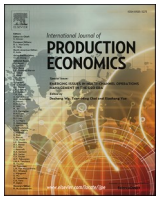
Int. J. Production Economics 232 (2021) 107940

Contents lists available at ScienceDirect

International Journal of Production Economics

journal homepage: http://www.elsevier.com/locate/ijpe



Dissections of input and output efficiency: A generalized stochastic frontier model

Subal C. Kumbhakar [a,b], Mike G. Tsionas [c,d,*]

[a] Department of Economics, State University of New York, Binghamton, NY, 13902, USA
[b] Inland Norway University of Applied Sciences, Lillehammer, Norway
[c] Montpellier Business School, 2300 Avenue des Moulins, 34080, Montpellier, France
[d] Lancaster University Management School, LA1 4YX, United Kingdom

ARTICLE INFO

JEL Classification:
C11
C13

Keywords:
Input and output inefficiency
X-efficiency
Markov chain Monte Carlo
Endogeneity
Nonlinear error components

ABSTRACT

This paper considers a model that accommodates both output and input-specific inefficiency components (input slacks). We use a translog function to represent the underlying production technology in which the input slacks are generalized to have both deterministic (functions of exogenous variables) and stochastic components. Consequently, the composed error term becomes a nonlinear function of several error components, viz., a one-sided input slack vector (the dimension of which depends on the number of inputs), a one-sided output technical inefficiency and a two-sided random noise. Identification of two sets of one-sided errors is possible in a translog model because the vector of one-sided input slacks appears in additive form as well as interactively with the (log) inputs. Distributional assumptions on technical inefficiency and slacks also help in identification. Bayesian inference techniques are introduced, organized around Markov Chain Monte Carlo, especially the Gibbs sampler with data augmentation, to estimate these inefficiency components. For an empirical application we use a large unbalanced panel of the U.K. manufacturing firms. Slacks associated with labor and capital are found to be 2.35% and 10.74%, on average. Output (revenue) loss from technical inefficiency is, on average, 2.43%, while revenue loss from input slacks is, on average, 9.2%.

> "… firms and economies do not operate on an outer-bound production possibility surface consistent with their resources. Rather they actually work on a production surface that is well within that outer bound." (Leibenstein (1966), p. 413).

1. Introduction

The importance of efficient use of resources has long been recognized in economics. The mainstream neoclassical paradigm assumes that producers always operate efficiently. However, in reality the producers are not always efficient. Two otherwise identical firms never produce the same output, and costs and profit are not the same. This difference in output, cost, and profit can be explained in terms of technical and allocative inefficiencies, and some unforeseen exogenous shocks. While explaining X-efficiency Leibenstein (1979) argued that "… There may be inefficiency in (1) labor utilization, (2) capital utilization …." (1979, pp. 13–14). Lack of full managerial effort is another source of inefficiency. Although Leibenstein did not provide a specific mathematical formulation that can be directly taken to econometric estimation and testing, he provided many arguments and evidence to show that given the resources (inputs), produced output is never at its maximum. In general, X-efficiency theory is concerned with the under-utilization of resources. If resources are underutilized cost (profit, revenue) is increased (decreased). Costs of these inefficiencies are also reflected in lower productivity of inputs. This might in turn lead to slowing down productivity growth.

Zvi Griliches (1994) in his American Economic Association presidential address noted that: "Our theories tend to assume that we are, indeed, at the frontier and that we can only either move along it or try to shift it, the latter being a difficult and chancy business. In fact we may be far from our existing 'frontiers'." Harvey Leibenstein (1966) ideas about X-efficiency, or more correctly X-inefficiency, did not get much of a sympathetic ear from us." (Griliches 1994, p. 16). According to Griliches the notion of X-inefficiency is indicative of unexploited profit

* Corresponding author. Lancaster University Management School, LA1 4YX, United Kingdom.
E-mail addresses: kkar@binghamton.edu (S.C. Kumbhakar), m.tsionas@lancaster.ac.uk (M.G. Tsionas).

https://doi.org/10.1016/j.ijpe.2020.107940
Received 13 November 2015; Received in revised form 2 September 2020; Accepted 25 September 2020
Available online 17 October 2020

opportunities and real economic growth will depend on ways of closing the gap between the observed state and the "frontier."

This study follows Leibenstein (1966, 1979) and considers a model which accommodates several aspects of inefficiency that Leibenstein alluded to. In particular the focus is on modeling both neutral (not associated with any particular input) as well as input-specific inefficiencies. The neutral part is often called managerial inefficiency while the input-specific components are viewed as slacks.[1] For example, if a worker shirks (puts less than 100% of his/her effort) we can say that labor is not fully utilized and therefore, given everything else, labor use can be reduced without reducing output(s). To paraphrase Leibenstein (1966), labor inefficiency (slack) can arise because labor contracts are never completely specified and workers cannot be forced to work with full effort (which is unobserved). Similarly, machines might not be used at their full capacity because workers handling the machines might not use them in the best possible way. In general, if an input is not used at its full capacity, it is said to have a slack. Finally, if managerial effort is lower because of, for example, non-competitive market conditions less output will be produced, ceteris paribus. In Leibenstein's terminology these are all X-inefficiency.

To make the empirical model manageable from the statistical inference point of view, the study shows how to decompose X-inefficiency in to a neutral managerial inefficiency (output inefficiency) and non-neutral input-specific technical inefficiency (ISTI) components or input slacks.[2] The distinguishing feature of the model is that output inefficiency and input slacks are separated. To be consistent with the terminology used in the stochastic frontier (SF) models, output inefficiency and technical inefficiency are interchangeably used. Similarly, the non-neutral ISTI components and input slacks are synonymous.

The standard SF models (Kumbhakar and Lovell (2000), Kumbhakar et al. (2020)) estimate inefficiency for each production unit. Often times, researchers examine factors that can influence these inefficiencies. However, these models fail to go in depths and examine how much of the inefficiency comes from what input (input-specific slacks). As a result of this the managers do not know how much of inefficiency comes from say labor. Knowing ISTI will help the managers to take steps in reducing slacks associated with the most inefficient input. If labor inefficiency is high, the manager of a production unit might consider supervising the workers more so that labor is used more efficiently. Thus ISTI provides more information about the sources of inefficiency which is useful for designing policies in reducing inefficiency.

One can view output and input inefficiency from the directional distance function (DDF) angle in which both the magnitude and directions are estimated using data. That is, contrary to the DDF in our model it is not necessary to assume directions to be exogenously given.[3] The unique feature of the paper is formulation of a generalized stochastic frontier model which combines both output inefficiency and non-radial (input-specific) technical inefficiency and show how to estimate them econometrically using a flexible production technology. The composed error term in such a model becomes a nonlinear function of several error components, viz., the one-sided input slack vector (the dimension of which depends on the number of inputs), a one-sided output technical inefficiency and two-sided random noise. Identification of two sets of one-sided errors is possible in a translog model because the vector of one-sided input slacks appears in additive form as well as interactively with the (log) inputs. In addition, this study accounts for endogeneity of inputs and provides Bayesian diagnostics for "weak instruments" in this context. Moreover, it is shown that all the existing efficiency models can be derived as special cases of the generalized model. The study proposes using Bayesian inference techniques organized around Markov Chain Monte Carlo, especially the Gibbs sampler with data augmentation, to estimate these inefficiency components. For an empirical application a translog production function is estimated using a large unbalanced panel of the U.K. manufacturing firms.

The outline of the rest of the paper is as follows. Section 2 contains a discussion of some ISTI models that are used in the existing literature and shows the generality of the new model. The formal model with input slacks and technical inefficiency is introduced in Section 3 using the translog production function. The econometric model and inference techniques are presented in Section 4. Section 5, addresses the issue of endogenous regressors. Results from estimating a translog production function with and without endogeneity corrections using an unbalanced panel of 582 British manufacturing firms, previously analyzed by Nickell (1996) and Nickell et al. (1997), is discussed in Section 5. Section 6 summarizes the main results.

2. Early literature and rudiments of a formal output and input inefficiency model

The concept of ISTI goes back to Kopp (1981) who discussed it graphically but never estimated it econometrically. Kumbhakar (1988) proposed an estimable econometric model along this route and but his analysis was confined within the Cobb-Douglas production function framework. Furthermore, neither Kopp nor Kumbhakar did separate technical inefficiency from input slacks in their models. Our idea of separating the ISTI from technical inefficiency helps us to get the off-the-shelf measure of technical inefficiency as a special case (when slacks are zero), and the ISTI of Kopp and Kumbhakar (when technical inefficiency is zero).

The notion of ISTI is also closely related to the idea of 'effective input' proposed in Koop et al. (2000, hereafter KOS) which is defined as the product of actual input and the 'efficiency factor'.[4] The efficiency factors in KOS are input-specific and are deterministic functions of input-specific vectors of exogenous variables (say z). That is, the z variables are input-specific in KOS. In our terminology (as explained in the next section) 'effective input' is slack adjusted input and the 'efficiency factor' is related to ISTI. Furthermore, it has both deterministic and stochastic components. The deterministic components are functions of the same exogenous variables. That is, unlike KOS this study does not require the z variables to be input-specific in our ISTI model, and in the extreme case (without any exogenous variables) our ISTI is simply a vector of one-sided random variables. The Cobb-Douglas production function used in KOS cannot identify the efficiency factors unless these are functions of input-specific exogenous variables. The ISTI model can identify the input-specific efficiency factors because a flexible functional form is used. Since a stochastic component is allowed in ISTI, the (composed) error term in the translog ISTI model becomes a nonlinear function of several error components, viz., the one-sided input slack vector (the dimension of which depends on the number of inputs), a

[1] The word 'slack' is used in Data Envelopment Analysis (which is based on programming technique), which is an alternative to stochastic frontier analysis. For details see Ray (2004) and references cited in there.

[2] The stochastic frontier literature, as it stands now, does consider either only output technical inefficiency or input technical inefficiency (radial) which assumes proportional reduction in all inputs. There are currently no models with input slacks with or without output technical inefficiency in the SF literature.

[3] Although there are many applications in which the directions are assumed to be exogenous, there are approaches in which directions are assumed (Färe et al. (2013), Hampf and Krüger (2015), Atkinson and Tsionas (2016), among others.

[4] This idea in a slightly different form can be found in Sato and Beckmann (1968), Burmeister and Dobell (1969), Beckmann and Sato (1969). They call it a factor augmenting model of technical change in which input efficiency factors are augmented by exponential functions of time and are deterministic as in KOS. This is also the framework used in Gollop and Roberts (1981). However, none of these models allow observation-specific input efficiency.

one-sided technical inefficiency and a two-sided random noise. Identification of two sets of one-sided errors is possible in a translog model because the vector of one-sided input slacks appears in additive form as well as interactively with the (log) inputs. On the other hand, the one-sided technical inefficiency terms appear only additively. Distributional assumptions on technical inefficiency and slacks also help in identification. These issues will be clear after the formal model is presented in Section 3. Since input slacks in the models are new it is possible to test formally the validity of the new specification against only neutral component, and the presence of ISTI in only one input using marginal likelihoods. Furthermore, endogeneity of inputs is addressed following Tran and Tsionas (2013). The study finds significant and economically important differences in the estimated inefficiency after correcting for endogeneity.

Bernstein et al. (2004) specified a production technology in terms of inputs that are scaled by input-specific parameters purporting to measure input-specific technical inefficiency. The product of these parameters and actual inputs generate efficiency adjusted input levels. They assert that these parameters capture net efficiencies from changes in input usage and adjustment costs. Their measures are input-specific but not firm-specific. Gollop and Roberts (1981) proposed a similar measure within the context of a cost function by introducing input price-specific functions that scale individual input prices and hence presumably measure input-specific technical inefficiency. These are allowed to vary over time in a deterministic fashion and are firm-invariant.

Chambers, Chung, and Färe (1998) introduced the DDF to model the direction towards the frontier from an inefficient production plan by expanding the outputs and contracting the inputs. DDF is specified as $\delta(X,Y|g_X,g_Y,\Psi)=sup\{\delta|(X-\delta g_X,Y+\delta g_Y)\in\Psi\}$, which, projects the output and input vectors $(X,Y)\in\Re^{J+M}$ onto the technology frontier in a direction determined by a vector $d=(-g_X,g_Y)$, where $(g_X,g_Y)\in\Re_+^{J+M}$ and Ψ is the feasible production set. The DDF $\delta(X,Y|g_X,g_Y,\Psi)$ shows that given the direction vector $(-g_X,g_Y)$, inputs X can be reduced by $\delta.g_X$ and simultaneously outputs can be increased by $\delta.g_Y$. Generality of the DDF lies in the fact that it allows outputs are to be expanded and inputs to be contracted simultaneously. Thus, the traditional output and input-oriented measures become special cases of DDF. Although the superiority of DDF is advocated in terms of its flexibility in choosing the direction, in their empirical model the direction vector $(-g_X,g_Y)$ is chosen exogenously. Since δ is a scalar, this means that in the end all the outputs are increased by an equal amount (δ) and all the inputs are decreased by the same amount (δ). Thus, the DDF loses much of its steam when it comes to empirical application. Furthermore, as noted by Hudgins and Primont (2007) only two parametric functions, namely, the logarithmic-transcendental and the quadratic (without an intercept term) forms satisfy the requirements of the DDF. None of these functions are popular empirically.

Table 1 summarizes the contribution of some papers that focus on either ISTI and/or decomposition of productivity into input-specific components.

The preset study considers a specification that looks similar to DDF but it does not restrict it in choosing the direction *a priori*, and the directions are not same for all pairs of inputs and outputs. Instead of focusing on directions, the focus is on proportional reduction in inputs and increase in outputs which are input and output-specific. Note that directions are important to determine reduction in inputs and increase in outputs. If these proportions are determined, the direction can be automatically determined. Thus, in the present formulation the focus is not on direction but proportional reduction in inputs and increase in outputs. Since we deal with proportions, our measures are not affected by units of measurement and one can use the popular functional forms such the translog.

