INTERFACE

royalsocietypublishing.org/journal/rsif

Machine-learned patterns suggest that diversification drives economic development

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

[1]Global Health and Social Medicine, Harvard Medical School, Boston, MA 02115, USA
[2]Growth Lab at Harvard University, Cambridge, MA, USA
[3]Center for International Development, Harvard Kennedy School, Cambridge, MA 02138, USA
[4]Santa Fe Institute, Santa Fe, NM 87501, USA

CDB, 0000-0003-2553-8862; AG-L, 0000-0001-8320-0857

Research

Cite this article: Brummitt CD, Gómez-Liévano A, Hausmann R, Bonds MH. 2020 Machine-learned patterns suggest that diversification drives economic development. *J. R. Soc. Interface* **17**: 20190283.
http://dx.doi.org/10.1098/rsif.2019.0283

Received: 20 April 2019
Accepted: 25 November 2019

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
biocomplexity

Keywords:
modelling, economic development, statistical learning, non-parametric regression, complex systems, dynamical systems

Author for correspondence:
Charles D. Brummitt
e-mail: brummitt@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4765319.

We combine a sequence of machine-learning techniques, together called Principal Smooth-Dynamics Analysis (PriSDA), to identify patterns in the dynamics of complex systems. Here, we deploy this method on the task of automating the development of new theory of economic growth. Traditionally, economic growth is modelled with a few aggregate quantities derived from simplified theoretical models. PriSDA, by contrast, identifies important quantities. Applied to 55 years of data on countries' exports, PriSDA finds that what most distinguishes countries' export baskets is their diversity, with extra weight assigned to more sophisticated products. The weights are consistent with previous measures of product complexity. The second dimension of variation is proficiency in machinery relative to agriculture. PriSDA then infers the dynamics of these two quantities and of *per capita* income. The inferred model predicts that diversification drives growth in income, that diversified middle-income countries will grow the fastest, and that countries will converge onto intermediate levels of income and specialization. PriSDA is generalizable and may illuminate dynamics of elusive quantities such as diversity and complexity in other natural and social systems.

Computers can learn to predict the future using vast datasets too large for humans to grasp. The logic behind a machine's prediction, however, is often inscrutable [1], and accuracy, not insight, is often the goal. Scientists, meanwhile, generate new theories using intuition or derive them incrementally from existing models. Recent advances in interpretable machine learning are enabling the generation of theories to be automated: machines have identified mathematical laws in physical and biological systems that took many years for scientists to solve manually [2–6]. But while pendulums and fluid dynamics may follow elegant governing equations, systems comprising capricious humans may not. Thus, in social sciences, it is challenging for machines to teach humans better theories.

We introduce a method for identifying interpretable patterns in high-dimensional time series called *Principal Smooth-Dynamics Analysis* (PriSDA), and apply it to a unifying question within the social sciences: why some countries are rich and others are poor. PriSDA ingests time-series data from a system characterized by potentially many dimensions, and it identifies a small number of dynamical equations involving smoothing splines that model how the system changes over time. The resulting model is optimized for accuracy yet is readily interpretable.

Traditionally, one builds an economic theory by manually choosing a dimension-reduced representation of a complex process, such as 'utility' or an aggregated value such as gross domestic product. Then one constructs models of how those few quantities change over time, and finally one estimates

THE ROYAL SOCIETY
PUBLISHING

the derived model with data. These models often neglect important patterns, such as the emergence of the tremendous diversity of goods or why rich countries trade with other rich countries [7]. Agnostic of explicit traditions of economics, what theory of economic growth would an automated economist find?

Here, we analyse machine-reduced dimensions of countries' export baskets (corrected for population size) using principal components analysis (PCA) [8], and then we train generalized additive models (GAMs) [9] to predict yearly changes in export baskets and in *per capita* income. The results indicate the fundamental importance of an economy's (complexity-weighted) diversity, both as a summary statistic—diversity captures more variance of export baskets across time than any other direction—and as a predictor of growth. The method, PriSDA, is generally applicable and may illuminate the emergence of diversity and complexity in other biological, physical and social systems.

1. Methods

The PriSDA method identifies equations that predict changes over time (using non-parametric regression, here GAMs) in a complex system described using dimension reduction and potentially other aggregate quantities. Crucial to its success is providing it data that allows large and small units (here, national economies) to be comparable yet allow for absolute growth.

1.1. Multidimensional characterization of the productive capabilities of economies

There is growing interest in multi-dimensional metrics of economic development and poverty [10,11]. We track macro-level multidimensional economic development based on annual exports, for which there are high-quality data for all countries over the past half-century. A country's exports indicate its international competitive advantages, and unlike domestic production, exports share a common classification system. Indices of the complexity or competitive fitness of economies have been constructed using exports data [12–14], a literature summarized in electronic supplementary material, SM-1 and discussed more below.

Instructions for accessing the data and details on its preprocessing are in electronic supplementary material, SM-2 and SM-3. Because data on exports are noisy, products are aggregated into 59 categories (electronic supplementary material, SM-3.1). The value of a country c's exports of a product p in year t, denoted X_{cpt}, tends to correlate positively with the size of the country's population, P_{ct}. To account for that relationship, X_{cpt} is divided by an expectation according to a null model of a country's expected value of an export given that country's population. To remove the effects of global price shocks, this quantity is divided by the total value of the world's exports of that product, which is also normalized by a null model that predicts global export value using global population. We call the resulting quantity the *absolute advantage* of a country c in a certain product p in year t, denoted $\mathcal{R}_{cpt}$:

$$\mathcal{R}_{cpt} = \frac{X_{cpt}/\mathbb{E}\,[X_{cpt}|P_{ct}]}{\sum_c X_{cpt}/\mathbb{E}\,\left[\sum_c X_{cpt}|\sum_c P_{ct}\right]}. \tag{1.1}$$

Details are in electronic supplementary material, SM-3.2. We consider countries as length-59 vectors of $\mathcal{R}_{cpt}$ across all products; a two-dimensional projection of two trajectories of $\mathcal{R}_{cpt}$ is shown in figure 1*a*. Unlike relative quantities such as revealed comparative advantage [15], a country can grow its absolute advantage arbitrarily. For example, in 2016, Belgium and The Netherlands were the only countries that 'punched above their weight' (i.e. had $\mathcal{R}_{cpt} > 1$) for all 59 products p.

