INSTITUT DE STATISTIQUE

BIOSTATISTIQUE ET

SCIENCES ACTUARIELLES

(ISBA)

DISCUSSION PAPER

2017/30

Cross-Section Dependence and Latent Heterogeneity to Evaluate the Impact of Human Capital on Country Performance

Mastromarco, C. and L. Simar

Cross-Section Dependence and Latent Heterogeneity to Evaluate the Impact of Human Capital on Country Performance

Camilla Mastromarco* Léopold Simar§

September 08, 2017

Abstract: Human Capital has been recognized as the most important force behind economic growth of countries. However, the effect of this important growth factor on economic growth remains ambiguous due to endogeneity and latent heterogeneity. By using a dataset of 40 countries over 1970-2007, we estimate the global frontier and explore the channels under which human capital and time affect the production process and its components: impact on the attainable production set (input-output space), and the impact on the distribution of efficiencies. We extend existing methodological tools - robust frontier in non parametric location-scale models - to examine these interrelationships. We use a flexible nonparametric two-step approach on conditional efficiencies to eliminate the dependence of production inputs/outputs on common factors. We emphasize the usefulness of "pre-whitened" inputs/outputs to obtain more reliable measure of productivity and efficiency to better investigate the impact of human capital on the catching-up productivity process. Then, we take into account the problem of unobserved heterogeneity and endogeneity in the analysis of the influence of human capital on the production process by extending the instrumental nonparametric approach proposed by Simar, Vanhems and Van Keilegom (JoE 2016) to account also for cross section and time dependence.

JEL: C14, C13, C33, D24, O47.

Keywords: Human Capital, Latent Heterogeneity and Endogeneity, Cross-Section Dependence, Heterogeneous Panels, Robust Frontier, Nonparametric Location-Scale, Time-Varying Efficiency Measures, Environmental Factors.

*Dipartimento di Scienze dell'Economia, Universitá degli Studi del Salento. Centro ECOTEKNE, Via per Monteroni - 73100 LECCE - Italy. Phone +39832298779, fax +390832298757, E-mail: camilla.mastromarco@unisalento.it.

§Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain. Voie du Roman Pays 20 B - 1348 Louvain-La-Neuve - Belgium. Email: leopold.simar@uclouvain.be. Research support from IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

1 Introduction

US economic slowdown in the second half of the nineties and the slowdown following 2001 - more marked in Europe than in US -, leading many to question the recipe for endogenous self-sustained growth. The understanding of the sources of growth may mirror the larger debate between the neoclassical and new growth theories, but economists generally agree that this recent economic decline has largely been caused by the weak growth in TFP (total factor productivity), i.e., that part of the rise in productivity which is neither due to the increase in capital nor to the rise in the labour.

The "new" growth theory of Lucas (1988), Romer (1990a) and Barro (1990) and Barro (1997) has human capital playing an important role in productivity growth because human capital can help in explaining an economy's capacity to absorb new technologies (Abromovitz 1986, Cohen and Levinthal 1989, Kneller 2005, Kneller and Stevens 2006). For example, Benhabib and Spiegel (1994)' empirical study suggests that human capital plays a role in economic growth by helping in the adoption of technology from abroad and in creating the appropriate domestic technology. Hence, according to these studies, human capital is the most important force behind economic growth of countries.

Moreover, in response to the question of how technology diffusion affects economic growth, there has been an emerging empirical literature examining the nexus between the channels of technology diffusion and human capital in promoting economic growth. The evidence on this issue is mixed as seen in the differing conclusions of Miller and Upadhyay (2000, 2002) and Olofsdotter (1998). However, a consistent feature of all the empirical studies on this issue is the use of a cross- country regressions framework on a sample of developed and developing countries. Cross-country regressions cannot control for the unobservable heterogeneity which can arise, for example, from the different institutions in the various countries. Furthermore, Rodriguez (2006) argued that policy analysis within the growth-regression framework can carry considerable risks from the misspecification bias that come from using such a specification when it is not valid.

Therefore, the effect of this important growth factor - human capital - on economic growth remain ambiguous due to endogeneity and latent heterogeneity. In order to avoid the pitfalls of these parametric cross-regression studies we use an alternative empirical methodology, robust frontier in non parametric location-scale models to estimate the global frontier of 40 countries over 1970-2007, to answer our research question of whether a countries' productivity is affected by their existing levels of human capital.

We also control for the effect of openness on productivity. In particular we include in our analysis FDI (foreign direct investment) as the most important openness channel for technology diffusion. FDI leads to increases in productivity by spurring competition and transferring technology. New foreign competition arrivals provide domestic firms an incentive to use existing resources more efficiently which increases their productivity. Consequently, foreign firms have to invest even more in order to keep up with their technological advantage (Glass and Saggi 1998). FDI can also increase productivity through the transfer of technology. This occurs with the adoption of new technology brought by foreign multinational companies, imports of high- technology inputs, and the skills acquired by the local labour force as they are educated and trained by the foreign firms (for the empirical evidence on the effect of FDI on productivity see Mastromarco and Simar 2015).[1]

In addition, by using nonparametric frontier methodology we can also study the channels under which human capital affects the production process and its components: impact on the attainable production set (input-output space), and the impact on the distribution of efficiencies. We use a flexible nonparametric two-step approach on conditional efficiencies to eliminate the dependence of production inputs/outputs on common and external factors. We emphasize the usefulness of "pre-whitened" inputs/outputs to eliminate cross-section dependence and observed heterogeneity and to obtain more reliable measure of productivity and efficiency to better investigate the impact of human capital on the catching-up productivity process. Then, we take into account the problem of unobserved heterogeneity and endogeneity in the analysis of the influence of human capital on the production process by extending the instrumental nonparametric approach proposed by Simar et al. (2016) to account also for cross section and time dependence.

Note that in our previous paper Mastromarco and Simar (2015) the focus is on the time dependence of conditional efficiency frontier. Here we propose a different method. The main advantage of the approach used here, is that it allows to consider cross sectional dependence and latent heterogeneity which are very important in the analysis of the effect of human capital on productivity.

The paper is organized as follows. In the next section we give the main ideas of the chosen methodology and of the basic concepts needed to reach our objective. Then Section 3 will provide the technical econometric details for estimating the pieces of our model. Section 4 describe our data set and the empirical results we obtain in our application. Section 5 concludes and summarizes our main findings.

[1]Using these arguments, Borensztein et al. (1998), De Mello (1999) and Xu (2000) conclude that FDI increases an economy's productive efficiency. Javorcik (2004) argues that FDI can also raise productivity growth through vertical spillovers rather than horizontal spillovers.

2 The Methodology

We consider a production process that involves the production of an output $Y \in \mathbb{R}_+$ (the level of production by using inputs $X \in \mathbb{R}_+^p$ (the factors of production). In our setup, Y will be measured by the GDP of a country and we will consider the three factors of production: Capital-X_K, Labor-X_L and Human Capital-X_H, where X_H is measured as the average years of education in the population. In the empirical section below we describe how these data are collected (we have 40 countries over the years 1970–2007). The optimal level of production is given by a global frontier, which is defined as the maximum attainable level of outputs that can be reached by using a level X of the factors. The efficiency of each observation can then be measured by its distance to its corresponding (optimal) frontier point.

We assume that Human Capital X_H is linked to some unobserved (latent) factor of heterogeneity. As recognized in the literature (see e.g. Simar et al. 2016), omitting such latent factor may induce some endogeneity problem for the estimation of the production frontier (see details below in section 3.1). We would like to explore the channels under which this latent factor may influence the production process of a country: impact on the attainable production set and impact on the distribution of efficiencies. This latent factor V can be identified by some instrumental variable W and we use a nonseparable nonparametric model for the link between X_H and the latent factor V, as suggested in Simar et al. (2016) (see also, e.g. Matzkin 2003). Our instrument W will be Life Expectancy. So we use the following model

$$X_H = \phi(W, V), \tag{2.1}$$

which is nonseparable in V. Here, W has to be correlated to X_H but is independent of V. As explained in details in Simar et al. (2016), the unobserved variable V can be viewed as the part of the X_H which is independent of W. We also impose a classical assumption of monotonicity of ϕ with respect to V and we assume, without loss of generality, that V is uniformly distributed on $[0, 1]$. Under these assumptions V is identified by the conditional distribution of X_H given W:

$$V = F_{X_H|W}. \tag{2.2}$$

It is useful to notice that ϕ is unknown and since V is uniform on $[0, 1]$, ϕ can be interpreted as a quantile function. The choice of a uniform distribution for V is just a matter of scaling V.

In our empirical application, human capital X_H impacts the productivity of a country. However, there may exist unobserved characteristics of the countries that may influence both the existing level of education and the productivity, as the quality of institutions. To take into account this latent heterogeneity we use life expectancy as instrument. This unobservable heterogeneity, correlated with human capital, might be interpreted as absorptive capability, defined as the potential to master new knowledge embodied in innovation, which acts as free disposal input. The innovation has similar attributes as a public good, indeed it is non rival and non exclusive. Given these features, there exists the problem to protect the property rights of innovations which would guarantee profits to the innovators and stimulate new efforts to innovate. Hence, it is very likely that, this unobservable factor is capturing the quality of institutions (e.g. very high in US, very low in Mexico) and the difference in property rights systems among countries. So to summarize, in our setup, we might expect that V captures some measure of "innovation" in the country. We will see in our data set how these potential interpretations receive some empirical evidences, by looking to the links of the estimated versions of V with some appropriate indicators.

The latent factor V is unobserved but can indeed be predicted for each observation by an estimate of the preceding conditional distribution. We have

$$\widehat{V}_{it} = \widehat{F}_{X_H|W}(X_{H,it}|W_{it}), \text{ for } i = 1, \dots, n, t = 1, \dots, s, \tag{2.3}$$

where some nonparametric estimator of the conditional distribution $\widehat{F}_{X_H|W}$ will be described below.

For the nonparametric frontier estimation, we have a panel of data (X_{it}, Y_{it}), $i = 1, \dots, n$ and $t = 1, \dots, s$. A simple (but naive) way to investigate the production process, and the performance of each observation, would be to analyze how far are the observations from a simple unconditional frontier, locus of optimal level of production y for a given level of the input factors x. According to the literature, a probability formulation of the production process defines the unconditional frontier function as

$$\tau(x) = \sup\{y|S_{Y|X}(y|X \leq x) > 0\}, \tag{2.4}$$

where the survival function $S_{Y|X}(y|X \leq x) = \mathbb{P}(Y \geq y|X \leq x)$. So $\tau(x)$ is the upper support for Y among units (countries) using less input factors than x.[2] This probability formulation allows also to define a less extreme benchmark than the full frontier for providing robust versions of the frontier that will be less sensible to extreme or outlying data. We select here the order-m frontier introduced by Cazals et al. (2002), where for a given value of m, we

[2]As usual, inequalities on vectors have to be understood component wise.

have

$$\tau_m(x) = \mathbb{E}[\max(Y_1, \ldots, Y_m)|X \leq x] \tag{2.5}$$

where $Y_1, \ldots, Y_m$ are m independent copies of Y drawn in the conditional population of units with $X \leq x$, i.e., using less that x input factors. So we have (when Y is univariate)

$$\begin{aligned}
\tau_m(x) &= \int_0^\infty [1 - (F_{Y|X}(y|X \leq x))^m] dy \\
&= \tau(x) - \int_0^{\tau(x)} F^m_{Y|X}(y|X \leq x) dy
\end{aligned} \tag{2.6}$$

since $F_{Y|X} = 1 - S_{Y|X}$, the conditional cdf of Y given $X \leq x$, is equal to 1 when $y \geq \tau(x)$.[3] It is known that when $m \to \infty$, the order-m frontier converges to the full frontier $\tau(x)$. But for finite m it is a less extreme benchmark frontier, since it involves an expectation. In practice we will use large values of m to provide robust estimates of the production frontier (see Daouia et al. 2012, for the theoretical justifications and see below how to fix m in practice). We can also look at the case $m = 1$ for exploring the average production level for units using less than x inputs.

