Sorting, Matching and Economic Complexity

Muhammed A. Yildirim

CID Faculty Working Paper No. 392
March 2021

Working Papers
Center for International Development
at Harvard University

Sorting, Matching and Economic Complexity[*]

Muhammed A. Yıldırım
Koç University & Growth Lab, CID at Harvard University

March 1, 2021

Abstract

Assignment models in trade predict that countries with higher productivity levels are assortatively matched to industries that make better use of these higher levels. Here, we assume that the driver of productivity differences is the differential distribution of factors among countries. Utilizing such a structure, we define and estimate the average factor level (AFL) for countries and products using only the information about the production patterns. Interestingly, our estimates coincide with the complexity variables of (Hidalgo and Hausmann, 2009), providing an underlying economic rationale. We show that AFL is highly correlated with country-level characteristics and predictive of future economic growth.

JEL Codes: F10, F11, F14, O41, O47, O50.
Keywords: International Trade, Supermodularity, Ricardian Model, Assignment Models, Sorting, Complexity, Economic Complexity.

[*]We would like to thank Daniel P. Stock, Sebastian Bustos and Ricardo Hausmann for their invaluable comments. We thank the members of the Growth Lab at Harvard University for their feedback.

1 Introduction

Countries differ greatly in terms of their productivity levels (Lucas, 1988). According to the Ricardian theory of trade, the productivity levels (i.e., the absolute advantage) drives the income differences between nations. Here, we assume that the productivity differences are driven by differential distribution of production factors across countries. We can think of the factors as individuals at different skill levels. Using an assignment model in trade, we show that an average factor level present in a country could be estimated through ability of a country making different products with comparative advantage. Interestingly, the resultant productivity estimation matches the Economic Complexity Index (ECI) introduced in Hidalgo and Hausmann (2009) and Hausmann et al. (2014). Similar to ECI, the productivity levels we estimate exhibit high levels of correlations with country-level measures such as GDP per capita, human capital and institutional measures. Furthermore, the productivity level estimates are also predictive of a country's future growth.

Our work builds on Costinot (2009), where the comparative advantage patterns of countries are linked to supermodularity in a multi-factor environment. The supermodularity implies that higher-indexed factors are relatively more productive at higher-indexed industries compared to a lower-indexed factor. For instance, the productivity gap between an engineer and an unskilled worker would be much bigger in the computer industry than it would be in the construction industry. The supermodularity enforces that higher-indexed factors would be assortatively matched to higher-indexed industries. Since the countries differ in terms of their availability of factors, a country's production pattern would be shaped by the distribution of factors present within the country. In particular, we assume that production factors are distributed normally in each country. Consequently, we can write a maximum likelihood function to estimate average factor levels for countries and matching factor levels for products. The Average Factor Level (AFL) of a country captures its productivity level. Although the Ricardian theory predicts that productivity determines the level of income, this relationship is not perfect in practice. For instance, endowments of natural resource products explain income levels that are not aligned with the underlying productivity in a country. On the other hand, countries that increase their average productivity often see increases in their income.

In the empirical assessment of the model we show that indeed country productivity levels are correlated with current levels of income, and the residual from this relationship predicts the subsequent economic growth. This finding is scrutinized using a wide range of different controls and robustness checks, and in all cases the relationship between the average factor level of a country and its economic growth survives.

The first applications of supermodular functions in economics, which lead to assortative matching of individuals to jobs, have been studied in labor markets to understand the distribution of income and emerging inequality resulting from complementarities.[1] Inspired by this literature, sorting and assortative matching have been studied in international trade. Grossman and Maggi (2000) introduce complementarity into the tasks of production to relate talent distribution of a country results to its specialization pattern. Antràs et al. (2006) study the matching of tasks and skills in the context of offshoring globalization. Ohnsorge and Trefler (2007) explore the effects of two-dimensional worker-skill heterogeneity across countries and its consequences on industrial structure, international trade and income distribution. We build on the model introduced in Costinot (2009) and Costinot and Vogel (2015), whereas Costinot and Vogel (2010) use this structure to investigate the implications of assortative matching on income inequality between countries. Grossman et al. (2017) study the complementarity between factors of production which results in a rich relationship between income distribution and international trade. Other than assignment models, the capability approach by Sutton and Trefler (2016) gives rise to an assortative matching between countries and industries.

Country- and product-level measures have been used by policymakers extensively. In particular, our measure is intimately linked with the economic complexity variables, ECI and PCI (Hidalgo and Hausmann, 2009; Hausmann et al., 2014).[2] These variables are calculated without explicit microfoundations, however; instead, they use an iterative correction algorithm that is shown to be equivalent to an eigenvector problem. In the approach introduced in this paper, the proposed economic framework arrives to the same eigenvector equation albeit with an underlying economic model. Hence, our paper is the first micro-founded model

[1] See Roy (1951), Becker (1973), Heckman and Sedlacek (1985), Heckman and Honoré (1990), Borjas (1987), Teulings (1995, 2005), Eeckhout and Kircher (2010a,b, 2011, 2018), Abowd et al. (1999) and Card et al. (2013) for important studies in this strand of literature.

[2] See Hidalgo (2021) for a recent review which also highlights policy applications.

that organically gives rise to these complexity measures. The success of elusive ECI and PCI measures in explaining country and product sophistication grabbed attention of researchers to interpret what these measures really capture. There have been previous efforts to give mechanical or mathematical interpretations to the ECI algorithm. For instance, Mealy et al. (2019) show that the algorithm leading to ECI is equivalent to a celebrated spectral clustering algorithm by Shi and Malik (2000). Schetter (2019) shows a link between supermodularity and the eigenvector equation for ECI, establishing a structural ranking of economic complexity. Here, our approach directly links these measures to the underlying distribution of production factors. Recently, additional measures inspired by the economic complexity variables that incorporate only the supply-side limitations into the model have emerged (Bustos and Yildirim, forthcoming).

The rest of the paper is organized as follows. In the next section, we introduce the model and develop our estimations. In Section 3, we show empirical results at the country level and relate our measure to other country characteristics. In Section 4, we conclude.

2 Model

We follow the model by Costinot (2009). We index countries by $c \in \{1, \ldots, N_C\}$ and industries by $i \in \{1, \ldots, N_I\}$. We assume there is a continuum of factors, indexed by $f \in F \subset \mathbb{R}$, employed in production. Productivity of factor f in industry i is given by $A(i, f) > 0$. Each worker is matched with a single factor and the total labor endowment of country c is given by L_c. Output of industry i in country c is:

$$Y_{c,i} = \int_F A(i, f) L(c, i, f) df \tag{1}$$

where $L(c, i, f)$ denotes the amount of factor f used in industry i in country c. We assume no trade costs between countries, hence, the global price of good i is p_i. The cost of producing good i is only the wage compensation to workers with different factor levels. The wage of a worker in country c with factor f is $w(c, f)$. We denote factors employed in industry i by $F_i \subset F$. Assuming perfect competition, the marginal cost of production is equal to the price.