To put products on equal footing with each other, we centre and scale $\mathcal{R}_{cpt}$ by the mean and standard deviation of $\mathcal{R}_{cpt}$ across all countries and across all years $t \leq 1988$ (figure 1*b*). We call the resulting quantity *scaled absolute advantage* and denote it by R_{cpt}. Scaled absolute advantage is the number of standard deviations above the pre-1988 mean of all countries' absolute advantage in that product. $R_{cpt} > 0$ means that country c excels at producing and exporting product p in year t. Making products comparable—by dividing by the product's global export market in $\mathcal{R}_{cpt}$ and by centring and scaling each product in R_{cpt}—enables detecting how expertise in one product enables developing expertise in another, regardless of the sizes of the markets of those products.

1.2. Reducing dimensions of export baskets

We reduce dimensions using PCA [8] because the resulting dimensions are interpretable and because summing exports reduces the noise in export data. Other methods are discussed in electronic supplementary material, SM-1.

Figure 2 shows the *loadings* (weights) of the first three principal components on the 59 products. The *score* of a country's export basket on the kth principal component—denoted ϕ_k—is the dot product (illustrated in figure 1*c*) of the country's export basket, $(R_{cpt})_{p\in\mathcal{P}}$, with that principal component's loading vector, drawn as a row of rectangles in figure 2*a*. We interpret this PCA in the Results section below.

1.3. Inferring dynamics of export baskets

To understand patterns in economic development, next we examine how two summary measures of an export basket, ϕ_0 and ϕ_1, interact with *per capita* income [16] (transformed logarithmically): $\texttt{GDPpc} \equiv \log_{10}(\text{GDP } \textit{per capita})$. Because excelling at exports reflects the capabilities and know-how within a country [12,17,18], by inferring a model of how the triple (ϕ_0, ϕ_1, $\texttt{GDPpc}$) changes over time, we aim to shed light on fundamental economic patterns.

The three variables (ϕ_0, ϕ_1, $\texttt{GDPpc}$) are aggregate descriptions of an economy, so we expect them to change smoothly over time. A natural choice for a smooth model are cubic smoothing splines [9, ch. 3 and 4]. We chose this class of GAMs because they are interpretable yet predict as well as other general-purpose learning methods like gradient boosting (electronic supplementary material, table SM-4). This method provides us with the following system of dynamical equations:

$$g(\Delta\phi_0(t)) = c_0 + s_{00}(\phi_0(t)) + s_{01}(\phi_1(t)) + s_{02}(\texttt{GDPpc}(t)) \tag{1.2a}$$

$$g(\Delta\phi_1(t)) = c_1 + s_{10}(\phi_0(t)) + s_{11}(\phi_1(t)) + s_{12}(\texttt{GDPpc}(t)) \tag{1.2b}$$

$$\text{and} \quad g(\Delta\texttt{GDPpc}(t)) = c_2 + s_{20}(\phi_0(t)) + s_{21}(\phi_1(t)) + s_{22}(\texttt{GDPpc}(t)), \tag{1.2c}$$

where Δ takes the expected difference in time, $\Delta f(t) \equiv \mathbb{E}[f(t+1) - f(t)]$; the c_i are intercept terms; the $s_{i,j}$ are cubic smoothing splines with smoothing strength chosen using nested-in-time cross validation (electronic supplementary material, SM-5.1.1); and the link function $g(x) \equiv \text{sign}(x)\,|x|^{1/2}$ is applied to make the residuals' distributions closer to a normal (electronic supplementary material, SM-5.2). The goodness of fit ($R^2 \approx 0.04$) and the GAM's competitiveness with other statistical learning methods are discussed in electronic supplementary material, SM-5.1.2.

The terms of (1.2) are plotted in figure 3, where one can compare not only the shapes but also the magnitudes. The GAM (1.2) can be understood as a dynamical model inferred from the data, with which we can attempt to predict the future; it can also be

