Matrix Completion of World Trade

Giorgio Gnecco (giorgio.gnecco@imtlucca.it)
IMT School for Advanced Studies, Lucca
Federico Nutarelli
IMT School for Advanced Studies, Lucca
Massimo Riccaboni
IMT School for Advanced Studies, Lucca

Research Article

Keywords: Economic complexity, revealed comparative advantage, matrix completion, nuclear norm regularization, GENEPY

Posted Date: November 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1030693/v1

Matrix Completion of World Trade

Giorgio Gnecco[1,*], Federico Nutarelli[1], and Massimo Riccaboni[1]

[1]IMT School for Advanced Studies, AXES Research Unit, Lucca, 55100, Italy
[*]giorgio.gnecco@imtlucca.it

ABSTRACT

This work applies Matrix Completion (MC) – a class of machine-learning methods commonly used in the context of recommendation systems – to analyze economic complexity. MC is applied to reconstruct the Revealed Comparative Advantage (RCA) matrix, whose elements express the relative advantage of countries in given classes of products, as evidenced by yearly trade flows. A high-accuracy binary classifier is derived from the MC application, with the aim of discriminating between elements of the RCA matrix that are, respectively, higher/lower than one. We introduce a novel Matrix cOmpletion iNdex of Economic complexitY (MONEY) based on MC, and related to the degree of predictability of the RCA entries of different countries (the lower the predictability, the higher the complexity). Differently from previously-developed economic complexity indices, MONEY takes into account several singular vectors of the matrix reconstructed by MC, whereas other indices are based only on one/two eigenvectors of a suitable symmetric matrix, derived from the RCA matrix. Finally, MC is compared with a state-of-the-art economic complexity index (GENEPY), showing that the false positive rate per country of a binary classifier constructed starting from the average entry-wise output of MC is a proxy of GENEPY.

Keywords: Economic complexity, revealed comparative advantage, matrix completion, nuclear norm regularization, GENEPY

Introduction

Since the early 2000s, building Ymetrics for measuring economic complexity has been a set goal. Starting from the Economic Complexity Index (ECI) developed by Hidalgo and Hausmann (2009)[4], it has become clear how most traditional economic growth theories often shrank internal socio-economic dynamics of countries through strict assumptions, restricting the analysis to a small subset of pre-determined factors. Unlike traditional growth theories, economic complexity measures are based on a data-driven approach and are generally agnostic about the determinants of countries' competitiveness. For instance, the ECI seeks to explain the knowledge accumulated by a country and expressed in all the economic activities present in that country. More and more refined measures of economic complexity have become available in the last few years. In a recent review, Hidalgo (2021)[6] identifies two main streams of the literature on economic complexity: the first involves metrics of so-called relatedness, whereas the second concerns economic complexity metrics, which apply dimensionality reduction techniques based, e.g., on Singular Value Decomposition (SVD). Metrics of relatedness measure the affinity between an activity and a location, while methods related to dimensionality reduction search for the best combination of factors explaining the structure of a given specialization matrix.

According to the principle of relatedness, the probability that a location c (e.g., a country) enters or exits an economic activity p (e.g., a sector) is influenced by the presence of related activities in that location. This poses, however, deeper questions about the role played by similar countries in determining the likelihood that the location c enters the economic activity p. Furthermore, while the principle of relatedness attempts to model the probability of entering an activity p by the location c, it does not provide hints about whether c will enter p successfully or not. Besides, there is a strong connection between the concept of production function – a function connecting economic inputs to outputs – and economic complexity via the SVD factorization of a suitable specialization matrix $\mathbf{R}$. Therefore, SVD is used to learn the singular vectors (factors) that best explain the structure of $\mathbf{R}$. The ECI index is closely related to the leading singular vectors of that specialization matrix (Hidalgo, 2021)[6], i.e. the truncated SVD of the matrix. These are also the leading eigenvectors of the product of the specialization matrix with its transpose. Usually, scholars select one of the first two eigenvectors (i.e., the ones associated with the two largest eigenvalues) because it carries out the maximum amount of information. Recently, Sciarra et al. (2020)[9] combined information coming from the first two eigenvectors into a unique index called GENeralised Economic comPlexitY (GENEPY). Nevertheless, it is worth noticing that, by doing this, the other eigenvectors are neglected and, together with them, further information which could potentially better explain economic complexity. Therefore, it looks reasonable to explore a suitable way to carefully select some other most informative eigenvectors (or, more in general, singular vectors) beyond the first two.

In this respect, the present paper exploits a class of machine-learning methods called Matrix Completion (MC) to extract the information coming from a subset of entries of the specialization matrix, to predict remaining entries of that matrix. Such information is encoded by the singular values and singular vectors of a suitable "approximating matrix" constructed

automatically by MC. The approach adopted here differs from (truncated) SVD (Hidalgo, 2021)[6], because in our framework, in order to predict a subset of its entries, the specialization matrix is assumed to be only partially observed (although the entries artificially obscured are later used as a ground truth, for validation and testing purposes). Moreover, we exploit the difficulty in predicting via MC some entries of the specialization matrix to quantify, in an aggregate way, the set of unobservables that make a country more competitive than expected. Such unobservables form the unexplained part of the specialization matrix. For the first time in the literature on economic complexity, a measure of the degree of predictability via MC of the entries of the specialization matrix corresponding to different countries is, then, used to define a novel index of economic complexity for countries (however, our proposed framework can be easily extended to the case of products). More precisely, our main idea is to adopt MC to infer information about the Relative Comparative Advantage (RCA), or disadvantage, of a country in a given trade category of products. Such information is collected, for each year, in a matrix, $\mathbf{RCA} \in \mathbb{R}^{C \times P}$, where C is the number of countries considered, and P is the number of products examined (at a given aggregation level). In formulas, one has

$$RCA_{c,p} := \frac{\frac{D_{c,p}}{\sum_{p'=1}^{P} D_{c,p'}}}{\frac{\sum_{c'=1}^{C} D_{c',p}}{\sum_{c'=1}^{C}\sum_{p'=1}^{P} D_{c',p'}}}, \tag{1}$$

where $D_{c,p}$ is the return in international dollars of the exports of the product p by country c. In case one among $D_{c,p}$, $D_{c,p'}$, $D_{c',p}$, and $D_{c',p'}$ in Eq. (1) is not available, one gets $RCA_{c,p} = NaN$. In this case, as a pre-processing step, that $RCA_{c,p}$ value can be replaced by 0.

