Supplementary Material
Machine-learned patterns suggest that diversification drives economic development
Journal of the Royal Society Interface

Charles D. Brummitt[1,*], Andrés Gómez-Liévano[2], Ricardo Hausmann[2,3,4], and Matthew H. Bonds[1]

[1]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, USA
[2]Growth Lab, Harvard University, Cambridge, MA 02138, USA
[3]John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138, USA
[4]Santa Fe Institute, Santa Fe, NM 87501
[*]Corresponding author: brummitt@gmail.com

December 14, 2019

SM-1 Related work

Recent work in economics has embraced the multidimensional nature of an economy by compressing information about the products that the economy exports. The "complexity index" [1, 2], "fitness" [3, 4, 5, 4, 6], and the "entropic measure of production diversification" [7] are notable examples. The first two of these measures can be defined in many ways [8, 9, 10, 11]; what they all share in common is essentially an attempt to quantify "competitiveness" by aggregating (e.g., averaging or adding) the difficulty of producing the products in an export basket. They all tackle an ambitious challenge: to describe an economy's complexity with just one number.

Other work is investigating how compressible economies and societies are. Machado and Mata [12] reduce the dimensions of four time-series (per-capita income, exports relative to income, school enrollment, and lifetime expectancy) to two dimensions using multidimensional scaling. Hruschka et al. [13] create multidimensional models of wealth by reducing the dimensions of responses to household surveys about ownership of assets such as TVs, land, and electricity. Turchin et al. [14] find that societies across millennia tended to move along a common, low-dimensional trajectory in which the complexity of social organization steadily increased over time.

The existence of common patterns in the trajectories of economies can enable forecasts using simple models. For example, economists have fit Markov chains [15] (and continuous versions of them [16] and variants of Markov chains with constraints from growth theory [17]) to time-series data on per-capita incomes. They used the stationary distribution to predict whether countries will converge to similar incomes, or whether they diverge to different "convergence clubs". References [15, 16, 17] predict a bimodal distribution of incomes in the future. What is needed are models inferred from high-dimensional data [18, p. 42 in Sec. 4.1]. Stochastic methods have also been used to model changes in global exports of many products [19].

Our approach was inspired by recent advances in statistical machine learning aimed at identifying governing laws of motion in data. One method, called SINDy [20], expands features using a hand-picked library of functions and then selects among them using sparse regression. It has since been extended to partial differential equations [21], differential equations with rational terms [22], information criteria [23], and control problems [24]. SINDy has proved successful in discovering laws of physics and microbiology, where we can expect polynomials and other simple functions. We found SINDy challenging to work well with noisy economic data with significant outliers, and great care must be taken in choosing the library of functions so that iterated predictions of the future do not diverge. Other approaches to system identification have used symbolic regression and genetic algorithms [25, 26]; least angle regression [27]; and nested hierarchies of models of smooth, nonlinear dynamics [28].

Forecasting economic time-series has a long history, with the method of choice often being autoregressive–moving-average (ARMA) models [29]. Like the "diffusion index" (or "factor augmented forecasts") [30], we use principal components to reduce dimensions.

SM-2 Data on exports

Our data has three main stages, which we will refer to as the *raw* data, the *cleaned and standardized* data, and the *final, aggregated* data.

1. The raw data is the data one can download freely from the United Nations' Commodities Trade Statistics website;

2. the cleaned and standardized data is after the raw data has been expressed using standard classifications across years, and problems of the reliability of the raw records addressed and corrected;

3. the final aggregated data is after we have removed countries and products, and then aggregated into higher level product codes, all with the goal of having reliable statistics.

Under a "Premium Site License" that Harvard has with the United Nations' Commodities Trade Statistics (COMTRADE), we provide our clean and standardized data, free to download, at Dataverse through the following link: https://doi.org/10.7910/DVN/B0ASZU

COMTRADE, the original source of our raw data, is the repository of the official trade transactions between importers and exporters. Products traded are codified in three different commodity classifications, but we express all transactions using the Standard International Trade Classification (SITC) system, Revision 2, because it covers the longest span of time. The cleaned and standardized data that we provide through the link above consists of approximately 8 million rows, each representing what a country exported of a 4-digit coded product in a year. Countries are coded following the International Organization for Standardization (ISO). We have a total of 231 unique country codes, 781 unique product codes, and 53 years (1962–2016).

One of the main issues with the raw data is that different countries use different classifications, and even when an importer and an exporter use the same classification, they may use different revisions.[1] Each transaction in the raw data reports the code as was originally submitted by each party. Hence, to analyze the data one has to standardize the records into a single classification. COMTRADE provides concordance tables that can be used to express a product from one classification to another (`http://unstats.un.org/unsd/cr/registry/regdnld.asp`). But since concordance tables are typically "many-to-many" mappings rather than "one-to-one", the act of re-expressing data from one classification to another introduces additional noise because one must make some arbitrary decisions for how to split the data.

As a consequence, to get the cleaned and standardized data, the raw data has been transformed through a long process of *correction* of reported transactions, *cleaning* of misreported records, and *standardization* of country and product codes. The process is described in detail in [31]. The general approach to do this is referred to in the literature as "mirroring" [32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. Mirroring consists of reconciling between what exporters and importers report, since each transaction should in principle be reported twice. But the difference between previous efforts for creating a trade dataset for research (e.g., the National Bureau of Economic Research [NBER] dataset, and the Centre d'Études Prospectives

[1] SITC codes have had four revisions: SITC Rev. 1 in 1961, Rev. 2 in 1975, Rev. 3 in 1988, and Rev. 4 in 2006.

et d'Informations Internationales [CEPII] BACI dataset) and that of Bustos and Yildirim [31] is that the latter accounts for transaction costs and restrictions implicit in trade reports, and they develop indices of reliability for importers, exporters, and products, which enable them to correctly impute exports of small and developing countries. Thus, the dataset of Bustos and Yildirim [31] is more complete because it increases the number of countries with available data and additional country-product combinations (see [31]), even at very disaggregated levels of the product classification.

SM-3 Preprocessing the exports data

Preprocessing the exports data occurs in five steps described below:

1. Remove some products and countries (Sec. SM-3.1)
2. Normalize by population and by global exports (itself normalized by global population) (Sec. SM-3.2)
3. Apply a logarithmic transformation that preserves zero values and that preserves the number of values above 1 (Sec. SM-3.3)
4. Center and scale for each product (Sec. SM-3.4)
5. Reduce dimensions (Sec. SM-3.5)

SM-3.1 Filtering countries and products

First, we filter the data by removing small countries and products that are not exported widely enough. The filters are similar to those in [8] with some differences. One difference is that we avoid path dependence of the filters: we take the union of the countries and products selected by each filter, and then we remove those countries and products all at once. Another difference is that we chose not to set to zero all export values below a certain small threshold (such as US$5000) so that we do not discard information; we let the models handle noisy, small values rather than choose an arbitrary threshold. The last difference is that we remove products that have first digit in their SITC classification equal to either `3` (*Fuels, lubricants & related materials*) or `9` (*Other*), which includes products such as zoo animals, coins, and gold). We remove fossil fuels because we are interested in how wealth results from a process of development of skills and capabilities rather than lotteries of geographic endowments of resources.

The steps below completely specify our filtering of products and countries. Countries are specified by their ISO-3166-1 alpha-3 country codes, while products are specified using the SITC classification, both found in [45].

1. Initialize `CountriesToRemove` $= \varnothing$ and `ProductsToRemove` $= \varnothing$ (the empty set).
2. **Remove countries with a small population**: Select the countries with population less than 1.25 million in 2008. This selection results in the following list of 81 countries:

   `CountriesToRemove` := `CountriesToRemove` ∪ {ABW, AIA, AND, ANS, ANT, ASM, ATA, ATF, ATG, BHR, BHS, BLZ, BMU, BRB, BRN, BTN, BVT, CCK, COK, COM, CPV, CXR, CYM, CYP, DJI, DMA, ESH, FJI, FLK, FRO, FSM, GIB, GNQ, GRD, GRL, GUM, GUY, IOT, ISL, KIR, KNA, LCA, LUX, MAC, MDV, MHL, MLT, MNE, MNP, MSR, MUS, MYT, NCL, NFK, NIU, NRU, PCN, PLW, PYF, SGS, SHN, SLB, SMR, SPM, STP, SUR, SWZ, SYC, TCA, TKL, TLS, TON, TUV, TWN, UMI, VAT, VCT, VGB, VUT, WLF, WSM}
3. **Remove countries with little total export value**: Select the countries with total export value smaller than 1 billion USD in 2008. This selection results in the following 81 countries:

`CountriesToRemove` := `CountriesToRemove` ∪ {AFG, AIA, AND, ARM, ASM, ATA, ATF, ATG, BDI, BEN, BFA, BLZ, BRB, BTN, BVT, CAF, CCK, COK, COM, CPV, CXR, CYM, DJI, DMA, ERI, ESH, FJI, FLK, FRO, FSM, GIB, GMB, GNB, GRD, GRL, GUM, GUY, HTI, IOT, KIR, KNA, LCA, LSO, MDV, MNE, MNP, MSR, MWI, MYT, NER, NFK, NIU, NPL, NRU, PCN, PLW, PSE, PYF, RWA, SGS, SHN, SLB, SLE, SMR, SOM, SPM, STP, SYC, TCA, TGO, TKL, TLS, TON, TUV, UMI, VAT, VCT, VGB, VUT, WLF, WSM}

4. **Remove countries that export very few products**: Select countries with zero export value for at least 95% of products in some year. This selection results in the following list of 52 countries:

   `CountriesToRemove` := `CountriesToRemove` ∪ {AIA, ATA, ATF, BDI, BTN, BVT, CCK, COK, COM, CPV, CXR, ERI, ESH, FLK, FSM, GNB, GNQ, GUF, HMD, IOT, KIR, LAO, LCA, MDV, MHL, MNG, MNP, MRT, MTQ, NFK, NIU, NPL, NRU, PCI, PCN, PYF, RWA, SGS, SSD, STP, SYC, TCA, TLS, TON, TUV, UMI, VGB, VIR, VUT, WLF, WSM, YEM}

5. **Remove war-torn countries**: Add Afghanistan (AFG), Iraq (IRQ), and Chad (TCD) to the set of countries to remove:

   `CountriesToRemove` := `CountriesToRemove` ∪ {AFG, IRQ, TCD}

6. **Remove all products in the categories of fossil fuels and miscellaneous**: Add to the set of products to remove all the products with first digit (in the SITC classification scheme) equal to `3` (fossil fuels) or `9` (miscellaneous products such as art and coins):

   `ProductsToRemove` := `ProductsToRemove` ∪ $\{3^*, 9^*\}$

   Here, 3^* means any product code that begins with `3`.

