AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-022-00448-6

ORIGINAL PAPER

Hierarchical clustering and matrix completion for the reconstruction of world input–output tables

Rodolfo Metulini[1] · **Giorgio Gnecco**[2] · **Francesco Biancalani**[2] · **Massimo Riccaboni**[2]

Received: 31 December 2020 / Accepted: 28 April 2022

Abstract

Multi-regional input–output (I/O) matrices provide the networks of within- and cross-country economic relations. In the context of I/O analysis, the methodology adopted by national statistical offices in data collection raises the issue of obtaining reliable data in a timely fashion and it makes the reconstruction of (parts of) the I/O matrices of particular interest. In this work, we propose a method combining hierarchical clustering and matrix completion with a LASSO-like nuclear norm penalty, to predict missing entries of a partially unknown I/O matrix. Through analyses based on both real-world and synthetic I/O matrices, we study the effectiveness of the proposed method to predict missing values from both previous years data and current data related to countries similar to the one for which current data are obscured. To show the usefulness of our method, an application based on World Input–Output Database (WIOD) tables—which are an example of industry-by-industry I/O tables—is provided. Strong similarities in structure between WIOD and other I/O tables are also found, which make the proposed approach easily generalizable to them.

Keywords Matrix completion · LASSO-like nuclear norm penalty · Panel data analysis · Hierarchical clustering · Input–output tables

✉ Rodolfo Metulini
rmetulini@unisa.it

[1] Department of Economics and Statistics (DISES), University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, SA, Italy

[2] Laboratory for the Analysis of CompleX Economics Systems (AXES), IMT School for Advanced Studies, Piazza S. Francesco, 19, 55100 Lucca, Italy

Published online: 02 June 2022

1 Introduction

The world economy is characterized by the interdependence of all countries with respect to their industrial activities. It can be modeled as a network in which each "country–sector" pair exchanges goods and services with any other (and even the same) such pair, at a different extent and using different technologies (Cerina et al. 2015). The structure of this network may provide information about such interdependencies of national economies and their changes over time. The literature usually refers to the study of these interdependencies with the term "input–output (I/O) analysis." I/O tables portray flows of what is produced by some economic agents and used by other economics agents (either as intermediate or final consumption). They are arranged into two types, either according to relationships between industries (industry-by-industry I/O tables) or according to relationships between products (product-by-product I/O tables). Many world I/O tables for selected countries are periodically produced by, e.g., the World Input–Output Database (WIOD, Timmer et al. 2015, 2016), the EORA Global Supply Chain initiative (Lenzen et al. 2012), EXIObase (Tukker et al. 2013), the Full International and Global Accounts for Research in input–output analysis (FIGARO, Rémond-Tiedrez and Rueda-Cantuche 2019) from EUROSTAT and (OECD 2018).[1] By analyzing I/O matrices, several authors (Cerina et al. 2015; McNerney et al. 2013; Zhu et al. 2015) studied the network topology of inter-industry flows and others (Liang et al. 2016) investigated its scaling patterns. Some studies highlighted the presence of significant asymmetry between inflow data (i.e., what a "country–sector" pair uses from other "country–sector" pairs) and outflow data (i.e., what a "country–sector" pair sells to other "country–sector" pairs), and the emergence of a clustering pattern among either countries or sectors. This can derive, e.g., from the presence of similar production technologies (i.e., similar dependence of two specific countries or sectors on other "country–sector" pairs (Carvalho 2009). This issue was taken into account, for instance, by Zhu et al. (2018), who used an ad hoc measure to detect patterns of similarity and to group together similar countries and similar sectors, and by Oliva et al. (2016), who applied a spectral clustering approach to group similar sectors of the Italian economy.

Our work has to do with the reconstruction of I/O matrices. The interest in this topic arises if one takes into account some aspects related to the methodology of data collection adopted by national statistical offices. Indeed, based on the work by Percoco et al. (2006) and Wen et al. (2014) provide the following motivations to apply a reconstruction method to I/O matrices:

- Data collection for I/O matrices is typically based on direct methods, e.g., on surveys (made at different times and in different countries). This leads to classical sampling errors;

[1] Other economic applications of network studies include, e.g., the cases of ownership networks (Riccaboni et al. 2021), banking networks (Iori et al. 2008), international trade networks and foreign direct investment networks (Fagiolo et al. 2008; Metulini et al. 2017; Sgrignoli et al. 2015).

- For large surveys, errors from the inference design can easily arise;
- The elements of an I/O matrix change over time. However, collecting them in a timely fashion (for instance, once for each country every year) is an almost impossible task, due to limited resources. For this reason, historical data are typically used to approximate the current I/O matrix.

Direct compilation of I/O tables does not rely only on surveys but also on administrative registers (e.g., firms' registers, accountancies of companies in official registers, public budget documents, information on trade customs, taxes authorities' reports, households' registers), censuses (population, firms' directories) and, among others, sectorial reports providing global economic information about specific sectors in an industry or in a country. It follows that, in order to compile an I/O table, many different statistical sources are contrasted (sometimes even contradictory ones). This is a huge time-consuming and resources-consuming operation, and the time gap between the publication of an I/O table and its reference year is one of the main reasons why also indirect methods exist. This was recently reviewed by Valderas-Jaramillo et al. (2019).

In this work, taking advantage from the similarity patterns discussed above, we apply, as an indirect method, matrix completion (MC, Hastie et al. 2015; Mazumder et al. 2010; Negahban and Wainwright 2012) to I/O submatrices associated with suitable groups of countries, by judiciously clustering them together, via a proper clustering method. We recall here that MC refers to a set of advanced statistical methods that can be used to predict unobserved entries of a matrix in terms of the set of the remaining observed entries (more technical details about MC are reported later in Sect. 2.2). In this way, the partially observed matrix is "completed" by the predictions produced by MC. Our choice of MC as an imputation method is based on the growing interest in using such method in the literature on economics (see Athey and Imbens 2019), and on the excellent performance achieved by MC in several applications (see, e.g., Hastie et al. 2015). We show that using a selected group of similar countries permits to increase the effectiveness of MC. This is done by comparing the results obtained by MC when, on the contrary, the selected group is made by highly dissimilar countries. To the best of the authors' knowledge, this is the first article in which MC is applied to I/O tables in connection with clustering.

The aforementioned similarity patterns have important consequences on the structure of the I/O matrices, justifying the application of MC. In fact, due to the presence of countries that share similar technologies for producing the same goods, an I/O matrix might be low-rank, or, at least, might be well-approximated by a low-rank matrix (in the sense that a few singular values would dominate all the other ones, i.e., the singular values' distribution would decay quickly to 0). This low-rank (approximation) property suggests, among other possible statistical or machine learning techniques, the adoption of MC to reconstruct potentially missing entries in an I/O matrix. Moreover, satisfying such a property is a necessary condition for obtaining good MC results (see Appendix 1 for a discussion on this issue). Nevertheless, the application of MC to a full I/O matrix is not straightforward, since elements in different blocks of that matrix can have quite different orders of magnitude. Having analyzed different real-world I/O matrices, we found that: (1) within-country

values are way larger than cross-country ones; (2) I/O matrices are sparse, because of many cross-country zero entries, and (3) there is a clear separation between large-to-large countries' values and small-to-small countries' ones. This suggests performing a pre-processing step, in which some blocks are removed from the full I/O matrices. In particular, we focus our analysis on bilateral trade blocks (Dietzenbacher et al. 2013; Arto et al. 2019). Their investigation constitutes a relevant problem in multi-country and multi-regional I/O tables.

Based on cross-country subsets of real-world I/O matrices, we performed a panel analysis using MC based on a LASSO-like nuclear norm penalty (Mazumder et al. 2010; Negahban and Wainwright 2012). This permits, through a suitable choice of the regularization parameter (based on a validation set), to select the number of nonzero singular values to be kept in the reconstructed matrix. The specific selection of countries was generated by the output of hierarchical clustering, whose application was based on a dissimilarity measure (the Average Absolute Correlation Distance or AACD, see later) highly related to the successive application of MC. Robustness of the results produced by hierarchical clustering was evaluated by considering synthetic counterparts of real-world I/O matrices, presenting a structure that is common to many I/O tables. This was done also in order to generalize the approach to all kinds of I/O matrices.

In summary, the main goal of our analysis is to compare the performance of MC when it is applied to I/O subtables made, respectively, by similar or dissimilar blocks (the first case corresponding to a subset of countries belonging to the same cluster, the second one to a subset of countries belonging to different clusters). It is worth remarking that our specific choice of the dissimilarity measure has been guided by its relationship with the regularization term in the objective function of the optimization problem associated with MC.

In more detail, this paper proposes a two-step methodological approach where, in the first step, judiciously selected groups of countries—in terms of either i) the distribution by country–industry from where they buy inputs, or ii) the distribution by country–industry to which they sell outputs—are retrieved using a hierarchical cluster analysis (Revelle 1979). This choice is preferable to other clustering techniques (e.g., *k*-means clustering MacQueen 1967) to group countries according to their I/O exchanges where the set of countries themselves is not a priori partitioned into a certain number of groups, but their complex structure suggests a hierarchy of clusters (see also Sect. 2.2 for other technical motivations behind this choice). In the second step, an I/O submatrix associated with this selection of countries is analyzed based on a LASSO-like formulation of the MC optimization problem. In particular, our approach is applied to a 5-year panel of (cross-country subsets of) I/O matrices for subsets of countries selected after performing hierarchical clustering, where a known part of the matrix associated with a specific year (i.e., the latest one) has been artificially obscured. The validation/testing phase is based on the Root Mean Square Error (RMSE) and on the Symmetric Mean Absolute Percentage Error (SMAPE) between actual and estimated values of the obscured part of the matrix, which is not provided as input to the MC algorithm. Ad hoc analyses have been also performed in order to i) evaluate the improvement that could be achieved by a suitable pre-processing of raw data (by eliminating domestic blocks); ii) select the

proper number of clusters; iii) evaluate the MC performance if the clustering failed in the first step. According to the latter point, we apply MC to subsets of either similar (cluster does not fail) or dissimilar (cluster fails) countries.