Nonparametric estimators of $\tau(x)$ are then obtained by plugging in (2.4) an empirical version of the survival function providing the FDH estimators introduced in Deprins et al. (1984). The statistical properties have been derived in Park et al. (2000) and Daouia et al. (2010). To summarize, under wide regularity conditions, the asymptotic distribution is linked to the Weibull but suffers from the "curse of dimensionality" with rates of convergence $n^{1/(p+q)}$ depending on the number of variables: p inputs and q outputs (in our case, $p+q = 4$).

Nonparametric estimates of $\tau_m(x)$ are obtained in the same way by plugging the empirical survival in equation (2.5). These order-m shares interesting properties. The main advantage is that they have better asymptotic properties, a parametric $\sqrt{n}$-rate of convergence and an asymptotic normal distribution but also they are much more robust to extreme or outlying data points as discussed in details in Cazals et al. (2002) and Daraio and Simar (2005). See also Daouia and Gijbels (2011) for the analysis of these estimators from a theory of robustness perspective. Simar and Wilson (2015) offer a recent survey on all these concepts with a detailed description of the statistical properties of the most popular nonparametric estimators.

In the presence of a panel of data, and due to our particular macroeconomic set-up, we have to adapt these concepts and base rather our analysis by using conditional frontiers as

[3] Multivariate versions have also been derived in Cazals et al. (2002) and in Daraio and Simar (2005).

developed in Cazals et al. (2002), Daraio and Simar (2005) and more recently by Florens et al. (2014). For handling the time dimension and the cross sectional dependence (CSD) due to the influence of globalization factors (e.g., technological shocks and financial crises) on the economic performance of countries under analysis, we envelop the effect of CSD on the production process. For doing so, we assume that the production process is function of some unobserved time-varying factors. We follow here Pesaran (2006) and Bai (2009) and we consider $F_t = (t, X_{.t}, Y_{.t})$ as proxy for the unobserved time-varying factors.[4] We will therefore analyze production frontier that are conditional to these time-varying factors F_t. Finally, we also include in the analysis observed environmental factors that may influence the production process. We have seen above that FDI is such a factor which allows for controlling the effect of openness of the economy on productivity, in what follows, the FDI variable will be denoted by Z.

To summarize we have data on $(X_{it}, Y_{it}, F_t, Z_{it}, W_{it})$ for $i = 1, \ldots, n$ and $t = 1, \ldots, s$ where a generic $X = (X_K, X_L, X_H)$. From the observations $(X_{H,it}, W_{it})$ we will predict the values $\widehat{V}_{it}$. Then we will first estimate for any value of (x, y, f, z) the conditional production frontier as the maximal achievable output level y for a country using inputs $X \leq x$ and confronted to the time-varying factors $F = f$ and to the observable environmental conditions $Z = z$ (FDI). From the literature on conditional frontier models (see Cazals et al. 2002, Daraio and Simar 2005) this frontier is defined as

$$\tau(x|f, z) = \sup\{y | S_{Y|X,F,Z}(y|x, f, z) > 0\}, \tag{2.7}$$

thus the frontier level, given a value of (x, z, f), is the upper support point of the following conditional survival function

$$S_{Y|X,F,Z}(y|x, f, z) = \mathbb{P}(Y \geq y | X \leq x, F = f, Z = z). \tag{2.8}$$

We remark that $\tau(x|f, z)$ is a function of (x, f, z) since Y is univariate.[5] We note also, as explained in Cazals et al. (2002), that the conditioning on X is through an inequality (to insure monotonicity of the frontier in the input level) but conditioning on the other factors are, as in usual conditional models, through equalities. Once we have the conditional frontier,

[4] Here we use the standard notation where a dot in a subscript means that we averaged over this index. Pesaran (2006) shows that the unobserved common factor can indeed be consistently proxied by cross-sectional averages of dependent and independent variables, as $n, s \to \infty$, and $s/n \to K$ (where $i = 1 \ldots n$ and $t = 1 \ldots s$) for some constant $K \in [0, \infty)$.

[5] For multivariate outputs, the similar expressions can be used for defining appropriate conditional efficiency scores; the optimal production set is then a surface. See e.g. Daraio and Simar (2005) for details.

we can define the performance (efficiency) of a unit operating at level (x, y) and confronted to conditions (f, z) by its distance to the frontier. We can use an additive directional measure of inefficiency in the output direction defined as

$$\delta(x, y|f, z) = \sup\{\delta > 0 | S_{Y|X,F,Z}(y(1 + \delta)|x, f, z) > 0\} = \tau(x|f, z) - y, \tag{2.9}$$

where we see this unit (x, y) is efficient if $\delta(x, y|f, z) = 0$. Here, for projecting a curent value (x, y) on the frontier, we use the directional vector $d_x = 0$ for the inputs and $d_y = y$ for the output (see for basic definitions, e.g. Färe and Grosskopf 2000, Simar and Vanhems 2012, for the probabilistic characterization we use here).

For estimating the conditional frontiers, conditional to (F, Z), we will follow the strategy of Florens et al. (2014) that use in a first step a flexible nonparametric location scale model to whiten the inputs X and the output Y from the effect of business cycles and globalization shocks (proxied by the common factors F) and from the effect of FDI, Z. The details are given in the next section, but the idea is to build "pure" inputs ε_x and "pure" output ε_y, cleaned from these effects. The efficient frontier in these new units is given by

$$\varphi(e_x) = \sup\{e_y | S_{\varepsilon_y|\varepsilon_x}(e_y|e_x) = \mathbb{P}(\varepsilon_y \geq e_y | \varepsilon_x \leq e_x) > 0\}, \tag{2.10}$$

gives the optimal level in the pure units. As in Florens et al. (2014) we define a measure of pure inefficiency for a unit with current values (e_x, e_y) by using the directional output distance between this point and its projection to the efficient frontier $(e_x, \varphi(e_x))$

$$\rho(e_x, e_y) = \varphi(e_x) - e_y. \tag{2.11}$$

The main advantage is that this measure of inefficiency has been cleaned from the effect of (F, Z) and allows a more fair comparison of the performance of each country. We will show in the next section that from these "pure" objects, we can recover, if wanted, the frontier $\tau(x|f, z)$ and the conditional inefficiency $\delta(x, y|f, z)$ in the original units for all values of (x, y, f, z).

At the next step, we will estimate the conditional frontier where in addition we condition on the identified latent factor V. For $V = v$ this is defined as

$$\tau(x|f, z, v) = \sup\{y | S_{Y|X,F,Z,V}(y|x, f, z, v) > 0\}, \tag{2.12}$$

where here $S_{Y|X,F,Z,V}(y|x, f, z, v) = \mathbb{P}(Y \geq y | X \leq x, F = f, Z = z, V = v)$. As above conditioning on (F, Z, V) the measure of inefficiency of a unit operating at level (x, y) is given by $\delta(x, y|f, z, v) = \tau(x|f, z, v) - y$.

For the estimation here, we have to be coherent with all the above assumptions, including the location-scale model for whiten the inputs and output. We explain below that this can be achieved by assuming a similar model for V, in other words, we will also whiten V from the effects of business cycles and globalization shocks and from the effect of FDI. This provide ε_v, the "pure" version of the latent factor V. Now using the methodology suggested by Simar et al. (2016) we can easily determine the conditional frontier in the pure inputs/output space $(\varepsilon_x, \varepsilon_y)$, conditionally on $\varepsilon_v = e_v$:

$$\varphi(e_x|e_v) = \sup\{e_y|S_{\varepsilon_y|\varepsilon_x,\varepsilon_v}(e_y|e_x, e_v) > 0\}, \tag{2.13}$$

and $S_{\varepsilon_y|\varepsilon_x,\varepsilon_v}(e_y|e_x, e_v) = \mathbb{P}(\varepsilon_y \geq e_y|\varepsilon_x \leq e_x, \varepsilon_v = e_v)$ is the conditional survival of ε_y given that $\varepsilon_x \leq e_x$ and that $\varepsilon_v = e_v$. Accordingly, a pure measure of inefficiency for a unit with current values (e_x, e_y) and facing the value of the pure latent factor e_v is given by

$$\rho(e_x, e_y|e_v) = \varphi(e_x|e_v) - e_y. \tag{2.14}$$

Again, we will show in the next section that if corresponding values of $\tau(x|f, z, v)$ and $\delta(x, y|f, z, v)$ in the original units are wanted, they can be immediately derived from $\varphi(e_x|e_v)$ and $\rho(e_x|e_v)$, repectively.

It is the comparison of the estimators of these two conditional frontiers in "pure" units, $\varphi(e_x)$ and $\varphi(e_x|e_v)$ (or of their corresponding "pure" efficiency measures $\rho(e_x)$ and $\rho(e_x|e_v)$) that will allow to explore the effect of V on the production process. We will choose and adapt the methodology suggested by Bădin et al. (2012), and Daraio and Simar (2014), as described in details in the next section.

Finally, we know that some extreme values (or outliers) in the data may influence the estimation of the frontier and may hide the real effect of ε_v on the frontier levels (see Figures 5.3 and 5.4 in Daraio and Simar 2007, for some classroom examples). So we will, as in Florens et al. (2014) and as in Simar et al. (2016) use the robust, order-m versions of the conditional frontiers. We will first evaluate the order-m frontier in the "pure" inputs and output units cleaned from the effect of (F, Z). To summarize

$$\varphi_m(e_x) = \mathbb{E}[\max(\varepsilon_{y,1}, \ldots, \varepsilon_{y,m})|\varepsilon_x \leq e_x], \tag{2.15}$$

is the simple order-m frontier of ε_y given that $\varepsilon_x \leq e_x$ and $\varepsilon_{y,1}, \ldots, \varepsilon_{y,m}$ are m iid copies of ε_y. Its computation goes along the lines of (2.5)–(2.6) by using the survival function $S_{\varepsilon_y|\varepsilon_x}$ defined in (2.10). Here again, a measure of pure order-m inefficiency for a unit with current value in pure units (e_x, e_y) is given by the distance

$$\rho_m(e_x, e_y) = \varphi_m(e_x) - e_y. \tag{2.16}$$

Note that here the values of $\rho_m(e_x, e_y)$ are not restricted to be positive, since the current value (e_x, e_y) can be above the order-m frontier.