7. **Remove products exported by few countries**: Select products not exported by at least 80% of countries in at least one year. This selection results in the following 78 product codes:

   `ProductsToRemove` := `ProductsToRemove` ∪ { 0019, 0115, 0451, 0452, 0742, 2114, 2223, 2226, 2231, 2232, 2234, 2235, 2512, 2516, 2518, 2613, 2634, 2652, 2654, 2655, 2659, 2685, 2712, 2714, 2741, 2742, 2784, 2814, 2816, 2860, 2872, 2876, 3223, 3224, 3231, 3341, 3342, 3343, 3344, 3415, 3510, 4233, 4236, 4241, 4244, 4245, 5163, 5223, 5249, 5323, 5828, 6112, 6113, 6121, 6344, 6546, 6642, 6674, 6727, 6741, 6750, 6784, 6793, 6831, 6880, 7187, 7433, 7521, 7524, 7911, 7912, 7913, 7914, 7924, 7931, 8821, 8941, 9110 }

8. **Remove products with little global exports**: Select products with global exports < 10 million in some year. This selection results in the following 37 products:

   `ProductsToRemove` := `ProductsToRemove` ∪ { 0019, 0742, 1122, 2114, 2232, 2235, 2239, 2634, 2652, 2711, 2714, 3224, 3415, 4311, 5223, 5323, 5828, 6112, 6113, 6121, 6122, 6349, 6546, 6642, 6646, 6674, 6741, 6750, 6880, 6912, 7187, 7213, 7433, 7521, 7524, 8941, 9110 }

9. **Remove products with little market share**: Select products whose market share is below the fifth percentile in year 2008. This selection results in the following 39 products:

   `ProductsToRemove` := `ProductsToRemove` ∪ { 0129, 0742, 2114, 2231, 2232, 2235, 2440, 2614, 2632, 2640, 2652, 2654, 2655, 2659, 2685, 2686, 2687, 2712, 2714, 2742, 2923, 3231, 3415, 4233, 4314, 6112, 6121, 6518, 6545, 6576, 6593, 6642, 6880, 6932, 7163, 7511, 7521, 7612, 7631 }

In the end, these filters remove 121 products (listed in Tables SM-1, SM-2, and SM-3) and the following 112 countries:

Afghanistan (AFG); American Samoa (ASM); Andorra (AND); Anguilla (AIA); Antarctica (ATA); Antigua and Barbuda (ATG); Armenia (ARM); Aruba (ABW); Bahamas (BHS); Bahrain (BHR); Barbados (BRB); Belize (BLZ); Benin (BEN); Bermuda (BMU); Bhutan (BTN); Bouvet Island (BVT); British Indian Ocean Territory (IOT); British Virgin Islands (VGB); Brunei (BRN); Burkina Faso (BFA); Burundi (BDI); Cape Verde (CPV); Cayman Islands (CYM); Central African Republic (CAF); Chad (TCD); Christmas Island (CXR); Cocos (Keeling) Islands (CCK); Comoros (COM); Cook Islands (COK); Cyprus (CYP); Djibouti (DJI); Dominica (DMA); Equatorial Guinea (GNQ); Eritrea (ERI); Falkland Islands (FLK); Faroe Islands (FRO); Fiji (FJI); French Guiana (GUF); French Polynesia (PYF); French South Antarctic Territory (ATF); Gambia (GMB); Gibraltar (GIB); Greenland (GRL); Grenada (GRD); Guam (GUM); Guinea-Bissau (GNB); Guyana (GUY); Haiti (HTI); Heard Island and McDonald Islands (HMD); Holy See (Vatican City) (VAT); Iceland (ISL); Iraq (IRQ); Kiribati (KIR); Laos (LAO); Lesotho (LSO); Luxembourg (LUX); Macau (MAC); Malawi (MWI); Maldives (MDV); Malta (MLT); Marshall Islands (MHL); Martinique (MTQ); Mauritania (MRT); Mauritius (MUS); Mayotte (MYT); Micronesia (FSM); Mongolia (MNG); Montenegro (MNE); Montserrat (MSR); Nauru (NRU); Nepal (NPL); Netherlands Antilles (ANT); New Caledonia (NCL); Niger (NER); Niue (NIU); Norfolk Island (NFK); Northern Mariana Islands (MNP); Pacific Island (US) (PCI); Palau (PLW); Palestine (PSE); Pitcairn Islands (PCN); Rwanda (RWA); Saint Helena (SHN); Saint Kitts and Nevis (KNA); Saint Lucia (LCA); Saint Pierre and Miquelon (SPM); Saint Vincent and the Grenadines (VCT); Samoa (WSM); San Marino (SMR); Sao Tome and Principe (STP); Seychelles (SYC); Sierra Leone (SLE); Solomon Islands (SLB); Somalia (SOM); South Georgia South Sandwich Islands (SGS); South Sudan (SSD); Suriname (SUR); Swaziland (SWZ); Taiwan (TWN); Timor-Leste (TLS); Togo (TGO); Tokelau (TKL); Tonga (TON); Turks and Caicos Islands (TCA); Tuvalu (TUV); United States Minor Outlying Islands (UMI); Vanuatu (VUT); Virgin Islands (VIR); Wallis and Futuna (WLF); Western Sahara (ESH); Yemen (YEM).

Final dataset After removing countries and products, we have a dataset of 138 countries, 665 products at the 4-digit level, and 6377 distinct (country, year) pairs. Summing export values at the 2-digit level results in 59 products. Merging this exports data with population data from the World Bank [46] and from [45] drops 240 (country, year) samples, resulting in 6137 distinct (country, year) pairs.

This dataset has on average 92% of the global population (minimum 86%, maximum 96%) and 77% of global trade (minimum 67%, maximum 85%). Time-series of those values are plotted in Figure SM-1.

SM-3.2 Normalize export values by population and by global exports

To make small and large countries comparable, we divide the value of a country c's exports of a product p in year t, denoted X_{cpt}, by a null model of a country's expected value of its exports of that product given that country's population, $\mathbb{E}\left[X_{cpt} \mid P_{ct}\right]$. To remove the effects of global price shocks, we divide this quantity by the total value of the world's exports of that product, which we also normalize by a null model that predicts global export value using global population. Formally, for each country c in a set of 123 countries $\mathcal{C}$ and for each product p in the set of 59 products $\mathcal{P}$, we define the *absolute advantage* of country c in product p as

$$\mathcal{R}_{cpt} := \frac{X_{cpt} / \mathbb{E}\left[X_{cpt} | P_{ct}\right]}{\sum_c X_{cpt} / \mathbb{E}\left[\sum_c X_{cpt} \middle| \sum_c P_{ct}\right]} \tag{SM-1}$$

SM-3.2.1 Null models of export values based on population size

Countries with more people tend to export more, but typically not in proportion to their population size. To allow for product-specific variation in the relationship between exports and population, we assume that the expectations in (SM-1) follow power laws of population size.

A country with population double that of another country typically exports more, but rarely does it export twice as much. For intuition, consider a disk-shaped country with its population distributed evenly

Table SM-1: 121 removed products (part 1)

0019	Live animals of a kind mainly used for human food, nes
0115	Meat of horses, asses, mules and hinnies, fresh, chilled or frozen
0129	Meat and edible meat offal, nes, in brine, dried, salted or smoked
0451	Rye, unmilled
0452	Oats, unmilled
0742	Mate
1122	Other fermented beverages, nes (cider, perry, mead, etc)
2114	Goat and kid skins, raw, whether or not split
2223	Cotton seeds
2226	Rape and colza seeds
2231	Copra
2232	Palm nuts and kernels
2234	Linseed
2235	Castor oil seeds
2239	Flour or meals of oil seeds or oleaginous fruit, non-defatted
2440	Cork, natural, raw and waste
2512	Mechanical wood pulp
2516	Chemical wood pulp, dissolving grades
2518	Chemical wood pulp, sulphite
2613	Raw silk (not thrown)
2614	Silk worm cocoons and silk waste
2632	Cotton linters
2634	Cotton, carded or combed
2640	Jute, other textile bast fibres, nes, raw, processed but not spun
2652	True hemp, raw or processed but not spun, its tow and waste
2654	Sisal, agave fibres, raw or processed but not spun, and waste
2655	Manila hemp, raw or processed but not spun, its tow and waste
2659	Vegetable textile fibres, nes, and waste
2685	Horsehair and other coarse animal hair, not carded or combed
2686	Waste of sheep's or lambs' wool, or of other animal hair, nes
2687	Sheep's or lambs' wool, or of other animal hair, carded or combed
2711	Animal or vegetable fertilizer, crude
2712	Natural sodium nitrate
2714	Potassium salts, natural, crude
2741	Sulphur (other than sublimed, precipitated or colloidal)
2742	Iron pyrites, unroasted
2784	Asbestos
2814	Roasted iron pyrites
2816	Iron ore agglomerates
2860	Ores and concentrates of uranium and thorium
2872	Nickel ores and concentrates; nickel mattes, etc

Table SM-2: 121 removed products (part 2)

2876	Tin ores and concentrates
2923	Vegetable plaiting materials
3221	Anthracite, not agglomerated
3222	Other coal, not agglomerated
3223	Lignite, not agglomerated
3224	Peat, not agglomerated
3231	Briquettes, ovoids, from coal, lignite or peat
3232	Coke and semi-coke of coal, of lignite or peat; retort carbon
3330	Crude petroleum and oils obtained from bituminous materials
3341	Gasoline and other light oils
3342	Kerosene and other medium oils
3343	Gas oils
3344	Fuel oils, nes
3345	Lubricating petroleum oils, and preparations, nes
3351	Petroleum jelly and mineral waxes
3352	Mineral tars and products
3353	Mineral tar pitch, pitch coke
3354	Petroleum bitumen, petroleum coke and bituminous mixtures, nes
3413	Petroleum gases and other gaseous hydrocarbons, nes, liquefied
3414	Petroleum gases, nes, in gaseous state
3415	Coal gas, water gas and similar gases
3510	Electric current
4233	Cotton seed oil
4236	Sunflower seed oil
4241	Linseed oil
4244	Palm kernel oil
4245	Castor oil
4311	Processed animal and vegetable oils
4314	Waxes of animal or vegetable origin
5163	Inorganic esters, their salts and derivatives
5223	Halogen and sulphur compounds of non-metals
5249	Other radio-active and associated materials
5323	Synthetic tanning substances; tanning preparations
5828	Ion exchangers of the condensation, polycondensation etc
6112	Composition leather, in slabs, sheets or rolls
6113	Calf leather
6121	Articles of leather use in machinery or mechanical appliances, etc
6122	Saddlery and harness, of any material, for any kind of animal
6344	Wood-based panels, nes
6349	Wood, simply shaped, nes

Table SM-3: 121 removed products (part 3)

6518	Yarn of regenerated fibres, put up for retail sale
6545	Fabrics, woven of jute or other textile bast fibres of heading 2640
6546	Fabrics of glass fibre (including narrow, pile fabrics, lace, etc)
6576	Hat shapes, hat-forms, hat bodies and hoods
6593	Kelem, Schumacks and Karamanie rugs and the like
6642	Optical glass and elements of optical glass (unworked)
6646	Bricks, tiles, etc of pressed or moulded glass, used in building
6674	Synthetic or reconstructed precious or semi-precious stones
6727	Iron or steel coils for re-rolling
6741	Universal plates of iron or steel
6750	Hoop and strip of iron or steel, hot-rolled or cold-rolled
6784	High-pressure hydro-electric conduit of steel
6793	Steel and iron forging and stampings, in the rough state
6831	Nickel and nickel alloys, unwrought
6880	Uranium depleted in U235, thorium, and alloys, nes; waste and scrap
6912	Structures and parts of, of aluminium; plates, rods, and the like
6932	Barbed iron or steel wire: fencing wire
7163	Rotary converters
7187	Nuclear reactors, and parts thereof, nes
7213	Dairy machinery, nes (including milking machines), and parts nes
7433	Free-piston generators for gas turbines and parts thereof, nes
7511	Typewriters; cheque-writing machines
7521	Analogue and hybrid data processing machines
7524	Digital central storage units, separately consigned
7612	Television receivers, monochrome
7631	Gramophones and record players, electric
7911	Rail locomotives, electric
7912	Other rail locomotives; tenders
7913	Mechanically propelled railway, tramway, trolleys, etc
7914	Railway, tramway passenger coaches, etc, not mechanically propelled
7924	Aircraft of an unladen weight exceeding 15000 kg
7931	Warships
8821	Chemical products and flashlight materials for use in photografy
8941	Baby carriages and parts thereof, nes
9110	Postal packages not classified according to kind
9310	Special transactions, commodity not classified according to class
9410	Animals, live, nes, (including zoo animals, pets, insects, etc)
9510	Armoured fighting vehicles, war firearms, ammunition, parts, nes
9610	Coin (other than gold coin), not being legal tender
9710	Gold, non-monetary (excluding gold ores and concentrates)

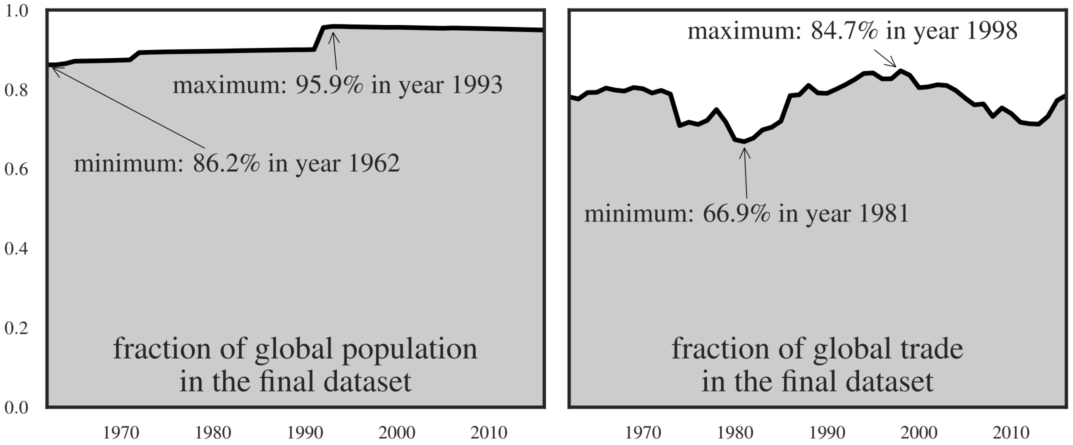

Figure SM-1: Fraction of global population and global trade in the dataset after the filters described in Sec. SM-3.1 are applied.

