Bureau
d'économie
théorique
et appliquée
(BETA)
UMR 7522

Documents de travail

« Fixed cost, variable cost, markups and returns to scale »

Auteurs

Xi Chen, Bertrand M. Koebel

Document de Travail n° 2013 – 13

Septembre 2013

Faculté des sciences économiques et de gestion
Pôle européen de gestion et d'économie (PEGE)
61 avenue de la Forêt Noire
F-67085 Strasbourg Cedex

Secétariat du BETA
Géraldine Manderscheidt
Tél. : (33) 03 68 85 20 69
Fax : (33) 03 68 85 20 70
g.manderscheidt@unistra.fr
www.beta-umr7522.fr

Fixed cost, variable cost, markups and returns to scale

Xi Chen*

Bertrand M. Koebel**

September 2013

Abstract. 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 estimates both fixed and variable components of the cost function and studies how fixed and variable costs interact and affect firms' behavior in terms of price setting and returns to scale.

Keywords: identification, imperfect competition, returns to scale, unobserved heterogeneity.

JEL Classification: C33, D24, E10, J23, L60.

* BETA, CNRS, Université de Strasbourg. Email: xchen@unistra.fr.

** Corresponding author: Bureau d'Economie Théorique et Appliquée (BETA), CNRS, Université de Strasbourg, 61 avenue de la Forêt Noire, 67085 Strasbourg Cedex (France). Tel (+33) 368 852 190. Fax (+33) 368 852 071. Email: koebel@unistra.fr.

We would like to thank Pierre Dehez, Andreas Irmen, François Laisney, Clemens Puppe and the participants of seminars in Aix-en-Provence, Cergy, Luxembourg, Maastricht, Nancy, Trier and Strasbourg for their helpful comments.

1. Introduction

A long tradition going back to Viner (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. If some variable inputs are substitutable to fixed inputs, then this sharp distinction vanishes. This paper extends the microeconomic foundations of production analysis by allowing each input to have a fixed and a variable part.

Empirical specifications of production and cost functions are also shaped by this dichotomy between fixed and variable inputs. Some specifications consider fixed costs to be the cost of the fixed inputs. Others, like the Cobb-Douglas, the CES, and even flexible functional forms like the Translog, assume that fixed costs are nonexistent. We propose a generalization of the Translog functional form which is compatible with inputs having both a fixed and a variable part. Our empirical results support the extended Translog specification and show that the fixed cost is significant and neglecting it yield estimation biases, especially on the markup and the rate of returns to scale. Fixed costs, although not functionally dependent on the output level, are correlated with output, and should be explicitly considered to avoid these estimation biases. Our findings are compatible with the predictions of models with heterogenous technologies (see e.g. Acemoglu and Shimer (2000) and Cabral (2012)), in which there is a trade-off between production functions having a large fixed cost and low variable cost and those with the converse configuration.

Despite the challenging result of Baumol and Willig (1981, p.405) according to which fixed costs "do not have the welfare consequences normally attributed to barriers to entry", there is a quite large literature on fixed inputs. Fixed costs are useful for explaining coordination failure (Murphy et al., 1989) and international trade (Krugman, 1979, Melitz, 2003). Blackorby and Schworm (1984, 1988) and Gorman (1995) have shown that fixed inputs hamper the aggregation of production (and cost) functions, whereas a fixed cost does not represent an aggregation problem. Fixed costs are also

[1] In the words of Viner (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"."

considered in general equilibrium theory with imperfect competition, see for instance Dehez et al. (2003). Contributions in the field of industrial organisation on the reasons and consequences of fixed (and sunk) cost, are so numerous that we cannot survey them here. Berry and Reiss (2007) discuss some important issues on identification and heterogeneity of fixed costs. Differences between fixed and sunk cost are commented by Wang and Yang (2004) and Sutton (2007).

We mainly contribute to the literature in production analysis. One objective is to characterize and estimate both fixed and variable components of the cost function, to investigate their heterogeneity over firms and study how fixed costs affect their behavior in terms of price setting and returns to scale. Microeconomic textbooks present alternative characterizations of fixed costs. We follow Baumol and 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 the production technology which generates the fixed cost, an issue which is usually neglected when dealing with fixed cost. It is well known (see Mas Colell et al., 1995, p.135) that fictitious inputs can be used for imposing constant returns to scale on arbitrary technologies. This paper shows that the fixed cost of production can be represented as the cost of fictitious (unobserved) inputs. We first characterize the production technology which generates the traditional fixed cost and show that it is quite restrictive and given by $y = F\left(x_v + x_f\right)$ where x_v denotes the vector of variable inputs and x_f the fixed inputs. As total input x can always be additively split into two categories, the structure F may be considered as perfectly general. However, two physically similar inputs may be technologically different and we propose to extend the production function to $y = G\left(x_v, x_f\right)$. This extended production technology generates a fixed cost which is not equal to the cost of inputs x_f, and identification of fixed inputs is no longer possible. However, the amount of inputs which allows to initiate production is well identified.

Our theoretical contribution also requires extending the econometric toolbox for estimating cost functions. First, usual cost function specifications are not compatible with a flexible specification of the fixed cost. For approximating a cost function with a fixed cost component, we have to go beyond (locally) flexible cost functions, and develop a cost specification which is a valid approximation at two points: around the actual point

of production and around the breakup point which allows a firm to start production. Second, as the inputs x_v and x_f cannot be observed, we have to amend the traditional estimation method by introducing unobserved and correlated heterogeneity in the fixed and variable cost specification. We extend Swamy's (1970) random coefficient estimator to our nonlinear setup. The empirical part of this paper uses panel data for US manufacturing sectors in order to estimate the height and the type of fixed cost as well as their implications in terms of markup pricing, returns to scale and technical change.

In Sections 2 and 3 we explore two definitions of fixed costs and their microeconomic foundations. Sections 4, 5, and 6 discuss econometric issues related to fixed costs: biases when they are neglected, specification issues, and unobserved heterogeneity. Section 7 reports the empirical results, obtained for 462 US manufacturing industries observed over the years 1958 to 2005.

2. Defining fixed costs and fixed inputs

The definition of fixed costs is central in economics and is briefly discussed in most introductory microeconomic textbooks.[2] One difficulty with most definitions is that they do not highlight the relationship between the fixed cost and the fixed inputs. Are fixed inputs physically fixed? Do fixed inputs correspond to nonoptimal choices? This section shows that it is not necessarily the case: a fixed cost can arise in a context where all inputs are optimally adjusted.

Most economists agree that the fixed cost u corresponds to the part of the cost which does not vary with the level of production:

$$c(w,y) = u(w) + v(w,y), \tag{1}$$

where w denotes the input prices and y the output level. Function v corresponds to the variable cost of production and satisfies $v(w,0) = 0$. Any cost function can uniquely be written in this way by defining

$$\begin{aligned} u(w) &\equiv c(w,0) \\ v(w,y) &\equiv c(w,y) - c(w,0). \end{aligned} \tag{2}$$

[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 commented in Diewert's (2008) contribution on cost functions.

We will comment the following alternative definitions for the fixed cost and fixed inputs.

Definitions 1. For an active firm, the *fixed cost* is

a) the accounting cost of the inputs which are physically fixed.

b) the cost of the inputs required for producing an arbitrarily small amount of output.

Definition 1 does not require (at this stage) that the level of the fixed cost is optimal (so it does not necessarily correspond to the minimal value of the accounting cost). In D1b the inputs required for initiating production could be physically fixed but it is not necessary the case. Since the cost function is related to input demands x° by the accounting relationship $c(w,y) = w^{\top}x^{\circ}(w,y)$ for any $y \geq 0$, we obtain the level of fixed cost compatible with D1b as

$$u(w) = \lim_{y \to 0^{+}} c(w,y) = \lim_{y \to 0^{+}} w^{\top}x^{\circ}(w,y).$$

This shows that D1b implies that the fixed cost does not change with the production level, but can change with w. Whereas it is straightforward to define *variable inputs* as inputs whose level can be adjusted to minimize their accounting cost and can *possibly* be set to zero, the definition of fixed inputs is more involved, as they are not *necessarily* optimal, nor can they *necessarily* be set to zero.

We show that the fixed cost $u(w)$ does not necessarily correspond to the cost of the fixed inputs, but that it also includes a part of the cost of variable inputs when they are sufficiently complementary to the fixed inputs. 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 and these have been largely ignored in the literature.

3. A microeconomic framework for fixed costs

The main result of this section characterizes an extended production function able to describe fixed inputs in a more general way than the existing literature. A shortcoming

of the traditional restricted cost function (see Subsection 3.1), is that it relies on a partition of all inputs into two *disjoint* categories: variable and fixed inputs. Actually, similar inputs can be used for different types of production activities. Engineers, for instance, can be allocated to production or to research and development activities. While engineers' production increases the *current* output level, it is not the case when they are allocated to research and development, which withdraws them from production (like in Aghion and Howitt, 1992, for instance). 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. Before presenting the extended production and cost function, we shortly overview traditional production analysis.

3.1 On the limitations of traditional production analysis

For modelling fixed inputs, production analysis relies on a partition of the input vector x into two *disjoint* categories: those which can be adjusted (variable inputs, denoted $\widetilde{x}$) and those which are fixed or quasi-fixed ($\overline{x}$):[3]

$$x = \begin{pmatrix} \widetilde{x} \\ \overline{x} \end{pmatrix} \geq 0. \tag{3}$$

The corresponding input prices are denoted by $\left(\widetilde{w}^{\top}, \overline{w}^{\top}\right)^{\top}$. The output level is given by $y = F(x)$ where $F : \mathbb{R}^J \to \mathbb{R}$ denotes the production function which is increasing in x. The *restricted* variable cost function is defined as:

$$V_r\left(\widetilde{w}, \overline{x}, y\right) = \min_{\widetilde{x} \geq 0} \left\{ \widetilde{w}^{\top} \widetilde{x} : F\left(\widetilde{x}, \overline{x}\right) \geq y \right\}.$$

The properties of the restricted cost functions have been investigated by Lau (1976) and Browning (1983). For empirical implementations see e.g. Caves et al. (1981), Pindyck and Rotemberg (1983) and Morrison (1988). The total restricted cost function is given by

$$V_r\left(\widetilde{w}, \overline{x}, y\right) + \overline{w}^{\top} \overline{x}, \tag{4}$$

where the last term denotes the fixed cost. In the long-run, all the fixed inputs can be adjusted at their optimal level and this defines the long-run or *unrestricted* cost function:

$$c\left(w, y\right) = \min_{\overline{x} \geq 0} \left\{ V_r\left(\widetilde{w}, \overline{x}, y\right) + \overline{w}^{\top} \overline{x} \right\} = \widetilde{c}\left(w, y\right) + \overline{c}\left(w, y\right), \tag{5}$$

[3] 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.

where $\widetilde{c}(w,y) = V_r\left(\widetilde{w}, \overline{x}^*(w,y), y\right)$ represents the long run variable cost, and $\overline{c}(w,y) = \overline{w}^\top \overline{x}^*(w,y)$ is the long run fixed cost. Function $\overline{x}^*$ denotes the optimal level of fixed inputs, which, without further restrictions on V_r, depends on the production level. As a consequence, this approach violates (in the long run) both definitions given in D1. More than that, in the long run it is not possible to identify $\widetilde{c}$ separately from $\overline{c}$, unless we make strong *a priori* assumptions on which inputs are fixed in the short run. A further drawback of technology F appears when we impose that V_r be a variable cost function, namely $V_r(\widetilde{w}, \overline{x}, 0) = 0$. This restriction implies that there are no fixed cost in the long run: $\overline{x}^*(w,0) = 0$ (unless we impose a positive lower bound to $\overline{x}$).