In the paper, MC is applied several times (starting from different training subsets of suitably discretized RCA values associated with several countries and products, excluding originally NaN values) to estimate the expected RCA values of pairs of countries c and products p that have not been used in the training phase. To fulfill this task, the adopted MC technique is based on a soft-thresholded SVD, which selects each time – via a suitable regularization technique – the subset of most informative singular values and corresponding singular vectors. The predictions provided by MC are then exploited to construct two surrogate incidence matrices, one of which is used to compute a novel index of economic complexity, and the other one is used as an input to the GENEPY algorithm (Sciarra et al., 2020)[9].

The work contributes to the literature on economic complexity in three ways: (i) it applies for the first time MC to assess the complexity of countries; (ii) it defines a novel index of economic complexity based on MC; (iii) it builds up a comparison with a state-of-the-art index of economic complexity (GENEPY), revealing a high correlation between the output of GENEPY when it is applied to the original incidence matrix and the false positive rate of a binary classifier derived by the repeated application of MC. The results of our analysis show that MC performs well in estimating the RCA of countries. Supported by the high quality predictions of MC, we propose a novel Matrix cOmpletion iNdex of Economic complexitY (MONEY) for countries, which exploits the accuracy of their RCA predictions derived from the repeated applications of MC. Such accuracy is expressed in terms of a suitably weighted Area Under the Curve (AUC), one for each country examined. The MONEY index ranks countries according to their degree of predictability, taking into account also the complexity of the products. Specifically, the larger the AUC for a specific country and the larger the average with respect to a subset of the products of that country of the MC performance in estimating the discretized RCA values of country-product pairs, the less complex that country. Using MC to construct the proposed index helps to solve the shortcoming of GENEPY, i.e., the fact that, differently from MC, GENEPY takes into account only the information coming from two eigenvectors. Moreover, the GENEPY index computed using the MC surrogate incidence matrix reveals interesting discrepancies in terms of economic complexity with respect to the original GENEPY, i.e., the one calculated starting from the incidence matrix associated with the observed **RCA** matrix. By considering multiple years of data (see the Supplemental), we find a strong and significant positive correlation between the false positive rate of the binary classifier derived from thresholding the average output of MC and the original GENEPY index.

Predicting the revealed competitive advantage of countries: a matrix completion approach

In this work, we apply Matrix Completion (MC) techniques to study economic complexity. This class of machine-learning methods has been popularized by the so-called Netflix competition; see the Supplemental for further details on MC and Hastie et al. (2015)[4], Alfakih et al. (2000)[1], and Cai et al. (2010)[2] for some of its applications. This paper uses MC to estimate the expected revealed competitive advantage (RCA) of countries c and products p. The specific MC method adopted in the paper consists in completing a partially observed matrix $\mathbf{A} \in \mathbb{R}^{C \times P}$ (which is derived from the **RCA** matrix in our case), by minimizing a suitable trade-off between the reconstruction error of the known portion of that matrix and a penalty term, which penalizes a high nuclear norm of the reconstructed (or completed) matrix. This is formulated via the following optimization

problem (Mazumder et al., 2010)[7]:

$$\underset{\mathbf{Z}\in\mathbb{R}^{C\times P}}{\text{minimize}}\left(\frac{1}{2}\sum_{(c,p)\in\Omega^{\text{tr}}}(A_{c,p}-Z_{c,p})^2+\lambda\|\mathbf{Z}\|_*\right), \tag{2}$$

where Ω^{tr} is a training subset of pairs of indices (c,p) corresponding to positions of known entries of the partially observed matrix $\mathbf{A}\in\mathbb{R}^{C\times P}$, $\mathbf{Z}\in\mathbb{R}^{C\times P}$ is the completed matrix (to be optimized), $\lambda\geq 0$ is a regularization constant (chosen by a suitable validation method), and $\|\mathbf{Z}\|_*$ is the nuclear norm of the matrix $\mathbf{Z}$, i.e., the sum of all its singular values. The reader if referred to the Supplemental for further technical details on the optimization problem (2) and on the algorithm we used to solve it.
While MC has already found many applications in many fields (e.g., movie recommendation, sensor engineering, econometrics), to the best of our knowledge, this is the first time it is used to analyze economic complexity. More precisely, we applied MC to define a novel complexity index to be compared with state-of-the-art complexity indices.
In our application of MC to economic complexity, the MC optimization problem (2) was solved several times by a specific algorithm previously developed for that purpose, named Soft Impute in Mazumder et al. (2010)[7] (see the Supplemental), for different choices of the regularization parameter λ and of the subset Ω^{tr} (detailed later in this section). Then, two MC surrogates $\overline{\mathbf{M}}^{(MC)}$ and $\hat{\mathbf{M}}^{(MC)}$ of the incidence matrix $\mathbf{M}\in\mathbb{R}^{C\times P}$ were generated (the reader is referred to the Supplemental for details on how the incidence matrix $\mathbf{M}$ is defined, starting from the **RCA** matrix). On one hand, $\overline{\mathbf{M}}^{(MC)}$ was exploited to evaluate the performance of MC by changing a suitable threshold. This allowed to build up performance measures that compose the MONEY index. On the other hand, $\hat{\mathbf{M}}^{(MC)}$ was used as input to the GENEPY algorithm, to construct a counterfactual $\widehat{\text{GENEPY}}^{(MC)}\in\mathbb{R}^{C\times P}$ to be compared with the GENEPY index computed using the original incidence matrix $\mathbf{M}$. In the following, we describe our approach of applying MC to the reconstruction of the **RCA** matrix for the case in which the products were aggregated at the 4-digits level in the Harmonized System Codes 1992 (HS-1992). Consistently with the literature (Sciarra et al., 2018)[8], we constructed the matrix $\mathbf{A}$ (one of the inputs to the optimization problem (2)) by discretizing the elements of the **RCA** matrix. For the sake of brevity, we refer to the MC application to the definition of a measure of complexity of the countries. To get a measure of complexity of the products, it is enough to replace the matrix $\mathbf{A}$ with its transpose (see also the Supplemental for some related results).