Results show the effectiveness of the proposed method to predict missing values in the current I/O matrix from both previous years' data and current data related to countries similar to the one for which current data are obscured. In contrast, the effectiveness reduces, as expected, if similar countries are replaced by ones belonging to quite different clusters. This conclusion holds both for the real-world and the synthetic data examined.

The rest of the manuscript begins with presenting the proposed methodological approach in Sect. 2 and reports the specific application to WIOD and simulated matrices and its results in Sect. 3. Section 4 is dedicated to future research directions and conclusions. Further technical details are reported in the appendix.

2 Methods

2.1 The input–output model

The traditional I/O matrix (Leontief 1986) depicts inter-industry relationships within an economy (or country), showing how the output from one sector becomes an input to another sector (or to itself), or it contributes to the final demand. Row indices represent inputs (in nominal monetary values) from an industrial sector, while column indices represent intermediate outputs to a given sector or needed to produce a final output. This table shows how dependent each sector is on every other sector, both as a customer of outputs from other sectors, and as a supplier of inputs.

Suppose an economy with n sectors and l final outputs is given, the assumption of constant returns to scale can be made, and sectors use inputs in fixed proportions. Fix also a specific year. In that year, each sector i produces a monetary value x_i of good i. Let $z_{i,j}$ be the value that sector i sells to sector j in that year, and let $f_{i,j}$ be the value that sector i sells to the final user in that year, to produce the final output j. In matrix notation, if one lets

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}, \ \mathbf{Z} = \begin{bmatrix} z_{1,1} & \cdots & z_{1,n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ z_{n,1} & \cdots & z_{n,n} \end{bmatrix}, \ \mathbf{F} = \begin{bmatrix} f_{1,1} & \cdots & f_{1,l} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ f_{n,1} & \cdots & f_{n,l} \end{bmatrix}, \tag{1}$$

one can write $\mathbf{x} = \mathbf{Z}\mathbf{i}_n + \mathbf{F}\mathbf{i}_l$, where $\mathbf{i}_n \in \mathbb{R}^{n\times 1}$ and $\mathbf{i}_l \in \mathbb{R}^{l\times 1}$ are column vectors made of all ones.

Available I/O tables can also report the multi-national structure of intra- and inter-industries (products) exchanges. In this case, let m be the number of countries considered. Then, analogously as in Eq. (1), one sets

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}^m \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{Z}^{1,1} & \cdot & \cdot & \cdot & \mathbf{Z}^{1,m} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \mathbf{Z}^{h,k} & & \cdot \\ \cdot & & & \cdot & \cdot \\ \mathbf{Z}^{m,1} & \cdot & \cdot & \cdot & \mathbf{Z}^{m,m} \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} \mathbf{F}^{1,1} & \cdot & \cdot & \cdot & \mathbf{F}^{1,m} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \mathbf{F}^{h,k} & & \cdot \\ \cdot & & & \cdot & \cdot \\ \mathbf{F}^{m,1} & \cdot & \cdot & \cdot & \mathbf{F}^{m,m} \end{bmatrix}, \tag{2}$$

where the generic (column) block $\mathbf{x}^m$ of the vector $\mathbf{x}$ in Eq. (2) can be expressed as

$$\mathbf{x}^m = \begin{bmatrix} x_1^m \\ \cdot \\ \cdot \\ \cdot \\ x_n^m \end{bmatrix}. \tag{3}$$

In this extended framework, one can analogously write $\mathbf{x} = \mathbf{Z}\mathbf{i}_{nm} + \mathbf{F}\mathbf{i}_{lm}$, where $\mathbf{i}_{nm} \in \mathbb{R}^{nm\times 1}$ and $\mathbf{i}_{lm} \in \mathbb{R}^{lm\times 1}$ are column vectors made of all ones and $\mathbf{x}$, $\mathbf{Z}$ and $\mathbf{F}$ are those in Eq. (2).

In the above, the generic block $\mathbf{Z}^{h,k}$ of the matrix $\mathbf{Z}$ in Eq. (2) represents the I/O subtable where h is the country in input and k is the country in output. Such a block can be expressed as

$$\mathbf{Z}^{h,k} = \begin{bmatrix} z_{1,1}^{h,k} & \cdots & z_{1,n}^{h,k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ z_{n,1}^{h,k} & \cdots & z_{n,n}^{h,k} \end{bmatrix}, \tag{4}$$

whereas the generic block $\mathbf{F}^{h,k}$ of the matrix $\mathbf{F}$ in Eq. (2) can be expressed as

$$\mathbf{F}^{h,k} = \begin{bmatrix} f_{1,1}^{h,k} & \cdots & f_{1,l}^{h,k} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ f_{n,1}^{h,k} & \cdots & f_{n,l}^{h,k} \end{bmatrix}. \tag{5}$$

The "transition" matrix $\mathbf{T} = [\mathbf{Z}|\mathbf{F}]$ is obtained from Eq. (2). Here, $\mathbf{T} \in \mathbb{R}^{mn\times m(n+l)}$ is a matrix whose row and column indices refer to an ordered "country, intermediate/final output" pair.

Since hierarchical clustering will be applied—depending on either i input or ii output criteria—to specific I/O submatrices, it is worth specifying two types of submatrices of $\mathbf{T}$ obtained by combining several blocks of the form $\mathbf{T}^{h,k} = [\mathbf{Z}^{h,k}|\mathbf{F}^{h,k}]$. As an example, for $h \neq 1, m$, let

$$\mathbf{T}^{h,\cdot} = \left[\mathbf{T}^{h,1}| \cdots |\mathbf{T}^{h,h-1}|\mathbf{T}^{h,h+1}| \cdots |\mathbf{T}^{h,m}\right] \tag{6}$$

be the transition submatrix related to what all sectors of country h sell to all sectors/final users of all countries with the exception of country h; and similarly, for $k \neq 1, m$, let

$$\mathbf{T}^{\cdot,k} = \begin{bmatrix} \mathbf{T}^{1,k} \\ \cdots \\ \mathbf{T}^{k-1,k} \\ \mathbf{T}^{k+1,k} \\ \cdots \\ \mathbf{T}^{m,k} \end{bmatrix} \tag{7}$$

be the transition submatrix related to what all sectors/final users of country k buy from all sectors of all countries with the exception of country k. Similar definitions obviously hold for $h, k \in \{1, m\}$. The submatrices $\mathbf{T}^{h,\cdot}$ have n rows and $(m-1)(n+l)$ columns, whereas the submatrices $\mathbf{T}^{\cdot,k}$ have $(m-1)n$ rows and $n+l$ columns. About the first kind of submatrix (criterion i of clustering), for example, the first block $\mathbf{T}^{h,1}$ expresses what all sectors/final users of country 1 buy from all sectors of country h (say, Italy). About the second kind of submatrix (criterion ii of clustering), for example, the first block $\mathbf{T}^{1,k}$ expresses what all sectors of country 1 sell to all sectors/final users of country k (say, Italy).

2.2 Clustering and matrix completion

It is recalled here that clustering is an unsupervised learning technique whose goal consists in partitioning a data set into several subsets (called clusters), by aiming to make more similar data points belong to the same cluster, while trying to assign less similar data points to distinct clusters[2]Aggarwal and Reddy (2014). Among several clustering methods, we choose a hierarchical method because, in I/O tables, the set of countries is not a priori partitioned into a certain number of groups and because their complex structure suggests a hierarchy of clusters. By using a set of pair-wise dissimilarities for m objects (i.e., countries), hierarchical clustering first assigns each object to its own cluster, then it proceeds iteratively by joining at each stage the two most similar clusters, continuing until there is just a single cluster. At each stage, distances between clusters are recomputed by the Lance–Williams formula (Lance and Williams 1967), based either on the complete or on the Ward linkage criterion

[2] Clustering techniques were recently applied to I/O matrices. For instance, Oliva et al. (2016) adopted a hierarchical clustering approach, based on the recursive application of spectral clustering, which is a clustering technique suitable for data organized according to a graph structure. The analysis was done at the sector level, for the WIOD submatrix obtained by keeping only rows and columns related to Italy, for the period 1995–2011. As another example, Zhu et al. (2018) proposed a network-based measure of similarity (and the resulting unweighted average distance clustering), to compare the (upstream and downstream) Global Value Chains (GVCs) between any pair of countries, for each sector and each year available at the time in the WIOD database. It is worth mentioning that also global and local spatial clustering techniques have been developed in the literature to, respectively, test whether a clustering structure is present in the analyzed region and to identify the locations of clusters (see, e.g., the exhaustive review provided by Aldstadt 2010).