Hence, for all factors employed in industry i we obtain the following relationship:

$$p_i = w(c,f)/A(i,f), \qquad \forall f \in F_i \tag{2}$$

The positiveness of function A guarantees that all production factors f would be employed in some industry by accepting an appropriate wage. From this simple setup, we arrive at the following proposition regarding wage equalization whose proof follows directly from Equation 2:

Proposition 1 *Given that there are no technology differences between countries and equal prices for each goods, the wage associated with each factor is the same for all countries, i.e.,* $w(c,f) = w(f)$ *for all* $c \in \{1,\ldots,C\}$.

We assume that the productivity function $A(i,f)$ is log-supermodular. Mathematically, this property could be stated as:

$$\frac{A(i',f')}{A(i,f')} \geq \frac{A(i',f)}{A(i,f)}, \quad \text{for all } i' \geq i \text{ and } f' \geq f. \tag{3}$$

This property implies that factors with higher indices, f', are relatively more productive in industries with higher indices, i'. Given the factor price equalization and the log-supermodularity, we can now analyze which factors are employed in industry i. In the lemma below, we show that factors are assigned to industries in an assortative manner, i.e., higher-indexed factors are assigned to higher-indexed industries:

Lemma 1 *Suppose* $i' > i$. *Then any factor* $f' \in F_{i'}$ *is greater than or equal to* $\forall f \in F_i$.

Proof. The proof is by the means of contradiction. Suppose that $f' < f$. Since f' is assigned to industry i', we know that its wage in that industry is higher than what it would have received in industry i:

$$p_i A(i,f') < p_{i'} A(i',f') = w(f').$$

The opposite is true for factor f:

$$p_{i'} A(i',f) < p_i A(i,f) = w(f).$$

By multiplying these equations:

$$A(i,f')A(i',f) < A(i,f)A(i',f')$$

which contradicts with the supermodularity property because $i < i'$ and $f' < f$. ■

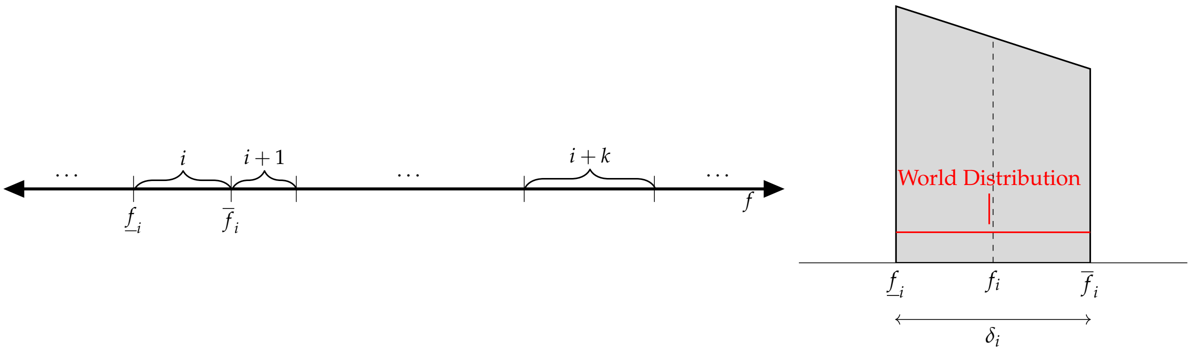
A corollary of this lemma is that factors are allocated to industries in continuous chunks:

Corollary 1 *Suppose there are two factors, $\{\underline{f}, \overline{f}\} \subset F_i$ with $\underline{f} < \overline{f}$. Then, any factor $\underline{f} < f < \overline{f}$ will also be assigned to produce in industry i, i.e., $f \in F_i$.*

Proof. Suppose f' is employed in industry $i' > i$. However, $\overline{f} > f$ is employed in industry i, which contradicts with Lemma 1. If f' is employed in industry $i' < i$, employment of $\underline{f} < f$ in industry i gives us a contradiction. ■

Combining the above lemma and its corollary, production factors are assigned to industries in an increasing and continuous fashion. Therefore, for an industry i, $\exists$ a lowest-indexed factor $\underline{f}_i$ and a highest-indexed factor $\overline{f}_i$ such that $F_i = \left[\underline{f}_i, \overline{f}_i\right)$.[3]

Figure 1: Factors and Industries



(a) Allocation of factors to industries

(b) Approximations

NOTES: (a) With the supermodularity assumption, factors can be assigned to industries assortatively. (b) This figure shows the expected output of a country in a product (area under the black line) and the world output (area under the red line). If δ_i is small, these areas could be approximated by an area of the trapezoid.

The equilibrium can be found given any Walrasian demand structure. As long as demand for any industry is non-zero, our results will hold. The demand determines $\underline{f}_i$ and $\overline{f}_i$ for each industry i.

[3] Technically, $\underline{f}_i$ and $\overline{f}_i$ could be employed in two industries, but we will consider them as exceptions.

The country heterogeneity is driven by the differential distribution of factors. Workers in country c are randomly assigned a factor from the distribution $g_c \sim N(\mu_c, \sigma^2)$ with μ_c specifying the average factor level of the country. We take σ to be common for all countries to simplify the estimation.[4] We can write the expected output of country c in industry i as:

$$Y_{c,i} = L_c \int_{\underline{f}_i}^{\overline{f}_i} w(f) g_c(f) df = L_c p_i \int_{\underline{f}_i}^{\overline{f}_i} A(i,f) g_c(f) df. \tag{4}$$

We assume that μ_c levels themselves are realizations from an underlying normal distribution g_μ with $\mu_c \sim N(0, \sigma_\mu^2)$. Hence, the world distribution of factors follows $g_W \sim N(0, \sigma_\mu^2 + \sigma^2)$. Consequently, the expected world output in industry i is:

$$Y_i = L p_i \int_{\underline{f}_i}^{\overline{f}_i} A(i,f) g_W(f) df \tag{5}$$

where L is the population of the world.

Let's define $f_i \equiv (\underline{f}_i + \overline{f}_i)/2$ and $\delta_i \equiv \overline{f}_i - \underline{f}_i$. Assuming a small δ_i and $A(i,f) \approx A(i,f_i)$ within this interval, we can approximate Equation 4 with the area of a trapezoid as shown in Figure 1b:[5]

$$Y_{c,i} \approx L_c p_i A(i,f_i) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(f_i - \mu_c)^2}{2\sigma^2}\right] \delta_i \tag{6}$$

The distribution of countries have a larger variance than the factor distribution of individual countries (i.e., $\sigma_\mu >> \sigma$). Hence, the world output in industry i is approximately:

$$Y_i \approx L p_i A(i,f_i) \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{f_i^2}{2\sigma_\mu^2}\right] \delta_i. \tag{7}$$

The share of country c in industry i is:

$$\frac{Y_{c,i}}{Y_i} \approx \frac{L_c}{L} \frac{\sigma_\mu}{\sigma} \exp\left[-\frac{(f_i - \mu_c)^2}{2\sigma^2}\right] \tag{8}$$

By dividing the share of country c in industry i relative to its population share in the world we

[4] Heterogeneity in this measure could be incorporated into the model as an extension.

