ANNALS OF ECONOMICS AND STATISTICS
NUMBER 127, SEPTEMBER 2017

Fixed Cost, Variable Cost, Markups and Returns to Scale

Xi CHEN
ANEC/STATEC, Luxembourg

Bertrand M. KOEBEL
Bureau d'Économie Théorique et Appliquée (BETA), CNRS, Université de Strasbourg

This paper derives the structure of a production function which is necessary and sufficient for generating a fixed cost. We extend the classical production function in order to allow each input to have a fixed and a variable part. We characterize and estimate both the fixed and variable components of the cost function and study how fixed and variable costs interact and affect firms' behavior in terms of price setting and returns to scale.*

I. Introduction

A long tradition going back to VINER, J. [1931] considers that fixed costs correspond to the cost of fixed inputs.[1] However, splitting the whole set of inputs into two disjoint sets (with either fixed or variable inputs) does not provide a faithful description of many economically interesting technologies. This paper extends the microeconomic foundations of production analysis by allowing each input to have a fixed and a variable part. Economically, the fixed cost does not only reflect the cost of fixed inputs, but also the choice of the technology among a set of alternative possibilities for initiating production. We show that most functional specifications used in the literature are not suitable for modeling fixed costs, and propose a new cost function specification which is the sum of two (locally) flexible functional forms, one for fixed costs and one for variable costs. If inputs have two components (fixed and variable) but only total input is observed by econometricians, this introduces two unobserved and correlated heterogeneity terms that should be controlled for. Our empirical results support the extended Translog specification: the fixed cost is often significant, and neglecting it yields estimation biases,

1. In the words of VINER, J. [1931], p. 26: "It will be arbitrarily assumed that all of the factors can for the short-run be sharply classified into two groups, those which are necessarily fixed in amount, and those which are freely variable. [...] The costs associated with the fixed factors will be referred to as the "fixed costs"."

*JEL: C33, D24, L60 / KEY WORDS: Identification, Imperfect Competition, Unobserved Heterogeneity.

especially on the markup and the rate of returns to scale. It turns out that the traditional Translog specification is not particularly suitable for estimating returns to scale (KLETTE, T. J. [1999]), for evaluating benefits related to privatization (CHIRWA, E. W. [2004]), mergers (BRITO *et al.* [2013]) or vertical integration (GARCIA *et al.* [2007]).

The purpose of this paper is to characterize and estimate both the fixed and variable components of the cost function, to investigate their heterogeneity over industries and to study how fixed costs affect producers' behavior in terms of price setting and returns to scale. We follow BAUMOL, W. J., and R. D. WILLIG [1981], (p. 406) and consider the long-run fixed cost as the magnitude of the total long-run cost function when the production level tends to zero. This paper derives an extended production technology that generates the fixed cost (an issue which is usually neglected when dealing with fixed cost). We show that the extended production function takes the form $y = G(x_v, x_f)$ where the fixed part x_f of each input makes it possible to initiate production, whereas the variable part x_v is necessary to reach production level y. This means that two physically similar inputs may be technologically different. For instance, if capital is physically fixed and energy is fully variable, but capital cannot be run without say 1000 KWh of energy, then the part of the energy input which is necessary to run the fixed capital input becomes fixed. It is the production technology which determines whether inputs are variable or fixed and which part of each input is fixed or variable. This remark has important implications for the specification of fixed and variable cost functions. We show that the textbook formulation of production technology, in which fixed and variable inputs are disjoint, is obtained as a special case of our extended production function when $G(x_v, x_f)$ takes the form $F(x_v + x_f)$.

The empirical part of this paper uses panel data for 462 US manufacturing sectors observed between 1958 and 2005. We estimate the fixed costs and analyze their implications in terms of markup pricing, returns to scale and technical change. SECTION II explores the definition and microeconomic foundations of fixed costs. SECTION III compares the extended production function with the traditional one. SECTIONS IV discusses econometric issues related to fixed costs: biases when they are neglected, unobserved heterogeneity and specification issues. SECTION V provides an empirical application. Concluding remarks are given in SECTION VI.

II. A Microeconomic Framework for Fixed Costs

The definition of fixed costs is central in economics, and is briefly discussed in most introductory microeconomics textbooks (e.g. VARIAN, H. R. [1992], p. 64).[2] In this study,

2. It seems somewhat surprising, however, that the New Palgrave dictionary of economics has no entry for the term "fixed cost". The term is also not mentioned in DIEWERT'S [2008] contribution on cost functions.

we adopt the definition of fixed costs given by Baumol, W. J., and R. D. Willig [1981], where we only focus on active firms with a positive level of output.[3]

Definition 1. For an active firm, the *fixed cost* is the cost of producing an arbitrarily small amount of output.

Definition 1 does not require (at this stage) that the level of the fixed cost be optimal (the case where some inputs cannot be changed in the "short-run" is presented later). Definition 1 implies that fixed costs do not change with the production level.[4]

Despite the challenging result of Baumol, W. J., and R. D. Willig [1981], (p. 405) according to which fixed costs "do not have the welfare consequences normally attributed to barriers to entry", there is quite a large literature on fixed costs. Fixed costs are useful for explaining coordination failure (Murphy *et al.* [1989]) and international trade (Krugman, P. [1979]; Melitz, M. [2003]). Blackorby, C., and W. Schworm [1984], [1988] and Gorman, W. M. [1995] have shown that fixed inputs hamper the aggregation of production (and cost) functions, whereas a fixed cost does not represent an aggregation problem (when fixed inputs are efficiently allocated). Fixed costs are also considered in general equilibrium theory with imperfect competition (see for instance Dehez *et al.* [2003]).

The traditional literature considers that in the short-run, firms cannot instantaneously adjust their production factors to their optimal level. Morrison, C. J. [1988] introduces a short-run cost function with quasi-fixed inputs and investigates empirically the long-run implications of fixed input adjustment. In the literature on production function estimation, Olley, S., and A. Pakes [1996] also distinguish between variable and fixed inputs, and use the timing of input adjustments to construct moment conditions. Typically, they consider labor as a pure variable input and the capital stock as a quasi-fixed input. Their approach, however, does not make it possible to identify the fixed and variable components of each input separately. Contributions in the field of industrial organization on the reasons and consequences of fixed inputs are so numerous that we cannot survey them here. See Berry, S., and P. Reiss [2007] who discuss some important issues on the identification and heterogeneity of fixed inputs.

While the fixed input and its dual counterpart – the fixed cost – have been used interchangeably in some debates, the relationship between them is not always obvious. The main result of this section characterizes an extended production function that explicitly links the fixed cost to the fixed input. This extended production function is also able to describe input fixity in a more general way than those considered in the literature.

3. For inactive firms, the total cost can be always reduced to zero by closing down the production plant and exit the market.

4. Definition 1 introduces a jump in the cost function near the zero production level. Adding further discontinuities in the technology for larger production levels would not change the analysis. In this case the cost function still includes only two parts: the fixed cost, which does not vary with output, and the variable cost. Empirically, discontinuities in these functions are captured by including time in the list of explanatory variables.

Input fixity depends both on the time period of analysis and on which input can be adjusted during that time period. Thus, it is crucial to distinguish firms' short-run behavior from their long-run behavior. In the following subsections, we study the fixed cost by explicitly introducing a timing in the input adjustment.

II.1. *The Short-Run Fixed Cost*

Let us introduce an extended production function G as $y = G(x_v, x_f)$ where $G : \mathbb{R}_+^J \times \mathbb{R}_+^J \to \mathbb{R}_+$. Unlike the traditional approach, the input set is not disjoint in the sense that a given input, say energy, can appear twice as an argument of G: once in vector x_v and once in x_f; so their marginal productivities can differ. This specification implies that similar inputs can be used for different types of production activities. Engineers, for instance, can be allocated to production, administrative tasks or research and development activities. While engineers' production increases the current output level, this is not the case when they are allocated to administrative tasks, which withdraws them from production. Similarly, computers can be used either for logistics, production management or accounting, activities which do not have the same impact in terms of production and cost.

Technology G is more general than that usually considered in production analysis. A formal comparison of the extended and traditional production functions is given in Section III. For simplicity, we assume that:

Assumption 1 (A1). The production function G is single valued and twice continuously differentiable. Moreover,

(i) G is nondecreasing and quasi-concave in x_v
(ii) G is nondecreasing in x_f
(iii) G is nondecreasing in x_f for some values of x_v
(iv) the Hessian matrix of G with respect to x_v is not singular.

Assumption A1(i) restates the properties commonly presented in microeconomic textbooks. Although A1(ii) regarding the monotonicity of the production function with respect to the fixed inputs is standard (e.g. Varian, H. R. [1992], p. 7), we prefer to relax it in A1(iii), and require instead that G can be nondecreasing in x_f for some, but not all values of x_v.[5] This allows for technologies which are not uniformly superior to the alternative production technologies, but only for some specific range of output. Indeed, if instead of A1(iii) we require A1(ii) and assume that G is nondecreasing in x_f for all x_v this excludes the interesting case where the alternative technology $x_f' > x_f$ is more productive for large values of the variable inputs $x_v' > x_v$, but less productive for low values of x_v. In analytical terms A1(ii) is not compatible with

$$G(x_v', x_f') \geq G(x_v', x_f) \qquad \text{and} \qquad G(x_v, x_f') < G(x_v, x_f)$$

5. We are indebted to a referee who highlighted this point to us.

for $x_v' > x_v$. In fact, Assumption A1(ii) implies strong restrictions on the alternative technological choices available to the firm, which seem difficult to reconcile with modern production technologies, compatible with the occurrence of irreversibilities, lock-in effects and adjustment costs. Assumption A1(iv) ensures that the implicit theorem can be applied.

In the short-run, before production actually starts, firms are exogenously assigned a production technology G and a fixed level of x_f. In order to start production, the level of fixed input x_f must belong to the following set of fixed input requirements.

Definition 2. In terms of the extended production function G, the *fixed input requirement set* X_G is defined as:[6]

$$X_G \equiv \lim_{\varepsilon \to 0^+} \{z \geq 0 : G(0,z) = \varepsilon\}. \tag{1}$$

Definition 2 introduces the set of all fixed input combinations required for starting production. We assume that $X_G \neq \varnothing$. Then, given a fixed level of $x_f \in X_G$, firms start to produce y by employing variable inputs x_v.

Whereas G has well-known properties w.r.t. (x_v, x_f), things become different once we consider total inputs $x = x_f + x_v$. FIGURE 1 below illustrates this point for $J = 2$, in the (x_1, x_2)-plane. Both axes represent the *total* quantity of each input: $x_1 = x_{v1} + x_{f1}$. FIGURE 1 represents the isoquants for technology G and two different fixed input vectors x_f^0 and x_f^1. The bold line represents the fixed input requirement set. With technology G, the level of fixed inputs determines the substitution possibilities between the variable inputs. Although we have not introduced any distinction between *ex-ante* and *ex-post* technologies in our model, FIGURE 1 resembles those typically obtained with putty-putty (or putty-clay or clay-clay) technologies (see e.g. FUSS, M. A. [1977]). The similarity is due to the fact that we split x into two (fixed and variable) non-additive components in the production function. With technology G, a particular fixed input level x_f coincides with a particular production technology and a specific substitution pattern between variable inputs. In FIGURE 1, the isoquant corresponding to x_f^0 characterizes inputs which can easily be substituted for each other, whereas for x_f^1 substitution becomes more difficult. Note that, for a given output level, the isoquants for G corresponding to the fixed input level x_f^0 can cross those obtained for x_f^1. For instance, at point A the production level y^0 can be produced using two types of technologies, each exhibiting a specific substitution pattern.