(Murtagh and Legendre 2014). According to hierarchical clustering, differently from (e.g.) the k-means clustering method, one can achieve different partitions of objects depending on the level of resolution one is looking at. Moreover, despite k-means clustering is less computationally expensive compared to hierarchical clustering, it requires strict assumptions regarding the homoscedasticity and the spherical variance of the variables, and that each cluster has—a priori—roughly an equal number of objects. Hierarchical clustering performs well even when those assumptions are not satisfied.

To measure how (dis)similar any two objects are, several (dis)similarity measures have been developed in the literature. Among them, the most commonly used are the l_1 norm of the difference of data points (also called Manhattan distance), and the l_2 norm of their difference (also called Euclidean distance). In this work, we use the Average Absolute Correlation Distance (AACD) as a dissimilarity measure for clustering. In other words, the absolute value of the Pearson's correlation coefficient between the j-th corresponding non-constant columns $\mathbf{b}_j^{c_1}$ and $\mathbf{b}_j^{c_2}$ of blocks[3] associated with two different countries c_1 and c_2 is evaluated, then it is averaged with respect to the columns, and subtracted from 1. In formulas, one has

$$\mathrm{AACD}_{c_1,c_2} = 1 - \frac{\sum_{j=1}^{n+l} |\mathrm{corr}(\mathbf{b}_j^{c_1}, \mathbf{b}_j^{c_2})|}{n+l}. \tag{8}$$

Our choice of the dissimilarity measure is motivated by the fact that AACD is highly related to the specific formulation of the regularization term in the optimization problem modeling MC, which is reported later in this subsection. Indeed, in a sense, it quantifies the average linear dependence of corresponding columns of blocks associated with different countries.[4] The adoption of this dissimilarity measure also (1) provides an additional motivation for the application of hierarchical clustering instead of a different clustering technique, since such a distance does not satisfy the triangle inequality, which is not required by hierarchical clustering but is required, e.g., by k-means clustering; (2) operates as a data pre-processing step because, differently from the l_1 and l_2 norms, it is not affected by each country's average dimension. It is worth remarking that, even if countries in I/O tables present a spatial dimension, we are not going to use any kind of spatial clustering technique, due to the following reasons: (1) in this work, the clustering step aims at selecting a subset of similar countries, in terms of a suitable dissimilarity measure (AACD) derived from the comparison of different portions of I/O subtables, in order to make the successive application of MC to I/O subtables easier. In other words, we are not

[3] The specific structure of the blocks considered is reported, e.g., in Table 5 for the case of inputs from Italy, and in Table 7 for the case of outputs from Italy. Each column $\mathbf{b}_j^c$ is composed of all the data related to a specific country c (different from Italy) and the intermediate/final output j in the years 2010–2013. The last year (2014) is not considered for the computation of the correlation because parts of its data refer to the validation/test set related to the successive MC application.

[4] In a first version of our analysis, the l_2 norm of the difference of data points associated with different countries was adopted as dissimilarity measure. Nevertheless, it was verified numerically that AACD produces better clustering results for what concerns the successive application of MC. This is likely due to the fact that the l_2 norm evaluates as dissimilar two highly correlated blocks whose entries have different sizes. The same remark holds for the l_1 norm.

interested specifically in finding the presence of a spatial pattern; (2) volumes of I/O exchanges across countries are determined, possibly to different extents, both by network relationships and by spatial contiguity. Both of them are taken indirectly into account by using data coming from I/O subtables to express the dissimilarity measure.

In this work, we apply hierarchical clustering to all countries in the I/O matrix with the exception of a specific country h, in terms of what they use from country h. Similarly, we also apply hierarchical clustering to all countries in the I/O matrix with the exception of a specific country k, in terms of what they sell to country k. In other words, we consider, for various years, the submatrices $\mathbf{T}^{h,\cdot}$ and $\mathbf{T}^{\cdot,k}$ defined in Sect. 2.1 to compare any two countries c_1 and c_2 (different, respectively, from h and k) in terms of what they use from (or what the sell to) the specific country h or k. For illustrative purposes, in the simulations reported in the application, the choice $h = k$ is made, and Italy is selected as such a specific country.

Clustering is often used as a preliminary data pre-processing step to a successive supervised learning task. In the present context, clustering is used as a pre-processing step for MC applied to a suitable submatrix of an I/O table.[5] The idea is that MC is expected to perform better if the submatrix refers to countries belonging to the same cluster. This expectation is based on one of the reasons provided in the literature as a motivation for the effectiveness of MC (Hastie et al. 2015), which is summarized as follows. Given a subset of observed entries of a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, MC works by finding a suitable low-rank approximation (say, with rank r) of $\mathbf{M}$, by assuming the following model:

$$\mathbf{M} = \mathbf{C}\mathbf{G}^T + \mathbf{E}\,, \tag{9}$$

where $\mathbf{C} \in \mathbb{R}^{m \times r}$, $\mathbf{G} \in \mathbb{R}^{n \times r}$, whereas $\mathbf{E} \in \mathbb{R}^{m \times n}$ is a matrix of modeling errors. The rank-r approximating matrix $\mathbf{C}\mathbf{G}^T$ is found by solving a suitable optimization problem (see, e.g., Eq. (11) reported later). Equation (9) can be written element-wise as

$$M_{i,j} = \sum_{l=1}^{r} C_{i,l} G_{j,l} + E_{i,j}\,. \tag{10}$$

Often, $C_{i,l}$ is interpreted as the degree of membership of row i of matrix $\mathbf{M}$ to some "latent" cluster l (for a total of r such clusters), and $G_{j,l}$ as the prediction of an element in column j of matrix $\mathbf{M}$, conditioned on its row i belonging to the l-th cluster.[6]

[5] Up to the authors' knowledge, MC was applied in the literature to I/O matrices only in two works. In more detail, (Wen et al. 2014) used it as a pre-processing step for robust linear optimization, for a problem whose coefficient matrix is a partially observable I/O table, whereas (Xu et al. 2014) applied MC to predict zero entries in an I/O matrix, based on a set of observed entries mainly made by zeros. Another interesting application of MC to networks (though not related to I/O matrices) was made by Nguyen et al. (2019), who used this method to recover sensor maps in 2-D or 3-D Euclidean spaces from local or in any case partial sets of pair-wise distances.

[6] It is worth mentioning that estimates $\hat{\mathbf{C}}$ and $\hat{\mathbf{G}}$ of the matrices $\mathbf{C}$ and $\mathbf{G}$ can be obtained as a by-product of the singular value decomposition $\hat{\mathbf{M}} = \mathbf{U}(\hat{\mathbf{M}})\mathbf{\Sigma}(\hat{\mathbf{M}})\Big(\mathbf{V}(\hat{\mathbf{M}})\Big)^T$ of the matrix $\hat{\mathbf{M}}$ produced as output by an MC algorithm (e.g., one can set $\hat{\mathbf{C}} := \mathbf{U}(\hat{\mathbf{M}})$ and $\hat{\mathbf{G}} := \mathbf{V}(\hat{\mathbf{M}})\Big(\mathbf{\Sigma}(\hat{\mathbf{M}})\Big)^T$).

In our application, $\mathbf{M}$ is composed of several cross-country blocks (coming from I/O tables in different years), whereas i and j refer, respectively, to an input sector of a country and an output sector/final user of another country. Moreover, l may be interpreted as a specific "latent" cluster, possibly discovered by the MC algorithm.

It is worth observing that, in order for MC to work properly in the case of an I/O table (possibly partially observed for a set of consecutive years), it can be useful to apply it to its suitable submatrix determined by a pre-preprocessing step of clustering. Intuitively, missing blocks of an I/O table that, thanks to historical data in the past years, are expected to be similar to the other observed blocks in the current year, could be reconstructed more effectively than missing blocks in the current year that, again based on historical data in the past years, are expected to be less similar to the other observed blocks in the same year. Another reason to focus the analysis on a submatrix of an I/O table is that solving the MC problem becomes computationally more expensive as the size of the matrix $\mathbf{M}$ increases.

In the work, we consider the following formulation for the MC optimization problem, which was investigated theoretically by Mazumder et al. (2010):

$$\underset{\hat{\mathbf{M}}\in\mathbb{R}^{m\times n}}{\text{minimize}}\left(\frac{1}{2}\sum_{(i,j)\in\Omega^{\text{tr}}}\left(M_{i,j}-\hat{M}_{i,j}\right)^2+\lambda\|\hat{\mathbf{M}}\|_*\right), \tag{11}$$

where Ω^{tr} (which, using a machine-learning expression, may be called training set) is a subset of pairs of indices (i, j) corresponding to positions of known entries of $\mathbf{M}$, $\hat{\mathbf{M}}$ is the completed matrix (to be optimized), $\lambda \geq 0$ is a regularization constant, and $\|\hat{\mathbf{M}}\|_*$ is the nuclear norm of the matrix $\hat{\mathbf{M}}$, i.e., the sum of all its singular values. The regularization constant λ controls the trade-off between fitting the known entries of the matrix $\mathbf{M}$ and achieving a small nuclear norm. The latter requirement is often related to getting a small rank of the obtained optimal solution[7] $\hat{\mathbf{M}}^{\circ}_{\lambda}$ of the optimization problem (11), which follows by geometric arguments similar to the ones typically adopted to justify how the classical LASSO (Least Absolute Shrinkage and Selection Operator) penalty term achieves effective feature selection in linear regression (Tibshirani 1996).