1. For the matrix $\mathbf{A}\in\mathbb{R}^{C\times P}$ (where $C=119$ is the number of countries, and $P=1243$ is the number of products), the MC optimization problem (2) was solved $N=1000$ times by the Soft Impute algorithm, based on various choices for the training/validation/test sets (and, as already mentioned, for the regularization parameter λ).
2. For each such repetition $n=1,\ldots,N$, the sets above were constructed as follows. First, a (pseudo)random permutation of the rows of $\mathbf{A}$ was generated. Then, a subset S_n of these rows was considered, by including in it the first row in the permutation and the successive $s\%\simeq 25\%$ rows. In this way, the resulting number of elements of the set S_n was $|S_n|=30$. Next, for each row in S_n, its elements belonging to all the groups except group "0" were obscured independently with probability $p_{\text{missing}}=0.3$. The (indices ot the) remaining entries of the matrix $\mathbf{A}$ (excluding the ones belonging to the group "0") formed the training set (denoted by Ω^{tr_n}). The obscured entries in one of the $|S_n|$ rows (say, row $h\in\{1,\ldots,|S_n|\}$) formed the test set (denoted by $\Omega^{\text{test}_{n,h}}$), whereas the obscured entries in the remaining $|S_n|-1$ rows formed the validation set (denoted by $\Omega^{\text{val}_{n,h}}$).
3. For each repetition n, the generation of the validation and test sets from the set S_n was made $|S_n|$ times, each time with a different selection of the row h associated with the test set (and, as a consequence, also of the $|S_n|-1$ rows associated with the validation set). Hence, the same training set was associated with $|S_n|$ different pairs of validation and test sets (the number of repetitions $N=1000$ and the percentage $s\%\simeq 25\%$ were selected in order to associate each row with the test set a sufficiently large number of times, with high probability; in particular, with these choices, the average number of times each row was associated with the test set was about 250). In this way, for each choice of S_n and of the regularization parameter λ, the MC optimization problem (2) was solved once instead of $|S_n|$ times, thus improving the computational efficiency. Finally, by construction, each time there was no overlap between the training, validation, and test sets.
4. To avoid overfitting, for each choice of the training set Ω^{tr_n}, the optimization problem (2) was solved for 30 choices λ_k for λ, exponentially distributed as $\lambda_k=2^{(k-1)/2}$ for $k=1,\ldots,30$. The resulting completed and post-processed matrix was indicated as $\mathbf{Z}^{(n)}_{\lambda_k}$. Then, for each λ_k and each of the $|S_n|$ selections of the validation sets associated with the same training set, the Root Mean Square Error (RMSE) of matrix reconstruction on that validation set was computed as

$$RMSE^{\text{val}_{n,h}}_{\lambda_k}:=\sqrt{\frac{1}{|\Omega^{\text{val}_{n,h}}|}\sum_{(c,p)\in\Omega^{\text{val}_{n,h}}}\left(A_{c,p}-Z^{(n)}_{\lambda_k,c,p}\right)^2}, \tag{3}$$

then the choice $\lambda_{k^\circ(n,h)}$ minimizing $RMSE^{\text{val}_{n,h}}_{\lambda_k}$ for $k=1,\ldots,30$ was found. Finally, the RMSE of matrix reconstruction on

the related test set was computed in correspondence of the so-obtained optimal value $\lambda_{k^\circ(n,h)}$ as

$$RMSE^{\text{test}_{n,h}}_{\lambda_{k^\circ(n,h)}} := \sqrt{\frac{1}{|\Omega^{\text{test}_{n,h}}|}\sum_{(c,p)\in\Omega^{\text{test}_{n,h}}}\left(A_{c,p}-Z^{(n)}_{\lambda_{k^\circ(n,h)},c,p}\right)^2}. \tag{4}$$

5. For each choice of n and h, the MC predictions contained in the matrix $\mathbf{Z}^{(n)}_{\lambda_{k^\circ(n,h)}}$ were used to build a binary classifier. More precisely, each time an element $A_{c,p}$ of the matrix $\mathbf{A}$ was in the test set, such element was attributed to the class 0 (corresponding to the case $0 \leq RCA < 1$) when its MC prediction from $\mathbf{Z}^{(n)}_{\lambda_{k^\circ(n,h)}}$ was lower than 0, otherwise it was attributed to the class 1 (corresponding to the case $RCA \geq 1$). Finally, the average classification of the element $A_{c,p}$ (with respect to all the test sets to which that element belonged) was indicated as $\overline{A}^{(MC)}_{c,p} \in [0,1]$, whereas its most frequent classification (either 0 or 1) was indicated as $\hat{A}^{(MC)}_{c,p}$. A random assignment between 0 and 1 was made to deal with ties. In the (unlikely) case the element $A_{c,p}$ appeared in none of the test sets, both $\overline{A}^{(MC)}_{c,p}$ and $\hat{A}^{(MC)}_{c,p}$ were chosen to be equal to 0 (due to the choice $p_{\text{missing}} = 0.3$, each element $A_{c,p}$ not associated with the group "0" appeared in the test set on average about 75 times; so, the probability that one such element appeared in none of the test sets was negligigle).
6. A first MC surrogate $\overline{\mathbf{M}}^{(MC)} \in \mathbb{R}^{119\times 1243}$ of the incidence matrix $\mathbf{M}$ was defined as follows:

$$\overline{M}^{(MC)}_{c,p} \doteq \begin{cases} 0, & \text{if } RCA_{c,p} = NaN, \\ \overline{A}^{(MC)}_{c,p}, & \text{otherwise}. \end{cases} \tag{5}$$

Similarly, a second MC surrogate $\hat{\mathbf{M}}^{(MC)} \in \mathbb{R}^{119\times 1243}$ of the incidence matrix $\mathbf{M}$ was defined as follows:

$$\hat{M}^{(MC)}_{c,p} \doteq \begin{cases} 0, & \text{if } RCA_{c,p} = NaN, \\ \hat{A}^{(MC)}_{c,p}, & \text{otherwise}. \end{cases} \tag{6}$$

7. Finally, $\overline{\mathbf{M}}^{(MC)}$ was combined with several thresholds from 0 to 1 in the first part of the construction of the proposed MONEY index (see the next section for details). Instead, $\hat{\mathbf{M}}^{(MC)}$ was provided as input to the GENEPY algorithm (replacing the original incidence matrix $\mathbf{M}$). In this way, a counterfactual GENEPY index, indicated as $\widehat{\text{GENEPY}}^{(MC)}$, was generated.