6. The notation 0^+ denotes a positive number arbitrarily close to zero.

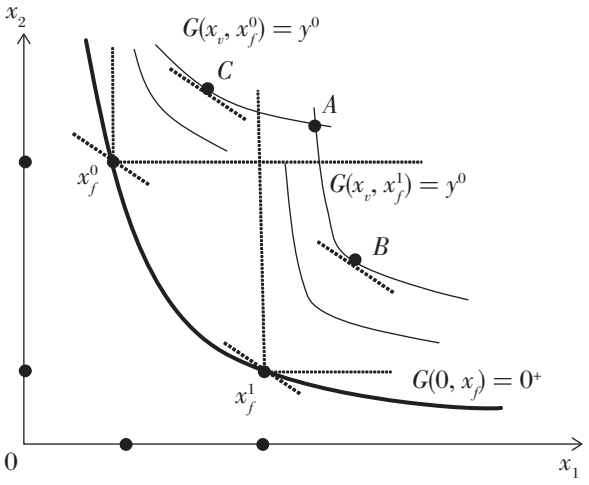

FIGURE 1. – Isoquants for Technology G

FIGURE 1 also illustrates that if fixed inputs are neglected, production function G is not necessarily quasi-concave in x (at point A). Moreover, the input bundles can be located in the zone violating quasi-concavity in x, so the cost function will not necessarily be concave in w. In the context of fixed cost, simultaneously imposing concavity in w and $x_f = 0$ on the cost function may end up with worse estimates than extending the cost function to be compatible with the occurrence of fixed cost (see LAU, L. J. [1978], and DIEWERT, W. E., and T. J. WALES [1987], for seminal contributions on concavity enforcement).

We now consider the short-run variable cost function:

$$v_r(w, x_f, y) = \min_{x_v} \left\{ w^\top x_v : G(x_v, x_f) \geq y \right\}. \tag{2}$$

The next properties are a straightforward extension of those available for the traditional cost function.[7]

Proposition 1. Under A1, if $x_f \in X_G$ then,
(i) $\lim_{y \to 0^+} v_r(w, x_f, y) = 0$,
(ii) v_r is nondecreasing in y
(iii) v_r is nonincreasing in x_f if G is nondecreasing in x_f.

Proposition 1 means that the short-run variable cost function v_r satisfies the properties of a variable cost function: it is vanishing for arbitrarily small production levels. As a consequence, the short-run total cost function is:

7. We only give those properties which are the most interesting for our purpose, see LAU, L. J. [1976] and BROWNING, M. J. [1983] for further properties.

$$c_r(w,x_f,y) = u_r(w,x_f) + v_r(w,x_f,y), \tag{3}$$

with the short-run fixed cost given by

$$u_r(w,x_f) \equiv \lim_{y\to 0^+} c_r(w,x_f,y) = w^\top x_f. \tag{4}$$

Note that both fixed and variable cost functions depend on the input prices of all inputs, and on the whole vector *xf*. This contrasts with traditional production theory (see below).

II.2. *The Long-Run Fixed Cost*

In the long-run, firms determine their level of fixed inputs endogenously by minimizing the short-run total cost function (3). For a given vector of (w,y), the inner solution for x_f^* is characterized by the equality between the shadow value of the fixed input and its market price:

$$-\frac{\partial v_r}{\partial x_{fj}}(w,x_f^*,y) = w_j. \tag{5}$$

This situation corresponds to what VINER, J. [1931] calls the *long-run*. For BLACKORBY, C., and W. SCHWORM [1984] it characterizes *efficiently allocated fixed inputs*.

Although there are no fixed inputs in the long-run, one important difference with VINER, J. [1931], (p. 28) is that we do not necessarily consider that "There will therefore be no costs which are technologically fixed in the long-run". If it turns out that $0 \in X_G$ and

$$v_r(w,0,y) \equiv \min_{x_v\geq 0}\{w^\top x_v : y \leq G(x_v,0)\} < v_r(w,x_f,y) + w^\top x_f,$$

for any $x_f > 0$ then $x_f^* = 0$ and there is no fixed cost in the long-run.[8] However, if $0 \notin X_G$ and $X_G \neq \varnothing$ then $u_r(w,x_f) > 0$ and a fixed cost occurs at any value of input prices w.

The long-run cost function is:

$$c(w,y) = c_r(w,x_f^*(w,y),y) = w^\top x_f^*(w,y) + v_r(w,x_f^*(w,y),y).$$

When the fixed inputs are optimal, the long-run cost function is always smaller than the short-run one. As the set X_G defined in (1) is independent of the input prices, it turns out that $x_f^*(w,0^+)$ is not necessarily included in X_G and so $v_r(w,x_f^*(w,0^+),0^+) \neq 0$. However, we achieve the additive decomposition:

$$c(w,y) = u(w) + v(w,y), \tag{6}$$

8. Here, the notation $x > 0$ means that all J components $x_j > 0$. In contrast $x \geq 0$ means that $x_j \geq 0$ for all j.

with $v(w,0^+)=0$ if the long-run fixed and variable cost functions are respectively defined by:

$$u(w) = v_r(w,x_f^*(w,0^+),0^+) + w^\top x_f^*(w,0^+) \tag{7}$$
$$v(w,y) = v_r(w,x_f^*(w,y),y) - v_r(w,x_f^*(w,0^+),0^+).$$

The optimal condition (5) illustrates the fundamental identification problem occurring when fixed inputs are optimally adjusted: the fixed cost generally differs from the cost $w^\top x_f^*(w,0^+)$ of the inputs which were fixed in the short-run. Indeed, after normalizing the variable and fixed cost function according to (7), we obtain the expression of the fixed cost function u, which comprises the part of the long-run variable cost function evaluated at $y=0^+$.

When fixed and variable inputs are imperfect substitutes for each other, the optimal amount of fixed input depends upon w, and $x_f^*(w,0^+)$ is not necessarily included in the input requirement set X_G. This means that, in the long-run, the set of fixed inputs cannot be determined ex-ante using only the definition of X_G. When x_f is optimally adjusted, it is no longer possible to separately identify x_f and x_v. Briefly, an input cannot be said to be fixed or variable *prima facie*, using only the physical properties of the inputs. It is the technology which in the last instance determines whether a given input is fixed or variable. Few technologies make it possible to obtain an optimal level of x_f^* independently of y. We characterize them below.

Proposition 2. Let us assume A1i, A1iv and that $x_v^*>0$ at the optimum. Let $K:\mathbb{R}_+^J\to\mathbb{R}_+^J$ and $F:\mathbb{R}_+^J\to\mathbb{R}_+$ both be increasing functions.

(i) The short-run cost function is given by

$$c_r(w,x_f,y) = u_r(w,x_f) + v(w,y), \tag{8}$$

with $\lim_{y\to 0^+} v(w,y)=0$ if and only if the production function is given by

$$G(x_v,x_f) = F(x_v + K(x_f)). \tag{9}$$

(ii) Assume that an inner solution to the long-run problem exists. In the long-run, the optimal level of x_f is independent of y if and only if the short-run cost function is (8) or the production function is (9).

Proposition 2 characterizes the cost and production functions yielding fixed inputs which do not depend upon y at the optimum. Note that requirement (9) does not impose that $K(x_f)$ be a unique aggregate fixed input. Here, the vector valued function K comprises J aggregates for the fixed inputs. The production function in Proposition 2 also aggregates additively fixed and variable inputs together since F depends upon $x_v + K(x_f)$. However, (9) is more general than $F(x_v + x_f)$, for which fixed and variable inputs are perfect substitutes.

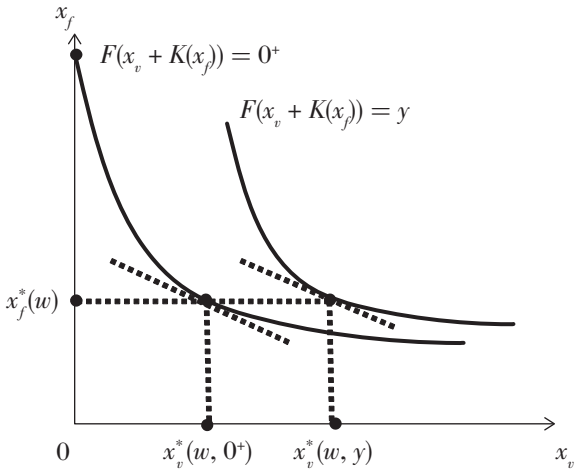

FIGURE 2. – Isoquants for Technology $F(x_v + K(x_f))$

FIGURE 2 provides an illustration in the one inputs case $(J = 1)$. It shows that x_f^* does not vary with y, unlike x_v^*. FIGURE 2 also shows how technology $F(x_v + K(x_f))$ differs from $F(x_v + x_f)$. With $F(x_v + K(x_f))$ there is perfect substitutability between the components of x_v and $K(x_f)$, but not between x_v and x_f. For a given input x_i, the slope $(\partial F / \partial x_{vi}) / (\partial F / \partial x_{fi})$ of the isoquant (FIGURE 2) is not restricted to be equal to -1 out of the optimum. Moreover, for two different inputs, x_h and x_i, the slope $(\partial F / \partial x_{fh}) / (\partial F / \partial x_{fi})$ of the isoquant is not restricted to be equal to $(\partial F / \partial x_{vh}) / (\partial F / \partial x_{vi})$ out of the optimum. Fixed inputs can be substituted for each other according to a pattern that is different to variable inputs.

III. On the Traditional Production Analysis

The previous section presents a technology G that is able to generate a fixed cost (as defined by Baumol and Willig) in both short- and long-run cases. The current section compares this technology with that usually considered in production analysis. It is shown that the traditional production function $y = F(x)$ with $F : \mathbb{R}_+^J \to \mathbb{R}_+$ increasing and quasi-concave in x, is obtained from G as a special case.

A shortcoming of the traditional approach is that it relies on a partition of inputs for generating fixed costs. This approach considers two disjoint categories of inputs: those which can be adjusted (variable inputs, denoted $\tilde{x}$) and those which are fixed or quasi-fixed $(\overline{x})$:

$$x = \begin{pmatrix} \tilde{x} \\ \overline{x} \end{pmatrix} \in \mathbb{R}_+^J. \tag{10}$$

The corresponding input prices are denoted by $(\tilde{w}^\top, \overline{w}^\top)^\top \in \mathbb{R}_{++}^J$.

The following proposition characterizes the short-run cost function when the technology is given by F with K being the identity. It shows that additive representations of G are behind technologies considered in traditional production analysis.