The optimization problem (11) can be also written as

$$\underset{\hat{\mathbf{M}}\in\mathbb{R}^{m\times n}}{\text{minimize}}\left(\frac{1}{2}\|\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{M})-\mathbf{P}_{\Omega^{\text{tr}}}(\hat{\mathbf{M}})\|_F^2+\lambda\|\hat{\mathbf{M}}\|_*\right), \tag{12}$$

where, for a matrix $\mathbf{Y}\in\mathbb{R}^{m\times n}$,

$$\left(P_{\Omega^{\text{tr}}}(\mathbf{Y})\right)_{i,j} := \begin{cases} Y_{i,j} & \text{if } (i,j)\in\Omega^{\text{tr}}, \\ 0 & \text{if } (i,j)\notin\Omega^{\text{tr}} \end{cases} \tag{13}$$

represents the projection of $\mathbf{Y}$ onto the set of positions of observed entries of the matrix $\mathbf{M}$, and $\|\mathbf{Y}\|_F$ denotes the Frobenius norm of $\mathbf{Y}$ (i.e., the square root of the summation of squares of all its entries).

[7] In the notation, we have highlighted the dependence of that optimal solution on λ.

It was shown by Mazumder et al. (2010) that the optimization problem (12) can be solved by applying Algorithm 1, named Soft Impute therein.[8] This is a state-of-the-art algorithm in the MC field.

In Algorithm 1, for a matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$, $\mathbf{P}^{\perp}_{\Omega^{\mathrm{tr}}}(\mathbf{Y})$ represents the projection of $\mathbf{Y}$ onto the complement of Ω^{tr}, whereas

$$\mathbf{S}_{\lambda}(\mathbf{Y}) := \mathbf{U}\mathbf{\Sigma}_{\lambda}\mathbf{V}^{T}, \tag{14}$$

being

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{T} \tag{15}$$

(with $\mathbf{\Sigma} = \mathrm{diag}[\sigma_1, \ldots, \sigma_r]$) the singular value decomposition of $\mathbf{Y}$, and

$$\mathbf{\Sigma}_{\lambda} := \mathrm{diag}[(\sigma_1 - \lambda)_+, \ldots, (\sigma_r - \lambda)_+], \tag{16}$$

with $t_+ := \max(t, 0)$.

It is worth mentioning that (Li and Zhou 2017) proposed a particularly efficient implementation of the operator $\mathbf{S}_{\lambda}(\cdot)$ defined in Eq. (14) (by means of the MATLAB function `svt.m` reported therein), which is based on the determination of only the singular values σ_i of $\mathbf{Y}$ that are larger than λ, and of their corresponding left-singular vectors $\mathbf{u}_i$ and right-singular vectors $\mathbf{v}_i$. Indeed, all the other singular values of $\mathbf{Y}$ are annihilated in $\mathbf{\Sigma}_{\lambda}$ (see Eq. 16).

Algorithm 1: Soft Impute [21]

Input: Partially observed matrix $\mathbf{P}_{\Omega^{\mathrm{tr}}}(\mathbf{M})$, regularization constant $\lambda \geq 0$, tolerance $tol \geq 0$, maximal number of iterations N^{it}

Output: Completed matrix $\hat{\mathbf{M}}_{\lambda} \in \mathbb{R}^{m \times n}$

1. Initialize $\hat{\mathbf{M}}$ as $\hat{\mathbf{M}}^{\mathrm{old}} = \mathbf{0} \in \mathbb{R}^{m \times n}$
2. Repeat for at most N^{it} iterations:
   (a) Set $\hat{\mathbf{M}}^{\mathrm{new}} \leftarrow \mathbf{S}_{\lambda}\left(\mathbf{P}_{\Omega^{\mathrm{tr}}}(\mathbf{M}) + \mathbf{P}^{\perp}_{\Omega^{\mathrm{tr}}}(\hat{\mathbf{M}}^{\mathrm{old}})\right)$
   (b) If $\frac{\|\hat{\mathbf{M}}^{\mathrm{new}} - \hat{\mathbf{M}}^{\mathrm{old}}\|_F^2}{\|\hat{\mathbf{M}}^{\mathrm{old}}\|_F^2} \leq tol$, exit
   (c) Set $\hat{\mathbf{M}}^{\mathrm{old}} \leftarrow \hat{\mathbf{M}}^{\mathrm{new}}$
3. Set $\hat{\mathbf{M}}_{\lambda} \leftarrow \hat{\mathbf{M}}^{\mathrm{new}}$

For this work, we combine the original MATLAB implementation of Soft Impute provided by Mazumder et al. (2010) with the MATLAB function `svt.m` developed by Li and Zhou (2017). Moreover, to avoid overfitting, we select the regularization constant λ via the following hold-out validation method. First, the set of positions of unobserved entries of the matrix $\mathbf{M}$ is divided randomly into a validation set Ω^{val} (about 25% of the positions of the unobserved entries) and a test set Ω^{test} (the positions of the remaining entries). In the present context of application of MC to I/O subtables, the union of the validation and test sets corresponds to a block which is artificially obscured (but which is still available as a ground truth), whereas the

[8] Compared to the original version, here we have included a maximal number of iterations N^{it}, which can be helpful to reduce the computational effort when one has to run the algorithm multiple times, e.g., for several values of the regularization constant λ.

training set corresponds to the positions of all the remaining entries of the submatrix considered. It is worth observing that, by the construction above, there is no overlap among the training, validation and test sets. Then, the optimization problem (12) is solved for several choices λ_k for λ, exponentially distributed as $\lambda_k = 2^{k/2-10}$, for $k = 1, \ldots, 40$. For each λ_k, the Root Mean Square Error (RMSE) of matrix reconstruction on the validation set is computed as

$$\mathrm{RMSE}^{\mathrm{val}}_{\lambda_k} := \sqrt{\frac{1}{|\Omega^{\mathrm{val}}|} \sum_{(i,j)\in\Omega^{\mathrm{val}}} \left(M_{i,j} - \hat{M}_{\lambda_k,i,j}\right)^2}, \tag{17}$$

then the choice λ_k° that minimizes $RMSE^{\mathrm{val}}_{\lambda_k}$ for $k = 1, \ldots, 40$ is found.[9] Finally, the RMSE of matrix reconstruction on the test set is computed in correspondence of the optimal value λ_k° as

$$\mathrm{RMSE}^{\mathrm{test}}_{\lambda_k^\circ} := \sqrt{\frac{1}{|\Omega^{\mathrm{test}}|} \sum_{(i,j)\in\Omega^{\mathrm{test}}} \left(M_{i,j} - \hat{M}_{\lambda_k^\circ,i,j}\right)^2}. \tag{18}$$

A similar expression holds for the RMSE of matrix reconstruction on the training set, in correspondence of the optimal value λ_k°:

$$\mathrm{RMSE}^{\mathrm{tr}}_{\lambda_k^\circ} := \sqrt{\frac{1}{|\Omega^{\mathrm{tr}}|} \sum_{(i,j)\in\Omega^{\mathrm{tr}}} \left(M_{i,j} - \hat{M}_{\lambda_k^\circ,i,j}\right)^2}. \tag{19}$$

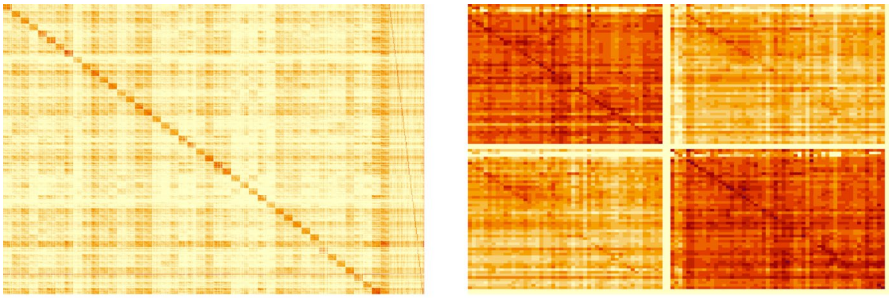
In our application of MC to submatrices of I/O tables, as a pre-processing step, the very few missing/negative entries of such submatrices (when present) are replaced by zeros before running Algorithm 1. The tolerance is chosen as $tol = 10^{-9}$. Moreover, when convergence is not achieved, in order to reduce the computational time, the algorithm is stopped after $N^{\mathrm{it}} = 500$ iterations. An additional post-processing step is included, thresholding to 0 any negative element (when present) of the completed submatrices.[10] In the following, in order to avoid introducing new notation, the expression $\hat{\mathbf{M}}_{\lambda_k}$ is actually used to denote each post-processed MC output.

[9] As an alternative method to define the optimal value for the regularization parameter, one could take into account different validation sets, then applying the one standard error rule (see Hastie et al. 2015). This is typically done by considering also different training sets (as in cross-validation), making the MC application more computationally intensive.

