) that the density of $\varepsilon$ can be written as

$$f(\varepsilon) = \int_0^\infty f_u(\varepsilon + u) f_u(u) c(F_v(\varepsilon + u), F_u(u)) du.$$

For commonly used copula families, this density does not have a close form expression similar to (2.2), even in the Normal Half Normal case, so a simulation-based approach would often need to be used, where we simulate many draws of $u$ and evaluate the sample analogue of the following expectation with respect to the distribution of $u$:

$$f(\varepsilon) = \mathbb{E}_d [f_u(\varepsilon + u) f_u(u) c(F_v(\varepsilon + u), F_u(u))].$$

Smith [81] found that ignoring the dependence can lead to biased estimates and discussed how one can test whether the independence assumption nested within this model is adequate. It is easy to see that the model in (2.2) is a special case of (2.3) when $c(\cdot, \cdot)$ is the independence (or product) copula.

From $f(\varepsilon)$, along with the assumption of independence over $i$, the log-likelihood function can be written as follows:

$$\ln \mathcal{L} = \ln \left( \prod_{i=1}^n f(\varepsilon_i) \right) = \sum_{i=1}^n \ln f(\varepsilon_i),$$

where $\varepsilon_i = y_i - m(x_i; \beta)$. The SFM can be estimated using the traditional MLE, if an analytic expression for the integrals is available, or the maximum simulated likelihood estimator (MSLE), if we need to use a simulation-based approach to evaluate the integrals of the density [60].\(^5\) The benefit of MLE/MSLE is that under the assumption of correct distributional specification of $\varepsilon$, the MLE is asymptotically efficient (i.e., consistent, asymptotically Normal, and its asymptotic variance reaches the Cramer–Rao lower bound).

---

\(^4\) With modern statistical software it is straightforward to sample from a wide swath of one-sided distributions that have been suggested in the SFA literature: Exponential, Gamma, Truncated Normal, Weibull, Beta, Uniform, Binomial, Generalized Exponential, etc.

\(^5\) Burns [21], who was Smith’s student, proposed using MSLE, and while predating Smith [81] actually attributes a working paper version.
A further benefit is that a range of testing options are available. For instance, tests related to \( \beta \) can easily be undertaken using any of the classic trilogy of tests: Wald, Lagrange multiplier, or likelihood ratio. The ability to readily and directly conduct asymptotic inference is one of the major benefits of SFA over data envelopment analysis (DEA).\(^6\)

Two main issues that practitioners face when confronting dependence are the choice of copula model and the assumed error distributions that best fit the data. As Wiboonpongse et al. ([95], p. 34) note “The impact of the independence assumption on technical efficiency estimation has long remained an open issue.” Analytical criteria such as AIC or BIC can be used for these purposes, see both [9] and [95] for detailed reviews.

More specifically, Wiboonpongse et al. [95] use MSLE and systematically consider several copula families including the Student-t, Clayton, Gumbel and Joe families as well as their relevant rotated versions. Their data is a cross section from coffee production in Thailand and they use AIC and BIC to determine which copula model is most appropriate. Wiboonpongse et al. [95] also assume the marginals of the two error components are Normal and Half Normal, then apply a range of copulas to inject dependence. In their empirical application they have a total of 111 observations. The Clayton copula is found to be the best and plots of technical efficiencies across 111 farmers for independence and the best fitting copula model found near uniformly lower TE scores (though not much different). Finally, it also appears that the ranks are preserved (see their Figure 3).

An unintended benefit of modelling dependence is that it may alleviate the “wrong skewness” issue that is common in the canonical Normal Half Normal SFM [77,89]. The wrong skewness issue arises when the OLS residuals display skewness that is of the wrong sign compared to that stemming from the frontier model (so positive when estimating a production frontier). For specific distributional pairs the model cannot separately identify the variance parameters for both \( v \) and \( u \). It was noted in [19] that the third central moment of the composed error is

\[
E[(\varepsilon - E[\varepsilon])^3] = E[(v - E[v])^3] - E[(u - E[u])^3] + 3\text{Cov}(u^2, v) - 3\text{Cov}(u, v^2) - 6E[u - E[u]]\text{Cov}(u, v).
\]

It is clear that the skewness of \( \varepsilon \) only depends on the skewness of \( u \) when \( v \) is assumed to be symmetric and independent of \( u \). Once \( u \) and \( v \) are allowed to be dependent and/or \( v \) is allowed to be asymmetric, then the skewness of the composed error does not have to align with the skewness of inefficiency. Thus, modelling dependence is one way in which some of the empirical vagaries of the SFM can be overcome [93].

### 2.2.1 Asymmetric dependence

A common feature of all of the papers that have allowed dependence in the SFM is the use of copulas that introduce symmetric dependence. Symmetric dependence assumes that the noise \( v \) and inefficiency components are treated equally in the SFM. However, a recent suggestion by Wei et al. [91] offers a set of copulas that allow for asymmetric dependence. As Wei et al. ([91], p. 57) note “[…]in practical situations, the inefficiency component \( u \) and the error component often play different roles in global inefficiency, and in such cases, the symmetric copulas are not suitable.” They define asymmetric copulas as those that have non-exchangeable and/or radial asymmetric properties [92].

Wei et al. [91] introduced the Skew Normal copula and used it to construct their SFM with dependence. An interesting feature of their general setup is that they allow both \( v \) and \( u \) to be asymmetric along with an asymmetric copula (see their Proposition 3.1). As in [95], Wei et al. [91] recommended selecting the copula model based on AIC/BIC. In their empirical application 31 out of 108 farms have the same efficiency rank.

---

\(^6\) This in no way suggests that inference cannot be undertaken when the DEA estimator is deployed; rather, the DEA estimator has an asymptotic distribution which is much more complicated that the MLE for the SFM, and so direct asymptotic inference is not available; bootstrapping techniques are required for many of the most popular DEA estimators [78,79].
(the bottom 5 are in complete agreement as are 4 of the top 5) across the standard SFM and the asymmetric copula SFM. The point estimates of technical efficiency however show large differences among the two competing models, which again provides evidence that ignoring dependence can have an undue influence on the point estimates of technical inefficiency.

3 Dependence via sample selection

Another way in which dependence can arise is through sample selection. By itself sample selection has only recently been a serious area of focus in the stochastic frontier literature. Several early approaches to deal with potential sample selection follow the two-step correction [37]. In the first stage, the probability of selection is estimated and the inverse Mill’s ratio is calculated for each observation. This estimated inverse Mill’s ratio is then included as a regressor in the final SFM. An example of this is the Finnish farming study [80]. This limited information two-step approach works in a standard linear regression framework because of linearity, which Greene [33] makes clear. However, as shown in [51], when inefficiency is present no two-step approach will work and full information maximum likelihood estimation is required.

Recognizing the limitations of direct application of the two-stage approach, both Kumbhakar et al. [51] and Greene [33] proposed alternative stochastic frontier selection models. The two approaches differ in how selection arises in the model. The Greene [33] model allows the choice of technology to be influenced by correlation between random error in the selection and frontier models, whereas Kumbhakar et al. [51] constructed a model where the choice of technology is based on some aspect of inefficiency, inducing a different form of sample selection. Beyond the difference in how selection arises, the sample selection stochastic production frontier models [51] and [33] are identical.

Sriboonshitta et al. [83] were the first to recognize that dependence could enter the sample selection model. They work with the Greene [33] stochastic frontier sample selection model and admit dependence into the composite error term. This is termed a double-copula because they have a copula in the sample selection equation and a copula in the SFM. See ([83], equation (20)) for the likelihood function.

Beyond a small set of simulations, Sriboonshitta et al. [83] applied the double copula sample selection SFM to 200 rice farmers from Kamphaeng Phet province, Thailand, in 2012 using a Cobb-Douglas production frontier and considered eight different copula functions (see their Table 4). Their preferred model based on the AIC is a Gaussian copula with 270-degree rotated Clayton model. They find a substantial difference in estimated TE scores between the Greene [33] model which assumes independence and their double-copula model (see their Figure 5). As Sriboonshitta et al. ([83], p. 183) note “[...]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.”

As a further extension of [83], Liu et al. [59] noted that “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.” Liu et al. [59] generalized the Greene [33] 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. The key feature is that the trivariate and double copula models rely on different assumptions concerning the joint distribution of v, u, and ξ (ξ here is the error in the selection equation). Liu et al. [59] made note of the decomposition

\[ f(v, u, ξ) = f(v)f(u|v)f(ξ|v, u), \]

and note that the double copula model assumes that \( f(ξ|v, u) = f(ξ|v - u) \), i.e., the distribution of ξ only depends on the composite error, not the individual pieces. This also implies that the double copula model and the trivariate copula model are nonnested.

Liu et al. [59] provided an application that focuses on Jasmine/non-Jasmine rice farming in Thailand. The data suggest u is Gamma distributed for the most preferred model. As with some of the earlier papers, Liu et al. ([59], p. 193) noted that “[...]both Greene’s model and the double-copula model appear to
overestimate technical efficiency. According to the [trivariate Gaussian copula] model, farmers also exhibit a wider range of production technical efficiency in Jasmine rice farming […].”