So, according to traditional production analysis, the only justification for fixed cost is that physically fixed inputs cannot be optimally adjusted (either for technical reasons or for lack of rationality). This view excludes a variety of interesting situations in which fixed and variable inputs are imperfect substitutes and play different roles in production.

3.2 Another view of the traditional production function

Instead of partitioning x into two disjoint types of inputs, let us assume that each input comprises a part which can be adjusted and a part which is fixed (in a sense that is clarified in Definition 2 below):

$$x = x_v + x_f, \tag{6}$$

with $x, x_v, x_f \in \mathbb{R}^J_+$. This generalizes (3) which is obtained as a special case when $x_v = \left(\widetilde{x}^\top, 0^\top\right)^\top$ and $x_f = \left(0^\top, \overline{x}^\top\right)^\top$. This subsection shows that the variable and fixed cost functions used in production analysis is generated from an additive production function

$$y = F\left(x_v + x_f\right), \tag{7}$$

which requires perfect substitutability between x_v and x_f.

As our purpose is to describe the production possibilities for a production level close to zero (in order to be consistent with D1b), we define the input requirement set as follow.

Definition 2. In terms of the traditional production function, the fixed cost is the cost associated to inputs belonging to the *input requirement set* X_F defined as

$$X_F \equiv \lim_{\varepsilon \to 0^+} \{z \geq 0 : F(z) = \varepsilon\}.$$

Definition 2 requires that the limiting isoquant X_F exists. Definition 2 is useful to characterize the fixed cost in terms of the production function F : it is easy to show that a fixed cost occurs if the set X_F does not include the point $x = 0$.[4] In order to be compatible with Definition 1b, we consider in Definition 2 the isoquant corresponding to the production level $\varepsilon > 0$, instead of $\varepsilon = 0$, because with most production functions compatible with a fixed cost, the condition $F(x) = 0$ characterizes a thick isoquant, in the sense that, if it is possible to produce nothing with something ($\exists x > 0 : F(x) = 0$), then it is also possible to produce nothing with even less (there exist $x' < x$ such that $F(x') = 0$). So, only the upper frontier of the set $\{z \geq 0 : F(z) = 0\}$ is interesting for identifying a fixed cost. Let us investigate the implications of this additive structure in terms of the restricted variable and total cost functions:

$$\begin{aligned} v_r(w, x_f, y) &= \min_{x_v \geq 0} \left\{ w^\top x_v : F(x_v + x_f) \geq y \right\} \\ c_r(w, x_f, y) &= v_r(w, x_f, y) + w^\top x_f. \end{aligned}$$

The restriction $x_v \geq 0$ is important here, because it can be optimal to use no variable inputs at all for some levels of x_f.

Proposition 1. Let $x_f \in X_F \neq \varnothing$ and $x_v \geq 0$. Then $c_r(w, x_f, 0) = w^\top x_f \geq 0$, $v_r(w, x_f, 0) = 0$. The restricted cost function c_r and the cost minimizing variable inputs x_v^* satisfy either

(i) $x_v^* > 0$ and for $y > 0$,

$$c_r(w, x_f, y) = C(w, y) > w^\top x_f \tag{8}$$

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

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

$$c_r(w, x_f, y) = V_r(\widetilde{w}, \overline{x}, y) + \overline{w}^\top \overline{x}, \tag{9}$$

where $\overline{x}$ is a subvector of x_f and $\widetilde{w}$ corresponds to the price subvector of $w = \left(\widetilde{w}^\top, \overline{w}^\top\right)^\top$ corresponding to $x_{v,j}^* > 0$.

The proof of this result is given in the Appendix. Proposition 1 states that the variable cost is zero when production vanishes. This result is driven by the additive structure of F which ensures that if there exists a point x such that $F(x) = 0$, then x_v

[4] For the purpose of exposition we assume that the minimum value of y included in the range of F is zero, but we could easily generalize to any other value.

can be set to zero in the additive decomposition $x = x_v + x_f$. In the case of Proposition 1(i), the production function F yields a cost function which is independent of the level of fixed input, and which is compatible with both Definitions 1a and 1b. A frustrating consequence of Proposition 1(i) is that fixed inputs can be seen as if they were set at their optimal level, as:

$$\frac{\partial c_r}{\partial x_f}\left(w, x_f, y\right) = 0 \Leftrightarrow -\frac{\partial v_r}{\partial x_f}\left(w, x_f, y\right) = w.$$

It is the perfect substitutability between the variable and the fixed inputs which is driving this result. Any mistake in adjusting x_f can be perfectly compensated by setting x_v optimally. In summary, technology F is not really suitable for modelling fixed inputs, as it lacks generality. Proposition 1(ii) gives the general formulation of the cost function corresponding to F when corner solutions for the variable inputs are allowed. The structure of the cost function (9) is the same as in (4) and is common in traditional production analysis (see Chambers, 1988, for instance). So we conclude this section by noting that production function F with an additive structure between x_v and x_f is behind the traditional theory of fixed and variable costs. This additive structure is restrictive and hides important features of production theory.[5]

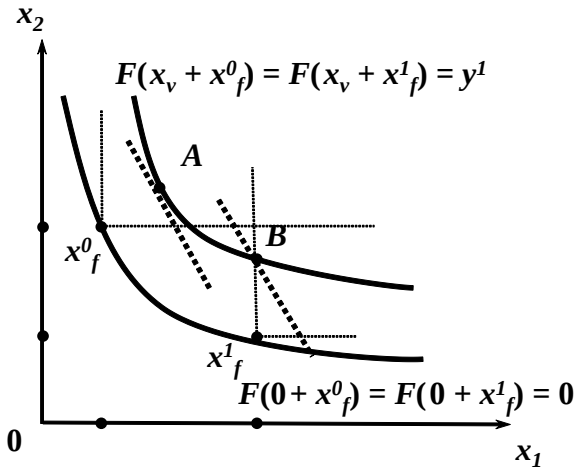


Figure 1: Isoquants for $F\left(x_v + x_f\right)$

Figure 1 illustrates Proposition 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

[5] One restriction is that

$$\frac{\partial^2 c_r}{\partial \widetilde{w}_j \partial \overline{w}_k}\left(w, x_f, y\right) = 0.$$

For given x_f, y, there is no substitutability between inputs j and k. This is too restrictive because, even for given x_f, inputs j and k can be substituted for each other because they have a fixed and a variable component.

x_f^0. At given input prices w, the firm minimizes its variable cost of producing y^1 at the interior point A. At this point, according to Proposition 1i, the cost function is given by $C(w,y)$. With another level of fixed inputs, however, the available set of variable inputs will be different. With x_f^1, the minimal variable cost for producing y^1 is achieved at B, on the boundary of the set $\left\{x_v \geq 0 : F\left(x_v + x_f^1\right) \geq y^1\right\}$. At point B we have $x_{v1}^* = 0$ and the optimal level of x_{2v}^* is restricted by the level of x_f^1.

3.3 An extended production function

Whereas from the accounting viewpoint both types of inputs x_v and x_f are similar (the cost of a unit of the j^{th} fixed and flexible input is w_j), technologically they should not be restricted to play similar roles as it is the case with $F\left(x_v + x_f\right)$. We now define an extended production function G as $y = G\left(x_v, x_f\right)$ where $G : \mathbb{R}_+^J \times \mathbb{R}_+^J \to \mathbb{R}_+$. For simplicity, we assume that G is single valued, continuously differentiable, increasing in its arguments and that $G\left(0, x_f\right) = 0$. In this context, the restricted variable cost function now becomes:

$$v_r\left(w, x_f, y\right) = \min_{x_v}\left\{w^\top x_v : G\left(x_v, x_f\right) \geq y\right\}. \tag{10}$$

Now, a given input, say capital, can appears twice in (10): once in vector x_v and once in x_f; their marginal productivities can be different. This overlapping structure is similar to the one considered by Blundell and Robin (2000) in consumer analysis. In contrast to their approach, we do not impose that x_v is separable from x_f (a structure which they call latent separability).

Leontief's (1947) aggregation theorem highlights the restrictions which are implicit in production function F. The number $2J$ of inputs x_v and x_f which appear in G can be reduced to the J aggregate inputs $x_v + x_f$ iff we have

$$\frac{\partial G}{\partial x_{v_i}} = \frac{\partial G}{\partial x_{f_i}}, \quad \forall i = 1, \ldots, J.$$

We do not assume in the sequel that these restrictions necessarily apply to G.

One difficulty with (10), is that if v_r is defined for any arbitrary levels of x_f, we can switch the notation from x_f to x_v and rewrite $v_r(w, x_v, y)$. So, in order to be able to identify x_f as the fixed inputs, we need to put more structure on v_r, and we do this by introducing restrictions derived from the definition of the fixed cost and inputs.

Definition 3. In terms of the extended production function G, the fixed cost is the

cost associated to inputs belonging to the *fixed input requirement set* X_G defined as

$$X_G \equiv \lim_{\varepsilon \to 0^+} \{z \geq 0 : G(0,z) = \varepsilon\}. \tag{11}$$

D3 defines the set of all fixed input combinations required for starting production. This definition is more general than D2, because it does not assume that fixed and variable inputs are perfectly substitutable. As for X_F, we impose that $x_v = 0$ belongs to the fixed input requirement set X_G, but get rid of additivity. The next result is a straightforward extension to technology G of those available for technology F.[6]

Proposition 2. If $x_f \in X_G$ then,

(i) $v_r(w, x_f, 0) = 0$, $v_r(w, x_f, y) > 0$ for any $y > 0$

(ii) v_r is increasing in y

(iii) v_r is decreasing in x_f.