Now for conditioning on the "pure" latent $\varepsilon_v = e_v$ we have for the conditional order-m frontier

$$\varphi_m(e_x|e_v) = \mathbb{E}[\max(\varepsilon_{y,1}, \ldots, \varepsilon_{y,m})|\varepsilon_x \leq e_x, \varepsilon_v = e_v]. \tag{2.17}$$

Here, its computation can be done along the lines of (2.6), using the conditional survival function $S_{\varepsilon_y|\varepsilon_x,\varepsilon_v}$ defined above in (2.13). Accordingly, the conditional pure inefficiency measure of order-m is given by

$$\rho_m(e_x, e_y|e_v) = \varphi_m(e_x|e_v) - e_y. \tag{2.18}$$

The way to estimate these measures and the asymptotic properties of the resulting estimators are described in Florens et al. (2014) and Simar et al. (2016) when conditioning on the latent factor. We summarize these details in the next section and show also how the corresponding measures can, if wanted, be easily derived in the original units of (X, Y) and V (e.g. to recover $\tau_m(x|f, z)$ and $\tau_m(x|f, z, v)$ and the corresponding inefficiencies).

As explained above for large values of m we have robust estimator of the full frontier, we will use these with large values of m to get insight on the role of V on the location of the frontier (check if V may influence the production set). We explain below how to select appropriate value of m in order to eliminate potential outliers. Then, as suggested in Bădin et al. (2012), we also estimate order-m frontier with small values of m, e.g. $m = 1$ to explore the effect of V on the position of the center of the distribution of the efficiencies, since $m = 1$ gives the conditional average of the production levels. All technical details and theoretical results are presented in the next section.

3 Technical Details

In this section, we summarize the concepts and notations introduced in Florens et al. (2014) and Simar et al. (2016) and how we adapt and extend them to include time, global factors and endogeneity. This is needed for applying the methodology summarized above and used in the empirical application.

3.1 Predicted values of the latent factor and endogeneity issues

As explained above under the nonseparable model (2.1) which link X_H to the latent factor by the way of the instrument W, V is identified by (2.2). So, given the data $(X_{H,it}, W_{it})$ the predicted values of V for each observation point is given by (2.3). A natural nonparametric estimator, as the one proposed by Simar et al. (2016) is defined by

$$\widehat{V}_{it} = \widehat{F}_{X_H|W}(X_{H,it}|W_{it}) = \frac{\sum_{j=1,u=1}^{n,s} \mathbb{1}\big(X_{H,ju} \leq X_{H,it}\big) K_{h_w}\big(W_{ju} - W_{it}\big)}{\sum_{j=1,u=1}^{n,s} K_{h_w}\big(W_{ju} - W_{it}\big)}, \tag{3.1}$$

where $K_{h_w}(\zeta) = (1/h_w)K(\zeta/h_w)$ and $K(\cdot)$ is a kernel with compact support on $[-1,+1]$, like e.g. Epanechnikov and h_w is a bandwidth. In practice this bandwidth can be determined by Least-Squares Cross Validation (LSCV) techniques that are now standard in nonparametric estimation (see e.g. Li et al. 2013).

As explained in Simar et al. (2016), if we ignore this latent factor when evaluating the performances of the countries we may create endogeneity problems. The argument goes as follows; if we consider the conditional frontier taking into account the latent heterogeneity factor V, we can write $Y = \tau(X|F, Z, V) - U$ where $U \geq 0$ is an additive measure of inefficiency. The distribution of U may depend on (X, F, Z, V) but by construction its support does not ($U \geq 0$). The latter is the concept of partial exogeneity of (X, F, Z, V) introduced in Cazals et al. (2016), where it is shown that in this case there is no severe consequence for the estimation of the conditional frontier by traditional methods (like conditional FDH/DEA type estimators); we keep consistency and rates of convergence. But if we ignore V and use rather $\tau(X|F, Z)$ as the frontier, we still can write $Y = \tau(X|F, Z) - \widetilde{U}$ but now we see that $\widetilde{U} = \tau(X|F, Z) - \tau(X|F, Z, V) + U$ and we may lose the partial exogeneity condition unless $\tau(X|F, Z) \equiv \tau(X|F, Z, V)$ with probability one. We remark that the latter is a reformulation of the "separability condition" for V introduced in Simar and Wilson (2007). We will suggest below a methodology for exploring these issues.

3.2 Nonparametric estimation of the conditional frontier $\tau(x|f,z)$

The conditional frontier $\tau(x|f,z)$ has been defined in (2.7) and a natural nonparametric estimator is obtained by plugging a nonparametric estimator of the conditional survival function $S_{Y|X,F,Z}(y|x,f,z)$ given in (2.8). This requires some smoothing relative to the conditioning variables (F,Z) but not for X since the condition on X is an equality and so empirical conditional cdf can be used when only X is concerned. These estimators have been analyzed in Cazals et al. (2002) and Daraio and Simar (2005) and we have

$$\widehat{S}_{Y|X,F,Z}(y|x,f,z) = \frac{\sum_{it} 1\!\!1(X_{it} \le x, Y_{it} \ge y) K_{h_f}(F_t - f) K_{h_z}(Z_{it} - z)}{\sum_{it} 1\!\!1(X_{it} \le x) K_{h_f}(F_t - f) K_{h_z}(Z_{it} - z)},$$

with the $K_{h_u}(\zeta) = (1/h_u)K(\zeta/h_u)$ where $K(\cdot)$ are traditional kernels (when needed multivariate kernels (like e.g. product kernels) with compact support and bandwidths vector h_u, with the usual notational convention that division by vectors is element-wise. We will not follow this route since Florens et al. (2014) have shown the advantage of using an alternate approach based on a flexible nonparametric location-scale models for "cleaning" the input Y and the outputs X from their dependence of (F,Z).

The approach can be seen as follows. Suppose the vector (X,Y,F,Z) follows the location-scale regression model:

$$\begin{cases} X_{it} &= \mu_x(F_t, Z_{it}) + \sigma_x(F_t, Z_{it})\varepsilon_{x,it} \\ Y_{it} &= \mu_y(F_t, Z_{it}) + \sigma_y(F_t, Z_{it})\varepsilon_{y,it} \end{cases}, \tag{3.2}$$

where μ_x, σ_x and ε_x have each 3 components (in our setup we have 3 input factors) and, for ease of notations, the product of vectors is component-wise. So the first equation in (3.2) represents 3 relations. Here $\mu_x(F,Z) = \mathbb{E}(X|F,Z)$ and element by element, $\sigma_x^2(F,Z) = \mathbb{V}(X|F,Z)$ and the same for Y, $\mu_y(F,Z) = \mathbb{E}(Y|F,Z)$ and $\sigma_y^2(F,Z) = \mathbb{V}(Y|F,Z)$. The location-scale model assume that $(\varepsilon_x, \varepsilon_y)$ is asymptotically independent of (F,Z). The approach suggested by Florens et al. (2014) is to estimate the full frontier $\rho(e_x)$, and their robust versions $\rho_m(e_x)$ and their corresponding inefficiency measures $\rho(e_x, e_y)$ and $\rho_m(e_x, e_y)$ without being affected by (F,Z). In other words, under model (3.2), comparison and ranking of observation units is legitimate, since the effect of the variables (F,Z) has been eliminated.

But if wanted we can recover the conditional frontier $\tau(x|f,z)$ and the inefficiency measures $\delta(x,y|f,z)$ as follows. Under the model assumptions, we have indeed from (2.7) and

(2.8)

$$\begin{aligned}
\tau(x|f,z) &= \sup\left\{y|\mathbb{P}\left(\frac{Y-\mu_y(F,Z)}{\sigma_y(F,Z)} \geq \frac{y-\mu_y(f,z)}{\sigma_y(f,z)}\Big|X \leq x, F=f, Z=z\right) > 0\right\} \\
&= \mu_y(f,z) + \sup\left\{e_y|\mathbb{P}\left(\varepsilon_y \geq e_y\big|\varepsilon_x \leq \frac{x-\mu_x(f,z)}{\sigma_x(f,z)}\right) > 0\right\}\sigma_y(f,z) \\
&= \mu_y(f,z) + \varphi\left(\frac{x-\mu_x(f,z)}{\sigma_x(f,z)}\right)\sigma_y(f,z), \qquad (3.3)
\end{aligned}$$

where $\varphi(e_x)$ is defined in the pure units space, see (2.10). In other words (3.3) indicates that the frontier function in the original units is a simple transformation of the frontier obtained in the new units. For the distance function, it is easy to verify that

$$\delta(x,y|f,z) = \sigma_y(f,z)\rho(e_x,e_y), \qquad (3.4)$$

which interestingly indicates that the distance in original units may be highly dependent of the variables (F,Z) though a scaling factor. This again advocate our approach of using “pure” measures of inefficiency.

The advantage is also from an estimation perspective: we only need estimators of the functions $\mu_\ell(f,z)$ and $\sigma_\ell(f,z)$ $(\ell = x,y)$ to obtain an estimate of $\tau(x|f,z)$ which require only smoothing in the center of the data. The fact that we avoid smoothing at the frontier is important, as in general, data are more sparse there and estimators can be sensitive to outliers. In addition, as pointed in Bădin et al. (2017), no easy ways are available to derive optimal bandwidths for estimating the support of a conditional distribution like $S_{Y|X,F,Z}(y|x,f,z)$. Note that in some applications, the practitioner may also analyze these additional tools like, e.g., $\mu_\ell(f,z)$ to appreciate marginally the mean behaviour of inputs and output as a function of (f,z).

We can also derive the link between the order-m conditional frontier and its analog in the “pure” units space:

$$\tau_m(x|f,z) = \mu_y(f,z) + \varphi_m\left(\frac{x-\mu_x(f,z)}{\sigma_x(f,z)}\right)\sigma_y(f,z), \qquad (3.5)$$

where $\varphi_m(e_x)$ was defined in (2.15). Similarly it is easy to show that

$$\delta_m(x,y|f,z) = \sigma_y(f,z)\rho_m(\varepsilon_x,\varepsilon_y), \qquad (3.6)$$

where ρ_m was defined in (2.16). We observe the same scaling factor as for the full frontier above.

In practice the estimation steps go as follows. We first clean the inputs and the output by estimating the models (3.2). We first local-linear models for estimating the functions μ_x and the function μ_y. From the squared residuals regressed on (F, Z) by a local constant estimator (to avoid negative variances) we derived the estimator of σ_ℓ, $\ell = x, y$. Finally the estimates of the "pure" inputs and output are obtained as

$$\widehat{\varepsilon}_{x,it} = \frac{X_{it} - \widehat{\mu}_x(F_t, Z_{it})}{\widehat{\sigma}_x(F_t, Z_{it})}, \tag{3.7}$$

$$\widehat{\varepsilon}_{y,it} = \frac{Y_{it} - \widehat{\mu}_y(F_t, Z_{it})}{\widehat{\sigma}_y(F_t, Z_{it})}, \tag{3.8}$$

where as before, for ease of notation, a ratio of two vectors has to be understood component wise.

We only need an estimator of the survival function $S_{\varepsilon_y|\varepsilon_x}(e_y|e_x)$ for obtaining estimators of the frontier and of the inefficiencies in the pure units, but this does not require any smoothing and can be obtained at parametric $\sqrt{N}$-rates (see Corollary 2,(i) in Florens et al. 2014) where $N = ns$ is the total number of observations. The empirical version of this survival is given by

$$\widehat{S}_{\varepsilon_y|\varepsilon_x}(e_y|e_x) = \frac{\sum_{it} \mathbb{I}(\widehat{\varepsilon}_{x,it} \leq e_x, \widehat{\varepsilon}_{y,it} \geq e_y)}{\sum_{it} \mathbb{I}(\widehat{\varepsilon}_{x,it} \leq e_x)} \tag{3.9}$$

Algorithms for computing $\widehat{\varphi}_m(e_x)$ or $\widehat{\rho}_m(e_x, e_y)$ are described in Simar and Vanhems (2012) and Daraio and Simar (2014).