[5] This approximation would fail if $\mu_c \in [\underline{f}_i, \overline{f}_i]$. In that case, we can put a lower bound on the area using $\underline{f}_i$ if $\mu_c \in [f_i, \overline{f}_i]$ or $\overline{f}_i$ if $\mu_c \in [\underline{f}_i], f_i$ instead of f_i.

arrive at a measure is known as Revealed per Capita Advantage (RpCA) (Bustos et al., 2012; O'Clery et al., forthcoming; Bustos and Yildirim, forthcoming):

$$\text{RpCA}_{c,i} \equiv \frac{Y_{c,i}/Y_i}{L_c/L} \approx \frac{\sigma_\mu}{\sigma} \exp\left[-\frac{(f_i-\mu_c)^2}{2\sigma^2}\right] \propto \exp\left[-\frac{(f_i-\mu_c)^2}{2\sigma^2}\right] \tag{9}$$

This equation implies that countries will be better in making the products whose factor levels closely match their own factor levels. In the estimation part below we will make us of this observation.

2.1 Estimation

The model introduced in the previous section implies that we can link the observed production patterns to underlying country and industry parameters. As a first step, we transform RpCA levels to binary variables and interpret Equation 9 as the probability of country c producing in industry i. In particular, we define a binary matrix $\boldsymbol{M}$ whose elements are denoted by M_{ci}:

$$M_{c,i} = \begin{cases} 1 & \text{if } \text{RpCA}_{c,i} \geq \tau \\ 0 & \text{Otherwise} \end{cases} \tag{10}$$

where τ is a threshold. From Equation 9, country c will make the product i with a probability decreasing with the distance between the country and the product levels, $\mu_c - f_i$:

$$Pr\{M_{ci} = 1\} \propto e^{-(\mu_c - f_i)^2/\sigma} \tag{11}$$

The parameter σ controls the sensitivity level of the probability on the distance between countries and products. Given the $\boldsymbol{M}$ matrix, we would like to find μ and f parameters that maximizes the following log-likelihood function:

$$\left[\{\hat{\mu_c}\}, \{\hat{f_i}\}, \hat{\sigma}\right] = \arg \max_{\{\mu_c\},\{f_i\},\sigma} \sum_c \sum_i \left[-M_{ci}\frac{(\mu_c - f_i)^2}{2\sigma^2} + (1-M_{ci})\ln(1 - e^{-(\mu_c - f_i)^2/2\sigma^2})\right] \tag{12}$$

The term $M_{ci} = 1$ indicates that the country is making the product, but $M_{ci} = 0$ is more prone to error because of thresholding or measurement errors. To analytically approximate the likelihood function, we only consider the first term in the summation. For most country-industry pairs, the term $(\mu_c - f_i)^2$ dominates to contribution from the exponential term, which would be close to 0. Ignoring the exponential term, on the other hand, opens up to the trivial solution where all country and product measures are equal. We introduce some restrictions below (inspired from Mealy et al. (2019)) to avoid these pitfalls and enables us to justify ignoring the exponential term. With this simplification, the log-likelihood function becomes:

$$\left[\{\hat{\mu_c}\}, \{\hat{f_i}\}, \hat{\sigma}\right] \approx \arg \max_{\{\mu_c\},\{f_i\},\sigma} \sum_c \sum_i -M_{ci} \frac{(\mu_c - f_i)^2}{2\sigma^2} \tag{13}$$

with following restrictions:

Restriction 1: $\sum_c \sum_i M_{ci}\hat{\mu}_c = 0 \Rightarrow \sum_c k_c\hat{\mu}_c = 0$ where $k_c \equiv \sum_i M_{ci}$ is the diversity of country c. Hence, the μ_c values are distributed around 0 with higher weights given to countries that have more presences and $\hat{\mu}_c = const. \neq 0$ is avoided.

Restriction 2: $\sum_c \sum_i M_{ci}\hat{f}_i = 0 \Rightarrow \sum_i k_i\hat{f}_i = 0$ where $k_i \equiv \sum_c M_{ci}$ is the ubiquity of i.

Restriction 3: $\sum_c \sum_i M_{ci}\hat{\mu}_c^2 = 1 \Rightarrow \sum_c k_c\hat{\mu}_c^2 = 1$ to normalize μ_c's. The choice of 1 as the normalization is arbitrary.

Restriction 4: $\sum_c \sum_i M_{ci}\hat{f}_i^2 = 1 \Rightarrow \sum_i k_i\hat{f}_i^2 = 1$.

The restrictions 2 and 4 have the same underlying logic as restrictions 1 and 3, respectively. Given these restrictions and the approximation to the log-likelihood function, the following proposition gives us an analytic solution for the optimal parameters.

Proposition 2 *Given the restrictions 1 to 4 above, $\hat{\mu}_c$ that maximizes the likelihood function in Equation 13 is given by the eigenvector corresponding to the second largest eigenvalue of $\tilde{\boldsymbol{M}} = \boldsymbol{D}^{-1}\boldsymbol{M}\boldsymbol{U}^{-1}\boldsymbol{M}'$, where $\boldsymbol{D}$ is the diagonal matrix whose entries are k_c and $\boldsymbol{U}$ is the diagonal matrix whose entries are k_i.*

Proof. Let λ_j be the Lagrange multiplier associated with j^{th} restriction. By taking the derivative of log-likelihood function with respect to $\hat{\mu}_c$ we obtain:

$$\sum_i M_{ci} \frac{(\hat{\mu}_c - \hat{f}_i)}{\hat{\sigma}^2} + \lambda_1 k_c + 2\lambda_3 k_c \mu_c = 0 \Rightarrow k_c\hat{\mu}_c(1 + 2\sigma^2)\lambda_3 + k_c\lambda_1\sigma^2 = \sum_i M_{ci}\hat{f}_i$$

If we sum both sides over c:

$$(1+2\sigma^2)\lambda_3 \sum_c k_c \hat{\mu}_c + \lambda_1 \sigma^2 \sum_c k_c = \sum_c \sum_i M_{ci} \hat{f}_i$$

Right hand side is 0 because of restriction 2. First term in the left hand side is 0 because of restriction 1. Hence $\lambda_1 = 0$. Similarly, by taking derivative with respect to $\hat{f}_i$ we can show that $\lambda_2 = 0$. Therefore:

$$\hat{\mu}_c = \frac{1}{(1+2\sigma^2)\lambda_3} \sum_i \frac{M_{ci}\hat{f}_i}{k_c} \quad \text{and} \quad \hat{f}_i = \frac{1}{(1+2\sigma^2)\lambda_4} \sum_{c'} \frac{M_{c'i}\hat{\mu}_c}{k_i} \tag{14}$$

Combining these two equations, we obtain:

$$\hat{\mu}_c = \frac{1}{(1+2\sigma^2)\lambda_3} \frac{1}{(1+2\sigma^2)\lambda_4} \sum_{c'} \sum_i \frac{M_{ci} M_{c'i}}{k_c k_i} \hat{\mu}_{c'} \tag{15}$$

This equation is an eigenvalue/eigenvector equation for $\tilde{\boldsymbol{M}}$ matrix: Hence:

$$\tilde{\boldsymbol{M}}\hat{\boldsymbol{\mu}} = \underbrace{(1+2\sigma^2)^2\lambda_3\lambda_4}_{\text{eigenvalue of } \tilde{\boldsymbol{M}}} \quad \underbrace{\hat{\boldsymbol{\mu}}}_{\text{eigenvector of } \tilde{\boldsymbol{M}}} \tag{16}$$