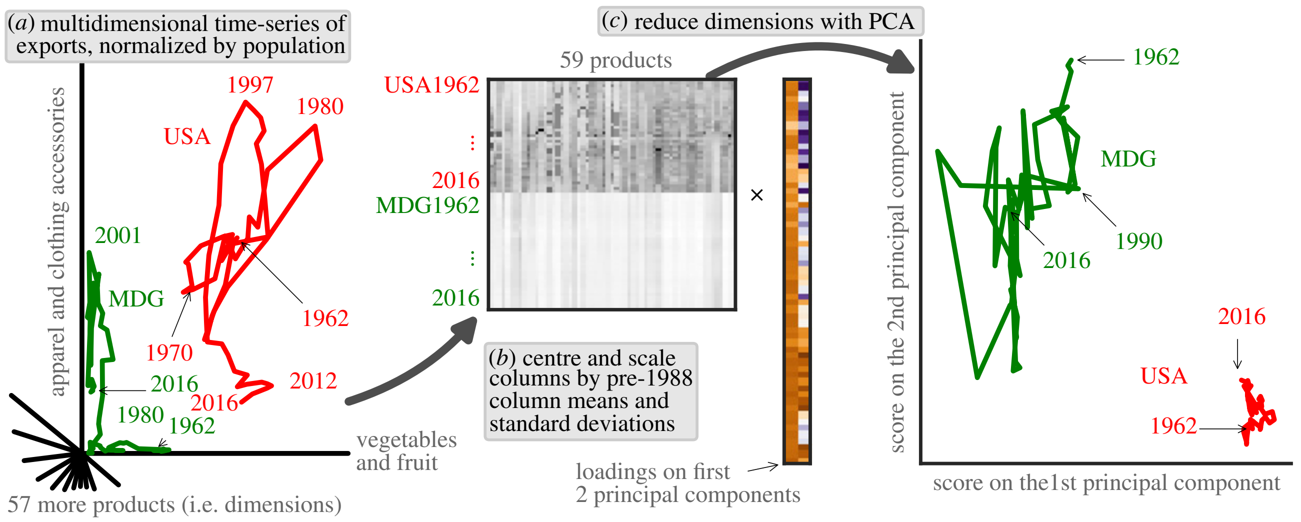

Figure 1. *Preprocessing and reducing dimensions of export baskets.* (*a*) Begin with time series of export values of 59 product categories, normalized by population (electronic supplementary material, Eq. (SM-1)) and logarithmically transformed (electronic supplementary material, Eq. (SM-5)) to make large and small countries comparable. For illustration, we show the trajectories of the United States (USA) and Madagascar (MDG). We are illustrating two dimensions here, but in reality the red and green curves live in 59 dimensions. (*b*) Centre and scale columns by their pre-1988 means and standard deviations. (*c*) Reduce dimensions with principal components analysis (PCA). Each country's score in a given principal component represents a certain linear combination of its export basket. Together, the scores on the first few principal components summarize the country's export basket with just a few numbers. (Online version in colour.)

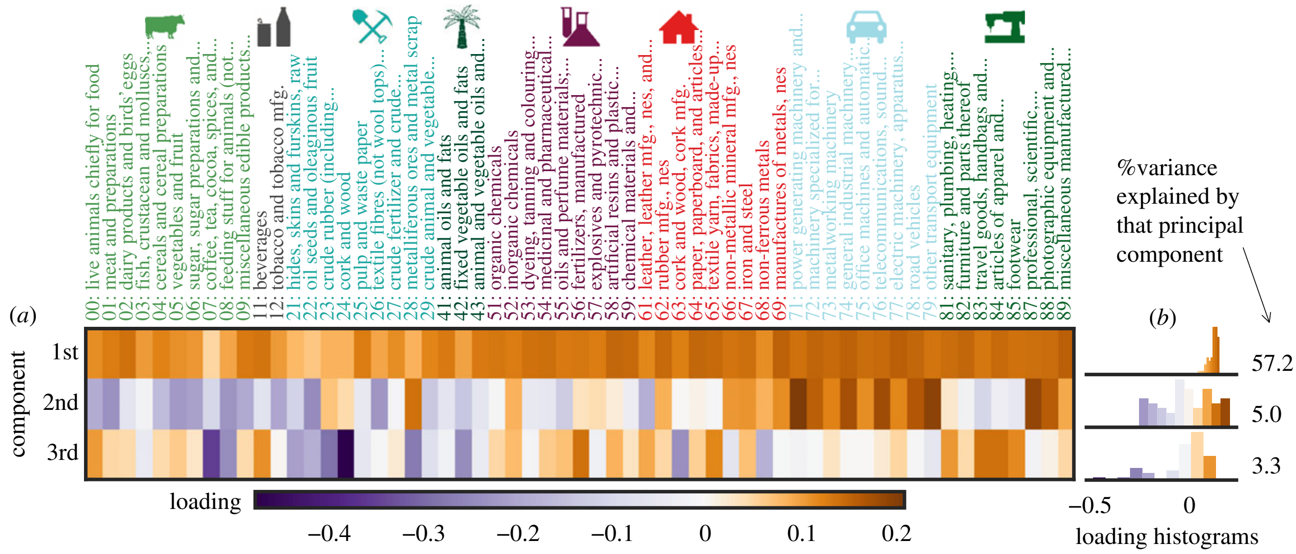

Figure 2. *The first three principal components are approximately (1) total absolute advantage summed across all products (with more weight on product codes above 50), (2) machinery minus agriculture, and (3) textiles and fertilizer minus coffee and cork.* In plot (*a*), the rows are principal components, the columns are products, and the rectangles' colours represent the loading (or 'weight') of that principal component on that product. The first component loads positively on all products. Thus, what distinguishes countries, above all, is their 'diversification' across products. The second component loads highly on machinery (product codes beginning with 7) and other manufactured goods, and it loads negatively on agricultural products. Thus, the direction in 59-dimensional product space orthogonal to the first component that most spreads out observations points towards machinery and away from agriculture. The third component loads positively on clothing and textile products and negatively on cork and wood (24) and coffee, tea and spices (07). The plots labelled (*b*) are histograms of loadings, across all 59 products, in the corresponding rows. (Online version in colour.)

understood as a histogram smoother that helps us see signal amid the noise.

2. Results

2.1. How a machine summarizes an economy

The first principal component is the direction in product space along which countries are most spread out in terms of variance [8]. We find this direction to be associated with total *per capita* exports and the diversification across products. This first component explains more than half of the variation in export baskets (57.2%) across countries and years.

Mathematically, a country's score ϕ_0 on the first principal component is a weighted sum of R_{cpt} (scaled absolute advantage) across all products. The weights are fixed once PCA has been fitted, but since export baskets change in time, scores ϕ_0 change from year to year (and from country to country). The weights of this principal component are all positive and are depicted in the top row of figure 2*a*. The score ϕ_0 is highly correlated with *per capita* exports summed across products, $\sum_p X_{cpt}/P_{ct}$ (Pearson