As proven in Florens et al. (2014) under wide regularity conditions (smoothness of the μ_ℓ and σ_ℓ functions, $\ell = x, y$), the resulting estimators of the full-frontier and of the order-m conditional frontiers obtained through the location-scale models (and through their estimates), are similar to the one that we would obtain by using $\widehat{S}_{Y|X,F,Z}(y|x, f, z)$. Here and below we are mostly interested to the robust order-m version and it can be proven that under wide regularity conditions, we have $\sqrt{N}$-rates of convergence and an asymptotic normal distribution (see Corollary 2,(ii) in Florens et al. 2014). This will be enough for what we need in our application for exploring the effect of V on the production process.

3.3 Nonparametric estimation of the conditional frontier $\tau(x|f,z,v)$

We might be willing to use the same attractive approach of the location-scale models to whiten also the output and the inputs from the influence of V. However the non-separable assumption defining V is not compatible with a location-scale model in V. So, here we have no choice and we follow the approach suggested in Simar et al. (2016). As shown below if we want to use the results of the preceding section and working in the "pure" inputs and output space, we need to whiten the latent factor V from the influence of (F,Z) and we will also a flexible location-scale model. So we add to the system (3.2) a new equation defining a "pure" version of the latent factor ε_v whitened from the influence of (F,Z):

$$V_{it} = \mu_v(F_t, Z_{it}) + \sigma_v(F_t, Z_{it})\varepsilon_{v,it} \tag{3.10}$$

where now the independence assumption is that $(\varepsilon_x, \varepsilon_y, \varepsilon_v)$ are jointly independent of (F,Z). The estimation of ε_v goes along the same lines as for ε_x and ε_y described above. We can now easily determine the conditional frontier in the pure units space, conditionally on $\varepsilon_v = e_v$:

$$\varphi(e_x|e_v) = \sup\{e_y | S_{\varepsilon_y|\varepsilon_x,\varepsilon_v}(e_y|e_x,e_v) > 0\}, \tag{3.11}$$

and $S_{\varepsilon_y|\varepsilon_x,\varepsilon_v}(e_y|e_x,e_v) = \mathbb{P}(\varepsilon_y \geq e_y|\varepsilon_x \leq e_x, \varepsilon_v = e_v)$ is the conditional survival of ε_y given that $\varepsilon_x \leq e_x$ and that $\varepsilon_v = e_v$.

It is easy to prove that the knowledge of $\varphi(e_x|e_v)$ allows to recover $\tau(x|f,z,v)$. Indeed we can write from (2.12)

$$\begin{aligned}
\tau(x|f,z,v) &= \sup\{y|\mathbb{P}(Y \geq y|X \leq x, F=f, Z=z, V=v) > 0\} \\
&= \sup\left\{y|\mathbb{P}\left(\frac{Y-\mu_y(F,Z)}{\sigma_y(F,Z)} \geq \frac{y-\mu_y(f,z)}{\sigma_y(f,z)}\Big| X \leq x, F=f, Z=z, V=v\right) > 0\right\} \\
&= \mu_y(f,z) + \sup\left\{e_y|\mathbb{P}\left(\varepsilon_y \geq e_y \Big| \varepsilon_x \leq \frac{x-\mu_x(f,z)}{\sigma_x(f,z)}, F=f, Z=z, V=v\right) > 0\right\}\sigma_y(f,z) \\
&= \mu_y(f,z) + \sup\left\{e_y|\mathbb{P}\left(\varepsilon_y \geq e_y \Big| \varepsilon_x \leq \frac{x-\mu_x(f,z)}{\sigma_x(f,z)}, \varepsilon_v = \frac{v-\mu_v(f,z)}{\sigma_v(f,z)}\right) > 0\right\}\sigma_y(f,z).
\end{aligned}$$

where the last equality is obtained because we assume ε_v is independent of (F,Z). Then we have the relation

$$\tau(x|f,z,v) = \mu_y(f,z) + \varphi\left(\frac{x-\mu_x(f,z)}{\sigma_x(f,z)} \Big| \frac{v-\mu_v(f,z)}{\sigma_v(f,z)}\right)\sigma_y(f,z). \tag{3.12}$$

We thus immediately obtain

$$\delta(x, y|f, z, v) = \sigma_y(f, z)\rho(e_x, e_y|e_v) \tag{3.13}$$

where the pure measure of conditional inefficiency $\rho(e_x, e_y|e_v)$ is defined in (2.14).

The robust order-m versions follow the same scheme. To summarize we have

$$\tau_m(x|f, z, v) = \mu_y(f, z) + \varphi_m\left(\frac{x - \mu_x(f, z)}{\sigma_x(f, z)}\middle|\frac{v - \mu_v(f, z)}{\sigma_v(f, z)}\right)\sigma_y(f, z), \tag{3.14}$$

and similarly

$$\delta_m(x, y|f, z, v) = \sigma_y(f, z)\rho_m(e_x, e_y|e_v). \tag{3.15}$$

Due to our objectives described above, we are mainly interested of the measures in the pure units space. The needed quantities will be estimated, as above, by plugging a nonparametric estimator of the survival function $S_{\varepsilon_y|\varepsilon_x,\varepsilon_v}$ obtained from the estimates $\{(\widehat{\varepsilon}_{y,it}, \widehat{\varepsilon}_{x,it}, \widehat{\varepsilon}_{v,it})\}$, $i = 1, \dots, n$ and $t = 1, \dots, s$. Note that here the nonparametric estimator of this conditional survival involves smoothing in the variable ε_v, and requires bandwidth selections. Simar et al. (2016) notice that we obtain better properties of the resulting estimator of the conditional frontier by determining the optimal bandwidth for estimating the joint conditional probabilities $H_{\varepsilon_y,\varepsilon_x|\varepsilon_v}(e_y, e_x|e_v) = \mathbb{P}(\varepsilon_y \geq e_y, \varepsilon_x \leq e_x|\varepsilon_v = e_v)$, i.e. where we only condition in ε_v. With this approach, there is also only one bandwidth to determine (and not one value for each value of e_x); see Appendix A in Simar et al. (2016) for the details and arguments.To be specific we have to compute the following estimator

$$\widehat{H}_{\varepsilon_y,\varepsilon_x|\varepsilon_v}(e_y, e_x|e_v) = \frac{\sum_{it} 1\!\!1(\widehat{\varepsilon}_{y,it} \geq e_y, \widehat{\varepsilon}_{x,it} \leq e_x)K_{h_v}(\widehat{\varepsilon}_{v,it} - e_v)}{\sum_{it} K_{h_v}(\widehat{\varepsilon}_{v,it} - e_v)}, \tag{3.16}$$

with the standard notation for the kernel K_{h_v} already introduced above and the bandwidth h_v determined by LSCV (see Li et al. 2013, for details). Then $\widehat{S}_{\varepsilon_y|\varepsilon_x,\varepsilon_v}$ may be obtained as

$$\widehat{S}_{\varepsilon_y|\varepsilon_x,\varepsilon_v}(e_y|e_x, e_v) = \frac{\widehat{H}_{\varepsilon_y,\varepsilon_x|\varepsilon_v}(e_y, e_x|e_v)}{\widehat{H}_{\varepsilon_y,\varepsilon_x|\varepsilon_v}(-\infty, e_x|e_v)}. \tag{3.17}$$

The practical computations for obtaining the estimators $\widehat{\rho}(e_x, e_y|e_v)$ and $\widehat{\rho}_m(e_x, e_y|e_v)$ are described in Daraio and Simar (2014).

Compared to the estimator of the preceding section, due to the smoothing in ε_v we loose the parametric $\sqrt{N}$ rate of convergence but it may be proven (see Cazals et al. 2002) that by conditioning on ε_v, the rate of convergence is now $\sqrt{Nh_v}$. Since the optimal bandwidths obtained by LSCV are order of $h_v = O\big(N^{-1/5}\big)$ this gives a rate of convergence for the conditional order-m estimators $\widehat{\varphi}_m(e_x|e_v)$ or $\widehat{\rho}_m(e_x, e_y|e_v)$ of the order $\sqrt{N^{4/5}}$ which is a bit lower than the rate $\sqrt{N}$ obtained for the unconditional objects $\widehat{\varphi}_m(e_x)$ or $\widehat{\rho}_m(e_x, e_y)$. As shown in Simar et al. (2016) the rates and the asymptotic normality is preserved when replacing the unobserved $\varepsilon_{v,it}$ by their estimates $\widehat{\varepsilon}_{v,it}$.

3.4 Exploring the effect of the latent factor on the production process

Updating the ideas of Daraio and Simar (2005) and Florens et al. (2014) to our framework here, the analysis of the global effect of the factor V (measured by its cleaned version ε_v) on the production process can be captured by the analysis of the differences (remember that here Y and so ε_y are univariate)

$$\begin{aligned} D(\varepsilon_x, \varepsilon_v) &= \varphi(\varepsilon_x) - \varphi(\varepsilon_x|\varepsilon_v), \\ &= \rho(\varepsilon_x, \varepsilon_y) - \rho(\varepsilon_x, \varepsilon_y|\varepsilon_v) \end{aligned}$$

as a function of ε_v, at various fixed levels of the inputs ε_x. For the output orientation we have here, $D(\varepsilon_x, \varepsilon_v) \geq 0$ but a global tendency of these differences to decrease with ε_v indicates a positive effect, the conditional frontier $\varphi(\varepsilon_x|\varepsilon_v)$ moves up to unconditional one $\varphi(\varepsilon_x)$ when ε_v increases, so ε_v acts as a freely disposable input. On the contrary, if the differences increase with ε_v indicates a negative effect of ε_v on the production process, the conditional frontier $\varphi(\varepsilon_x|\varepsilon_v)$ moves away from the unconditional one $\varphi(\varepsilon_x)$, so ε_v acts as an undesirable output.

As explained above we prefer to do the analysis on the robust order-m frontiers, i.e. To be specific define the following random variable

$$\begin{aligned} D_m(\varepsilon_x, \varepsilon_v) &= \varphi_m(\varepsilon_x) - \varphi_m(\varepsilon_x|\varepsilon_v), \\ &= \rho_m(\varepsilon_x, \varepsilon_y) - \rho_m(\varepsilon_x, \varepsilon_y|\varepsilon_v). \end{aligned} \tag{3.18}$$

When the effect of the frontier of the production set is wanted we will choose large m to have a robust estimate of the frontier (in our empirical application we took $m = 900$ to let the most extreme 1% of data points outside the order-m frontier, we see below how to justify this choice). Then, as suggested in Bădin et al. (2012), we will choose small m for looking to the effect of ε_v on the center of the distribution of the efficiencies; e.g. with $m = 1$

we look to the average production level. Some potential shift already observed for large m (robust frontier estimates) could be enhanced (or reduced) when looking to the difference in the center (with $m = 1$).