Proposition 3. Let $x_f \in X_G \neq \varnothing$. If

$$G(x_v, x_f) = F(x_v + x_f), \tag{11}$$

then, the short-run cost function c_r defined in (3) satisfies either:

(i) $x_v^*(w, x_f, y) = X_v^*(w, y) > 0$ and

$$c_r(w, x_f, y) = C(w, y) > w^\top x_f \tag{12}$$

(ii) $x_{v,j}^* = 0$ for some j, and

$$c_r(w, x_f, y) = \overline{w}^\top \overline{x} + V_r(\widetilde{w}, \overline{x}, y), \tag{13}$$

with $\overline{x}$ defined as in (10) and $V_r(\widetilde{w}, \overline{x}, y) = \min_{\tilde{x}} \{\widetilde{w}^\top \tilde{x} : F(\tilde{x}, \overline{x}) \geq y\}$.

Proposition 3 means that if the technology $G(x_v, x_f)$ takes the specific additive form $F(x_v + x_f)$, then the total inputs $x = x_v + x_f$ will endogenously split into two disjoint types of inputs: $x_v = (\tilde{x}^\top, 0^\top)^\top$ and $x_f = (0^\top, \overline{x}^\top)^\top$ so that $F(x_v + x_f) = F(\tilde{x}, \overline{x})$, which corresponds to the short-run production technology usually considered in microeconomic textbooks, see also VINER's quotation (footnote 1).

Proposition 3(i) states that the production function F may yield a short-run total cost function C which is independent of the level of fixed inputs. In this case, the restriction of the fixed quantity of x_f is not binding. It is the perfect substitution between the variable and the fixed inputs in $F(x_v + x_f)$ which is driving this result. Any restriction in adjusting x_f can be perfectly compensated for by setting x_v optimally.

Proposition 3(ii) shows that the traditional production function F can also generate a short-run fixed cost function depending on some fixed inputs $\overline{x}$ (a subvector of x_f). This is due to the fact that the optimal choice of x_v may be zero for some components, i.e., $x_{v,j}^* = 0$. Thus, the corresponding inputs are fixed, i.e., $x_j = x_{fj}$. This splits $x = x_v + x_f$ into two disjoint subvectors $\tilde{x}$ and $\overline{x}$ with imperfect substitution between them. The properties of this short-run cost functions have been investigated by LAU, L. J. [1976] and BROWNING, M. J. [1983]. For empirical implementations see, e.g., CAVES *et al.* [1981] and MORRISON, C. J. [1988].

FIGURE 3 illustrates Proposition 3. Now the isoquants are convex in (x_1, x_2) (compare point A of FIGURE 3 with point A of FIGURE 1). Endowed with a fixed input vector x_f^0, the variable inputs available to the firm and satisfying $x_v \geq 0$ are located in the north-east quad of x_f^0. At given input prices w, the firm minimizes its variable cost of producing y^1 at the interior point C. At this point, according to Proposition 3(i), the cost function is given by $C(w, y)$ and does not depend on x_f. With another level of fixed inputs, however, the available set of variable inputs will be different. With x_f^1,

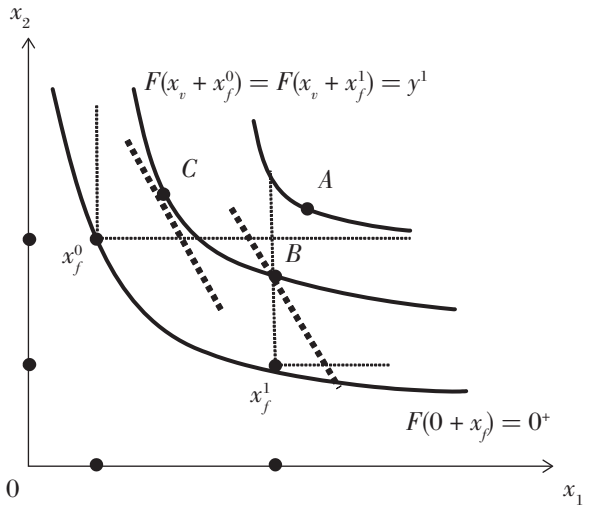

Figure 3. – Isoquants for $F(x)$

the minimal variable cost for producing y^1 is achieved at B, on the boundary of the set $\{x_v \geq 0 : F(x_v + x_f^1) \geq y^1\}$. At point B we have $x_{v1}^* = 0$ and the optimal level of x_{2v}^* is restricted by the level of x_f^1. This corresponds to case (ii) of Proposition 3.

Briefly, the traditional production function with J inputs yields a cost function with J input prices; however, the converse is not true: a cost function with J input prices can be generated by a technology with $2J$ inputs which are all technologically different. The short-run cost function (3) associated with technology G is more general than (12) or (13) associated with F.

IV. Econometric Treatment of Fixed Costs

In the previous sections we have examined the microeconomic framework that generates fixed costs and analyzed the short-run and long-run implications of fixed inputs. Here we examine the empirical content of these theoretical concepts. In the following subsections, we first discuss three drawbacks arising when fixed inputs or fixed costs are neglected. Then, we present two key features of our empirical model: (i) the treatment of unobserved cost heterogeneity; and (ii) the specification of functional forms for the fixed and variable cost functions.

IV.1. *Some Consequences of Neglecting Fixed Costs*

A first problem of disregarding the fixed inputs x_f is the oversimplification of various economic relationships, in particular the relationship between fixed inputs and pricing behavior. Let $p = P(y, z)$ denote the inverse output demand which depends on

exogenous parameters z and the firm's own production level. With market power, the firm's optimal pricing rule is characterized by:

$$\frac{\partial v_r}{\partial y}(w, x_f, y) = p\left(1 + \frac{\partial P}{\partial y}\frac{y}{P}\right). \tag{14}$$

This equation and the discussion above shows that a fixed input x_f has an impact on the marginal cost function unless v_r is independent of x_f. It also implies that there is a relationship between the fixed input and the markup $\eta \equiv \partial \ln P / \partial \ln y$ via the marginal cost.

The second problem of neglecting the fixed cost is that, even in the long-run, it will introduce a source of bias for analyzing the input demand system. By using Shephard's lemma, we obtain:

$$x^*(w, y) = \frac{\partial u}{\partial w}(w) + \frac{\partial v}{\partial w}(w, y).$$

From an empirical point of view, if the fixed cost u is neglected, then it enters the residual term which will correlate with w and bias the estimates.

Neglecting the fixed cost also leads to an underestimation of returns to scale. In order to show this point, we consider the long-run case and assume that the cost function is convex in y. By convexity we have,

$$\begin{aligned} u(w) = \lim_{y \to 0^+} c(w, y) &\geq c(w, y) + \frac{\partial c}{\partial y}(w, y)(0^+ - y) \\ \Rightarrow \frac{\partial c}{\partial y}(w, y)\frac{y}{c(w, y)} &\geq 1 - \frac{u(w)}{c(w, y)}. \end{aligned}$$

As the return to scale is the inverse of the cost elasticity with respect to the output, imposing zero fixed cost implies imposing decreasing returns to scale. The equation above also shows that, for a given level of costs and outputs, neglecting the fixed cost leads to an overestimation of the marginal cost. This will then cause an underestimation of the markup. In the same spirit, Crépon *et al.* [2005] show that the omission of labor market frictions in the empirical model also leads to an underestimation of the markup. In addition, as $\partial c / \partial y(w, y) = w^\top \partial x^* / \partial y(w, y)$, the overestimation of the marginal cost often coincides with the overestimation of the input demand sensitivity to output variations.[9] In Section V, these problems will be illustrated with an empirical example.

9. In general, the approaches that set the fixed cost equal to zero suffer from the omitted-variable bias in the estimation of technology parameters.

IV.2. *On Unobserved Cost Heterogeneity*

We intend to estimate our model using panel data, where individual observations are repeatedly recorded for several time periods. In the following discussion, we use subscript $n = 1,...,N$ to denote individual firms or sectors, and subscript $t = 1,...,T$ to denote time periods. A major advantage of using panel data is the possibility of modeling unobserved individual heterogeneity. In this subsection, we present the source of unobserved heterogeneities in our model and how to deal with them in the estimation.

In our most general model, the level of fixed inputs is not necessarily optimal and has an impact on both the fixed and variable cost:

$$c_r(w_{nt}, x_f, y_{nt}) = u_r(w_{nt}, x_f) + v_r(w_{nt}, x_f, y_{nt}). \tag{15}$$

There are two types of unobserved heterogeneity here: one due to the unobserved fixed inputs, x_f, and one due to the heterogeneous functional forms, u_r and v_r. It is important to highlight that it is not known *ex-ante* (i) which part of each input makes it possible to initiate production and (ii) which part increases the level of production. This is why we need a model for unobserved fixed input levels. More generally, in practice several technologies and ways to start producing a commodity exist (as do several ways of splitting x into two vectors x_f and x_v) and we do not have details concerning these technical possibilities.

Our objective is not to estimate individual-specific functions, but rather their conditional mean, given the values of the observed explanatory variables w_{nt} and y_{nt}, so we consider:

$$C(w_{nt}, y_{nt}) \equiv \mathrm{E}\left[c_r(w_{nt}, x_f, y_{nt}) \mid w_{nt}, y_{nt}\right].$$

Here the integration is over unobserved heterogeneity with respect to the joint distribution of x_f and functional heterogeneity in c_r.

From the theory developed in Section II.2 it turns out that it is not possible, in general, to identify

$$\mathrm{U}(w_{nt}) \equiv \mathrm{E}\left[u_r(w_{nt}, x_f) \mid w_{nt}\right], \tag{16}$$

$$\mathrm{V}(w_{nt}, y_{nt}) \equiv \mathrm{E}\left[v_r(w_{nt}, x_f, y_{nt}) \mid w_{nt}, y_{nt}\right]. \tag{17}$$

unless some restrictions on the technologies are taken for granted.

Assumption 2 (A2).

(i) The distribution of individual heterogeneity in the fixed and variable cost functions is independent of y;

(ii) The distribution of individual heterogeneity in the fixed and variable cost functions is independent of x_f;

(iii) The joint distribution of x_f and individual heterogeneity in the fixed and variable cost functions is independent of y.

Corollary 1. The empirical fixed and variable cost functions (16) and (17) are identified if either

(i) A2i is satisfied and the short-run cost function is (8) or the production function is (9) or

(ii) A2iii is satisfied.

Given that the restrictions required by Corollary 1 to identify **U** and **V** are unlikely to be satisfied, we normalize u_r and v_r using (7) and consider

$$U(w_{nt}) \equiv \mathrm{E}\Big[w_{nt}^{\top} x_f^{0+} + v_r(w_{nt}, x_f^{0+}, 0^+) \mid w_{nt}\Big], \tag{18}$$

$$V(w_{nt}, y_{nt}) \equiv \mathrm{E}\Big[v_r(w_{nt}, x_f, y_{nt}) - v_r(w_{nt}, x_f^{0+}, 0^+) \mid w_{nt}, y_{nt}\Big]. \tag{19}$$

The fixed inputs x_f^{0+} either satisfy $x_f^{0+} \in X_G$ (if they are exogenously fixed) or $x_f^{0+} = x_f^*(w, 0^+)$ (if they are optimally allocated) and the fixed inputs x_f either satisfy $x_f = x_f^{0+}$ (in the fixed input case) or $x_f = x_f^*(w, y)$ (in the case of optimally allocated fixed inputs). The expectation is taken over three types of unobserved heterogeneity: in the functional form of (u_r, v_r), in the fixed inputs (x_f, x_f^{0+}) and in the fact that we do not know whether they are optimally allocated or not. It follows from their definitions that functions U and V depend on observed variables and their functional forms are identical for all firms or sectors.