across space and with exports occurring at the border in proportion to the size of the perimeter. That country's exports increase with the square root of the population size. (This example is more extreme than reality: the exponent is ≈ 0.88 rather than 0.5.) Motivated by this intuition, we create a null model of exports by assuming that export value of a certain product, either by a certain country or by the whole world, grows with population size raised to some power, and that this exponent varies from one product to another. Specifically, we assume that

$$\mathbb{E}\left[X_{cpt}|P_{ct}\right] = \alpha_p \left(P_{ct}\right)^{\beta_p}, \tag{SM-2}$$

$$\mathbb{E}\left[\sum_c X_{cpt} \middle| \sum_c P_{ct}\right] = \gamma_p \left(\sum_c P_{ct}\right)^{\delta_p} \tag{SM-3}$$

With (SM-2) and (SM-3), our measure of a country c's *absolute advantage* in producing the product p in year t is

$$\mathcal{R}_{cpt} = \frac{X_{cpt}/\left(\alpha_p \left(P_{ct}\right)^{\beta_p}\right)}{\sum_c X_{cpt}/\left(\gamma_p \left(\sum_c P_{ct}\right)^{\delta_p}\right)}. \tag{SM-4}$$

This quantity $\mathcal{R}_{cpt}$ captures how proficient a country c is in exporting product p in year t, relative to an average country of its population size.

The distributions of the exponents β_p and δ_p are plotted in Figure SM-2. The exponents β_p have average value of 0.88; the minimum is 0.49 for the product *Dairy products and birds' eggs* (product code 02), and the maximum is 1.21 for the product *Crude rubber (including synthetic and reclaimed)* (product code 23). Thus, the export value of certain product tends to grow sublinearly with the population size, in accordance with the hypothetical disk-shaped country described above. Meanwhile, the exponents δ_p are much larger: the average (across all 59 products) is 5.37. The minimum is 2.50 for the product *Textile fibers (not wool tops) and their wastes (not in yarn)* (product code 26), and the maximum is 8.18 for the product *Office machines and automatic data processing equipment)* (product code 75). Thus, global exports of a product tend to grow superlinearly with global population.

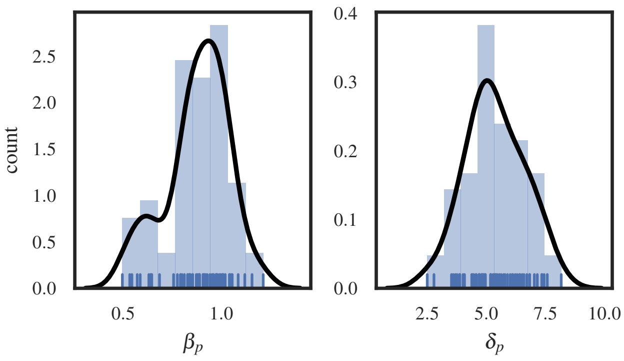

Figure SM-2: Distribution of exponents β_p and δ_p in the null models (SM-2) and (SM-3), respectively. For illustrative purposes, we draw in black a kernel density estimate with a Gaussian kernel.

SM-3.3 Logarithmically transforming data with lots of zeros in it

In this paper, we consider yearly export values X_{cpt} of 59 two-digit products. These export values range from zero to nearly a trillion US dollars per year. China, for example, has recently exported over $300 billion in *Electric machinery, apparatus and appliances, nes, and parts, nes* (product code 77) in one year. After normalizing by population and by global exports with (SM-1), the values are still rather heavy-tailed and range from 0 to 7.3×10^4 US dollars per year; see the left and middle panels of Figure SM-3.

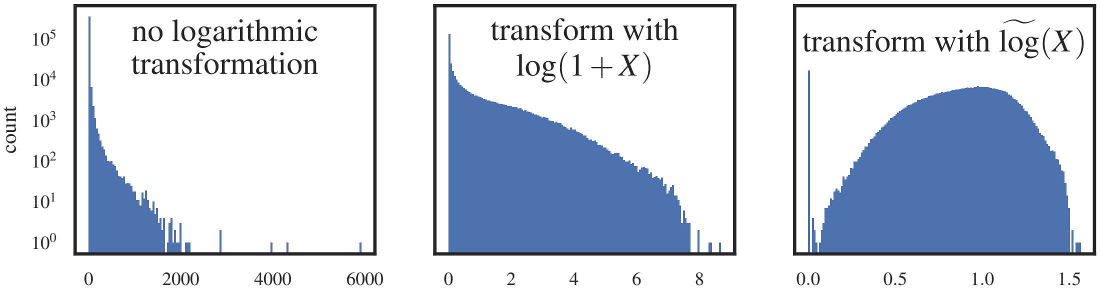

Figure SM-3: Histograms of the flattened data (SM-4) before it is logarithmically transformed (left panel), after it is logarithmically transformed with $\log(1+\cdot)$ (middle plot), and after it is logarithmically transformed with $\widetilde{\log}(\cdot)$ (right plot).

One way to logarithmically transform heavy-tailed data with zeros in it is to add one before applying the natural logarithm, so that zero maps to zero. However, we found that this transformation resulted in data that was approximately exponentially distributed rather than normally distributed, and we found that adding one introduces a scale in the data. To avoid these outcomes, we applied a different logarithmic transformation that is plotted in Fig. SM-4:

$$\widetilde{\log}(x) \equiv \begin{cases} 1 + s\log(x) & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases} \tag{SM-5}$$

where the scaling factor

$$s \equiv \lim_{z \to x_m} \frac{z-1}{\log(z)} = \begin{cases} 1 & \text{if } x_m = 1 \\ (x_m - 1) \,/\, \log x_m & \text{if } x_m \neq 1 \end{cases}, \tag{SM-6}$$

and x_m is the smallest positive value of all elements of the matrix X:

$$x_m \equiv \min \{x : x \in X, x > 0\} \,. \tag{SM-7}$$

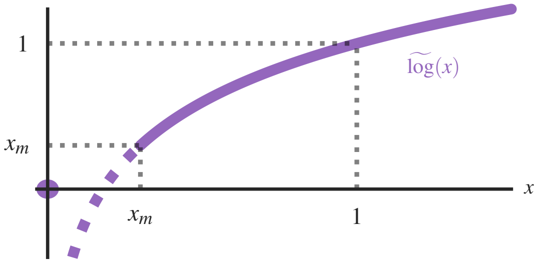

Figure SM-4: The logarithmic transformation (SM-5) used. It leaves unchanged zeros and whether values are above one.

The limit in (SM-6) ensures that s exists for all $x_m > 0$; in particular, $s = 1$ when $x_m = 1$. Note that

$$\widetilde{\log}(x_m) = x_m, \tag{SM-8}$$

$$\widetilde{\log}(1) = 1, \tag{SM-9}$$

$$\widetilde{\log}(x) \text{ is increasing.} \tag{SM-10}$$

Equations (SM-8) and (SM-9) are direct computations. Equation (SM-10) holds because $(z-1)\,/\,\log(z)$ is positive for $z > 0$. A consequence of (SM-9) and of (SM-10) is that

$$\widetilde{\log}(x) > 1 \text{ if and only if } x > 1. \tag{SM-11}$$

Statement (SM-11) is an important property for a logarithmic transformation of data like that studied here: because the data is normalized by dividing by the prediction of a null model, being above one (or not) is meaningful, so we wish our logarithmic transformation to preserve which values are above one and which values are below one.

SM-3.4 Centering and scaling

Next we pivot the data so that the rows are observations of a certain country in a certain year, and the columns are the values of $\mathcal{R}_{cpt}$ for each of the 59 many products p. We center and scale the columns using the pre-1989 column means and standard deviations:

$$R_{cpt} := \frac{\mathcal{R}_{cpt} - \mu\left(\{\mathcal{R}_{cpt} : 1962 \leq t \leq 1988, c \in \mathcal{C}\}\right)}{\sigma\left(\{\mathcal{R}_{cpt} : 1962 \leq t \leq 1988, c \in \mathcal{C}\}\right)} \tag{SM-12}$$

where μ denotes mean and σ denotes standard deviation. The column means and standard deviations, like all other preprocessing steps such as dimension reduction described next, are fit to data from year 1988 or earlier. That way, we can split the data into cross validation sets that are nested in time, and all preprocessing is done with the earliest set of data (years 1962 to 1988, inclusive).

SM-3.5 Reduce dimensions

Next we reduce dimensions using principal components analysis (PCA) [47]. Because the data was centered (see Section SM-3.4), PCA is equivalent to doing a truncated singular value decomposition. More insights from PCA applied to this exports data are given next in Sec. SM-4.

SM-4 Further analysis of the principal components

SM-4.1 Correlation between the loading on the first principal component and the Product Complexity Index

Recall from Fig. 2 that the first principal component loads positively on all products. But the loadings are not equal: the first principal component loads more on complex products like power generating machinery (product code `71`) that are produced by few countries, compared to simpler products like vegetables and fruit (product code `06`) that are produced by many countries. In fact, as shown in Fig. SM-5, these loadings are highly correlated with the Product Complexity Index [1], a notion of complexity (or knowledge intensity) of products based on the complexity (or knowledge intensity) of the countries that produce them. The second principal component is also correlated with the Product Complexity Index, but less so (Pearson correlation $\rho = 0.70$ versus $\rho = 0.81$).

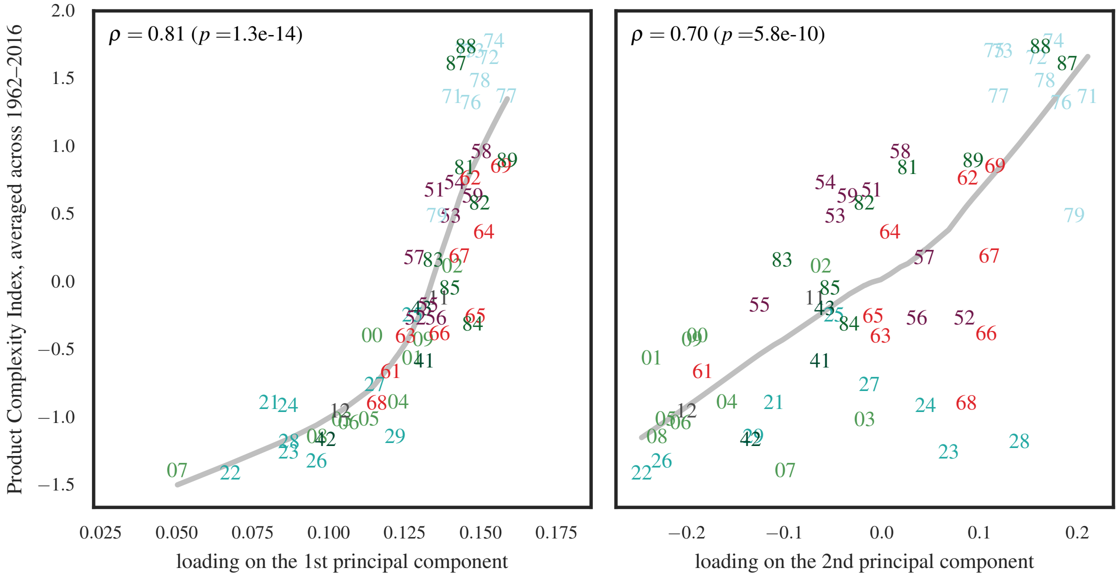

Figure SM-5: The loadings on the first principal component, and to a lesser degree the loadings on the second principal component, are highly correlated with the Product Complexity Index [1]. In these scatterplots, products are labeled by their 2-digit SITC product codes (available for download here), with colors denoting the first digit. To guide the eye, a locally weighted scatterplot smoothing (LOWESS) is shown in gray; this LOWESS was made using the package `seaborn` (DOI: https://doi.org/10.5281/zenodo.883859).