In our empirical analysis, we will fix the level of the inputs X_K and X_L jointly to their three quartile values to look to the effect of V on the differences $D_m(\varepsilon_x, \varepsilon_v)$ for three levels of these two factors of production: those with low level of both X_K and X_L, those with joint median level and finally those with joint high level. This is done of course on the pure units for Capital and Labor. To be specific, when fixing these level, we select for each of the three analysis, the data points such that $\varepsilon_x = Median(\varepsilon_x) \pm h_{\varepsilon_x}$, where h_{ε_x} is a bandwidth given by the normal reference rule and ε_x is only considered for Capital and Labor. Then we can analyse the corresponding nonparametric (local linear) regression of $D_m(\varepsilon_x, \varepsilon_v)$ on ε_v, for these subsamples.

We can also provide a marginal analysis, i.e. neglecting the effect of ε_x, and looking to the differences $D_m(\varepsilon_x, \varepsilon_v)$ as a function of ε_v only. Of course by doing so, we may hide some local effect or some interaction between ε_x and ε_v. We will illustrate these approaches in the empirical application below.

4 Empirical Application

4.1 The data and the variables

Our non parametric approach in constructing the worldwide production frontier does not require the specification of the production functional form, and also limit the problem of 'curse of dimensionality' at the second stage of our methodology (due to the use of order-m frontier). In addition, we provide an analysis which is robust to extreme or outlying data points that might hide some features of the production process. We consider the simplest production model with only four macroeconomic variables: aggregate output and three aggregate inputs (labour, capital and human capital).

Human capital is dealt differently from the other two more conventional production factors (capital and labour). Indeed, human capital may affect the production process in two way: it may influence the attainable set, as a production input (Mankiw et al. 1992), or the efficiency distribution as productivity factor (Lucas 1988, Romer 1990a).

The dataset is collected over the period, 1970-2007 (38 years) for a total of 40 countries using data from the Penn; 26 are developed OECD countries (Australia, Austria, Belgium, Canada, Chile, Hong Kong, Denmark, Finland, France, Germany, Greece, Ireland, Israel, Italy, Japan, Korea, Mexico, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Turkey, the United Kingdom and the United States) and 14 are developing countries (Argentina, Bolivia, Côte d'Ivoire, Dominican Republic, Honduras, Jamaica, Kenya, Malawi, Morocco, Philippines, Thailand, Venezuela, Zambia, Zimbabwe).[6]

We use data from the Penn World Tables (version 8) where output is the real gross domestic product $RGDP$ measured in million US dollars at 2005 constant prices. For labor input, we use the number of workers EMP. Capital stock which is our chosen input is then measured in million US dollars at 2005 constant prices. All three variables are transformed in logarithms and then rescaled to get a standard deviation of 1 before estimation.

For human capital we use the variable defined as "index of human capital per person, based on years of schooling (Barro and Lee 2012) and returns to education (Psacharopoulos 1994)". The purpose of this study is to re-exam the growth effect of human capital. We thus acknowledge the endogenous growth models of Lucas (1988) and Romer (1990a) that use a theoretical framework where persistent economic growth is conditional on the accumulation of human capital.[7] The empirical growth literature emphasises the latent heterogeneity and endogeneity of human capital in the economic process, due to unobserved characteristics, which may impact both the level of education and the level of output. This problem leads often to counter-intuitive results due to the difficulty to properly assess the impact of this important economic growth driver on country's output. Indeed, many empirical papers find not significant or even negative impact of human capital on economic growth.[8] To take

[6]The choice of countries depends on data availability, the constraint variable is human capital. Developed and developing countries are classified following the World Bank (2007) classification.

[7]The new endogenous growth theories (Romer 1990a, Aghion and Howitt 1992) describe human capital as the engine of growth through innovation. Grossman and Helpman (1991) show that the skill composition of the labor force matters for the amount of innovation in the economy. In particular, they obtain that an increase in the stock of skilled labor is growth-enhancing while an increase in the stock of unskilled labor can be growth-depressing. Benhabib and Spiegel (1994), Tallman and Wang (1994) find that the channel through which human capital positively affects output is through the efficiency enhancing effect. Recent contributions emphasise the different roles that different types of human capital may play in either backward or advanced economies (Caselli and Coleman, 2006), and the distinction between innovation activities and adoption of existing technologies from the (world) technology frontier (Acemoglu et al., 2006). In this context, low-skilled human capital appears better suited to the adoption of technology in low-income countries, while skilled human capital has a growth enhancing impact which increases with the level of development (Caselli and Coleman 2006, Vandenbussche et al. 2006).

[8]Romer (1990b), Benhabib and Spiegel (1994) and Barro (1997) find a significantly positive effect of schooling levels on output growth, while Cohen and Soto (2001) find no link. Temple (1999) and De La Fuente and Domenech (2001) find a significantly positive correlation between improvements in education and growth, while Barro and Sala-i-Martin (1995), Caselli et al. (1996), and Pritchett (2001) find no effect of schooling improvements on growth. Topel (1999) and Krueger and Lindahl (2001) find both education level and improvement effects on growth.

into account the unobserved heterogeneity and endogeneity of human capital we use life expectancy as instrument. This variable (from World Bank Indicators) is measured as "life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life".

For globalization factor we identify one of the most important channels: FDI inflows, measured as net inflows of foreign direct investment, which are then transformed as a ratio to GDP.[9] This external variable FDI might suffer from endogeneity bias. The endogeneity caused by reverse causality is still an open issue in the empirical studies investigating the relationship between total factor productivity (TFP) and FDI. We explicitly address this issue by eliminating in the first stage the dependence of the FDI on the production process. Furthermore, the global economy becomes increasingly integrated, all the individual countries are likely to be exposed more to global shocks. As explained above we follow Pesaran (2006) and Bai (2009) and consider $F_t = (t, X_{\cdot t}, Y_{\cdot t})$ as proxy for these common factors.

4.2 Effect of human capital on the production process

In our framework we consider that there is an unobserved heterogeneity, as quality of institutions (e.g. different property right systems) and different absorptive capability among countries, which determines the level of innovation and, hence, it affects our production data generation process. We assume that these latent factors are linked to one of our production factor and, specifically, to human capital variable.

To identify this non-observable factor which is related to the level of human capital, we use life expectation as instrument. Indeed, the human capital is linked to the life expectation but there exists a non-observable factor which may also influence the level of human capital in a country as quality of institutions. Therefore, by following Simar et al. (2016) we use life expectancy as an instrument W to identify the latent factor. We calculate the predicted value $\widehat{V}$ for each observation given by equation (3.1) as explained in section 3. We then estimate the conditional output-oriented efficiency, conditional on observed and non-observed factors.

The latent factor $\widehat{V}$ proxies for the aggregate effect of human capital not explained by life expectation - our instrument W -. In our macroeconomic context, this component would be related, for example, to the influence of institutions as difference in property rights systems or absorptive capability, defined as the potential to master new knowledge, among countries. Figure 1 shows the distributions and Table 1 summarizes descriptive statistics for the latent

[9] FDI is sourced from the World Bank World Development Indicators and Unctad, all the other data from PWT 8. The observation period is selected by the data availability.

factor for 40 countries. The highest values over the sample period are registered by USA, New Zealand and Australia, whereas France, Portugal, Italy and Spain stand at the bottom of the ranking. These low values are mainly due to the fact that, in these countries, the life expectancy is very high. We may infer that these countries benefit from the increase in human capital more in terms of highest life expectancy than in better quality of their institutions.

To explore if our predicted latent variable are somehow linked to absorptive capability, defined as the potential to master new knowledge and, hence, to the level of innovation in a country, and to quality of institutions, in Table 3 we report the correlations between $\widehat{V}$ and $R\&D$ (expenditure in R&D as %of GDP) and with 'The Worldwide Governance Indicators (WGI)" from the Wold Bank for the available countries and periods: voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law, and control of corruption.[10] As Table 3 exhibits the correlations are all positive and significant, and unveils that our estimated unobserved factor is significantly positive correlated with R&D expenditure (as % of GDP), and with these governance indicators, especially with Political Stability, Rule of Law and Control of Corruption. This evidence seems to confirm that our latent variable is somehow connected to the level of innovation and quality of institutions. To give a visual impression of the change in latent factor over time, average value for each year is displayed in Figure 2 for United States, New Zeeland, Australia, Morocco, Spain, Belgium and Italy.[11]

To take into account the impact on the production process of observed and non-observed common factors, we follow the two-stage estimation procedure described above which enables

[10] As reported in the World Bank web site www.govindicators.org "The Worldwide Governance Indicators (WGI) project reports aggregate and individual governance indicators for over 200 countries and territories over the period 1996 - 2015, for six dimensions of governance: 1. Voice and accountability (VA), the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and free media 2. Political stability and absence of violence (PV), perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including political violence and terrorism 3. Government effectiveness (GE), the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies 4. Regulatory quality (RQ), the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development 5. Rule of law (RL), the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, the police, and the courts, as well as the likelihood of crime and violence 6. Control of corruption (CC), the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests."

[11] We use the Hodrick and Prescott (1997) filter to smooth the time paths with a smoothing weight equal to 100.

us, in the first stage, to better capture the impacts of global shocks (such as FDI, trade policy and cycle fluctuations) and, hence, CSD on the world production frontier and technical efficiency. By applying the estimation of the models (3.2) to our transformed data, we obtain by equations (3.7) and (3.8) the "pure" versions of our inputs, $\widehat{\epsilon}_x$ (capital, labour and human capital) and of the output $\widehat{\epsilon}_y$ (GDP). Before looking for frontier estimates we have to verify if our assumption of asymptotic independence between the "pure" inputs-output $(\varepsilon_x, \varepsilon_y)$ and the conditioning variables (F_t, Z) is reasonable. Table 2 reports the correlations between $\varepsilon_x, \varepsilon_y$ and time factors $F_t = (t, X_{\cdot t}, Y_{\cdot t})$ and FDI. They are very small indicating that our first stage location-scale model has cleaned most of the effects of these variables and confirming that the influence of FDI and the cross section dependence has been removed from our data.

The estimation of the world production frontier then follows in the second stage. The full frontier estimate is the FDH of the preceding cloud of points and was defined above as $\widehat{\varphi}(e_x|\widehat{\varepsilon}_v)$. For a robust version of the frontier we have to specify a value for m. We compute the order-m frontier for several values of m and look to the corresponding percentage of data points staying above the resulting $\widehat{\varphi}_m(e_x)$ (see e.g. Simar 2003, Daraio and Simar 2007). We know that this percentage decreases when m increases, converging to 0 when $m \to \infty$ since at a moment, for very large value of m we will observe $\widehat{\varphi}_m \equiv \widehat{\varphi}$, i.e. the FDH estimate. This percentage of points outside the frontier (with values $\widehat{\varepsilon}_{y,it} > \widehat{\varphi}_m(e_x|\widehat{\varepsilon}_v)$) as a function of m is shown in Figure 3. We see that as expected for small values of m this percentage decreases rapidly but around values near 900, the value of m has to increase a lot to get the remaining points outside the frontier at this stage, under the frontier. So the points that are outside the frontier for $m = 900$ are rather extreme and may be outlying relative to the rest of the clouds. In fact the "elbow" effect just described is more precisely identified near $m = 900$. So for the robust version of the full frontier, we select $m = 900$ (this leaves 1% of data points above the corresponding order-m frontier).