Since $\tilde{\boldsymbol{M}}$ is a row stochastic matrix, its largest eigenvalue is 1 and the eigenvector corresponding to this eigenvalue is a vector whose elements are all equal to each other. But because of the restrictions above, that cannot be the solution. Hence, the eigenvector and the eigenvalue minimizing the likelihood are the second largest eigenvalue of $\tilde{\boldsymbol{M}}$:

$$e_2 = (1+2\sigma^2)^2\lambda_3\lambda_4$$

and the eigenvector corresponding to this eigenvalue gives us the estimate $\hat{\boldsymbol{\mu}}$. ■

Surprisingly, $\tilde{\boldsymbol{M}}$ is the same matrix used in Hausmann et al. (2014) to define the complexity variables. This implies that $\hat{\mu}_c$ is equivalent to ECI for country c and it represents the average productive knowledge or factor level present in a country. Similarly, $\hat{f}_i$ is equivalent to the PCI level of industry i and it captures the average factor intensity assigned to industry i. As we argued above, the demand determines the factor assignments but as still f_i captures the ranking of factor-responsiveness of an industry. We call the $\hat{\mu}$ parameters the country Average Factor Level (AFL) and the $\hat{f}$ parameters the Product Factor Level (PFL).

Coming back to the first restriction, it implies that the average factor levels and production diversity are orthogonal. In the context of ECI, Kemp-Benedict (2014) shows that this restriction is satisfied for ECI, and that is why ECI cannot be thought of a generalization of the diversity of industrial activities carried out in a country.

Next, we will test some empirical implications of our model. Particularly, we will explore the features of the country level statistic, AFL, and relate it with a number of country characteristics, including economic growth.

3 Empirical Results

Our model ties the observed production patterns in countries to the underlying factor distributions. However, we do not have access to a uniform dataset with a detailed enough industry classification to capture comparable production patterns of countries. We bypass this shortcoming with a simple assumption that all countries export the same share of their output in an industry. Therefore, we can use the rich international trade data –cleaned by Bustos and Yildirim (2020) from the raw data provided by UN COMTRADE– for our estimations. We choose the SITC-4 Revision 2 classification, which available consistently between 1962 and 2018, to maximize temporal coverage.[6] We limit ourselves to 125 countries present in the Hausmann et al. (2014). To eliminate spurious presences due to re-exports or reporting errors, we require at least three years of consecutive non-zero trade levels. Following Bustos and Yildirim (forthcoming), we remove industries for which do not observe trade throughout the sample period, and also drop the smallest industries that constitute cumulatively less than 0.5% of the world trade, leaving us with 660 industries.

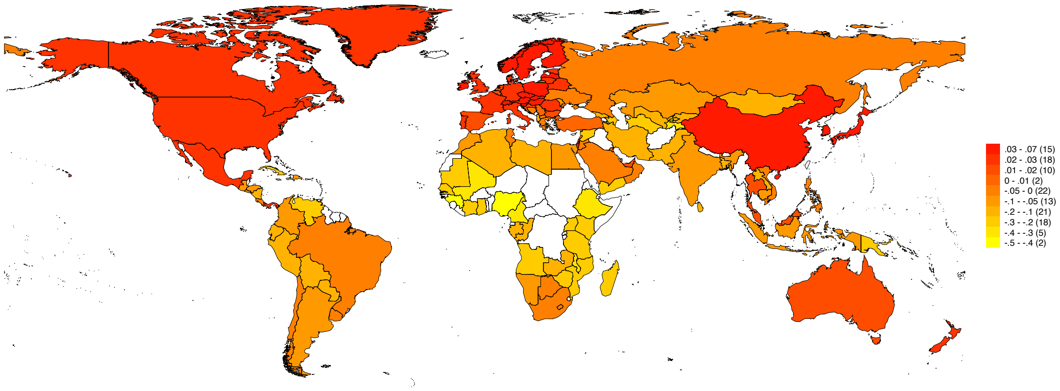
First, we calculate the continuous RpCA values and convert them to binary presence data to obtain M matrix. Following Hausmann et al. (forthcoming) and O'Clery et al. (forthcoming), we select the threshold in Equation 10 to be $\tau = 0.25$.[7] With this threshold, an industry is called to be present in a country if the country's share in the industry is at least a quarter

[6] The dataset is available at https://dataverse.harvard.edu/dataverse/atlas.

[7] In the Appendix, we provide results with $\tau = 0.5$.

of its share in the world population. Using M matrix, we estimate the Average Factor Level (AFL) of each country and factor level matching each product (PFL) using the eigenvector formulation given in Equation 16.

Figure 2: Distribution of Country Factor Levels (AFL)



NOTES: Country Factor Level (AFL) measure for year 2015. We use $\tau = 0.25$ to determine industry presences in Equation 10.

Figure 2 shows the calculated AFL measure for each country. The highest AFL is observed for Japan, followed by South Korea and China. Western European and North American countries are ranked towards the top of AFL rankings. Guinea, Nigeria and Mali, on the other hand, have the lowest AFLs.

How is AFL related to other country-level characteristics such as output, human capital, physical capital and institutionals? In Table 1 we show correlations with country characteristics, starting with variables related to outputs levels. AFL shows a high correlation level, 79.1%, with GDP per capita of a country.[8] We would like to highlight that no price nor level information goes into calculation of AFL as we only use the presence of products in a country's export baskets to arrive at this measure. The second variable that we use is the exports per capita with 74.4% correlation with AFL. Third, higher levels of AFL is associated with lower levels of concentration exports calculated as the Herfindahl–Hirschman Index. Fourth, trade openness, measured as the ratio of imports and exports as a share of GDP, is also positively correlated with AFL. Our fifth variable is the natural reserve products as a share of

[8] See the caption of Table 1 for data sources.

exports. We include this variable to control for countries that have high income per capita due to presence of natural resources, such as oil. This variable does not exhibit a significant level of correlation with AFL. Overall, we observe that AFL is higher for countries that have higher income, exports, diversity and openness.

AFL also shows significant correlations with human capital variables. Years of schooling show more than 70% correlation with AFL, suggesting an underlying relationship between education and AFL. More urbanized countries have higher AFL levels with 73% correlation. R&D spending as a part GDP show 55% correlation with AFL. Population share under 15, which is indicative of high dependency in a country, has a high negative correlation of -76%. However, correlation values decrease when we explore the relationship between AFL and physical capital variables. Investment as a share of GPD have 40% correlation whereas price level of capital formation has close to 13%. Interestingly, FDI does not exhibit any significant correlation with AFL. We speculate that this is due to composite nature of FDI: horizontal FDI is often dominated by Advanced Economy (AE)-AE interactions but vertical FDI is dominated by AE - Emerging Markets and Developing Economies (EMDE) interactions.

The last set of variables that we use is the institutional measures. We rely on two sources for the institutional variables, namely from the Freedom House (FH) and the Worldwide Governance Indicators (WGI) by the World Bank. AFL is significantly correlated with all variables that reflect institutional quality with expected signs. Interestingly, correlation levels are higher for WGI variables, which cover only more recent years starting in 1996. The highest correlation is observed for the Government Effectiveness measure.