Using these definitions, we rewrite the firm-specific cost function (15) as follows:

$$c_r(w_{nt}, x_f, y_{nt}) = \gamma^U(w_{nt}, x_f)\, U(w_{nt}) + \gamma^V(w_{nt}, x_f, y_{nt})\, V(w_{nt}, y_{nt}), \tag{20}$$

where the functions γ^U and γ^V are defined by:

$$\gamma^U(w_{nt}, x_f) \equiv \frac{w_{nt}^{\top} x_f^{0+} + v_r(w_{nt}, x_f^{0+}, 0^+)}{U(w_{nt})},$$

$$\gamma^V(w_{nt}, x_f, y_{nt}) \equiv \frac{v_r(w_{nt}, x_f, y_{nt}) - v_r(w_{nt}, x_f^{0+}, 0^+)}{V(w_{nt}, y_{nt})}.$$

These random terms satisfy

$$\mathrm{E}\Big[\gamma^U(w_{nt}, x_f) \mid w_{nt}\Big] = \mathrm{E}\Big[\gamma^V(w_{nt}, x_f, y_{nt}) \mid w_{nt}, y_{nt}\Big] = 1.$$

Given that x_f is unobserved in most data sets, we cannot directly use (20) for the empirical investigation. However, progress can be made if we introduce the following decomposition:

$$\gamma^U(w_{nt}, x_f) = \gamma_n^U + \varepsilon_{nt}^U; \qquad \gamma^V(w_{nt}, x_f, y_{nt}) = \gamma_n^V + \varepsilon_{nt}^V. \tag{21}$$

We assume that $\mathrm{E}[\varepsilon_{nt}^{U} \mid w_{nt}] = \mathrm{E}[\varepsilon_{nt}^{V} \mid w_{nt}, y_{nt}] = 0$, which implies that $\mathrm{E}[\gamma_{n}^{U} \mid w_{nt}] = \mathrm{E}[\gamma_{n}^{V} \mid w_{nt}, y_{nt}] = 1$ and allows some correlation between unobserved heterogeneity γ_{n}^{U}, γ_{n}^{V} and the explanatory variables w_{nt}, y_{nt}. The empirical model with the observed total cost level c_{nt} is given by:

$$c_{nt} = \gamma_{n}^{U} U(w_{nt}) + \gamma_{n}^{V} V(w_{nt}, y_{nt}) + e_{nt}. \tag{22}$$

where the composite residual term

$$e_{nt} = \varepsilon_{nt}^{U} U(w_{nt}) + \varepsilon_{nt}^{V} V(w_{nt}, y_{nt}) + \varepsilon_{nt}. \tag{23}$$

The additive error term ε_{nt} is i.i.d. and satisfies $\mathrm{E}[\varepsilon_{nt} \mid w_{nt}, y_{nt}] = 0$. As a consequence $\mathrm{E}[e_{nt} \mid w_{nt}, y_{nt}] = 0$.

Many empirical applications to industrial organization restrict heterogeneity to affect only the fixed cost but not the variable cost (see the survey of Berry, S., and P. Reiss [2007]). Bresnahan, T. F., and P. C. Reiss [1991] are a noteworthy exception (see especially their equation (4)), but they do not investigate the relationship between the fixed and variable cost. This road is followed here. In addition to the fact that both fixed and variable cost heterogeneities are considered in (22), an interesting feature of our empirical model is that the unobserved relationship between fixed and variable cost functions can be measured by the covariance between γ^{U} and γ^{V}. Note that the covariance between γ^{U} and γ^{V} can *a priori* take any value. However, we derive an important statistical relationship between the fixed and variable cost functions $\gamma^{U}U$ and $\gamma^{V}V$.

Proposition 4. Under Assumption A2ii, if the fixed inputs x_{f} are positive and are optimally allocated, then:

(i) the conditional covariance $\mathrm{cov}(\gamma^{U}, \gamma^{V} \mid w, y)$ is nonpositive;

(ii) the conditional variance matrix $\mathrm{V}[\gamma \mid w, y]$ is singular.

When the fixed inputs are unobserved we will not be able to estimate functions u_{r} and v_{r}, and we cannot test whether $\partial c_{r} / \partial x_{f} = 0$ is satisfied or not. However, we will be able to estimate $\mathrm{V}[\gamma \mid w, y]$ and $\mathrm{cov}[\gamma^{U}, \gamma^{V} \mid w, y]$. If the statistical test leads to rejection of the singularity of $\mathrm{V}[\gamma \mid w, y]$ or $\mathrm{cov}[\gamma^{U}, \gamma^{V} \mid w, y] \leq 0$, then we can deduce that either the fixed inputs are not optimally allocated (Proposition 4), or that the production technology has the specific structure given in (9). The level of the fixed cost $\gamma^{U}U$ and the level of the variable cost $\gamma^{V}V$ are likely to be positively correlated: both the fixed and the variable costs increase over time, and firms with a high fixed cost certainly produce more than smaller firms and also have a higher variable cost. Proposition 4, however, states that there is a trade-off – a negative correlation – between the fixed and the variable cost *for given values* of the explanatory variables (w, y). Such a trade-off cannot be directly observed in a data set, because it pertains to unobserved heterogeneity. With panel data, the issue of interrelated heterogeneity is often discarded,

although one exception is GLADDEN, T., and C. TABER [2009], who considered it in estimating linear wage equations.

The trade-off between fixed and variable costs is represented in FIGURE 4. For given input prices, higher fixed inputs yield a higher fixed cost u_h and make it possible to decrease the variable cost, so that beyond production level $\underline{y}$ it becomes costless to rely on technology h instead of n. The long-term cost function is not the lower envelope of these curves, because technologies and technological choices are heterogenous, and because fixed and variable costs are complementary. It is therefore not possible to produce $y > \underline{y}$ with a variable cost given by v_h and a fixed cost given by u_n. In this context, a positive correlation emerges between the level of fixed cost and the level of production, even though the fixed cost is *functionally* independent of the production level y.

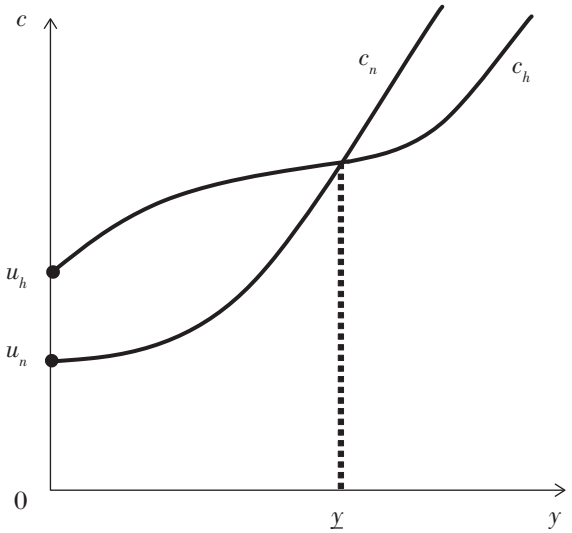

FIGURE 4. – Different Cost Functions for Firm h and n

IV.3. *On Flexible Functional Forms*

The empirical model (22) at this stage is not yet operational, because it depends on general functions. In order to fit the model to the data, we need to choose specific functional forms for U and V. In the 1970s and 1980s, several researchers proposed new parametric specifications for production technology, and introduced so-called flexible functional forms, which are able to locally approximate an arbitrary cost function. These functional forms, still widely used in production analysis, are, however, not adequate for modelling fixed costs: either they completely exclude fixed costs, or they specify them in an inflexible way. This arises because the traditional approaches do not explicitly specify the fixed cost function separately from the variable cost function.

In their seminal paper, DIEWERT, W. E., and T. J. WALES [1987] introduced several cost functions, many of which can be written as:

$$C^{DW}(w,y,t) = a_w^{\top} w + (\alpha_w^{\top} w)\, a_t t + V^{DW}(w,y,t). \tag{24}$$

The variable t is now introduced for denoting technical change. The subscripts n and t are dropped here to simplify notations. The variable cost function satisfies $V^{DW}(w,0,t) = 0$. This identifies the fixed cost as $U^{DW}(w,t) = a_w^{\top} w + (\alpha_w^{\top} w)\, a_t t$, where a_w, α_w, a_t denote technology parameters. The fixed cost function is therefore linear in w and t and is not a flexible specification in the sense of DIEWERT, W. E., and T. J. WALES [1987]. The same can be shown for the variable cost specification V^{DW}.

Let us now consider the Translog functional form (CHRISTENSEN *et al.* [1971]) with technology parameters denoted by vector β:

$$\begin{aligned} C^{TL}(w,y,t) = \exp(&\beta_0 + \beta_w^{\top} \ln w + \beta_y \ln y + \beta_t t \\ &+ \frac{1}{2} \ln w^{\top} B_{ww} \ln w + \ln w^{\top} B_{wy} \ln y + \ln w^{\top} B_{wt} t \\ &+ \frac{1}{2} \beta_{yy} (\ln y)^2 + \beta_{yt} t \ln y + \frac{1}{2} \beta_{tt} t^2), \end{aligned} \tag{25}$$

where β_w, B_{wy} and B_{wt} are $(J \times 1)$ vectors and B_{ww} is a $(J \times J)$ matrix. Let ι denote a $(J \times 1)$ vector of ones. The Translog function has $3 + 2J + (J+1)\,J\,/\,2$ free parameters which satisfy

$$\iota^{\top} \beta_w = 1, \quad \iota^{\top} B_{wt} = \iota^{\top} B_{wy} = 0, \quad \iota^{\top} B_{ww} = 0, \quad B_{ww} = B_{ww}^{\top}, \tag{26}$$

in order to ensure that C^{TL} is linearly homogenous in w and symmetric. This functional form is still widely used in applied econometrics (see KOEBEL *et al.* [2003], for instance). One of the main drawbacks of the Translog functional form is that it is not suitable for modelling fixed costs.

Proposition 5. The Translog functional form implies a fixed cost that is either zero or infinite (in which case C^{TL} is decreasing in y for some values of y).

This result shows that the Translog cost function is badly behaved in some regions, especially when production is close to zero. This result is clearly related to the "flip-flop" property highlighted by RÖLLER, L.-H. [1990] in the multi-output context. Proposition 5 points out a paradox: although the Translog specification is flexible (DIEWERT, W. E., and T. J. WALES [1987], Theorem 1), it excludes fixed costs (in the best case). The reason for this apparent contradiction is to be found in the limitations of the flexibility requirement, which only requires that the Translog cost function be a local approximation, in some neighborhood of y, but not necessarily in the neighborhood of $y = 0^{+}$, which defines the fixed cost. In contrast, we introduce a functional form that is flexible at two points.