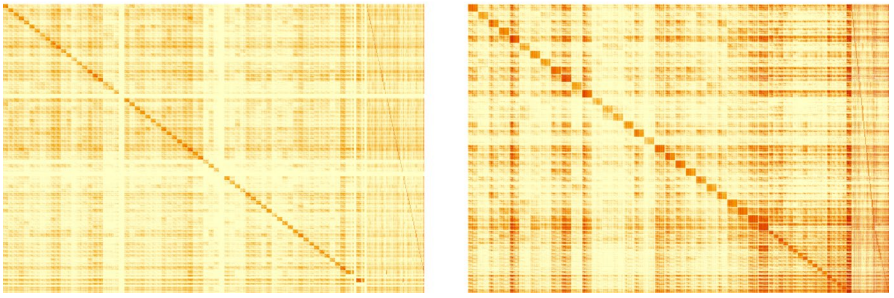
[10] It is worth noticing that, without its pre-processing and post-processing steps, the method described here is applicable also in the case of matrices having both positive and negative elements (see, e.g., the United Nations handbook on supply and use tables and input–output tables (United Nations 2018) for some motivations behind the possible existence of negative entries in real-world I/O tables, beside the presence of positive entries). Nevertheless, in our specific application, the presence of a very small percentage of negative entries translates approximately into an a priori knowledge about non-negativity of the I/O table. Hence, the aim of the two steps introduced is to possibly improve the MC predictions, based on such a knowledge. By the way, we have found in all our numerical results that improved MC predictions were obtained in this way.

Table 1 Percentage of zero values in the I/O tables, by year

	Year				
Source	2010	2011	2012	2013	2014
WIOD	17.904%	17.924%	17.788%	17.854%	17.821%
OECD	21.480%	21.195%	21.440%	21.354%	21.230%
FIGARO	39.293%	38.408%	37.244%	37.563%	37.270%



(a) WIOD (left). Zoom of the WIOD submatrix composed of rows and columns of Italy and Spain (right).



(b) OECD (left). FIGARO (right).

Fig. 1 Colored visualization of the elements of the complete I/O tables (year: 2013). For a better visualization, a logarithmic scale is used in the subfigures

Finally, as a measure of the MC performance, we also use a second metric, which is known in the literature as Symmetric Mean Absolute Percentage Error (SMAPE).[11] Differently from the RMSE, it takes into account the relative error of reconstruction. Its definition for the validation set is as follows:

$$\text{SMAPE}^{\text{val}}_{\lambda_k} := \frac{100}{|\Omega^{\text{val}}|} \sum_{(i,j)\in\Omega^{\text{val}}} \frac{\left|M_{i,j} - \hat{M}_{\lambda_k,i,j}\right|}{\left|M_{i,j}\right| + \left|\hat{M}_{\lambda_k,i,j}\right|} \tag{20}$$

[11] As an alternative, one could use a more recent weighted version of the SMAPE, known as SWAPE (Valderas-Jaramillo et al. 2019).

(the constant 100 is used[12] to make the metric range from 0 to 100; when both the numerator and the denominator are equal to 0, the ratio is assumed to be equal to 0, too). Similar definitions hold for the test set and the training set. Again, the metric is first evaluated on the validation set for different choices of λ_k, then it is computed on both the training and test sets in correspondence of the value of λ_k that minimizes the SMAPE on the validation set. Differently from the RMSE, this metric is not directly related to the optimization problem (11) solved by MC, nor to the choice of AACD as the dissimilarity measure used by hierarchical clustering in the present article. Hence, for this metric, differently from the RMSE, one does not necessarily expect an improvement in MC performance when moving from "dissimilar" to "similar" countries.

3 Application of clustering and matrix completion

In this section, we present an application of the proposed methodological approach to WIOD data. Before presenting the application to real data, some aspects need to be considered. In order to generalize our method to any kind of I/O tables (whether they are either industry-by-industry or product-by-product ones), in Sect. 3.2 we show how the structure of WIOD tables—described in Sect. 3.1—is similar to those of alternative I/O tables. Then, in Sect. 3.3 some simulation results are reported, based on synthetic I/O matrices generated from the raw WIOD tables, in order to discuss the benefits resulting from applying MC to proper pre-processed data and to determine the optimal number of clusters for the choice of similar and dissimilar countries. Later, in Sects. 3.4 and 3.5 we provide full details on how we operatively apply the method to real data (specifically, Sect. 3.4 reports a simple example of MC application without its combination with the clustering step, whereas Sect. 3.5 includes the clustering step). Additional analyses are performed in the remaining subsections. In more details, Sect. 3.6 repeats the analysis of Sect. 3.5 for the case of a synthetic dataset, whereas Sect. 3.7 shows the dependence of the results obtained in Sect. 3.5 with respect to changes in the choices of the validation and test sets. Finally, Sect. 3.8 examines the MC performance when MC is applied to an I/O subtable containing both intra-country and inter-country blocks.

It is important to clarify why we use WIOD matrices in our application. According to Timmer et al. (2015), WIOD represents a real improvement over other existing databases (such as EORA, EXIObase and OECD tables) for several reasons: i) its data are extrapolated by certified national statistical institutions, ii) to determine data from the rest of the world, data from United Nations (UN), International Monetary Funds (IMF) and other international institutions are used, iii) all versions of WIOD are available for free from the website www.wiod.org. Compared with other I/O datasets, we choose WIOD tables for two additional reasons: first, because they are characterized by a quite large coverage period; second, because their size is quite representative of the ones of the other tables (i.e., both the number of countries

[12] In some references, it is replaced by the constant 200.

considered and the sector disaggregation are neither too small, nor too large with respect to the other tables). The latter issue is also important for computational reasons, because the application of MC is typically slow for large matrices.

3.1 Data

The WIOD database was constructed and developed in the seventh framework program funded by the European Commission in 2009, and is licensed under a Creative Commons Attribution 4.0 International-license. From a technical point of view, WIOD tables are built up from public databases coming from different national and international statistics' offices. Currently, there exist two releases of WIOD: the 2013 and the 2016 release. The latest release covers the period between 2000 and 2014 and 43 among the most relevant countries in the world: EU-28 (including the UK), Australia, Brazil, Canada, Switzerland, China, Indonesia, India, Japan, South Korea, Mexico, Norway, Russia, Turkey, Taiwan, USA.[13] The yearly tables are split into 56 different macro-industries, classified according to the International Standard Industrial Classification Revision 4 (ISIC Rev. 4), and their pair-wise combinations.[14] Moreover, 5 final aggregated outputs—still classified according to ISIC Rev. 4—are reported in the tables. Finally, an estimation for the remaining non-covered part of the world economy (called "Rest Of the World," ROW) is reported (details are provided by Timmer et al. 2015, 2016). Thereby, using WIOD can help to perform excellent and detailed input/output analyses (some very recent applications being provided, e.g., by Bhattacharya et al. 2020; Chen et al. 2019; Wang et al. 2020; Xu and Liang 2019).

In the following subsection, we put WIOD tables in comparison with OECD and FIGARO ones. The latest release of OECD inter-country input–output (ICIO) tables dates back to 2018. ICIO tables report yearly data from 2005 to 2015 among 64 countries (including ROW) and 36 industries (products). FIGARO tables, also known as EU inter-country Supply, Use and input–output tables (EU IC-SUIOTs) are available on a yearly basis from 2010 to 2019 and display exchanges among EU economies, the UK and the USA in 64 industries (products).

3.2 Characterization of the data

Here, we provide a descriptive analysis of WIOD tables in comparison with OECD and FIGARO ones (whose data are available in the same time span of WIOD tables) in terms of their within-country and cross-country value distributions, level of sparsity (i.e., percentage of zeros in each subtable) and separation between values of transactions between so-called large-to-large and small-to-small country pairs. These analyses are made on industry-by-industry tables, as the product-by-product ones depart from the former to just a little extent (Pearson's correlation coefficient

[13] In WIOD tables, countries are represented by ISO-3166-1 alpha-3 codes.

[14] In contrast, the 2013 release covers the period 1995–2011, considering only 40 countries and 38 macro-sectors.

Table 2 Distribution of the average exchanges from Italy (in input), by country of output (first row) and of the within-country standard deviation (second row)

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Averages by country	0.034	0.488	0.837	1.567	1.780	10.080
Within-country st. dev.	0.354	4.034	6.917	10.254	13.097	47.461

All countries excluding Italy. WIOD data, year: 2010

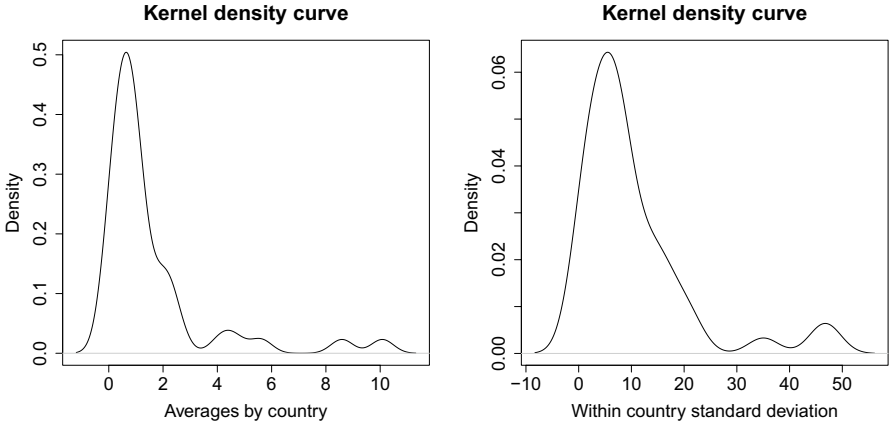


Fig. 2 Kernel density curves of the distribution of the average exchanges from Italy (in input), by country of output (`averages by country`) and of the within-country standard deviation (`within-country standard deviation`). All countries excluding Italy. WIOD data, year: 2010

Table 3 Simulation results for the choice of the optimal number of clusters

Direction	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Input	14	21	21	21.070	21	28
Output	16	22	22	21.971	22	28