Given these high level of correlations with country level characteristics that are often associated with economic growth, we test whether AFL explains some aspects of country growth dynamics that cannot be captured by these variables. In particular, as shown in Table 1, AFL is highly correlated with GDP per capita, but the deviations from this relationship might have implications for future growth since AFL captures the productive capacity of a country. In relation to economic growth, our hypothesis is that if a country's AFL is higher than the one implied by that its per capita income level, the excess AFL will trigger higher growth, whereas lower AFL would do the opposite. To test this hypothesis, we run growth regressions following Barro (1991) and Mankiw et al. (1992). We add variables that are found by Moral-Benito

Table 1: Correlations with country characteristics.

	Correlation with CFL			
Variable	Coef		s.e.	obs
Output				
GDP pc, log	0.791	***	0.043	5,487
Exports pc, logs	0.744	***	0.046	6,021
Concentration of exports	-0.439	***	0.076	6,129
Openness	0.207	***	0.056	5,637
NR exports (% GDP)	-0.133		0.085	5,684
Human Capital				
Years of schooling, log	0.704	***	0.056	5,560
R&D (% of GDP)	0.551	***	0.071	1,628
Urban share	0.725	***	0.052	6,129
Population share under 15	-0.760	***	0.046	6,129
Physical Capital				
Investment (% of GDP)	0.404	***	0.067	5,817
Price level of capital formation	0.129	**	0.050	5,817
FDI (% of GDP)	0.043		0.047	5,032
Institutions				
Level of Democracy (FH)	0.481	***	0.064	5,102
Civil Liberties (FH)	-0.527	***	0.056	5,102
Political Rights (FH)	-0.511	***	0.058	5,102
Freedom Status (FH)	-0.486	***	0.059	5,102
Government Effectiveness (WGI)	0.719	***	0.052	2,362
Control of Corruption (WGI)	0.650	***	0.050	2,362
Rule of Law (WGI)	0.672	***	0.048	2,362

Note: Each cell reports the correlation between the variable in the row with the average factor level of a country. To calculate correlation level for the whole sample, first we standardize each variable for each year. Then we run the following regression:

$$\hat{x}_{c,t} = \beta\widehat{\text{AFL}}_{c,t} + \delta_t + \varepsilon_{c,t}$$

where $\hat{x}_{c,t}$ represents the standardized country characteristic, $\widehat{\text{AFL}}_{c,t}$ represents the standardized AFL, δ_t is the year fixed effects to capture yearly fluctuations. GDP per capita, exports per capita, openness, urban share, population share under 15 and Foreign Direct Investment (FDI) variables are either directly taken from or calculated as a ratio from the variables from the World Development Indicators (WDI) by the World Bank. Concentration of exports and Natural resource exports as a share of GDP variables are calculated from the UNCOMTRADE data. Years of schooling variable is provided by Barro and Lee (2013). Investment and price level of capital formation variables are obtained from Penn World Table version 9.1 (Feenstra et al., 2015). Level of Democracy, Civil Liberties, Political Rights and Freedom Status are provided by the Freedom House (FH). Government Effectiveness, Control of Corruption and Rule of Law variable are taken from the Worldwide Governance Indicators (WGI) prepared by the World Bank. Robust standard errors in parentheses clustered by country. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

(2012) to be associated with higher growth rates to our repertoire of control variables. We define the annualized growth rate for period between years t and $t + \Delta t$ as:

$$g_{c,t \to t+\Delta t} = \left(\frac{y_{c,t+\Delta t}}{y_{c,t}} \right)^{-1/\Delta t} - 1. \tag{17}$$

where $y_{c,t}$ denotes the GDP per capita of country c at time t. Using the growth rate as the dependent variable, we estimate the following regression:

$$g_{c,t \to t+\Delta t} = \beta_y \times \ln y_{c,t-1} + \beta_c \times \text{AFL}_{c,t} + \beta_x \chi_{c,t} + \delta_t + \varepsilon_{c,t} \tag{18}$$

where δ_t is the year fixed effect and $\chi_{c,t}$ is a vector of country-year level control variables specified at the initial year, t. The control variables for countries are added in groups to observe the changes in β_c coefficient. In a final specification, we use country fixed-effects instead of country characteristics. Following Barro (1991), we use the lagged income level, $\ln y_{c,t-1}$, to avoid biased estimates, specifically when we include country fixed-effects instead of country characteristics (Nickell, 1981). We use time intervals of $\Delta t = 5$ or 10 years with non-overlapping years between 1965-2015. We use standardized versions of independent variables to be able to compare the coefficients across specifications.

In Table 2, we report the results of these growth regressions. In columns 1 and 6, we only include AFL, lagged income per capita and year fixed-effects as independent variables for estimating 5 and 10 year growth rates, respectively. For both growth rates, we observe significant and economically meaningful contribution from AFL; with one standard deviation increase in AFL translating to an additional 1.15 and 1.25 percent growth over 5 and 10 year time-spans, respectively. In the second and seventh columns, we include a small set of variables to control for physical capital (investments a share of GDP variable), human capital (years of schooling) and population growth. Consistent with the high level of correlations present in Table 1, the coefficient for AFL decreases because of the multi-collinearity, however AFL still explains a significant portion of economic growth.

In columns 3 and 8 of Table 2, we extend the control variables by incorporating additional measures such as investment distortions (captured by investment price levels), the share of

Table 2: Predicting economic growth

	5 year growth rate					10 year growth rate				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Income per capita lagged, log	-1.22***	-1.36***	-1.58***	-2.21***	-4.40***	-1.38***	-1.55***	-1.88***	-2.40***	-4.46***
	(0.21)	(0.20)	(0.29)	(0.36)	(1.06)	(0.22)	(0.22)	(0.33)	(0.38)	(0.93)
Avg. Factor Level (AFL)	1.15***	0.74***	0.47**	0.54**	0.62**	1.25***	0.84***	0.46**	0.59**	0.70**
	(0.22)	(0.18)	(0.19)	(0.21)	(0.31)	(0.23)	(0.21)	(0.22)	(0.23)	(0.29)
Investment (% of GDP)		0.40***	0.24**	0.23**			0.26**	0.18*	0.16	
		(0.11)	(0.10)	(0.10)			(0.11)	(0.11)	(0.11)	
Years of schooling		0.24	0.07	0.10			0.38**	0.20	0.16	
		(0.14)	(0.15)	(0.16)			(0.15)	(0.15)	(0.17)	
Population growth (%)		-0.52***	-0.33***	-0.26**			-0.48***	-0.17	-0.07	
		(0.14)	(0.11)	(0.12)			(0.12)	(0.13)	(0.13)	
Investment price			-0.44***	-0.41***				-0.32***	-0.25**	
			(0.07)	(0.08)				(0.08)	(0.11)	
Population share under 15			-0.90***	-1.16***				-0.94***	-1.22***	
			(0.24)	(0.26)				(0.26)	(0.30)	
Urban share			0.02	0.14				0.22	0.22	
			(0.18)	(0.20)				(0.21)	(0.22)	
Openness			0.35***	0.26**				0.24*	0.20	
			(0.10)	(0.12)				(0.13)	(0.14)	
Natural Res. Exp. (% of GDP)			0.05	0.14				-0.17	-0.00	
			(0.13)	(0.13)				(0.12)	(0.15)	
Level of Democracy (FH)				-0.13					-0.06	
				(0.14)					(0.16)	
Civil Liberties (FH)				-0.14					-0.05	
				(0.15)					(0.18)	
Political Rights (FH)				-0.21					-0.15	
				(0.21)					(0.23)	
Freedom Status (FH)				0.00					-0.12	
				(0.37)					(0.42)	
Observations	826	826	826	706	826	400	400	400	348	400
R^2	0.18	0.25	0.31	0.35	0.44	0.24	0.32	0.38	0.41	0.62
Year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Country FE	-	-	-	-	yes	-	-	-	-	yes