Definition 3. A *two-point Flexible Functional Form (2FFF)* for a cost function provides a second order approximation to an arbitrary twice continuously differentiable cost function C at point where $y > 0^+$ and at $y = 0^+$.

We have seen that a production technology with fixed costs, can be represented by *two* different production technologies: one for initiating production (using only fixed inputs), and one for reaching the output level y. It therefore becomes quite natural to specify both technologies in a flexible way. Similarly, the cost function is additively separable in two parts: one part corresponding to the cost at the near zero output level and one part reflecting the production cost of the output. If our objective is to provide an approximation of the production technology, both parts should be treated with equal importance, and we suggest here that a flexible functional form be used for both the fixed and variable cost functions. Definition 3 implies that a 2FFF cost function is the sum of two 1FFF fixed and variable cost functions. The 2FFF for a cost function of J inputs has in total $1 + 3(J+1) + J(J+1)$ free parameters, with $(J+1) + J(J+J)/2$ parameters for the fixed cost function and $(J+2) + (J+1)(J+2)/2$ parameters for the variable cost function. We provide a proof of this claim in the Appendix. A similar concept of 2FFF has been used in a different context by Diewert, W. E., and D. Lawrence [2002], in order to avoid trending elasticities.

For our empirical model, we assume that the fixed and variable cost functions U and V in (22) take flexible functional forms, respectively denoted by U^Q and V^{TL}. As seen in Proposition 5, the traditional Translog cost function C^{TL} satisfies $\lim_{y\to 0^+} C^{TL}(w,y,t) = 0$ (in the case where $\beta_{yy} \leq 0$) and is a suitable variable cost function. We therefore specify $V^{TL}(w,y,t)$ by using the parametric function, $C^{TL}(w,y,t)$ defined in (25). As the Translog function is not compatible with the occurrence of a fixed cost, we use a normalized quadratic form for the fixed cost function U^Q:

$$U^Q(w,t;\alpha) = \alpha_w^\top w + \frac{1}{2}\frac{w^\top A_{ww} w}{\theta^\top w} + \alpha_{wt}^\top w t + \frac{1}{2}\alpha_{tt}(\theta^\top w)t^2. \tag{27}$$

We achieve parameter identification and impose symmetry in w using the following $J + (J-1)J/2$ parametric restrictions on U^Q:

$$\iota^\top A_{ww} = 0, \quad A_{ww} = A_{ww}^\top. \tag{28}$$

We follow Diewert, W. E., and T. J. Wales [1987] and define θ as a vector of constant weights chosen so that $\theta^\top w$ corresponds to the Laspeyres index for total cost, normalized to one in the first period. Finally, there are $1 + J + (J+1)J/2$ free parameters left in U^Q.

V. An Empirical Investigation

We use the NBER-CES manufacturing industry database for our empirical study.[10] This database records annual information on output y_{nt}, output price p_{nt}, and the input levels x_{nt}, together with input price indices w_{nt}, for 462 US manufacturing industries (at the six-digit NAICS aggregation level) and covers the period 1958 to 2005. This information is available for four inputs: capital, labor, energy and intermediate inputs. Bartelsman, E., and W. Gray [1996] present the construction of the database and descriptive statistics. Chen, X. [2016] reports details on the computations made for generating the depreciation rate, the interest rate, the user cost of capital, and total labor expenditure, including fringe benefits.

V.1. *Estimation Models*

Our benchmark model is given by (22) with a composite residual term defined in (23). The fixed and variable functions take the specific parametric forms (27) and (25), respectively. Our parameters of interest are α, β, γ_n^U and γ_n^V for $n = 1,...,N$. In this empirical application, there are four types of inputs $(J = 4)$. Given the parametric restrictions (26) and (28), the fixed cost function has 15 free technology parameters in α, and the variable cost function has 21 free technology parameters in β.

Since we have $\mathrm{E}[\gamma_n^U \mid w,t] = \mathrm{E}[\gamma_n^V \mid w,y,t] = 1$, it is natural to impose the normalization conditions:

$$\frac{1}{N}\sum_{n=1}^{N}\gamma_n^U = \frac{1}{N}\sum_{n=1}^{N}\gamma_n^V = 1. \tag{29}$$

These two restrictions allow us to identify all the $2(N-1)$ industry specific (γ_n^U, γ_n^V) parameters, which represent industry-specific deviations in percentage from the average fixed and variable cost functions. For instance, if the estimated value of γ_n^U is significantly above one and the estimated value of γ_n^V is significantly below one, this indicates that sector n incurs more fixed and less variable costs than average.

We estimate the model using the Nonlinear Least-Squares (NLS) estimator. The NLS estimation of this model is consistent because the conditional mean of the residual term equals zero, i.e., $\mathrm{E}[e_{nt} \mid w_{nt}, y_{nt}, t] = 0$.[11] In this empirical application, we compare our benchmark model with two alternatives: Model I corresponds to the classical Translog cost function, which includes only the variable cost function V^{TL} and assumes that $U^Q \equiv 0$ and $\gamma_n^V = 1$ for all $n = 1,...,N$; while Model II extends Model I by including the fixed cost U^Q, but sticks to the assumption that $\gamma_n^U = \gamma_n^V = 1$. This restriction is removed in Model III (our benchmark model), which adds industry specific parameters

10. An updated version of the dataset can be downloaded at: http://www.nber.org/data/nberces.html

11. $\mathrm{E}[e_{nt} \mid w_{nt}, y_{nt}, t] = U(w_{nt},t)\mathrm{E}[\varepsilon_{nt}^U \mid w_{nt}, y_{nt}, t] + V(w_{nt}, y_{nt}, t)\mathrm{E}[\varepsilon_{nt}^V \mid w_{nt}, y_{nt}, t] + \mathrm{E}[\varepsilon_{nt} \mid w_{nt}, y_{nt}, t] = 0$

γ_n^U and γ_n^V. All models are estimated in levels (the dependent variable is c_{nt} as in (22)) and in first difference (the dependent variable is $c_{nt} - c_{n,t-1}$), with the purpose of avoiding problems due to non-stationary data. In the case of Model III, a large number of parameters are estimated. Thus, in order to gain efficiency, we use a system approach that estimates the cost function together with input demand equations obtained using Shephard's lemma. This approach allows us to include more observation points in the estimation.

V.2. *Estimation Results*

Instead of reporting estimates for all parameters of the cost function, we only select some informative estimated coefficients and statistics. An important coefficient is the parameter β_{yy}, which is crucial for Proposition 5. A further important matrix is A_{ww}, which describes how fixed inputs can be substituted for each other, and vanishes when the fixed cost is linear in w. This restriction is often imposed in empirical contributions, for example, (24) considered by DIEWERT, W. E., and T. J. WALES [1987]. We also report post-estimation statistics such as the share of the fixed cost in the total cost $\gamma_n^U U / C$; the ratio of the output price to the predicted marginal cost of production $p / (\partial C / \partial y)$, which measures the markup; the elasticity of costs with respect to output $\varepsilon(C, y) \equiv \partial \ln C / \partial \ln y$ (corresponding to the inverse of the rate of returns to scale); and the semi-elasticity $\partial \ln C / \partial t$, representing technological change. TABLE I contains the estimation results for the three models.

Model I in both level and first difference yield quite similar estimation results. The t-values of the model in first difference are somewhat lower than when the estimation is in levels. With both specifications returns to scale are found to be significantly decreasing and there is evidence for technological regression as the cost increases over time (for given input prices and output level). There is one important difference, however: the median value of the markup is 1.35 when estimation is in level and 2.14 in first difference. This is due to the fact that the estimation in first difference allows us only to identify the cost function up to a constant term (over n and t), which could typically represent an omitted fixed cost. Hence, first difference allows implicitly for the occurrence of some fixed cost, which explains the difference in the markup between the two variants of Model I.

Regarding the results of Model II, we find that the fixed cost function is significantly different from zero and nonlinear in input prices (indicated by the Wald test in the fourth line of TABLE I). Model II still yields a positive estimate of $\partial \ln C / \partial t$: technological regress is statistically significant for the estimation in first difference, but it is not as deep as with the Translog specification of Model I. Unlike Model I, returns to scale are now found to be increasing. This is in line with our discussion in SECTION IV.1, however, the rate of returns to scale seems too high to be plausible. The estimated level of the fixed cost is also quite unrealistic, and certainly does not represent more than 50% of total cost in most cases. Briefly, the theoretical advantages of Model II seem not to be supported by the empirical results, which are as surprising as those of the traditional

Table I. – Summary of Estimation Results

	Model	I	I	II	II	III	III	
	Specification	Level	FD	Level	FD	Level	FD	%
β_{yy}		–0.07 (–22.1)	–0.01 (–2.5)	–0.09 (–21.4)	–0.01 (–1.4)	–0.12 (–161.0)	–0.08 (–61.6)	100
$H_0 : A_{ww} = 0$	p-value	–	–	0.00	0.00	0.00	0.00	
$\gamma_n^U U^Q / C$	1^{st} quartile	–	–	0.09 (7.1)	0.42 (15.5)	0.00 (0.0)	0.12 (0.7)	
	median	–	–	0.49 (30.6)	0.68 (26.1)	0.09 (2.1)	0.19 (1.6)	48.5
	3^{rd} quartile	–	–	0.94 (198.6)	0.89 (66.1)	0.15 (5.6)	0.31 (3.2)	
$p / (\partial C / \partial y)$	1^{st} quartile	1.18 (13.0)	1.83 (14.2)	1.26 (12.5)	1.90 (12.2)	1.15 (1.3)	1.40 (1.2)	
	median	1.35 (22.0)	2.14 (18.9)	1.70 (20.5)	2.31 (16.5)	1.33 (3.7)	1.74 (1.9)	59.3
	3^{rd} quartile	2.60 (32.0)	4.07 (26.7)	6.44 (29.9)	5.60 (22.3)	1.55 (8.7)	2.27 (3.3)	
$\varepsilon(C,y)$	1^{st} quartile	1.06 (8.8)	1.08 (4.9)	0.11 (108.2)	0.15 (47.8)	0.88 (–4.9)	0.80 (–1.8)	
	median	1.19 (22.5)	1.21 (13.5)	0.71 (14.3)	0.40 (23.0)	0.99 (–0.2)	0.92 (–0.5)	26.5
	3^{rd} quartile	1.55 (46.0)	1.53 (30.9)	1.01 (0.6)	0.65 (13.6)	1.12 (1.6)	1.02 (0.1)	
$\partial \ln C / \partial t$	1^{st} quartile	0.015 (10.4)	0.021 (9.8)	–0.016 (–6.6)	0.002 (0.5)	–0.006 (–4.9)	–0.005 (–1.1)	
	median	0.025 (16.2)	0.033 (14.7)	0.002 (0.9)	0.010 (2.4)	–0.002 (–2.0)	–0.001 (–0.2)	18.4
	3^{rd} quartile	0.034 (20.7)	0.045 (20.3)	0.016 (7.4)	0.020 (4.1)	0.000 (0.3)	0.003 (0.5)	

Notes: Row 3 reports the estimated parameter values and the corresponding t-value for the hypothesis that the parameter is equal to zero (in parentheses). Row 4 gives the p-value for the hypothesis test $H_0 : A_{ww} = 0$. Rows 5 to 16 report the median value of the corresponding statistic over all observations as well as the 1st and 3rd quartiles. In parenthesis we report the t-value for the null hypothesis that respectively $\gamma_n^U U^Q / C = 0$, $p / (\partial C / \partial y) = 1$, $\varepsilon(C,y) = 1$ and $\partial \ln C / \partial t = 0$. The last column gives the percentage of observations for which the t-test statistic cannot reject the corresponding null hypothesis (only reported for Model 3 in FD).