$N = 1000$ replications of hierarchical clustering with complete linkage criterion and AACD as dissimilarity measure, on I/O synthetic matrices derived from WIOD original data. Adopted criterion: $\frac{\text{WSS}}{\text{TSS}}$. More precisely, we chose the minimum number of clusters such that $\frac{\text{WSS}}{\text{TSS}} < 0.5$. Considering Italy in input and all other countries in output (`input`), and considering Italy in output and all other countries in input (`output`). Years: from 2010 to 2013 (4 years)

equal to 0.9958 for FIGARO, year: 2010), and are reported for the years from 2010 to 2014, which is the time span considered in our panel analysis. After having removed ROW rows and columns, and also rows related to taxes and value added, WIOD tables count for 2408 rows (which is the product of 56 intermediate industries and 43 countries) and 2623 columns (which is the result of 56 intermediate industries plus 5 final outputs, multiplied by 43 countries). The 56 macro-industries (which are produced by the aggregation of various micro-industries) are reported in the following order: primary industry appears in the first 4 positions, followed

Table 4 Structure of the WIOD submatrix used for the example reported in Fig. 3a

I/O, year	
FRA/ITA, 2010	ITA/FRA, 2010
FRA/ITA, 2011	ITA/FRA, 2011
FRA/ITA, 2012	ITA/FRA, 2012
FRA/ITA, 2013	ITA/FRA, 2013
FRA/ITA, 2014	**ITA/FRA, 2014**

Each block is made of 56 rows for the inputs (corresponding to the 56 intermediate sectors considered by WIOD), 56 columns for the intermediate outputs (one for each intermediate sector) and other 5 columns for the final products. All the entries contained in the block highlighted in bold are obscured and reconstructed by MC

by secondary industry (18 positions), and finally by tertiary industry (34 positions). About the sparsity of WIOD I/O matrices, Table 1 shows that, consistently over the years, the percentage of zeros is between 17% and 18%. Moreover, results are consistent if compared to those of OECD tables while the number of zeros of FIGARO is slightly larger.

Figure 1 shows, as an example, a colored visualization of the elements of the 2013 I/O tables where each colored rectangle corresponds to the exchange between country–industry pairs. In its subfigures, final consumption is reported on the right extremes. The figure sheds lights on the fact that, consistently over the three compared I/O tables, the largest values (depicted in red) are concentrated in the domestic blocks (main diagonal blocks). Indeed, industries usually tend to consume (with respect to trade) products coming from their home country, for reasons such as higher proximity and safety (i.e., less uncertainty in terms of price, and more regularity in terms of supplies). Moreover, flows from a specific "country–industry" pair to the same "country–industry" pair (main diagonal of the tables) are generally much larger than the other flows (this holds especially for the case of secondary macro-sectors, as one can see from the right chart of Fig. 1a representing exchanges within and between Italy and Spain for the case of WIOD). This issue is partially motivated by outsourcing, but especially by the fact that in some industries (e.g., manufacturing industries) there are several concatenated products in the production line, which in the case of WIOD tables are aggregated in the same macro-sector. Moreover, it can be noticed from the figure that secondary industrial products (e.g., ore, iron, oil, metals, technical equipment, products from manufacturing industry in general) are more open to international trade (see the corresponding parts of the main diagonals of the off-diagonal blocks, also called international trade blocks), whereas services are less traded internationally. Finally, by looking indifferently to one of the I/O tables in Fig. 1 and by comparing either single country blocks by row (supplying countries) or single country blocks by column (receiving countries), it is possible to notice how some countries are similar to each other according to how they are dependent with respect to other specific countries.

Overall, real-world I/O tables provided by different institutions share very similar characteristics and motivate us to work just on one data source (WIOD tables, in our specific selection) and to generalize the results obtained using it over different I/O tables.

Fig. 3 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 4. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 4. **c** Singular values' distribution of the WIOD submatrix reported in Table 4, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion). **d** Colored visualization of the elements of the WIOD submatrix reported in Table 4, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (color figure online)

3.3 Clustering step and simulations

As discussed in Sect. 3.2, the main diagonal blocks in the I/O tables (also called domestic blocks, since each of them refers to trade inside the same country), and especially their entries which refer to exchanges within the same sector and the same country, are characterized by much larger values than the other blocks, which may lead to problems for an effective application of MC, as the quite different orders of magnitude could make it difficult for MC to have a good generalization capability on both kinds of blocks.[15] Computational reasons (i.e., the need of performing a singular value decomposition step at each iteration of the Soft Impute algorithm) suggest to apply MC to a submatrix associated with a small subset of countries, as this reduces the size of that submatrix. Moreover, as discussed in Sect. 2.2, we argue that MC is more effective when it is applied to an I/O subtable made of "similar" blocks (with respect to the case of "dissimilar" blocks). For this reason, we apply MC to submatrices of WIOD tables obtained by excluding systematically the main diagonal blocks,[16] where the selection of the countries associated with the submatrices is made by means of hierarchical cluster analysis. In this context, the choice of the number of clusters is crucial and so, in order to validate hierarchical clustering, we perform some simulation exercises, based on synthetic data.[17] We choose to generate our synthetic I/O matrices by adding a matrix of normally distributed random terms ϵ to the subset of interest of the WIOD dataset, where each element of the matrix $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma_\epsilon), \forall i, \forall j$, σ_ϵ being a Gamma(α, β) such that different generated synthetic matrices display different levels of variability. Specifically, we choose $\alpha = 1$ and $\beta = 1$.

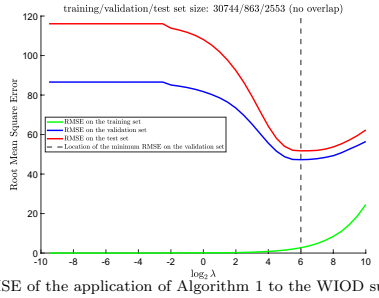
To find the correct number of groups in terms of similarity either with respect to input from Italy or with respect to output to Italy, we simulate:[18] (1) $N = 1000$ synthetic I/O matrices of dimension 56×2562, where 56 is the number of industries in Italy, and 2562 is the product of the 61 industries (final sectors included) and the

[15] As shown later in Sect. 3.8, indeed, the performance of MC on an I/O submatrix which considers simultaneously the blocks of intra-country and inter-country exchanges (whose orders of magnitude are highly different) is quite bad.

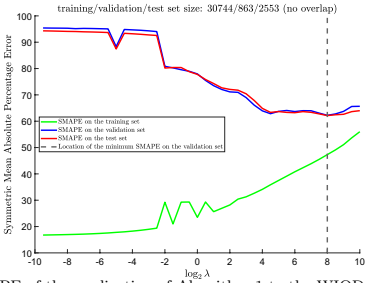
[16] This is not necessarily a limitation of the proposed method. Indeed, as an alternative, one could exclude systematically the off-diagonal blocks (this is left for future research). The important issue is that the blocks kept have entries with similar orders of magnitude.

[17] Methods for generating synthetic I/O matrices were investigated, e.g., by Wang et al. (2015), which used a cubic polynomial with coefficients generated from a standardized normal distribution. Pavia et al. (2009) added either a normally or a uniformly distributed disturbance term to six heterogeneous origin–destination matrices, whereas (Fernandez-Vazquez 2016) also added random terms to the elements of the I/O tables.

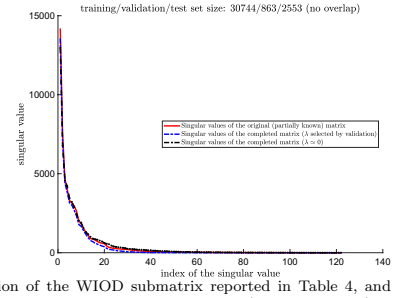
[18] It is worth clarifying that, while the generalization of the simulation results to other I/O tables holds, the generalization to other reference countries different from Italy does not necessarily hold.



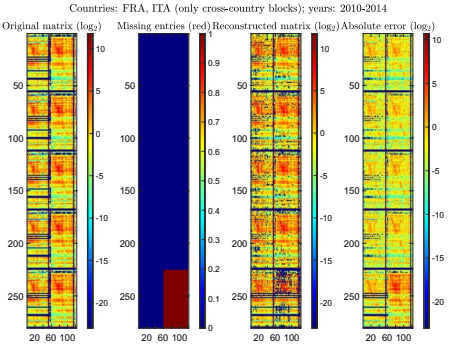
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 4.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 4.



(c) Singular values' distribution of the WIOD submatrix reported in Table 4, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion).



(d) Colored visualization of the elements of the WIOD submatrix reported in Table 4, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error.