Note: The dependent variable is the geometric growth rate of GDP per capita, over 5 and 10 year periods, measured using constant dollars as of 2010. We use the lagged income as a control. All other control variables are measured for the initial year. We normalize all variables to have a mean of zero and standard deviation of one in each year. The variables definitions and resources are given in the caption of Table 1. Robust standard errors in parentheses clustered by country. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

population that are in need of direct support (captured by population share under 15 years old), urbanization, trade openness and natural resource exports (to proxy for income obtained from extractive industries). The explanatory power of AFL measure declines to 0.47 and 0.46 for 5 and 10 year growth predictions, but the coefficients are still significant at 1% significance level. In columns 4 and 9, we add variables capturing the institutional quality of countries, namely level of democracy, civil liberties, political rights and freedom status variables from Freedom House. We use these variables because they have the longest temporal coverage, starting from 1972. Even after controlling with these extensive set of variables, the coefficient for AFL is still economically meaningful and statistically significant. Finally, in columns 5 and 10, we control for all country-level hetereogeneity by including country fixed-effects in our estimations. AFL survives through this very stringent test, as well.

It is important to note that we are not testing the explanatory power of control variables in economic growth. In literature, many of these variables have been shown to be associated with growth, but in our exercise, the sign and the significance of coefficients associated with these variables might change due to severe multi-colinearity. The methodology we employ here is akin to a "kitchen sink" approach to expose AFL to the highest scrutiny. The coefficient for AFL variable remains significant across all specifications, indicating a strong positive relation between economic growth and AFL.

Is the relationship between AFL and economic growth causal? To answer this, we use the Granger non-casuality test for panels proposed by Dumitrescu and Hurlin (2012). First, we de-trend the by calculating the residual from the following regression:

$$\mathrm{AFL}_{c,t} = \beta_y \times \ln y_{c,t-1} + \delta_t + \widehat{\mathrm{AFL}}_{c,t} \tag{19}$$

where $\widehat{\mathrm{AFL}}_{c,t}$ is the residual. We use the income per capita from the previous year to eliminate bias in the prediction. We conjecture that higher levels of $\widehat{\mathrm{AFL}}_{c,t}$ should trigger economic growth. We use the 5-year growth rate defined in Equation 17 as the dependent variable and test whether there is a casual link between AFL residual and growth. We create a balanced panel with 78 countries for this purpose and use xtgcause package in Stata to test the null hypothesis that excess AFL does not Granger-cause growth. We include a single-lag into the underlying VAR estimation. We run 1000 bootstraps to obtain a p-value of 3.4% for the null-hypothesis, indicating that we can reject it with greater than 95% confidence.[9]

As a final robustness check, we would like to test whether the presence of a country's own exports while estimating AFL creates endogeneity. To this end, we remove a country, c, from our estimation of PFL and re-calculate AFL from these PFL levels. We perform this exercise in two ways. In the first case, we remove the exports and population of country c from the trade and population data, respectively, and create $\hat{M}^c$ matrix with $N_C - 1$ rows. In the second case, we remove the row corresponding to our country of interest from the original M matrix to create $\hat{M}^c$. We re-estimate PFL from $\hat{M}^c$ and calculate AFL using the original M matrix via Equation 14. We repeat this exercise for all countries in our dataset. Consequently, we arrive at an AFL vector that does not include self-information in the estimation process. In machine learning this type of robustness exercise is referred to as leave-one-out procedure. Table 3 shows our results. The results do not change significantly between columns corresponding to the same specification. Hence, we conclude that there is no unintended endogenous mechanism that give rise to the observed results.

4 Conclusions

Here, by assuming a supermodular relationship between production factors and industries, we obtain an estimate of average factor level of a country. Interestingly, this estimate coincides with the celebrated economic complexity framework introduced in Hidalgo and Hausmann (2009) and Hausmann et al. (2014). Therefore, our approach could be thought of a rationalization of the widely used complexity methodology through an underlying economic model for the first time. Measures like AFL and PFL give policymakers tools that they could use as a gauge to understand the productive structures of their countries or locations. This is corroborated by the widespread adaption of enigmatic ECI and PCI variables in policy circles.

Supermodularity is a powerful tool capturing the complementarities in the production process. Of course, there could be some extensions in usage of supermodularity. Supermodularity could be embedded into multi-dimensional factor structure using Eaton and Kortum

[9] Associated W-bar, Z-bar and Z-bar tilde statistics are 2.39, 8.66 and 7.49, respectively.

Table 3: Robustness of country economic growth results.

	5 year growth rate						10 year growth rate					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Income per capita lagged, log	-1.58***	-4.39***	-1.58***	-4.38***	-1.58***	-4.37***	-1.87***	-4.45***	-1.87***	-4.40***	-1.86***	-4.40***
	(0.29)	(1.06)	(0.29)	(1.04)	(0.29)	(1.05)	(0.33)	(0.94)	(0.33)	(0.93)	(0.33)	(0.93)
AFL	0.47**	0.62**					0.45**	0.69**				
	(0.19)	(0.31)					(0.22)	(0.29)				
AFL (Rem. Exports)			0.49**	0.63**					0.46**	0.66**		
			(0.19)	(0.29)					(0.22)	(0.29)		
AFL (Rem. Binary)					0.48**	0.62**					0.46**	0.66**
					(0.19)	(0.29)					(0.22)	(0.30)
Observations	825	825	825	825	825	825	399	399	399	399	399	399
R^2	0.30	0.44	0.31	0.44	0.31	0.44	0.38	0.61	0.38	0.61	0.38	0.61
Year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Country FE	-	yes	-	yes	-	yes	-	yes	-	yes	-	yes
Controls	yes	-	yes	-	yes	-	yes	-	yes	-	yes	-

Note: The dependent variable is the geometric growth rate of GDP per capita, over 5 and 10 year periods, measured using constant dollars of 2010. In columns 3, 4, 9 and 10 we use AFL calculated by removing a country's exports and population from the world export and population. In columns 5, 6, 11 and 12, we remove country's one-by-one from the M matrix. In odd numbered columns, we use the following control variables: Investment (% of GDP), Years of schooling, Population growth (%), Investment price, Population share under 15, Urban share, Openness, NNRR exports (% of GDP). The specification for odd-numbered columns are the same as column 3 and 8 of Table 2. Even numbered columns include country-fixed effects. We normalize all variables to have a mean of zero and standard deviation of one in each year. The variables definitions and resources are given in the caption of Table 1. Robust standard errors in parentheses clustered by country. *** p<0.01, ** p<0.05, * p<0.1

(2002)'s framework. This would result in a supermodular export patterns for countries as pointed out by Costinot and Vogel (2015). Hence, this will also enable us to incorporate trade costs into our model. A supermodular function, $f(x, y)$ divided by the multiplication of any positive valued functions of x and y would result in a supermodular function by definition. Hence, the supermodularity of exports is extended to the supermodularity of Balassa (1965)'s Revealed Comparative Advantage (RCA) measure. In the original formulation of the complexity variables in Hidalgo and Hausmann (2009), the RCA measure is used to calculate ECI and PCI variables. Schetter (2019) shows that the eigenvector corresponding to the second largest eigenvalue of the matrix shown in Equation 16 sorts the countries consistent with the underlying supermodular function. Nevertheless, the advantage of using RpCA here is that we can tie its level to underlying country and product characteristics.