The production technology is specified in terms of a transformation function which is augmented to include inefficiency in Y and X, i.e., $F(\Lambda\odot Y,\Theta\odot X)=A$, where $\Lambda_m\geq 1$ indicates inefficiency in output Y_m and $\Theta_j\leq 1$ indicates inefficiency in the use of input j. Note that here Λ and Θ are output- and input-specific, and $\odot$ represents the Hadamard product. We call it generalized stochastic frontier (GSF) model, where the stochastic part of it comes from A specified as $A=A_0e^v$. Here v is a random noise term with zero mean.

The idea of input and output inefficiency is illustrated in Fig. 1 with one input and one output. Point A shows the observed input output combination which is below the production frontier $Y=f(X)$. If the optimal input output combination (determined from, say, profit maximization behavior) is given by the point B, the presence of input inefficiency is indicated from the fact that $X^o<X^a$ and inefficiency in the production of output is indicated by $Y^o>Y^a$. That is, by fully eliminating input inefficiency input use can be reduced by (X^a-X^o) and production of output can be increased by(Y^o-Y^a). Alternatively, an inefficient firm could simultaneously increase its output by $(\Lambda-1)\times 100$ percent and decrease its input use X by $(1-\Theta)\times 100$ percent. Note that in terms of the figure $\Lambda=Y^o/Y^a\geq 1$ and $\Theta=X^o/X^a\leq 1$.

Input and output inefficiency in Fig. 1 can be easily related to reduction in input use and increase in output in the DDF formulation, presented in Fig. 2. Note that $X-g_X\delta$ in Fig. 2 is the optimal amount of X, which in our framework of Fig. 1 is indicated by $X\Theta$. That is, $X-g_X\delta=\Theta X\Rightarrow X(1-\Theta)=g_x\delta\equiv\Delta X$, which implies $\Delta X/X=1-\Theta$. Therefore, if $X-g_X\delta$ is known, so is Θ. Similarly, we can write $Y-g_Y\delta=\Lambda Y$ which gives $\Lambda-1=g_Y\delta/Y$. Thus, from Λ one can recover $g_Y\delta$ and vice-versa. From this, it is clear that the two representations are quite similar. The main difference is that the DDF is expressed in levels of inputs and outputs, while the GSF model is expressed in logarithms of inputs and outputs. Note that optimal point (direction) in both figures is determined by profit maximization behavior.

Similar to the DDF, the SF models used in the literature can be shown as special cases of GSF. For example, if $\Theta_j=1\ \forall j$ and $\Lambda_m=\Lambda\ \forall m$ the GSF model reduces to the output-oriented technical inefficiency model. Similarly, if $\Lambda_m=1\ \forall m$ and $\Theta_j=\Theta\ \forall j$ the GSF model reduces to the input-oriented technical inefficiency model. Finally, if $\Lambda_m=\Lambda\ \forall m$, $\Theta_j=\Theta\ \forall j$ and $\Lambda\times\Theta=1$, the GSF model reduces to the hyperbolic inefficiency model. Another special case (in a single output model) is when $\Lambda=1$. This would be a model to estimate input slacks, given output. The model is used in a cost function set-up by Kumbhakar and Tsionas (2012). However, they did not estimate input efficiency. In the Data Envelopment Analysis literature input-specific inefficiency is estimated in Mahlberg and Sahoo (2011), Kapelko et al. (2015), Oude Lansink et al. (2002), and Skevas and Oude Lansink (2014).

In summary, the present study uses a translog production function to estimate a generalized efficiency model with input slacks and technical

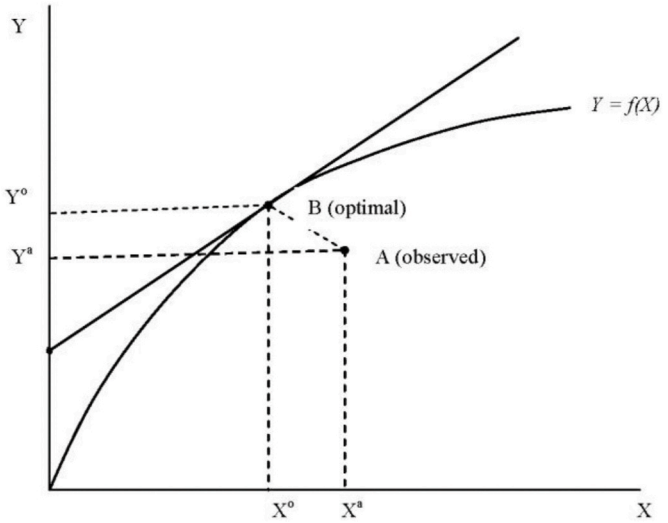


Fig. 1. Input and output inefficiency.

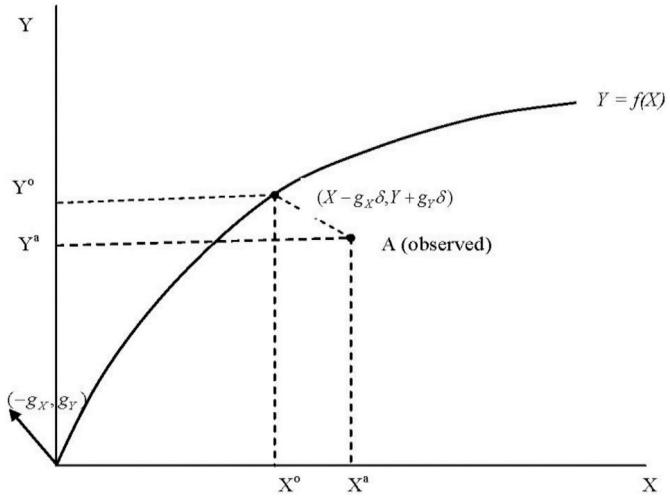


Fig. 2. Directional distance function.

inefficiency. The inputs slacks have both deterministic (functions of exogenous variables) and stochastic components. Furthermore, these are observation-specific. Separating input slacks from technical inefficiency necessitates the use of a flexible functional form such as the translog in the generalized model. This enables an examination of the role of the exogenous factors on effectiveness (X-efficiency/productivity) of inputs and test whether technical inefficiency is neutral (the way it is modeled in the standard stochastic frontier literature following Aigner et al. (1977) and Meeusen and van den Broeck (1977)). Thus, the generalized model provides an extension of the Kumbhakar (1988) and the KOS models. It also provides both economic and econometric generalizations to Caselli and Coleman (2006) (in which restrictive functional forms and assumptions are used). To estimate the model, proposed are Bayesian inference techniques organized around Markov Chain Monte Carlo, especially the Gibbs sampler with data augmentation, to estimate the inefficiency components. The model is of special interest to managers and managerial economists as it provides more information on efficiency. Most stochastic frontier models provide an overall efficiency measure which is hard to interpret as it does not say how it relates to individual inputs. The present model deviates from standard practice and focuses explicitly on input-specific efficiency so that management knows the sources of inefficiency in production. In fact, the model clearly shows the channels through which inefficiency affects output.

3. Formalizing the GSF model

3.1. Production function with input and output inefficiency

Given the concepts discussed in Fig. 1, a mathematical formulation of the production function is written as follows:

$$\Lambda_i Y_i = A_i f(X_{1i}\Theta_{1i}, ..., X_{Ji}\Theta_{Ji}) \Rightarrow \tilde{Y}_i = A_i f\left(\tilde{X}_{1i}, ..., \tilde{X}_{Ji}\right), \quad i = 1, ..., n, \tag{1}$$

where Y is a single output, X is a vector of J inputs and $f(.)$ is the production technology. The subscript i indexes firm and j indexes inputs. The term A_i is the shift parameter and is also captures stochastic nature of production, $\Lambda_i \geq 1$ captures firm-specific output efficiency. The input- and firm-specific efficiency factors are captured by $\Theta_{ji}(0 \leq \Theta_{ji} \leq 1)\ \forall j, i$. Thus, $\tilde{X}_{ji} = X_{ji}\Theta_{ji}$ is input j in efficiency (slack adjusted) units. Similarly, $\tilde{Y}_i = Y_i\Lambda_i$is output in efficiency unit. Alternatively,$1 - \Theta_{ji} \geq 0$ is the percentage (when multiplied by 100) by which the input j is over used by firm i, and $(\Lambda_i - 1) \geq 0$ is the percent by which output could be increased. The existence of input slacks can explain differences in input productivity (effectiveness) among firms. Thus, this formulation of the production function captures both (output) technical efficiency and input slacks (existence of which can explain differences in effectiveness of inputs). If there are no slacks in any of the inputs $(\Theta_{ji} = 1\ \forall j, i)$, one gets back to the standard output-oriented stochastic frontier model in which $\ln \Lambda \geq 0$ captures output inefficiency. Note that it is not required to have panel data to estimate such a model. The question is whether one can separate these two types of inefficiencies, given that this formulation is neither discussed in Kopp (1981), Kumbhakar (1988) nor in KOS. This is addressed in the context of a flexible (translog) production function in the next section.

Although the model in (1) looks similar to the KOS model, there are some important differences between these two models. The KOS model does not include output technical inefficiency. However, the concept of 'effective input' in KOS is same as our 'slack adjusted input'. The differences are in the specification of the Θ_{ji} functions, which are more flexible in our model. These are discussed in the next section. The new model has some features that are similar to Gollop and Roberts (1981) and Bernstein et al. (2004), although none of these models introduced observation-specific (i.e., firm-specific in a cross-section and both firm-specific and time-varying in a panel) input slack (effectiveness) measures. The slack measures in the present study, contain both deterministic and stochastic components which are not introduced in any of the previous papers. Consequently, the new model generalizes all the previous models – both frontier and non-frontier specifications.

3.2. The econometric model and inference

The present study uses a flexible functional form (translog) to represent the production technology, i.e.,

$$y_i = \ln A_i + u_{0i} + \underset{(J\times 1)}{\alpha'}\left(\underset{(J\times 1)}{x_i} + \underset{(J\times 1)}{\theta_i}\right) + \frac{1}{2}(x_i + \theta_i)'\underset{(J\times J)}{B}(x_i + \theta_i), \tag{2}$$

where

$$\ln A_i = \alpha_0 + \delta_0'\tilde{z}_i + \varepsilon_{0i}, \quad \ln\left(\Lambda_i^{-1}\right) = u_{0i} \geq 0$$

$$\Rightarrow \ln A_i - \ln \Lambda_i = \alpha_0 + \delta_0'\tilde{z}_i + \varepsilon_{0i} + u_{0i} = \delta' z_i + \varepsilon_{0i} + u_{0i} \tag{3a}$$

$$\underset{(J\times 1)}{\theta_i} \equiv \ln \underset{(J\times 1)}{\Theta_i} = \underset{(J\times m)}{\Delta}\underset{(m\times 1)}{z_i} + \underset{(J\times 1)}{u_i}, \tag{3b}$$

$$\theta_i \leq 0_{(J\times 1)}, \quad u_{0i} \leq 0, \quad \text{for each } i = 1, ..., n, \tag{3c}$$

$z_i = [1, \tilde{z}_i']'$, $y_i = \ln Y_i$, $\ln X_{ji} = x_{ji}$ and $x_i = (x_{1i}, ..., x_{Ji})'$, $\alpha_{(J\times 1)} = (\alpha_1, ..., \alpha_J)'$ and B is a $J \times J$symmetric matrix. The model in (2) and (3) constitutes the translog GSF model. The z_i variables are introduced as neutral technology shifters via $\ln A_i$ and determinants of input inefficiency ($\ln \Theta_i$) via (3b). The negative values of $\theta_{ji}(\approx 1 - A_{ji})$, multiplied by 100, can be interpreted as the percentage over-use of input j for firm i. Output technical inefficiency is captured by the u_{0i} term. The usual stochastic noise component (ε_{0i}) captures exogenous shocks in the production technology. In the absence of the slacks, the generalized inefficiency model reduces to the standard stochastic frontier model of Aigner et al. (1977) and Meeusen and van den Broeck (1977). The

presence of the slacks can be econometrically tested. One can also test whether inefficiency is only in the output (only technical inefficiency) or in the slacks (no technical inefficiency).[5]

One can justify the claim that the model specified in (2) and (3) is much more general by arguing that it clearly shows the channels through which inefficiency affects output. If one believes that there is inefficiency and part of it can be explained by some exogenous variables (z) variables, then it is natural to think of the channels through which inefficiency affects output. A *direct* channel through which the z variables affect output in a neutral fashion (neutral to the regular input variables (x)) and enter into the production function through the shift parameter (namely the $\delta' z_i$ part in $\ln A_i$in considered in (3a)). The effect of unobserved variables affecting output through this channel is captured by the one-sided random term u_{0i}in (3a).[6] In addition, these z variables can also affect output *indirectly* by affecting input productivity which is captured by the slacks in (3b). This indirect channel is 'new' and has not been used in the traditional SF models (with or without the z variables) which are special cases of our general model.

Several comments on the GSF model in (2) and (3) are in order, particularly in relation to the KOS model. First, equation (2) is a translog whereas it is a Cobb-Douglas in KOS. Second, in the new model the same z variables can affect 'effectiveness' of *all* inputs whereas in KOS the z variables are input-specific (i.e., effectiveness of each input is explained by a separate set of z variables). Third, in the new model the input effectiveness parameters contain both deterministic and stochastic components whereas these are completely deterministic in KOS. Thus, even if there are no z variables, the new model can provide estimates of ISTI in a translog model. This, however, is not the case in the KOS model.