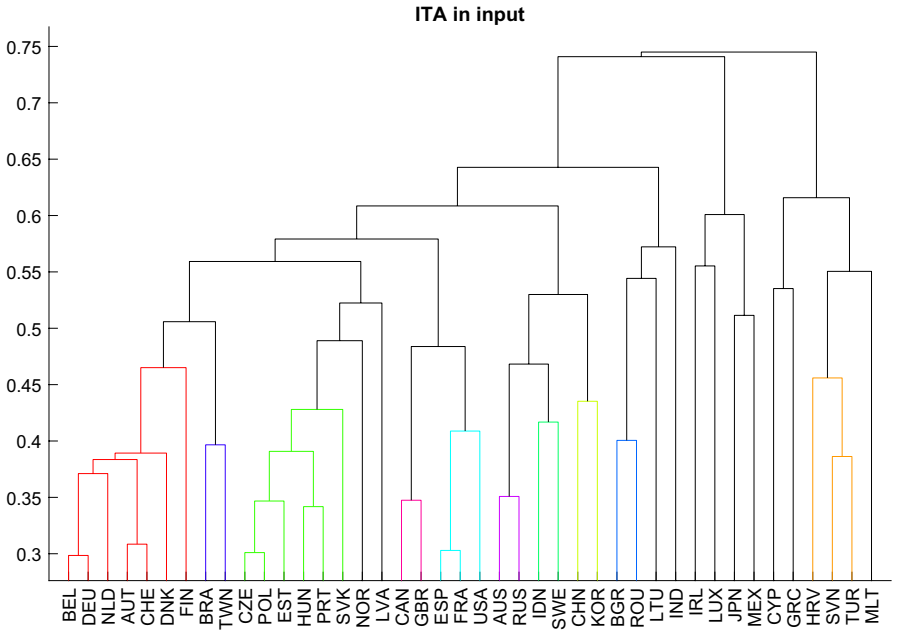


Fig. 4 Dendrograms of output countries with Italian sectors in input, based on WIOD tables (stacked over the years 2010–2013). Hierarchical clustering performed with the AACD dissimilarity measure (y-axis) and complete linkage. 21 desired groups (compare with Table 3). Countries in the same cluster are depicted with the same color. Countries in singleton clusters are highlighted in black

number of countries with the exception of Italy, (2) $N = 1000$ synthetic I/O matrices of dimension 2352×61, where 2352 is the product of the 56 industries and the number of countries with the exception of Italy and 61 is the number of industries in Italy (final sectors included). To select the number of clusters[19] we consider the ratio $\frac{\text{WSS}}{\text{TSS}}$, where *WSS* is the "Within-cluster sum of squares" and *TSS* is the "Total sum of squares." More in detail, the optimal number of clusters is the minimum such number for which $\frac{\text{WSS}}{\text{TSS}} < K$, where K is a cutoff that we set to 0.5.

It is worth observing that the volume of inputs (outputs) taken from (given to) a specific country is not equally distributed all along other countries. Table 2 and Fig. 2 show, for the case of inputs from Italy in 2010, a strong dispersion both in terms of averages by countries and in terms of within-countries standard deviations. This evidence further motivated us to use AACD as a dissimilarity measure for hierarchical clustering. Table 3 reports the results of the simulation using (stacked) years from 2010 to 2013.

According to the results in Table 3, we obtain a fairly good clustering with around 21/22 groups.[20] Based on these results, in Sect. 3.5 we consider as similar those

[19] Alternative criteria for the choice of the optimal number of clusters are the Elbow method (Thorndike 1953), the average silhouette method (Rousseeuw 1987) and the "Gap" index (Tibshirani et al. 2001).

[20] It is worth noting here that the optimal number of clusters for the matrix where Italy is in input is a bit smaller than the optimal number of clusters for the matrix where Italy is in output. This difference may be due to the asymmetry between inflow and outflow trade data, motivated by the fact that countries, typically, import goods that they do not have, and they export goods that they produce.

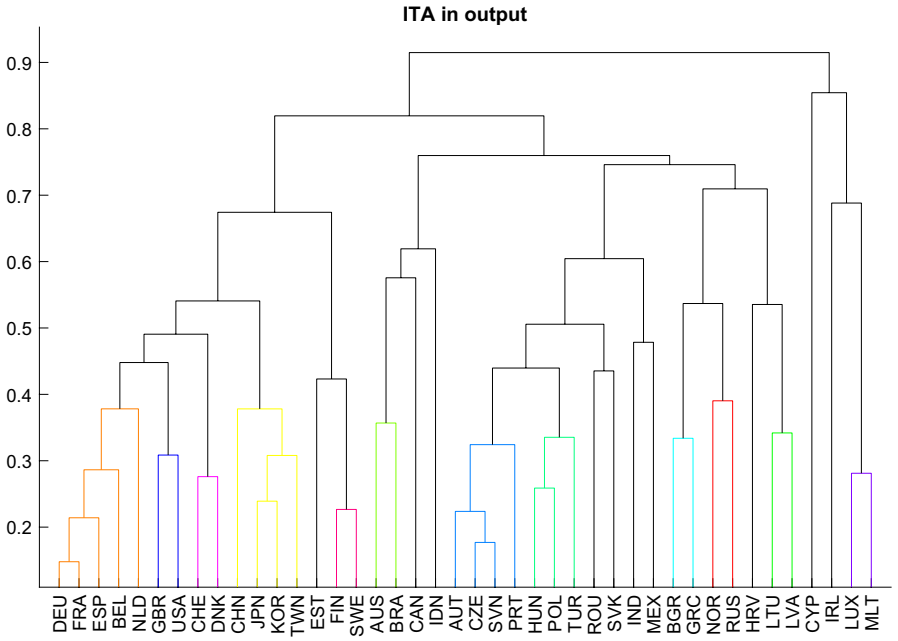


Fig. 5 Dendrograms of input countries with Italian sectors in output, based on WIOD tables (stacked over the years 2010–2013). Hierarchical clustering performed with the AACD dissimilarity measure (y-axis) and complete linkage. 22 desired groups (compare with Table 3). Countries in the same cluster are depicted with the same color. Countries in singleton clusters are highlighted in black (color figure online)

Table 5 Structure of the WIOD submatrix used for the example reported in Fig. 6a

I/O, year			
ITA/AUT, 2010	ITA/BEL, 2010	ITA/DEU, 2010	ITA/NDL, 2010
ITA/AUT, 2011	ITA/BEL, 2011	ITA/DEU, 2011	ITA/NDL, 2011
ITA/AUT, 2012	ITA/BEL, 2012	ITA/DEU, 2012	ITA/NDL, 2012
ITA/AUT, 2013	ITA/BEL, 2013	ITA/DEU, 2013	ITA/NDL, 2013
ITA/AUT, 2014	ITA/BEL, 2014	ITA/DEU, 2014	ITA/NDL, 2014

Similar comments as in Table 4 apply

Table 6 Structure of the WIOD submatrix used for the example reported in Fig. 7a

I/O, year			
ITA/AUS, 2010	ITA/BEL, 2010	ITA/JPN, 2010	ITA/MLT, 2010
ITA/AUS, 2011	ITA/BEL, 2011	ITA/JPN, 2011	ITA/MLT, 2011
ITA/AUS, 2012	ITA/BEL, 2012	ITA/JPN, 2012	ITA/MLT, 2012
ITA/AUS, 2013	ITA/BEL, 2013	ITA/JPN, 2013	ITA/MLT, 2013
ITA/AUS, 2014	ITA/BEL, 2014	**ITA/JPN, 2014**	ITA/MLT, 2014

Similar comments as in Table 4 apply

Fig. 6 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5. **c** Singular values' distribution of the WIOD submatrix reported in Table 5, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion). **d** Colored visualization of the elements of the WIOD submatrix reported in Table 5, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (color figure online)

countries belonging to the same cluster in a configuration with 21 groups (when Italy is in input) or 22 groups (when Italy is in output).

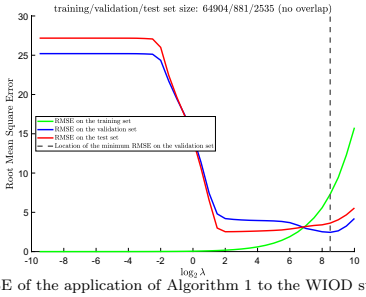
Finally, MC has been applied to I/O subtables associated with suitable groups of "fictitious" similar and dissimilar countries obtained from synthetic matrices, and the performance has turned out to be similar to the one obtained in the case of real data, which is discussed extensively in Sect. 3.5. To ease the reading of the work and give more focus on the combined application of hierarchical clustering and MC to real-world data, the results obtained for the case of synthetic matrices are reported in Sect. 3.6, after presenting in detail the case of real-world matrices in the next two subsections.

3.4 Application to prediction of a subset of entries of an I/O table, based on historical data

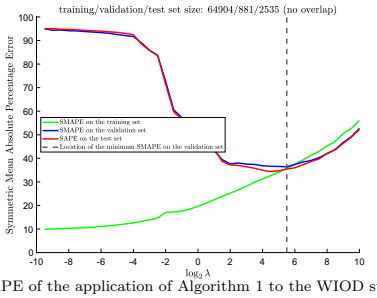
In this subsection, we consider the application of MC to the prediction of a subset of current entries of an I/O table, based on historical data: more precisely, both past entries and other available current entries coming from other portions of such an I/O table are used to make predictions. In more detail, the following situation is considered. Starting from WIOD tables relative to some consecutive years, the information associated with a subset of countries is reported in a matrix, keeping only the off-main diagonal blocks, as done in the previous subsection. Then, for one of these ordered pairs of countries, the information about the last year is obscured, and one tries to reconstruct it by MC. As an example, Table 4 refers to the case in which the countries considered are France and Italy, and the years analyzed are 2010, 2011, 2012, 2013 and 2014. All the entries related to France imports in 2014 coming from Italy (i.e., input sectors are from Italy and intermediate/final outputs are from France) are obscured,[21] then such entries are reconstructed by MC.[22] The rationale behind this application is that WIOD tables are obtained by combining information coming from different sources, and these are not necessarily synchronized. So, one

[21] Here, the term "obscured" means "not observed." It is important to remark that this is conceptually different from "set to 0" (although the obscured entries may be set to 0 at the initialization phase of an MC algorithm, as in the case of Soft Impute). To clarify how the MC optimization problem (12) deals with the obscured entries of the partially observed matrix $\mathbf{M}$, it is enough to look at the form of its objective function, which takes into account only the observed elements $M_{i,j}$ of that matrix. So, any initial assignment of values to the unobserved entries is irrelevant.