4 Dependence in panel SFM

When repeated observations of the firms are available, then we can allow for richer models that incorporate unobserved components and various other dependence structures. Most importantly, we can extract information about likely time trends in inefficiency and time constant firm-specific characteristics. Pitt and Lee [69] seem to be the first to extend the cross-sectional SFM to a panel structure, and Schmidt and Sickles [76] were the first to propose a panel-specific methodology for SFA.

4.1 A benchmark specification

The benchmark panel SFM can be written as follows:

\[ y_{it} = m(x_{it}; \beta) + c_i - \eta_i - u_{it} + v_{it} = m(x_{it}; \beta) - \alpha_i + e_{it}. \] (4.1)

This model differs from (2.1) in many ways. All observed variables and error terms inherited from (2.1) now have a double-index for both firms, \( i \), and time, \( t = 1, \ldots, T \). In addition, the model contains the so-called firm-specific heterogeneity, \( c_i \), and the time-invariant component of inefficiency \( \eta_i \). Compared with (2.1), \( c_i \) encapsulates any unobserved factors that affect output (other than inputs) without changing over time such as unmeasured management or operational specifics of the firm. If such factors are present, the dependence between \( c_i \) and \( x_i \) causes omitted variable biases and invalidates inference based on cross-sectional SFM. In panel models, when ignored, such factors can serve as common sources of dependence in the error term \( e_i \) which also leads to invalid inference.

Another distinguishing feature of (4.1) is the presence of \( \eta_i \), a component of inefficiency which is time-invariant. This means that inefficiency is composed of both time-invariant and time variant components, which are sometimes interpreted as long-run and short-run inefficiency. Since both \( c_i \) and \( \eta_i \) are unobserved, it will generally be difficult to decompose \( \alpha_i \) into its subsequent firm-specific and time-invariant inefficiency components.

Classical panel methods (i.e., methods that assume that \( \eta_i \) and \( u_{it} \) do not exist) allow for various forms of dependence between \( c_i \) and \( x_{it} \). For example, estimation under the fixed effects (FE) framework allows \( x_{it} \) to be correlated with \( c_i \) and uses a within transformation to obtain a consistent estimator of \( \beta \). Alternatively, estimation in the random effects (RE) framework assumes that \( x_{it} \) and \( c_i \) are independent and uses OLS or GLS. The difference between OLS and GLS arises due to the fact that the variance-covariance matrix of the composed error term \( c_i + v_{it} \) is no longer diagonal, and so, feasible GLS is asymptotically efficient.

The early work on panel SFM assumed inefficiency to be time-invariant. This allowed handling dependence within panels using classical panel methods such as FE and RE estimation [76]. The standard time-invariant SFM is

\[ y_{it} = \beta_0 + x_{it}'\beta + v_{it} - \eta_i = (\beta_0 - \eta_i) + x_{it}'\beta + v_{it} = c_i + x_{it}'\beta + v_{it}, \] (4.2)

where \( c_i = \beta_0 - \eta_i \). Under the FE framework, the serially independent inefficiency \( \eta_i \), \( i = 1, \ldots, n \), is allowed to have arbitrary dependence with \( x_{it} \). In cases in which there are time-invariant variables of interest in the production model, one can use the RE framework, which also requires no distributional assumptions on \( v \) and \( \eta \) and can be estimated with OLS or GLS. Alternatively, in such cases, one can rely on distributional assumptions as in [69], where \( v_{it} \) is assumed to follow a Normal distribution and \( \eta_i \) Half Normal.
Table 1: Selection of panel data methods

| Paper                          | Serial dependence in inefficiency | Other dependence allowed |
|--------------------------------|----------------------------------|--------------------------|
| Schmidt and Sickles [76]       | None                             | Between $\eta_i$ and $x_{it}$ |
| Cornwell et al. [26]           | $u_{it} = c_{0i} + c_{1i}t + c_{2i}t^2$ | Between $\eta_i$ and $x_{it}$ |
| Kumbhakar [47]                 | $u_{it} = [1 + \exp(y_{it} + y_{2i}^2)]^{-1}u_i$ | None |
| Battese and Coelli [13]        | $u_{it} = \exp(\gamma(t-T)\gamma(t))u_i$ | None |
| Lee and Schmidt [57]           | $u_{it} = \eta_i\beta$          | Between $\eta_i$ and $x_{it}$ |
| Battese and Coelli [14]        | None                             | $u_{it} \sim N(\sigma_\Delta^2, \sigma_\tau^2)$ |
| Kumbhakar and Heshmati [48]    | $u_{it} = \eta_i + \tau_t$      | None |
| Greene [31]                    | None                             | Between $c_i$ and $x_{it}$ |
| Greene [32]                    | None                             | Between $c_i$ and $x_{it}$ |
| Wang and Ho [90]               | None                             | $u_{it} \sim N(0, \sigma_{\tau_i}^2)$ |
| Badunenko and Kumbhakar [11]   | $u_{it} = \eta_i + \tau_t$      | $\eta_i \sim N(0, \sigma_{\eta_i}^2)$ |
|                                |                                  | $\sigma_{\eta_i} = \eta_i \exp(z_{it}\theta_1)$ |
|                                |                                  | $\tau_i \sim N(0, \sigma_{\tau_i}^2)$ |
|                                |                                  | $\sigma_{\tau_i} = \sigma_\tau \exp(z_{it}\theta_2)$ |

Table 1 contains a summary of the classical SFMs allowing for specific forms of serial dependence in $u_{it}$. It also lists any additional dependence structures permitted in these different models such as dependence between $c_i$ and $x_{it}$. See [67] and [50] for a detailed discussion of these methods.

4.2 Quasi MLE

If there is no $(c_i, \eta_i)$ or $(c_i, \eta_i)$ is assumed to be part of $u_{it}$ or $\nu_{it}$, which are independent of $x$, then we can view the panel model as a special case of the cross-sectional model (2.1), only with the double-index $it$. The MLE method described in Section 2 applies in this case but it uses the sample likelihood obtained leveraging the assumption of independence over both $i$ and $t$, not just $i$. Because independence over $t$ is questionable in panels, this version of MLE is often referred to as quasi-MLE (QMLE).

Let $\theta = (\beta', \sigma^2, \lambda)'$ and let $f_{it}$ denote the density of the composed error term evaluated at $v_{it} = \nu_{it} - u_{it}$. Then,

$$f_{it}(\theta) = f(v_{it}) = f(y_{it} - x_{it}\beta)$$

and the QMLE of $\theta$ can be written as

$$\hat{\theta}_{\text{QMLE}} = \arg\max_{\theta} \sum_i \sum_t \ln f_{it}(\theta).$$

The QMLE is known to be consistent even if there is no independence over $t$ but to obtain the correct standard errors one needs to use the so-called “sandwich,” or misspecification-robust, estimator of the QMLE asymptotic variance matrix. QMLE is known to be dominated in terms of precision by several other estimators that use the dependence information explicitly (and correctly). However, an appeal of the QMLE in this setting is that assuming independence is more innocuous in the sense that it does not lead to estimation bias, only to a lack of precision, when compared to a misspecification of the type of dependence that can lead to distinct biases.
Amsler et al. [5] proposed several estimators that model time dependence in panels. One such estimator can be obtained in the Generalized Method of Moments (GMM) framework. Let \( s_\theta(\theta) \) denote the score of the density function \( f_\theta(x) \), i.e.,

\[
s_\theta(\theta) = \nabla_\theta \ln f_\theta(x),
\]

where \( \nabla_\theta \) denotes the gradient with respect to \( \theta \). Then, the QMLE of \( \theta \) solves

\[
\sum_i \sum_t s_\theta(\hat{\theta}_{QMLE}) = 0
\]

and is identical to the GMM estimator based on the moment condition

\[
\mathbb{E} \sum_t s_\theta(\theta) = 0, \tag{4.6}
\]

where expectation is with respect to the distribution of \( \epsilon \). However, under time dependence, summation (over \( t \)) of the scores in (4.6) is not the optimal weighting. The theory of optimal GMM suggests using correlation of \( s_i \) over \( t \) by applying the GMM machinery to the \( T \) score functions written as follows:

\[
\begin{bmatrix}
\vdots \\
\varepsilon_{it} \\
\varepsilon_{iT}
\end{bmatrix}
\]

where, as before, \( \nabla_\theta \approx \nabla \ln f_\theta(x) \) is the pdf of the composed error term evaluated at \( x_{it} \) and \( f_\theta(x) \) is the corresponding cdf. Once the joint density is obtained we can construct a log-likelihood and run MLE. If we let the copula density have a scalar parameter \( \rho \), then the sample log-likelihood can be written as follows:

\[
\ln \mathcal{L}(\theta, \rho) = \sum_{i=1}^n \ln c(F_{it}(\theta), \ldots, F_{iT}(\theta); \rho) + \ln f_{it}(\theta) + \ldots + \ln f_{iT}(\theta). \tag{4.7}
\]

The first term in the summation is what distinguishes this likelihood from QMLE – an explicit modelling of dependence between the composed errors at different \( t \).