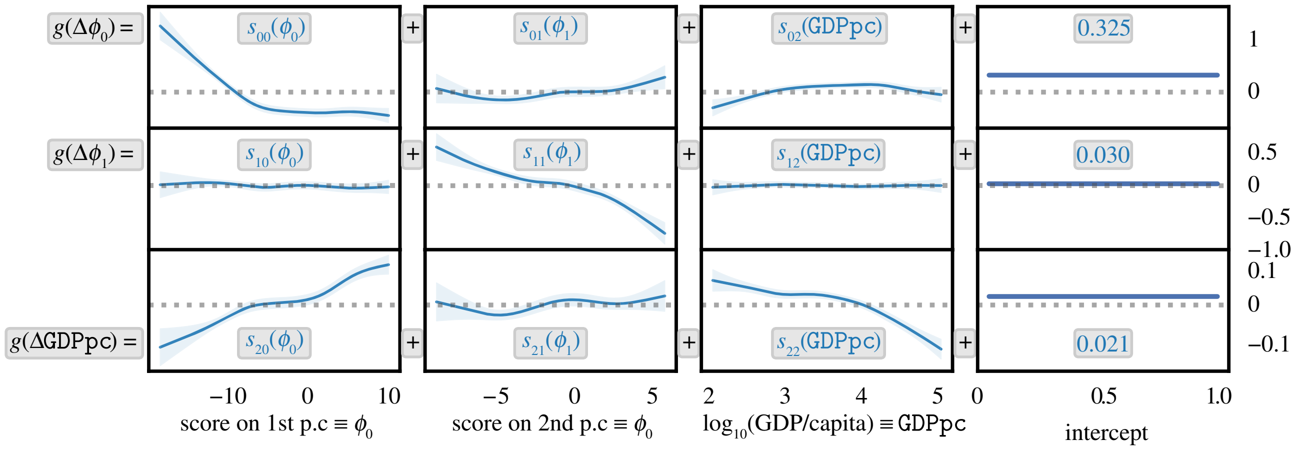

Figure 3. *Export baskets tend to diversify and converge to a balance of agriculture and manufactured goods.* Shown are *partial dependence plots* of the three equations in (1.2). Each blue (solid) curve is an additive contribution to the quantity written in black on the left-hand side of this figure, which is a link function g applied to the expected yearly change in one of the three variables ϕ_0, ϕ_1, GDPpc. (See the text after (1.2) for the definition of g.) In each plot, the quantity being plotted is written in blue within the plot. Adding the blue expressions across a row gives the right-hand sides of (1.2). The shaded regions show the 95% CI. Each equation has an intercept, c_i, shown in the right column. The plots on the diagonal have negative trends, suggesting convergence. Interestingly, income is not associated with changes in export baskets, but ϕ_0 appears to drive GDPpc: diversifying precedes income growth. (Replacing ϕ_0 with diversification preserves the positive trend in the bottom-left plot, but replacing it with *per capita* exports greatly weakens that relationship; see electronic supplementary material, Figs. SM-13 and SM-14.) (Online version in colour.)

$\rho = 0.82$; electronic supplementary material, Fig. SM-6). This correlation, however, is trivial and unsurprising given that ϕ_0 is a positively weighted sum across products. However, the finding that the weights are all positive, together with the fact that this principal component explains almost 60% of the variation, is not trivial and is of economic significance.

More surprising is the association between the score on the first component, ϕ_0, and diversity, the number of distinct product categories exported in large numbers (see electronic supplementary material, SM-4.2.2). Diversity is more correlated with ϕ_0 than *per capita* exports are (Pearson $\rho = 0.67$ versus $\rho = 0.46$). The values of ϕ_0 remain highly correlated with *per capita* exports and diversity even after controlling for other covariates (namely, Worldwide Governance Indicators and measures of educational attainment; see electronic supplementary material, Fig. SM-6–SM-10). Thus, in the dimension defined by the first principal component, countries are separated by not just how much they export in total but how diversified those exports are.

The score ϕ_0 on the first principal component also emphasizes the complexity of products: notice in the top row of figure 2*a* that the loadings are not uniform: they are about twice as large on the more complex products. These variations in the loadings, in fact, are highly correlated with the Product Complexity Index [12] (Pearson $\rho = 0.81$; electronic supplementary material, Fig. SM-5). Thus, the score ϕ_0 captures the diversification [12] of an economy, with an emphasis on more complex goods. Relationships between ϕ_0 and other quantities are characterized in electronic supplementary material, SM-4.

The next direction that most spreads out export baskets across countries and across time—conditional on being orthogonal to the first component—loads highly on machinery and negatively on agricultural products (figure 2*a*, middle row). Thus, after knowing a country's complexity-weighted diversity ϕ_0, the next characteristic that most spreads out countries is how much more they export in machinery relative to agricultural goods. This second principal component explains 5% of the variance, 11 times less than that explained by ϕ_0.

Garments have long been considered to be the first sector to industrialize in a country, including in England during the industrial revolution and in many East Asian countries since the 1960s [19]. However, these products are not an important direction of variation of export baskets between 1962 and 1988 in the first and second principal components. Only in the third principal component are textile products substantially loaded (figure 2*a*, bottom row). Because this component only explains 3.3% of the variance of export baskets, hereafter, we focus on the first two principal components.

2.2. Complexity-weighted diversity predicts growth

The partial dependence plots in figure 3 show how yearly changes in ϕ_0, ϕ_1, and GDPpc are predicted by sums of one-dimensional functions of those same variables. The rows of figure 3 depict the three equations (1.2*a*)–(1.2*c*). Note that *per capita* income is not a strong predictor of changes in export baskets as measured by ϕ_0 and ϕ_1 (see the top and middle plots in the third column of figure 3). By contrast, the score ϕ_0 on the first principal component is associated with significant growth in income, even though ϕ_0 and ϕ_1 were defined independently of income. The fact that ϕ_0, the complexity-weighted diversity of an economy, seems to drive income, and not the reverse, is consistent with the hypothesis that income is the outcome: income emerges from the productive capabilities of an economy, captured here by ϕ_0 and ϕ_1.

If the 'off-diagonal' terms $s_{01}(\phi_1) + s_{02}(\texttt{GDPpc})$ in (1.2*a*) were absent, then ϕ_0 would settle onto a value near -10. This amount is approximately the value of ϕ_0 of the poorest countries in 2016, such as Liberia, Angola and the Democratic Republic of the Congo. However, the large intercept in (1.2*a*) (the top-right plot in figure 3) suggests a general positive tendency to diversify, regardless of the country's absolute advantage in machinery relative to agriculture (ϕ_1).