SM-4.2 Interpreting a country's score on the first principal component

SM-4.2.1 Pair-wise correlations

Figure SM-6 shows that a country's score on the first principal component, ϕ_0, is highly correlated with export value per capita [Pearson correlation $\rho = 0.82$, Fig. SM-6(D)]. However, compared to per-capita exports, ϕ_0 is more correlated with the Economic Complexity Index (ECI) and with the diversification of an export basket. In fact, ϕ_0 is central among all these quantities [as visualized in the graph diagram in Fig. SM-6(K)]: each of the three variables per-capita exports, ECI, and diversification is more similar to ϕ_0 than to any of the other variables. Because ϕ_0 is similar to per-capita exports but also captures aspects of economic complexity (ECI) and diversification, we interpret ϕ_0 as "complexity-weighted diversity".

Here, we consider the notion of diversification of exports used in [1, Equation 3], namely the number of products p such that the revealed comparative advantage RCA_{cpt} exceeds one:

$$\texttt{diversity}_{ct} := \{p : \mathrm{RCA}_{cpt} > 1\} . \tag{SM-13}$$

where

$$\mathrm{RCA}_{cpt} \equiv \frac{X_{cpt} / \sum_p X_{cpt}}{\sum_c X_{cpt} / \sum_{cp} X_{cpt}} .$$

Figure SM-6(G) indicates that export baskets with the highest score ϕ_0 on the first principal component tend to have RCA_{cpt} larger than one for approximately half of the 59 2-digit products, while the export baskets with the lowest ϕ_0 tend to have RCA_{cpt} larger than one for fewer than 10 out of the 59 2-digit products. Thus, the direction in the space of products in which export baskets over the past 50 years are most spread out is, loosely speaking, one that distinguishes undiversified, small export baskets from diversified, large ones.

SM-4.2.2 Intuition behind the correlations

Per-capita exports and the score ϕ_0 on the first principal component As shown in Fig. 2 and SM-5, the loadings of the first principal component are positive and range from 0.05 to 0.15. This homogeneity of the loadings means that the scores ϕ_0 captures an *average scaled absolute advantage*.

Consider a country with large export value per capita. It is likely that for certain product categories this country exports more than what one would expect for a country of its population size. Thus, this country's scaled absolute advantage is high for some products. Because ϕ_0 is a weighted sum of scaled absolute advantage with all positive weights, it follows that ϕ_0 is high. Therefore, we expect ϕ_0 to be highly correlated with the logarithm of exports per capita, as confirmed by $\rho = 0.82$ in Fig. SM-6(D).

However, large exports does not imply diversified exports. A country could simply export a lot of a small number of products. Next we motivate why the score ϕ_0 on the first principal component is correlated with diversification.

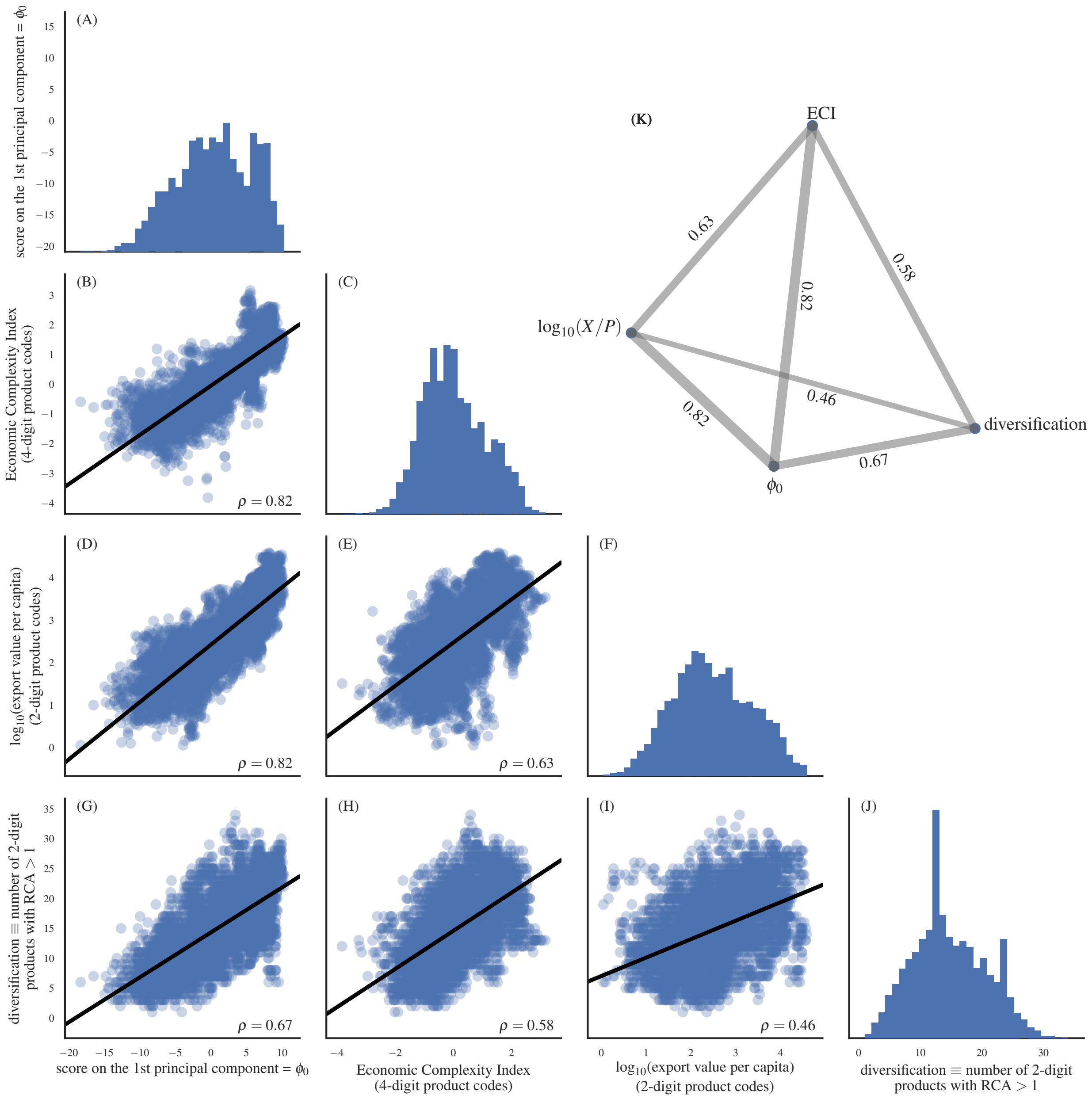

Figure SM-6: **In terms of Pearson correlations ρ, the score ϕ_0 on first principal component is equally similar to per-capita exports and to the economic complexity index [$\rho = 0.82$, panels (B), (D)], but it is also more similar to diversification than either of those quantities are [$\rho = 0.67$ versus $\rho = 0.46, 0.58$, panels (G), (I), (H)].** These correlations motivate our interpretation of ϕ_0 as "complexity-weighted diversification". In the scatterplots, the disks show each (country, year) sample, while the black line shows a least-squares regression. The diagonal shows histograms with 30 bins each. The graph diagram in panel (K) uses multidimensional scaling to illustrate the correlations; the line widths are proportional to ρ; this diagram merely guides the eye. The Economic Complexity Index [1] is taken from the same source as the data copied from the Atlas at Harvard's Center for International Development (see Sec. SM-2) and is computed from product codes at the 4-digit level for all products. All the other data in this figure is from the dataset analyzed in this paper, with countries and products filtered and aggregated at the 2-digit level as described in Sec. SM-3.1. To connect with other literature, diversification is computed here using revealed comparative advantage (SM-13).

Diversification and the score ϕ_0 on the first principal component Using a pair of approximations, we will motivate why ϕ_0 is correlated with the diversification of an export basket. Define the matrix M by

$$M_{cpt} = \begin{cases} 1, & \text{if } R_{cpt} > 0 \\ 0, & \text{if } R_{cpt} \leq 0 \end{cases} . \tag{SM-14}$$

This thresholding function can be roughly approximated by

$$M_{cpt} \approx \frac{1}{2}\left(1 + R_{cpt}\right). \tag{SM-15}$$

Although the approximation (SM-15) becomes quite rough for large $|R_{cpt}| > 2$, it is reasonably accurate for the small values of R_{cpt} that are frequently observed: the mean of R_{cpt} is 0.092; the standard deviation is 0.92; the inter-quartile range is $[-0.41, 0.75]$; and the 2.5^{th} and 97.5^{th} percentiles are -2.17 and 1.47.

Next, define diversification in terms of scaled absolute advantage as

$$d_{ct} = \sum_p M_{cpt}. \tag{SM-16}$$

Using the approximation (SM-15), we have

$$d_{ct} \approx \sum_p \frac{1}{2}\left(1 + R_{cpt}\right). \tag{SM-17}$$

Meanwhile, recall that the score on the first principal component of a country c in year t is

$$\phi_0(c,t) = \sum_p R_{cpt} l_0(p), \tag{SM-18}$$

where $l_0(p)$ is the loading of the first principal component on product p. The loadings of the first principal component across the 59 products, $\{l_0(p) : p \in \mathcal{P}\}$, are closely centered around their mean of 0.13, with a standard deviation of 0.024. Thus,

$$\phi_0(c,t) \approx 0.13 \times \sum_p R_{cpt}, \tag{SM-19}$$

The right-hand sides of (SM-17) and (SM-19) have Pearson correlation 1 because they are affine transformations of $\sum_p R_{cpt}$. Because the approximations in (SM-17) and (SM-19) are imperfect, the correlation of the left-hand sides of those two equations is 0.95. Figure SM-6(G) reports a correlation of 0.67 between ϕ_0 and diversification defined in terms of revealed comparative advantage (RCA) rather than in terms of scaled absolute advantage R_{cpt}.

SM-4.2.3 Regressions of ϕ_0

To investigate whether the score on the first principal component captures information beyond these three quantities Economic Complexity Index, log-exports, and diversification, we use the following datasets:

Worldwide Governance Indicators (WGI) from `http://info.worldbank.org/governance/wgi/index.aspx#home`. According to the source, this dataset comprises "aggregate and individual governance indicators for over 200 countries and territories over the period 1996–2016, for six dimensions of governance: Voice and Accountability, Political Stability and Absence of Violence, Government Effectiveness, Regulatory Quality, Rule of Law, and Control of Corruption."

Barro-Lee Educational Attainment Data from `http://barrolee.com/data/Lee_Lee_v1.0/LeeLee_v1.dta` or `http://www.barrolee.com/data/BL_v2.2/BL2013_MF1599_v2.2.csv`, which reports "educational attainment data for 146 countries in 5-year intervals from 1950 to 2010". It also provides information about the distribution of educational attainment of the adult population over age 15 and over age 25 by sex at seven levels of schooling: no formal education, incomplete primary, complete primary, lower secondary, upper secondary, incomplete tertiary, and complete tertiary. Average years of schooling at all levels—primary, secondary, and tertiary—are also measured for each country and for regions in the world.

International Data on Cognitive Skills from `http://hanushek.stanford.edu/sites/default/files/publications/hanushek%2Bwoessmann.cognitive.xls` which was studied in [48].

The question is: how much do the quantities and indicators in these datasets explain ϕ_0?

Figures SM-7, SM-8, SM-9, and SM-10 show the results of the standardized coefficients for different univariate and multivariate regressions. While all regressors predict ϕ_0 to some extent when we carry out univariate regressions, when all are put together only exports per capita, diversity [Eq. (SM-16)], government effectiveness, and rule of law survive. In the multivariate regressions done per year, the coefficients for exports per capita and diversity are consistently significant and positive, and have similar magnitudes. In light of these regressions and of the relationship between the loadings on the first principal component with product complexity (Fig. SM-5), in the main text we refer to the score ϕ_0 on the first principal component as "complexity–weighted diversity".