It is possible to rewrite (2) after substituting (3) as

$$y_i = \alpha_0 + \delta_0' \tilde{z}_i + \alpha' x_i + \frac{1}{2} x_i' B x_i + \left[\theta_i' q_i + \frac{1}{2} \theta_i' B \theta_i + \varepsilon_{0i} + u_{0i} \right] \tag{4}$$

where $q_i = \alpha + Bx_i$ is the output elasticity vector ($\partial y_i / \partial x_i$). In this form, the model consists of the usual translog production frontier (with a neutral shift via the z variables) and a composed error term that is a nonlinear function of several error components, viz.,

$$v_i = \left[\theta_i' q_i + \frac{1}{2} \theta_i' B \theta_i \right] + u_{0i} + \varepsilon_{0i} = \left[\left(\theta_i' \alpha + \frac{1}{2} \theta_i' B \theta_i \right) + \theta_i' B x_i \right] + u_{0i} + \varepsilon_{0i} \tag{5}$$

The composed error term v_i in (5) consists of (i) a quadratic function of the non-positive input slack vector θ_i (the dimension of which depends on the number of inputs), (ii) non-positive technical inefficiency component u_{0i}, and (iii) the random noise component ε_{0i}. That is, the composed error term v_i is a function of two sets of one-sided error components (θ_i and u_{0i}) and a two-sided error term ε_{0i}. Furthermore, it is nonlinear in θ_i and depends on the input quantities as well. The question is, whether one can identify all these components, especially, the non-positive slack vector θ_i from the non-positive technical inefficiency component u_{0i}. The answer is in the affirmative because the θ_i vector appears in both linear $\theta_i' \alpha$ and quadratic forms ($\frac{1}{2}\theta_i' B \theta_i$). It also appears interactively with the x_i vector $\theta_i' B x_i$. On the other hand, the technical inefficiency component u_{0i} appear only linearly. Furthermore, distributional assumptions on θ_i, $u_{0i}$and ε_{0i} will help in identification.

Some special cases of the model might be of interest. If the percentage over-use of all the inputs are exactly the same (i.e.$\theta_{ji} = \theta_i$, $j = 1, ..., J$) and $u_{0i} = 0$, then one obtains a stochastic translog production frontier model with input-oriented technical inefficiency (Kumbhakar and Tsionas (2006)), in which the composed error term is $\left(\theta_i l' \alpha + \frac{1}{2} \theta_i^2 l' B l \right) + \theta_i l' x_i + \varepsilon_{0i}$where θ_i is a scalar and l is a $J \times 1$ column vector of ones. On the other hand, in the absence of input slacks ($\theta_{ji} = 0 \; \forall \, j = 1, ..., J$) the model in (4) reduces to a standard translog stochastic production frontier function with output-oriented technical inefficiency in which the composed error term is the sum of u_{0i} and ε_{0i}. Note that the conditional expectation of the composed error is $E(v_i | x_i, z_i) = \overline{\theta}_i q_i + \frac{1}{2} tr[B(\Lambda_i + \overline{\theta}_i \overline{\theta}_i')] + \overline{u}_{0i}$, where $\overline{\theta}_i = \Delta z_i + \overline{u}_i$, $\overline{u}_i = E(u_i | x_i, z_i)$, $\overline{u}_{0i} = E(u_{0i} | x_i, z_i)$ and$\Lambda_i = Cov(\theta_i)$, for each $i = 1, ..., n$. Since it is non-zero and depends on the data, the least squares estimation of (4) will clearly yield inconsistent estimates. It should be noted that Λ_i is not the same as the scale matrix Ω which is the pre-truncated covariance of$u_i | z_i \sim N_J(0, \Omega)$.

Assumptions.

For estimation purposes, following assumptions on the stochastic error components are made.

Assumption 1. $\varepsilon_{0i} \sim iid \; N(0, \sigma^2)$,

Assumption 2. $u_{0i} \sim iidN(0, \sigma_u^2)$, $u_{0i} \leq 0$,

Assumption 3. $u_i | z_i \sim N_J(0, \Omega)$, $u_i \leq -\Delta z_i$ and are distributed independently of $(\varepsilon_{0i}, u_{0i})$conditional on z_i.

Assumption 4. Furthermore, ε_{0i} and $u_{0i}$are independent of one another. Here, $N_J(0, \Omega)$ denotes a J-variate normal distribution with mean vector zero, and variance-covariance matrix Ω.

Assumption 4b. Equivalently, it can be assumed that $\theta_i | z_i \sim N_J(\Delta z_i, \Omega)$, $\theta_i \leq 0_{(J \times 1)}$.

Given these assumptions, we can rewrite the model in (4) compactly as[7]:

$$y_i = w_i' \delta + v_i, \tag{6}$$

where.

$$w_i = \begin{bmatrix} 1 \\ \tilde{z}_i \\ x_i \\ vech(x_i \otimes x_i') \end{bmatrix} \text{ and } \delta = \begin{bmatrix} \alpha_0 \\ \delta_0 \\ \alpha \\ vech(B) \end{bmatrix}.$$

The joint pdf of u_i is:

$$p(u_i | z_i) = C_i(\Omega, \Delta)^{-1} (2\pi)^{-J/2} |\Omega|^{-1/2} \exp\left(-\frac{1}{2} u_i' \Omega^{-1} u_i \right), u_i \leq -\Delta z_i,$$

$$for \; each \; i = 1, ..., n$$

where the integrating constant is:

$$C_i(\Omega, \Delta) = P(u_i \leq -\Delta z_i \; | u_i \sim N_J(0, \Omega)) = \int_{-\infty}^{S_{i1}} ... \int_{-\infty}^{S_{iJ}} (2\pi)^{-J/2} |\Omega|^{-1/2} \exp\left(-\frac{1}{2} u_i' \Omega^{-1} u_i \right) du_i,$$

and $S_{i1}, ..., S_{iJ}$ denote the elements of $-\Delta z_i$. This integrating constant is a J dimensional multivariate normal integral and it involves the z_i terms so that it is data-dependent. Moreover, it depends on the parameters in Δ and Ω. The importance of these issues will become clear in the course of the discussion.

The pdf of θ_i is:

[5] Since the concept is new, we used the profit maximization argument to explain the concept of non-radial inefficiency associated with input use and production of output. So, the graph we used is for illustration of the concept. However, we did not setup a model that include the first-order conditions of profit maximization which require information on output and input prices. This will involve a system approach and a different estimation approach. Also, we don't have price information in the data.

[6] It is possible to extend the model further and let the z_i variables affect output via the output efficiency terms Λ_i.

[7] By *vec*() we denote the usual vectorization operator. The operator *vech*() stacks the *different* elements of row vectors of a certain matrix.

$$p(\theta_i|z_i) = C_i(\Omega,\Delta)^{-1}(2\pi)^{-J/2}|\Omega|^{-1/2}\exp\left[-\frac{1}{2}(\theta_i-\Delta z_i)'\Omega^{-1}(\theta_i-\Delta z_i)\right] \quad (7)$$

Now consider the joint pdf of $(y_i, u_{0i}, \theta_i|x_i, z_i)$,

$$p(y_i,u_{0i},\theta_i|x_i,z_i) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left[-\frac{\left(y_i-w_i'\delta-\theta_i'q_i-\frac{1}{2}\theta_i'B\theta_i-u_{0i}\right)^2}{2\sigma^2}\right]\times$$

$$\left(\frac{\pi\sigma_u^2}{2}\right)^{-1/2}\exp\left(-\frac{u_{0i}^2}{2\sigma_u^2}\right)\cdot C_i(\Omega,\Delta)^{-1}(2\pi)^{-J/2}|\Omega|^{-1/2}\exp\left[-\frac{1}{2}(\theta_i-\Delta z_i)'\Omega^{-1}(\theta_i-\Delta z_i)\right]$$

The conditional distribution of y for the ith observation is:

$$p\ \ y_i \mid x_i,z_i\ \ = \int_{-\infty}^{0}\int_{\mathcal{A}} p(y_i,u_0,\theta_i \mid x_i,z_i)d\theta_i du_0\ , \quad (8)$$

where $\mathcal{A} = -\mathbb{R}_+^J$.

Analytical integration of (8) with respect to θ_i is not possible, although the univariate integral with respect to u_{0i} can be computed analytically (Kumbhakar and Lovell, 2000, p. 96). Unfortunately, this leads to further nonlinearities involving the standard normal distribution function, and it is not advisable to proceed along this path. The MCMC procedure is described in Part 1 of the Technical Appendix.

Part 2 of Appendix provides a discussion of computing elasticities in the context of the new model.

4. Endogenous regressors

So far, it has been assumed that the inputs are exogenous. To address possible endogeneity of the input variables (inputs being correlated with the composed error term in (4)), one can keep the basic structure of the model the same (equation (4)) but assume availability of panel data. In turn, one can express the input vector x_{it} as:

Assumption 5.

$$x_{it} = \gamma_o + \underset{(J\times J)}{\Gamma^*}x_{i,t-1} + \underset{(J\times m)}{E}\tilde{z}_{i,t-1} + \underset{(J\times 1)}{\xi_{it}} \equiv \Gamma G_{it-1} + \xi_{it} \quad (12)$$

which is a **panel vector auto-regression (VAR)** for x_{it} including the variables in $\tilde{z}_{it}$ which are assumed strictly exogenous in the sense that their expectations conditional on all error terms are zero. The endogeneity is captured by allowing for correlation between ξ_{it} and ε_{0it} as in Tran and Tsionas (2013), viz.,

Assumption 6.

$$\begin{bmatrix}\tilde{\xi}_{it}\\ \varepsilon_{0it}\end{bmatrix} = \begin{bmatrix}\Omega_\xi^{-1/2}\xi_{it}\\ \varepsilon_{0it}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}I_J & \sigma_\varepsilon\rho\\ \sigma_\varepsilon\rho' & \sigma_\varepsilon^2\end{bmatrix}\right) \quad (13a)$$

where σ_ε^2 is the variance of ε_{0it}, ρ is a $J\times 1$ vector of correlation coefficients and Ω_ξ is the $J\times J$ covariance matrix of ξ_{it}. Using the Cholesky decomposition one can write:

$$\begin{bmatrix}\tilde{\xi}_{it}\\ \varepsilon_{0it}\end{bmatrix} = \begin{bmatrix}I_J & 0\\ \sigma_\varepsilon\rho' & \sigma_\varepsilon\sqrt{1-\rho'\rho}\end{bmatrix}\begin{bmatrix}\tilde{\xi}_{it}\\ \tilde{\omega}_{it}\end{bmatrix} \quad (13b)$$

where.

Assumption 7. $\tilde{\omega}_{it} \sim N(0,1)$ is independent of $\tilde{\xi}_{it}$. With these, one can we rewrite (4) as follows:

$$y_{it} = \delta' z_{it} + \alpha'(x_{it}+\theta_{it}) + \frac{1}{2}(x_{it}+\theta_{it})'B(x_{it}+\theta_{it}) + \eta'(x_{it}-\Gamma G_{it-1}) + \omega_{it} + u_{0it}, \quad (14)$$

where $\eta = \sigma_\varepsilon\Omega_\xi^{-1/2}\rho$ and $\omega_{it} \sim N(0,\ \sigma_\varepsilon^2(1-\rho'\rho))$. Conditional on θ_{it} (14) is a standard stochastic frontier model. This equation has to be embedded in the context of the panel VAR in (12) from which identification of Γ and Ω_ξ is possible. Compared to the methods developed in the previous section, (14) is a minor extension in which updates of the elements of Γ are made from the joint conditional posterior distribution defined by (14) and (12) .[8] Draws from the conditional posterior distribution of η are straightforward and draws from the conditional posterior distribution of ρ are implicit in this construction.

Additional details are provided in Part 3 of Appendix.

5. Empirical results

The new model proposed in the paper is estimated using an unbalanced panel of 582 British manufacturing firms previously analyzed by Nickell (1996) and Nickell and Nicolitsas (1999). Details on the construction of the variables used here can be found in these two papers. The total number of observations in the data set is 5273 and the number of yearly observations for each firm ranges from 5 to 13. For estimation purposes, the following translog production function is used:

$$\log Y_{it} = f(\log L_{it}, \log K_{it}) + v_{it}$$

where Y_{it} represents firm sales[9] (deflated by producer price index of the three-digit industry in which the firm belongs), L_{it} is labor (number of employees), and K_{it} is capital. The function f (.) is assumed to be translog. Although this data set does not have information on other inputs like materials, use of it is made because this data has some interesting z variables. In this sense the application might be viewed as an illustration of the new technique. Moreover, Nickell (1996) and Nickell and Nicolitsas (1999) are well known for addressing the relationship between competition and financial pressure on firm performance. Because of this particular data set is quite useful as a benchmark to apply the new techniques. This study follow the spirit of the analysis in Nickell (1996) and Nickell and Nicolitsas (1999), and uses the following variables as the determinants of input slacks: concentration ratio (CR), the market share (MS), import penetration (IMP), trend (TREND) and financial pressure (FP). Market share is defined as the ratio of firm sales

[8] Given η this involves drawings from a multivariate normal distribution. Given $\mathbf{\Gamma}$ all the parameters in (12), in turn, can be drawn from a multivariate normal given θ_{it}.

[9] Ideally one should use materials in estimating the production function when the dependent variable is sales. Unfortunately, materials is not available in the data set and this has not been included in Nickell (1996), Nickell and Nicolitsas (1999) and some other studies that used this data. Omission of materials is likely to affect results and thus our results should be viewed as illustrative.

over industry total sales. Financial pressure is defined as the ratio of interest payments over the sum of profits before tax, depreciation and interest payments. Concentration ratio is five-firm concentration ratio in terms of sales. Import penetration is the ratio of imports over home demand (sales + imports – exports). All these z variables are used in explaining capital- and labor-specific slacks (i.e., θ_K, θ_L). Using only a time trend would make the analysis somewhat close (ex post) to the deterministic analysis in Caselli and Coleman (2006).

Regression parameters (δ and $\Delta_0 = vec(\Delta)$) are *a priori* (or marginally) independent $N(\mathbf{0}, 10^3\mathbf{I})$ variables with the respective dimensions. For the scale parameters (σ^2 and σ_u^2) we use proper priors with $\underline{q} = \underline{q}_u = 0.001$ and $\underline{n} = \underline{n}_u = 1$. The same priors were selected for σ_λ^2 and σ_τ^2. For Ω, our Wishart prior has $\underline{n}_\Omega = 1$, and $\underline{\Sigma} = 0.001\cdot\mathbf{I}$.