Exporting more agricultural than machinery products (i.e. $\phi_1 < 0$) is typically associated with a slight drag on ϕ_0 (second column, top plot in figure 3); once machinery and agriculture balance ($\phi_1 \approx 0$), that drag on ϕ_0 vanishes.

Once a country has complexity-weighted diversification $\phi_0 > 0$, it can expect significant growth in income (first column, left plot of figure 3). Interestingly, simply exporting more *per capita*, regardless of the allocation across products, is not associated with growth: when ϕ_0 is substituted with total *per capita* exports, the relationship with income growth flattens (electronic supplementary material, Fig. SM-13). Exporting more kinds of goods matters: replacing ϕ_0 with another notion of diversity [12] preserves the positive relationship with income growth (electronic supplementary material, Fig. SM-14).

We note, in addition, that the lack of clear association between s_{21} (ϕ_1) and growth of GDPpc implies that there are weak returns to specialization. In fact, ϕ_1 tends towards zero, regardless of the other two variables, meaning that countries tend towards a diversified export basket that balances agriculture with machinery. These results suggest that export baskets tend to increasingly resemble one another. Next, we examine this convergence in more detail.

2.3. Two-dimensional projections of the data and of the learned dynamics

In figure 4, we compare the model (1.2) with the empirical data. This figure projects the data onto (ϕ_0, ϕ_1) (top row) and onto (ϕ_0, GDPpc) (bottom row). The left-hand column shows empirical data, with some countries' trajectories highlighted. The right-hand column visualizes the vector field of (1.2) as a 'stream plot', with the third variable not plotted taken to be the pre-1988 mean. That is, the arrows are the expected movement for countries whose third variable (the one not plotted) equals the pre-1988 mean; for other countries, the arrows approximate their expected movement.

The data in figure 4*a* show that countries tend to move from left to right (they export more and diversify) and towards the middle of the vertical axis (they move towards $\phi_1 \approx 0$, a balance between agriculture and machinery). The grey streamlines of the inferred model in figure 4*b* confirm this pattern, suggesting that countries converge towards the trajectory like that of Thailand's (purple, labelled THA). In the bottom row, the data in figure 4*c* show that development success stories like South Korea (KOR), Thailand and China (CHN) share a common trajectory of increasing ϕ_0 and income. The inferred model's streamlines in figure 4*d* suggest that poor countries will follow in their footsteps, but also that income in the richest countries may fall. The 'J' shape in figure 4*c*,*d* suggests that growth only takes over after diversification reaches a critical value.

It is interesting to compare the bottom row of figure 4 with the analogous plots in figs 1 and 2 of Cristelli *et al.* [20], who constructed a 'fitness' index of a country's intangible capabilities using an iterative calculation involving export baskets. (A summary of this index and of others like it is in electronic supplementary material, SM-1.) Here, ϕ_0 plays a role analogous to the fitness index. Our method automatically recovered the behaviour found in [20] that countries tend to diversify before their *per capita* income grows. Whereas the coarse-grained dynamics of Cristelli *et al.* [20] indicate that chaos and turbulence characterize poorer countries, the splines (1.2) chose more smoothness in that region.

2.4. Stream plots of export baskets at different levels of income

In figure 5, we vary GDPpc across three values, the 10th, 50th and 90th percentiles of *per capita* income in year 1988. As a country's *per capita* income rises, the map of how its export basket moves through the space of products (as described by ϕ_0, ϕ_1) morphs from the plot on the left to the plot on the right. The colours denote the model's predicted change in *per capita* income (equation (1.2*c*)). In the plot on the left, we see that the poorest countries tend towards a fixed point: what little they export ($\phi_0 \approx -8$) tends towards a balance between agriculture and machinery (ϕ_1 tends to zero). Countries with *per capita* income near the median ($2764 per year) tend to grow their complexity-weighted diversification ϕ_0 (note the trend to the right in the middle plot of figure 5), a pattern that continues for the richest countries (right-hand plot of figure 5). It appears that one need not be very rich to begin to diversify.

This movement in product space (ϕ_0, ϕ_1) appears to maximize expected short-run increases in income, according to (1.2) (electronic supplementary material, Fig. SM-18). High-income countries tend to be best at moving towards higher income (except for brief periods), and China has been exceptional at it since 1990.

2.5. Long-run predictions of *per capita* income

Research on economic complexity has focused on growth predictions as validation [12,20] and, recently, research [21] has benchmarked these predictions against those of the International Monetary Fund. We found that predicting the change of export baskets simultaneously with the change of *per capita* income was inherently a hard problem across different statistical learning methods (electronic supplementary material, SM-5.1.2). Instead, we found our low-dimensional model to be better suited for generating interpretable, qualitative insights rather than making competitive predictions. With this caveat in mind, we investigate the model's long-run predictions by iterating 1-year predictions starting from 2016 data.

Figure 6 shows the model's long-run predictions of growth in *per capita* income as a function of (A) *per capita* income and (B) ϕ_0 in 2016. The model predicts that the diverse, middle-income countries today will significantly catch up to the richest ones, growing at an annual rate of 2%. Meanwhile, it predicts that poor countries (such as Liberia) and middle-income countries with low diversity ϕ_0 (such as Angola) are predicted to grow between 0 and 1% annually. Rich countries like Norway are predicted to barely grow at all. The next economic success stories, according to this model, are those with intermediate income and diversification today. These results are consistent with one of the 'New Kaldor facts' [22] that rich countries grow more slowly than middle-income countries.