In a GMM framework, the MLE that maximizes (4.7) is identical to the GMM estimator based on the moment conditions

\[
\mathbb{E} \left[ \nabla_\theta \ln c(\theta, \rho) + \nabla_\theta \ln f_{it}(\theta) + \ldots + \nabla_\theta \ln f_{iT}(\theta) \right] = 0, \tag{4.8}
\]
where \(c(\theta, \rho) = c(F_{it}(\theta), \ldots, F_{Tt}(\theta); \rho)\). Again, efficiency improvement is, in some circumstances, possible if we instead use the optimal GMM machinery on the moment conditions

\[
\mathbb{E} \left[ \begin{array}{c}
\nu_0 \ln f_{\theta_{11}}(\theta) \\
\nu_0 \ln f_{\theta_{1T}}(\theta) \\
\nu_0 \ln c_{\theta_{11}}(\theta, \rho) \\
\nu_0 \ln c_{\theta_{1T}}(\theta, \rho)
\end{array} \right] = 0.
\]

However, this improvement may now come at the price of a bias as the copula-based moment conditions may be misspecified causing inconsistency of GMM and offsetting any benefit of higher precision. So assuming a wrong kind of time dependence may be worse than assuming independence (over time). Prokhorov and Schmidt [70] explored these circumstances.

The alternative copula-based specification is to form the joint distribution of \((u_{t1}, \ldots, u_{Tt})\) rather than that of \((\varepsilon_{t1}, \ldots, \varepsilon_{Tt})\). A challenge of this specification is that a \(T\)-dimensional integration will be needed to form the likelihood in this case. Let \(f_u(u; \theta, \rho)\) denote the copula-based joint density of the one-sided error vector and let \(f_{\nu_{it}}(u; \theta)\) denote the marginal density of an individual one-sided (Half Normal) error term. Then,

\[
f_u(u; \theta, \rho) = c(F_{u}(u_{t1}; \theta), \ldots, F_{u}(u_{Tt}; \theta); \rho),
\]

where \(F_{U}(u; \theta) \equiv \int f_u(s; \theta)ds\) is the cdf of the Half Normal error term.

To form the sample likelihood we need the joint density of the composed error vector \(\varepsilon\). Given the density of \(u\) and assuming, as before, that \(v \perp u\), this density can be obtained as follows:

\[
f_{\varepsilon}(\varepsilon; \theta, \rho) = \int_{0}^{\infty} \cdots \int_{0}^{\infty} f_{\varepsilon}(\varepsilon + u; \theta)f_u(u; \theta, \rho)du_{t1} \cdots du_{Tt} = \mathbb{E}_{u_{\theta, \rho}}\phi(\varepsilon + u),
\]

where \(\mathbb{E}_{u_{\theta, \rho}}\) denotes the expectation with respect to \(f_u(u; \theta, \rho)\) and \(f_{\varepsilon}(v; \theta) = \phi(v)\) is the multivariate Normal pdf of \(v\), where all \(v\)'s are independent and have equal variance \(\sigma^2_v\).

Similar to the previous section, this integral has no analytical form. Additionally, this is a \(T\)-dimensional integral, which is computationally strenuous to evaluate using numerical methods. However, it has the form of an expectation over a distribution we can sample from and this, as before, permits application of MSLE, where we simulate the \(u\)'s and estimate \(f_{\varepsilon}(\varepsilon)\) by averaging \(\phi(\varepsilon + u)\) over the draws. To be precise, let \(S\) denote the number of simulations. The direct simulator of \(f_{\varepsilon}(\varepsilon)\) can be written as follows:

\[
\hat{f}_{\varepsilon}(\varepsilon; \theta, \rho) = \frac{1}{S} \sum_{i=1}^{S} \phi(\varepsilon + u_i(\theta, \rho)),
\]

where \(u_i(\theta, \rho)\) is a draw from \(f_u(u; \theta, \rho)\) constructed in (4.9). Then, a simulated log-likelihood can be obtained as follows:

\[
\ln L^s(\theta, \rho) = \sum_{i} \ln \hat{f}_{\varepsilon}(\varepsilon_i; \theta, \rho),
\]

where, as before, \(\varepsilon_i = (\varepsilon_{t1}, \ldots, \varepsilon_{Tt})\) and \(\varepsilon_{tit} = y_{tit} - x_{tit}\beta\).

This method is a multivariate extension to the simulation-based estimation of univariate densities discussed earlier. An important additional requirement is the ability to sample from the copula; see ([64], Ch. 2) for a discussion of how to sample from the copula to allow dependence. Other than that, similar asymptotic arguments suggest that MSLE is asymptotically equivalent to MLE [29].

5 Dependence due to endogeneity

A common assumption in the SFM is that \(x\) is either exogenous or independent of both \(u\) and \(v\). If either of these conditions are violated, then all of the estimators discussed so far will be biased and most likely
inconsistent. Yet, it is not difficult to think of settings where endogeneity is likely to exist. For example, if shocks are observed before inputs are chosen, then producers may respond to good or bad shocks by adjusting inputs, leading to correlation between \( x \) and \( v \). Alternatively, if managers know they are inefficient, they may use this information to guide their level of inputs, again, producing endogeneity. In a regression model, dealing with endogeneity is well understood. However, in the composed error setting, these methods cannot be simply transferred over, but require care in how they are implemented [6].

To incorporate endogeneity into the SFM in (2.1), we set \( m(x_i; \beta) = \beta_0 + x_i' \beta_1 + x_i' \beta_2 \), where \( x_i \) are our exogenous inputs, and \( x_i \) are the endogenous inputs, where endogeneity may arise through correlation of \( x_i \) with \( u \), \( v \), or both. To deal with endogeneity we require instruments, \( w \), and identification necessitates that the dimension of \( w \) is at least as large as the dimension of \( x_i \). The natural assumption for valid instrumentation is that \( w \) is independent of both \( u \) and \( v \).

Why worry about endogeneity? Economic endogeneity means that the inputs in question are choice variables and chosen to optimize some objective function such as cost minimization or profit maximization. Statistical endogeneity arises from simultaneity, omitted variables, and measurement errors. For example, if the omitted variable is managerial ability, which is part of inefficiency, inefficiency is likely to be correlated with inputs because managerial ability affects inputs. This is the Mundlak argument for why omitting a management quality variable (for us inefficiency) will cause biased parameter estimates. Endogeneity can also be caused by simultaneity meaning that more than one variable in the model are jointly determined. In many applied settings, it is not clear what researchers mean when they attempt to handle endogeneity inside the SFM. An excellent introduction into the myriad of influences that endogeneity can have on the estimates stemming from the SFM can be found in [63]. Mutter et al. [63] used simulations designed around data based on the California nursing home industry to understand the impact of endogeneity of nursing home quality on inefficiency measurement.

The simplest approach to accounting for endogeneity is to use a corrected two-stage least squares (C2SLS) approach, similar to the common correct ordinary least squares (COLS) approach that has been used to estimate the SFM. This method estimates the SFM using standard two-stage least squares (2SLS) with instruments \( w \). This produces consistent estimators for \( \beta_1 \) and \( \beta_2 \), but not \( \beta_0 \), as this is obscured by the presence of \( E[u] \) (to ensure that the residuals have mean zero). The second and third moments of the 2SLS residuals are then used to recover estimators of \( \sigma_u^2 \) and \( \sigma_v^2 \). Once \( \sigma_u^2 \) is determined, the intercept can be corrected by adding \( \frac{\sigma_u^2}{n} \). See Section 4.1 of [6] for details.

This represents a simple avenue to account for endogeneity, and it does not require specifying how endogeneity enters the model, i.e., through correlation with \( v \), with \( u \) or both. However, as with other corrected procedures based on calculations of the second and third (or even higher) moments of the residuals, from [66] and [89], if the initial 2SLS residuals have positive skew (instead of negative), then \( \sigma_u^2 \) cannot be identified and its estimator is 0. Furthermore, the standard errors from this approach need to be modified for the estimator of the intercept to account for the step-wise nature of the estimation.

### 5.1 A likelihood framework

Likelihood-based alternatives allow for explicit modelling and estimation of the dependence structure that underlies endogeneity. This has recently been studied by Kutlu [54], Karakaplan and Kutlu [44], Tran and Tsionsas [87,88], and Amsler et al. [6]. Our discussion here follows [6] as their derivation of the likelihood relies on a simple conditioning argument as opposed to the earlier work relying on the Cholesky decomposition or alternative approaches. While all approaches lead to a likelihood function, the conditioning idea of Amsler et al. [6] is simpler and more intuitive.