For implementing the Gibbs sampler see the discussion in Appendix A regarding the various tuning parameters and the number of iterationsM_1, M_2, and M. Overall, the present study used ten million passes (after using Geweke's diagnostic, (Geweke, 1999)). The computation of marginal likelihood for these models involves a total of $n$integrals of dimension $J+1$. Presented are two sets of results, viz., with and without endogeneity corrections. In Table 2, presented are summary statistics of the inefficiency and slack measures together with input elasticities. The model with endogeneity correction refers to the system based on (12) and (14) and results therefrom are referred to as system in various tables/figures.

Mean output technical inefficiency u_0 based on the endogeneity corrected model is found to be 2.43% with a standard deviation 1.8%. This means, on average, sales revenue is 2.43% lower due to technical inefficiency. The interesting finding is the presence of slacks in both labor and capital, which led to their over-use. Mean over-use of labor and capital (negative values of θ) are 2.35% and 10.74% (with standard deviations 3.79% and 2.91%). Thus, relative to labor, capital is over-used by nearly a factor of four. These results are reported in Table 2 (in summary form) and the details are in Fig. 2c which plots these inefficiencies for all observations. It can be seen that slacks in capital are distributed symmetrically but have wide variation. On the other hand, the slack in labor has a long right tail but there are fewer firms for which slacks in labor are as high as slacks in capital.

Results from the model that ignores endogeneity of the inputs are quite different. Mean output technical inefficiency u_0 is found to be 3.73% with a standard deviation 2.48%. This means, on average, sales revenue is 3.73% lower due to technical inefficiency. Mean over-use of labor and capital are 10.23% and 19.23% (with standard deviations 8.4% and 9.5%). Thus, capital is over-used nearly double of labor. Fig. 2a plots these inefficiencies for all observations. It can be seen that slacks in capital have wide variation. On the other hand, the slack in labor has long right tail but there are fewer firms for which slacks in labor are as high as slacks in capital.

Existence of input slacks indicates that inputs are not used with 100% effectiveness. As a result of this produced output will be less than maximum possible. This potential reduction in output for our translog model is given by the expression $\theta_i' q_i + \frac{1}{2}\theta_i' B\theta_i$ in (4). The resulting distribution of output loss is given in Fig. 2b in which we also include technical inefficiency since it is interpreted as output loss due to managerial inefficiency. Thus, both the components in Fig. 2b measure output loss and can be compared. Output loss due to input slacks is, on average, 15.10% with a standard deviation of 7.2%. This, when added to technical inefficiency, gives a total output loss of 18.83%, on average, due to both types of inefficiency. The distribution of output loss for

Table 1
Some relevant papers related to ISTI.

Author/Year	Model Characteristics	Method/Approach
Kopp (1981)	Introduced the concept of ISTI.	Graphical and numerical illustration of the idea.
Kumbhakar (1988)	Used a Cobb-Douglas production function to capture the idea.	Used SFA with half-normal distributions on the ISTI components and normal distribution on the noise term. Estimate ISTI for each observation.
Kopp et al. (KOS) (1988)	ISTI is specified as a deterministic function of exogenous variables which are specific to each input.	Used Bayesian econometric approach to estimate ISTI for each observation.
Bernstein et al. (2004)	ISTI is modeled via input-specific but invariant across firms.	Used an econometric model with parameters varying with inputs.
Gollop and Roberts (1981)	ISTI is modeled by scaling input prices. The scaling factors are input specific but invariant across firms.	Used an econometric cost function.
Chambers et al. (CCF) (1998)	Used directional distance function to model input and output slacks. Directions are set exogenously (not estimated).	Used deterministic DEA approach.
Skevas and Oude Lansink (2014)	Measured the composition of productivity growth of pesticides and the environmental impacts of pesticides.	Dynamic DEA and bootstrap regression model.
Mahlberg and Sahoo (2011)	Used the non-radial input-specific slacks based on directional Russell measure of inefficiency.	DEA
Kapelko et al. (2015)	Estimated input-specific productivity growth and input-specific technological change, technical efficiency change.	DEA
Oude Lansink et al. (2002)	Computed overall technical and input-specific technical efficiency measures of conventional and organic farms in Finland.	DEA
Kumbhakar and Tsionas (2020)	Generalizes the idea in Kumbhakar (1988) in several important ways. Details are in the text of this paper.	Bayesian SF approach

Table 2
Estimates of inefficiency and other economic measures.

	With endogeneity (system)		Without endogeneity	
	mean	s.d.	mean	s.d.
u_0 (technical inefficiency)	0.024	0.018	0.037	0.025
θ_L (labor inefficiency)	0.023	0.038	0.102	0.084
θ_K (capital inefficiency)	0.107	0.029	0.193	0.094
Labor elasticity	0.519	0.021	0.519	0.157
Capital elasticity	0.414	0.042	0.420	0.135
Returns to scale	0.933	0.131	0.940	0.069

Note: Technical inefficiency and input slacks are sample averages of the latent variables u_0, θ_K and θ_L. Input elasticities and returns to scale are derived from (4).

Table 3
Estimates of the Δ coefficients.

	Labor post. mean post.s.d.	Capital post. mean post.s.d.
constant	−0.814 0.130	0.173 0.172
CR	−1.245 0.178	−0.226 0.141
MKSH	1.117 0.162	−0.317 0.224
FP	1.227 0.335	0.154 0.078
IMP	1.448 0.355	0.243 0.013
TREND	−0.023 0.022	−0.013 0.017

Note: Posterior standard deviations are computed using a Newey-West robust covariance estimator with L = 10 lags. The estimated posterior s.d. were similar to those obtained from thinning.

system approach with endogeneity correction is reported in Fig. 2d. Output loss due to input slacks is, on average, 9.2% with a standard deviation of 4.3%. This, when added to technical inefficiency, gives a total output loss of 11.63%, on average, due to both types of inefficiency.

Since these estimates are available for each observation, it might be interesting to examine them for each firm. To conserve space, these measures are reported for the 5 worst and the 5 best firms in Fig. 3. It can be seen that these measures vary substantially across the best and worst firms. Fig. 3a and b plot the posterior distributions of output technical inefficiency for the 5 worst and best firms. These figures show that the distributions of the worst firms vary substantially. On the contrary, the posterior distributions of inefficiency for the best 5 firms are very similar. The same conclusion holds for the posterior distributions of slacks in capital (Fig. 3c and d) and slacks in labor (Fig. 3e and f). The posterior distributions of output losses from slacks in capital and labor for the 5 worst and 5 best firms are shown in Fig. 3g and h. In this case, the distributions for the worst as well as the best 5 firms are quite different.

The average input elasticities from $Elas_i = \frac{\partial E(y_i|z_i,x_i)}{\partial x_i} = \alpha + B(x_i + \bar{\theta}_i)$ are 0.519 for labor and 0.414 for capital (with standard deviations 0.021 and 0.042 respectively). These input elasticities are often interpreted as share of input costs to total value of output (revenue). Thus, on average, labor cost share is more than that of capital. The input elasticities are uniquely related to returns to scale (RTS), viz., RTS is the sum of these input elasticities. The RTS average is around 0.933 (with standard deviation 0.131) (see Table 1). In Nickell (1996) the labor elasticities were 0.73 (t-statistic = 8.3) and 0.17 respectively but the t-statistic is not reported.

The distributions of input elasticities and RTS are reported in Fig. 4a and b for the models without and with endogeneity correction. These distributions are symmetric with positive values (within the [0,1] interval), as expected from theory. It can be seen from Fig. 4a and b that most of the firms are operating below unitary RTS (efficient scale size). There are more firms with decreasing RTS than increasing RTS, which means that there are potential for some firms to expand since they are operating below their efficient scale size. Also note that although capital and labor elasticities from the models with and without endogeneity correction are quite different, the estimated RTS (sum of these two elasticities) are quite similar.

Estimates of the coefficients associated with the z variables in the input slack functions (θ_K and θ_L) in (3b) are reported in Table 3. Since $\ln X_j - \ln\tilde{X}_j = -\theta_j \geq 0$ the Δcoefficients (when multiplied by -1) are marginal effects of the z variables on input slacks (input over-use). Thus, an increase in concentration ratio, CR (less competition) increases slacks in both labor and capital. That is, if the market is less competitive inputs will be over-used. This result can be used to support the quiet life hypothesis of a monopolist – lower work effort is associated with less competition. All other z variables reduce slacks in both labor and capital (except for market share in capital). The negative coefficients on the TREND variable show that input over-use have decreased over time.

Table 4
Marginal effects of the z variables on posterior expectations of θ and lnY.

	θ_L post. mean post. s.d.	θ_K post. mean post. s.d.	Output loss post. mean post.s.d.
CR	−0.312 0.012	−0.282 0.1772	0.392 0.077
MKSH	−0.850 0.1215	−0.671 0.136	1.0695 0.1643
FP	−0.0155 0.0111	−0.115 0.0172	−2.9839 0.4126
IMP	0.127 0.0444	0.232 0.0165	0.2516 0.1175
TREND	0.0022 0.0017	−0.0012 0.0015	0.2029 0.2420

Notes: Reported are the posterior means and posterior s.d. of the marginal effects of the z variables shown in the first column on E(θ|y,X,Z) from their MCMC draws of the Bayesian posterior distribution for the model in (4) with endogeneity correction. The derivatives are calculated numerically using the formula of E(θ|y,X,Z) from Tallis (1961).

Table 5
Bayes factors for slacks.

ISTI vs. SFM	2271.16
ISML vs. SFM	404.22
ISMK vs. SFM	501.91

Note: ISTI is the input-slack model, SFM is the traditional stochastic frontier model, ISML and ISMK are the input-slack models allowing for slacks *only* in labor and capital, respectively. The table reports the Bayes factor in favor of ISTI, ISML and ISML against the traditional SFM.

To focus more on the importance of the z variables on input-specific slacks, one can first compute the posterior means and posterior s.d. of the θ parameters from their MCMC draws of the Bayesian posterior distribution for the model in (4) with endogeneity correction (equation (14)). Then, one can calculate the marginal effects of each of the z variable on the means of posterior of θ. These derivatives are calculated numerically after obtaining $E(\theta|y,X,Z)$ from the formulae in Tallis (1961). The results are reported in Table 4.

Since Nickell (1996) incorporated the z variables into the production function, a direct comparison with our results is not straightforward. His findings are as follows. The coefficient of market share at the period t-2 was −1.30 (t = 2.5), the coefficient of concentration ratio was −0.061 (t = 2.00), for import penetration it was −0.023 (t = 0.80), for size it was 0.39 (t = 2.50) and for rents it was 0.39 (t = 2.50), see Table 2 in Nickell (1996, p. 638). These results are marginally significant judging from their t-statistics. In our framework, the effect of these variables are input as well as firm-specific unlike Nickell (1996) formulation. Despite this fact from the last column of our Table 4 it turns our that marginal effects are material and are tightly estimated as posterior standard deviations are much lower than posterior means. Moreover, signs for concentration ratio and market share differ from those reported by Nickell (1996).

It can be seen from Table 4 that, on average, higher concentration ratio, market shares, and financial pressure decrease input-slacks and therefore increases productivity of both inputs. When CR increases (competition decreases), we expect less pressure to reduce slacks for both labor and capital. Similarly, when the market share of a particular firm increases and the firm is in a monopolistic situation, firm managers

might not be under pressure to increase work effort. This argument goes in line with *'quiet life' hypothesis* of Hicks (1935). Finally, if financial pressure increases it is likely to force the workers to work hard (thereby reducing slacks) and increase productivity. An increase in import penetration, on average, had a negligible increase in slack in labor. However, the increase in slack in capital is much higher. This might be due to the fact that firms that are heavily dependent on imports are likely to keep excess reserve of capital. Finally, change in slacks associated with labor and capital over time are almost negligible.

In the last column of Table 4, reported are the marginal effects of these z variables on E(lnY) computed from (11), holding inputs unchanged. Since none of these z variables is in logarithmic terms, the marginal effects in (11) can be interpreted as semi-elasticities because y is in logs. Note that these z variables affect output directly (captured by δ_0, the first term in (11)) and indirectly by affecting input productivity via slacks. These indirect effects are more complex and are captured by the last two terms in (11). More specifically, a z variable affects output indirectly by affecting the means and variances of slacks associated with both capital and labor. In Table 4 and Fig. 5, reported are the total (overall) effect for each of the z variables on $E(\ln Y)$. The concentration ratio has a positive overall effect on expected output $E(\ln Y)$. Market share has a negative overall effect on expected output. Financial pressure and import penetration have a small positive overall effect on $E(\ln Y)$. Finally, productivity change (effect on mean output over time), on average is positive but quite small. Details on these marginal effects can be seen from Fig. 5 where the entire distributions in the sample are presented.

The model proposed in the paper differs from the existing models in terms of the input slacks. In the absence of such slacks, the model will reduce to the standard translog stochastic frontier production model with output inefficiency. Thus, it is desirable to test the appropriateness of the new model. Bayes factors in favor of the hypothesis of no input slacks against three competing models are reported in Table 5. Since the dimensionality of the problem is small $(J = 2)$ computing Bayes factors is not cumbersome. The results in Table 5 show that the null hypothesis (the traditional frontier model, SFM, which includes only radial technical inefficiency) is rejected (following Jeffrey's rule, i.e., Bayes factor in excess of 100:1 provides "decisive evidence" in favor of the alternative hypothesis) against the ISTI, ISML and ISMK where ISTI allows slacks in both capital and labor; is the input-specific slack model, ISML and ISMK are the input-slack models allowing for slacks only in labor and capital, respectively.