In order to assess the prediction capability of the binary classifier associated with MC (see Step 5 above), for each row (country) c of $\mathbf{A}$, we also computed the false positive rate fpr_c and the false negative rate fnr_c as the average classification error frequency, respectively, of the true negative/true positive examples in all the test sets associated with that row (where the "negative class" refers to the class 0 associated with $0 \leq RCA < 1$, and the "positive class" to the class 1 associated with $RCA \geq 1$).

The Matrix cOmpletion iNdex of Economic complexitY (MONEY)

In this section, we introduce our proposed economic complexity index, called Matrix cOmpletion iNdex of Economic complexitY (MONEY), whose construction is based on MC.

The MONEY index is built starting from the matrix $\overline{\mathbf{M}}^{(MC)}$ introduced in the section above. It is based on constructing a binary classifier for each country by combining the corresponding row of $\overline{\mathbf{M}}^{(MC)}$ with a threshold, then assessing the performance of the resulting MC classifications at the level of each country. First, for the binary classifier associated with each country, a Receiver Operating Characteristic (ROC) curve (denoted as ROC_c) is constructed, based on a country-dependent threshold. The corresponding Area Under the Curve (AUC) is denoted as AUC_c. We remind the reader that, for a binary classifier, the ROC curve expresses the trade-off between fall-out (false positive rate) and sensitivity (true positive rate) of that classifier, as a function of its threshold. It is recalled here that the true positive rate is equal to 1 minus the false negative rate. In general, ROC curves closer to the top-left corner indicate a better performance. As a baseline ("Bench."), a random guessing binary classifier is associated with a ROC curve with points lying along the diagonal indicated, e.g., in Fig. 1 (for which the true positive rate is equal to the false positive rate). The closer a ROC curve to the diagonal in the ROC space, the worse the performance of the associated binary classifier. It is worth reminding the reader that ROC curves do not depend on class frequencies. This makes them useful for evaluating classifiers predicting rare events as in the case of very high RCA values. We also remind the reader that the AUC measures the area of the entire two-dimensional region underneath the entire ROC curve from $(0,0)$ to $(1,1)$. The AUC is exploited in the literature to provide an aggregate measure of performance across all possible classification thresholds. Formally, it represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, assuming that "positive" ranks higher than "negative" (Fawcett, 2006[3]).

In more details, for each country c, the elements of the c-th row of the matrix $\overline{\mathbf{M}}^{(MC)}$ are compared with a threshold to construct the associated binary classifier. The elements belonging to the same row of the original incidence matrix $\mathbf{M}$ are taken as ground truth. The discrimination threshold is varied from 0 to 1, using a step size equal to 0.01. All the elements of $\overline{\mathbf{M}}^{(MC)}$ are used as dataset, except the ones having the same indices as the originally *NaN* values in the **RCA** matrix. This allows to form a binary classifier for each threshold and for each country. The idea now is to exploit the AUC_c of the binary classifiers associated with the countries in order to provide a measure of complexity of such countries, based on the degree of predictability of the corresponding rows. Specifically, countries with lower AUC_c may be considered as more complex, being harder for MC to predict their RCA entries. The AUC_c alone, however, does not capture the reasons why MC performed poorly (or, vice versa, adequately). As an example, consider the three following hypothetical scenarios. Assume that MC performs poorly on a country c by attributing $RCA \geq 1$ to a product p when its true RCA was smaller than 1, and assigned correctly a RCA smaller than 1 to all the other countries for the same product p (Scenario 1). Consider now the two following similar scenarios for which, for the same product p and the same country c, MC performs poorly on the country c by attributing $RCA \geq 1$ to the product p when its true RCA was smaller than 1, and it attributed either correctly (Scenario 2) or incorrectly (Scenario 3) $RCA \geq 1$ to all the other countries for the same product. It is reasonable to suppose that, all other things being equal, the country c to which MC assigned $RCA \geq 1$ for the product p in Scenario 1 is more complex than the same country to which MC assigned $RCA \geq 1$ for the product p in Scenarios 2 and 3. In fact, while in Scenario 2, MC could have been driven to predict, for country c, a RCA of p larger than or equal to 1 by the presence of several RCA entries larger than or equal to 1 for the other countries, this is not the case for Scenario 1. Scenario 3 is more unlikely to occur, since, as it is shown later in the next section and in the Supplemental, MC has typically a quite satisfying prediction capability in its specific application to the **RCA** matrix. In this case, it is not possible to conclude that country c is more complex than the other countries, since MC is wrongly attributing $RCA \geq 1$ to p, for all such countries.

The example above suggests us that, by adopting the AUC_c alone as a complexity measure, country c would be classified as equally complex in Scenarios 1, 2 and 3 (assuming the AUC_c being equal in all these cases). In order to correct for this, we propose a refined complexity measure, based on weighting the AUC_c for each country c. The rationale of the proposed complexity measure is that not only less predictable countries (according to MC) are more complex, but one should also take into account the product dimension when comparing the MC predictions obtained for different countries, controlling for the quality of each prediction. More precisely, it is proposed to associate a weight w_c to each country c, which is constructed in such a way that the AUC_c's of countries with an higher share of "rare" false positives are weighted less (since they are less predictable). In more details, the proposed complexity measure is constructed as follows.

a First, the MC analysis made for the countries is repeated for the products, still referring to the same year. This is obtained simply by replacing at the beginning of the analysis the **RCA** matrix with its transpose. Analogously, the matrices $\hat{\mathbf{M}}^{(MC)}$ and $\overline{\mathbf{M}}^{(MC)}$ are replaced by similarly constructed matrices $(\hat{\mathbf{M}}^{\top})^{(MC)}$ and $(\overline{\mathbf{M}}^{\top})^{(MC)}$. In particular, each element of the latter matrix represents the average MC prediction for the corresponding product-country pair.