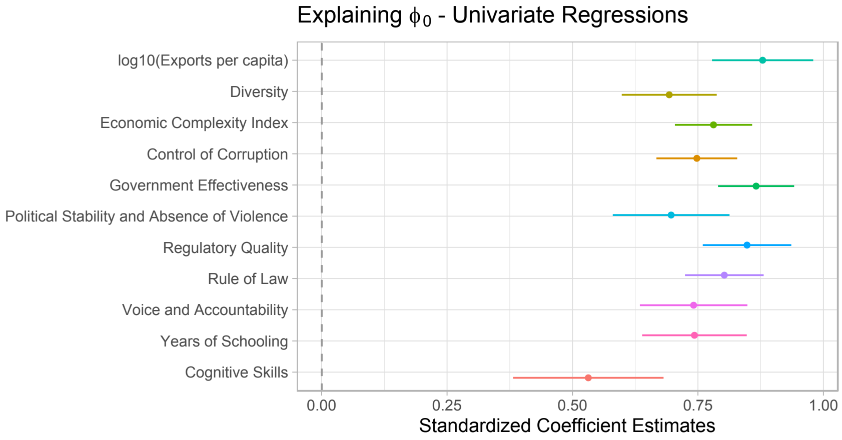

Figure SM-7: **Coefficients of the predictors from univariate regressions.** All regressions included year-specific fixed-effects; errors are clustered by country; and error bars reflect 95% confidence intervals. The estimates are for standardized coefficients (i.e., the variables are standardized to have zero mean and unit variance).

SM-4.3 More ways of interpreting the first three principal components

One-digit product codes To help interpret the principal components, in Fig. SM-11 we group the 59 products at the 1-digit level and plot the mean and standard deviation. Recall from Sec. SM-3.1 that we removed product category `3` (i.e., all products with SITC product code that begins with `3`) to avoid focusing on endowments of fossil fuels.

Most and least loaded 2-digit products in the second and third principal components Figure SM-12 shows the top 10 most and least loaded products in the second and third principal components. These

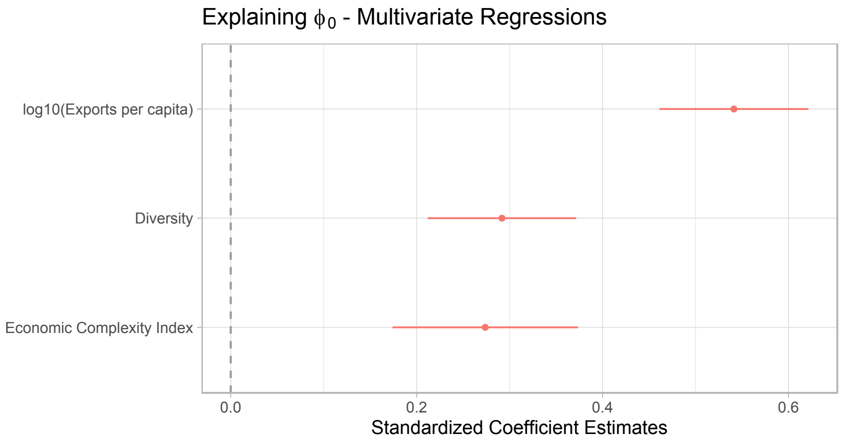

Figure SM-8: **Coefficients of the predictors from a multivariate regression only including exports per capita, diversity and economic complexity index.** This regression included year-specific fixed-effects; errors are clustered by country; and error bars reflect 95% confidence intervals. The estimates are for standardized coefficients.

“top 10 lists” aid the interpretation of the first three principal components.

SM-4.4 Substituting per-capita exports or diversification for the score ϕ_0 on the first principal component suggests that diversification, not simply a rise in total exports per capita, precedes economic growth

Rich countries are usually big exporters, have diversified economies, and produce complex products. Hence, these quantities correlate positively with each other, which makes it particularly difficult to interpret the meaning of scores resulting from PCA. We find in the main text that high levels in ϕ_0 precede growth in income. But the scores of ϕ_0 are correlated with both exports per-capita and product diversification, so this relationship could mean either that high levels of export per-capita precede growth, or that high levels of diversification precede growth, or both. To better understand what ϕ_0 represents in our analysis and what it reveals about economic development, here we substitute another variable for it in the GAM: either per-capita export value, or another definition of “product diversification”.

SM-4.4.1 Exports per capita are less strongly associated with growth in incomes compared to ϕ_0

Figure SM-13 shows a partial dependence plot (like Fig. 3) from a model fitted to the same dataset except that ϕ_0 is replaced by total export value per capita. Note in particular the bottom-left plot: The 95% confidence interval of the relationship between economic growth and export value per capita contains zero or is slightly below zero, suggesting a weak relationship between rises exports (no matter the product) and economic growth. Contrast this flat relationship with the positive trend in the bottom-left plot of Fig. 3.

SM-4.4.2 Replacing ϕ_0 with another notion of diversification results in qualitatively similar results

Next we tried replacing ϕ_0 by diversification as defined in Eq. [3] in [1]: the number of products with revealed comparative advantage (RCA) larger than one. The resulting GAM is approximately linear and behaves qualitatively similarly to the model in the main text; in fact, it appears to be a linear approximation of that

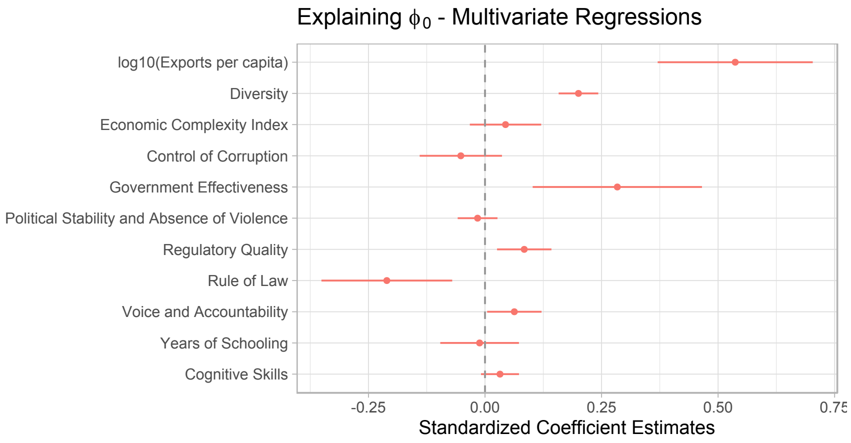

Figure SM-9: **Coefficients of the predictors from a multivariate regression including.** This regression included year-specific fixed-effects; errors are clustered by country; and error bars reflect 95% confidence intervals. The estimates are for standardized coefficients.

GAM. In light of this resemblance to the model in the main text, it seems reasonable to call ϕ_0 something akin to diversification; here, we call ϕ_0 "complexity-weighted diversification". The comparison between Figs. 3, SM-13, and SM-14 suggests that exporting a large diversity of complex products precedes economic growth.

SM-5 Details about the generalized additive model: Training and performance

SM-5.1 Cubic smoothing splines using the B-spline basis

Generalized additive models were estimated using the package `pyGAM 0.2.17` [49], which uses a B-spline basis, computed using De Boor recursion. The basis functions extrapolate linearly past the end-knots. Details on cubic smoothing splines are in [50, Chapters 3 and 4] and [51, Sec. 5.4].

SM-5.1.1 Nested-in-time cross validation

The generalized additive model (GAM) (1.2) has two hyperparameters: the smoothing strength λ that penalizes wiggliness, and the number of splines. Following the advice of [50], we tried relatively large values for the number of splines (uniformly distributed over $\{15, 16, \ldots, 60\}$) and let the smoothing penalty do the regularization. We sampled $\log_{10} \lambda$ uniformly over $[-3.0, 10.0]$.

To choose the best hyperparameters, we split the data into five training sets that are nested in time as follows. The task is to predict the change in the time-series between year $t-1$ and t given the value of the time-series at year $t-1$ (i.e., autoregression with lag 1). We put the earliest 39% of samples in the first training set, and then we partition the remaining samples into roughly equal-size sets. (Because countries appear and disappear, some care needs to be taken with time-series of different lengths; we use quantiles of the times of all the samples to find where to split the data.) The result is that the first training set is data with t between 1962 and 1988, and the corresponding test set is data with t between 1989 and 1995. The train–test splits are shown in Table SM-4.

The model is always tested on data from the future relative to the test set. With this cross validation scheme, the hyperparameters with best performance on the test sets were smoothing strength $\lambda = 2748.5$ and 37 splines. These values were used for each of the three equations in equation (1.2).

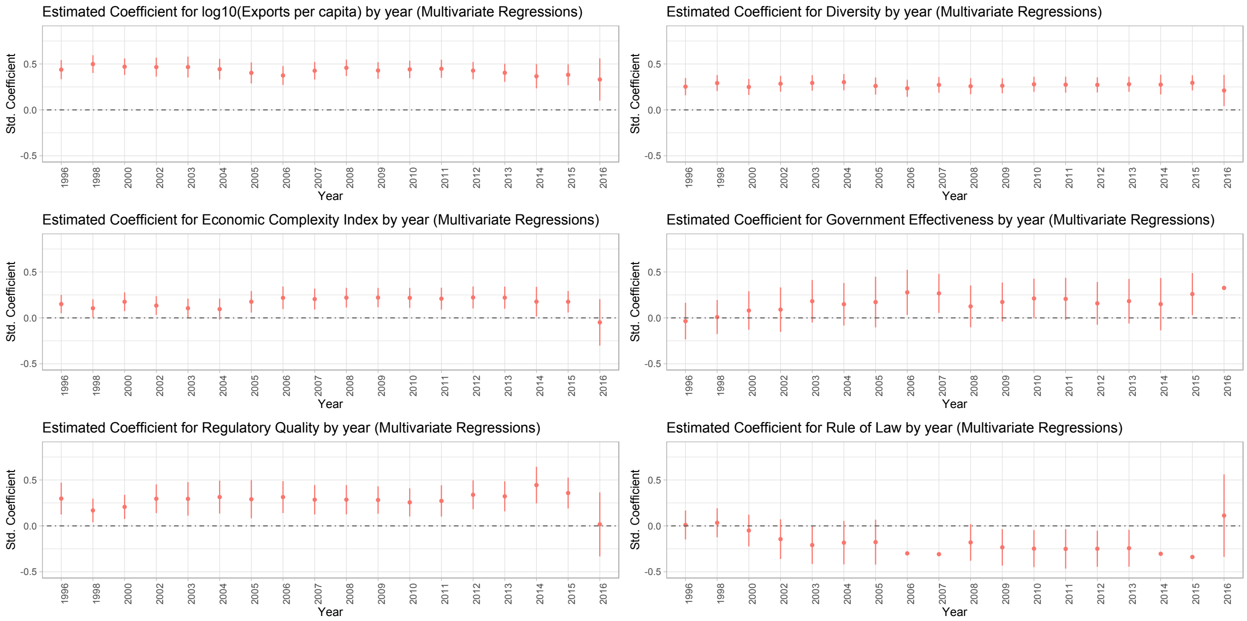

Figure SM-10: **Coefficients of the predictors from multivariate regressions carried separately by year.** The estimates are for standardized coefficients.

SM-5.1.2 The GAM outperforms a baseline model that predicts the average change in the test set

Table SM-4 shows the performance of the GAM (1.2) in terms of the coefficient of determination, R^2. The GAM does better than a simple baseline model that simply predicts the average change observed in the training set. (This baseline model predicted the test set more accurately than a baseline model that predicted the median of the training set.) However, the performance of the GAM has considerable room for improvement: its R^2, averaged across the three prediction problems (of predicting ϕ_0, ϕ_1, and `GDPpc`) is 0.049 on the training sets and -0.066 on the test sets. Another popular general-purpose learning method, gradient boosting, performed similarly to GAMs (see Table SM-4), but we found GAMs to be easier to interpret.

We tried several alternative modeling strategies. We rejected (a) neural networks because they excel on datasets larger than this one; (b) kernel ridge regression because of difficulty of interpreting the model; (c) decision trees and ensembles of them (random forests, gradient boosting) because we sought a model that changes smoothly as a function of the inputs. Future methodological work could investigate interactions among different variables in the model. Another potentially fruitful avenue for future work is to combine the dimension reduction and prediction into one pipeline that is trained simultaneously (rather than as two independent steps fitted independently, as done here).