Proposition 2 means that the restricted variable cost function v_r satisfies the properties of a variable cost function: it vanishes for arbitrarily small production levels. As a consequence, the restricted fixed cost is given by

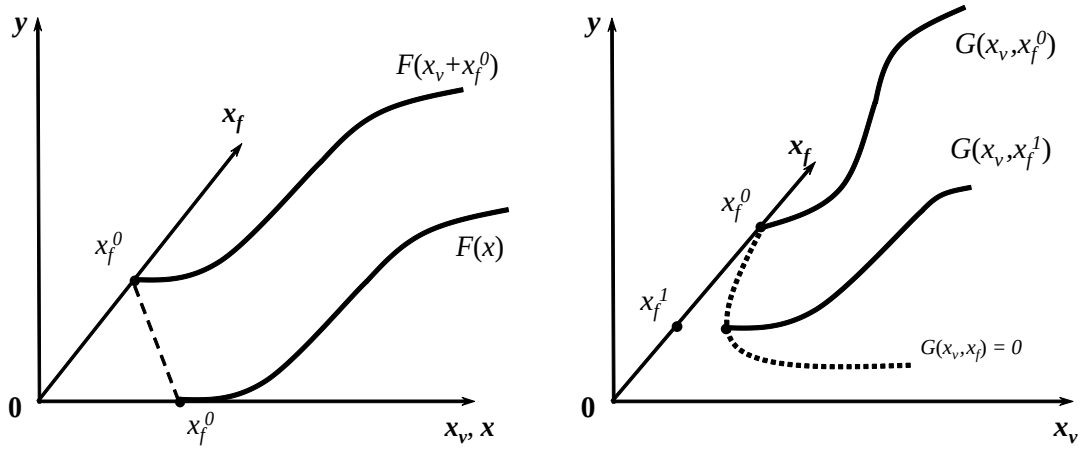
$$u_r(w, x_f) \equiv \lim_{y \to 0^+} c_r(w, x_f, y) = w^\top x_f,$$

and total restricted cost satisfies

$$c_r(w, x_f, y) = u_r(w, x_f) + v_r(w, x_f, y). \tag{12}$$

Both production technologies F and G are represented on Figure 2 in the case where a single input is decomposed into a fixed and a variable component. Figure 2a illustrates how the introduction of a fixed input x_f satisfying $F\left(x_f^0\right) = 0$ and the reparameterization $x \equiv x_v + x_f$ yield the technology $F\left(x_v + x_f^0\right)$. On Figure 2b, the isoquant corresponding to the startup production level $G(x_v, x_f) = \varepsilon$ is not a straight line, which opens the possibility to choose a fixed input different from x_f^0 as an admissible value for starting production. Input x_f^1 for instance, allows to start production with production function $G\left(x_v, x_f^1\right) \neq G\left(x_v, x_f^0\right)$, provided that x_v is sufficiently high for compensating the decline from x_f^0 to x_f^1.

[6] We only give the properties which are the more interesting for our purpose, see Lau (1976) and Browning (1983) for other properties.



(a) Technology F (b) Technology G

Figure 2: Fixed and variable inputs and production possibilities

We illustrate the usefulness of technology G with an example which also illustrates the claims of Proposition 2.

Example 1. The technology $G : \mathbb{R}^2_+ \to \mathbb{R}_+$ is given by

$$y = G\left(x_v, x_f\right) = \left(x_v + \beta x_f\right) x_f^{\alpha} - \gamma$$

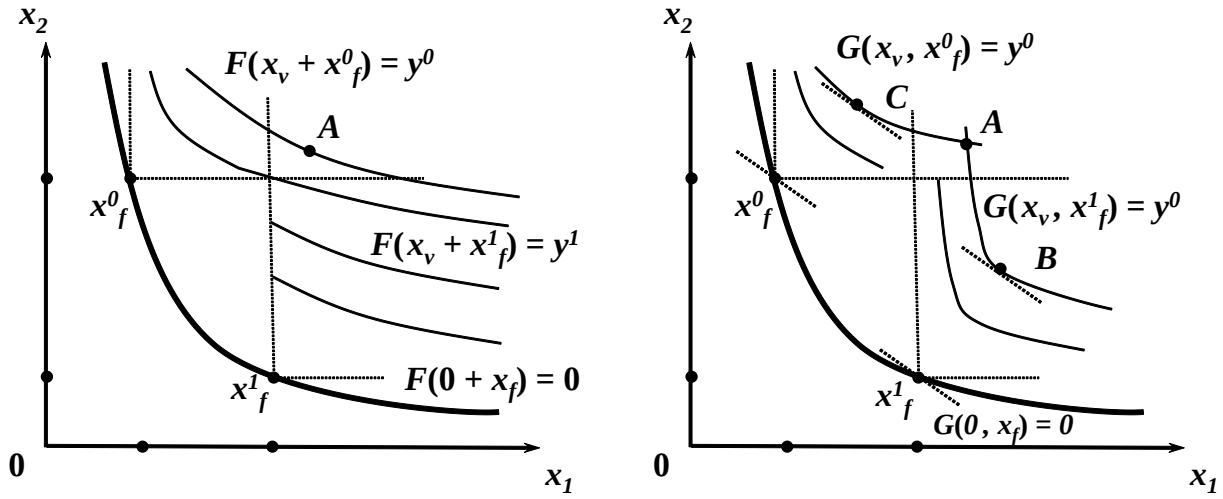
for $y \geq 0$. Here $x_f \in X_G \Leftrightarrow x_f = (\gamma/\beta)^{1/(\alpha+1)}$. This yields the restricted variable cost function

$$v_r\left(w, x_f, y\right) = w x_v^*\left(w, x_f, y\right) = w\left(y + \gamma\right) x_f^{-\alpha} - w\beta x_f = w\frac{y}{x_f^{\alpha}},$$

which satisfies $v_r\left(w, x_f, 0\right) = 0$ for $x_f = (\gamma/\beta)^{1/(\alpha+1)}$. The restricted fixed cost function is $u_r\left(w, x_f\right) = w x_f = w\left(\gamma/\beta\right)^{1/(\alpha+1)}$. For $\alpha = 0$ and $\beta = 1$ we obtain the traditional production function as a special case. Example 1 also illustrates that in both cases of exogenous (physically fixed) and endogenous input x_f, there is no conflict between D1a and D1b.

The structure of the isoquants of F and G is represented in Figure 3 for $J = 2$, in the (x_1, x_2)-plane (with $x_1 = x_{v1} + x_{f1}$). In Figure 3a the slopes of the isoquants corresponding to F only depend upon total input use $x = x_v + x_f$ and not upon the share of the fixed inputs x_f in the composite input x. At point A for instance, it is possible to produce y^0 using fixed input x_f^0 or x_f^1. Only the total input quantity matters and since $x_f^0 + x_v^0 = x_f^1 + x_v^1$ at point A, the choice of the fixed input is irrelevant. Note that, contrary to Fig. 2, the isoquants do not necessarily cross the axes on Fig. 3, because axes now report total input levels for two different inputs, and not just how a

given input is split into variable and fixed amounts.



(*a*) Isoquants for F (*b*) Isoquants for G

Figure 3: Fixed and variable inputs and substitution possibilities

Figure 3*b* represents in the (x_1, x_2)-plane the isoquants for technology G and two different fixed input vectors x_f^0 and x_f^1. With technology G, the choice of 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 3 resembles those typically obtained with putty-putty (or putty-clay or clay-clay) technologies (see e.g. Fuss, 1977). The similarity is due to the fact that we split x into two (fixed and variable) non-additive components. With technology G the choice of a particular fixed input level x_f coincides with a choice of a particular production technology and a specific substitution pattern between variable inputs. On Figure *3b*, the isoquant corresponding to x_f^0 characterizes inputs which can easily be substituted the one for the 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 one exhibiting a specific substitution pattern.

Figure 3*b* also illustrates that if fixed inputs are neglected, production function G is not necessarily quasi-concave in x (at point A). Moreover, optimal choices for input bundles can be located in the zone violating quasi-concavity in x and so the cost function

will not necessarily be concave in w. In the context of fixed cost, imposing simultaneously 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, 1978, and Diewert and Wales, 1987, for seminal contributions on concavity enforcement).

The next difficulty we have to deal with is related to the fact that the level of fixed inputs can be either exogenous or endogenous. Figure 3b depicts at point C a situation at which the variable inputs are optimal given the levels of fixed inputs x_f^0 and production level y^0, however, if x_f could be chosen, the firm would set them to x_f^1 and produce y^0 at point B. It is important to note that isoquant and isocost line are not necessarily tangent at the optimum level x_f^1 for $x_v = 0$.

Whereas variable inputs can by definition be adjusted for minimizing costs, the fixed inputs are not necessarily set at their optimal level. We say that a fixed input x_{fj} is *exogenous* when its actual level is not optimal in the sense that the equality between its shadow value and market price is violated:

$$-\frac{\partial v_r}{\partial x_{fj}}\left(w, x_f, y\right) \neq w_j, \tag{13}$$

for the observed values of $\left(w, x_f, y\right)$ and $x_{fj} > 0$. The extended framework based on $G\left(x_v, x_f\right)$ is useful as it allows to split the input x into a part x_v that is efficiently allocated, and a part x_f which is not necessarily so.[7]

In the long run, fixed inputs can be determined endogenously by the firm, and they may in some case be set to zero. Such a corner solution occurs at $x_f = 0$ if $0 \in X_G$ and:

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

for any $x_f > 0$. Equivalently, the choice $x_f^* = 0$ is (locally) optimal if at point $(w, 0, y)$ the increase in fixed cost is not compensated by a greater reduction of the variable cost:

$$w_j + \frac{\partial v_r}{\partial x_{fj}}\left(w, 0, y\right) > 0.$$

Then it is optimal to adopt a production structure without any fixed input. In many cases however, an inner solution for x_f^* exists. It is characterized by the equality between the shadow value of the fixed input and its market price:

$$-\frac{\partial v_r}{\partial x_{fj}}\left(w, x_f^*, y\right) = w_j. \tag{14}$$

[7] Common explanations for why the level of the fixed inputs is not optimal are related to (i) technological constraints, (ii) indivisibilities of the fixed inputs, (iii) allocative inefficiencies and (iv) intertemporal dependences.

Example 1 (continuation). For $v_r\left(w, x_f, y\right) = w\left(y+\gamma\right)x_f^{-\alpha} - w\beta x_f$, we find that (assuming $\alpha > 0$ and $\beta < 1$)

$$x_f^*\left(w, y\right) = \left(\frac{\alpha}{1-\beta}\left(y+\gamma\right)\right)^{\frac{1}{1+\alpha}}$$

which varies with the level of output. In the traditional case: $\alpha = 0$ and $\beta = 1$, the restricted variable cost function becomes $v_r\left(w, x_f, y\right) = wy$ and we obtain a corner solution $x_f^* = 0$, conformably to Section 3.1. The long-run variable cost function becomes:

$$v_r\left(w, x_f^*, y\right) = w\left(y+\gamma\right)\left(\frac{\alpha}{1-\beta}\left(y+\gamma\right)\right)^{\frac{-\alpha}{1+\alpha}} - w\beta\left(\frac{\alpha}{1-\beta}\left(y+\gamma\right)\right)^{\frac{1}{1+\alpha}}$$

and this does not necessarily vanish anymore for a production level going to zero:

$$v_r\left(w, x_f^*\left(w, 0\right), 0\right) = w\gamma\left(\frac{\alpha}{1-\beta}\gamma\right)^{\frac{-\alpha}{1+\alpha}} - w\beta\left(\frac{\alpha}{1-\beta}\gamma\right)^{\frac{1}{1+\alpha}}.$$

Example 1 illustrates the fundamental identification problem occurring when inputs are optimally adjusted: the fixed cost generally differs from the cost of the fixed inputs. Indeed, after normalizing the variable and fixed cost function according to (2), we obtain the fixed cost

$$u\left(w\right) = w^\top x_v^*\left(w, x_f^*\left(w, 0\right), 0\right) + w^\top x_f^*\left(w, 0\right).$$

When fixed and variable inputs can be imperfectly substituted for each other, the optimal amount of fixed input depends upon w and $x_f^*\left(w, 0\right)$ is not necessarily included in the input requirement set X_G. This means that the level of fixed input cannot be determined ex-ante using only the definition of X_G. When x_f can be adjusted, it is no longer possible to separately identify x_f and x_v. Fortunately, definition D1b of the fixed cost is fully compatible with this situation, but D1a is violated: $u\left(w\right) \neq w^\top x_f^*\left(w, 0\right)$. Briefly, an input cannot be said to be fixed or variable *prima facie*, using only physical properties of the inputs. It is the technology which in last instance determines whether a given input is fixed or variable. This explains why D1b which relies on the technology provide the more general definition of the fixed cost. Few technologies allow to obtain an optimal level of x_f^* independent of y. We characterize them below.