Consider the stochastic frontier system:

\[
y_i = x_i \beta + \epsilon_i
\]

\[
x_i = w_i \Gamma + \eta_i,
\]

where
where \( \mathbf{x}_i = (x_{i1}, x_{i2}) \), \( \mathbf{\beta} = (\beta_1, \beta_2) \), \( \mathbf{w}_i = (x_{i3}, \mathbf{q}_i) \) is the vector of instruments, \( \eta_i \) (different from \( \eta_i \) in Sections 3.1–3.2) is uncorrelated with \( \mathbf{w}_i \) and endogeneity of \( x_{i2} \) arises through \( \text{cov}(\varepsilon_i, \eta_i) \neq 0 \). Here simultaneity bias (and the resulting inconsistency) exists because \( \eta_i \) is correlated with either \( v_i, u_i \), or both.

We start with the case of dependence between \( \eta_i \) and \( v_i \) while \( u_i \) is independent of \( (\mathbf{\eta}_i, \mathbf{v}_i) \). Assume that, conditional on \( \mathbf{w}_i \), \( \mathbf{\psi}_i = (\mathbf{v}_i, \mathbf{\eta}_i)' \sim N(\mathbf{0}, \Omega) \), where

\[
\Omega = \begin{bmatrix}
\sigma_v^2 & \Sigma_{\mathbf{\eta} \mathbf{\eta}} \\
\Sigma_{\mathbf{\eta} \mathbf{\eta}} & \Sigma_{\mathbf{\eta} \mathbf{\eta}} \mathbf{\eta}_i
\end{bmatrix}.
\]

To derive the likelihood function, [6] condition on the instruments, \( \mathbf{w} \). Doing this yields \( f(y|\mathbf{x}_i, \mathbf{w}) = f(y|\mathbf{x}_i, \mathbf{w}) \cdot f(\mathbf{x}_i|\mathbf{w}) \). With the density in this form, the log-likelihood follows suite: \( \ln \mathcal{L} = \ln \mathcal{L}_1 + \ln \mathcal{L}_2 \), where \( \ln \mathcal{L}_1 \) corresponds to \( f(y|\mathbf{x}_i, \mathbf{w}) \) and \( \ln \mathcal{L}_2 \) corresponds to \( f(\mathbf{x}_i|\mathbf{w}) \). These two components can be written as

\[
\ln \mathcal{L}_1 = -(n/2) \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mathbf{e}_i' \mathbf{e}_i + \sum_{i=1}^{n} \ln \Phi(-\lambda_i, \mathbf{e}_i / \sigma)
\]

\[
\ln \mathcal{L}_2 = -(n/2) \ln |\Sigma_{\mathbf{\eta} \mathbf{\eta}}| - 0.5 \sum_{i=1}^{n} \eta_i' \Sigma_{\mathbf{\eta} \mathbf{\eta}}^{-1} \eta_i,
\]

where \( \mathbf{e}_i = y_i - \beta_0 - \mathbf{x}_i \mathbf{\beta} - \mathbf{\mu}_i, \mathbf{\mu}_i = \Sigma_{\mathbf{\eta} \mathbf{\eta}}^{-1} \mathbf{\eta}_i, \sigma^2 = \sigma_u^2 + \sigma_c^2, \lambda_c = \sigma_u / \sigma_c, \) and \( \sigma_v^2 = \sigma_v^2 - \Sigma_{\mathbf{\eta} \mathbf{\eta}}^{-1} \Sigma_{\mathbf{\eta} \mathbf{\eta}} \mathbf{\eta}_i \). The subtraction of \( \mathbf{\mu}_i \) in \( \ln \mathcal{L}_1 \) is an endogeneity correction while it should be noted that \( \ln \mathcal{L}_2 \) is nothing more than the standard likelihood function of a multivariate normal regression model (as in (5.2)). Estimates of the model parameters \( (\mathbf{\beta}, \sigma_u^2, \sigma_c^2, \mathbf{\Gamma}, \Sigma_{\mathbf{\eta} \mathbf{\eta}}) \) and \( \Sigma_{\mathbf{\eta} \mathbf{\eta}} \) can be obtained by maximizing \( \ln \mathcal{L} \).

While direct estimation of the likelihood function is possible, a two-step approach is also available [54]. However, as pointed out by both Kutlu [54] and Amsler et al. [6], this two-step approach will have incorrect standard errors. Even though the two-step approach might be computationally simpler, it is, in general, different from full optimization of the likelihood function of Amsler et al. [6]. This is due to the fact that the two-step approach ignores the information provided by \( \mathbf{\Gamma} \) and \( \Sigma_{\mathbf{\eta} \mathbf{\eta}} \) in \( \ln \mathcal{L}_1 \). In general, full optimization of the likelihood function is recommended as the standard errors (obtained in a usual manner from the inverse of the Fisher information matrix) are valid.\(^7\)

5.2 A GMM framework

An insightful avenue to model dependence due to endogeneity in the SFM that differs from the traditional corrected methods or maximum likelihood stems from the GMM framework as proposed by Amsler et al. [6], who used the insights of Hansen et al. [35]. Similar to our discussion on the use of GMM in panel estimation, the idea is to use the first-order conditions for maximization of the likelihood function under exogeneity as a GMM problem:

\[
E[\mathbf{e}_i^2 / \sigma^2 - 1] = 0,
\]

\(^7\) Typically, the standard errors can be obtained either through the use of the outer product of gradients (OPG) or direct estimation of the Hessian matrix of the log-likelihood function. Given the nascently of these methods it has yet to be determined which of these two methods is more reliable in practice, though in other settings both tend to work well. One caveat for promoting the use of the OPG is that since this only requires calculation of the first derivatives, it can be more stable (and more likely to be invertible) than calculation of the Hessian. Also note that in finite samples, the different estimators of covariance of MLE estimator can give different numerical estimates, even suggesting different implications on the inference (reject or do not reject the null hypothesis). So, for small samples, it is often advised to check all feasible estimates whenever there is suspicion of ambiguity in the conclusions (e.g., when a hypothesis is rejected only at say around the 10% of significance level).
\[ E \left[ \frac{\varepsilon_i \phi_i}{1 - \Phi_i} \right] = 0, \quad (5.4) \]
\[ E \left[ x_i \varepsilon_i / \sigma + \lambda x_i \frac{\phi_i}{1 - \Phi_i} \right] = 0, \quad (5.5) \]

where \( \phi_i = \phi \left( \frac{u_i}{\sigma} \right) \) and \( \Phi_i = \Phi \left( \frac{u_i}{\sigma} \right) \). Note that these expectations are taken over \( x_i \) and \( y_i \) (and by default, \( \varepsilon_i \)) and solved for the parameters of the SFM.

The key here is that these first-order conditions (one for \( \sigma^2 \), one for \( \lambda \), and the vector for \( \beta \)) are valid under exogeneity and this implies that the MLE is equivalent to the GMM estimator. Under endogeneity however, this relationship does not hold directly. But the seminal idea of Amsler et al. [6] is that the first-order conditions (5.3) and (5.4) are based on the distributional assumptions on \( v \) and \( u \), not on the relationship of \( x \) with \( v \) and/or \( u \). Thus, these moment conditions are valid whether \( x \) contains endogenous components or not. The only moment condition that needs to be adjusted is (5.5). In this case, the first-order condition needs to be taken with respect to \( w \), the exogenous variable, not \( x \). Doing so results in the following amended first-order condition:
\[ E \left[ w_i \varepsilon_i / \sigma + \lambda w_i \frac{\phi_i}{1 - \Phi_i} \right] = 0, \quad (5.6) \]

where \( \phi_i \) and \( \Phi_i \) are identical to those in (5.5). It is important to acknowledge that this moment condition is valid when \( \varepsilon_i \) and \( w_i \) are independent. This is a more stringent requirement than the typical regression setup with \( E[\varepsilon_i | w_i] = 0 \). As with the C2SLS approach, the source of endogeneity for \( x_i \) does not need to be specified (through \( v \) and/or \( u \)).

### 5.3 An economic model of dependence

From an economic theory perspective, there are grounds to model dependence between \( u_i \) and \( \eta_i \), not only between \( v_i \) and \( \eta_i \). A system similar to (5.1)–(5.2) arises as a result of appending the SFM with the first-order conditions of cost minimization\(^8\) ([52], Chapter 8):
\[ \min P'x, \quad \text{s.t. } y = m(x; \beta) + v - u, \quad (5.7) \]

for input prices \( P \), the first-order conditions in this case are
\[ \frac{m_j(x; \beta)}{m(x; \beta)} = \frac{P_j}{P_1}, \quad j = 2, \ldots, J, \quad (5.8) \]

where \( m_j(x; \beta) \) is the partial derivative of \( m(x; \beta) \) with respect to \( x_j \). These first-order conditions are exact, which usually does not arise in practice, rather, a stochastic term is added which is designed to capture allocative inefficiency. That is, our empirical first-order conditions are
\[ \frac{m_j(x; \beta)}{m(x; \beta)} = \frac{P_j}{P_1} \eta_{ij} \quad \text{for } j = 2, \ldots, J, \]
where \( \eta_{ij} \) captures allocative inefficiency for the \( j \)th input relative to input one (the choice of input to compare is without loss of generality).