We also develop a test for weak instruments in the Bayesian framework we used. Our focus is the panel vector autoregression in (12) where $[x_{i,t-1}', z_{i,t-1}']' \equiv G_{it}$ were assumed exogeneous. Although this cannot be disputed, at least on economic theory grounds, it can be debated whether these instruments are, in fact, "weak". Let us consider a standard regression model, say $y_t = \beta x_t + \zeta_t$ where ζ_t is an error term that is not, necessarily, orthogonal to x_t. Suppose there is a variable, $z_t$that is orthogonal to the error. It is not difficult to show that the so-called Instrumental Variables (IV) estimator $b_{IV} = \frac{\sum_{t=1}^{n} z_t y_t}{\sum_{t=1}^{n} z_t x_t}$is always consistent under our assumptions. The problem, however, is that when the correlation between x_t and z_t is low then the denominator in b_{IV} is low and, as a result, the IV estimator will behave erratically. In the context of the panel vector autoregression in (12), we rely on the minimum R^2 across all equations of (12) for each MCMC. If the values of R^2 are "low" then we have weak instruments in the sense that $x_{it}$and $G_{it}$are only weakly correlated. The marginal posterior density of minimum R^2 across all MCMC draws is reported in Fig. 6 from which it is seen that $x_{it}$and $G_{it}$are highly correlated. Since the R^2 statistic may not be enough, we report three alternative F-tests for weak instruments, viz., Olea and Pflueger (2013), Stock and Yogo (2005) and Kleibergen (2007) (see also Kleibergen, 2004; Stock et al., 2002). The rough guide is that if the F values exceed 10 (see Table 1 in Olea and Pflueger, 2013) the instruments are considered strong. The posterior (across MCMC draws) densities of the various F-tests well exceed this critical value so we are relatively confident that, for the data at hand, there is no weak instruments problem.

6. Conclusion

Although input slacks are an essential part of X-inefficiency, the efficiency literature has focused exclusively on the interpretation of X-inefficiency as output (managerial) inefficiency thereby ignoring the essential part of Leibenstein's argument that input slacks are important determinants of the X-inefficiency. In this paper a formal approach is proposed to jointly model input slacks and output technical inefficiency. The present study considers a model that accommodates both a neutral and input-specific inefficiency component. The neutral part is technical inefficiency while the input-specific components are labeled as slacks. The inputs slacks have both deterministic (functions of exogenous variables) and stochastic components. More generally, it can be argued that if inefficiency exists then it is natural to think of the channels through which it affects output. In the proposed model some observed exogenous variables (z) and an unobserved one-sided random variable affect output neutrally. This is called the *direct channel* and it is what the traditional frontier models considers. The study goes beyond this and argues, in the spirit of Leibenstein, that these observed z variables along with some unobserved variables can also affect output *indirectly* by affecting input effectiveness (productivities) which are captured by the input slacks. This *indirect channel* is 'new' and has not been used in the traditional stochastic frontier models. The new model also works if no z variables are available to explain input effectiveness.

The present study used a flexible functional form (the translog) to represent the underlying production technology. The composed error term in such a model becomes a nonlinear function of several error components, viz., the one-sided input slack vector (the dimension of which depends on the number of inputs), one-sided neutral technical inefficiency and two-sided random noise. Identification of two sets of one-sided errors is possible in a translog model because the vector of one-sided input slacks appears in additive form as well as interactively with the (log) inputs. Distributional assumptions on technical inefficiency and slacks also help in identification. Bayesian inference techniques organized around Markov Chain Monte Carlo are used, especially the Gibbs sampler with data augmentation, to estimate these inefficiency components. Statistical inference in the new model is provided using MCMC techniques that are efficient and easy to work with. Since input slacks in the new models are new one can also test formally the validity of the new specification against only neutral component, and the presence of ISTI in only one input using marginal likelihoods. Furthermore, proper account is made for endogeneity of inputs and the study provided a test for weak instruments.

The paper also provides an empirical application of the new methodology to a large data on U.K. manufacturing. The study documents significant differences in the estimated inefficiency after correcting for endogeneity. Slacks associated with labor and capital are found to be

2.35% and 10.74%, on average. Mean output loss from neutral inefficiency is 2.43% (with a standard deviation of 1.83%), while output loss from input slacks is, on average, 9.2% with a standard deviation of 4.3%. Since output is real sales, these findings suggest that sales revenue, on average, is reduced by 11.63% due to both types of inefficiency. In terms of future research, it would be important to extend the model to the case of input-oriented or output-oriented (and perhaps other directional) distance functions. Although this extension is straightforward econometrically, it is, nevertheless, important for applied research. Another extension would be to incorporate firm effects and perhaps distinguish them from inefficiency.

Appendix A

LIST OF ABBREVIATIONS AND SHORT DESCRIPTION OF SOME OF THE CONCEPTS USED

Abbreviations:

CR concentration ratio
DDF directional distance function
DEA data envelopment analysis
FP financial pressure
GSM generalized stochastic frontier.
KOS Koop et al. (2000), see References.
IMP import penetration
ISTI input-specific technical inefficiency
MCMC Markov Chain Monte Carlo
MS market share
RTS returns to scale
SF(M) stochastic frontier (model)
TREND time trend
VAR vector autoregression
IV Instrumental Variables

SHORT DESCRIPTIONS OF SOME OF THE CONCEPTS USED

Bayesian methods. Given a set of data, the parameters θ can be estimated using either sampling-theory methods like maximum likelihood, method of moments, etc., or Bayesian techniques. In Bayesian analysis, it is assumed that there is prior information about the parameters. It uses the Bayes' theorem which states that the posterior distribution of the parameter is proportional to the product of the prior and the likelihood.

Data envelopment analysis. DEA is a linear programming analysis that can be used to provide efficiency scores for each decision-making units or firm given observations on their inputs and outputs. DEA is based on linear programming analysis, and it abstracts from the presence of noise in the data. **Stochastic frontier models**, on the other hand, take proper account of noise but use parametric assumptions about the functional form that relates inputs to outputs, unlike DEA which is based on piecewise linear approximation.

Markov Chain Monte Carlo. MCMC is a set of techniques designed to produce a long sample of (not necessarily independent) draws from the posterior distribution of a parameter θ in a given model. MCMC techniques, in effect, provide access to complicated posterior distributions which often involve multivariate integrals that cannot be computed in closed form. One such MCMC technique is the Gibbs sampler. The Gibbs sampler produces a sequence of draws that converges to the posterior by providing random numbers from the full list of conditional distributions of the parameter vector θ.

LIST OF THE MAIN SYMBOLS

Y_i output $i = 1, \ldots, n$, $n =$ number of firms.
$\Lambda_i \geq 1$ firm-specific output efficiency.
A_i shifting parameter of production function.
$(\Theta_{1i}, \ldots, \Theta_{Ji})$ input-specific efficiency factors, $j = 1, \ldots, J$inputs.
$\Theta_{Ji} \in (0, 1]$
$\underset{(J\times 1)}{\theta_i} \equiv \ln \underset{(J\times 1)}{\Theta_i}$
z_i vector of variables representing shifters of technology (viz. determinants of A_i).
$\underset{(J\times 1)}{x_i}$ vector representing log inputs of production, $\ln X_{ji} = x_{ji}$
y_i represents log of output.
$\tilde{z}_i$ represents shifting variables of the frontier.
z_i is vector $\tilde{z}_i$plus a column of ones for the intercept.
v_i represents the composed error term.

ε_{0i} represents the error term in $\ln A_i$.
$u_{0i} \leq 0$ represents output-specific technical inefficiency.
$q_i = \alpha + Bx_i$ is the output elasticity vector ($\partial y_i/\partial x_i$), where vector and matrix α, Bcontain parameters of the production function.
$\overline{\theta}_i = \Delta z_i + \overline{u}_i$,
$\overline{u}_i = E(u_i|x_i, z_i)$,
$\overline{u}_{0i} = E(u_{0i}|x_i, z_i)$
$\Lambda_i = Cov(\theta_i)$, for each $i = 1, ..., n$.
$N_J(0, \Omega)$ denotes a J-variate normal distribution with mean vector zero, and variance-covariance matrix Ω.
l is a $J \times 1$ column vector of ones.
δ and $\Delta_0 = vec(\Delta)$are regression parameters.

Appendix B. Markov Chain Monte Carlo computations

Given the structural parameters ϕ, consider the augmented parameter vector $\Phi = \begin{bmatrix} \phi \\ u_0 \\ vec(\Theta) \end{bmatrix}$. The notation $\Phi_{-\omega}$ denotes the parameter vector δ, except the element(s) of ω, where ω is a proper subvector of Φ. In this appendix we provide some details for the MCMC scheme we have used to perform the computations.

For the regression parameters δ of the translog, we have:

$$\delta \Big| \Phi_{-\delta}, y, X, Z \sim N\left(\widehat{\delta},\ \widehat{V}_\delta\right),$$

where $\widehat{\delta} = \left[W(\Theta)' W(\Theta) + \sigma^2 \underline{V}_\delta^{-1}\right]^{-1} \left[W(\Theta)'(y - u_0) + \sigma^2 \underline{V}_\delta^{-1} \underline{\delta}\right]$, and $\widehat{V}_\delta = \sigma^2 \left[W(\Theta)' W(\Theta) + \sigma^2 \underline{V}_\delta^{-1}\right]^{-1}$.

For the scale parameters, σ_u^2 and σ^2, we obtain:

$$\frac{u_0' u_0 + \underline{q}_u}{\sigma_u^2} \Big| \Phi_{-\sigma_u^2}, y, X, Z \sim \chi^2\left(n + \underline{n}_u\right),$$

$$\frac{S + \underline{q}}{\sigma^2} \Big| \Phi_{-\sigma^2}, y, X, Z \sim \chi^2\left(n + \underline{n}\right),$$

where $S = (y - W(\Theta)\delta - u_0)'(y - W(\Theta)\delta - u_0)$.

Here, $\chi^2(\nu)$ denotes a chi-square distribution with ν degrees of freedom. We note that a standard property of the *chi*-square distribution is: $\frac{Q}{\sigma^2} \sim \chi^2(\nu)$ for a random variable σ^2 whose sum of squares in the sample is Q and the number of observations is ν. This is also known as type II inverted *gamma* distribution.

The n elements of u_0 are conditionally independent in the posterior, and their posterior conditional distribution is given by:

$u_{0i}|\Phi_{-u_0}, y, X, Z \sim N_-(\widehat{u}_{0i}, \sigma_*^2)$, for each $i = 1, ..., n$,where $\sigma_*^2 = \frac{\sigma^2 \sigma_u^2}{\sigma^2 + \sigma_u^2}$, $\widehat{u}_{0i} = \sigma_*^2 \frac{\varepsilon_i}{\sigma^2}$, $\varepsilon_i = y_i - w_0'(\theta_i)\delta$, and $N_-(\widehat{u}_{0i}, \sigma_*^2)$ denotes a normal distribution with the indicated moments, truncated to $(-\infty, 0]$[10].

Regarding the latent variables θ_i, they are conditionally independent in the posterior, and we have:

$$p(\theta_i|\Phi_{-\theta_i}, y, X, Z) \propto \exp\left[-\frac{1}{2} \frac{\left(e_i - \theta_0' q_i - \frac{1}{2}\theta_0' B \theta_i\right)^2}{\sigma^2} - \frac{1}{2}(\theta_i - \Delta z_i)' \Omega^{-1} (\theta_i - \Delta z_i) \right]$$

where $\tilde{e}_i = y_i - w_i'\delta - u_{0i}$, $e_i = \tilde{e}_i - \overline{\theta}_0' B \overline{\theta}_i$, and $q_i = \alpha + Bx_i$ (as in (4)). The distribution is nonstandard due to the presence of the term $\theta_i' B \theta_i$. We form a proposal (or "approximating") distribution as follows. Suppose $\overline{\theta}_i$ is a known "reasonable" approximation to θ_i. Then the above conditional posterior kernel is approximately:

$$Q(\theta_i) \propto \exp\left[-\frac{1}{2} \frac{\left(e_i - \theta_i' q_i\right)^2}{\sigma^2} - \frac{1}{2}(\theta_i - \Delta z_i)' \Omega^{-1} (\theta_i - \Delta z_i) \right],$$

After completing the square, this can be shown to be distributed as $N_J(\widehat{\theta}_i, \widehat{V}_i)$ subject to $\theta_i \leq 0_{(J\times 1)}$, where $\widehat{\theta}_i = (q_i q + \sigma^2 \Omega^{-1})^{-1}(e_i q_i + \sigma^2 \Omega^{-1} \Delta z_i)$ and $\widehat{V}_i = \sigma^2 (q_i q_i' + \sigma^2 \Omega^{-1})^{-1}$ for $i = 1, ..., n$.

We use this $N_J(\widehat{\theta}_i, \widehat{V}_i)$ as a proposal distribution for generating a candidate vector $\theta_i^{(c)}$. Denote by $f_N(\widehat{\theta}_i, \widehat{V}_i)$ the density of this proposal. The candidate vector $\theta_i^{(c)}$ is generated from $N_J(\widehat{\theta}_i, \widehat{V}_i)$, subject to the constraint that $\theta_i^{(c)} \leq 0_{(J\times 1)}$. With probability $\min\left(1, \frac{p(\theta_i^{(c)}|\Phi_{-\theta_i}, y, X, Z)/Q(\theta_i^{(c)})}{p(\theta_i^{(0)}|\Phi_{-\theta_i}, y, X, Z)/Q(\theta_i^{(0)})}\right)$, the

[10] Drawing from this distribution is quite standard (see for example Kumbhakar and Tsionas (2005)).

candidate vector is accepted, else we repeat the existing draw, $\theta_i^{(0)}$. See Tierney (1994). Drawing from a multivariate normal distribution subject to inequality constraints has been considered before (Geweke, 1991).[11]

To determine a "reasonable" approximation $\overline{\theta}_i$, for each $i = 1,...,n$, we start by setting $\overline{\theta}_i = 0$, which corresponds to full input efficiency, and run the MCMC scheme for M_1 iterations. At the end of the M_1 iterations we set $\overline{\theta}_i$ equal to the means of the θ_is. We rerun the MCMC for another M_1 iterations and obtain the new means. The new means should be closer to a "reasonable" approximation of $\overline{\theta}_i$. The Gibbs sampler is run again[12] for M_2 iterations but now we use as proposals distributions of the form $N_J(\widehat{\theta}_i,\ h_i\widehat{V}_i)$, and adjust the h_is every 100 iterations to obtain acceptance rates close to 25–30%%. After completion, we take the proposals as given, and run a final MCMC scheme for M iterations to obtain final results (without any other tuning of the parameters). In practice we have used $M_1 = M_2 = M$(one million iterations). In our previous work (Kumbhakar and Tsionas, 2006) we have used a simplified procedure for a different but somewhat related model. In the context of the present model, we assumed $\overline{\theta}_i = 0$, and used the resulting J-variate normal as a proposal distribution. The procedure developed here is somewhat more complicated, but much more accurate.[13] It remains to propose draws from the conditional posterior distributions of Ω, and $\Delta_0 = vec(\Delta)$. The conditional posterior distribution of Ω, is the following:

$$p(\Omega|\Phi_{-\Omega},y,X,Z) \propto \left[\prod_{i=1}^{n} C_i(\Omega,\Delta)\right]^{-1} |\Omega|^{-\left(n+\underline{n}_{\Omega}\right)/2}\cdot\exp\left[-\frac{1}{2}tr\Omega^{-1}\left(\Sigma+\underline{\Sigma}\right)\right],$$

where $\Sigma = (\Theta - Z\Delta')(\Theta - Z\Delta')'$, as we defined previously. If it were not for the factor $K(\Omega,\Delta) = [\prod_{i=1}^{n} C_i(\Omega,\Delta)]^{-1}$, the conditional distribution would be Wishart with parameters $n+\underline{n}_{\Omega}$, and $\Sigma+\underline{\Sigma}$.