Proposition 3. Assume that the technology G is increasing and quasi-concave in x_v, 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 restricted cost function is given by

$$c_r\left(w, x_f, y\right) = u_r\left(w, x_f\right) + v\left(w, y\right), \tag{15}$$

with $v\left(w, 0\right) = 0$ if and only if the production function is given by

$$G\left(x_v, x_f\right) = F\left(x_v + K\left(x_f\right)\right). \tag{16}$$

(ii) The optimal level of x_f is independent of y if and only if the restricted cost function is (15) or the production function is (16).

Proposition 3 characterizes the cost and production functions which generate a fixed cost. Requirement (16) is less stringent than separability of G in x_f because it does not impose that $K\left(x_f\right)$ be a unique aggregate fixed input. Here, the vector valued function K comprises J aggregates for the fixed inputs. Proposition 3 also aggregates additively some fixed and variable inputs together since F depends upon $x_v + K\left(x_f\right)$. As can be seen by comparing (16) and $F\left(x_v + x_f\right)$, the former is also more general than the traditional production function F for which fixed and variable inputs are perfect substitutes. Figure 4 provides an illustration in the two inputs case ($J = 1$). It shows that x_f does not vary with y, contrary to x_v^*.[8] Figure 4 gives the decomposition of variable input x_v into a fully variable component $x_v^*\left(w, y\right) - x_v^*\left(w, 0\right)$ which can be set to zero when there is no production, and $x_v^*\left(w, 0\right)$ which has to be used for starting production. It also shows how technology $F\left(x_v + K\left(x_f\right)\right)$ differs from $F\left(x_v + x_f\right)$. With $F\left(x_v + K\left(x_f\right)\right)$ there is perfect substitutability between the components of x_v and $K\left(x_f\right)$, but not between x_v and x_f. For a given input x_i, the slope $\left(\partial F/\partial x_{vi}\right)/\left(\partial F/\partial x_{fi}\right)$ of the isoquant (Figure 4) is not restricted to be equal to -1 out of the optimum. Moreover, for two different inputs, x_h and x_i, the slope $\left(\partial F/\partial x_{fh}\right)/\left(\partial F/\partial x_{fi}\right)$ of the isoquant is not restricted to be equal to $\left(\partial F/\partial x_{vh}\right)/\left(\partial F/\partial x_{vi}\right)$ out of the optimum. Fixed inputs can be substituted according to a different pattern than variable inputs.

[8] We also see why the corner solution $x_v^* = 0$ has to be excluded, because at this point the level of x_f^* can vary with y.

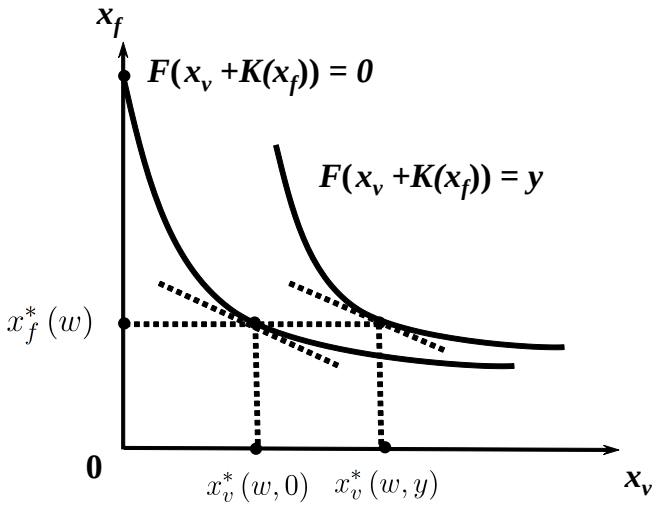


Figure 4: The fixed and variable inputs decomposition

We conclude this section by emphasizing that, even though a separate identification of x_f and x_v is not possible without additional restrictions, it is possible to identify uniquely $x_f^*(w,0) + x_v^*(w,0)$ as well as the level of the fixed cost. If we assume that decomposition (1) is not unique, then there exist $\widetilde{u} \neq u$ and $\widetilde{v} \neq v$ such that:

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

with $\widetilde{v}(w,0) = v(w,0) = 0$. However, the equality

$$u(w) + v(w,y) = \widetilde{u}(w) + \widetilde{v}(w,y)$$

is satisfied for any (w,y) iff $u(w) = \widetilde{u}(w)$ (obtained for $y = 0$) and $v(w,y) = \widetilde{v}(w,y)$, and this proves unicity. It is also interesting to note, that although fixed cost cannot be observed, because the situation in which firms produce an output level close to zero is hypothetical, the level of fixed cost is well identified empirically and can be estimated.

4. Some consequences of neglecting fixed costs

This section discusses three drawbacks arising when fixed inputs are neglected. A first problem of disregarding 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 macroeconomic parameters z and the firm's own production level. With market power, the firms' the optimum

is characterized by:

$$\frac{\partial v_r}{\partial y}\left(w, x_f, y\right) = p\left(1 + \frac{\partial P}{\partial y}\frac{y}{P}\right). \tag{17}$$

This equation and the discussion above shows that a fixed input x_f has an impact on the marginal cost function unless c_r has the specific structure given in (15). 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.

Neglecting the fixed cost is a source of bias. By Shephard's lemma, we have

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

If the fixed cost is neglected, then it enters the residual term which will be correlated with w, which may bias the estimates.

From a theoretical viewpoint, neglecting the fixed cost by setting u (or u_r) equal to zero may lead to 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,

$$\begin{aligned} & c\left(w, 0\right) \geq c\left(w, y\right) + \frac{\partial c}{\partial y}\left(w, y\right)\left(0 - y\right) \\ \Rightarrow \quad & \frac{\partial c}{\partial y}\left(w, y\right)\frac{y}{c\left(w, y\right)} \geq 1 - \frac{c\left(w, 0\right)}{c\left(w, y\right)}. \end{aligned}$$

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 given level of costs and outputs, neglecting the fixed cost leads to an overestimation of the marginal cost, which will also cause an underestimation of the markup. As $\partial c / \partial y\left(w, y\right) = w^{\top} \partial x^* / \partial y\left(w, y\right)$, overestimating the marginal cost often coincides with the overestimation of the input demand sensitivity to output variations. In addition, from an empirical viewpoint, setting the fixed cost equal to zero introduces an omitted-variable bias in the estimation of technology parameters. In the following sections, we discuss the empirical issues raised by the estimation of the fixed cost, including suitable functional forms for cost functions, and the treatment of cost heterogeneity with unobserved levels of x_f.

5. On flexible functional forms

In the 1970's and 1980's, several researchers proposed new parametric specifications for the production technology, and introduced so-called flexible functional forms, which are able to approximate locally an arbitrary cost function. These functional forms, still widely used in production analysis, are not adequate for modelling fixed costs: either they completely exclude fixed costs, or specify them in an inflexible way. The variable t is now introduced for denoting technical change.

In their seminal paper, Diewert and Wales (1987) have introduced several cost functions, many of which can be written as

$$C^{DW}(w,y,t) = a_w^\top w + \left(\alpha_w^\top w\right) a_t t + V^{DW}(w,y,t), \tag{18}$$

with $V^{DW}(w,0,t) = 0$. This identifies the fixed cost as $U^{DW}(w,t) = a_w^\top w + \left(\alpha_w^\top w\right) a_t t$, where a_w, α_w, a_t denote technological parameters. So, the fixed cost function is linear in w and t and is not a flexible specification (in the sense of Diewert and 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 given by β:

$$\begin{aligned} C^{TL}(w,y,t) \;=\; & \exp(\beta_0 + \beta_w^\top \ln w + \beta_y \ln y + \beta_t t \\ & + \frac{1}{2} \ln w^\top B_{ww} \ln w + \ln w^\top B_{wy} \ln y + \ln w^\top B_{wt} t \\ & + \frac{1}{2}\beta_{yy} (\ln y)^2 + \beta_{yt} t \ln y + \frac{1}{2}\beta_{tt} t^2), \end{aligned} \tag{19}$$

where the notation is as in Koebel et al. (2003). One of the main drawbacks of the Translog functional form is that it is not suitable for modelling fixed cost.

Proposition 4. 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, and especially when production is close to zero, which defines the fixed cost of production. This proposition illustrates that the Translog is only able to approximate locally an unknown cost function, but not globally, and justifies the specification of alternative functional forms for the purpose of estimating a fixed cost. Proposition 4 points out a paradox: although the Translog specification is flexible (Diewert and Wales, 1987,

Theorem 1), it excludes fixed costs. The reason for this apparent contradiction is to be found in the limitations of the flexibility requirement, which just requires that the cost function be a local approximation, in some neighborhood of y, but not necessarily at the neighborhood of $y = 0$ which defines the fixed cost. In the sequel we rely on a functional form which is flexible at two points.

Definition 4. A *two-points 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 cost, can be represented by *two* different production technologies: one for initiating production $H\left(x_f\right) \equiv G\left(0, x_f\right)$ (using only fixed inputs), and one for reaching the output level y, and given by $G\left(x_v, x_f\right)$. So it becomes quite natural to specify both technologies in a flexible way. Similarly, the cost function is additively separable in two parts: one part u corresponding to the cost at zero output level and one part, v, reflecting the production cost of the output. So if our objective is to provide an approximation of the production technology, both parts should be treated with equal importance, and we suggest here to use a flexible functional form for both the fixed and variable cost functions. Definition 4 implies that a 2FFF cost function is the sum of two 1FFF fixed and variable cost functions U and V.

Diewert and Wales (1987, p.45-46) define a one point (1FFF) flexible cost function at the point $\left(w^0, y^0, t^0\right)$ as one being 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":

$$\begin{aligned} C\left(w^0, y^0, t^0\right) &= C^0\left(w^0, y^0, t^0\right) \\ \nabla C\left(w^0, y^0, t^0\right) &= \nabla C^0\left(w^0, y^0, t^0\right) \\ \nabla^2 C\left(w^0, y^0, t^0\right) &= \nabla^2 C^0\left(w^0, y^0, t^0\right), \end{aligned} \tag{20}$$

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 $J\left(J + 1\right)/2 + 2J + 3$ free equations. The requirements (20) have to be fulfilled at a single point y^0 which can be chosen to be

positive, so the 1FFF definition is compatible with the absence of fixed cost. This explains why the Translog is flexible although $U \equiv 0$. This drawback of 1FFF explains why we consider 2FFF.

A 2FFF for a cost function has enough free parameters for satisfying the following $1 + (J+1) + (J+1)^2 + 1 + (J+2) + (J+2)^2$ equations:

$$\begin{aligned} U\left(w^0, t^0\right) &= U^0\left(w^0, t^0\right), \\ \nabla U\left(w^0, t^0\right) &= \nabla U^0\left(w^0, t^0\right), \\ \nabla^2 U\left(w^0, t^0\right) &= \nabla^2 U^0\left(w^0, t^0\right), \end{aligned} \tag{21}$$

and for $y^0 > 0$,

$$\begin{aligned} V\left(w^0, y^0, t^0\right) &= V^0\left(w^0, y^0, t^0\right), \\ \nabla V\left(w^0, y^0, t^0\right) &= \nabla V^0\left(w^0, y^0, t^0\right), \\ \nabla^2 V\left(w^0, y^0, t^0\right) &= \nabla^2 V^0\left(w^0, y^0, t^0\right). \end{aligned} \tag{22}$$

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:

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

It turns out 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

$$V\left(w^0, 0, t^0\right) = 0.$$

Note that (21) and (22) imply (20), but not conversely.