The idea behind allocative inefficiency is that firms could be fully technically efficient, and still have room for improvement due to over or under use of inputs, relative to another input, given the price ratio. On the other hand, firms can be technically inefficient because of allocative inefficiency and vice versa so independence between \( u \) and \( \eta \) is hard to justify. Additionally, if firms are cost minimizers and one

\(^{8}\) It is possible to treat a subset of \( x \) as endogenous, i.e., \( x = (x_1, x_2) \), where \( x_1 \) is endogenous and \( x_2 \) is exogenous.
estimates a production function, the inputs will be endogenous as these are choice variables to the firm. In this case, input prices can serve as instruments.

Combining the SFM, under the Cobb–Douglas production function, with the information in the \( J - 1 \) conditions in (5.8) with allocative inefficiency built in, results in the following system:

\[
y_i = x_i \beta + \epsilon_i \tag{5.9}
\]
\[
x_{it} - x_{i0} = \ln(\beta_1) - \ln(\beta_j) + p_j - p_{i0} + \eta_{ij}, \quad j = 2, \ldots, J, \tag{5.10}
\]

where \( x_{ij} \) is the log of input \( j \) of firm \( i \), \( p_j \) is the log of input \( j \) price, \( \beta_j \) is the coefficient on input \( j \) in (5.9), and \( \eta = (\eta_{12}, \ldots, \eta_{1j}) \) are the allocative inefficiencies for \( J - 1 \) inputs with respect to input one. See Schmidt and Lovell [74,75] for details.

5.4 A copula-based approach

Amsler et al. [6] used copulas to obtain a joint distribution for \( u, v, \) and \( \eta \), whereas Amsler et al. [8] developed a new copula family for \( u \) and \( \eta \) with properties that reflect the nature of allocative (symmetric) and technical (one-sided) inefficiencies. Here we provide the derivation of a copula-based likelihood for the most general case that allows us to model dependence between all the components of \( (u, v, \eta) \).

We keep the Half Normal marginal for \( u \), Normal marginals for the elements of \( \psi = (v, \eta')' \) as before, and assume a copula density \( c(\cdot, \cdot, \cdot) \). Amsler et al. [6] used the Gaussian copula, which implies that the joint distribution of \( \psi \) is Normal but this is largely done for convenience. This gives the joint density of \( (u, v, \eta) \):

\[
f_{u,v,\eta}(u, v, \eta) = c(F_u(u), F_v(v), F_\eta(\eta_2), \ldots, F_\eta(\eta_J))f_u(u)f_v(v)f_\eta(\eta_2)\cdots f_\eta(\eta_J).
\]

However, we need the joint density of \( \epsilon \) and \( \eta \) in order to form a sample log-likelihood. This density can be obtained by integrating \( u \) out of \( f_{u,v,\eta}(u, v, \eta) \) as follows:

\[
f_{\epsilon,\eta}(\epsilon, \eta) = \int_0^\infty f_{u,v,\eta}(u, \epsilon + u, \eta)du = \int_0^\infty \left[ \int \frac{f_{u,v,\eta}(u, \epsilon + u, \eta)}{f_u(u)} f_u(u)du \right] = \mathbb{E}_u \left[ \int \frac{f_{u,v,\eta}(u, \epsilon + u, \eta)}{f_u(u)} f_u(u)du \right]. \tag{5.11}
\]

Again, we can use simulation techniques to evaluate this density. Specifically, given \( S \) draws of \( u_s \), \( s = 1, \ldots, S \), the direct simulator can be written as

\[
\hat{f}_{\epsilon,\eta}(\epsilon, \eta) = \frac{1}{S} \sum_{s=1}^S \frac{f_{u,v,\eta}(u_s, \epsilon + u_s, \eta)}{f_u(u_s)},
\]

and this leads to MSLE using the log-likelihood

\[
\ln L^\pm = \sum_{i=1}^n \ln \hat{f}_{\epsilon,\eta}(\epsilon_i, \eta_i), \tag{5.12}
\]

where \( \epsilon_i = y_i - x_i'\beta \) and \( \eta_i = x_{i0} - w_i'\Gamma \) as follows from the system in (5.1)–(5.2).

The MSLE will produce estimates of all the parameters of the model, that is, \( \beta, \Gamma, \sigma^2, \alpha^2 \), variances of \( \eta \) and whatever copula parameters appear in \( c \). This permits modelling and testing the validity of independence assumptions between all error terms in the system including the assumption of exogeneity.

5.5 Dependence on determinants of inefficiency

To conclude this section, we consider the extension to a setting when inefficiency depends on covariates and some of these determinants of inefficiency may be endogenous [7,55]. These models can be estimated
using traditional instrumental variable methods. However, given that the determinants of inefficiency enter the model nonlinearly, nonlinear methods are required.

Amsler et al. [7] considered the model

\[ y_i = x_i' \beta + v_i - u_i = x_i' \beta + v_i - u_i' e^{z_i' \delta}, \]  

(5.13)

where \( u' \) is the baseline inefficiency and \( u \) has the property that the scale of its distribution (relative to the distribution of \( u' \)) changes depending on the determinants \( z \) (the so-called scaling property). The covariates \( x_i \) and \( z_i \) are partitioned as

\[ x_i = \begin{bmatrix} x_{ii} \\ x_{ji} \end{bmatrix}, \quad z_i = \begin{bmatrix} z_{ii} \\ z_{ji} \end{bmatrix}, \]

where \( x_{ii} \) and \( z_{ii} \) are exogenous and \( x_{ji} \) and \( z_{ji} \) are endogenous. The set of instruments used to combat endogeneity are defined as

\[ w_i = \begin{bmatrix} x_{ii} \\ z_{ii} \\ q_i \end{bmatrix}, \]

where \( q_i \) are the traditional outside instruments. Identification of all the parameters requires that the dimension of \( q_i \) be at least as large as the dimension of \( x_i \) plus the dimension of \( z_i \) (the rank condition).

In the model of Amsler et al. [7], endogeneity arises through dependence between a variable in the model (\( x_i \) and/or \( z_i \)) and noise, \( v \). That is, both \( x \) and \( z \) are assumed to be independent of baseline inefficiency \( u' \). Given that \( E[u_i] \) is not constant, the COLS approach to deal with endogeneity proposed by Amsler et al. [6] cannot be used here. To develop an appropriate estimator, add and subtract the mean of inefficiency to produce a composed error term that has mean 0,

\[ y_i = x_i' \beta - \mu' e^{z_i' \delta} + v_i - (u_i' - \mu') e^{z_i' \delta}. \]  

(5.14)

Proper estimation through instrumental variables requires that the following moment condition holds

\[ E[v_i - (u_i' - \mu') e^{z_i' \delta} | w_i] = 0. \]  

(5.15)

The nonlinearity of these moment conditions would necessitate use of nonlinear two-stage least squares (NL2SLS) [4].

Latruffe et al. [55] have a similar setup to Amsler et al. [7], using the model in (5.13), but develop a four-stage estimator for the parameters; additionally, only \( x_i \) is treated as endogenous. Latruffe et al.’s [55] approach is based on [23] using the construction of efficient moment conditions. The vector of instruments proposed in [55] is defined as

\[ w_i(y, \delta) = \begin{bmatrix} x_{ii} \\ q_i' y \\ z e^{z_i' \delta} \end{bmatrix}, \]  

(5.16)

where \( q_i' y \) captures the linear projection of \( x_i \) on the external instruments \( q_i \). The four-stage estimator is defined as

Step 1 Regress \( x_i \) on \( q_i \) to estimate \( y \). Denote the OLS estimator of \( y \) as \( \hat{y} \).

Step 2 Use NLS to estimate the SFM in (5.13). Denote the NLS estimates of \( (\beta, \delta) \) as \( (\hat{\beta}, \hat{\delta}) \). Use the NLS estimate of \( \delta \) and the OLS estimate of \( y \) in Step 1 to construct the instruments \( w_i(\hat{y}, \hat{\delta}) \).

Step 3 Using the estimated instrument vector \( w_i(\hat{y}, \hat{\delta}) \), calculate the NL2SLS estimator of \( (\beta, \delta) \) as \( (\tilde{\beta}, \tilde{\delta}) \). Use the NL2SLS estimate of \( \delta \) and the OLS estimate of \( y \) in Step 1 to construct the instruments \( w_i(\tilde{y}, \tilde{\delta}) \).

Step 4 Using the estimated instrument vector \( w_i(\tilde{y}, \tilde{\delta}) \), calculate the NL2SLS estimator of \( (\beta, \delta) \) as \( (\hat{\beta}, \hat{\delta}) \).
This multi-step estimator is necessary in the context of efficient moments because the actual set of instruments is not used directly, rather \( w(\gamma, \delta) \) is used, and this instrument vector requires estimates of \( \gamma \) and \( \delta \). The first two steps of the algorithm are designed to construct estimates of these two unknown parameter vectors. The third step then is designed to construct a consistent estimator of \( w(\gamma, \delta) \), which is not done in Step 2 given that the endogeneity of \( \epsilon_i \) is ignored (note that NLS is used as opposed to NL2SLS). The iteration from Step 2 to Step 3 does produce a consistent estimator of \( w(\gamma, \delta) \), and as such, Step 4 produces consistent estimators for \( \beta \) and \( \delta \). While Latruffe et al. [55] proposed a set of efficient moment conditions to handle endogeneity, the model of Amsler et al. [7] is more general because it can handle endogeneity in the determinants of inefficiency as well. Finally, the presence of \( z \) is attractive since this allows the researcher to dispense with distributional assumptions on \( v \) and \( u \).