Table 4 summarizes descriptive statistics for the conditional pure efficiency (2.18) for the 40 countries. We find that the most efficient countries conditioned on latent factor of human capital, and cleaned by the effect of economic cycle and other globalised factor, as FDI, over the sample period, are Germany, Israel and Spain, while the least efficient are Finland, New Zealand and Zimbabwe.

To assess the influence of cleaned latent variable $\widehat{\varepsilon}_v$ defined as "the part of the human capital which is not related to the life expectancy" on the production process, we investigates the differences of conditional and unconditional efficiency measures for full and partial frontier as discussed in Section 3.4. The results of the potential effects of the latent variable $\widehat{\varepsilon}_v$ are shown in Figure 4 for the diffferences (3.18) for $m = 900$ (robust version of the full differences) and for $m = 1$ to assess the influence of $\widehat{\varepsilon}_v$ on the average of the inefficient distribution.

The main messages of these pictures is as follows. To investigate the effect of $\widehat{\varepsilon}_v$ on the shift of the frontier, we have to analyze the differences for robust version (Figure 4 left panel) of the full frontier. We see for the left panel a nonlinear shape with the level of the differences very low and near to zero that, in our setup, we can interpret that we have not shifts of the frontiers when human capital increases. So, $\widehat{\varepsilon}_v$ does not act on the shift of the boundary. Hence, from this evidence, $\widehat{\varepsilon}_v$ appears not to play an important role in accelerating the technological change (shifts in the frontier).

The Figure 4 (right panel) allows to identify some changes in the distribution of the inefficiencies due to $\widehat{\varepsilon}_v$. Globally, we can see some changes in the shape of the clouds of points and we observe a clear decreasing trend of differences with respect to $\widehat{\varepsilon}_v$. The level of these differences is different from zero. Combining this result with the previous on the shift of the technology, we could interpret as the fact that $\widehat{\varepsilon}_v$, which captures absorptive capacity, thus the capability to assimilate innovations, induces catching-up to the production frontier, but not necessarily shifts of it. This seems to suggest that countries benefit from new technology only when they have the ability to exploit it. Countries can switch to a better technology if they accumulate the technology-specific expertise (Greenwood and Jovanovic 2001, Helpman and Rangel 1999).

In addition, since there may be some interaction with the values of the inputs, in Figure 5 we fix the values of Labor and Capital simultaneously at their three univariate quartiles, looking to the effect of human capital on the production process for small, medium and large countries, in term of these inputs.[12] To facilitate the interpretation of the pictures we computed as usual in this kind of analysis the nonparametric regression line of the differences on $\widehat{\varepsilon}_v$. The main messages of these pictures is as follows. To investigate the effect of $\widehat{\varepsilon}_v$ on the shift of the frontier, we have to analyze the differences for the robust version of the full frontier. First, we see for left and right panels an inverted-U shaped of the regression lines for small and medium countries, and a more linear shape for the large sized countries. Second and importantly, the level of the differences is changing with the size of the countries: high level, for the small, degreasing for medium and very low, for large countries (near the

[12]We could provide more pictures with more values and combinations of the inputs but to save space, we limit our analysis to these three combinations.

values zero). In our setup this can be interpreted as follows. We might have some shift of the frontier when human capital increases. So, human capital accumulation acts on the shift of the boundary. Hence, from this evidence, human capital appears to play an important role in accelerating the technological change (shifts in the frontier). This is particularly true for small and medium countries (large values of differences). The bottom panels of Figure 5 deserves also some comments, they allow, when compared to the top panels, to identify some changes in the distribution of the inefficiencies due to $\widehat{\varepsilon}_v$. Globally, we can see changes in the shape of the clouds of points and of the regression lines. Now the level of the differences are similar for small, medium and large countries. So the shift of the technology we have identified above, is compensated by the fact that for small, medium and large size countries, there is much less dispersion, ending up with similar values for the average production levels. This could be interpreted as he fact that human capital induces some shift of the production frontier, but mainly acts as driven factor of technology catching-up process.

Efficiency is the most important growth component for convergence analysis of countries that are below the technological frontier because it reflects "the process of imitation and transmission of existing knowledge" (Romer 1986). Quah (1997), Mankiw et al. (1992), Barro and Sala-i-Martin (1995) argue that slow convergence in the level of output per worker is caused by slow technological catch-up. This latent variable which captures the part of human capital not linked to life expectancy, i.e. might capture absorptive capability, might increase efficiency and, hence, convergence. This occurs with the adoption of foreign technology through technology licensing or technology purchase, imports of high technology capital goods, and the skills acquired by the local labour force (Borensztein et al. 1998, De Mello 1999, Xu 2000).

Our findings support the convergence evolution of output among countries with respect to human capital. Our latent variable which captures the part of human capital not linked to life expectancy, i.e. might measure absorptive capability of a country, thus the capability to assimilate innovations, influences efficiency distribution and it acts as a transmission channel to diffuse technology. Hence, it induces catching-up to the production frontier, but not necessarily shifts of it.

5 Conclusion

The recent economic slowdown first in USA during late nineties and then in Europe in 2001 leads the economists to question the recipe for endogenous self-sustained economic growth.

Economic growth literature emphasizes the importance of human capital in spurring productivity growth. Moreover, the productivity analysis recognizes the importance of considering the spillover effects of global shocks and business cycles due to increasing globalization and interconnection among countries.

So far all studies analysing effect of human capital on productivity of countries have produced ambiguous empirical results due to endogeneity and latent heterogeneity. Many empirical studies have been on the stream of parametric modelling which suffers of misspecification problems when the data generating process is unknown, as usual in the applied studies.

In order to avoid the pitfalls of these parametric cross-regression studies we propose an alternative empirical methodology, robust frontier in non parametric location-scale models for accommodating simultaneously the problem of model specification uncertainty, potential endogeneity and cross-section dependence in modelling technical efficiency in frontier models.

The neglect of a latent factor, or unobserved global factors (hence cross section dependence) or an observed one will cause an endogeneity problem (Simar et al. 2016). We combine two different methodology, the two-step procedure advanced by Florens et al. (2014), which enables us to deal with both observed heterogeneity and cross section dependence by combining location scale model and conditional efficiency estimation and Simar et al. (2016) which handles the endogeneity due to unobserved heterogeneity.

Our non parametric approach to estimate conditional efficiency does not require any parametric assumption regarding technology or efficiency term. Moreover, the assumption of complete homogeneity of considered units is not needed. Therefore the economic units under investigation, can potentially consist of different groups of population governed by different distributional laws of the generation of input-output mix and on efficiency. This is an advantage in our sample formed by developed and developing countries which most likely have different distributions of efficiency scores.

Moreover our frontier model enables us to see whether the effect of environmental/global variables on productivity occurs via technology change or efficiency. We can then quantify the impact of environmental/global factors on efficiency levels and make inferences about the contributions of these variables in affecting efficiency.

The "new" growth theory of Romer (1990a), Lucas (1988) and Barro (1997) has human capital playing an important role in growth because human capital can help in explaining an economy's capacity to absorb new technologies (Abromovitz 1986, Cohen and Levinthal 1989, Kneller 2005, Kneller and Stevens 2006, Mastromarco and Ghosh 2009). Our paper extends previous studies on similar topics by investigating this channel in full nonparametric framework which avoids some restrictive and often unverifiable prior assumptions on functional relationships and distributions.

We focus on the effect of human capital on economic performance of 40 countries over the period 1970-2007. In a cross-country framework, production inefficiencies can be identified as the distance of the individual country's production from the frontier as proxied by the maximum output of the reference country (regarded as an empirical counterpart of an optimal production boundary). Hence, efficiency improvement will represent productivity catch-up via technology diffusion because inefficiencies generally reflect a sluggish adoption of new technologies (Ahn and Sickles 2000).

Our findings prove that human capital plays an important role in accelerating the technological catch-up (increase in the efficiency) but not on the technological changes (shifts in the frontier). This result seems to confirm the theoretical hypothesis that countries benefit from new technology (technological catch-up) only when they have the ability to exploit it, hence only when they have high level of absorptive capability. Countries can switch to a better technology if they accumulate the technology-specific expertise (Greenwood and Jovanovic 2001, Helpman and Rangel 1999).

References

Abromovitz, M.: 1986, Catching up, forging ahead, and falling behind, *Journal of Economic History* **46**, 386–406.

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

Ahn, S. C. and Sickles, R. C.: 2000, Estimation of long-run inefficiency levels: A dynamic frontier approach, *Econometric Reviews* **19**, 461 – 492.

Bai, J.: 2009, Panel data models with interactive fixed effects, *Econometrica* **77**, 1229–1279.

Barro, R.: 1990, Government spending in a simple model of endogenous growth, *Journal of Political Economy* **98**, S103–S125.

Barro, R. J.: 1997, *Determinants of Economic Growth*, The MIT Press, Cambridge, MA.

Barro, R. J. and Lee, J.: 2012, A new data set of educational attainment in the world, 1950-2010, *Journal of Development Economics* **104**, 184–198.

Barro, R. J. and Sala-i-Martin, X.: 1995, *Economic Growth*, McGraw-Hill, New York.

Benhabib, J. and Spiegel, M.: 1994, The role of human capital in economic development. Evidence from aggregate cross-country data, *Journal of Monetary Economics* **34**, 143–173.

Borensztein, E., De Gregorio, J. and Lee, L.-W.: 1998, How does foreign direct investment affect economic growth?, *Journal of International Economics* **45**, 115–135.

Bădin, L., Daraio, C. and Simar, L.: 2012, How to measure the impact of environmental factors in a nonparametric production model?, *European Journal of Operational Research* **223**, 818–833.

Bădin, L., Daraio, C. and Simar, L.: 2017, A bootstrap based approach for bandwidth selection in estimating conditional efficiency measures. Working Paper, Institut de Statistique Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Caselli, F. and Coleman, W. J.: 2006, The world technology frontier, *American Economic Review* **96**, 499–522.

Caselli, F., Esquivel, G. and Lefort, F.: 1996, Reopening the convergence debate: A new look at cross-country growth empirics, *Journal of Economic Growth* **1**, 363–389.

Cazals, C., Fève, F., Florens, J. and Simar, L.: 2016, Nonparametric instrumental variables estimation for efficiency frontier, *Journal of Econometrics* **190**, 349–358.

Cazals, C., Florens, J. and Simar, L.: 2002, Nonparametric frontier estimation: A robust approach, *Journal of Econometrics* **106**, 1–25.

Cohen, W. and Levinthal, D.: 1989, Innovation and learning: Two faces of R&D, *Economic Journal* **107**, 139–149.

Daouia, A., Florens, J. and Simar, L.: 2010, Frontier estimation and extreme value theory, *Bernoulli* **16**, 1039–1063.

Daouia, A., Florens, J. and Simar, L.: 2012, Regularization of non-parametric frontier estimators, *Journal of Econometrics* **168**, 285–299.

Daouia, A. and Gijbels, I.: 2011, Robustness and inference in nonparametric partial-frontier modeling, *Journal of Econometrics* **161**, 147–165.

Daraio, C. and Simar, L.: 2005, Introducing environmental variables in nonparametric frontier models: A probabilistic approach, *Journal of Productivity Analysis* **24**, 93–121.

Daraio, C. and Simar, L.: 2007, *Advanced Robust and Nonparametric Methods in Efficiency Analysis. Methodology and Applications*, Springer, New York.