6. Econometric treatment of cost heterogeneity

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

$$c_r\left(w, x_f, y, t\right) = u_r\left(w, x_f, t\right) + v_r\left(w, x_f, y, t\right),$$

which is somewhat embarrassing as we do not observe the level of x_f, but only total input quantity x. However, our objective is not to estimate firm specific functions v_r and u_r but rather their conditional mean given the value of the observed explanatory variables w, y and t, so we consider:

$$\begin{aligned} V(w,y,t) &\equiv \mathrm{E}\left[v_r\left(w,x_f,y,t\right)|w,y,t\right], \\ U(w,t) &\equiv \mathrm{E}\left[u_r\left(w,x_f,t\right)|w,t\right]. \end{aligned}$$

Here integration is over unobserved heterogeneity with respect to the joint distribution of x_f and the individual cost functions v_r and u_r. Using these definitions, we rewrite the model as follow:

$$c_r\left(w,x_f,y,t\right) = \gamma^U\left(w,x_f,t\right)U(w,t) + \gamma^V\left(w,x_f,y,t\right)V(w,y,t), \tag{23}$$

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

$$\gamma^U\left(w,x_f,t\right) \equiv \frac{u_r\left(w,x_f,t\right)}{U(w,t)}, \qquad \gamma^V\left(w,x_f,y,t\right) \equiv \frac{v_r\left(w,x_f,y,t\right)}{V(w,y,t)},$$

and satisfy $\mathrm{E}\left[\gamma^U|w,t\right] = \mathrm{E}\left[\gamma^V|w,y,t\right] = 1$. 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 5. Under the assumptions that, (a) individual heterogeneity in the fixed and variable cost functions is independent of x_f; (b) the fixed inputs x_f are positive and are optimally allocated; then:

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

(ii) the conditional variance matrix $\mathrm{V}\left[\gamma|w,y,t\right]$ 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 $\mathrm{V}\left[\gamma|w,y,t\right]$ and $\mathrm{cov}\left[\gamma^U,\gamma^V|w,y,t\right]$. If the statistical test leads to rejection of the singularity of $\mathrm{V}\left[\gamma|w,y,t\right]$ or $\mathrm{cov}\left[\gamma^U,\gamma^V|w,y,t\right] \leq 0$, then we can deduce that either the fixed inputs are not optimally allocated (Proposition 5), or that the production technology has the specific structure given in (16). The level of the fixed cost $\gamma^U U$ and the level of the variable cost $\gamma^V V$ are certainly positively correlated with any dataset: both the fixed and the variable cost increase over time, and firms with a high fixed cost certainly produce more than smaller firms and also have a higher variable cost. Hence

the positive correlation between $\gamma^U U$ and $\gamma^V V$. Proposition 5, however, states that there is a tradeoff – a negative correlation – between the fixed and the variable cost *for given values* of the explanatory variables (w, y, t). Such a tradeoff cannot be directly observed in a dataset, because it pertains to unobserved heterogeneity.With panel data, the issue of interrelated heterogeneity is often discarded, one exception is Gladden and Taber (2009) who considered it in estimating linear wage equations. In contrast to Gladden and Taber (2009), we derive the sign of the covariance from a structural nonlinear model.

Let us now explain our strategy for estimating this covariance along with other statistics of interest. We have to explicitly introduce the parameters in the notations of the cost function and rewrite the observed cost level c_{nt} as follow:

$$c_{nt} = \gamma_{nt}^U U\left(w_{nt}, t; \alpha\right) + \gamma_{nt}^V V\left(w_{nt}, y_{nt}, t; \beta\right) + e_{nt}, \tag{24}$$

where $n = 1, \ldots, N$ denotes the sector, $t = 1, \ldots, T$ represents time. The random term e_{nt} is iid, satisfies $\mathrm{E}\left[e_{nt} | w_{nt}, y_{nt}, t\right] = 0$ and has constant variance σ_c^2. It is also assumed that e_{nt} is uncorrelated with $\gamma_{nt} \equiv \left(\gamma_{nt}^U, \gamma_{nt}^V\right)^\top$ and any right hand side regressors. Equivalently, we can write our empirical model as:

$$c_{nt} = U\left(w_{nt}, t; \alpha\right) + V\left(w_{nt}, y_{nt}, t; \beta\right) + \varepsilon_{nt}^c. \tag{25}$$

with the composite error term:

$$\varepsilon_{nt}^c \equiv \left(\gamma_{nt}^U - 1\right) U\left(w_{nt}, t; \alpha\right) + \left(\gamma_{nt}^V - 1\right) V\left(w_{nt}, y_{nt}, t; \beta\right) + e_{nt}. \tag{26}$$

Note that $\mathrm{E}\left[\varepsilon_{nt}^c | w_{nt}, y_{nt}, t\right] = 0$. We also assume that

$$\mathrm{V}\left[\gamma_{nt} | w, y, t\right] \equiv \Sigma = \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix}, \tag{27}$$

and $\mathrm{V}\left[\gamma_{nt}\gamma_{ms}^\top | w, y, t\right] = 0$, for any $n \neq m$ and $t \neq s$. This model is an extension of Swamy's (1970) random coefficient model to our nonlinear setup with individual and time varying random coefficients. The values of γ_{nt} can be considered as incidental parameters, because they are not fundamentally interesting (and cannot be identified). Their distribution however is informative. The joint distribution of γ_{nt} reflects the way the variable and fixed cost vary together. The covariance between γ_{nt}^U and γ_{nt}^V allows to discriminate between the case of optimally and nonoptimally allocated fixed input and whether fixed cost has an impact on the marginal cost of production and the markup via (17). The parameters of interest are the technology parameters $\theta \equiv \left(\alpha^\top, \beta^\top\right)^\top$ and

the variance matrix Σ.

In principle, all estimates of the technology parameters θ and the covariance matrix can be obtained simultaneously by solving (numerically) the likelihood maximization or the nonlinear least squares problem.[9] However, these objective functions are highly nonlinear in θ, and it turns out that nonlinear numerical algorithms often do not converge to a solution. We avoid these numerical problems, and use a two-stage estimation procedure. First, the technological parameters θ are consistently estimated (without identification of Σ and σ_c^2) by minimizing the sum of squared residuals:

$$\widehat{\theta} = \arg\min_{\alpha,\beta} \sum_{n,t} \left[c_{nt} - U\left(w_{nt}, t; \alpha\right) - V\left(w_{nt}, y_{nt}, t; \beta\right)\right]^2.$$

As the random term ε_{nt}^c exhibits heteroscedasticity and serial correlation, we rely on the Newey-West (1987) estimator for estimating the variance matrix $\mathrm{V}\left[\widehat{\theta}\right]$.

In the second-stage, two equivalent estimation methods are again available: Maximum Likelihood (ML) and Least Squares (LS). The conditional variance of $\hat{\varepsilon}_{nt}^c$ can be expressed as (using (26)):

$$\begin{aligned} & \mathrm{E}\left[(\hat{\varepsilon}_{nt}^c)^2 \,|w_{nt}, y_{nt}, t\right] \equiv \triangle_{nt}(\sigma_c^2, \Sigma, \widehat{\theta}) \qquad (28) \\ = \;& \sigma_c^2 + \sigma_U^2 U^2\left(w_{nt}, t; \widehat{\alpha}\right) + \sigma_V^2 V^2\left(w_{nt}, y_{nt}, t; \widehat{\beta}\right) + 2\sigma_{UV} U\left(w_{nt}, t; \widehat{\alpha}\right) V\left(w_{nt}, y_{nt}, t; \widehat{\beta}\right). \end{aligned}$$

It turns out that the parameters σ_c^2, σ_U^2, σ_V^2 and σ_{UV} of (28) can be estimated by an OLS regression of the squared NLS residuals

$$(\widehat{\varepsilon}_{nt}^c)^2 = \left[c_{nt} - U\left(w_{nt}, t; \widehat{\alpha}\right) - V\left(w_{nt}, y_{nt}, t; \widehat{\beta}\right)\right]^2 \qquad (29)$$

on a constant, $\widehat{U}^2$, $\widehat{V}^2$ and $\widehat{U}\widehat{V}$. If we assume that the heterogeneity vector γ_{nt} and the error term e_{nt} follow some parametric distribution, then the estimated covariance matrix can be obtained by maximizing the likelihood function. Both second-stage estimation methods are asymptotically equivalent, but their estimation outcomes may differ: first, because the ML is more efficient than OLS if the distribution of the random terms is

[9] The NLS estimator of $\left(\theta, \sigma_c^2, \Sigma\right)$ could be obtained (in one step) by minimizing the following sum of squared residuals:

$$\sum_{n,t} \left[\varepsilon_{nt}^{c2}\left(\theta\right) - \sigma^2 - \sigma_U^2 U^2\left(w_{nt}, t; \alpha\right) - \sigma_V^2 V^2\left(w_{nt}, y_{nt}, t; \beta\right) - 2\sigma_{UV} U\left(w_{nt}, t; \alpha\right) V\left(w_{nt}, y_{nt}, t; \beta\right)\right]^2.$$

An alternative estimator of parameters θ, σ_c^2 and Σ is the maximum likelihood estimator. Under the normality assumption of the random term $\varepsilon_{nt}^c \sim N\left(0, \triangle_{nt}\left(\theta, \sigma_c^2, \Sigma\right)\right)$, we can write

$$\log L(\theta, \sigma_c^2, \Sigma) = -\frac{1}{2} \sum_{n,t} \left\{\log\left(2\pi\right) + \log \triangle_{nt}\left(\theta, \sigma_c^2, \Sigma\right) + \triangle_{nt}\left(\theta, \sigma_c^2, \Sigma\right)^{-1} \left(\varepsilon_{nt}^c\left(\theta\right)\right)^2\right\}.$$

well specified; second, because the covariance matrix Σ is not restricted to be positive-definite in the OLS regression, but this restriction is imposed in most ML estimation algorithms.[10] As this matrix may well be singular (Proposition 5), we prefer the OLS approach.

Our estimation approach can be viewed as a sequential two-stage M-estimation, where in the first-stage $\widehat{\theta}$ is obtained by solving a NLS problem and then, given $\widehat{\theta}$, the estimates $\widehat{\sigma}^2, \widehat{\Sigma}$ are obtained by OLS. This second stage estimator is simple and consistent if the first-stage estimator is consistent for θ, see Cameron and Trivedi (2005, Section 6.6). However, the asymptotic distribution of $\hat{\Sigma}$ given the estimation of $\widehat{\theta}$ is difficult to establish. Hence we use the panel bootstrap for deriving the standard deviations of the second-stage estimator.[11]

7. The empirical investigation

In this section, we first summarize the empirical models and strategies, we then present briefly the data set and discuss the estimation results.