Translog (Model I). Unobserved heterogeneity in fixed costs could be one explanation for these results. When estimated in first difference, Model II becomes compatible with some *additive* unobserved heterogeneity. However, as the cost function has to be homogeneous of degree one in input prices, heterogeneity cannot be additive, but must interact with input prices. We therefore turn to Model III, which allows for (multiplicative) individual heterogeneity.

The empirical results for Model III show that once unobserved heterogeneity is controlled for, the results become plausible, and stand often in contradiction with those of Models I and II. The only regularity between Models I, II and III is that the estimated

coefficient of β_{yy} is negative (and significant in all cases except one), which implies that the limit of the classical Translog variable cost function is zero as y goes to 0. Our preferred specification (III-FD) emphasizes that fixed costs are quite heterogenous over industries and significant for 48.5% of the observations. The median value of the fixed cost share is 19% (but not significant at the 5% threshold). For 25% of the observations, the fixed cost represents more than 31% of the total cost. The returns to scale are increasing ($\varepsilon(C,y) \leq 1$) for 26.5% of the observations, and constant for the bulk of the observations. Regarding the markup, the estimates indicate that for about 40% of the observations it is close to one, whereas 60% of the firms have significant market power. The median value of $p / \partial C / \partial y$ is 1.74 and is lower than those obtained in Model I and II. Regarding the rate of technical change, in Model III the cost tends to decrease by 0.1% over time. Results are similar for the estimations in level and in first difference, but less often significant when estimation is in first difference: $\partial \ln C / \partial t$ is found to be significantly negative for only 18.4% of the observations. Altogether, our results are in line with those obtained by Diewert, W. E., and K. J. Fox [2008] and Bartelsman, E., and W. Gray [1996] who also found modest empirical evidence for technical change in US manufacturing. Our interpretation is that the deterministic trend only partially captures technical progress, and that one important part of technical change is stochastic and embodied in the unobserved fixed inputs x_f. These fixed inputs contribute to increasing the fixed cost and decrease the variable cost, and, as a consequence of our approach, this correlated random component of technical change is captured by the negative correlation between γ_n^U and γ_n^V.

We examine the empirical correlation between γ_n^U, γ_n^V and shares of fixed cost, returns to scale, markups and rate of technological change. Figure 5 depicts the estimated coefficients of γ_n^U and γ_n^V. The homogeneity hypothesis, $\gamma_n^U = \gamma^U$ and $\gamma_n^V = \gamma^V$, is statistically rejected by a likelihood-ratio test. The correlation between $\hat{\gamma}_n^U$ and $\hat{\gamma}_n^V$ is found to be negative, and significantly different from zero, a result that is in line with Proposition 5; industries with lower than average fixed costs generally have higher than average variable costs and vice versa. Only 162 out of 462 parameters γ_n^U are found to be significantly different from zero, confirming that several industries operate with negligible fixed costs. The correlation between γ_n^U and γ_n^U is equal to -0.13 and significantly different from zero at the 1% threshold (for Model III in level, the correlation is -0.41). According to this result, the separable structure (8) of Proposition 2, that implies no interaction between fixed and variable cost, is statistically rejected: technology $G(x_v, x_f)$ fits the data better than $F(x_v + K(x_f))$. The fixed cost function is found to be nonlinear in w, as the substitution matrix A_{ww} is statistically different from zero (see row 4 of Table I).

Table II reports the empirical correlations between different estimated statistics obtained from Model III-FD. We include concentration data available for a few years, and restrict this analysis for the year 2002 (462 observations are available). Many correlation coefficients are smaller than 0.20. However, we find noteworthy exceptions.

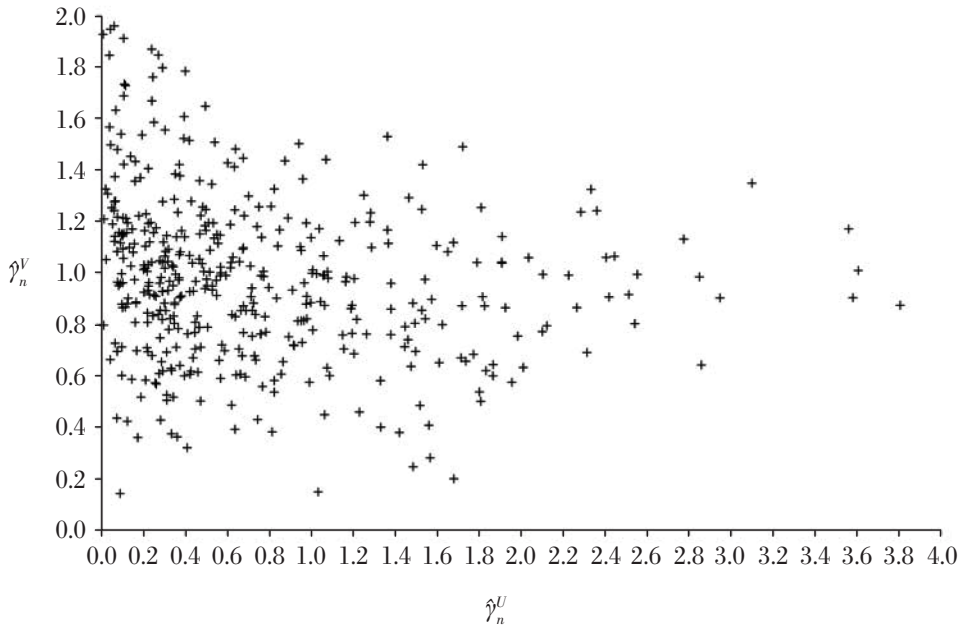

Figure 5. – Scatterplot of $\hat{\gamma}_n^U$ and $\hat{\gamma}_n^V$

The fixed-cost heterogeneity $\hat{\gamma}_n^U$ (column 2) is positively correlated with the output level, which typically yields biased estimates in models (like II) neglecting this interdependence. Firms producing massively have an incitation to increase their fixed cost infrastructure in order to decrease their variable costs and produce more efficiently. This result helps to understand why specifications neglecting the fixed cost (or including an inflexible parameterization of the fixed cost) are likely to overestimate the marginal cost of production and underestimate the markup and the rate of returns to scale. The omission of the fixed cost leads to attributing neglected variations in fixed costs (which, according to Table II, are positively correlated with output) to the variable cost function, which is increasing in y. As in the case of an omitted variable bias, the variable cost function (and especially its partial derivative w.r.t. y) will catch up the part of the fixed cost function which is correlated with production, and as a result, the marginal cost of production will be biased upwards. The positive correlation $corr(\gamma_n^U; y_n) = 0.83$ helps to understand the gap between the results obtained with the standard and extended Translog specifications (see Table I). In Model I, the neglected fixed cost causes the rate of returns to scale to be lower and the markups to be moderate in comparison to Models II and III.

The variable cost heterogeneity (column 3) is negatively correlated with the markup and both concentration statistics ($cr_{20,n}, H_n$). The share of fixed cost is highly correlated with the returns to scale: the higher the fixed cost-share, the lower $\varepsilon(C, y)$ and the higher the returns to scale. This coincides with our discussion of Section IV.1 on

Table II. – Correlation Matrix, Model III, FD

	γ_n^U	γ_n^V	$\gamma_n^U \frac{U}{C}$	$\frac{p}{\partial C / \partial y}$	$\varepsilon(C,y)$	$\frac{\partial \ln C}{\partial t}$	y_{nt}	N_n	cr_n
γ_n^V	–0.13								
$\gamma_n^U \frac{U}{C}$	–0.02	–0.04							
$\frac{p}{\partial C / \partial y}$	0.04	–0.55	0.19						
$\varepsilon(C,y)$	–0.05	0.04	–0.99	–0.15					
$\frac{\partial \ln C}{\partial t}$	0.01	–0.03	0.99	0.17	–0.99				
y_n	0.83	–0.07	–0.06	0.03	–0.02	–0.03			
N_n	0.17	0.25	–0.09	–0.11	0.06	–0.07	0.11		
$cr_{20,n}$	0.00	–0.32	0.18	0.23	–0.16	0.16	0.03	–0.56	
H_n	0.02	–0.20	0.08	0.16	–0.07	0.07	0.10	–0.31	0.77

Notes: N_n denotes the number of active firms, $cr_{20,n}$ the concentration ratio for the 20 largest firms and H_n the Herfindahl-Hirschman index of concentration.

the dangers of neglecting fixed cost. The high correlation between the fixed cost share and the rate of technological change is surprising, and means that firms with higher fixed costs, have, on average, a smaller productivity growth. This result is due to the fact that our estimated fixed cost function is not decreasing with t but increasing, i.e. $\partial \ln U / \partial t \geq 0$.

We also investigate the link between the fixed cost, the size and the concentration of industries. Table II reports correlations between the fixed cost, the output level, the number of active firms N_n within industry n, the concentration ratio for the 20 largest firms $cr_{20,n}$, and the Herfindahl-Hirschman index H_n.[12] These results suggest that industries with a higher fixed cost are likely to be more concentrated as $\text{corr}(\gamma_n^U U / C, cr_{20,n}) = 0.18$. We find some negative correlation between variable cost heterogeneity and industrial concentration and positive correlation with the markup.

12. The number of compagnies and the concentration data for 2002 are obtained from the U.S. Census Bureau.

VI. Conclusion

This paper investigates technologies in which fixed inputs can be imperfectly substituted for variable inputs, and we propose extended production and cost functions compatible with the occurrence of a fixed cost. Some available flexible specifications, like the Translog cost function, restrict the fixed cost to be equal to zero. Our extended specification of the Translog is compatible with arbitrary levels of fixed cost, and allows for interactions between fixed and variable costs. The empirical findings highlight the importance of fixed cost: they are significant for about 50% of the industries and represent about 20% of total cost. The estimates also support our extended framework, which explains why industries with higher fixed costs, on average have lower variable costs, higher returns to scale and markups. In line with our theoretical prediction, we find that the classical Translog cost specification yields biased estimates, and that joint unobserved heterogeneity in fixed and variable cost has to be controlled for in order to avoid further estimation biases.

Acknowledgements

We would like to thank Claude d'Aspremont, Pierre Dehez, Rodolphe Dos Santos Ferreira, Andreas Irmen, Gauthier Lanot, François Laisney, Jacques Mairesse, Pierre Mohnen, Clemens Puppe and the participants of seminars in Aix-en-Provence, Cambridge, Cergy, Freiburg, Konstanz, Luxembourg, Maastricht, Nancy, Trier and Umeå Strasbourg for their helpful comments. We are also indebted to two anonymous referees for their remarks which contributed to improve the quality of the paper. We would like to thank David Horner for careful proof reading.