We assume that the factor distribution in countries follow a normal distribution with the same standard deviation, which gives us the functional form $\exp[-(\mu_c - f_i)^2/(2\sigma^2)]$ to estimate AFL and PFL. By relaxing this assumption we can also target other moments of the data, such as the diversity and ubiquity (Hidalgo and Hausmann, 2009; Hausmann and Hidalgo,

2011) or the nestedness property (Bustos et al., 2012). In particular, one might consider country specific standard deviations or non-symmetric distributions, with longer left tails than right tails.

A significant chunk of the data we ignored in our estimations is the zeroes of the M matrix, which correspond to the absent products. Our measures of AFL and PFL maximize the likelihood of observed presences in this matrix. The advantage of this approach is that we obtain an analytical solution to the problem. One extension could be incorporating zeroes and solving the likelihood problem numerically. The second extension could be to estimate directly the non-binarized versions of RpCA values directly.

We believe that bridging successful policy tools with an underlying economic model and we achieve this for the complexity variables here. We think that this opens us up to other extensions both in theoretical and empirical aspects.

References

Abowd, John M, Francis Kramarz, and David N Margolis, "High wage workers and high wage firms," *Econometrica*, 1999, *67* (2), 251–333.

Antràs, Pol, Luis Garicano, and Esteban Rossi-Hansberg, "Offshoring in a knowledge economy," *The Quarterly Journal of Economics*, 2006, *121* (1), 31–77.

Balassa, Bela, "Trade liberalisation and "revealed" comparative advantage," *The Manchester School*, 1965, *33* (2), 99–123.

Barro, Robert J, "Economic Growth in a Cross Section of Countries," *The Quarterly Journal of Economics*, 1991, *106* (2), 407–443.

_ and Jong Wha Lee, "A new data set of educational attainment in the world, 1950–2010," *Journal of Development Economics*, 2013, *104*, 184–198.

Becker, Gary S, "A theory of marriage: Part I," *Journal of Political Economy*, 1973, *81* (4), 813–846.

Borjas, George J, "Self-Selection and the Earnings of Immigrants," *American Economic Review*, 1987, pp. 531–553.

Bustos, Sebastián and Muhammed A Yildirim, "Uncovering trade flows," 2020. Unpublished mimeo, available upon request.

_ and _ , "Production Ability and economic growth," *Research Policy*, forthcoming.

_ , Charles Gomez, Ricardo Hausmann, and César A Hidalgo, "The dynamics of nestedness predicts the evolution of industrial ecosystems," *PloS one*, 2012, *7* (11), e49393.

Card, David, Jörg Heining, and Patrick Kline, "Workplace heterogeneity and the rise of West German wage inequality," *The Quarterly journal of economics*, 2013, *128* (3), 967–1015.

Costinot, Arnaud, "An elementary theory of comparative advantage," *Econometrica*, 2009, *77* (4), 1165–1192.

_ and Jonathan Vogel, "Matching and inequality in the world economy," *Journal of Political Economy*, 2010, *118* (4), 747–786.

_ and _ , "Beyond Ricardo: Assignment Models in International Trade," *Annual Review of Economics*, 2015, *7* (1), 31–62.

Dumitrescu, Elena-Ivona and Christophe Hurlin, "Testing for Granger non-causality in heterogeneous panels," *Economic Modelling*, 2012, *29* (4), 1450–1460.

Eaton, Jonathan and Samuel Kortum, "Technology, geography, and trade," *Econometrica*, 2002, *70* (5), 1741–1779.

Eeckhout, Jan and Philipp Kircher, "Sorting and decentralized price competition," *Econometrica*, 2010, *78* (2), 539–574.

_ and _ , "Sorting versus screening: Search frictions and competing mechanisms," *Journal of Economic Theory*, 2010, *145* (4), 1354–1385.

_ and _ , "Identifying sorting – in theory," *The Review of Economic Studies*, 2011, *78* (3), 872–906.

_ and _ , "Assortative Matching With Large Firms," *Econometrica*, 2018, *86* (1), 85–132.

Feenstra, Robert C, Robert Inklaar, and Marcel P Timmer, "The next generation of the Penn World Table," *American Economic Review*, 2015, *105* (10), 3150–82.

Grossman, Gene M and Giovanni Maggi, "Diversity and trade," *American Economic Review*, 2000, *90* (5), 1255–1275.

_ , Elhanan Helpman, and Philipp Kircher, "Matching, Sorting, and the Distributional Effects of International Trade," *Journal of Political Economy*, 2017, *125* (1), 224–264.

Hausmann, Ricardo and César A Hidalgo, "The network structure of economic output," *Journal of Economic Growth*, 2011, *16* (4), 309–342.

_ , _ , Sebastián Bustos, Michele Coscia, Alexander Simoes, and Muhammed A. Yıldırım, *The Atlas of Economic Complexity: Mapping Paths to Prosperity*, The MIT Press, 2014.

_ , Daniel Stock, and Muhammed A Yildirim, "Implied Comparative Advantage," *Research Policy*, forthcoming.

Heckman, James J and Bo E Honoré, "The empirical content of the Roy model," *Econometrica*, 1990, pp. 1121–1149.

_ and Guilherme Sedlacek, "Heterogeneity, aggregation, and market wage functions: an empirical model of self-selection in the labor market," *Journal of Political Economy*, 1985, *93* (6), 1077–1125.

Hidalgo, César A, "Economic complexity theory and applications," *Nature Reviews Physics*, 2021, pp. 1–22.

_ and Ricardo Hausmann, "The building blocks of economic complexity," *Proceedings of the National Academy of Sciences of the United States of America*, 2009, *106* (26), 10570–10575.

Kemp-Benedict, Eric, "An interpretation and critique of the Method of Reflections," 2014. MPRA Paper No. 60705.

Lucas, Robert E., "On the mechanics of economic development," *Journal of Monetary Economics*, 1988, *22* (1), 3–42.

Mankiw, N Gregory, David Romer, and David N Weil, "A Contribution to the Empirics of Economic Growth," *The Quarterly Journal of Economics*, 1992, *107* (2), 407–437.

Mealy, Penny, J. Doyne Farmer, and Alexander Teytelboym, "Interpreting economic complexity," *Science Advances*, 2019, *5* (1), eaau1705.

Moral-Benito, Enrique, "Determinants of economic growth: a Bayesian panel data approach," *Review of Economics and Statistics*, 2012, *94* (2), 566–579.

Nickell, Stephen, "Biases in dynamic models with fixed effects," *Econometrica: Journal of the econometric society*, 1981, pp. 1417–1426.