b Then, a threshold t is applied to the elements of the matrix $(\overline{\mathbf{M}}^{\top})^{(MC)}$. For each value of that threshold, one constructs a matrix $\left(\overline{\mathbf{M}}_t^{\top}\right)^{(MC)} \in \{0,1\}^{P\times C}$, being each entry of it equal to 1 whenever the corresponding element in the matrix $(\overline{\mathbf{M}}^{\top})^{(MC)}$ is higher than or equal to t, otherwise being it equal to 0.

c At this point, for each product p and each threshold t, one computes the quantity

$$ftot_{p,t} := fpr_{p,t} \times \frac{N_p}{P_p + N_p}, \tag{7}$$

being $fpr_{p,t}$ the false positive ratio for the classifications associated with that product (determined by the comparison between $\left(\overline{\mathbf{M}}_t^{\top}\right)^{(MC)}$ and $\mathbf{M}^{\top}$, restricted to the entries associated with that product) and $\frac{N_p}{N_p+P_p}$ the proportion of entries with true $RCA < 1$ with respect to all the entries associated with that product (i.e., 119). Besides, the average $\overline{ftot}_p$ of $ftot_{p,t}$ with respect to t is computed.

d Then, for each country c, the weight w_c is defined as follows:

$$w_c := \frac{\sum_{p=1}^{P} (\hat{M}^{\top})_{p,c}^{(MC)} \times \overline{ftot}_p}{\sum_{p=1}^{P} (\hat{M}^{\top})_{p,c}^{(MC)}}. \tag{8}$$

In other words, for each country c, the weight w_c is the average of $\overline{ftot}_p$ with respect to all the products p for which one predicts $RCA \geq 1$ through the surrogate incidence matrix $(\hat{\mathbf{M}}^{\top})^{(MC)}$.

e Finally, the MONEY index for each country c is computed as:

$$MONEY_c := 1 - w_c \times AUC_c. \tag{9}$$

Results

Global performance of matrix completion

In the following, the diagnostic ability of MC is illustrated. Likewise in the section above, the matrix $\overline{\mathbf{M}}^{(MC)}$ was combined with a threshold to construct a binary classifier (in this case, however, differently from that section, the threshold did not depend on the country). The discrimination threshold was varied from 0 to 1, using a step size equal to 0.01. All the elements of $\overline{\mathbf{M}}^{(MC)}$ were used as dataset, except the ones having the same indices as the originally *NaN* values in the **RCA** matrix. The ground truth was provided by the corresponding elements of the original incidence matrix $\mathbf{M}$. Figure 1 shows the resulting ROC curve. Similarly, ROC_c curves for a random sample of countries are displayed in Fig. 2.

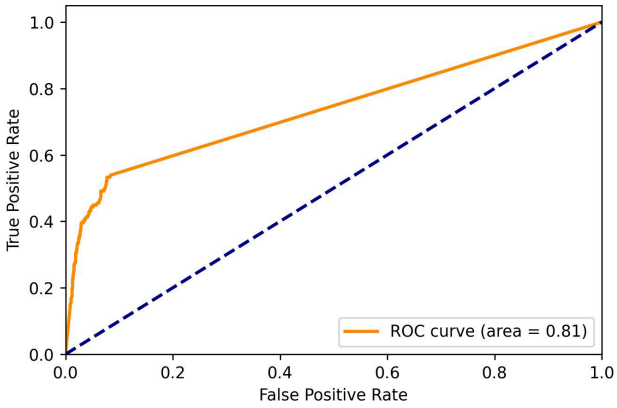


Figure 1. Global ROC curve constructed starting from the matrix $\overline{\mathbf{M}}^{(MC)}$, for the year 2018. "Bench." stands for the line passing through the origin with slope 1.

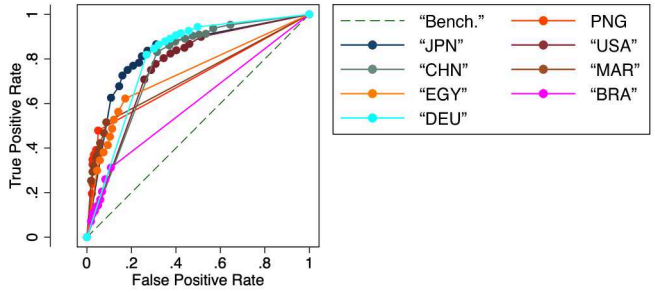
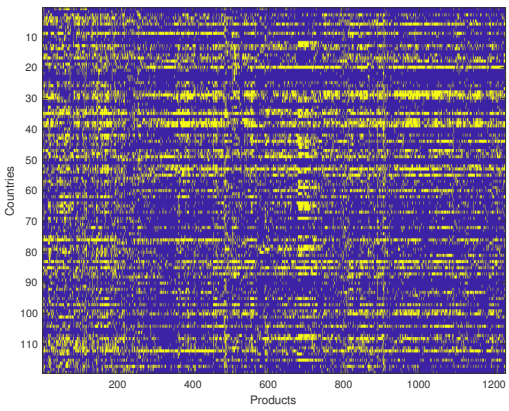
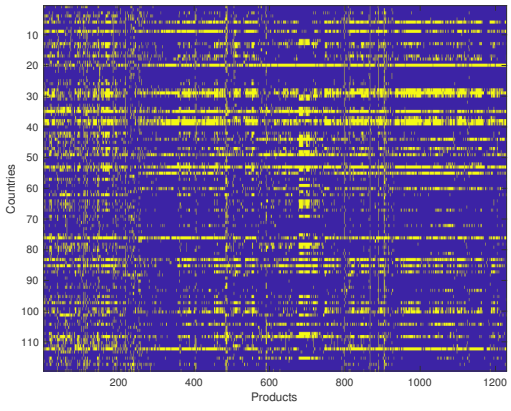


Figure 2. ROC_c curves constructed starting from the matrix $\overline{\mathbf{M}}^{(MC)}$, for a sample of countries and for the year 2018. "Bench." stands for the line passing through the origin with slope 1.