SM-5.2 Quantile–quantile plots and raising the response variables to the 1/2 power

Figure SM-15 shows a quantile-quantile (QQ) plot of the residuals of generalized additive models of the form (1.2) with the response transformed by $g(x) \equiv \operatorname{sign}(x)|x|^{1/2}$ (bottom row) or not (top row). The residuals are conditioned on the fitted model coefficients and scale parameter. The closer the QQ-plot is to a straight line, the better the distributional assumptions are satisfied. The QQ-plots were made using the function `qq.gam` in the package `mgcv` by Simon Wood [52].

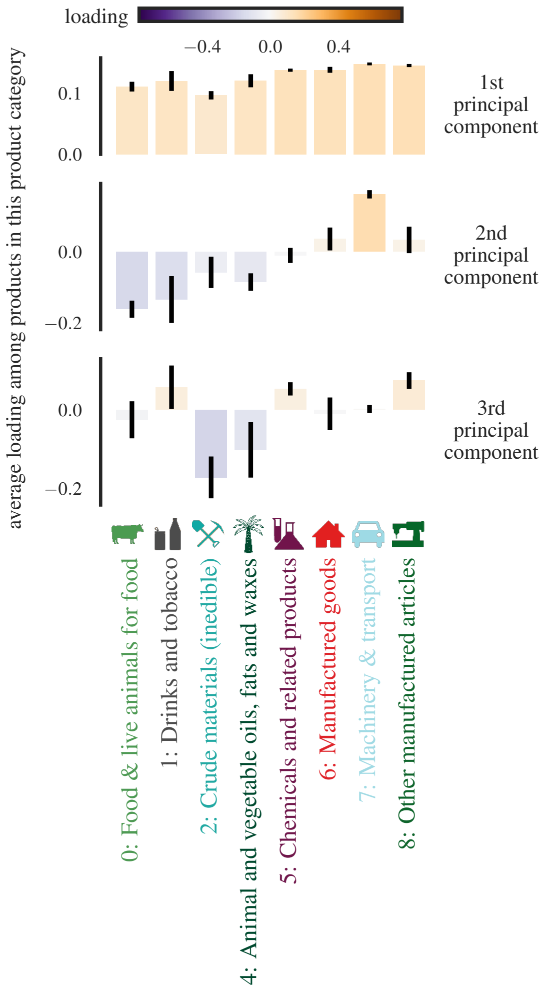

Figure SM-11: PCA loadings grouped and averaged at the 1-digit level. Error bars show the standard deviation of the loadings divided by the square root of the number of 2-digit products.

Because the QQ-plots are closer to a straight line when we transform the response with $g\,(x) \equiv \text{sign}(x)|x|^{1/2}$ (compare bottom row and top row in Fig. SM-15), we transform the response variables with the invertible function $g\,(x) \equiv \text{sign}(x)|x|^{1/2}$. (An example of such a transformation based on the results of `qq.gam` is given on page 230 in Section 5.2.1 in [50].) When making iterated predictions (as in Figure 6), we invert $g\,(x)$ in order to feed the response back into the model as a predictor variable.

SM-5.3 Errors averaged by country

This model tends to be more accurate for more developed countries, as shown in Figure SM-16. Figure SM-16 shows the squared errors averaged by country, for predicting export baskets that have been dimension-reduced with PCA (left column) and for predicting the export baskets themselves (right-column).

The trajectories of poorer countries in Figures 4 and 5 in the main text appear to be laminar. By contrast, Cristelli et al. [6] found that the trajectories of the poorest countries are turbulent when they analyzed yearly changes in “fitness” and per-capita incomes. The dynamics in our model (1.2) are laminar because a large

Table SM-4: Train–test splits used in cross validation, and coefficient of determination (R^2) averaged across the three prediction problems (predict annual changes in ϕ_0, ϕ_1, and `GDPpc`) for the GAM (1.2), a baseline ("Dummy") model that predicts the average change observed in the training set, linear regression, and gradient-boosting regression. The model's task is to predict year t using data about year $t-1$. The GAM behaves comparably with gradient boosting but is more interpretable.

		Average R^2 across 3 targets			
		Dummy	Linear	GAM	GBR
Average	Train	0	0.031	0.049	0.059
	Test	-0.080	-0.076	-0.066	-0.063
Split 0	Train 1962–1988	0	0.027	0.048	0.068
	Test 1988–1995	-0.026	0.028	0.031	0.022
Split 1	Train 1962–1995	0	0.037	0.056	0.068
	Test 1995–2001	-0.051	-0.031	-0.030	-0.039
Split 2	Train 1962–2001	0	0.035	0.053	0.059
	Test 2001–2006	-0.118	-0.132	-0.132	-0.109
Split 3	Train 1962–2006	0	0.031	0.048	0.054
	Test 2006–2011	-0.008	-0.033	-0.018	-0.018
Split 4	Train 1962–2011	0	0.025	0.042	0.045
	Test 2011–2016	-0.213	-0.236	-0.202	-0.187

smoothing strength is chosen in cross validation (Sec. SM-5.1.1). However, the greater predictability of richer countries (Fig. SM-16) is consistent with the finding of Cristelli et al. [6] that richer countries move in a more laminar, predictable path through the space defined by per-capita income and by a summary measure of export baskets.

SM-5.4 Alignment of changes in export baskets with the gradient of per-capita incomes

SM-5.4.1 Countries tend to "hill climb" to higher incomes

Do economies' export baskets change in ways that lead to rising incomes? To explore that question, we plot in the left column of Figure SM-17 the direction in (ϕ_0, ϕ_1) that would most increase per-capita incomes [i.e., the gradient $(s'_{20}(\phi_0), s'_{20}(\phi_1))$. For comparison, in the right-hand column of Figure SM-17 we plot the typical movement in (ϕ_0, ϕ_1) according to the fitted model (1.2). In these plots on the right-hand column, at each point in a fine grid of points, we find the `GDPpc` of the closest sample to that grid point. This procedure results in more wiggles in the streamlines compared to when `GDPpc` is fixed at a certain value, as in Figure 5 in the main text.

By comparing the left- and right-hand plots in Figure SM-17, we see how well economies tend to "hill climb" toward higher per-capita incomes according to the model (1.2). Except for two extreme points where few observations are found (very high and very low ϕ_1), countries do tend to move along the gradient of per-capita income.

Figure SM-18 shows the cosine similarity of countries' movement in (ϕ_0, ϕ_1) and the gradient of how Δ`GDPpc` depends on (ϕ_0, ϕ_1). By this measure, countries with higher incomes tend to be better hill climbers (i.e., high cosine similarity), and China has become an unusually good hill climber since the late 1990s, while Madagascar has only recently moved slightly aligned with the gradient of per-capita incomes.

References

[1] Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 10570–10575 (2009). URL https://www.pnas.org/content/106/26/10570. https://www.pnas.org/content/106/26/10570.full.pdf.

[2] Hausmann, R. *et al. The Atlas of Economic Complexity: Mapping Paths to Prosperity* (Puritan Press, 2011).

[3] Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. A New Metrics for Countries' Fitness and Products' Complexity. *Scientific Reports* **2**, 482–7 (2012). URL http://www.nature.com/articles/srep00723.

[4] Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G. & Pietronero, L. Measuring the Intangibles: A Metrics for the Economic Complexity of Countries and Products. *PloS one* **8**, 70726 (2013). URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2013PLoSO...870726C&link_type=EJOURNAL.

[5] Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. Economic complexity: Conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamics and Control* **37**, 1683–1691 (2013). URL http://www.sciencedirect.com/science/article/pii/S0165188913000833. Rethinking Economic Policies in a Landscape of Heterogeneous Agents.

[6] Cristelli, M., Tacchella, A. & Pietronero, L. The Heterogeneous Dynamics of Economic Complexity. *PloS one* **10**, e0117174–15 (2015). URL http://dx.plos.org/10.1371/journal.pone.0117174.

[7] Teza, G., Caraglio, M. & Stella, A. L. Growth dynamics and complexity of national economies in the global trade network. *Scientific Reports* **8**, 1–8 (2018). URL http://dx.doi.org/10.1038/s41598-018-33659-6.

[8] Albeaik, S., Kaltenberg, M., Alsaleh, M. & Hidalgo, C. A. Improving the Economic Complexity Index (2017). URL http://arxiv.org/abs/1707.05826v3. ArXiv:1707.05826v3, 1707.05826v3.

[9] Gabrielli, A. *et al.* Why we like the ECI+ algorithm (2017). URL https://arxiv.org/abs/1708.01161v1. ArXiv:1708.01161, 1708.01161.

[10] Albeaik, S., Kaltenberg, M., Alsaleh, M. & Hidalgo, C. A. 729 new measures of economic complexity (Addendum to Improving the Economic Complexity Index) (2017). URL http://arxiv.org/abs/1708.04107v1. ArXiv:1708.04107v1, 1708.04107v1.

[11] Pietronero, L. *et al.* Economic Complexity: "Buttarla in caciara" vs a constructive approach (2017). URL https://arxiv.org/abs/1709.05272v1. ArXiv:1709.05272, 1709.05272.

[12] Machado, J. A. T. & Mata, M. E. Analysis of World Economic Variables Using Multidimensional Scaling. *PloS one* **10**, e0121277–17 (2015). URL http://dx.plos.org/10.1371/journal.pone.0121277.

[13] Hruschka, D. J., Hadley, C. & Hackman, J. Material wealth in 3D: Mapping multiple paths to prosperity in low- and middle- income countries. *PloS one* **12**, e0184616–18 (2017). URL http://dx.plos.org/10.1371/journal.pone.0184616.

[14] Turchin, P. *et al.* Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E144–E151 (2018). URL https://www.pnas.org/content/115/2/E144.

[15] Quah, D. T. Empirical cross-section dynamics in economic growth. *European Economic Review* **37**, 426–434 (1993). URL http://www.sciencedirect.com/science/article/pii/0014292193900315.

[16] Quah, D. T. Convergence empirics across economies with (some) capital mobility. *Journal of Economic Growth* **1**, 95–124 (1996). URL `http://www.jstor.org/stable/40215883`.

[17] Azariadis, C. & Stachurski, J. A forward projection of the cross-country income distribution. *Institute of Economic Research, Kyoto University, Discussion Paper No* **570** (2003). URL `https://pdfs.semanticscholar.org/82c0/ddc18900bc6a05a360b21a5276a28aebcbc9.pdf`.

[18] Azariadis, C. & Stachurski, J. Poverty Traps. In Aghion, P. & Durlauf, S. (eds.) *Handbook of Economic Growth*, 295–384 (Elsevier, 2005). URL `http://hdl.handle.net/11343/34380`.

[19] Caraglio, M., Baldovin, F. & Stella, A. L. Export dynamics as an optimal growth problem in the network of global economy. *Scientific Reports* 1–10 (2016). URL `http://dx.doi.org/10.1038/srep31461`.

[20] Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 3932–3937 (2016). URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1517384113`.

[21] Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Science Advances* **3**, e1602614 (2017). URL `http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1602614`.

[22] Mangan, N. M., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **2**, 52–63 (2016). URL `https://ieeexplore.ieee.org/document/7809160`.

[23] Mangan, N. M., Kutz, J. N., Brunton, S. L. & Proctor, J. L. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**, 20170009 (2017). URL `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2017.0009`.

[24] Brunton, S. L., Proctor, J. L. & Kutz, J. N. Sparse identification of nonlinear dynamics with control (sindyc). *IFAC-PapersOnLine* **49**, 710 – 715 (2016). URL `http://www.sciencedirect.com/science/article/pii/S2405896316318298`. 10th IFAC Symposium on Nonlinear Control Systems NOLCOS 2016.

[25] Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 9943–9948 (2007). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.0609476104`.

[26] Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.1165893`.

[27] Zhang, L. & Li, K. Forward and backward least angle regression for nonlinear system identification. *Automatica* **53**, 94–102 (2015). URL `http://dx.doi.org/10.1016/j.automatica.2014.12.010`.

[28] Daniels, B. C. & Nemenman, I. Automated adaptive inference of phenomenological dynamical models. *Nature Communications* **6**, 1–8 (2015). URL `http://dx.doi.org/10.1038/ncomms9133`.