Daraio, C. and Simar, L.: 2014, Directional distances and their robust versions: Computational and testing issues, *European Journal of Operational Research* **237**, 358–369.

De La Fuente, A. and Domenech, R.: 2001, Schooling data, technical diffusion, and the neoclassical model, *American Economic Review Papers and Proceedings* **90**, 323–327.

De Mello, L. R. J.: 1999, Foreign direct investment-led growth: Evidence from time series and panel data, *Oxford Economic Papers* **51**, 133–151.

Deprins, D., Simar, L. and Tulkens, H.: 1984, Measuring labor inefficiency in post offices, *in* M. Marchand, P. Pestieau and H. Tulkens (eds), *The Performance of Public Enterprises: Concepts and measurements*, North-Holland, pp. 243–267.

Färe, R. and Grosskopf, S.: 2000, Theory and application of directional distance functions, *Journal of Productivity Analysis* **13**, 93–103.

Florens, J., Simar, L. and van Keilegom, I.: 2014, Frontier estimation in nonparametric location-scale models, *Journal of Econometrics* **178**, 456–470.

Glass, A. J. and Saggi, K.: 1998, International technological transfer and technology gap, *Journal of Development Economics* **55**, 369–398.

Greenwood, J. and Jovanovic, B.: 2001, Accounting for growth, *in* T. U. o. C. P. NBER (ed.), *New Directions in Productivity Analysis*, NBER, pp. 179–222.

Grossman, G. and Helpman, E.: 1991, Quality ladders in the theory of growth, *Review of Economic Studies* **58**, 43–61.

Helpman, E. and Rangel, A.: 1999, Adjusting to a new technology: Experience and training,, *Journal of Economic Growth* **4**, 359–383.

Hodrick, R. and Prescott, E. C.: 1997, Postwar u.s. business cycles: An empirical investigation, *Journal of Money, Credit and Banking* **29**, 1–16.

Javorcik, B. S.: 2004, Does foreign direct investment increase the productivity of domestic firms? in search of spillovers through backward linkages, *American Economic Review* **94**(4), 605–627.

Kneller, R.: 2005, Frontier technology, absorptive capacity and distance, *Oxford Bulletin of Economics and Statistics* **67**, 1–24.

Kneller, R. and Stevens, P. A.: 2006, Frontier technology and absorptive capacity: Evidence from OECD manufacturing industries, *Oxford Bulletin of Economics and Statistics* **68**, 1–21.

Krueger, A. and Lindahl, M.: 2001, Education for growth: Why and for whom?, *Journal of Economic Literature* **39**, 1101–1136.

Li, Q., Lin, J. and Racine, J.: 2013, Optimal bandwidth selection for nonparametric conditional distribution and quantile functions, *Journal of Business and Economic Statistics* **31**, 57–65.

Lucas, R. E.: 1988, On the mechanics of economic development, *Journal of Monetary Economics* **22**, 3–42.

Mankiw, G., Romer, D. and Weil, D.: 1992, A contribution to the empirics of economic growth, *Quarterly Journal of Economics* **107**, 407–437.

Mastromarco, C. and Ghosh, S.: 2009, Foreign capital, human capital, and efficiency: A stochastic frontier analysis for developing countries, *World Development* **37**, 489 – 502.

Mastromarco, C. and Simar, L.: 2015, Effect of fdi and time on catching-up: New insights from a conditional nonparametric frontier analysis, *Journal of Applied Econometrics* **30**, 826–847.

Matzkin, R.: 2003, Nonparametric estimation of nonadditive random functions, *Econometrica* **71**, 1339–1375.

Miller, S. M. and Upadhyay, M. P.: 2000, The effects of openness, trade orientation, and human capital on total factor productivity, *Journal of Development Economics* **63**, 399–423.

Miller, S. M. and Upadhyay, M. P.: 2002, Total factor productivity, human capital and outward orientation: Differences by stage of development and geographic regions. Department of Economics Working Paper Number 2002-03, University of Connecticut.

Olofsdotter, K.: 1998, Foreign direct investment, country capabilities and economic growth, *Weltwirtschaftliches Archiv* **134**, 534–547.

Park, B., Simar, L. and Weiner, C.: 2000, The FDH estimator for productivity efficiency scores: Asymptotic properties, *Econometric Theory* **16**, 855877.

Pesaran, M. H.: 2006, Estimation and inference in large heterogeneous panels with a multi-factor error structure, *Econometrica* **74 (4)**, 967–1012.

Pritchett, L.: 2001, Where has all the education gone?, *World Bank Economic Review* pp. 367–391.

Psacharopoulos, G.: 1994, Returns to investment in education: A global update., *World Development* **22**, 1325–1343.

Quah, D. T.: 1997, Empirics for growth and distribution: Stratification, polarization, and convergence clubs, *Journal of Economic Growth* **2**, 27–59.

Rodriguez, F.: 2006, Cleaning up the kitchen sink: On the consequences of the linearity assumption for cross-country growth empirics. Mimeo, Wesleyan University, Middletown, CT.

Romer, P. M.: 1986, Increasing returns and long-run growth, *Journal of Political Economy* **94**, 1002–1037.

Romer, P. M.: 1990a, Endogenous technological change, *Journal of Political Economy* **98**, S71–S102.

Romer, P. M.: 1990b, Human capital and growth: Theory and evidence, *Carnegie-Rochester Conference Series on Public Policy* **32**, 251–286.

Simar, L.: 2003, Detecting outliers in frontier models: a simple approach, *Journal of Productivity Analysis* **20**, 391–424.

Simar, L. and Vanhems, A.: 2012, Probabilistic characterization of directional distances and their robust versions, *Journal of Econometrics* **166**, 342–354.

Simar, L., Vanhems, A. and van Keilegom, I.: 2016, Unobserved heterogeneity and endogeneity in nonparametric frontier estimation, *Journal of Econometrics* **190**, 360–373.

Simar, L. and Wilson, P. W.: 2007, Estimation and inference in two-stage, semi-parametric models of production processes, *Journal of Econometrics* **136**, 31–64.

Tallman, E. and Wang, P.: 1994, Human capital and endogenous growth: Evidence from taiwan, *Journal of Monetary Economics* **34**, 101–124.

Temple, J.: 1999, The new growth evidence, *Journal of Economic Literature* **XXXVII**, 112–156.

Topel, R. H.: 1999, Labor markets and economic growth, *in* O. Ashenfelter and D. Card (eds), *The Handbook of Labor Economics*, Elsevier, The Netherlands: North Holland, Amsterdam, pp. 2943–2984.

Vandenbussche, J., Aghion, P. and Meghir, C.: 2006, Growth, distance to frontier and composition of human capital, *Journal of Economic Growth* **11**, 97–127.

Xu, B.: 2000, Multinational enterprises, technology diffusion, and host country productivity growth, *Journal of Development Economics* **62**, 477–493.

Country	Mean	Std. Dev.	Change (%)
USA	0.969	0.021	0.001
NZL	0.953	0.028	0.001
AUS	0.890	0.020	-0.001
BOL	0.887	0.087	0.011
KOR	0.861	0.065	0.004
IRL	0.827	0.023	-0.001
PHL	0.774	0.084	0.009
ISR	0.765	0.041	-0.003
CAN	0.739	0.066	0.005
ZWE	0.703	0.227	0.015
NOR	0.694	0.171	0.018
ARG	0.690	0.077	-0.001
ZMB	0.662	0.188	0.039
KEN	0.661	0.203	0.049
DNK	0.629	0.067	-0.004
NLD	0.598	0.075	0.005
BEL	0.592	0.039	-0.008
SWE	0.582	0.110	0.024
JAM	0.558	0.258	0.034
CHL	0.537	0.125	-0.009
JPN	0.511	0.105	0.007
DOM	0.456	0.088	0.009
FIN	0.429	0.110	-0.014
TUR	0.396	0.102	-0.014
CIV	0.341	0.240	0.073
AUT	0.319	0.124	-0.024
MWI	0.315	0.169	0.056
HND	0.312	0.089	0.009
MEX	0.309	0.120	0.038
GER	0.297	0.287	0.103
HKG	0.290	0.078	0.005
GBR	0.285	0.125	-0.032
GRC	0.273	0.081	-0.001
THA	0.232	0.068	0.017
VEN	0.188	0.129	-0.001
FRA	0.141	0.116	0.126
PRT	0.111	0.045	-0.057
ITA	0.103	0.043	-0.013
ESP	0.084	0.112	0.104
MAR	0.050	0.025	0.025

Table 1: *Estimated latent factor $\widehat{V}$ of 40 countries over 1970 till 2007: mean and standard deviation over time and change in % from 1970 to 2007.*

	ε_L	ε_K	ε_H	ε_Y	ε_V
Pearson correlations					
t	-0.0115	-0.0522	-0.0470	-0.0521	-0.0162
$L_{\cdot t}$	-0.0279	0.0142	-0.0026	-0.0063	0.0053
$K_{\cdot t}$	-0.0369	0.0212	-0.0171	0.0089	-0.0023
$H_{\cdot t}$	-0.0096	0.0086	-0.0067	0.0112	-0.0023
$Y_{\cdot t}$	-0.0353	0.0187	-0.0142	0.0104	-0.0023
FDI	0.0081	-0.0052	-0.0251	0.0066	0.0244
Spearman rank correlations					
t	-0.0127	-0.0690	-0.0476	-0.0533	-0.0188
$L_{\cdot t}$	-0.0278	0.0229	0.0070	-0.0095	0.0107
$K_{\cdot t}$	-0.0454	0.0308	-0.0193	0.0117	0.0046
$H_{\cdot t}$	-0.0132	0.0119	-0.0121	0.0131	0.0038
$Y_{\cdot t}$	-0.0379	0.0302	-0.0154	0.0140	0.0030
FDI	0.0273	-0.0168	-0.0229	0.0223	0.0262
Kendall correlations					
t	-0.0082	-0.0452	-0.0320	-0.0348	-0.0121
$L_{\cdot t}$	-0.0192	0.0151	0.0049	-0.0064	0.0065
$K_{\cdot t}$	-0.0316	0.0204	-0.0126	0.0081	0.0036
$H_{\cdot t}$	-0.0090	0.0072	-0.0080	0.0090	0.0027
$Y_{\cdot t}$	-0.0266	0.0202	-0.0101	0.0096	0.0024
FDI	0.0185	-0.0104	-0.0156	0.0146	0.0178

Table 2: *Correlation between* ε_L, ε_K, ε_H, ε_Y, ε_V *and the factor* $F_t = t, L_{\cdot t}, K_{\cdot t}, H_{\cdot t}, Y_{\cdot t}$ *and* FDI

	R&D	*Acc.*	*Pol.Stab.*	*Gov.Effect.*	*Reg.Qual.*	*RuleofLaw*	*ControlofCorr.*
Pearson correlations							
$\hat{V}$	0.2916	0.1642	0.2061	0.1887	0.1288	0.2008	0.1774
	(0.0000)	(0.0001)	(0.0013)	(0.0005)	(0.0123)	(0.0000)	(0.0004)
Spearman rank correlations							
$\hat{V}$	0.3540	0.2432	0.2007	0.2266	0.1569	0.2781	0.2295
	(0.0000)	(0.0001)	(0.0013)	(0.0005)	(0.0123)	(0.0000)	(0.0004)
Kendall correlations							
$\hat{V}$	0.2312	0.1575	0.1273	0.1379	0.0993	0.1776	0.1392
	(0.0000)	(0.0001)	(0.0013)	(0.0005)	(0.0123)	(0.0000)	(0.0004)