6 Estimation of individual inefficiency using dependence information

Once the parameters of the SFM have been estimated, estimates of firm level productivity and efficiency can be recovered. Observation-specific estimates of inefficiency are one of the main benefits of the SFM relative to neoclassical models of production. Firms can be ranked according to estimated efficiency; the identity of under-performing firms as well as those who are deemed best practice can also be gleaned from the estimated SFM. All of this information is useful in helping to design more efficient public policy or subsidy programs aimed at improving the market, for example, insulating consumers from the poor performance of heavily inefficient firms.

As a concrete illustration, consider firms operating electricity distribution networks that typically possess a natural local monopoly given that the construction of competing networks over the same terrain is prohibitively expensive.\(^9\) It is not uncommon for national governments to establish regulatory agencies which monitor the provision of electricity to ensure that abuse of the inherent monopoly power is not occurring. Regulators face the task of determining an acceptable price for the provision of electricity while having to balance the heterogeneity that exists across the firms (in terms of size of the firm and length of the network). Firms which are inefficient may charge too high a price to recoup a profit, but at the expense of operating below capacity. However, given production and distribution shocks, not all departures from the frontier represent inefficiency. Thus, precise measures designed to account for noise are required to parse information from \( e_i \) regarding \( u_i \).

Alternatively, further investigation could reveal what it is that makes these establishments attain such high levels of performance. This could then be used to identify appropriate government policy implications and responses or identify processes and/or management practices that should be spread (or encouraged) across the less efficient, but otherwise similar, units. This is the essence of the determinants of inefficiency approach discussed in previous section. More directly, efficiency rankings are used in regulated industries such that regulators can set tougher future cost reduction targets for the more inefficient companies, in order to ensure that customers do not pay for the inefficiency of firms.

The only direct estimate coming from the Normal Half Normal SFM is \( \tilde{\sigma}_U^2 \). This provides context regarding the shape of the Half Normal distribution on \( u_i \) and the industry average efficiency \( E[u] \), but not on the absolute level of inefficiency for a given firm. If we are only concerned with the average level of technical efficiency for the population, then this is all the information that is needed. Yet, if we want to know about a specific firm, then something else is required. The main approach to estimating firm-level

---

\(^9\) The SFA literature contains a fairly rich set of examples for the estimation and use of efficiency estimates in different fields of research. For example, in the context of electricity providers, see [36,45,53]; for banking efficiency, see [22] and references cited therein; for the analysis of the efficiency of national health care systems, see [30] and the review [40]; for analyzing efficiency in agriculture, see [13,14,20,58], to mention just a few.
inefficiency is the conditional mean estimator [42], commonly known as the JLMS estimator. Their idea was to calculate the expected value of $u_t$ conditional on the realization of composed error of the model, $e_t \equiv v_t - u_t$, i.e., $E[u_t|e_t]$. This conditional mean of $u_t$ given $e_t$ gives a point prediction of $u_t$. The composed error contains individual-specific information, and the conditional expectation is one measure of firm-specific inefficiency.

JLMS [42] shows that for the Normal Half Normal specification of the SFM, the conditional density function of $u_t$ given $e_t$, $f(u_t|e_t)$, is $N(\mu_i, \sigma_i^2)$, where

$$\mu_i = \frac{-\epsilon_i \sigma_u^2}{\sigma^2}$$  \hspace{1cm} (6.1)

and

$$\sigma_i^2 = \frac{\sigma_u^2 \sigma_v^2}{\sigma^2}.$$  \hspace{1cm} (6.2)

Given results on the mean of a Truncated Normal density it follows that

$$E[u_t|e_t] = \mu_i + \frac{\sigma_i \Phi(\frac{\mu_i}{\sigma_i})}{\Phi(\frac{\mu_i}{\sigma})}.$$  \hspace{1cm} (6.3)

The individual estimates are then obtained by replacing the true parameters in (6.3) with MLE (or MSMLE or GMM) estimates from the SFM.

Another measure of interest is the Afriat-type level of technical efficiency, defined as $e^{-u_t} = Y_t / e^{\mu_i + \sigma_i v_t}$ $\in [0, 1]$. This is useful in cases where output is measured in logarithmic form. Furthermore, technical efficiency is bounded between 0 and 1, making it somewhat easier to interpret relative to a raw inefficiency score. Since $e^{-u_t}$ is not directly observable, the idea of JLMS [42] can be deployed here, and $E[e^{-u_t}|e_t]$ can be calculated [12,56]. For the Normal Half Normal model, we have

$$E[e^{-u_t}|e_t] = e^{(-\mu_i + \sigma_i^2)} \frac{\Phi(\frac{\mu_i}{\sigma_i} - \sigma_i)}{\Phi(\frac{\mu_i}{\sigma})},$$  \hspace{1cm} (6.4)

where $\mu_i$ and $\sigma_i$ were defined in (6.1) and (6.2), respectively. Technical efficiency estimates are obtained by replacing the true parameters in (6.4) with MLE estimates from the SFM. When ranking efficiency scores, one should use estimates of $1 - E[u_t|e_t]$, which is the first-order approximation of (6.4). Similar expressions for the JMLS [42] and Battese and Coelli [12] efficiency scores can be derived under the assumption that $u$ is Exponential [49, p. 82], Truncated Normal [49, p. 86], and Gamma [49, p. 89]; see also [52].

An interesting and important finding from [5] and [6] is that when we allow for dependence of the kinds described in Sections 4 and 5, we can potentially improve estimation of inefficiency through the JLMS estimator. We focus on the case of endogeneity (Section 5) but the case of dependence over $t$ in panels (Section 4) is similar. The traditional predictor [42] is $\mathbb{E}(u_t|e_t)$. However, more information is available when dependence is allowed, namely via $\eta_t$. This calls for a modified JLMS estimator, $\mathbb{E}(u_t|e_t, \eta_t)$. Note that even though it is assumed that $u_t$ is independent from $\eta_t$, similar to [6], because $\eta_t$ is correlated with $v_t$, there is information that can be used to help predict $u_t$ even after conditioning on $e_t$.

Amsler et al. [6] showed that $\eta_t$ is independent of $(u_t, \tilde{e}_t)$:

$$\mathbb{E}(u_t|e_t, \eta_t) = \mathbb{E}(u_t|\tilde{e}_t, \eta_t) = \mathbb{E}(u_t|\tilde{e}_t),$$

and that the distribution of $u_t$ conditional on $\tilde{e}_t = y_t - \mathbf{x}_t^\prime \beta - \mu_i$ is $N(\mu_i, \sigma_i^2)$ with $\mu_i = -\sigma_i^2 \tilde{e}_t / \sigma^2$ and $\sigma_i^2 = \sigma_u^2 \sigma_v^2 / \sigma^2$, which is identical to the original JLMS estimator, except that $\sigma_i^2$ is replaced with $\sigma_i^2$ and $\tilde{e}_t$ taking the place of $e_t$. The modified JLMS estimator in the presence of endogeneity becomes

[10] JLMS [42] also suggested an alternative estimator based on the conditional mode.
\[ E(u_i | e_i, \eta_i) = \sigma \left( \frac{\phi(L_i)}{1 - \phi(L_i)} - \xi \right) \] with \( \xi = \lambda \tilde{e}_i / \sigma \). Note that \( E(u_i | e_i, \eta_i) \) is a better predictor than \( E(u_i | e_i) \) because \( \sigma^2 < \sigma_o^2 \). The improvement in prediction follows from the textbook identity for variances, where for any random vector \((X, Z)\), where \(X\) and \(Z\) are random sub-vectors, we have
\[
\mathbb{V}(X) = \mathbb{V}[E(X|Z)] + E(\mathbb{V}[X|Z]).
\]
In this case, by conditioning on both \(e_i\) and \(\eta_i\) the conditioning set is larger than simply conditioning on \(e_i\) and so it must hold that the unexplained portion of \(E(u_i | e_i, \eta_i)\) is smaller than that of \(E(u_i | e_i)\). It then holds that there is less variation in \(E(u_i | e_i, \eta_i)\) as a predictor than \(E(u_i | e_i)\), which is a good thing. A similar result is obtained by [5] in a panel setting, where the new estimator \(E(u_i | e_{i1}, \ldots, e_{iT})\) dominates the traditional estimator \(E(u_i | e_i)\) due to dependence over \(t\). While it is not obvious at first glance, one benefit of allowing for a richer dependence structure in SFM is that researchers may be able to more accurately predict firm-level inefficiency, though it comes at the expense of having to deal with a more complex model. This improvement in prediction may also be accompanied by narrower prediction intervals; however, this is not known as Amsler et al. [6] did not study the prediction intervals.