Correspondence:
Xi Chen
ANEC/STATEC, Luxembourg
E-mail: Xi.Chen@statec.etat.lu

Bertrand M. Koebel
Bureau d'Économie Théorique et Appliquée (BETA)
UMR 7522 CNRS, Université de Strasbourg
61 avenue de la Forêt Noire
67085 Strasbourg Cedex (France)
Tel (+33) 368 852 190
E-mail: koebel@unistra.fr

Appendix

Proof of Proposition 1.

(i) If $x_f \in X_G$ then the choice $x_v = 0$ is admissible and so

$$v_r(w, x_f, 0^+) = \min_{x_v \geq 0}\{w^\top x_v : G(x_v, x_f) \geq 0^+\} = 0.$$

(ii) For $y' > y$, and G nondecreasing in x_v, it implies that $\{x_v : G(x_v, x_f) \geq y'\} \subseteq \{x_v : G(x_v, x_f) \geq y\}$ and as a consequence

$$v_r(w, x_f, y') = \min_{x_v \geq 0}\{w^\top x_v : G(x_v, x_f) \geq y'\} \geq v_r(w, x_f, y).$$

(iii) Similarly, for $x'_f > x_f$ and G nondecreasing in (x_v, x_f), implies that $\{x_v : G(x_v, x_f) \geq y\} \subseteq \{x_v : G(x_v, x'_f) \geq y\}$ and as a consequence

$$v_r(w, x'_f, y) = \min_{x_v \geq 0}\{w^\top x_v : G(x_v, x'_f) \geq y\} \leq v_r(w, x_f, y).$$

□

Proof of Proposition 2.

Part (i), (9) $\Rightarrow$ (8). For an exogenous level of $x_f \in X_G$, we have

$$\begin{aligned} v_r(w, x_f, y) &= \min_{x_v \geq 0}\{w^\top x_v : y = F(x_v + K(x_f))\} \\ &= \min_{x_v \geq 0}\{w^\top x_v + w^\top K(x_f) : y = F(x_v + K(x_f))\} - w^\top K(x_f) \\ &= \min_{X \geq K(x_f)}\{w^\top X : y = F(X)\} - w^\top K(x_f) \\ &= v_y(w, y) - w^\top K(x_f). \end{aligned}$$

The last line follows from our assumption that $x_v^*(w, y) > 0$ at the optimum. Defining $v(w, y) \equiv v_y(w, y) - v_y(w, 0^+)$ ensures that $v(w, 0^+) = 0$. Defining $u_r(w, x_f) \equiv v_y(w, 0^+) - w^\top K(x_f) + w^\top x_f$ ensures that $c_r(w, x_f, y) = u_r(w, x_f) + v(w, y)$.

(9) $\Leftarrow$ (8). We can recover the convex hull of all inputs x_v producing y, for a given level of x_f, by solving

$$\min_w\{w^\top x_v - v_y(w, y) + w^\top K(x_f)\}.$$

As G is quasi-concave in x_v, this convex hull corresponds to the isoquants of G. The J first order conditions for an inner solution are given by

$$x_v + K(x_f) - \frac{\partial v_y}{\partial w}(w,y) = 0,$$

which can be solved with respect to w / w_J and y to obtain

$$y = F(x_v + K(x_f)).$$

Part (ii). Necessity. With (8), the first order conditions for an inner solution in x_f to the cost minimization problem are given by

$$\frac{\partial u_r}{\partial x_f}(w,x_f) = w,$$

and do not depend on y, and so do the solutions $x_f^*(w)$. With (9), the first order conditions for an inner solution in x_v are

$$w = \lambda \frac{\partial F}{\partial x_v}(x_v + K(x_f))$$
$$y = F(x_v + K(x_f)),$$

where λ denotes the Lagrange multiplier. Under A1iv, the solution in x_v to this system takes the form $x_v^*(w,x_f,y) = X^*(w,y) - K(x_f)$ and so the short-run cost function (8), with $v_y(w,y) \equiv w^\top X^*(w,y)$ and $u_r(w,x_f) = w^\top x_f - w^\top K(x_f)$. Then x_f^* is independent of y.

Sufficiency. If x_f^* depends only upon w at the optimum for any y, then the first order conditions for an inner solution (which fully characterize x_f^*), given by

$$\frac{\partial u_r}{\partial x_f}(w,x_f) + \frac{\partial v_r}{\partial x_f}(w,x_f,y) = 0$$

have to satisfy

$$\frac{\partial^2 v_r}{\partial x_f \partial y}(w,x_f,y) = 0$$

and so $c_r(w,x_f,y) = u_r(w,x_f) + v(w,y)$. □

Proof of Proposition 3.

(i) The variable inputs satisfy the nonnegativity constraints $x_v \geq 0$. If these constraints are not binding at the optimum, we can write

$$c_r(w, x_f, y) = \min_{x_v > 0} \{w^\top x_v + w^\top x_f : F(x_v + x_f) \geq y\}$$
$$= \min_{X > x_f} \{w^\top X : F(X) \geq y\} = C(w, y),$$

and by Shephard's lemma $x_v^*(w, x_f, y) = X_v^*(w, y)$.

(ii) If some constraints $x_{v,j} \geq 0$ are binding at the optimum, the total input x can be rewritten as

$$x = x_v + x_f = \begin{pmatrix} \tilde{x} \\ \bar{x} \end{pmatrix},$$

with $\tilde{x}_i = x_{v,i} + x_{f,i}$ for $x_{v,i} > 0$ and $\bar{x}_j = x_{f,j}$ for $x_{v,j} = 0$. Vector w is partitioned accordingly as $w = (\tilde{w}^\top, \bar{w}^\top)^\top$. Then

$$c_r(w, x_f, y) = \min_{x_v \geq 0} \{w^\top x_v + w^\top x_f : F(x_v + x_f) \geq y\}$$
$$= \min_{\tilde{x} > 0} \{\tilde{w}^\top \tilde{x} + \bar{w}^\top \bar{x} : F(\tilde{x}, \bar{x}) \geq y\}$$
$$= \min_{\tilde{x} > 0} \{\tilde{w}^\top \tilde{x} : F(\tilde{x}, \bar{x}) \geq y\} + \bar{w}^\top \bar{x} = V_r(\tilde{w}, \bar{x}, y) + \bar{w}^\top \bar{x}.$$

□

Proof of Corollary 1.

(i) This is a direct consequence of Proposition 2 and the definitions of (16) and (17). If the optimal level of the fixed input is $x_f^*(w)$ and does not depend upon y, then

$$\mathrm{E}[u_r(w_{nt}, x_f) \mid w_{nt}, y_{nt}] = \mathrm{E}[u_r(w_{nt}, x_f^*(w)) \mid w_{nt}, y_{nt}],$$

which is equal to $\mathrm{E}[u_r(w_{nt}, x_f^*(w)) \mid w_{nt}]$ under independence (A2i) between the distribution of u_r and y.

(ii) Under A2iii

$$\mathrm{E}[u_r(w_{nt}, x_f) \mid w_{nt}, y_{nt}] = \mathrm{E}[u_r(w_{nt}, x_f) \mid w_{nt}] = \mathrm{U}(w_{nt}).$$

□

Proof of Proposition 4.
There are two types of unobserved heterogeneities here: one due to unobserved x_f, and one due to heterogenous functional forms for u_r and v_r over individuals. For simplicity we use the subscript r to denote this heterogeneity. Let $f_{u|x}$ denote the conditional density function of $u_r(w,x_f,t) \mid x_f$. Under Assumption A2ii we can write $f_{u|x} = f_u$ where f_u denotes the marginal density of u_r. Let us define the average fixed and variable cost functions (over all firms in our sample) as

$$\bar{u}(w,x_f,t) \equiv \int u_r(w,x_f,t) f_u(r) dr$$
$$\bar{v}(w,x_f,y,t) \equiv \int v_r(w,x_f,y,t) f_v(r) dr.$$

These functions still depend on the unobserved heterogeneity in x_f, but individual heterogeneity in the cost functions u_r and v_r is integrated out. Let us also consider

$$\bar{\gamma}^U(w,x_f,t) \equiv \frac{\bar{u}(w,x_f,t)}{U(w,t)}, \qquad \bar{\gamma}^V(w,x_f,y,t) \equiv \frac{\bar{v}(w,x_f,y,t)}{V(w,y,t)},$$

and (we skip the arguments for simplicity)

$$\bar{c} = \bar{\gamma}^U U + \bar{\gamma}^V V.$$

(i) Using the optimality condition $\partial c_r / \partial x_f = 0$, and under A2ii, it follows that $\partial \bar{c} / \partial x_f = 0$. So, conditionally on observations (w,y,t), we write

$$\mathrm{cov}\left[\bar{\gamma}^U, \bar{\gamma}^V\right] = \mathrm{cov}\left[(\bar{c} - \bar{\gamma}^V V) / U, \bar{\gamma}^V\right] = \mathrm{cov}\left[-\bar{\gamma}^V V / U, \bar{\gamma}^V\right] = -\frac{V}{U} \mathrm{V}\left[\bar{\gamma}^V\right] \leq 0$$
$$\mathrm{V}\left[\bar{\gamma}^U\right] = \mathrm{V}\left[(\bar{c} - \bar{\gamma}^V V) / U\right] = \frac{V^2}{U^2} \mathrm{V}\left[\bar{\gamma}^V\right].$$

Under Assumption A2ii we can write

$$\begin{aligned}
\mathrm{cov}(\bar{\gamma}^U, \bar{\gamma}^V) &= \int (\bar{\gamma}^U - 1)(\bar{\gamma}^V - 1) f_x dx_f \\
&= \int \left(\int_{\mathcal{R}} \gamma^U f_{uv}(r) dr - 1 \right) \left(\int_{\mathcal{R}} \gamma^V f_{uv}(r) dr - 1 \right) f_x dx_f \\
&= \int \int_{\mathcal{R}} (\gamma^U - 1)(\gamma^V - 1) f_{uv}(r) f_x(x_f) dr dx_f \\
&= \int \int_{\mathcal{R}} (\gamma^U - 1)(\gamma^V - 1) f_{uv|x}(r \mid x_f) f_x(x_f) dr dx_f \\
&= \mathrm{cov}(\gamma^U, \gamma^V),
\end{aligned}$$

where the fourth equality follows from the fact that under A2ii we have the independence of individual heterogeneity with respect to the level of fixed inputs: $f_{uv|x}(r \mid x_f) = f_{uv}(r)$. Putting things together, we have $\text{cov}(\overline{\gamma}^U, \overline{\gamma}^V) = \text{cov}(\gamma^U, \gamma^V) \leq 0$.