3. Discussion

This investigation sits at the intersection of three recent developments in the quantitative social, natural, and physical

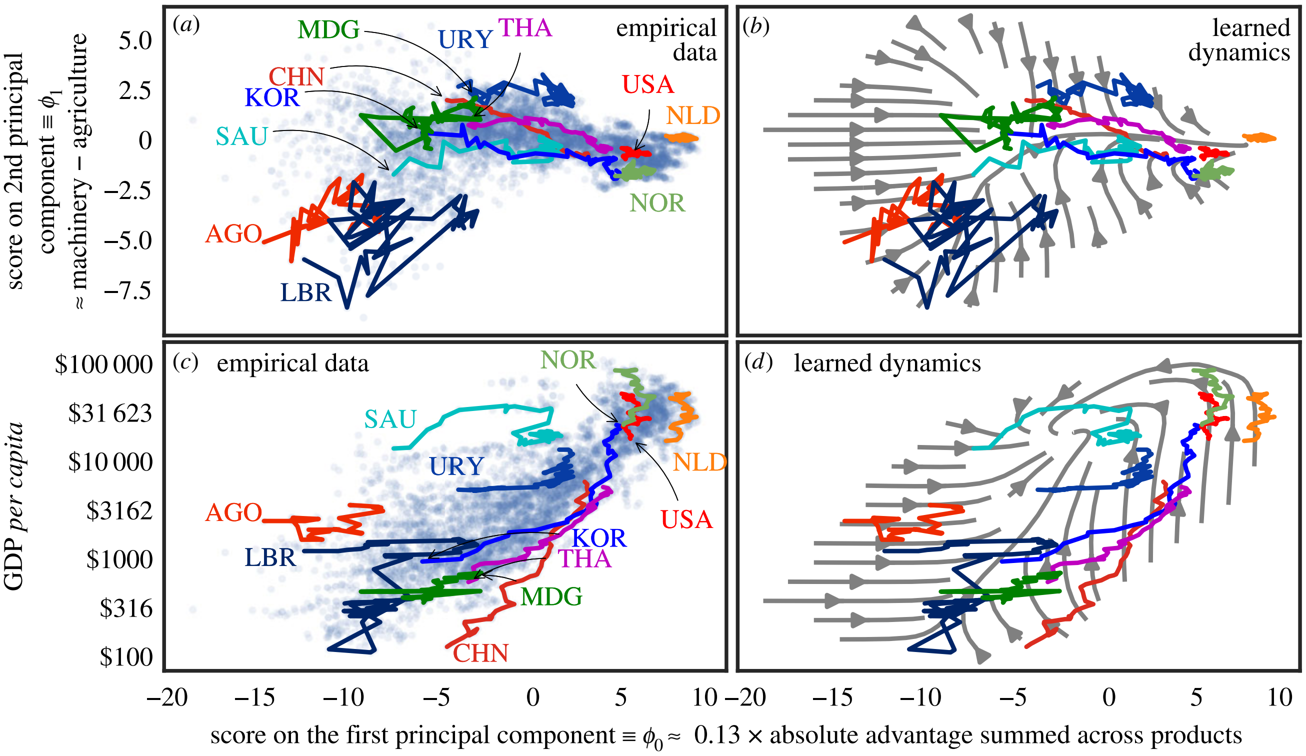

Figure 4. *The learned dynamics (1.2) predict that countries are converging.* The left column shows empirical data with blue dots; the right column shows predictions of the model (1.2) as stream plots. The empirical trajectories of eight countries over years 1962–2016 are superimposed on all four plots. Trajectories are labelled at the first available sample (year 1985 for Angola, 1962 for the rest). A country is represented by a triple (ϕ_0, ϕ_1, GDPpc), and the model (1.2) has been trained on this three-dimensional space, but here we show projections onto (ϕ_0, ϕ_1) in the top row and onto (ϕ_0, GDPpc) in the bottom row. (*a*) Countries tend to diversify (increase ϕ_0) and strike a balance between machinery and agriculture ($\phi_1 \approx 0$). (*c*) Development success stories (e.g. THA, KOR, CHN) share a common trajectory of increasing ϕ_0 and income. (*d*) Poor countries may follow in their footsteps, but income in the richest countries may stagnate or even fall. Countries are labelled with United Nations ISO-alpha3 codes. (Online version in colour.)

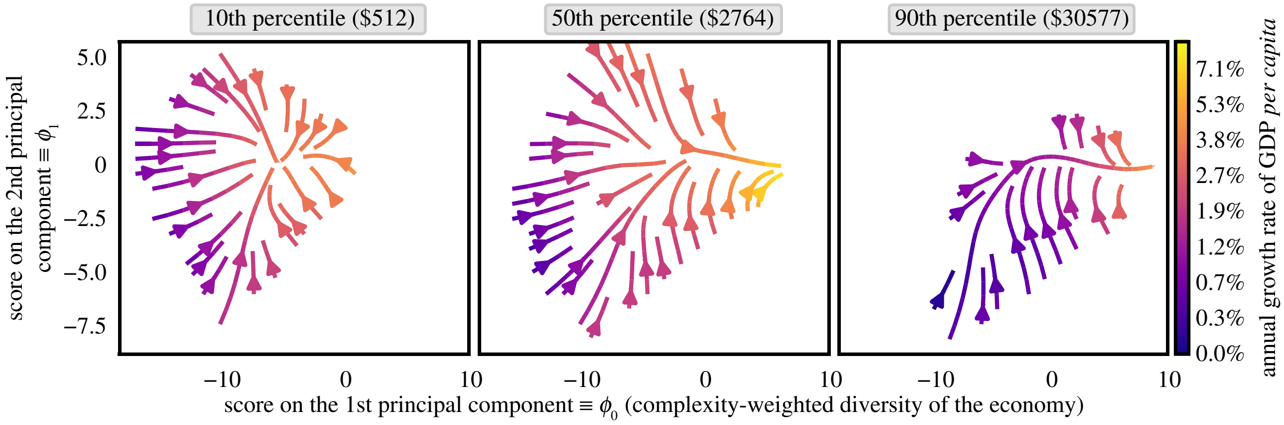

Figure 5. *Inferred dynamics of export baskets, at three levels of* per capita *income, predict convergence in the long run.* The streamlines show how a country's export basket, described by its scores (ϕ_0, ϕ_1) on the first two principal components, changes over time according to the GAM (1.2). From left to right, the plots correspond to GDP *per capita* at the 10th, 50th and 90th percentiles of *per capita* income among countries in the year 1988. Those percentiles are the value inserted into (1.2); we show streamlines at (ϕ_0, ϕ_1) pairs in the convex hull of all empirical samples (ϕ_0, ϕ_1, GDPpc) with GDPpc within 15% of the value shown at the top of the plot. The predicted yearly change in *per capita* income is plotted in colour. The model predicts that poor countries move towards a balance of agriculture and machinery before increasing their total exports. (Said formally, $\phi_1 \to 0$ in the left plot, and ϕ_0 increases substantially in the middle and right plots.) Eventually, all countries are predicted to become rich and to have diverse export baskets (high ϕ_0) that balance between agriculture and machinery ($\phi_1 \approx 0$). (Online version in colour.)