A prediction interval for \(E[u_i | e_i]\) was first derived by Taube [85] and also appeared in [39], [41], and [17] (see a discussion of this in [77]). The prediction interval is based on \(f(u_i | e_i)\). The lower \((L_i)\) and upper \((U_i)\) bounds for a \((1 - \alpha)\)100\% prediction interval are
\[
L_i = \mu_i + \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \left[ 1 - \Phi\left( -\frac{\mu_i}{\sigma_i} \right) \right] \sigma_i,
\]
\[
U_i = \mu_i + \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \left[ 1 - \Phi\left( -\frac{\mu_i}{\sigma_i} \right) \right] \sigma_i,
\]
where \(\mu_i\) and \(\sigma_i\) are defined in (6.1) and (6.2), respectively, and replacing them with their MLE estimates will give estimated prediction intervals for \(E[u_i | e_i]\). Using the above result from [6] that \(u_i\) conditional on \(\tilde{e}_i\) is \(N(\mu_i, \sigma_i^2)\) with \(\mu_i = -\alpha_i^2 \tilde{e}_i / \sigma^2\) and \(\sigma_i^2 = \alpha_i^2 \alpha_i^2 / \sigma^2\), one can easily obtain analogous prediction intervals for \(E(u_i | e_i, \eta_i)\). These new intervals will potentially be narrower.

7 Conclusion

In this article, we surveyed the workhorse SFM and various recent extensions that permit a wide range of dependence modelling within SFM. We discussed dependencies that arise in panels and that underpin endogeneity in production systems. Copulas play a key role in these settings because they naturally permit the construction of a likelihood, often based on simulated draws, while preserving the desired Half Normal distribution of technical inefficiency.

While this is not a survey of SFA applications, it is worth pointing out that SFA has become a popular tool to isolate inefficiency in the behavior of economic agents, e.g., banks and non-financial firms. As just a couple of examples, Koetter et al. [46] showed that ignoring inefficiency, i.e., assuming all banks in the US economy are on the frontier, leads to substantial downward bias in the estimates of the banks’ market power, as measured by the Lerner index of price mark-ups; Henry et al. [38] applied SFA to obtain a correct estimate of TFP at the country level in a panel of 57 developing economies over the 1970–1998 period. The list of applications is long and growing.

This survey’s goal was to provide a comprehensive overview of the state of the art in methods of dependence modeling (either directly with noise, through endogeneity, through sample selection or across time), but it is clear that many important issues still remain and this is an active area of research for the field. We remain excited about the potential developments in this area and the insights that they can shed in
applications. At present there is little cognizance into the direction of impact of unmodelled, or misspecified, dependence on efficiency scores.

Acknowledgments: Helpful comments from two anonymous referees, the Editor, as well as from Robert James, are gratefully acknowledged. Mamonov acknowledges the financial support by the MGIMO University, research grant number 1921-01-08. Prokhorov’s research for this article was supported by a grant from the Russian Science Foundation (Project No. 20-78-10113).

Conflict of interest: Artem Prokhorov is a member of the Editorial Advisory Board for this journal but had no involvement in the review of this manuscript or the final editorial decision.

References

[1] Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review*, 13(3), 568–598.
[2] Aigner, D., & Chu, S. (1968). On estimating the industry production function. *American Economic Review*, 58, 826–839.
[3] Aigner, D. J., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production functions. *Journal of Economics*, 61(1), 21–37.
[4] Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2, 105–111.
[5] Amsler, C., Prokhorov, A., & Schmidt, P. (2014). Using copulas to model time dependence in stochastic frontier models. *Econometric Reviews*, 33(5–6), 497–522.
[6] Amsler, C., Prokhorov, A., & Schmidt, P. (2016). Endogeneity in stochastic frontier models. *Journal of Econometrics*, 190, 280–288.
[7] Amsler, C., Prokhorov, A., & Schmidt, P. (2017). Endogeneity environmental variables in stochastic frontier models. *Journal of Economics*, 199, 131–140.
[8] Amsler, C., Prokhorov, A., & Schmidt, P. (2021). A new family of copulas, with application to estimation of a production frontier system. *Journal of Productivity Analysis*, 55, 1–4.
[9] Amsler, C., & Schmidt, P. (2021). A survey of the use of copulas in stochastic frontier models. In C. F. Parmeter & R. C. Sickles (Eds.), *Advances in efficiency and productivity analysis* (pp. 125–138). Cham, Switzerland: Springer Nature.
[10] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2), 171–178.
[11] Badunenko, O., & Kumbhakar, S. C. (2017). Economies of scale, technical change and persistent and time-varying cost efficiency in Indian banking: Do ownership, regulation and heterogeneity matter? *European Journal of Operational Research*, 260, 789–803.
[12] Battese, G. E., & Coelli, T. J. (1988). Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. *Journal of Econometrics*, 38, 387–399.
[13] Battese, G. E., & Coelli, T. J. (1992). Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India. *Journal of Productivity Analysis*, 3, 153–169.
[14] Battese, G. E., & Coelli, T. J. (1995). A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, 20(1), 325–332.
[15] Battese, G. E., & Corra, G. S. (1977). Estimation of a production frontier model: With application to the pastoral zone off Eastern Australia. *Australian Journal of Agricultural Economics*, 21(3), 169–179.
[16] Benabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3), 141–164.
[17] Bera, A. K., & Sharma, S. C. (1999). Estimating production uncertainty in stochastic frontier production function models. *Journal of Productivity Analysis*, 12(2), 187–210.
[18] Bloom, N., Lemos, R., Sadun, R., Scur, D., & Van Reenen, J. (2016). International data on measuring management practices. *American Economic Review*, 106(5), 152–156.
[19] Bonanno, G., De Giovannni, D., & Domma, F. (2015). The “wrong skewness” problem: A re-specification of stochastic frontiers. *Journal of Productivity Analysis*, 47(1), 49–64.
[20] Bravo-Ureta, B. E., & Rieger, L. (1991). Dairy farm efficiency measurement using stochastic frontiers and neoclassical duality. *American Journal of Agricultural Economics*, 73(2), 421–428.
[21] Burns, R. (2004). *The simulated maximum likelihood estimation of stochastic frontier models with correlated error components*. Sydney, Australia: The University of Sydney.
[22] Case, B., Ferrari, A., & Zhao, T. (2013). Regulatory reform and productivity change in indian banking. *The Review of Economics and Statistics*, 95(3), 1066–1077.
[23] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(2), 305–334.

[24] Chen, Y.-Y., Schmidt, P., & Wang, H.-J. (2014). Consistent estimation of the fixed effects stochastic frontier model. *Journal of Econometrics*, 181(2), 65–76.

[25] Coelli, T. J. (1995). Estimators and hypothesis tests for a stochastic frontier function: A Monte Carlo analysis. *Journal of Productivity Analysis*, 6(4), 247–268.

[26] Cornwell, C., Schmidt, P., & Sickles, R. C. (1990). Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics*, 46(2), 185–200.

[27] Dugger, R. (1974). An application of bounded nonparametric estimating functions to the analysis of bank cost and production functions. Ph.D. thesis. University of North Carolina, Chapel Hill.

[28] ElMehdi, R., & Hafner, M. (2014). Inference in stochastic frontier analysis with dependent error terms. *Mathematics and Computers in Simulation*, 102, 104–116.

[29] Gouriéroux, C., & Monfort, A. (1991). Simulation based inference in models with heterogeneity. *Annales de l’Économie et de Statistique*, 20/21, 69–107.

[30] Greene, W. (2004). Distinguishing between heterogeneity and inefficiency: Stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics*, 13(9), 959–980.

[31] Greene, W. H. (2005a). Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis*, 23(1), 7–32.

[32] Greene, W. H. (2005b). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, 126(2), 269–303.

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

[34] Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2), 23–48.

[35] Hansen, C., McDonald, J. B., & Newey, W. K. (2010). Instrumental variables estimation with flexible distributions. *Journal of Business and Economic Statistics*, 28, 13–25.

[36] Hattori, T. (2002). Relative performance of U.S. and Japanese electricity distribution: An application of stochastic frontier analysis. *Journal of Productivity Analysis*, 18(3), 269–284.

[37] Heckman, J. J. (1976). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.

[38] Henry, M., Kneller, R., & Milner, C. (2009). Trade, technology transfer and national efficiency in developing countries. *European Economic Review*, 53(2), 237–254.

[39] Hjalmarsson, L., Kumbhakar, S. C., & Heshmati, A. (1996). DEA, DFA, and SFA: A comparison. *Journal of Productivity Analysis*, 7(2), 303–327.

[40] Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, 17(10), 1107–1128.

[41] Horrace, W. C., & Schmidt, P. (1996). Confidence statements for efficiency estimates from stochastic frontier models. *Journal of Productivity Analysis*, 7, 257–282.

[42] Jondrow, J., Lovell, C. A. K., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical efficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19/2(3), 233–238.

[43] Kantorovich, L. (1939). *Mathematical methods of organizing and planning production*. Leningrad: Publishing House of Leningrad State University.