Random number generation from Wishart distributions is straightforward. However, the term $K(\Omega,\Delta)$ complicates things considerably, the distribution above is no longer a Wishart, and it is not a member of other known families. Random number generation is accommodated using a "Metropolis step"). Given a current draw, say $\Omega^{(0)}$, suppose $\Omega^{(c)}$ is a candidate matrix draw from the Wishart distribution, with parameters $n+\underline{n}_{\Omega}$, and $\Sigma+\underline{\Sigma}$. Since $K(\Omega^{(0)},\Delta)$ has already been computed,[14] the candidate is accepted with probability $\min(1,\ R)$, where $R = \frac{K(\Omega^{(c)},\Delta)}{K(\Omega^{(0)},\Delta)}$. The only overhead is the computation of $1/K(\Omega^{(c)},\Delta) = \prod_{i=1}^{n} C_i(\Omega^{(c)},\Delta)$, and therefore the computation of the multivariate normal integrals $C_i(\Omega^{(c)},\Delta)$, for each $i = 1,...,n$.

For the regression coefficients in Δ of the determinants of θ_is, we obtain:

$$p(\Delta|\Phi_{-\Delta},y,X,Z) \propto K(\Omega,\Delta)\cdot\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\theta_i-\Delta z_i)'\Omega^{-1}(\theta_i-\Delta z_i)\right]\cdot p(\Delta),$$

where $p(\Delta)$ is the induced matric-variate analogue of the prior on $\Delta_0 = vec(\Delta) \sim N\left(\underline{\Delta}_0,\underline{V}_{\Delta_0}\right)$. Ignoring, for the moment, the product of multivariate normal integrals in $K(\Omega,\Delta)$, it is clear that an approximation to the conditional posterior distribution above, would be

$$Q(\Delta) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\theta_i-\Delta z_i)'\Omega^{-1}(\theta_i-\Delta z_i)\right]\cdot p(\Delta),$$

The equations $\theta_i = \Delta z_i + \Xi_i$, $\Xi_i \sim N_J(0,\Omega)$, for $i = 1,...,n$, which relate to the likelihood generated by $Q(\Delta)$, can be written equivalently as: $\theta_{(j)} = Z\Delta_{(j)} + \xi_j$, for $j = 1,...,J$, where Z is the familiar $n \times m$ matrix of data on the determinants of input efficiencies, $\theta_{(j)}$ is an $n \times 1$ vector containing the elements of the jth column of Θ, $\Delta_{(j)}$ is a vector of dimension $m \times 1$ containing the elements of the jth row of matrix Δ, and ξ_j is an $n\times 1$ vector of error terms containing the elements of the jth row of Ξ. Given our assumptions on Ξ_i, we have $\xi_j \sim N_n(0,\ \Omega \otimes I_n)$. This is the multivariate regression model analyzed by Tiao and Zellner (1964).

In the case of "non-informative priors", $p(\Delta_{(j)},\Omega) \propto |\Omega|^{-(J+1)/2}$, and the same set of regressors in each equation of the multivariate regression model, the marginal posterior conditional distribution of $\Delta_{(j)}$ is: $\Delta_{(j)}|\Phi_{\Delta_{(j)}},y,X,Z \sim N_J(\widehat{\Delta}_{(j)},V_j)$, where $\widehat{\Delta}_{(j)} = (Z'Z)^{-1}Z'\theta_{(j)}$, $V_j = \omega_{jj}(Z'Z)^{-1}$, and $\Omega \equiv [\omega_{ij},i,j = 1,...,J]$. See equation (3.6) in Tiao and Zellner (1964), or equation (3.4) for the entire vector of $\Delta_{(j)}$s (for all $j = 1,...,J$). Similar expressions hold, of course, when informative priors are placed on $\Delta_0 = vec(\Delta)$, or equivalently the $\Delta_{(j)}$s ($j = 1,...,J$) as we do here. See also Zellner (1971, pp. 240–242).

The quantities $\widehat{\Delta}_{(j)}$ and V_j, is computed by least squares regressions (on an equation by equation basis) conditional on Ω. In turn, it is straightforward to form a proposed vector $\Delta_{(j)}^{c}$, for each $j = 1,...,J$, and finally form a proposed matrix, say $\Delta^{(c)}$, of dimension $J \times m$. Given an existing draw $\Delta^{(0)}$, operation of a "random walk Metropolis step" involves accepting the proposed draw will probability $\min\left(1,\ \frac{K(\Omega,\Delta^{(c)})}{K(\Omega,\Delta^{(0)})}\right)$. Based on the existing value of the multivariate integrals involved in $K(\Omega,\Delta)$ when $\Delta = \Delta^{(0)}$, this computation involves evaluating the multivariate normal integrals involved in the integrating constant of $K(\Omega,\Delta^{(c)})$.

This completes the discussion on a single step of a Gibbs sampling scheme with data augmentation on the latent variables u_0, and $(\theta_i, i = 1,...,n)$, where $-\theta_i \in \mathbb{R}_+^J$.

[11] We thank Gary Koop for providing us with a Gauss subroutine and John Geweke for providing the original Fortran 77 program.

[12] Alternatively, one can linearize the term around a certain point θ_0. Linearization around zero involves ignoring the quadratic form and proceeding as we remarked.

[13] In the context of sampling-theory inference for nonlinear random effect models see, for example, Wolfinger (1993), Wolfinger and Lin (1997) and Kumbhakar and Tsionas (2005) for some other techniques of approximation in deriving proposal distributions in related models.

[14] Computation of multivariate normal integrals has been an active domain of research. Here, we have used the method of Drezner (1992).

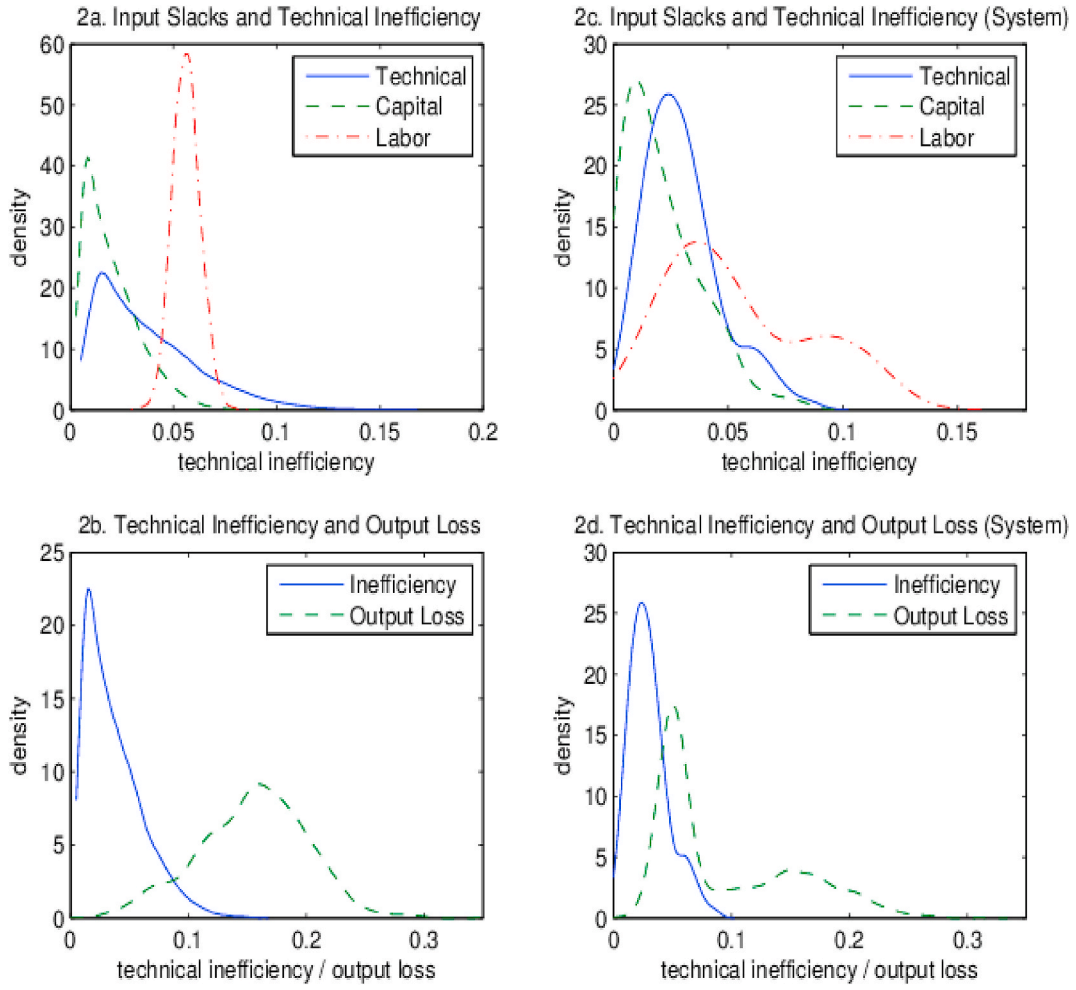


Fig. 2Input slacks and technical inefficiency and output losses. Note: Fig. 2a and c figures present the sample distributions of input slacks and technical inefficiency measures across all firms and time periods (with and with endogeneity correction). Fig. 2b and d presents the sample distributions of output losses due to input slacks and technical inefficiency measures across all firms and time periods (with and with endogeneity correction). System refers to the model with endogeneity correction.

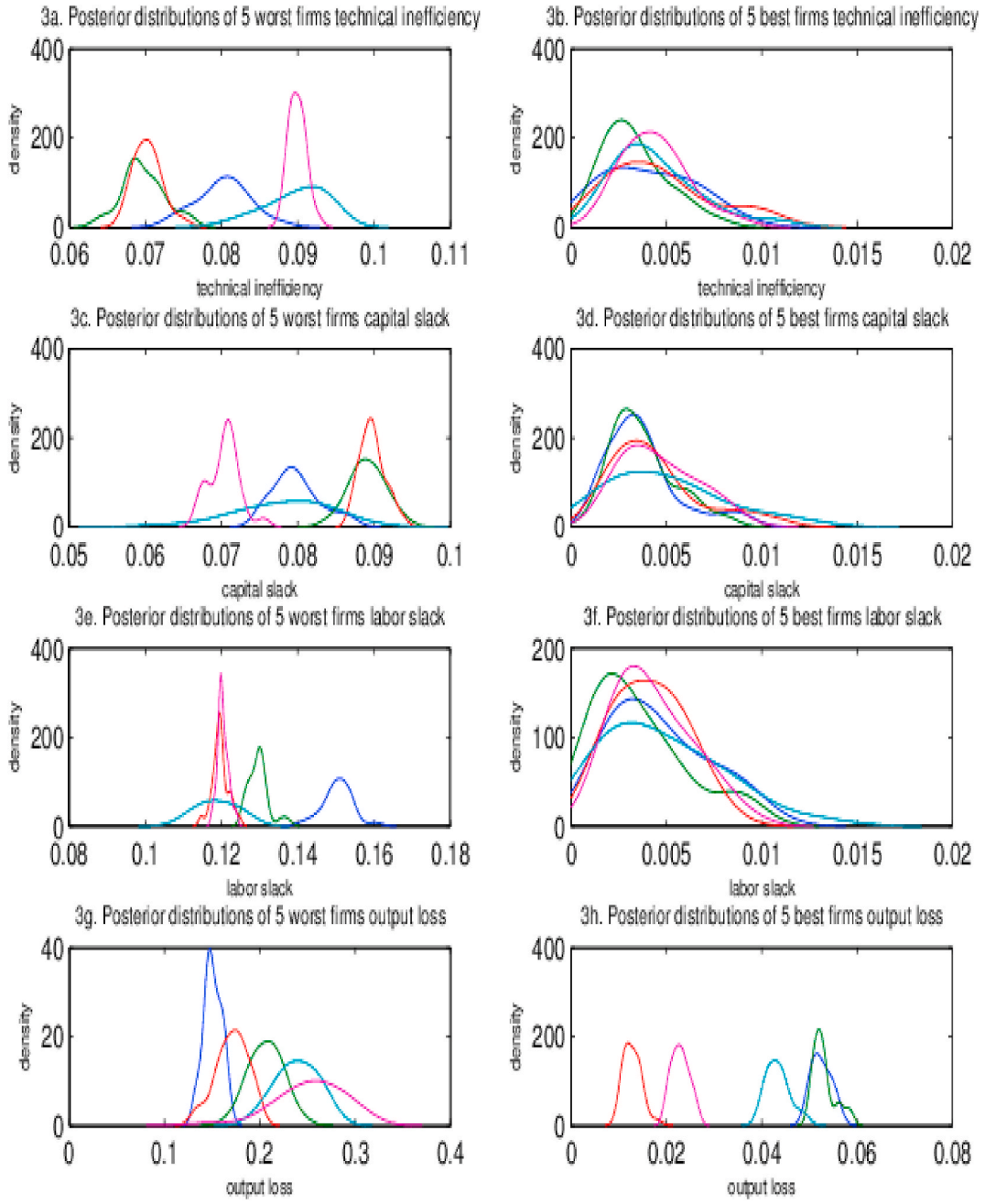


Fig. 3Posterior distributions of measures of interest for the 5 worst and 5 best firms. Note: Fig. 3a and b plot technical inefficiency distributions; 3c-3f plot capital and labor slacks; 3g and 3h plot output losses.