Table 3: *Correlation between* $\hat{V}$ *and R&D, the government quality indicators = Accountability, Political Stability, Government Effectiveness,s Regulatory Quality, Rule of Law, Control of Corruption. P-values for testing the hypothesis of no correlation in parentheses.*

Country	Pure Cond. Mean Eff.	STD Pure Cond Eff.	Change (%) in Pure Cond. Eff.
GER	1.0168	0.0362	0.0198
ISR	1.0241	0.0387	-0.0416
ESP	1.0235	0.0216	-0.0276
FRA	1.0250	0.0214	0.0041
SWE	1.0251	0.0366	-0.0009
CIV	1.0276	0.0280	0.0000
DNK	1.0279	0.0216	0.0180
ARG	1.0281	0.0248	-0.0407
CAN	1.0297	0.0243	-0.0576
KEN	1.0321	0.0441	0.0588
GRC	1.0323	0.0188	-0.0420
PRT	1.0327	0.0290	-0.0014
NLD	1.0329	0.0249	0.0058
MEX	1.0329	0.0363	0.0154
HND	1.0331	0.0317	0.0436
PHL	1.0323	0.0273	-0.0647
MAR	1.0345	0.0389	0.0977
CHL	1.0346	0.0374	0.0115
ZWE	1.0347	0.0350	-0.0055
GBR	1.0307	0.0364	-0.0143
TUR	1.0349	0.0285	0.0278
DOM	1.0316	0.0317	-0.0469
BEL	1.0363	0.0218	-0.0499
ITA	1.0367	0.0279	0.0195
USA	1.0375	0.0388	0.0278
IRL	1.0400	0.0391	0.0001
THA	1.0409	0.0287	-0.0211
AUS	1.0411	0.0310	-0.0617
MWI	1.0432	0.0397	0.0772
NOR	1.0381	0.0354	-0.0186
JAM	1.0439	0.0378	-0.0624
HKG	1.0452	0.0258	0.0622
VEN	1.0445	0.0299	-0.0226
KOR	1.0460	0.0258	-0.0604
JPN	1.0480	0.0390	0.0287
BOL	1.0495	0.0314	-0.0825
AUT	1.0502	0.0251	0.0184
FIN	1.0470	0.0308	0.0042
NZL	1.0500	0.0406	0.0622
ZMB	1.0584	0.0330	0.0650

Table 4: *Note: Mean, standard deviation and percentage change of pure conditional mean efficiency; see equation* (2.18).

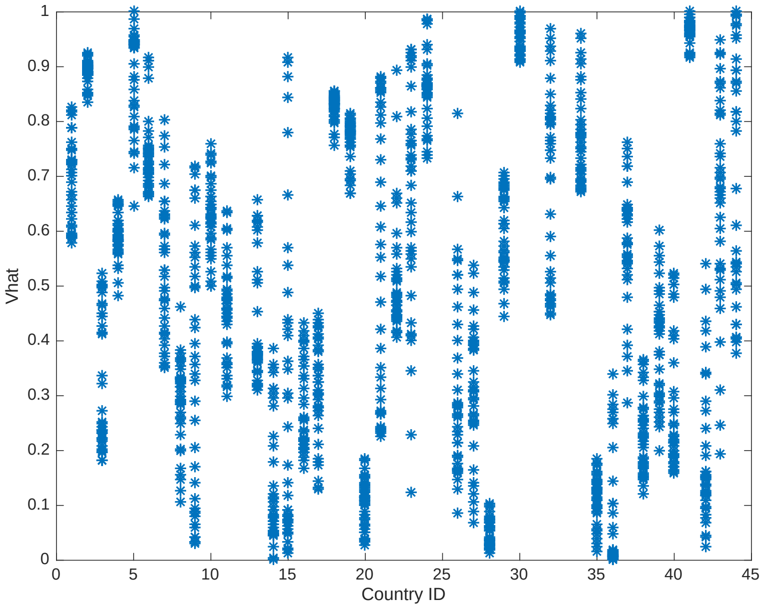

Figure 1: *Distribution of latent factor* $\widehat{V}$ *for the countries under analysis.*

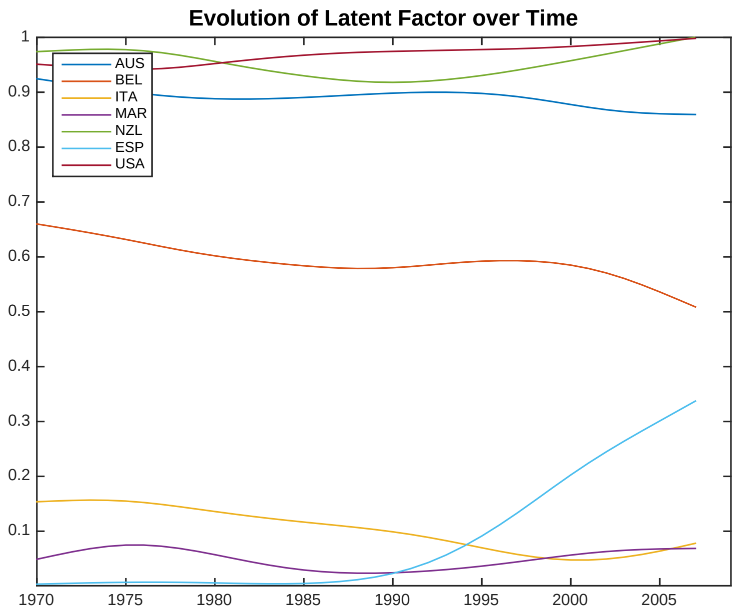

Figure 2: *Evolution over time of latent factor* $\widehat{V}$ *for Australia (AUS), Belgium (BEL), Italy (ITA), Morocco (MAR), New Zealand (NZL), Spain (ESP) and United States (USA). Note also that we use the Hodrick and Prescott (1997) filter to smooth the time paths with a smoothing weight equal to 100.*

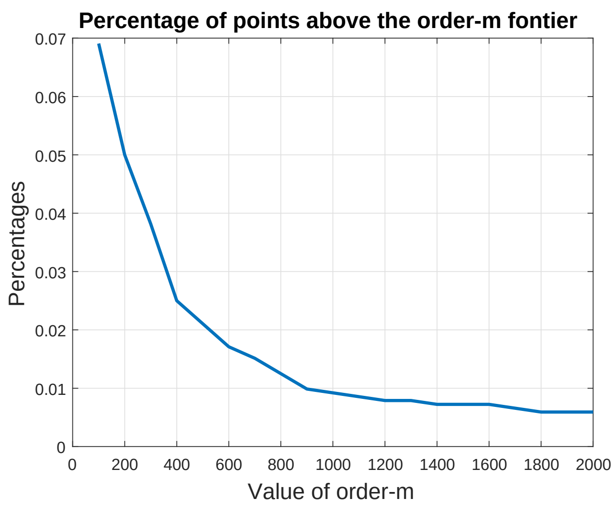

Figure 3: *Percent of points outside the m-frontier at each value of m.*

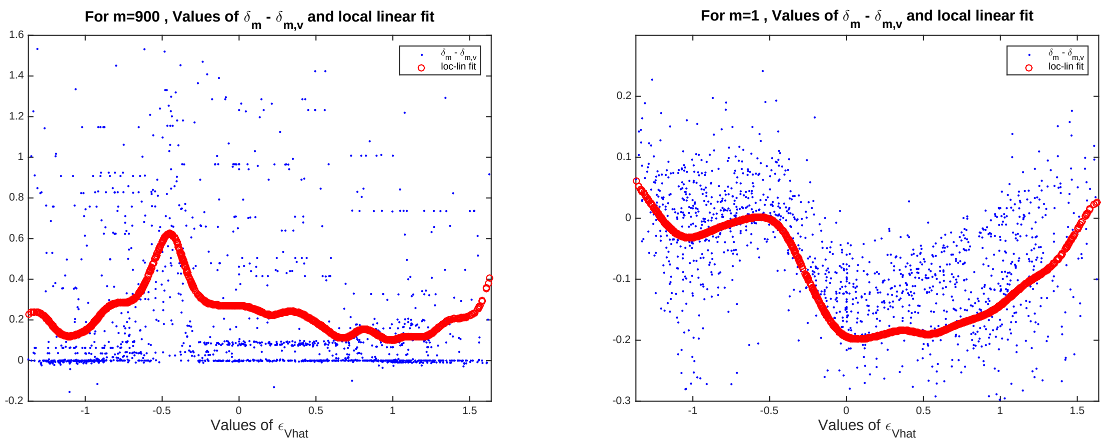

Figure 4: *Estimated differences of marginal and conditional efficiency of order-*$m = 900$ *frontier (left panel) and order-*$m = 1$ *frontier (right panel). The sample size n=1520*

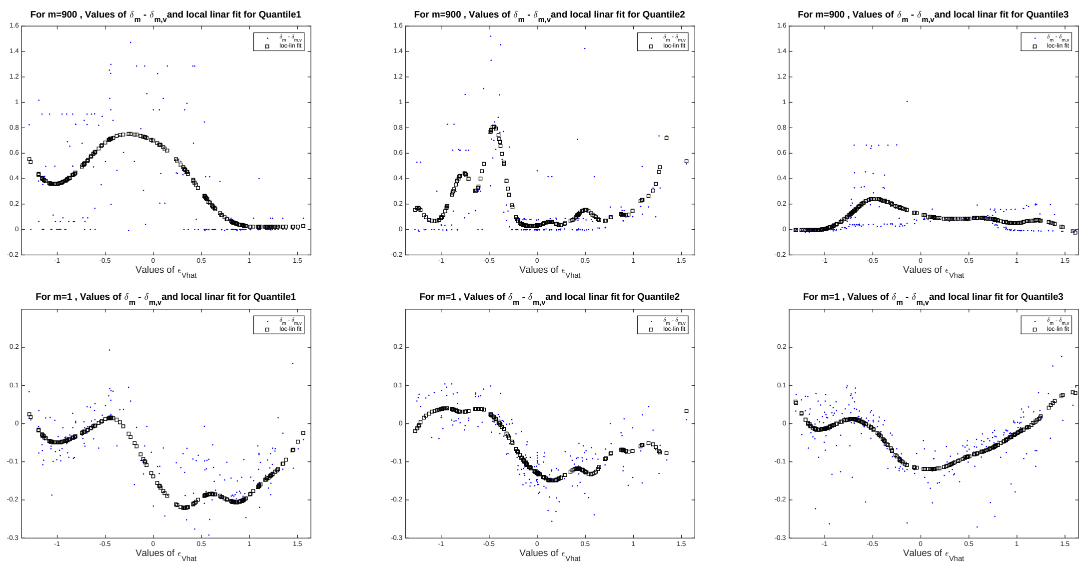

Figure 5: *The first three top panels represent the estimated differences of marginal and conditional efficiency of order-*$m = 900$ *as a function of* $\widehat{\varepsilon}_V$ *with the two inputs (labour and capital) fixed at their three quartiles, from left to right,* $Q_{0.25}, Q_{0.50}, Q_{0.75}$*;, the bottom panels are for* $m = 1$.