[29] Box, G., Jenkins, G., Reinsel, G. & Ljung, G. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics (Wiley, 2015). URL `https://books.google.com/books?id=EjXHCQAAQBAJ`.

[30] J Bai, S. N. Forecasting economic time series using targeted predictors. *Journal of Econometrics* **146**, 304–317 (2008). URL `http://dx.doi.org/10.1016/j.jeconom.2008.08.010`.

[31] Bustos, S. & Yildirim, M. A. Uncovering trade flows (2018). Forthcoming.

[32] Bhagwati, J. On the underinvoicing of imports1. *Bulletin of the Oxford University Institute of Economics & Statistics* **27**, 389–397 (1964). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0084.1964.mp27004007.x/abstract.

[33] Naya, S. & Morgan, T. The accuracy of international trade data: The case of southeast asian countries. *Journal of the American Statistical Association* **64**, 452–467 (1969). URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10500987.

[34] Yeats, A. J. On the accuracy of partner country trade statistics. *Oxford Bulletin of Economics and Statistics* **40**, 341–361 (1978). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0084.1978.mp40004004.x/abstract.

[35] Yeats, A. J. On the accuracy of economic observations: Do sub-saharan trade statistics mean anything? *The World Bank Economic Review* **4**, 135–156 (1990). URL https://academic.oup.com/wber/article/4/2/135/1643142/On-the-Accuracy-of-Economic-Observations-Do-Sub.

[36] Rozanski, J. & Yeats, A. On the (in)accuracy of economic observations: An assessment of trends in the reliability of international trade statistics. *Journal of Development Economics* **44**, 103–130 (1994). URL http://www.sciencedirect.com/science/article/pii/0304387894000085.

[37] Gehlhar, M. Reconciling bilateral trade data for use in GTAP. *GTAP Technical Papers* (1996). URL http://docs.lib.purdue.edu/gtaptp/11.

[38] Makhoul, B. & Otterstrom, S. M. Exploring the accuracy of international trade statistics. *Applied Economics* **30**, 1603–1616 (1998). URL http://dx.doi.org/10.1080/000368498324689.

[39] Beja, E. L. Estimating trade mis-invoicing from china: 20002005. *China & World Economy* **16**, 82–92 (2008). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1749-124X.2008.00108.x/abstract.

[40] Ferrantino, M. J. & Wang, Z. Accounting for discrepancies in bilateral trade: The case of china, hong kong, and the united states. *China Economic Review* **19**, 502–520 (2008). URL http://www.sciencedirect.com/science/article/pii/S1043951X08000163.

[41] Barbieri, K., Keshk, O. M. & Pollins, B. M. Trading data: Evaluating our assumptions and coding rules. *Conflict Management and Peace Science* **26**, 471–491 (2009). URL http://dx.doi.org/10.1177/0738894209343887.

[42] Gaulier, G. & Zignago, S. BACI: International trade database at the product-level (the 1994-2007 version). SSRN Scholarly Paper ID 1994500, Social Science Research Network (2010). URL https://papers.ssrn.com/abstract=1994500.

[43] Guo, D. *Mirror Statistics of International Trade in Manufacturing Goods: The Case of China* (United Nations Industrial Development Organization, 2010). URL https://books.google.com/books/about/Mirror_Statistics_of_International_Trade.html?id=IGaQtwAACAAJ.

[44] Ferrantino, M. J., Liu, X. & Wang, Z. Evasion behaviors of exporters and importers: Evidence from the U.S.–China trade data discrepancy. *Journal of International Economics* **86**, 141–157 (2012). URL http://www.sciencedirect.com/science/article/pii/S0022199611000924.

[45] Exports data from the Center for International Development, Harvard University. Export values and population (2016). URL https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FB0ASZU.

[46] World Bank. GDP per capita, indicator NY.GDP.PCAP.KD, expressed in constant 2010 USD (2016). URL https://data.worldbank.org/indicator/NY.GDP.PCAP.KD.

[47] Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. *Nature Publishing Group* **14**, 641–642 (2017). URL `http://dx.doi.org/10.1038/nmeth.4346`.

[48] Hanushek, E. A. & Woessmann, L. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth* **17**, 267–321 (2012). URL `https://doi.org/10.1007/s10887-012-9081-x`.

[49] Servén, D. & Brummitt, C. pyGAM: Generalized Additive Models in Python (version 0.2.17) (2018). URL `https://pypi.python.org/pypi/pygam/0.2.17`. DOI: 10.5281/zenodo.1226652.

[50] Wood, S. *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC Texts in Statistical Science (Taylor & Francis, Boca Raton, FL, 2006). URL `https://books.google.com/books?id=GbzXe-L8uFgC`.

[51] Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning.* Springer Series in Statistics (Springer New York, New York, NY, 2009), second edn. URL `http://link.springer.com/10.1007/978-0-387-84858-7`.

[52] Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* **73**, 3–36 (2011). URL `https://www.jstor.org/stable/41057423`.

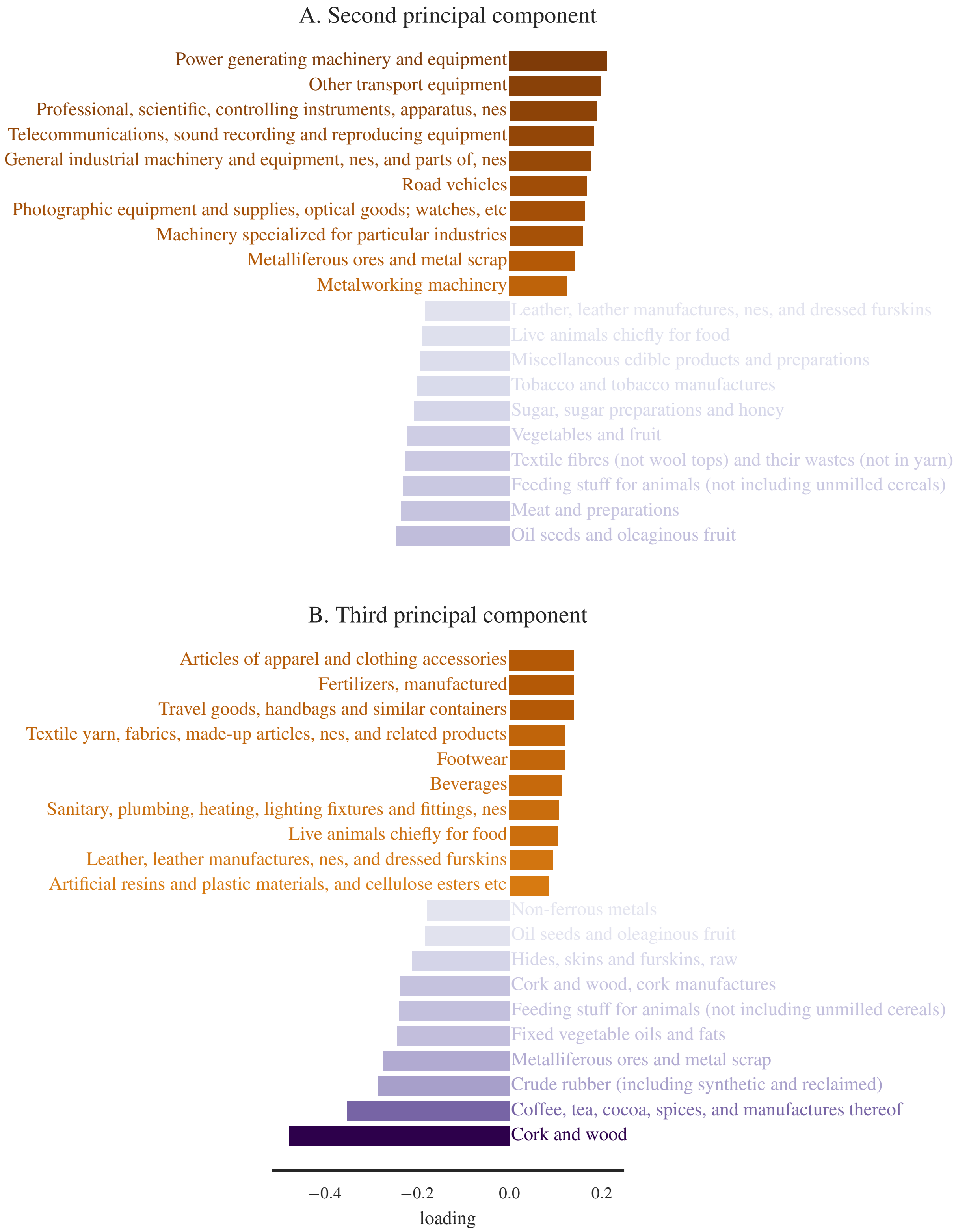

Figure SM-12: Most and least loaded products in the second and third principal components for the data on absolute advantage (SM-1).

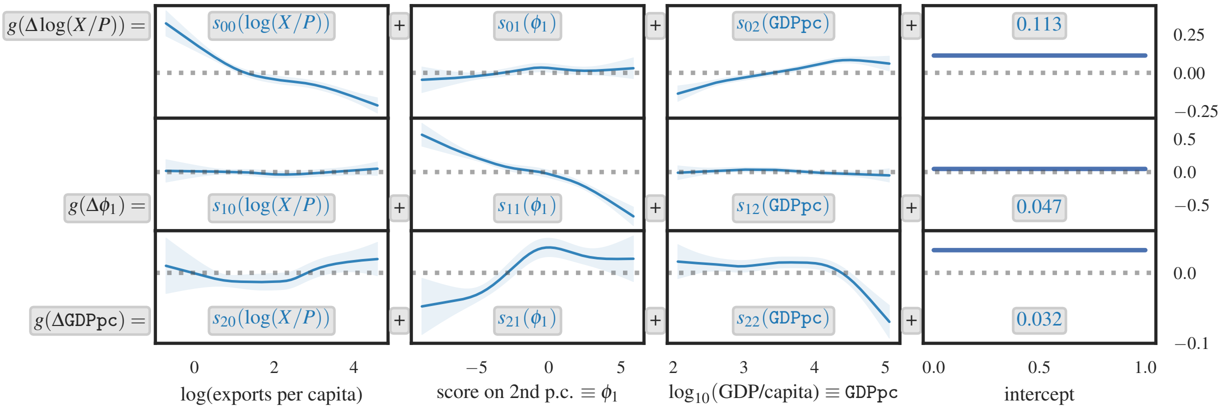

Figure SM-13: Substituting total export value per capita X/P for ϕ_0 results in a flat relationship between X/P and growth in income (bottom-left plot). Compare this partial dependence with Fig. 3 and Fig. SM-14.

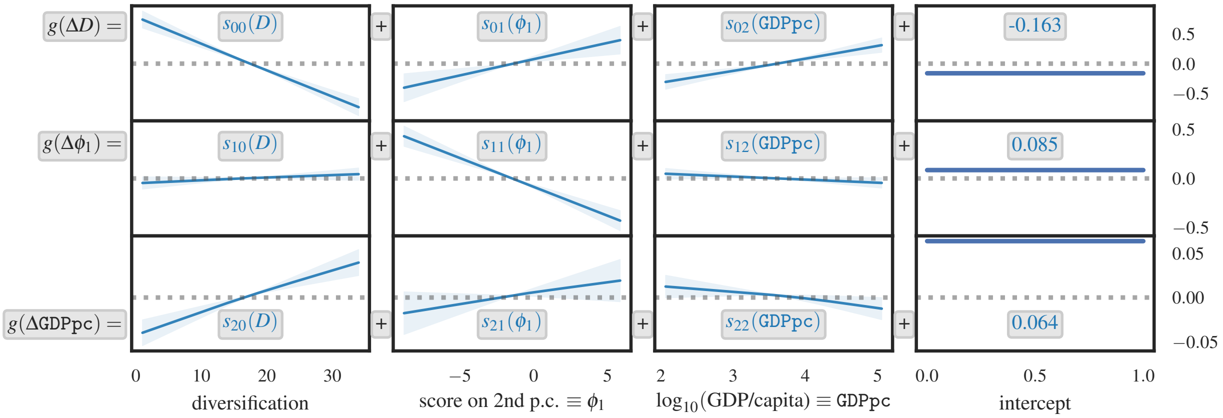

Figure SM-14: Replacing ϕ_0 with the definition of diversification defined in Eq. [3] in [1] (the number of products with revealed comparative advantage (RCA) larger than one) results in a qualitatively similar model that is more linear than the one discussed in the main text. Compare this partial dependence with Fig. 3 and Fig. SM-13.