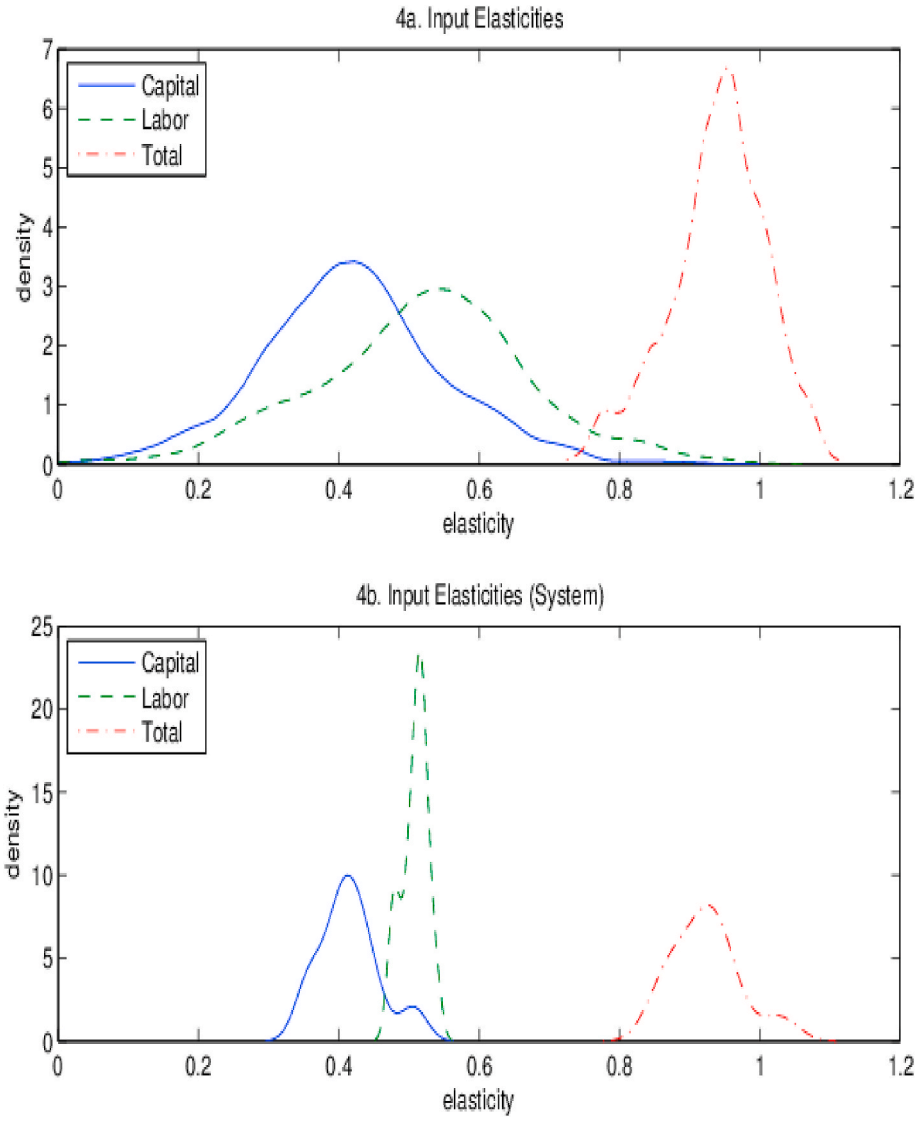


Fig. 4Distributions of input elasticities. Note: The figure presents the sample distributions of input elasticities for capital and labor. Total refers to sum of capital and labor elasticities which measure scale elasticity (returns to scale).

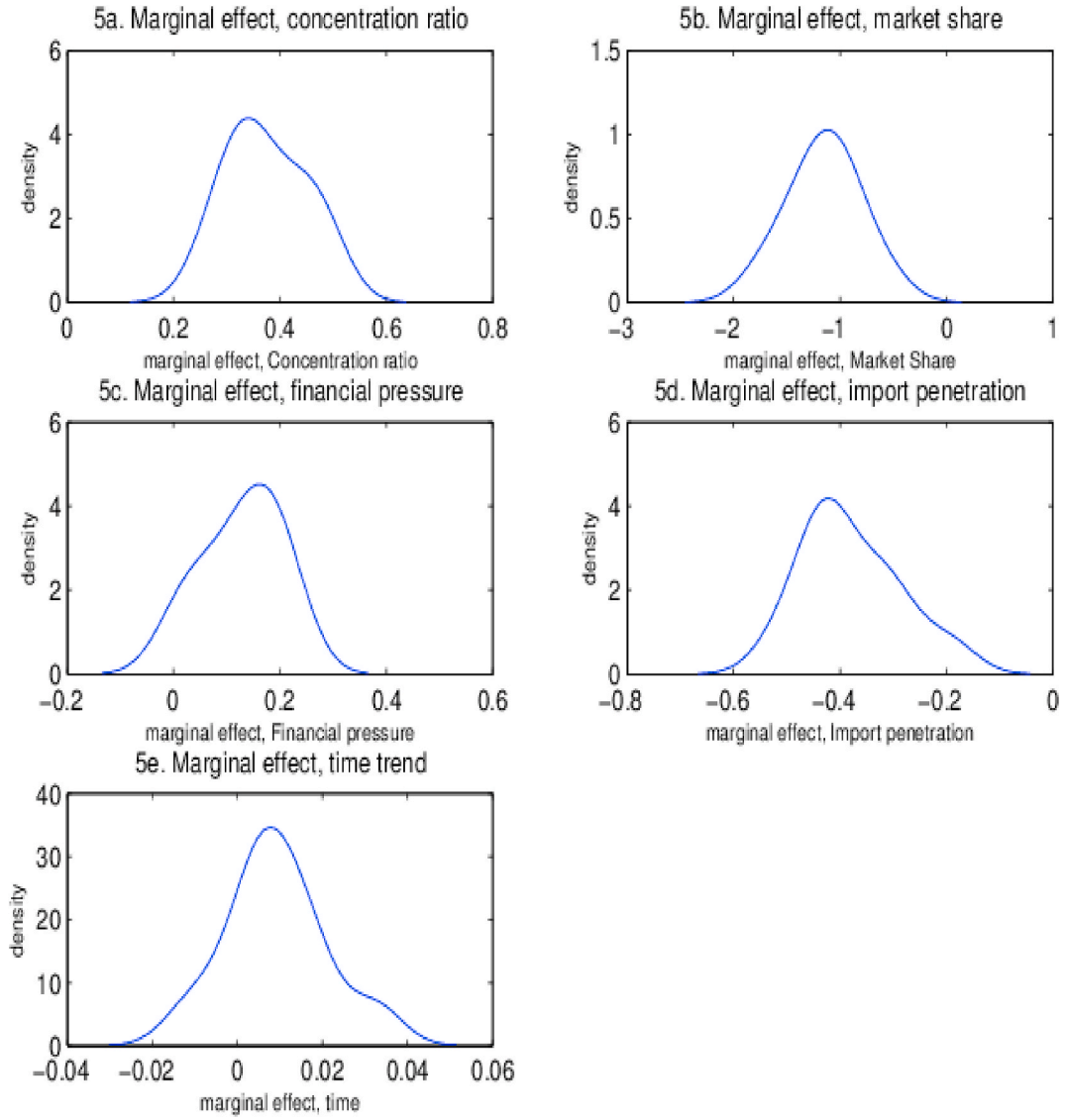


Fig. 5Posterior distributions of marginal effect of z variables on expected output. Note: Fig. 5a–d presents marginal effects of concentration ratio, market share, financial pressure, import penetration, and time trend. These marginal effects are computed based on the expression for the expectations of truncated multivariate normal variables in Tallis (1961) using numerical derivatives.

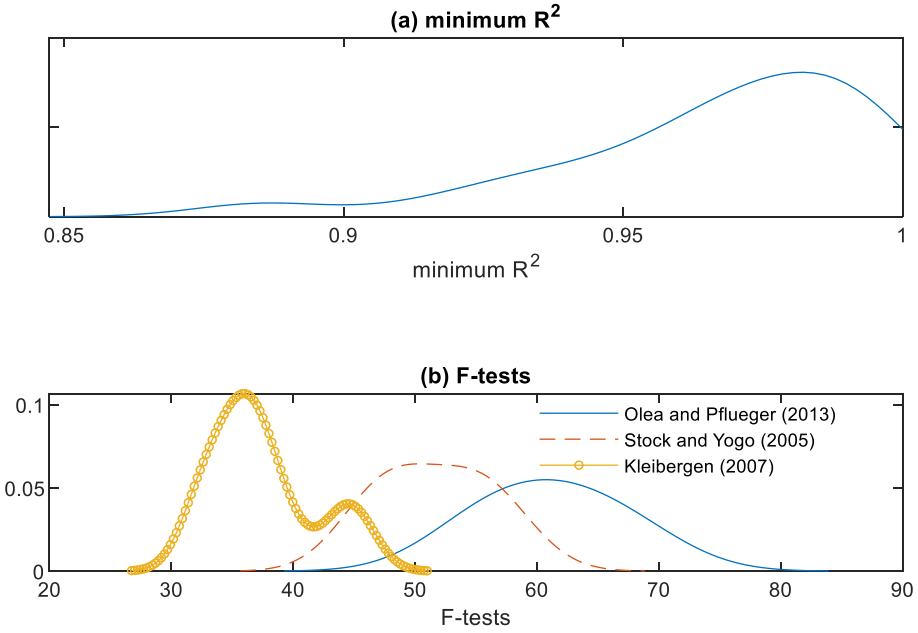


Fig. 6In panel (a) we report the minimum R^2 across equations of the panel autoregression in (12).

In panel(b) we report *F*-tests (across MCMC draws) according to Olea and Pflueger (2013), Stock and Yogo (2005), and Kleibergen (2007). See also Kleibergen (2004) and Stock et al. (2002). We repeat here equation (12) which is

$$x_{it} = \gamma_o + \underset{(J\times J)}{\Gamma^*} x_{i,t-1} + \underset{(J\times m)}{E}\tilde{z}_{i,t-1} + \underset{(J\times 1)}{\xi_{it}} \equiv \Gamma G_{it-1} + \xi_{it},$$

and the production function in equation (14):

$$y_{it} = \delta' z_{it} + \alpha'(x_{it}+\theta_{it}) + \frac{1}{2}(x_{it}+\theta_{it})' B(x_{it}+\theta_{it}) + \eta'(x_{it}-\Gamma G_{it-1}) + \omega_{it} + u_{0it}.$$

Define

$$\zeta_{it}(D_{it},\varpi) = \begin{bmatrix} x_{it} - \Gamma G_{it-1} \\ y_{it} - \delta' z_{it} - \alpha'(x_{it}+\theta_{it}) - \frac{1}{2}(x_{it}+\theta_{it})' B(x_{it}+\theta_{it}) - \eta'(x_{it}-\Gamma G_{it-1}) - u_{0it} \end{bmatrix},$$

which is the vector of errors of GSFM and the reduced form in (12). In this expression, D_{it}denotes observed data and ϖ is the vector of all structural parameters in the model.

To test the validity of instruments we test for the following orthogonality conditions:

$$(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\zeta_{it}(D_{it},\varpi)\otimes G_{i,t-1} \equiv (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi) = \mathbf{0}.$$

There are $d = (J+1)(J+m) - (J+m) = J(J+m)$ overidentifying restrictions as we have $(J+1)(J+m)$ orthogonality conditions and $J+m$instruments. In the context of generalized method of moments (GMM) estimation testing these overidentifying restrictions (which is equivalent to

testing for the validity of instruments) is performed using Hansen's statistic

$$\mathcal{J}(\varpi) = nT[(nT)^{-1}\{\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)]\}'\Omega_{nT}(\varpi)^{-1}\{\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)]\},$$

where $\Omega_{nT}(\varpi) = nT)(nT)(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)\mathbf{h}(D_{it},\varpi)']^{-1}$.When evaluated at the GMM parameter estimates, say ϖ_{GMM}, it is known that

$$\mathcal{J} \to \chi^2_d$$

, viz. the chi-square distribution with d degrees of freedom. The notation

$$\mathcal{J}(\varpi)$$

is somewhat misleading as the statistic also depends on the latent variables of the model but we do not wish to overburden notation. In a Bayesian context we cannot, of course, use the GMM estimates. Instead,

$$\mathcal{J}(\varpi)$$

is evaluated for each MCMC draw of ϖ(and the latent variables). Therefore, we can derive its marginal posterior density from which it is possible to examine formally whether

$$\mathcal{J}(\varpi)$$

is zero.

Before proceeding, we perform two tests for the validity of instruments. The first is as we described and it refers to the entire system as we consider both the GSFM and the reduced form in (12). The second, is an orthogonality test between the GSFM "residuals" $\omega_{it}\varpi) \equiv y_{it} - \delta' z_{it} - \alpha' x_{it} + \theta_{it}) - \frac{1}{2}(x_{it} + \theta_{it})' B(x_{it} + \theta_{it}) - \eta'(x_{it} - \Gamma G_{it-1}) - u_{0it}$.Therefore, this second test refers only to orthogonality between the instruments and the error of the GSFM. The two posterior densities are reported in Fig. 7.