(ii) Similarly, the variance matrices satisfy $\text{V}[\overline{\gamma}] = \text{V}[\gamma]$ and so

$$\text{V}[\gamma] = \begin{bmatrix} \dfrac{V^2}{U^2}\text{V}\left[\overline{\gamma}^V\right] & -\dfrac{V}{U}\text{V}\left[\overline{\gamma}^V\right] \\ -\dfrac{V}{U}\text{V}\left[\overline{\gamma}^V\right] & \text{V}\left[\overline{\gamma}^V\right] \end{bmatrix},$$

whose determinant is zero. □

Proof of Proposition 5.

We rewrite C^{TL} as

$$C^{TL}(w,y,t) = b(w,t)y^{\beta_y + \ln w^\top B_{wy} + \frac{1}{2}\beta_{yy}\ln y + \beta_{yt}t},$$

with

$$b(w,t) \equiv \exp\left(\beta_0 + \beta_w^\top \ln w + \beta_t t + \frac{1}{2}\ln w^\top B_{ww} \ln w + \ln w^\top B_{wt} t + \frac{1}{2}\beta_{tt}t^2\right) > 0.$$

If $\beta_{yy} \leq 0$, then

$$\lim_{y \to 0^+} C^{TL}(w,y,t) = 0, \tag{30}$$

whereas if $\beta_{yy} > 0$,

$$\lim_{y \to 0^+} C^{TL}(w,y,t) = +\infty.$$

The cost function is nondecreasing in $y > 0$ iff

$$\frac{\partial C^{TL}}{\partial y}(w,y,t) = (\beta_y + \ln w^\top B_{wy} + \beta_{yy}\ln y + \beta_{yt}t)\frac{C^{TL}(w,y,t)}{y} \geq 0.$$

If $\beta_{yy} > 0$, then

$$\lim_{y \to 0^+} \frac{\partial C^{TL}}{\partial y}(w,y,t) < 0,$$

and $\partial C^{TL} / \partial y$ becomes positive only for y sufficiently large. □

Two-point flexible functional forms

DIEWERT, W. E., and T. J. WALES [1987], (p. 45-46) define a one point (1FFF) flexible cost function at the point (w^0, y^0, t^0) as one which is able to approximate an arbitrary cost function C^0 locally, where C^0 is continuous and homogeneous of degree one in w. This definition is satisfied if and only if C has "enough free parameters so that the following $1 + (J+2) + (J+2)^2$ equations can be satisfied":

$$\begin{aligned} C(w^0, y^0, t^0) &= C^0(w^0, y^0, t^0) \\ \nabla C(w^0, y^0, t^0) &= \nabla C^0(w^0, y^0, t^0) \\ \nabla^2 C(w^0, y^0, t^0) &= \nabla^2 C^0(w^0, y^0, t^0), \end{aligned} \tag{31}$$

where the ∇C (respectively $\nabla^2 C$) denotes the first (second) order partial derivatives with respect to all arguments of C. Since the Hessian is symmetric and C is linearly homogeneous in w, this system includes only $J(J+1)/2 + 2J + 3$ free equations. The requirements (31) have to be fulfilled at a single point y^0 which is chosen to be positive, so the 1FFF definition is compatible with the absence of fixed cost. This explains why the Translog is flexible although $u^{TL} \equiv 0$. This drawback of 1FFF explains why we consider a 2FFF.

A cost function is a 2FFF if it has enough free parameters to satisfy the following $1 + (J+1) + (J+1)^2 + 1 + (J+2) + (J+2)^2$ equations:

$$\begin{aligned} U(w^0, t^0) &= U^0(w^0, t^0), \\ \nabla U(w^0, t^0) &= \nabla U^0(w^0, t^0), \\ \nabla^2 U(w^0, t^0) &= \nabla^2 U^0(w^0, t^0). \end{aligned} \tag{32}$$

The function U^0 denotes an arbitrary fixed cost function, and U is our fixed cost specification. Similarly, our variable cost specification has to satisfy the following criteria, for $y^0 > 0$,

$$\begin{aligned} V(w^0, y^0, t^0) &= V^0(w^0, y^0, t^0), \\ \nabla V(w^0, y^0, t^0) &= \nabla V^0(w^0, y^0, t^0), \\ \nabla^2 V(w^0, y^0, t^0) &= \nabla^2 V^0(w^0, y^0, t^0). \end{aligned} \tag{33}$$

Since U is linearly homogeneous in w, and its Hessian is symmetric, this imposes the following additional restrictions $2 + J + (J+1)J/2$ on U:

$$w^\top \frac{\partial U}{\partial w}(w,t) = U(w,t), \qquad w^\top \frac{\partial^2 U}{\partial w \partial t}(w,t) = \frac{\partial U}{\partial t}(w,t),$$

$$w^\top \frac{\partial^2 U}{\partial w \partial w^\top}(w,t) = 0, \qquad \nabla^2 U(w,t) = \nabla^2 U(w,t)^\top$$

It turns out that the fixed cost function U has at least $(J+1)+J(J+1)/2$ free parameters in order to be flexible. Similarly, the variable cost function V must have at least $(J+2)+(J+1)(J+2)/2$ free parameters. In total, a 2FFF cost function must have at least $1+3(J+1)+J(J+1)$ free parameters. Moreover, in order to identify V as a variable cost function, we impose

$$V(w^0,0,t^0)=0.$$

Note that (32) and (33) imply (31), but not conversely.

References

Bartelsman, E., and W. Gray (1996): "The NBER Manufacturing Productivity Database", Technical Working Paper 205, National Bureau of Economic Research.

Baumol, W. J., and R. D. Willig (1981): "Fixed Costs, Sunk Costs, Entry Barriers, and Sustainability of Monopoly", *The Quarterly Journal of Economics*, **96**, 405-431.

Berry, S., and P. Reiss (2007): "Empirical Models of Entry and Market Structure", in Armstrong, M., and R. Porter (ed.), *Handbook of Industrial Organization*, **3**, Elsevier.

Blackorby, C., and W. Schworm (1984): "The Structure of Economies with Aggregate Measures of Capital: A Complete Characterization", *Review of Economic Studies*, **51**, 633-650.

Blackorby, C., and W. Schworm (1988): "The Existence of Input and Output Aggregates in Aggregate Production Functions", *Econometrica,* **56**, 613-643.

Brito, D., P. Pereira, and J. J. S. Ramalho (2013): "Mergers, Coordinated Effects and Efficiency in the Portuguese Non-Life Insurance Industry", *International Journal of Industrial Organization*, **31**, 554-568.

Bresnahan, T. F., and P. C. Reiss (1991): "Entry and Competition in Concentrated Markets", *Journal of Political Economy*, **99**, 977-1009.

Browning, M. J. (1983): "Necessary and Sufficient Conditions for Conditional Cost Functions", *Econometrica*, **51**, 851-857.

Caves, D. W., L. R. Christensen, and J. A. Swanson (1981): "Productivity Growth, Scale Economies, and Capacity Utilization in U.S. Railroads, 1955-1974", *American Economic Review,* **71**, 994-1002.

Chen, X. (2017): "Biased Technical Change, Scale, and Factor Substitution in U.S. Manufacturing Industries", *Macroeconomic Dynamics*, **21**, 488-514.

Chirwa, E. W. (2004): "Industry and Firm Effects of Privatization in Malawian Oligopolistic Manufacturing", *The Journal of Industrial Economics*, **52**, 277-290.

Christensen, L. R., D. W. Jorgenson, and L. J. Lau (1971): "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica*, **39**, 255-256.

Crépon, B., R. Desplatz, and J. Mairesse (2005): "Price-Cost Margins and Rent Sharing: Evidence from a Panel of French Manufacturing Firms", *Annales d'Économie et de Statistique*, **79/80**, 583-610.

Dehez, P., J. H. Drèze, and T. Suzuki (2003): "Imperfect Competition à la Negishi, Also with Fixed Costs", *Journal of Mathematical Economics*, **39**, 219-237.

Diewert, W. E. (2008): "Cost Functions", *The New Palgrave Dictionary of Economics*, Second Edition, Durlauf, S. N., and L. E. Blume (eds.), Palgrave Macmillan.

Diewert, W. E., and K. J. Fox (2008): "On the Estimation of Returns to Scale, Technical Progress and Monopolistic Markups", *Journal of Applied Econometrics*, **145**, 174-193.

Diewert, W. E., and D. Lawrence (2002): "The Deadweight Costs of Capital Taxation in Australia", in *Efficiency in the Public Sector*, Kevin J. Fox (ed.), Kluwer Academic Publishers, 103-167.

Diewert, W. E., and T. J. Wales (1987): "Flexible Functional Forms and Global Curvature Conditions", *Econometrica*, **55**, 43-68.

Fuss, M. A. (1977): "The Structure of Technology Over Time: A Model for Testing the "Putty-Clay" Hypothesis", *Econometrica*, **45**, 1797-1821.

Garcia, S., M. Moreaux, and A. Reynaud (2007): "Measuring Economies of Vertical Integration in Network Industries: An Application to the Water Sector", *International Journal of Industrial Organization*, **25**, 791-820.

Gladden, T., and C. Taber (2009): "The Relationship between Wage Growth and Wage Levels", *Journal of Applied Econometrics*, **24**, 914-932.

Gorman, W. M. (1995): *Separability and Aggregation, Collected Works of W. M. Gorman*, Volume I, Blackorby, C., and A. F. Shorrocks (eds.), Clarendon Press: Oxford, 1995.

Klette, T. J. (1999): "Market Power, Scale Economies and Productivity: Estimates from a Panel of Establishment Data", *The Journal of Industrial Economics*, **47**, 451-476.

Koebel, B., M. Falk, and F. Laisney (2003): "Imposing and Testing Curvature Conditions on a Box-Cox Cost Function", *Journal of Business and Economic Statistics*, **21**, 319-335.

KRUGMAN, P. (1979): "Increasing Returns, Monopolistic Competition, and International Trade", *Journal of International Economics*, **9**, 469-479.

LAU, L. J. (1976): "A Characterization of the Normalized Restricted Profit Function", *Journal of Economic Theory*, **12**, 131-163.

LAU, L. J. (1978): "Testing and Imposing Monoticity, Convexity and Quasi-Convexity Constraints", in Fuss, M., and D. McFadden (eds.). *Production Economics: A Dual Approach to Theory and Applications*, **1**, North Holland, 409-453.

MELITZ, M. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity", *Econometrica*, **71**, 1695-1725.

MORRISON, C. J. (1988): "Quasi-Fixed Inputs in U.S. and Japanese Manufacturing: A Generalized Leontief Restricted Cost Function Approach", *Review of Economics and Statistics*, **70**, 275-287.

MURPHY, K., A. SHLEIFER, and R. VISHNY (1989): "Industrialization and the Big Push", *Journal of Political Economy*, **97**, 1003-26.

OLLEY, S., and A. PAKES (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, **64**, 1263-1297.

RÖLLER, L.-H. (1990): "Modelling Cost Structure: The Bell System Revisited", *Applied Economics*, **22**, 1661-1674.

VARIAN, H. R. (1992): *Microeconomic Analysis* (3rd ed.), Norton & Co.

VINER, J. (1931): "Costs Curves and Supply Curves", *Zeitschrift für Nationalölkonomie*, **3**, 23-46.