[44] Karakaplan, M. U., & Kutlu, L. (2013). *Handling endogeneity in stochastic frontier analysis*. Unpublished manuscript.

[45] Knittel, C. R. (2002). Alternative regulatory methods and firm efficiency: Stochastic frontier evidence from the U.S. electricity industry. *The Review of Economics and Statistics*, 84(3), 530–540.

[46] Koetter, M., Kolari, J. W., & Spierdijk, L. (2012). Enjoying the Quiet Life under Deregulation? Evidence from Adjusted Lerner Indices for U.S. Banks. *The Review of Economics and Statistics*, 94(2), 462–480.

[47] Kumbhakar, S. C. (1990). Production frontiers, panel data, and time-varying technical inefficiency. *Journal of Econometrics*, 46(1), 201–211.

[48] Kumbhakar, S. C., & Heshmati, A. (1995). Efficiency measurement in Swedish dairy farms: An application of rotating panel data, 1976–88. *American Journal of Agricultural Economics*, 77(3), 660–674.

[49] Kumbhakar, S. C., & Lovell, C. A. K. (2000). *Stochastic frontier analysis*. Cambridge, United Kingdom: Cambridge University Press.

[50] Kumbhakar, S. C., & Parmeter, C. F. (2019). Implementing generalized panel data stochastic frontier estimators. In E. G. Tsions (Ed.), *Panel data econometrics: Theory and empirical applications*, Chapter 9. London, United Kingdom: Elsevier.

[51] Kumbhakar, S. C., Tsonias, E. G., & Sipiäinen, T. (2009). Joint estimation of technology choice and technical efficiency: An application to organic and conventional dairy farming. *Journal of Productivity Analysis*, 31(2), 151–161.

[52] Kumbhakar, S. C., Wang, H.-J., & Horncastle, A. (2015). *A practitioner’s guide to stochastic Frontier analysis using stata*. Cambridge, United Kingdom: Cambridge University Press.
[53] Kuosmanen, T. (2012). Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. *Energy Economics*, 34, 2189–2199.

[54] Kutlu, L. (2010). Battese-Coelli estimator with endogenous regressors. *Economics Letters*, 109, 79–81.

[55] Latruffe, L., Bravo-Ureta, B. E., Carpentier, A., Desjieux, Y., & Moreira, V. H. (2017). Subsidies and technical efficiency in agriculture: Evidence from European dairy farms. *American Journal of Agricultural Economics*, 99, 783–799.

[56] Lee, L.-F., & Tyler, W. G. (1978). The stochastic frontier production function and average efficiency: An empirical analysis. *Journal of Econometrics*, 7, 385–389.

[57] Lee, Y., & Schmidt, P. (1993). A production frontier model with flexible temporal variation in technical efficiency. In K. L. H. Fried & S. Schmidt (Eds.), *The measurement of productive efficiency*. Oxford, United Kingdom: Oxford University Press.

[58] Lien, G., Kumbhakar, S. C., & Hardaker, J. B. (2017). Accounting for risk in productivity analysis: an application to Norwegian dairy farming. *Journal of Productivity Analysis*, 47(3), 247–257.

[59] Liu, J., Sriboonchitta, J., Wiboonpongse, A., & Denœux, T. (2021). A trivariate Gaussian copula stochastic frontier model with sample selection. *International Journal of Approximate Reasoning*, 137, 181–198.

[60] McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5), 995–1026.

[61] Meeusen, W., & van den Broeck, J. (1977a). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2), 435–444.

[62] Meeusen, W., & van den Broeck, J. (1977b). Technical efficiency and dimension of the firm: Some results on the use of frontier production functions. *Empirical Economics*, 2(2), 109–122.

[63] Mutter, R. L., Greene, W. H., Spector, W., Rosko, M. D., & Mukamel, D. B. (2013). Investigating the impact of endogeneity on inefficiency estimates in the application of stochastic frontier analysis to nursing homes. *Journal of Productivity Analysis*, 39(1), 101–110.

[64] Nelsen, R. (2006). *An introduction to copulas*, 2nd ed. New York City, NY: Springer Science and Business Media.

[65] O’Hagan, A., & Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1), 201–203.

[66] Olson, J. A., Schmidt, P., & Waldman, D. A. (1980). A Monte Carlo study of estimators of stochastic frontier production functions. *Journal of Econometrics*, 13, 67–82.

[67] Parmeter, C. F., & Kumbhakar, S. C. (2014). Efficiency analysis: a primer on recent advances. *Foundations and Trends in Econometrics*, 7(3–4), 191–385.

[68] Parmeter, C. F., & Zelenyuk, V. (2019). Combining the virtues of stochastic frontier and data envelopment analysis. *Operations Research*, 67, 1628–1658.

[69] Pitt, M. M., & Lee, L.-F. (1981). The measurement and sources of technical inefficiency in the Indonesian weaving industry. *Journal of Development Economics*, 9(1), 43–64.

[70] Prokhorov, A., & Schmidt, P. (2009). Likelihood-based estimation in a panel setting: robustness, redundancy and validity of copulas. *Journal of Econometrics*, 153(1), 93–104.

[71] Richmond, J. (1974). Estimating the efficiency of production. *International Economic Review*, 15(2), 515–521.

[72] Schmidt, P. (1976). On the statistical estimation of parametric frontier production functions. *The Review of Economics and Statistics*, 58(2), 238–239.

[73] Schmidt, P., & Lin, T.-F. (1984). Simple tests of alternative specifications in stochastic frontier models. *Journal of Econometrics*, 24(3), 349–361.

[74] Schmidt, P., & Lovell, C. (1979). Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers. *Journal of Econometrics*, 9(3), 343–366.

[75] Schmidt, P., & Lovell, C. (1980). Estimating stochastic production and cost frontiers when technical and allocative inefficiency are correlated. *Journal of Econometrics*, 13(1), 93–100.

[76] Schmidt, P., & Sickles, R. C. (1984). Production frontiers and panel data. *Journal of Business and Economic Statistics*, 2(2), 367–374.

[77] Simar, L., & Wilson, P. W. (2010). Inferences from cross-sectional, stochastic frontier models. *Econometric Reviews*, 29(1), 62–98.

[78] Simar, L., & Wilson, P. W. (2013). Estimation and Inference in Nonparametric Frontier Models: Recent developments and perspectives. *Foundations and Trends in Econometrics*, 5(2), 183–337.

[79] Simar, L., & Wilson, P. W. (2015). Statistical Approaches for Nonparametric Frontier Models: A Guided Tour. *International Statistical Review*, 83(1), 77–110.

[80] Sipiläinen, T., & Oude Lansink, A. (2005). Learning in switching to organic farming. *Nordic Association of Agricultural Scientists NFJ Report*, 1(1), 169–172.

[81] Smith, M. (2008). Stochastic frontier models with dependent error components. *The Econometrics Journal*, 11(1), 172–192.

[82] Solow, R. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, 39(3), 312–320.

[83] Sriboonchitta, S., Liu, J., Wiboonpongse, A., & Denœux, T. (2017). A double-copula stochastic frontier model with dependent error components and correction for sample selection. *International Journal of Approximate Reasoning*, 80, 174–184.
[84] Stiglitz, J. E., & Greenwald, B. C. (1986). Externalities in economies with imperfect information and incomplete markets. *Quarterly Journal of Economics, 101*(2), 229–264.

[85] Taube, R. (1988). *Möglichkeiten der effizienzmessung von öffentlichen verwaltungen*. Berlin: Duncker & Humbolt GmbH.

[86] Timmer, C. P. (1971). Using a probabilistic frontier production function to measure technical efficiency. *The Journal of Political Economy, 79*(4), 776–794.

[87] Tran, K., & Tsionas, M. (2015). Endogeneity in stochastic frontier models: Copula approach without external instruments. *Economics Letters, 133*(C), 85–88.

[88] Tran, K. C., & Tsionas, E. G. (2013). GMM estimation of stochastic frontier models with endogenous regressors. *Economics Letters, 118*, 233–236.

[89] Waldman, D. M. (1982). A stationary point for the stochastic frontier likelihood. *Journal of Econometrics, 18*(1), 275–279.

[90] Wang, H.-J., & Ho, C.-W. (2010). Estimating fixed-effect panel stochastic frontier models by model transformation. *Journal of Econometrics, 157*(2), 286–296.

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

[92] Wei, Z., & Kim, D. (2018). On multivariate asymmetric dependence using multivariate skew-normal copula-based regression. *International Journal of Approximate Reasoning, 92*, 376–391.

[93] Wei, Z., Zhu, X., & Wang, T. (2021). The extended skew-normal-based stochastic frontier model with a solution to “wrong skewness” problem. *Statistics, 55*, 1387–1406.

[94] Weinstein, M. (1964). The sum of values from a normal and a truncated normal distribution (with some additional material, pp. 469–470). *Technometrics, 6*(4), 104–105.

[95] Wiboonpongse, A., Liu, J., Sriboonchitta, S., & Denœux, T. (2015). 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.