sciences: (1) the roles of complexity and diversity as drivers of economic growth [12,23]; (2) identifying universal, low-dimensional patterns of complex human systems over time [24]; and (3) using machine learning to uncover governing laws of biological and physical systems [4–6].

We accordingly developed a new method, PriSDA, by applying tools from statistical learning—namely, dimension reduction and GAMs—to identify stylized patterns in economic development. Our measure of countries' proficiencies in exporting 59 product categories allows for small and

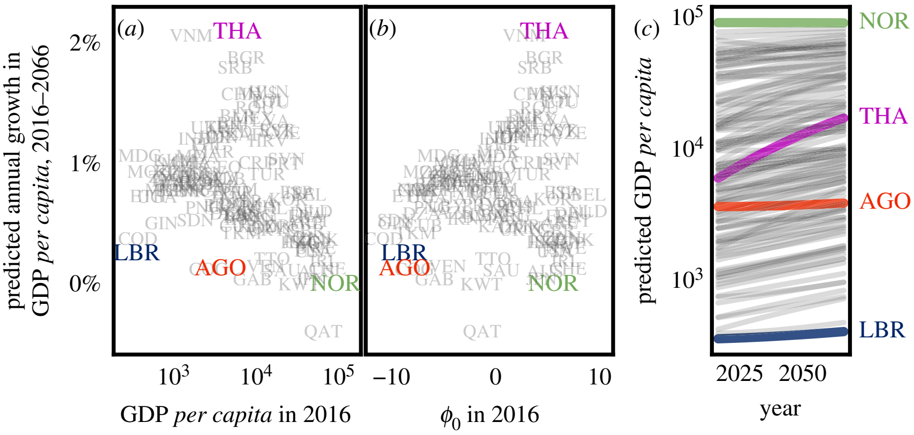

Figure 6. *Catch-up of the diverse, middle-income countries.* Shown are predicted annual growth rates of *per capita* income (in constant 2010 USD per person per year) over the next 50 years as a function of (*a*) current *per capita* income and (*b*) current score ϕ_0 on the first principal component. (*c*) Predicted trajectories of *per capita* income. Highlighted are four countries representative of four groups: low-income countries predicted to grow little (Liberia, LBR); middle-income countries with high diversity (high ϕ_0) today predicted to grow a lot (Thailand, THA); middle-income countries with low diversity (low ϕ_0) predicted to grow little (Angola, AGO); and high-income countries predicted to grow little (Norway, NOR). The GAM (1.2) predicts the highest growth in income for economies that currently have intermediate income (annual growth ≈ 1.5% to 2% for countries with yearly *per capita* income between \$1000 and \$20 000) and lower growth rates for poorest countries (0 to 1% growth) and the richest countries (0 to 0.5% growth). (Online version in colour.)

large countries to be comparable, adjusts for global shocks, and can account for absolute economic growth. Given these data, PriSDA found a complexity-weighted measure of diversity, and it approximately recovered the product complexity index [12].

Our analysis generated two core insights. First, diversity appears to drive *per capita* income; merely exporting a lot *per capita* has a weaker association with income growth. Second, countries are not predicted to split into rich and poor clubs, nor into manufacturing hubs and agricultural hubs, but instead to converge on the same increasingly diverse basket of goods (and capabilities). We hope that future research reconciles these patterns of diversification with the specialization predicted by Ricardian theories of comparative advantage [25,26]. The most dynamic economies of the twenty-first century are predicted to be middle-income economies that are somewhat diversified across products. The least diversified countries have dismal prospects for economic growth, consistent with previous findings [23].

The importance of this approach rests on its applicability beyond the specific case studied here. In general, systems whose evolution is described by a multiplicity of properties are amenable to analysis such as the one we propose here. For example, it is known that wildfires typically reduce the number of species that inhabit an ecosystem, but then species recolonize over time as diversity rises in a process called ecological succession [27]. The composition of species in a system can also converge due to migration [28]. PriSDA could reveal other patterns, still unknown, in such ecological systems. These commonalities suggest the possibility of general theories of complex systems, uncovered by machines less tied to disciplinary paradigms.

Rapidly advancing ways for machines to learn interpretable models bode well for followup studies. Because the model (1.2) is additive, it does not capture interactions, so the effect of ϕ_1 on ϕ_0, for example, is assumed to be independent of ϕ_0. Future work could relax that assumption, using methods like GA^{2}M [29]. High-dimensional data could be fitted with GAMs that both smooth the data and select terms, such as GAMSEL [30] or SPLAM [31]. PriSDA takes a step towards this broader goal of using machines to generate fundamental theories of complex natural and social systems.

Data accessibility. The data, available at Dataverse, were analysed using open-source software, including NumPy, pandas, SciPy and pyGAM [32]. Code to reproduce this work is available at GitHub. Details on the standardization and preprocessing of this dataset are in electronic supplementary material, SM-2 and SM-3.

Authors' contributions. C.D.B. and A.G.L. performed research. C.D.B. wrote programming and statistical code. All authors designed research, wrote the paper and gave final approval for publication.

Competing interests. The authors declare that they have no competing interests.

Funding. C.D.B. and M.H.B. acknowledge funding from the James S. McDonnell Foundation for the Postdoctoral Award and the Scholar Award (respectively) in Complex Systems.