7.1 The empirical models and estimation strategies

For the empirical fixed and variable cost functions U and V, we assume Translog functional forms denoted by U^{TL} and V^{TL}. As seen in Proposition 4, the traditional Translog cost function C^{TL} satisfies $C^{TL}(w, 0, t) = 0$ and is not compatible with the occurrence of a fixed cost (in the best case where $\beta_{yy} \leq 0$). It is, however, quite simple to generalize the Translog specification by adding a fixed cost function to the variable Translog cost function (the two-points flexible form):

$$C^{TL}(w, y, t; \alpha, \beta) = U^{TL}(w, t; \alpha) + V^{TL}(w, y, t; \beta),$$

where

$$U^{TL}(w, t; \alpha) = \exp\{\alpha_0 + \alpha_w^\top \ln w + \alpha_t t + \frac{1}{2} \ln w^\top A_{ww} \ln w + \ln w^\top A_{wt} t + \frac{1}{2}\alpha_{tt} t^2\}, \qquad (30)$$

[10] It can be shown that the first order conditions of ML are identical to the moment conditions of OLS.

[11] We assume that the errors are i.i.d. over individuals (but not over time). The panel bootstrap performs a classical paired bootstrap that resamples only over n and not over t.

and

$$V^{TL}(w,y,t;\beta) = \exp\{\beta_0 + \beta_w^\top \ln w + \beta_y \ln y + \beta_t t + \frac{1}{2}\ln w^\top B_{ww} \ln w + \ln w^\top B_{wy} \ln y + \ln w^\top B_{wt} t + \frac{1}{2}\beta_{yy}(\ln y)^2 + \beta_{yt} t \ln y + \frac{1}{2}\beta_{tt} t^2\}.$$

We impose linear homogeneity and symmetry in w using the following $2 + J + (J+1)J/2$ parametric restrictions on U^{TL}:

$$\iota^\top \alpha_w = 1, \quad \iota^\top A_{wt} = 0, \quad \iota^\top A_{ww} = 0, \quad A_{ww} = A_{ww}^\top. \tag{31}$$

There are $1 + J + (J+1)J/2$ free parameters left in U^{TL}. Similarly, the variable cost function V^{TL} has $3 + 2J + (J+1)J/2$ free parameters which satisfy

$$\iota^\top \beta_w = 1, \quad \iota^\top B_{wt} = \iota^\top B_{wy} = 0, \quad \iota^\top B_{ww} = 0, \quad B_{ww} = B_{ww}^\top. \tag{32}$$

Note that the logarithmic transformation of the total cost function is not useful anymore for linearizing the nonlinear Translog specification (unless $U^{TL} \equiv 0$). For $J = 4$, the fixed cost function has 15 free parameters to which are added the 21 free parameters of the variable cost function.

Given the two-points flexible specification, we estimate the parameters α and β by using NLS based on (25) in the first-stage. The second-stage consists in the estimation of the variance matrix Σ and σ_c^2 by using OLS based on (28) and (29). The classical Translog cost function which includes only the variable cost function V^{TL} (and assumes that $U^{TL} \equiv 0$) is also considered for comparison. We consider further empirical models that include the system estimation by adding the input demand equations (obtained by applying Shephard's lemma to C^{TL}), as well as the model estimated in first-differences. Substantial gains in efficiency can be realized by system estimation, because more observations are available. The first-difference estimation model is more robust against non-stationarity of the series and unobserved individual fixed effects. Henceforth, Model I denotes the single equation model without any fixed cost. Model II is the baseline model where the cost function includes both a fixed and a variable part (the two-points flexible form). More efficient frameworks are Model III (in level) and Model IV (in difference), which include the cost and the input demand functions. We note that the choice of starting values is crucial for reaching the optimum in the case of

system NLS.[12]

7.2 The data and empirical results

We use the NBER-CES manufacturing industry database for our empirical study.[13] This database records annual information on output y_{nt}, output price p_{nt}, and the input levels x_{nt}, together with input prices indices w_{nt}, for 462 U.S. manufacturing industries (at the six-digit NAICS aggregation level) and covers the period 1958 to 2005. See Chen (2012) for descriptive statistics and details on the computations made for generating the depreciation rate, interest rate, and the user cost of capital. Information is available for four inputs: capital, labor, energy and intermediate materials.

We begin by commenting the first-stage estimation results for models I to IV. Instead of reporting estimates for all Translog parameters, we only select some informative estimated coefficients and statistics. An important coefficient is the parameter β_{yy}, which is crucial for Proposition 5. Given the estimated Translog coefficients, we compute statistics such as the share of the fixed cost in the total cost U/C, the ratio of the output price to the predicted marginal cost of production $p/\left(\partial C/\partial y\right)$ which measures the markup, and the rate of returns to scale $1/\varepsilon\left(C,y\right)$, where $\varepsilon\left(C,y\right) \equiv \partial \ln C/\partial \ln y$ denotes the elasticity of costs with respect to output.

As mentioned in Section 5, neglecting the fixed cost is a source of bias. By comparing the estimation outcomes of Model I and Model II, we note that the results of the two models differ with respect to several key points. First, the parameters of the fixed cost function (α) in Model II are significantly different from zero, which indicates the existence of fixed costs in the production process. Second, the model without a fixed cost (Model I) suggests that the industries exhibit decreasing returns to scale, but the model with a fixed cost (Model II) suggests increasing returns to scale. The bias on the degree of returns to scale is due to the overestimation of the elasticity of cost and neglect of the fixed cost (see Section 4). Finally, the overestimation of marginal costs by Model I leads to underestimation of the markup: the median of $p/\left(\partial C/\partial y\right)$ in Model

[12] For the single equation estimation (Model I and II), the starting values are set arbitrarily to zero. For the system estimation in levels (Model III), the starting values are the estimates obtained from Model II. For the system estimation in first differences (Model IV), the starting values are obtained from the estimation of the cost function in first-differences.
[13] The dataset can be downloaded at: http://www.nber.org/data/nbprod2005.html

Table 1. Summary of estimation results

	Model	I	II	III	IV
U/C	1st q	-	0.19	0.27	0.48
	median	-	0.52	0.51	0.76
	3rd q	-	0.91	0.77	0.94
$p/\left(\partial C/\partial y\right)$	1st q	1.19	1.32	1.37	1.58
	median	1.36	1.87	1.78	2.63
	3rd q	2.47	6.60	2.74	5.73
$1/\varepsilon\left(C,y\right)$	1st q	0.69	1.01	1.01	1.10
	median	0.89	1.40	1.37	2.07
	3rd q	0.98	6.04	2.71	7.07
β_{yy}	coeff	-0.05	-0.14	-0.15	-0.32
	t-value	-2.05	-4.21	-3.11	-3.85
σ_U^2	coeff	-	0.54	30.83	1.09
	t-value	-	1.67	1.29	0.15
σ_V^2	coeff	-	0.02	0.27	0.12
	t-value	-	1.95	1.96	1.27
σ_{UV}	coeff	-	-0.32	-13.04	-1.19
	t-value	-	-0.91	-1.43	-0.11
σ_c^2	coeff	3.7e+6	1.9e+6	1.3e+6	2.1e+6
	t-value	-	2.09	0.24	0.66

Notes: Rows 2 to 11 report the estimated parameter values and the corresponding t-statistic for the hypothesis that the parameter is equal to zero. Rows 12 to 20 report the median value of the corresponding statistic over all observations as well as the 1st and 3rd quartiles.

I is about 36% lower than the one predicted by Model II.[14]

Table 1 also shows that empirical results obtained from models II to IV exhibit some regularities. First, the estimated coefficient of β_{yy} is significantly negative in all cases, which implies that the limit of the classical Translog variable cost function is zero as y approaches 0. Second, all models predict that the fixed cost represents a considerable share of total cost. The median of estimated shares U/C varies in the range between 51% and 76%. Third, the estimation results also suggest that the industries exhibit increasing returns to scale. The median of the rate of returns to scale, $\varepsilon\left(C,y\right)^{-1}$ ranges between 1.4 and 2.1. Fourth, there is a significant difference between the selling price and the predicted marginal cost of production, the median of estimated markup varies from 1.8 to 2.6. However, we note that the results of Model IV (with data in first-differences) differ quantitatively from those of Model II and Model III (with data expressed in levels).

[14] We also reestimate Model I after appending a linear fixed cost term $w^{\top}\widetilde{\alpha}$ in the specification. The corresponding empirical results are not reported, but lie in between those obtained for Model I and II.

Now, we focus on the fixed cost share (U/C), in particular on its evolution over time. These series (averaged over all industries) are depicted on Figure 5. We note that for all the empirical models, the fixed cost shares are decreasing over time. This may reflect firms' efforts to increase production flexibility. The series generated by model II (where the input demands system is not included in the estimation), exhibit a structural break around 1980. For other models, the decline of fixed cost shares over time is smoother. However, the decrease is less significant in the first-differenced model (Model IV).

When it comes to the second-stage estimation, the estimates of σ_U^2, σ_V^2, σ_{UV} and σ_c^2, are somewhat more divergent across the models. However, we see that the variance of the fixed cost heterogeneity γ^U is always larger than the variance of the variable cost heterogeneity γ^V. The covariance between heterogeneities is found to be negative and the covariance matrix Σ is close to singular for all models, which is conform to what we expect from Proposition 6. The second-stage estimation results, however, are not precisely estimated and are not statistically significant. This result may be due to our overly restrictive assumption of random heterogeneity in the fixed and variable cost function specification (23). Economically, this heterogeneity may well be correlated with further explanatory variables which are individual specific (like for instance the level of production, the type of industry, etc.). So we conduct further analysis in the next subsection.

Figure 5. Fixed costs shares over time

7.3 Estimation with industry specific dummies

Although Models II to IV with random heterogeneity yield some interesting results on the scope of fixed cost and returns to scale, the interaction between fixed and variable costs was not precisely estimated. This may due to the fact that heterogeneity is not purely random but correlated with sectorial characteristics as the level of production or the level of fixed and variable cost. We pursue the investigation a step further and introduce individual-specific dummies into Model IV. The most flexible specification replaces γ_{nt}^U and γ_{nt}^V in regression (24) by $2N$ individual-specific parameters. In order to limit overparameterization, we introduce instead dummies for more broadly defined groups of industries. There are different ways to define these groups, for instance, in the spirit of Mundlak's (1978) correlated random coefficient model, individuals can be grouped w.r.t. the average value of their covariates. For the industry database, however, a more natural clustering criterion, is to group the 462 manufacturing sectors available at the six-digit NAICS level into 20 three-digit NAICS sectors. See Table 2 for a list of

the 3-digit industries.[15]

Formally, we introduce the multiplicative dummy variables γ_j^U and γ_j^V for $j = 1, ..., 20$ in place of the random parameters of (24) which becomes:

$$c_{nt} = \gamma_j^U U^{TL}(w_{nt}, t; \alpha) + \gamma_j^V V^{TL}(w_{nt}, y_{nt}t; \beta) + e_{nt}. \tag{33}$$

Since the Translog cost function also includes the terms α_0 and β_0, all the parameters cannot be identified separately, unless we consider two additional restrictions. Since by construction, we have $\mathrm{E}\left[\gamma^U|w,t\right] = \mathrm{E}\left[\gamma^V|w,y,t\right] = 1$, it is natural to impose the normalization conditions:

$$\frac{1}{20}\sum_{j=1}^{20}\gamma_j^U = \frac{1}{20}\sum_{j=1}^{20}\gamma_j^V = 1,$$

which allow to identify all parameters. In this case, the estimated parameters γ_j^U and γ_j^V represent the industry-specific deviation in percentage from the average. For instance, if the estimated value of γ_j^U is significantly above one and the estimated value of γ_j^V is significantly below one, this indicates that the industry group j incurs more fixed and less variable costs than average. In this framework, the interaction between the fixed and variable components of the cost function is characterized by the variation of γ_j^U and γ_j^V over industry groups. We examine the empirical correlation between γ_j^U and γ_j^V along with group-specific shares of fixed cost, degree of returns to scale, markups and rate of technical change.