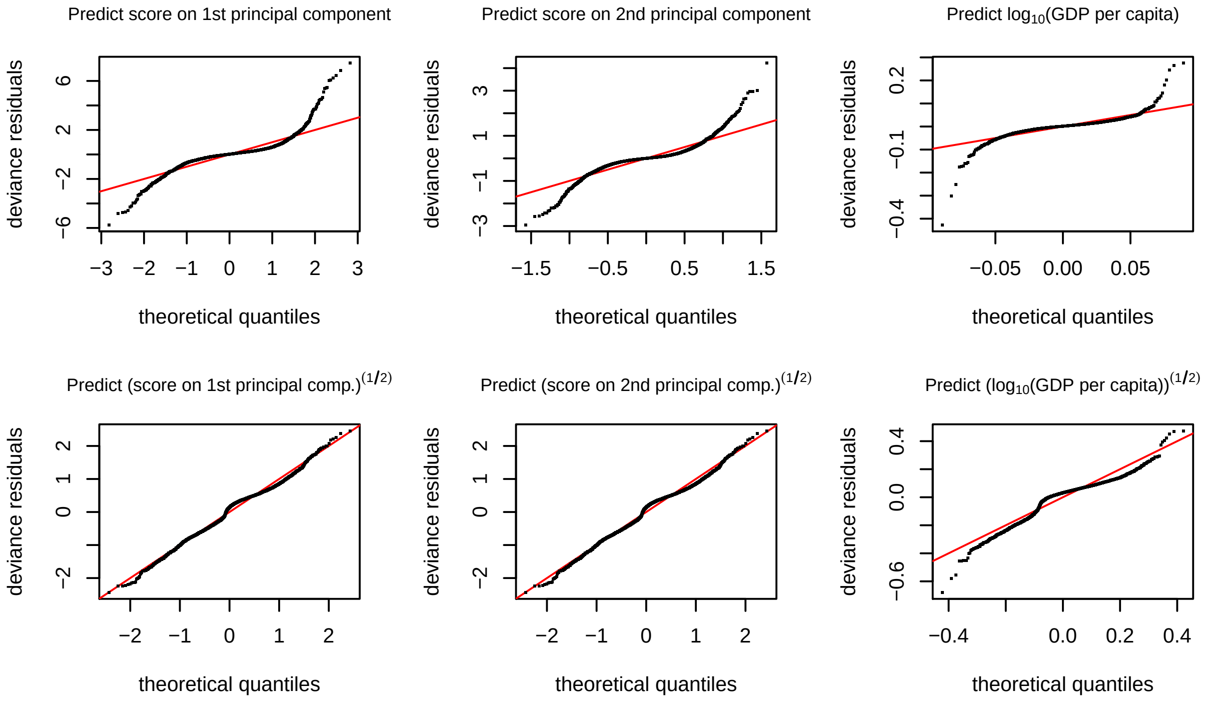

Figure SM-15: Quantile-quantile plots for the task of predicting yearly changes $\phi_0(t+1) - \phi_0(t)$, $\phi_1(t+1) - \phi_1(t)$, $\texttt{GDPpc}(t+1) - \texttt{GDPpc}(t)$ (top row, left to right) and for predicting those yearly transformed by $g\,(x) \equiv \text{sign}(x)|x|^{1/2}$ (bottom row). In the top row, we see significant improvement in how close the deviance residuals are to the straight red line.

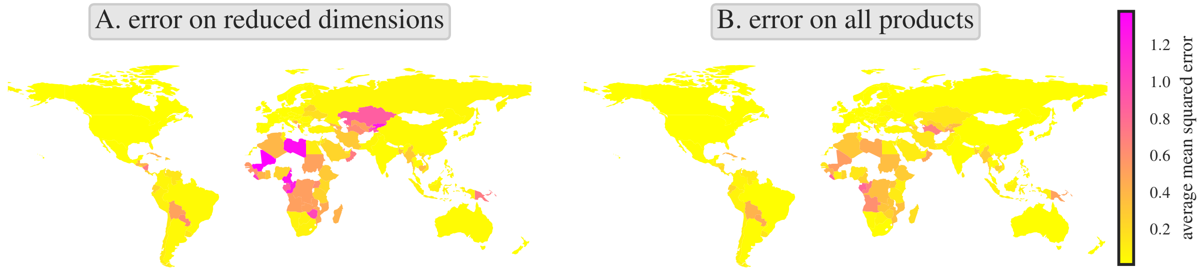

Figure SM-16: **The inferred model tends to make larger errors in predicted changes of export baskets and per-capita incomes of low-income countries, especially in Africa.** Plotted are the squared errors in predicting export baskets and per-capita incomes, averaged across columns [i.e., across $(\phi_0, \phi_1, \texttt{GDPpc})$] and then averaged across time. The left-hand column shows the error on the reduced dimensions $(\phi_0, \phi_1, \texttt{GDPpc})$, while the right-hand column shows the errors after the principal component scores (ϕ_0, ϕ_1) are inverted back to the original dimensions corresponding to 59 products.

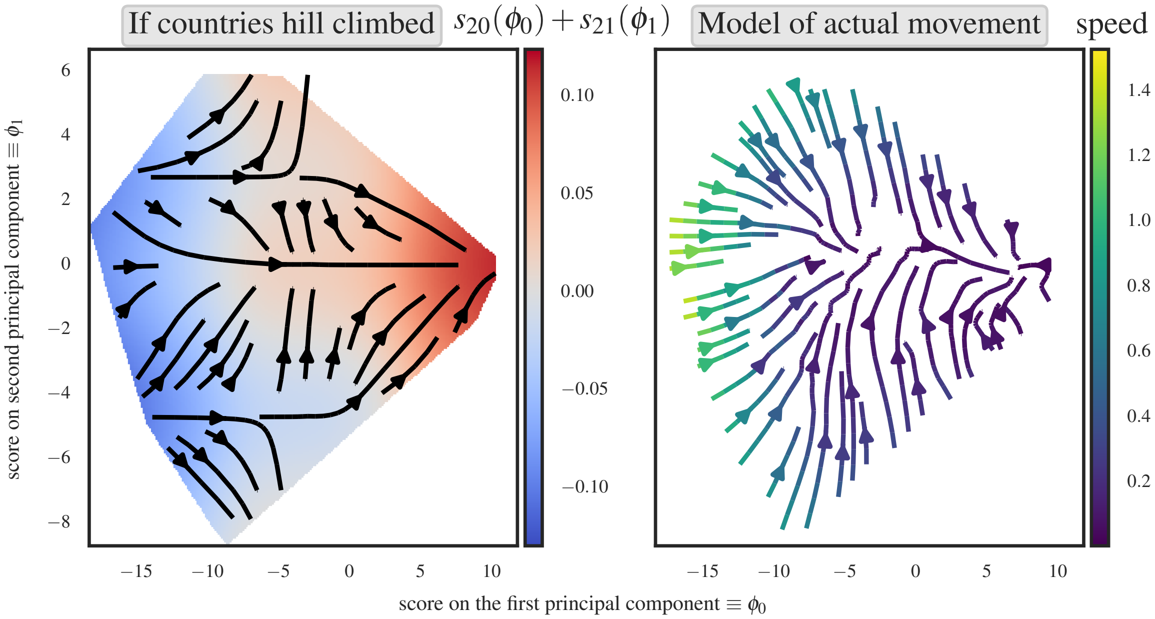

Figure SM-17: **Hill climbing: countries tend to change their export baskets to maximize short-run gains in per-capita incomes.** In the left column, the contribution of (dimension-reduced) export baskets to the changes in per-capita incomes, $s_{20}(\phi_0) + s_{21}(\phi_1)$, is plotted using colors in a blue–red spectrum. The black streamlines show the gradient of that mapping; they mark the direction in which a country would change its ϕ_0 (roughly speaking, its export diversity) and ϕ_1 (roughly speaking, its exports of agriculture minus machinery) to maximize next year's per-capita income, according to the fitted model. The right column shows a smoothed version of how countries actually move through (ϕ_0, ϕ_1), with colors denoting the speed of movement, $\sqrt{(\Delta\phi_0)^2 + (\Delta\phi_1)^2}$. For each rectangle in a fine grid of rectangles covering the diagram, we find the per-capita income $\widetilde{g}$ of the sample with closest (ϕ_0, ϕ_1) to a corner $(\widetilde{\phi_0}, \widetilde{\phi_1})$ of that rectangle, and then we plot the predicted movement in (ϕ_0, ϕ_1) evaluated at $(\widetilde{\phi_0}, \widetilde{\phi_1}, \widetilde{g})$ according to the cubic-spline model (1.2).

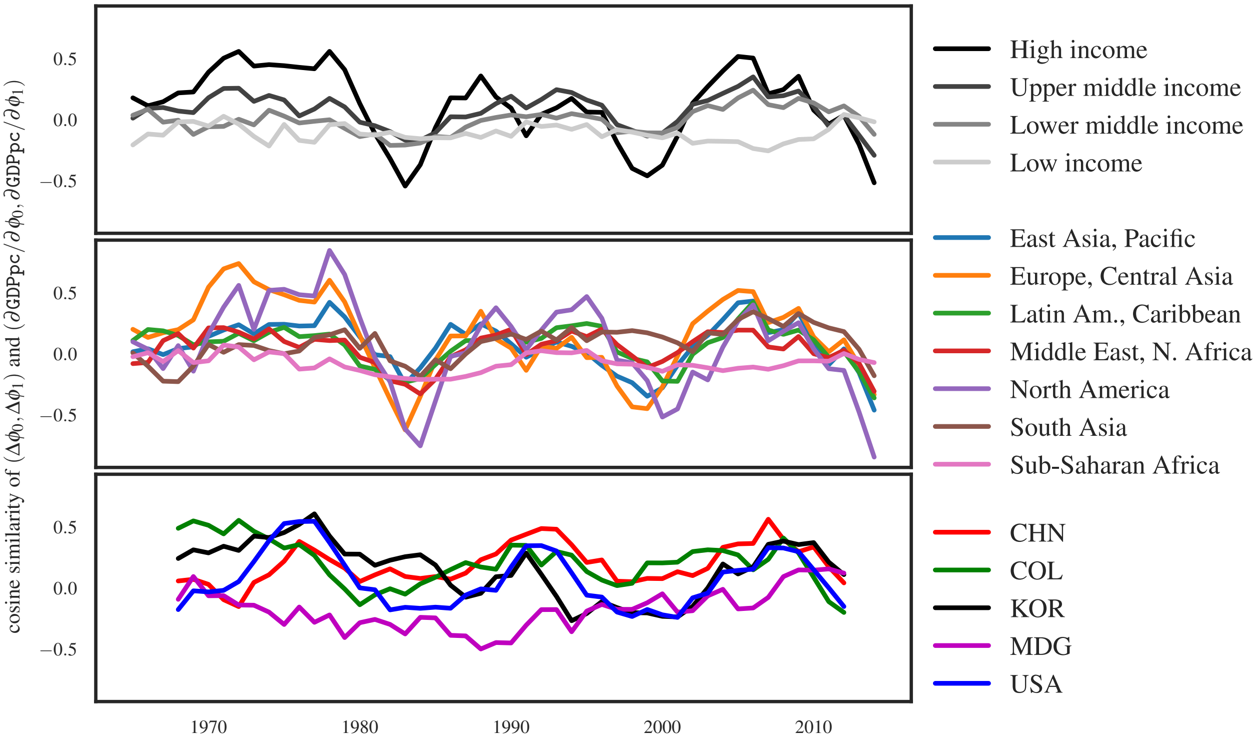

Figure SM-18: **Alignment of economies' changes in export baskets with the direction that would most increase per-capita incomes.** Plotted is the cosine similarity between countries' movement in the first two principal components, $(\Delta\phi_0, \Delta\phi_1)$ and the gradient of the change in per-capita incomes with respect to the scores on the first two principal components, $(s'_{20}(\phi_0), s'_{21}(\phi_1))$. A centered rolling average is applied to reduce noise (with window size 5 in the first two rows and size 10 in the third row). Income groups in the top row are from the World Bank.