O'Clery, Neave, Muhammed A Yildirim, and Ricardo Hausmann, "Productive Ecosystems and the Arrow of Development," *Nature Communications*, forthcoming.

Ohnsorge, Franziska and Daniel Trefler, "Sorting it out: International trade with heterogeneous workers," *Journal of Political Economy*, 2007, *115* (5), 868–892.

Roy, Andrew Donald, "Some thoughts on the distribution of earnings," *Oxford Economic Papers*, 1951, *3* (2), 135–146.

Schetter, Ulrich, "A Structural Ranking of Economic Complexity," 2019. CID Research Fellow & Graduate Student Working Paper 119.

Shi, Jianbo and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, 2000, *22* (8), 888–905.

Sutton, John and Daniel Trefler, "Capabilities, wealth, and trade," *Journal of Political Economy*, 2016, *124* (3), 826–878.

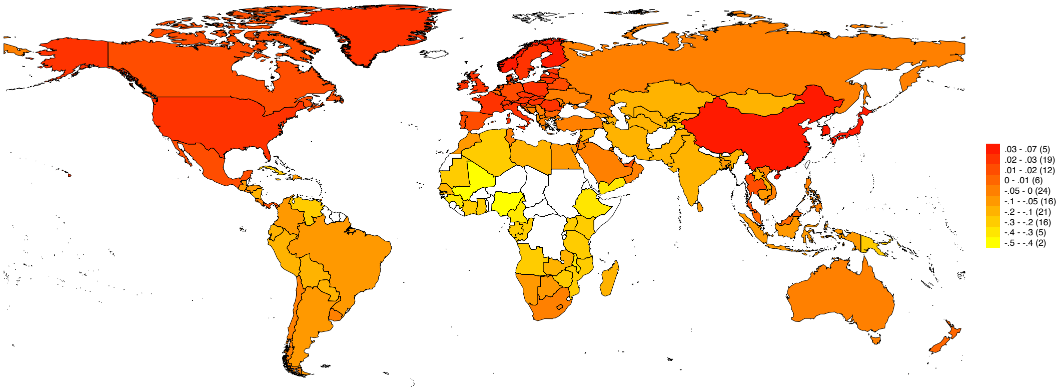
Teulings, Coen N, "The wage distribution in a model of the assignment of skills to jobs," *Journal of Political Economy*, 1995, *103* (2), 280–315.

_ , "Comparative advantage, relative wages, and the accumulation of human capital," *Journal of Political Economy*, 2005, *113* (2), 425–461.

A Appendix

A.1 Additional Figures and Tables

Figure A.1: Distribution of Country Factor Levels (AFL)



NOTES: Distribution of AFL for $\text{RpCA}_{c,i} > \tau = 0.5$ threshold to determine presences in Equation 10 for year 2015.

Table A.1: Correlations with country characteristics.

	Correlation with eci			
Variable	Coef		s.e.	obs
Output				
GDP pc, log	0.787	***	0.042	5,479
Delta GDP pc, log	0.079	**	0.034	5,437
Exports pc, logs	0.736	***	0.046	5,986
Delta Exports pc, logs	0.077	***	0.024	5,980
Concentration of exports	-0.432	***	0.074	6,092
Openness	0.207	***	0.055	5,625
NR exports (% GDP)	-0.146	*	0.085	5,672
Human Capital				
Years of schooling, log	0.701	***	0.056	5,529
R&D (% of GDP)	0.560	***	0.071	1,628
Urban share	0.713	***	0.051	6,092
Population share under 15	-0.765	***	0.044	6,092
Physical Capital				
Investment (% of GDP)	0.402	***	0.066	5,789
Price level of capital formation	0.131	**	0.051	5,789
FDI (% of GDP)	0.040		0.048	5,020
Institutions				
Level of Democracy (FH)	0.485	***	0.064	5,075
Civil Liberties (FH)	-0.530	***	0.056	5,075
Political Rights (FH)	-0.513	***	0.058	5,075
Freedom Status (FH)	-0.484	***	0.059	5,075
Government Effectiveness (WGI)	0.728	***	0.051	2,362
Control of Corruption (WGI)	0.661	***	0.050	2,362
Rule of Law (WGI)	0.683	***	0.047	2,362

Note: In this Table, we use a higher threshold of presence, i.e., $\text{RpCA}_{c,i} > 0.5$. The rest of the exercise is the same with Table 1. Please refer to its caption for more details. *** p<0.01, ** p<0.05, * p<0.1

Table A.2: Country economic growth

	5 year growth rate					10 year growth rate				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Income per capita lagged, log	-1.23***	-1.36***	-1.58***	-2.21***	-4.33***	-1.40***	-1.56***	-1.88***	-2.42***	-4.36***
	(0.21)	(0.20)	(0.29)	(0.36)	(1.06)	(0.23)	(0.22)	(0.33)	(0.37)	(0.94)
Avg. Factor Level (AFL)	1.16***	0.75***	0.46**	0.57**	0.55*	1.28***	0.87***	0.48**	0.65***	0.57*
	(0.22)	(0.19)	(0.21)	(0.22)	(0.31)	(0.23)	(0.21)	(0.22)	(0.23)	(0.30)
Investment (% of GDP)		0.39***	0.24**	0.22**			0.25**	0.18*	0.16	
		(0.11)	(0.10)	(0.10)			(0.11)	(0.11)	(0.11)	
Years of schooling		0.24	0.07	0.11			0.37**	0.20	0.15	
		(0.15)	(0.15)	(0.16)			(0.16)	(0.15)	(0.17)	
Population growth (%)		-0.51***	-0.33***	-0.26**			-0.47***	-0.18	-0.08	
		(0.14)	(0.11)	(0.12)			(0.12)	(0.13)	(0.13)	
Investment price			-0.44***	-0.41***				-0.32***	-0.26**	
			(0.07)	(0.08)				(0.08)	(0.11)	
Population share under 15			-0.90***	-1.14***				-0.92***	-1.18***	
			(0.24)	(0.26)				(0.27)	(0.30)	
Urban share			0.03	0.15				0.24	0.24	
			(0.19)	(0.20)				(0.21)	(0.22)	
Openness			0.34***	0.26**				0.23*	0.19	
			(0.10)	(0.12)				(0.13)	(0.14)	
Natural Res. Exp. (% of GDP)			0.05	0.14				-0.17	0.01	
			(0.13)	(0.13)				(0.12)	(0.15)	
Level of Democracy (FH)				-0.13					-0.06	
				(0.14)					(0.16)	
Civil Liberties (FH)				-0.13					-0.04	
				(0.15)					(0.18)	
Political Rights (FH)				-0.22					-0.16	
				(0.21)					(0.23)	
Freedom Status (FH)				-0.03					-0.13	
				(0.37)					(0.42)	
Observations	825	825	825	705	825	399	399	399	347	399
R-squared	0.18	0.25	0.30	0.35	0.44	0.25	0.32	0.38	0.41	0.61
Year FE	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
Country FE	-	-	-	-	yes	-	-	-	-	yes

Note: In this Table, we use a higher threshold of presence, i.e., $\text{RpCA}_{c,i} > 0.5$. The rest of the exercise is the same with Table 2 and please refer to its caption for more details. Robust standard errors in parentheses clustered by country. *** p<0.01, ** p<0.05, * p<0.1