We estimate the parameters of the extended Model IV and report the estimation results in Table 2. Column 3 and 4 of Table 2 report the estimated coefficients of γ_j^U and γ_j^V. Our estimation results indicate, for instance, that compared to the average, the industry group NAICS 311 (food) operates with 24% less fixed cost and 4% less variable cost than the average. We also note that industries with lower than average fixed cost generally have higher than average variable cost and conversely. Contrary to the above random effect models, the parameters reflecting cost heterogeneity are now statistically significant.

Columns 5 to 10 report the median (for each group) of the fixed cost share U/C, the markup $p/\left(\partial C/\partial y\right)$, returns to scale $1/\varepsilon\left(C,y\right)$, and technical change measured as

[15] At the three-digit NAICS level, there are actually 21 manufacturing industry groups. We merge the smallest (in terms of the number of subsectors) NAICS 324 industry group (petroleum and coal products manufacturing) with NAICS 325 industry group (chemical manufacturing).

$\partial \ln C/\partial t$, $\partial \ln U/\partial t$ and $\partial \ln V/\partial t$. For the NAICS 311 industry group, the estimates indicate that the fixed cost represents 25% of the total production cost, with almost constant returns to scale and a markup of 68%. In average over all industries, the results confirm former findings with strong evidence for fixed cost, increasing returns to scale, and markup pricing. We also find evidence for the conjecture brought forward in Section 4: industries with higher fixed cost also exhibit higher markups and returns to scale.

Regarding the rate of technical change, our results on $\partial \ln V/\partial t$ show that the variable cost is on average decreasing by 0.9% over time with little variance over industries. In contrast, the fixed cost increases with time i.e. $\partial \ln U/\partial t = 0.04$. Altogether, our results are in line with those obtained by Diewert and Fox (2008) who 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 (the x_f). These fixed inputs contribute to increase the fixed cost and decrease the variable cost and, as a consequence of our approach, this random component of technical change is captured by the negative correlation between γ_j^U and γ_j^V.

Table 2. Summary of estimation results with industry dummies

NAICS	industry groups	γ_j^U	γ_j^V	$\gamma_j^U U/C$	$p/\left(\partial C/\partial y\right)$	$1/\varepsilon(C,y)$	$\partial \ln C/\partial t$	$\partial \ln U/\partial t$	$\partial \ln V/\partial t$
311	Food	0.76 (7.36)	0.96 (26.40)	0.25	1.68	1.05	0.004	0.039	-0.009
312	Beve.&Toba.	1.22 (6.03)	0.37 (6.22)	0.62	4.42	2.07	0.020	0.037	-0.009
313	Textile	0.58 (5.02)	1.35 (18.15)	0.34	1.57	1.11	0.008	0.041	-0.009
314	Textile Prod.	0.51 (7.17)	1.10 (17.52)	0.41	2.14	1.19	0.015	0.042	-0.009
315	Apparel	0.50 (6.55)	1.51 (18.64)	0.28	1.57	1.01	0.008	0.044	-0.009
316	Leather	0.16 (3.76)	1.39 (11.83)	0.27	2.18	0.97	0.005	0.042	-0.010
321	Wood	1.00 (10.22)	1.01 (24.23)	0.41	1.74	1.29	0.009	0.038	-0.009
322	Paper	1.56 (10.03)	0.81 (16.49)	0.61	2.15	1.93	0.019	0.036	-0.010
323	Printing	0.70 (3.16)	1.15 (8.52)	0.31	1.50	1.13	0.005	0.038	-0.009
324-5	Petr.&Chem.	0.46 (2.91)	1.12 (24.15)	0.17	1.35	0.96	-0.001	0.035	-0.009
326	Plastic	1.44 (8.86)	1.03 (17.43)	0.52	1.74	1.62	0.012	0.037	-0.009
327	Mineral Prod.	0.64 (9.94)	1.07 (26.75)	0.43	1.83	1.31	0.011	0.036	-0.010
331	Primary Metal	1.08 (6.91)	0.87 (11.44)	0.47	1.85	1.46	0.011	0.036	-0.010
332	Fabricated Metal	0.72 (8.73)	1.13 (28.03)	0.32	1.47	1.13	0.006	0.037	-0.009
333	Machinery	0.81 (14.34)	0.85 (26.75)	0.42	1.88	1.33	0.010	0.036	-0.010
334	Computer	3.96 (9.62)	0.20 (4.10)	0.94	8.81	13.47	0.037	0.042	-0.009
335	Elec.Equipement	0.58 (7.36)	1.05 (32.20)	0.30	1.70	1.10	0.004	0.038	-0.009
336	Transportation	2.37 (10.41)	0.77 (26.46)	0.49	1.84	1.61	0.011	0.035	-0.009
337	Furniture	0.44 (5.06)	1.29 (22.58)	0.28	1.50	1.04	0.004	0.040	-0.009
338	Miscellaneous	0.50 (7.29)	0.97 (15.30)	0.45	2.16	1.32	0.014	0.040	-0.009
	Average	1.00	1.00	0.42	2.25	1.91	0.011	0.039	-0.009

Note: t-values are reported in parenthesis for $H_0 : \gamma_j^U = 0$ and for $H_0 : \gamma_j^V = 0$.

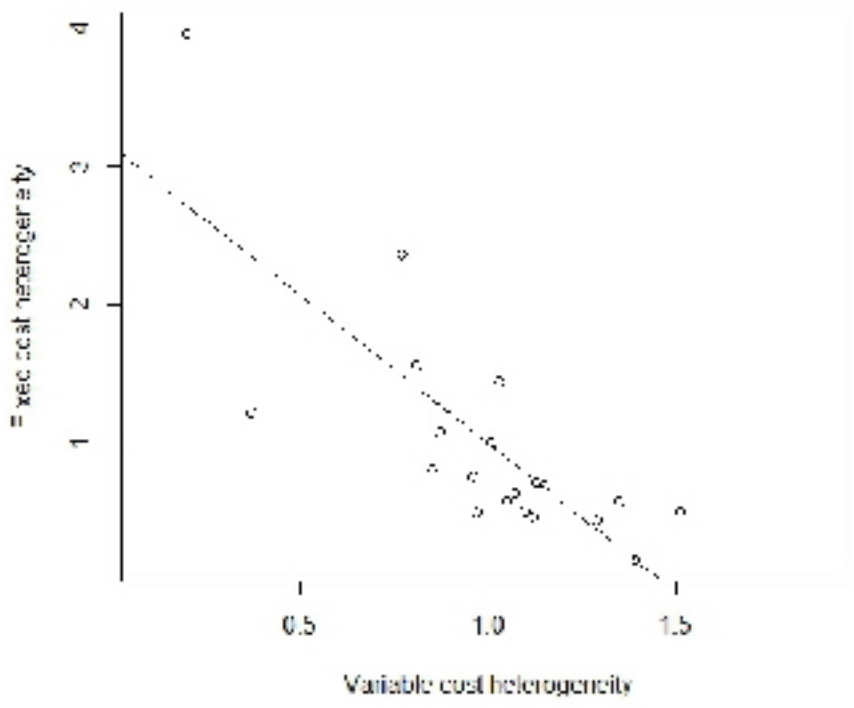


Figure 6. Scatterplot of $\widehat{\gamma}_j^U$ and $\widehat{\gamma}_j^V$.

Table 3 reports the empirical correlations between different estimated statistics. The main result is that the correlation between $\widehat{\gamma}_j^U$ and $\widehat{\gamma}_j^V$ is negative, and quite strong (-0.79). The scatterplot of $\widehat{\gamma}_j^U$ and $\widehat{\gamma}_j^V$ is depicted on Figure 6. These results are in line with Proposition 5. The extension of Model IV to include industry-specific fixed and variable cost heterogeneity now allows us to find more precise empirical results than those obtained with random heterogeneity. The separable structure of Proposition 3, (15), which implies no interaction between fixed and variable cost, is statistically rejected: technology $G\left(x_v, x_f\right)$ fits the data better than $F\left(x_v + K\left(x_f\right)\right)$ for any function K.

We also find that the fixed-cost heterogeneity is positively correlated with most of the statistics especially with the markup and the rate of returns. This coincides with our discussion of Section 4 on the dangers of neglecting fixed cost. Not surprisingly, the correlations involving γ_j^V have the opposite sign to those involving γ_j^U. The strong positive correlation between $\gamma_j^U U/C$ and $p/\left(\partial C/\partial y\right)$ seems to be contrary to the prediction made by the theory of contestable markets (Baumol, Panzar and Willig, 1982).

Table 3. Correlation matrix

	γ_j^U	γ_j^V	$\gamma_j^U \frac{U}{C}$	$\frac{p}{\partial C/\partial y}$	$\frac{1}{\varepsilon(C,y)}$	$\frac{\partial \ln C}{\partial t}$	$\frac{\partial \ln U}{\partial t}$	$\frac{\partial \ln V}{\partial t}$	$\overline{y}_j$	cr_j
γ_j^V	-0.79									
$\gamma_j^U \frac{U}{C}$	0.86	-0.84								
$\frac{p}{\partial C/\partial y}$	0.80	-0.77	0.83							
$\frac{1}{\varepsilon(C,y)}$	0.86	-0.67	0.79	0.95						
$\frac{\partial \ln C}{\partial t}$	0.81	-0.78	0.97	0.88	0.83					
$\frac{\partial \ln U}{\partial t}$	-0.07	0.31	0.01	0.28	0.26	0.20				
$\frac{\partial \ln V}{\partial t}$	0.22	0.02	-0.03	0.26	0.29	0.09	0.56			
$\overline{y}_j$	0.58	-0.58	0.29	0.27	0.28	0.18	-0.55	0.22		
cr_j	0.23	-0.39	0.20	0.23	0.07	0.20	-0.16	0.14	0.58	
H_j	0.24	-0.45	0.25	0.21	0.01	0.24	-0.20	0.17	0.61	0.92

However, it can be explained in the light of our framework: a higher fixed cost reduces the variable cost (at given level of production), a relationship which is reflected by the negative correlation between γ_j^U and γ_j^V. This negative correlation is in turn inherited by $\gamma_j^U U/C$ and $\partial C/\partial y$.

These results help 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 attribute neglected variations in fixed costs (which according to Table 3 are positively correlated with output) to the variable cost function which is increasing in y. Like 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 so, it will be biased upwards. The positive correlation $corr\left(\gamma_j^U; \overline{y}_j\right) = 0.58$ explains the gap between the results obtained with the standard and extended Translog specifications (see Table 1). In Model I, the neglected fixed cost is directly responsible for the low rate of returns to scale and moderate markups obtained with this specification.