As it is evident from Figs. 1 and 2, MC performed quite well on average both globally and for developed countries such as Japan, United States and Germany. Its performance was poorer (though still above the baseline) for countries that either provided less information on their trade flows or whose trade flows were extremely volatile (i.e., they alternated between products with extremely high RCA values and products with very low RCA values). Specifically, fnr_c was higher for the latter countries. Nonetheless, the average performance of MC over all the countries was high as depicted by the AUC reported in Fig. 1, which turned out to be about 0.81 for the binary classifier described in Step 5 of our proposed approach. As a further check, since the positive and negative labels were unbalanced in the original dataset (specifically, entries with $RCA < 1$ represented almost the 70% of the entire dataset), we also applied the Balanced Accuracy (BACC) index, which turned out to be 0.75. We recall that the BACC is a performance metric designed for binary classifiers in the case of unbalanced datasets. It is calculated as the average of the proportion of correctly classified elements of each class individually, and ranges from 0 (low balanced accuracy) to 1 (maximum balanced accuracy). Formally, it is equal to $(tpr + tnr)/2$, where t stands for "true". Figures 3a-3b display the original incidence matrix $\mathbf{M}$ as compared to the MC surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$ obtained at the HS-4 level of product aggregation. The two matrices display similar but not identical entries. On one hand, their similarity confirms the good MC prediction performance at a global level. On the other hand, their differences could be attributed to the high complexity of specific country/product pairs being predicted. In other words, there may be a discrepancy between the actual RCA value of a country/product pair and its potential RCA value, predicted by MC on the basis of similar country/product pairs.



(a) Original incidence matrix $\mathbf{M}$ at the HS-4 level.

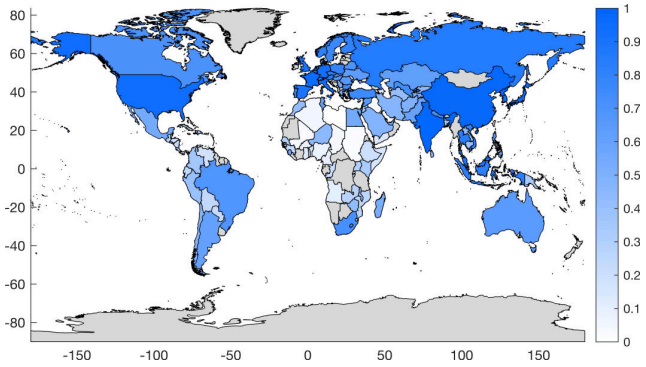


(b) Surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$ at the HS-4 level.

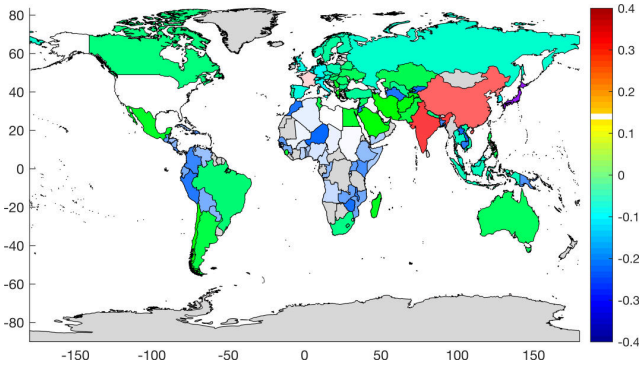
Figure 3. Original versus surrogate incidence matrix for the year 2018 at the HS-4 level.

Results related to the MONEY index

In this section we report the ranking of countries in terms of economic complexity as expressed by the MONEY index. In particular, we represent the countries according to their MONEY index (Fig. 4a), then we compare the obtained ranking with the one expressed by GENEPY (Fig. 4b). In Fig. 4a, countries are colored according to their MONEY values (normalized between 0 and 1), which are proportional to the shade of blue. In particular, the color map ranges from the least complex countries c (colored in white) to the most complex ones (colored in dark blue).



(a) Countries colored according to the MONEY index. Countries in darker shades of blue are associated with a lower MONEY, hence they are considered more complex. Countries colored in grey are not considered in the analysis.



(b) Countries colored according to the difference between the values of their MONEY and GENEPY indices. Countries in darker shades of blue are associated with a value of MONEY<GENEPY, vice versa countries in red.

Figure 4. Values of the MONEY index for the year 2018 at the HS-4 level of aggregation and their differences with respect to the corresponding values of the GENEPY index for the same year and the same level of aggregation.

It is worth observing that both the GENEPY and the proposed MONEY index arise from the attempt to reconstruct (in a different way for each method) a matrix related to trade flows. In the case of GENEPY, the matrix is a proximity matrix $\mathbf{N}$ derived from the incidence matrix $\mathbf{M}$ (see the Supplemental for the definition of the matrix $\mathbf{N}$), and its reconstruction is obtained as a nonlinear least-square estimate based on the components of the first two (normalized) eigenvectors of that matrix. Then, a successive evaluation on how the quality of the estimate changes by dropping specific components of such eigenvectors (the ones associated with a given country) is made. In our case, the matrix $\mathbf{A}$ is obtained as a discretization of the **RCA** matrix. Then, MC is applied several times to the matrix $\mathbf{A}$ to reconstruct a portion of that matrix which has been obscured, in the attempt to uncover a "latent" similarity between countries, which can be useful for the prediction of whether their RCA entries are lower than 1, or higher than or equal to 1. Another difference is that the matrix reconstruction on which GENEPY is based relies only on two eigenvectors of $\mathbf{N}$, whereas our method, being also based on MC, exploits a typically much larger number of left-singular/right-singular vectors to build the reconstructed matrix, for each application of MC. The choice of the