Acknowledgements. We thank L. Pietronero for highlighting the relationship between fitness and complexity-weighted diversity.

References

1. Voosen P. 2017 How AI detectives are cracking open the black box of deep learning. *Science*. (doi:10.1126/science.aan7059)
2. Bongard J, Lipson H. 2007 Automated reverse engineering of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **104**, 9943–9948. (doi:10.1073/pnas.0609476104)
3. Schmidt M, Lipson H. 2009 Distilling free-form natural laws from experimental data. *Science* **324**, 81–85. (doi:10.1126/science.1165893)
4. Daniels BC, Nemenman I. 2015 Automated adaptive inference of phenomenological dynamical models. *Nat. Commun.* **6**, 8133. (doi:10.1038/ncomms9133)
5. Zhang L, Li K. 2015 Forward and backward least angle regression for nonlinear system identification.

Automatica **53**, 94–102. (doi:10.1016/j.automatica.2014.12.010)

6. Brunton SL, Proctor JL, Kutz JN. 2016 Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl Acad. Sci. USA* **113**, 3932–3937. (doi:10.1073/pnas.1517384113)
7. Krugman PR 1993 *Geography and trade*. Gaston Eyskens Lecture Series. Cambridge, MA: MIT Press.
8. Lever J, Krzywinski M, Altman N. 2017 Principal component analysis. *Nat. Methods* **14**, 641–642. (doi:10.1038/nmeth.4346)
9. Wood S 2006 *Generalized additive models: an introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, FL: Taylor & Francis.
10. Alkire S, Foster J. 2011 Counting and multidimensional poverty measurement. *J. Public Econ.* **95**, 476–487. (doi:10.1016/j.jpubeco.2010.11.006)
11. Hruschka DJ, Hadley C, Hackman J. 2017 Material wealth in 3D: mapping multiple paths to prosperity in low- and middle-income countries. *PLoS ONE* **12**, e0184616. (doi:10.1371/journal.pone.0184616)
12. Hidalgo CA, Hausmann R. 2009 The building blocks of economic complexity. *Proc. Natl Acad. Sci. USA* **106**, 10 570–10 575. (doi:10.1073/pnas.0900943106)
13. Tacchella A, Cristelli M, Caldarelli G, Gabrielli A, Pietronero L. 2012 A new metrics for countries' fitness and products' complexity. *Sci. Rep.* **2**, 482–487. (doi:10.1038/srep00723)
14. Teza G, Caraglio M, Stella AL. 2018 Growth dynamics and complexity of national economies in the global trade network. *Sci. Rep.* **8**, 15230. (doi:10.1038/s41598-018-33659-6)
15. Balassa B. 1965 Trade liberalisation and 'revealed' comparative advantage. *Manchester Sch.* **33**, 99–123. (doi:10.1111/j.1467-9957.1965.tb00050.x)
16. World Bank. 2016 GDP per capita, indicator `NY.GDP.PCAP.KD`, expressed in constant 2010 USD.
17. Hausmann R, Hwang J, Rodrik D. 2006 What you export matters. *J. Econ. Growth* **12**, 1–25. (doi:10.1007/s10887-006-9009-4)
18. Hausmann R, Hidalgo CA. 2011 The network structure of economic output. *J. Econ. Growth* **16**, 309–342. (doi:10.1007/s10887-011-9071-4)
19. Birdsall NM, Campos JEL, Kim CS, Corden WM, MacDonald L, Pack H, Page J, Sabor R, Stiglitz JE. 1993 The East Asian miracle: economic growth and public policy. Technical Report 12351. The World Bank.
20. Cristelli M, Tacchella A, Pietronero L. 2015 The heterogeneous dynamics of economic complexity. *PLoS ONE* **10**, e0117174. (doi:10.1371/journal.pone.0117174)
21. Tacchella A, Mazzilli D, Pietronero L. 2018 A dynamical systems approach to gross domestic product forecasting. *Nat. Phys.* **14**, 861–865. (doi:10.1038/s41567-018-0204-y)
22. Jones CI, Romer PM. 2010 The New Kaldor facts: ideas, institutions, population, and human capital. *Am. Econ. J.: Macroecon.* **2**, 224–245. (doi:10.1257/mac.2.1.224)
23. Hidalgo CA, Klinger B, Barabási AL, Hausmann R. 2007 The product space conditions the development of nations. *Science* **317**, 482–487. (doi:10.1126/science.1144581)
24. Turchin P *et al.* 2018 Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. *Proc. Natl Acad. Sci. USA* **115**, E144–E151. (doi:10.1073/pnas.1708800115)
25. Arkolakis C, Costinot A, Rodriguez-Clare A. 2012 New trade models, same old gains? *Am. Econ. Rev.* **102**, 94–130. (doi:10.1257/aer.102.1.94)
26. Eaton J, Kortum S. 2012 Putting Ricardo to work. *J. Econ. Perspect.* **26**, 65–90. (doi:10.1257/jep.26.2.65)
27. Rosenzweig M. 1995 *Species diversity in space and time*. Cambridge, UK: Cambridge University Press.
28. Pickett S, White P. 2013 *The ecology of natural disturbance and patch dynamics*. Orlando, FL: Academic Press.
29. Lou Y, Caruana R, Gehrke J, Hooker G. 2013 Accurate intelligible models with pairwise interactions. In *Proc. 19th SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013*, pp. 623–631. New York, NY: ACM. (doi:10.1145/2487575.2487579)
30. Chouldechova A, Hastie T. 2015 Generalized additive model selection. (https://arxiv.org/abs/1506.03850)
31. Lou Y, Bien J, Caruana R, Gehrke J. 2016 Sparse partially linear additive models. *J. Comput. Graph. Stat.* **25**, 1126–1140. (doi:10.1080/10618600.2015.1089775)
32. Servén D, Brummitt C. 2018 pyGAM: generalized additive models in Python (version 0.2.17). (doi:10.5281/zenodo.1226652)