Regarding technical change, we find that $\partial \ln C/\partial t$ is positive and highly correlated with γ_j^U, γ_j^V, $\gamma_j^U U/C$ and $p/\left(\partial C/\partial y\right)$, which means that fixed cost and market power preclude productivity growth (as in Arrow, 1962). Surprisingly, neither $\partial \ln U/\partial t$ nor $\partial \ln V/\partial t$ are strongly correlated with market power. This paradox is solved if we go back to the definition of technical change, in which the share of fixed cost plays an important role:

$$\frac{\partial \ln C_j}{\partial t} = \frac{\partial \ln U}{\partial t}\frac{\gamma_j^U U}{C_j} + \frac{\partial \ln V}{\partial t}\left(1 - \frac{\gamma_j^U U}{C}\right),$$

and introduces correlation between $\partial \ln C_j/\partial t$ and γ_j^U and $\gamma_j^U U/C_j$. We also investigate the link between the fixed cost, the size and the concentration of industries. Table 3 also reports correlations between the fixed cost and the average output level (over time and subsectors within industry j), the concentration ratio for the 20 largest firms cr_j, and the Hirschman-Herfindahl index H_j.[16] We find a positive correlation between the fixed cost share and the industrial concentration. These results suggest that industries with a higher fixed cost and a lower variable cost, produce more in average, and are

[16] The concentration data for 2002 are obtained from the U.S. Census Bureau.

more concentrated.

8. Conclusion

This paper investigates technologies in which fixed inputs can be imperfectly substituted to variable inputs, and we propose extended production and cost functions compatible with the occurrence of a fixed cost. Many 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 the fixed and the variable cost. Our empirical findings highlight the importance of fixed cost which represent about 20% to 60% of total cost in the manufacturing industries and tend to decline to decline over time. Our estimates also supports our extended framework which explains why industries with higher fixed cost, in average have lower variable cost, higher returns to scale and markups. Conformably to our theoretical prediction, we also find that the classical Translog cost function underestimates the rate of returns to scale and the markup.

A natural extension of our framework would be to examine explicitly strategic interactions between firms in their joint decision on product price and production capacity (fixed cost). This would potentially allow to revisit the link between fixed cost and barriers to entry.

9. Appendix: proof of the results

Proof of Proposition 1. From the definition of X_F and $X_F \neq \varnothing$ it directly follows that

$$c_r\left(w, x_f, 0\right) = \min_{x_v \geq 0} \left\{ w^\top x_v + w^\top x_f : F\left(x_v + x_f\right) \geq 0 \right\} = w^\top x_f \geq 0,$$

and so $v_r\left(w, x_f, 0\right) = c_r\left(w, x_f, 0\right) - w^\top x_f = 0$.

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

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

where $v_r\left(w, x_f, y\right) \equiv \min_{x > x_f} \left\{ w^\top x : F\left(x\right) \geq y \right\} - w^\top x_f > 0$. Then $c_r\left(w, x_f, y\right) = C\left(w, y\right)$ and by Shephard's lemma $x_v^*\left(w, x_f, y\right) = X_v^*\left(w, y\right)$.

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

$$x = x_v + x_f = \begin{pmatrix} \widetilde{x} \\ \overline{x} \end{pmatrix},$$

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

$$\begin{aligned} c_r\left(w, x_f, y\right) &= \min_{x_v \geq 0}\left\{w^\top x_v + w^\top x_f : F\left(x_v + x_f\right) \geq y\right\} \\ &= \min_{\widetilde{x}>0}\left\{\widetilde{w}^\top \widetilde{x} + \overline{w}^\top \overline{x} : F\left(\widetilde{x}, \overline{x}\right) \geq y\right\} \\ &= \min_{\widetilde{x}>0}\left\{\widetilde{w}^\top \widetilde{x} : F\left(\widetilde{x}, \overline{x}\right) \geq y\right\} + \overline{w}^\top \overline{x} = V_r\left(\widetilde{w}, \overline{x}, y\right) + \overline{w}^\top \overline{x}. \end{aligned}$$

□

Proof of Proposition 2.

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

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

The assumption that G is single valued and increasing implies that $G\left(x_v, x_f\right) > 0$ for and $x_v > 0$ and $x_f \in X_G$. Then $v_r\left(w, x_f, y\right) = w^\top x_v^*\left(w, x_f, y\right) > 0$ for $y > 0$ because $w > 0$ at least one element of $x_v^*\left(w, x_f, y\right)$ is strictly positive.

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

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

(iii) Similarly, $x_f' > x_f$ and G increasing in $\left(x_v, x_f\right)$, implies that $\left\{x_v : G\left(x_v, x_f\right) \geq y\right\} \subset \left\{x_v : G\left(x_v, x_f'\right) \geq y\right\}$ and as a consequence

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

□

Proof of Proposition 3.

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

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

The last line follows from our assumption that $x_v^*\left(w, y\right) > 0$ at the optimum. Defining

$v(w,y) \equiv v_y(w,y) - v_y\left(w,0^+\right)$ ensures that $v\left(w,0^+\right)=0$. Defining $u_r\left(w,x_f\right) \equiv v_y\left(w,0^+\right) - w^\top K\left(x_f\right) + w^\top x_f$ ensures that $c_r\left(w,x_f,y\right) = u_r\left(w,x_f\right) + v(w,y)$.

Conversely, we can recover the convex hull of all inputs producing y, for a given level of x_f, by solving

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

The corresponding J first order conditions for an inner solution are given by

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

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

$$y = F\left(x_v + K\left(x_f\right)\right).$$

If G is quasi-concave in x_v, this convex hull corresponds to the isoquants of G.

Part (ii). Necessity. With (15), the first order conditions for an inner solution in x_f to the cost minimization problem are given by

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

and do not depend on y and so the solutions $x_f^*(w)$. With (16), the first order conditions for an inner solution in x_v are

$$\begin{aligned} w &= \lambda \frac{\partial F}{\partial x_v}\left(x_v + K\left(x_f\right)\right) \\ y &= F\left(x_v + K\left(x_f\right)\right), \end{aligned}$$

where λ denotes the Lagrange multiplier. The solution in x_v to this system takes the form $x_v^*\left(w,x_f,y\right) = X^*(w,y) - K\left(x_f\right)$ and so the restricted cost function (15), with $v_y(w,y) \equiv w^\top X^*(w,y)$ and $u_r\left(w,x_f\right) = w^\top x_f - w^\top K\left(x_f\right)$. Then x_f^* is independent of y.

Sufficiency. If x_f^* depends only upon w, then the first order conditions for an inner solution, given by

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

imply that

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

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

Proof of Proposition 4. 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

$$\lim_{y\to 0^+} C^{TL}(w,y,t) = 0, \tag{34}$$

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

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

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

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

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. □

Proof of Proposition 5. 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 for denoting this heterogeneity. Let $f_{u|x}$ denote the conditional density function of $u_r(w,x_f,t)|x_f$. Under Assumption (a) 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

$$\begin{aligned}
\overline{u}(w,x_f,t) &\equiv \int u_r(w,x_f,t) f_u(r)\, dr \\
\overline{v}(w,x_f,y,t) &\equiv \int v_r(w,x_f,y,t) f_v(r)\, dr.
\end{aligned}$$

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

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

and (we skip the arguments for simplicity)

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

Using the optimality condition $\partial c_r/\partial x_f = 0$, and Assumption (a), it follows that $\partial \overline{c}/\partial x_f = 0$. So, conditionnaly on observations (w,y,t), we write

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

(i) Under Assumption (a) we can write

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

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

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

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

whose determinant is zero. □

10. References

Acemoglu D. and R. Shimer, 2000, "Wage and Technology Dispersion," *The Review of Economic Studies*, 67, 585-607.

Aghion, P., and P. Howitt, 1992, "A model of growth through creative destruction," *Econometrica*, 60, 323-351.

Arrow, K., 1962, Economic welfare and the allocation of resources for invention, in: Nelson, R. (Ed.), *The Rate and Direction of Inventive Activity*, Princeton University Press, 609-625.

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.

Baumol W. J., J. C. Panzar and R. D. Willig, 1982, *Contestable Markets and the Theory of Industry Structure*, Harcourt Brace Jovanovich Inc.

Berry, S. and P. Reiss, 2007, "Empirical Models of Entry and Market Structure," in M. Armstrong and R. Porter (ed.), *Handbook of Industrial Organization*, vol. 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 Schworm, 1988, "The Existence of Input and Output Aggregates in Aggregate Production Functions," *Econometrica,* 56, pp.613-643.

Blundell R. and J.-M. Robin, 2000, "Latent Separability: Grouping Goods without Weak Separability," *Econometrica*, 68, 53-84.

Browning, M. J., 1983, "Necessary and sufficient conditions for conditional cost functions," *Econometrica*, 51, 851-857.

Cabral, L., 2012, "Technology uncertainty, sunk costs, and industry shakeout," *Industrial and Corporate Change*, 21, 539-552.

Cameron, A. C., P. K. Trivedi., 2005, *Microeconometrics: methods and applications*, Cambridge University Press.

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.

Chambers, R. G., 1988, *Applied production analysis: the dual approach*, Cambridge University Press.

Chen, X., 2012, "Estimation of the CES Production Function with Biased Technical Change: A Control Function Approach," *BETA Working paper No 2012-20*, Department of Economics, University of Strasbourg.

Christensen, L.R., D.W. Jorgenson, and L.J. Lau, 1971, "Conjugate duality and the transcendental logarithmic production function," *Econometrica* 39, 255-256.

Colander, D., 2004, "On the treatment of fixed and sunk costs in the principles textbooks," *The Journal of Economic Education*, 35, 360-364.

Dehez, P., J. H. Drèze, 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, S. N. Durlauf 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 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, pp.1797-1821.

Gladden, T, 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, edited by C. Blackorby and A. F. Shorrocks, Clarendon Press: Oxford, 1995.

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 M. Fuss and D. McFadden (eds.) *Production Economics: A dual Approach to Theory and Applications*, Volume 1, North Holland, 409-453.

Leontief, W., 1947, "Introduction to a theory of the tnternal structure of functional relationships," *Econometrica*, 15, 361–373.

Mas-Colell, A., M. D. Whinston, J. R. Green, 1995, *Microeconomic theory*, Oxford University Press.

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.

Mundlak, Y., 1978, "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69-85.

Murphy, K., A., Shleifer, and R. Vishny, 1989, "Industrialization and the Big Push," *Journal of Political Economy*, 97, 1003-26.

Newey, W. K., K. D. West, 1987, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, 55, 703-708.

Pindyck, R. S., and Rotemberg, J. J., 1983, "Dynamic Factor Demand and the Effects of

Energy Price Shocks," *American Economic Review*, 73, 1066-1079.

Sutton, J., 2007, "Market Structure: Theory and Evidence," in M. Armstrong and R. Porter (ed.), *Handbook of Industrial Organization*, vol. 3, Elsevier.

Swamy, P. A. V. B.,1970, "Efficient inference in a random coefficient regression model," *Econometrica*, 38, 311-323.

Viner, J., 1931, "Costs Curves and Supply Curves," *Zeitschrift für Nationalölkonomie*, 3, 23-46.

Wang, X. H. and B. Z. Yang, 2001, "Fixed and Sunk Costs Revisited," *The Journal of Economic Education*, 32, 178-185.

Wang, X. H. and B. Z. Yang, 2004, "On the treatment of fixed and sunk costs in principles textbooks: a comment and a reply," *The Journal of Economic Education*, 35, 365-369.