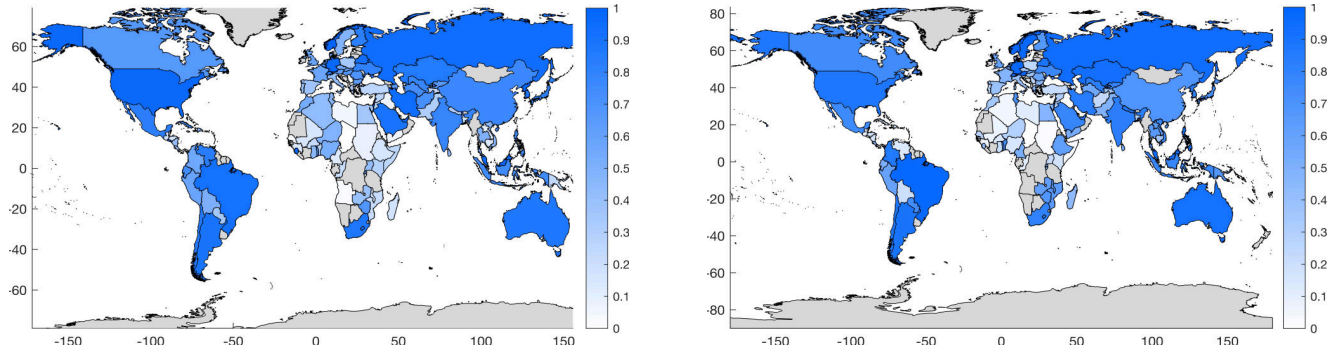
number of such pairs is made automatically by the adopted validation procedure. Moreover, a final evaluation of the quality of the reconstruction is made, by considering several test sets, on which the AUC_c's are based. A further quality assessment is provided by Table 1, which reports the number of G19+5 countries in the top 20, 30 and 40 positions, computed according to each among the GENEPY, $\widehat{\text{GENEPY}}^{(MC)}$ and MONEY indices, then divided by 24 (in Table 1 we considered countries within G20; however, since G20 countries comprise EU – except France, Italy and Germany, which are accounted separately –, that is an agglomerate of countries, we considered a group of 5 representative countries for EU, namely: Spain, Switzerland, Greece, Denmark and Hungary). It is evident from Table 1 that the largest ratio is obtained in correspondence of the proposed MONEY index.

	# of G19+5 countries in the top x positions/24		
Index	$x = 20$	$x = 30$	$x = 40$
GENEPY	42%	58%	75%
$\widehat{\text{GENEPY}}^{(MC)}$	55%	64%	71%
MONEY	55%	66%	79%

Table 1. Number of G19+5 countries in the top 20, 30 and 40 positions for the year 2018, computed according to each among the GENEPY, $\widehat{\text{GENEPY}}^{(MC)}$ and MONEY indices, then divided by 24.

MONEY index for the years 2005 and 2014

The present subsection provides robustness checks on the MONEY index. In particular, the whole analysis has been repeated for the years 2005 and 2014. For such years, the global AUC is respectively 0.76 for 2005 and 0.80 for 2014. For the same reason as reported in the main text – that is, the fact that the two categories $RCA > 1$ and $0 \leq RCA < 1$ are unbalanced – we also computed the BACC. The latter amounts to 0.74 for 2005 and to 0.75 for 2014.
Figures 5a-5b display the MONEY index for the years 2005 and 2014.



(a) Values of the MONEY index for the year 2005 at the HS-4 level of aggregation.

(b) Values of the MONEY index for the year 2014 at the HS-4 level of aggregation.

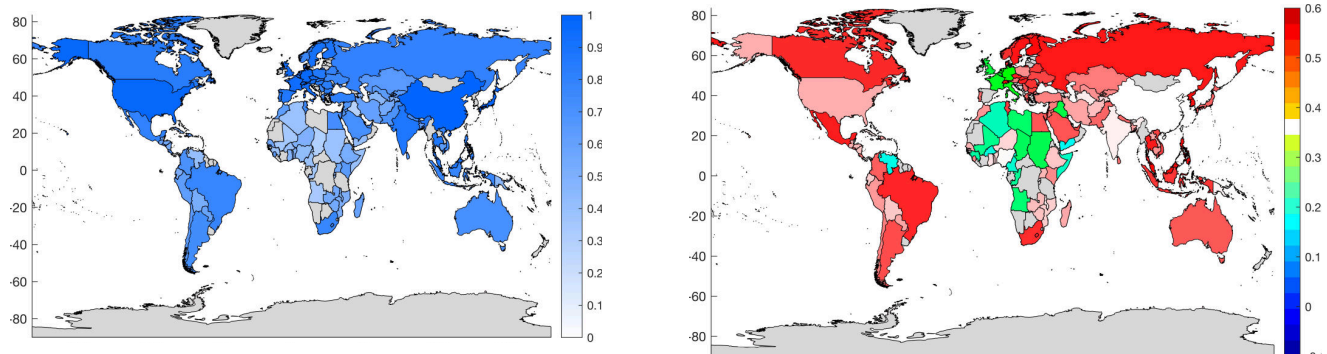
Figure 5. Values of the MONEY index for the years 2005 and 2014 at the HS-4 level of aggregation.

The figures display similar results to those obtained in the main analysis. However, some differences emerge. Specifically, Russia and East Asia appear to be more complex in 2005 and 2014 than in 2018. Such countries were indeed rapidly developing in those years and their growth rate was higher than the one of EU, which was slowed down by the 2008 financial crisis. The crisis had an impact also on USA. Its MONEY index was in fact lower in 2014 as compared to 2005. In the successive years, yet, EU and USA recovered from the crisis, and their complexity, as measured by the MONEY index, raised accordingly in 2018.

Differences in the GENEPY indices based on the original incidence matrix $\mathbf{M}$ and on $\hat{\mathbf{M}}^{(MC)}$

Figure 6a represents the GENEPY index computed based on the original incidence matrix $\mathbf{M}$. The interpretation is the same as in Fig. 4a. It is worth noticing that the GENEPY value computed based on the incidence matrix $\mathbf{M}$ and the $\widehat{\text{GENEPY}}^{(MC)}$ value based on its surrogate $\hat{\mathbf{M}}^{(MC)}$ are quite similar (see Fig. 6b for their difference). Hence, they provide analogous results in terms of the complexity of the countries, confirming the satisfactory prediction capability of MC for the specific learning task. Nevertheless, one can also notice that the two complexities differ in some countries. Such differences may be ascribed to surpluses/deficits of the actual complexities of such countries (i.e., the ones measured by GENEPY based on the original incidence matrix $\mathbf{M}$) with respect to the respective predicted complexities (i.e., the ones measured by $\widehat{\text{GENEPY}}^{(MC)}$, which is

based on the surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$).
To quantify the correlation between the GENEPY rankings computed based on $\mathbf{M}$ and $\hat{\mathbf{M}}^{(MC)}$, respectively, we evaluated their Kendall rank correlation coefficient τ_k. The statistical test produced $\tau_k \simeq 0.8$ with a p-value near 0, rejecting significantly the null hypothesis of independence between GENEPY and $\widehat{\text{GENEPY}}^{(MC)}$.