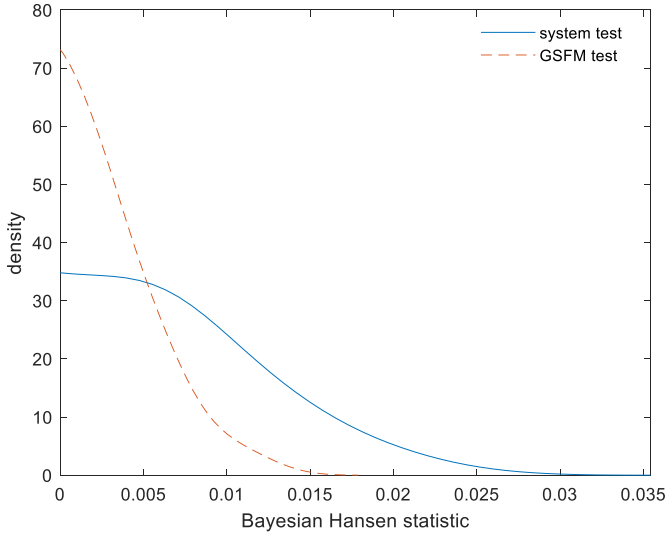


Fig. 7Marginal posterior densities of Bayesian Hansen's statistic

$$\mathcal{J}(\varpi)$$

.

In standard Bayesian fashion, all evidence about the validity of the model and the instruments are summarized in the marginal posterior densities of Bayesian Hansen's statistic

$$\mathcal{J}(\varpi)$$

. From the form of the marginal posterior densities, it appears that there is considerable probability density concentration around the origin so, naturally, we have no reason to doubt that

$$\mathcal{J}(\varpi)$$

is, in fact, zero. The results of the system test and the single equation test (based on the GSFM) are similar suggesting that the overidentifying restrictions of the reduced form are, in the light of the data, very likely to be correct.

We repeat here equation (12) which is

$$x_{it} = \gamma_o + \underset{(J\times J)}{\Gamma^*} x_{i,t-1} + \underset{(J\times m)}{E}\tilde{z}_{i,t-1} + \underset{(J\times 1)}{\xi_{it}} \equiv \Gamma G_{it-1} + \xi_{it},$$

and the production function in equation (14):

$$y_{it} = \delta' z_{it} + \alpha'(x_{it} + \theta_{it}) + \frac{1}{2}(x_{it} + \theta_{it})' B(x_{it} + \theta_{it}) + \eta'(x_{it} - \Gamma G_{it-1}) + \omega_{it} + u_{0it}.$$

Define

$$\zeta_{it}(D_{it},\varpi)=\begin{bmatrix} x_{it}-\Gamma G_{it-1} \\ y_{it}-\delta^{'}z_{it}-\alpha^{'}(x_{it}+\theta_{it})-\frac{1}{2}(x_{it}+\theta_{it})^{'}B(x_{it}+\theta_{it})-\eta^{'}(x_{it}-\Gamma G_{it-1})-u_{0it} \end{bmatrix},$$

which is the vector of errors of GSFM and the reduced form in (12). In this expression, D_{it}denotes observed data and ϖ is the vector of all structural parameters in the model.

To test the validity of instruments we test for the following orthogonality conditions:

$$(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\zeta_{it}(D_{it},\varpi)\otimes G_{i,t-1}\equiv(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)=\mathbf{0}.$$

There are $d=(J+1)(J+m)-(J+m)=J(J+m)$ overidentifying restrictions as we have $(J+1)(J+m)$ orthogonality conditions and $J+m$instruments. In the context of generalized method of moments (GMM) estimation testing these overidentifying restrictions (which is equivalent to testing for the validity of instruments) is performed using Hansen's statistic

$$\mathcal{J}(\varpi)=nT[(nT)^{-1}\{\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)]\}'\Omega_{nT}(\varpi)^{-1}\{\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)]\},$$

where $\Omega_{nT}(\varpi)=nT)(nT)(nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{h}(D_{it},\varpi)\mathbf{h}(D_{it},\varpi)']^{-1}$.When evaluated at the GMM parameter estimates, say ϖ_{GMM}, it is known that

$$\mathcal{J}\rightarrow\chi^{2}_{d}$$

, viz. the chi-square distribution with d degrees of freedom. The notation

$$\mathcal{J}(\varpi)$$

is somewhat misleading as the statistic also depends on the latent variables of the model but we do not wish to overburden notation. In a Bayesian context we cannot, of course, use the GMM estimates. Instead,

$$\mathcal{J}(\varpi)$$

is evaluated for each MCMC draw of ϖ(and the latent variables). Therefore, we can derive its marginal posterior density from which it is possible to examine formally whether

$$\mathcal{J}(\varpi)$$

is zero.

Before proceeding, we perform two tests for the validity of instruments. The first is as we described and it refers to the entire system as we consider both the GSFM and the reduced form in (12). The second, is an orthogonality test between the GSFM "residuals" $\omega_{it}\varpi)\equiv y_{it}-\delta^{'}z_{it}-\alpha^{'}x_{it}+\theta_{it})-\frac{1}{2}(x_{it}+\theta_{it})^{'}B(x_{it}+\theta_{it})-\eta^{'}(x_{it}-\Gamma G_{it-1})-u_{0it}$.Therefore, this second test refers only to orthogonality between the instruments and the error of the GSFM. The two posterior densities are reported in Fig. 8.

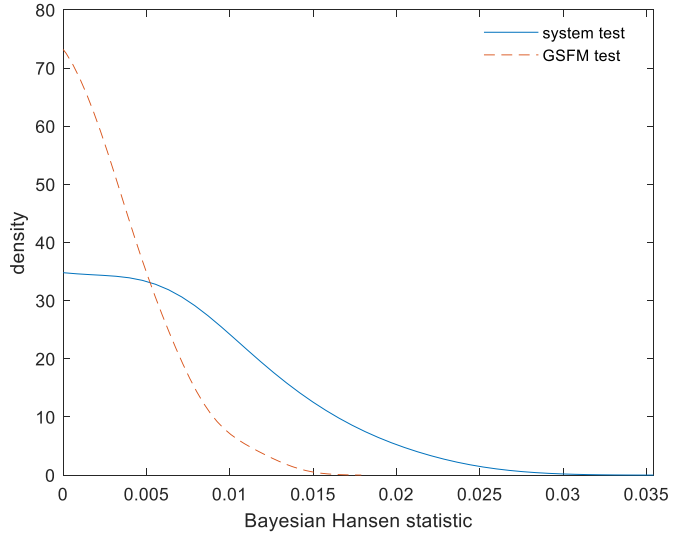


Fig. 8Marginal posterior densities of Bayesian Hansen's statistic

$$\mathcal{J}(\varpi)$$

.

In standard Bayesian fashion, all evidence about the validity of the model and the instruments are summarized in the marginal posterior densities of Bayesian Hansen's statistic

$$\mathcal{J}(\varpi)$$

. From the form of the marginal posterior densities, it appears that there is considerable probability density concentration around the origin so, naturally, we have no reason to doubt that

$$\mathcal{J}(\varpi)$$

is, in fact, zero. The results of the system test and the single equation test (based on the GSFM) are similar suggesting that the overidentifying restrictions of the reduced form are, in the light of the data, very likely to be correct.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijpe.2020.107940.

References

Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. J. Econom. 6, 21–37.

Atkinson, S., Tsionas, E., 2016. Directional distance functions: optimal endogenous directions. J. Econom. 190, 301–314.

Beckmann, M.J., Sato, R., 1969. Aggregate production function and types of technical progress: a statistical analysis. Am. Econ. Rev. 59, 88–101.

Bernstein, J.I., Mamuneas, T.P., Pashardes, P., 2004. Technical efficiency and U.S. manufacturing productivity growth. Rev. Econ. Stat. 86, 402–412.

Burmeister, E., Dobell, R., 1969. Disembodied technical change with several factors. J. Econ. Theor. 1, 1–8.

Caselli, F., Coleman II, W.J., 2006. The world technology frontier. Am. Econ. Rev. 96 (3), 499–522.

Drezner, Z., 1992. Computation of the multivariate normal integral. ACM Trans. Math Software 18, 470–480.

Färe, R., Grosskopf, Whittaker, G., 2013. Directional output distance functions: endogenous directions based on exogenous normalization constraints. J. Prod. Anal. 40, 267–269.

Geweke, J., 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In: Keramidas, E.M., Kaufman, S.M. (Eds.), Computing Science and Statistics: Proceedings of the 23rd Symposium in the Interface. Interface Foundation of North America, Fairfax, VA, 1991.

Geweke, J., 1999. Using simulation methods for Bayesian econometric models: inference, development and communication (with discussion and rejoinder). Econom. Rev. 18, 1–126.

Gollop, F.M., Roberts, M.J., 1981. The sources of economic growth in the U.S. electric power industry. In: Cowing, T., Stevenson, R. (Eds.), Productivity Measurement in Regulated Industries. Academic Press, New York.

Griliches, Z., 1994. Productivity, R&D, and the data constraint. Am. Econ. Rev. 84, 1–23.

Hampf, B., Krüger, J.J., 2015. Optimal directions for directional distance functions: an exploration of potential reductions of greenhouse gases. Am. J. Agric. Econ. 97, 920–938.

Hicks, J.R., 1935. Annual survey of economic theory: the theory of monopoly. Econometrica 3, 1–20.

Hudgins, L.B., Primont, D., 2007. Derivative properties of directional technology distance functions. In: Färe, R., Grosskopf, S., Primont, D. (Eds.), Aggregation, Efficiency, and Measurement. Studies in Productivity and Efficiency. Springer, Boston, MA.

Kapelko, M., Horta, I.M., Camanho, A.S., Oude Lansink, A., 2015. Measurement of input-specific productivity growth with an application to the construction industry in Spain and Portugal. Int. J. Prod. Econ. 166, 64–71.

Kumbhakar, S.C., Tsionas, G., 2020. On the estimation of technical and allocative efficiency in a panel stochastic production frontier system model: Some new formulations and generalizations. Eur. J. Oper. 287 (2), 762–775, 1 December 2020.

Kleibergen, F., 2004. Testing subsets of structural parameters in the instrumental variables regression model. Rev. Econ. Stat. 86 (1), 418–423.

Kleibergen, F., 2007. Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. J. Econom. 139, 181–216.

Koop, Gary, Osiewalski, J., Steel, M.F.J., 2000. Modeling the sources of output growth in a panel of countries. J. Bus. Econ. Stat. 18, 284–299.

Kopp, R.J., 1981. The measurement of productive efficiency: a reconsideration. Q. J. Econ. 96, 477–503.

Kumbhakar, S.C., 1988. Estimation of input-specific technical and allocative inefficiency in stochastic frontier models. Oxf. Econ. Pap. 40, 535–549.

Kumbhakar, S.C., Lovell, C.A.K., 2000. Stochastic Frontier Analysis. Cambridge University Press, New York.

Kumbhakar, S.C., Tsionas, E.G., 2005. The joint measurement of technical and allocative inefficiency: an application of Bayesian inference in nonlinear random effects models. J. Am. Stat. Assoc. 100 (471), 736–747.

Kumbhakar, S.C., Tsionas, E.G., 2006. Estimation of stochastic frontier production functions with input-oriented technical efficiency. J. Econom. 133, 71–96.

Kumbhakar, S.C., Parmeter, C.F., Zelenyuk, V., 2020. In: Ray, Chambers, Kumbhakar (Eds.), Stochastic Frontier Analysis: Foundations and Advances I, Forthcoming in *Handbook of Production Economics*, Springer Nature.

Leibenstein, H., 1966. Allocative efficiency vs. 'X-Efficiency'. Am. Econ. Rev. 56, 392–415.

Leibenstein, H., 1979. X -Efficiency: from concept to theory. Challenge Sept-Oct, 13–22.

Mahlberg, B., Sahoo, B.K., 2011. Radial and non-radial decomposition of Luenberger productivity indicator with an illustrative application. Int. J. Prod. Econ. 131, 721–726.

Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. Int. Econ. Rev. 18, 435–444.

Nickell, S., 1996. Competition and corporate performance. J. Polit. Econ. 104, 724–746.

Nickell, S., Nicolitsas, D., 1999. How does financial pressure affect firms? Eur. Econ. Rev. 43, 1435–1456.

Nickell, S., Nicolitsas, D., Dryden, N., 1997. What makes firms perform well? Eur. Econ. Rev. 41, 783–796.

Olea, J.S.L., Pflueger, 2013. A robust test for weak instruments. J. Bus. Econ. Stat. 31 (3), 358–369.

Oude Lansink, A., Pietola, K., Backman, S., 2002. Efficiency and productivity of conventional and organic farms in Finland 1994-1997. Eur. Rev. Agric. Econ. 29 (1), 51–65.

Ray, S.C., 2004. Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research. Cambridge University Press, New York, NY.

Sato, R., Beckmann, M.J., 1968. Neutral innovations and production functions. Rev. Econ. Stud. 35, 57–66.

Skevas, T., Oude Lansink, A., 2014. Reducing pesticide use and pesticide impact by productivity growth: the case of Dutch arable farming. J. Agric. Econ. 65 (1), 191–211.

Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Andrews, D.W.K., Stock, J.H. (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge, pp. 80–108.

Stock, J.H., Wright, J.H., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. J. Bus. Econ. Stat. 20 (4), 518–529.

Tallis, G.M., 1961. The moment generating function of the truncated multi-normal distribution. J. Roy. Stat. Soc. B 23, 223–229.

Tiao, G.C., Zellner, A., 1964. On the Bayesian estimation of multivariate regression. J. Roy. Stat. Soc. B 26, 277–285.

Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). Ann. Stat. 22, 1701–1762.

Tran, K.C., Tsionas, E.G., 2013. GMM estimation of stochastic frontier model with endogenous regressors. Econ. Lett. 118, 233–236.

Wolfinger, R.D., 1993. Laplace's approximation for nonlinear mixed models. Biometrika 80, 791–795.

Wolfinger, R.D., Lin, X., 1997. Two Taylor-series approximation methods for nonlinear mixed models. Comput. Stat. Data Anal. 25, 465–490.

Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. Wiley, New York.