(a) Countries colored according to the GENEPY index computed starting from the original incidence matrix $\mathbf{M}$ at the HS-4 level of aggregation.

(b) Difference between the GENEPY index in (a) and the $\widehat{\text{GENEPY}}^{(MC)}$ index based on its surrogate $\hat{\mathbf{M}}^{(MC)}$.

Figure 6. Original GENEPY values for countries for the year 2018 at the HS-4 level of aggregation, and their comparison with their $\widehat{\text{GENEPY}}^{(MC)}$ values. Countries colored in grey are not considered in the analysis.

It is worth noticing that, with a few exceptions (China, France, Italy, UK and Germany) the more complex the country according to GENEPY, the higher the difference between GENEPY and $\widehat{\text{GENEPY}}^{(MC)}$. Finally, Fig. 7 displays the false positive rate fpr_c for each country considered in the analysis, which turned out to produce a ranking of countries quite similar to the one generated by GENEPY ($\tau_k = 0.75$).

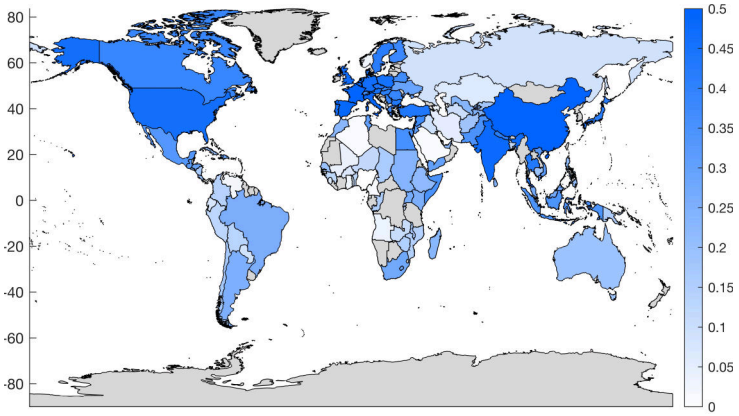


Figure 7. False positive rate fpr_c, reported proportionally to the shade of blue for the year 2018 and the product aggregation level HS-4.

Discussion

In the present work, we applied Matrix Completion (MC) to investigate in various ways the economic complexity of countries. First, we assessed a quite high accuracy of the MC predictions, when MC was applied to reconstruct the Revealed Comparative Advantage (RCA) matrix, which is at the basis of the construction of several existing economic complexity indices (see the Supplemental). Then, we proposed the Matrix cOmpletion iNdex of Economic complexitY (MONEY), based on the degree of predictability of the RCA entries associated with different countries. As an additional contribution, we combined MC with a recently-developed economic complexity index (GENEPY), to assess the expected economic complexity of countries. In the work, MC was exploited to infer the expected discretized RCA of a country c in a certain class of goods or services p. The MC technique employed is based on a soft-thresholded SVD. This, combined with the MC validation phase, allows to select automatically a suitable number of singular vectors to be used to reconstruct the discretized **RCA** matrix. In this way,

differently from previous economic complexity indices, the information extracted is not restricted to the first two singular vectors.

The results of our analysis highlighted a generally quite good performance of MC in discerning country-product pairs with RCA values greater than or equal to the critical threshold of 1, denoting the competitiveness of c in producing p. The outcomes were summarized by reporting the global ROC curve and comparing the heat-map of the true incidence matrix $\mathbf{M}$ and the one of its MC surrogate matrix $\hat{\mathbf{M}}^{(MC)}$, which was obtained from various applications of MC. Motivated by the high MC accuracy, we developed the MONEY index taking into account both the predictive performance of MC for each country (as measured by its AUC_c) and the product dimension. In other words, when constructing that index, each AUC_c was weighted by the average of the $\overline{ftot}_p$'s with respect to a subset of products associated with the specific country. As a further step, we applied the GENEPY algorithm first to the incidence matrix derived directly from the original **RCA** matrix, then to the MC surrogate incidence matrix $\hat{\mathbf{M}}^{(MC)}$. This allowed us to directly compare the values of the two GENEPY indices, thus assessing their potential discrepancies. On average, such discrepancies were higher for more complex countries according to the original GENEPY index.

References

1. Alfakih, A., & Wolkowicz, H. (2000). Matrix completion problems. In Handbook of semidefinite programming (pp. 533–545). Springer.

2. Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on optimization, 20(4), pp. 1956–1982.

3. Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

4. Hastie, T., Mazumder, R., Lee, J.D. and Zadeh, R., 2015. Matrix completion and low-rank SVD via fast alternating least squares. The Journal of Machine Learning Research, 16(1), pp. 3367–3402.

5. Hidalgo, C. A., & Hausmann, R (2009). The building blocks of economic complexity. Proceedings of the National Academy of Sciences, 106, pp. 10570–10575.

6. Hidalgo, C. A. (2021). Economic complexity theory and applications. Nature Reviews Physics, 3, pp. 92–113.

7. Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11, pp. 2287–2322.

8. Sciarra, C., Chiarotti, G., Laio, F., & Ridolfi, L. (2018). A change of perspective in network centrality. Scientific Reports, 8, article no. 15269.

9. Sciarra, C., Chiarotti, G., Ridolfi, G., & Laio, F. (2020). Reconciling contrasting views on economic complexity. Nature Communications, 11, article no. 3352.

Author contributions statement

M.R. conceived the idea of the work, G.G. and F.N. developed the proposed method and implemented it. All the authors analysed the results and reviewed the manuscript.

Additional information

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

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplemental.pdf