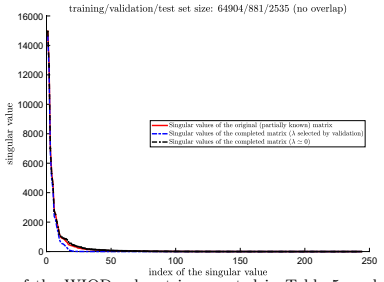
[22] The reader is referred to Appendix 2 for a technical motivation of the specific displacement of the blocks in Table 4. Such a motivation is obtained by comparing it with a possible alternative displacement.



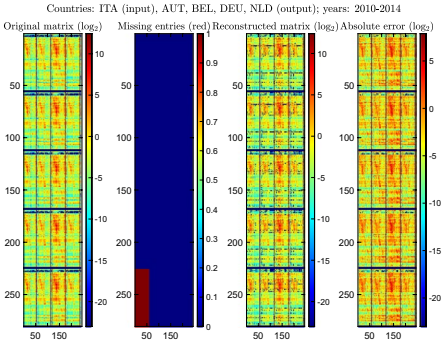
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5.



(c) Singular values' distribution of the WIOD submatrix reported in Table 5, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion).



(d) Colored visualization of the elements of the WIOD submatrix reported in Table 5, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error.

could combine the complete information available in the past with the partial one currently available, to predict currently missing elements.

In the following application of MC, our focus is not on the absolute RMSE, but on its percentage of reduction (in correspondence of the optimal value of the regularization parameter λ), with respect to a base case[23] (represented by $\lambda \simeq 0$). Figure 3a illustrates the results of the application of the MC Algorithm 1 to the partially observed WIOD submatrix reported in Table 4 (details about the construction of the training, validation and test sets are provided at the end of Sect. 2.2). The figure shows that, in this case, MC is able to reduce significantly the RMSE of reconstruction on the missing elements in 2014, when moving from the case $\lambda \simeq 0$ (for which the predictions of the missing elements are nearly equal to 0) to the optimal choice of λ. More precisely, for $\lambda \simeq 0$, one gets

$$\begin{aligned} \mathrm{RMSE}_{\lambda}^{\mathrm{val}} &= \sqrt{\frac{1}{|\Omega^{\mathrm{val}}|} \sum_{(i,j)\in\Omega^{\mathrm{val}}} \left(M_{i,j} - \hat{M}_{\lambda,i,j}\right)^2} \\ &\simeq \sqrt{\frac{1}{|\Omega^{\mathrm{val}}|} \sum_{(i,j)\in\Omega^{\mathrm{val}}} M_{i,j}^2} \simeq 86.5711 \end{aligned} \tag{21}$$

(since, in this case, one gets $\mathbf{S}_{\lambda}(\mathbf{Y}) \simeq \mathbf{Y}$ from Eq. (14), hence every time Step 2.a of Algorithm 1 is performed, one gets a matrix $\hat{\mathbf{M}}^{\mathrm{new}}$ whose entries are nearly equal to 0 in the positions corresponding to unobserved entries of $\mathbf{M}$). Instead, for the optimal value $\lambda^{\circ} = 2^6 = 64$ of λ (whose location is highlighted in the figure), one gets

$$\mathrm{RMSE}_{\lambda^{\circ}}^{\mathrm{val}} \simeq 47.2791\,,$$

obtaining a reduction of the RMSE of about 45%. A similar behavior is observed for the RMSE of matrix reconstruction on the test set, the reduction of such RMSE in this case being from 116.0874 (for $\lambda \simeq 0$) to 51.7544 (for $\lambda = \lambda^{\circ}$), which amounts at about 55%. Hence, a good generalization capability is observed, showing no overfitting occurred in the application of MC.[24] Results in terms of SMAPE are reported in Fig. 3b. The

[23] Our specific choice for the base case is related to the fact that the goal of our analyses in the next Sects. 3.5 and 3.6 is to assess if MC behaves better when it is applied to I/O subtables composed of data coming similar countries (i.e., belonging to the same cluster), with respect to the case in which it is applied to I/O subtables composed of data coming from dissimilar countries (i.e., belonging to different clusters). So, in a sense, we are comparing MC to itself, but in different situations. We believe that the relative improvement achieved by MC at the optimal choice of its regularization parameter with respect to this base case (the same for all the I/O subtables considered) is a fair way to address this specific comparison.

[24] In more detail, we argue that in this case there is no overfitting due to the two following reasons: the RMSE curves on the validation and test sets have a similar behavior, with approximately the same locations of the respective minimizers; the RMSEs on the three sets (training, validation and test sets) have quite similar orders of magnitude in correspondence of the optimal value of λ. It is also worth noticing that the RMSE curve on the training set is always a non-decreasing function of λ (as it can be easily proved following standard arguments from regularization theory). This remark does not extend to the SMAPE curve on the training set, since it is not part of the objective function of the MC optimization problem (11).

irregular behavior of the curves associated with the SMAPE is due to the fact that MC does not address directly the SMAPE criterion, whereas the RMSE on the training set is part of the objective function of the MC optimization problem (11). Moreover, Fig. 3c compares the singular values' distribution of the WIOD submatrix reported in Table 4, and the one of the completed submatrix produced as output by the algorithm, for both $\lambda \simeq 0$ and the optimal value of λ. It is evident from the figure that MC was able to reconstruct excellently the singular values' distribution of the original WIOD submatrix (part of which was not observed), due to the large overlap of the curves reported in the figure. Moreover, such distribution decays rapidly to 0, which, as already reported in the Introduction, is a necessary (but not sufficient) condition for a good performance of MC. Indeed, Eckart–Young theorem (see Appendix 1) provides an upper bound on the performance of MC, for a given number of singular values kept. Finally, Fig. 3d shows a colored visualization of the elements of the original WIOD submatrix, the positions of the missing entries (highlighted in red), the reconstructed submatrix obtained for the optimal value of the regularization constant λ, and the element-wise absolute value of the reconstruction error. It is worth recalling that the positions of the missing entries form the union of the validation and test sets, whereas the positions of the observed entries form the training set.[25] In this case, although the third column in Fig. 3d shows that MC looks able to reconstruct some pattern in the missing block of the matrix (with respect to the case $\lambda \simeq 0$, for which the missing block is predicted as a block of all negligible elements), the reconstruction error looks to be still large (fourth column), having a similar pattern as the corresponding original non-obscured block (first column). This is partly due to the fact that a 50% reconstruction error corresponds to a reduction by 1 in logarithmic scale with base 2. Improved results are reported in the next subsection (see Figs. 6a and 8a), where the choice of the WIOD subtable to which MC is applied is guided by hierarchical clustering.

3.5 Matrix completion applied to historical data for groups of similar/dissimilar countries determined by hierarchical clustering

In this subsection, using data from the WIOD latest release, we compare the application of MC to WIOD submatrices obtained using a pre-processing step based on hierarchical clustering.[26] The dissimilarity between any two countries c_1 and c_2 is computed as the AACD between the corresponding blocks of $\mathbf{T}$ in the WIOD table (stacked by considering several consecutive years), obtained by either choosing Italian sectors in input and intermediate/final outputs from the two countries c_1 and c_2 ($\mathbf{T}^{Italy,c_1}$ and $\mathbf{T}^{Italy,c_2}$, recalling the notation introduced in Sect. 2.1), or choosing

[25] The explicit partition of the set of missing entries into validation and test sets is not reported in the figure, since such partition is randomly generated.

[26] It is well known from the standard practice of projection of I/O tables that it is typically better to use a table in previous years to project missing pieces of current I/O tables than using current I/O tables of similar countries (Rueda-Cantuche et al. 2018; Valderas-Jaramillo et al. 2021). One reason is that the former gathers detailed country-specific information that is not expected to change in the short term. In this subsection, we use a combined approach, because we consider both the information coming from the same block in previous years, and the one coming from similar blocks in the current year.

Fig. 7 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6. **c** Singular values' distribution of the WIOD submatrix reported in Table 6, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion). **d** Colored visualization of the elements of the WIOD submatrix reported in Table 6, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (color figure online)

Table 7 Structure of the WIOD submatrix used for the example reported in Fig. 8a

I/O, year			
BEL/ITA, 2010	DEU/ITA, 2010	ESP/ITA, 2010	FRA/ITA, 2010
BEL/ITA, 2011	DEU/ITA, 2011	ESP/ITA, 2011	FRA/ITA, 2011
BEL/ITA, 2012	DEU/ITA, 2012	ESP/ITA, 2012	FRA/ITA, 2012
BEL/ITA, 2013	DEU/ITA, 2013	ESP/ITA, 2013	FRA/ITA, 2013
BEL/ITA, 2014	DEU/ITA, 2014	ESP/ITA, 2014	FRA/ITA, 2014

Similar comments as in Table 4 apply

Table 8 Structure of the WIOD submatrix used for the example reported in Fig. 9a

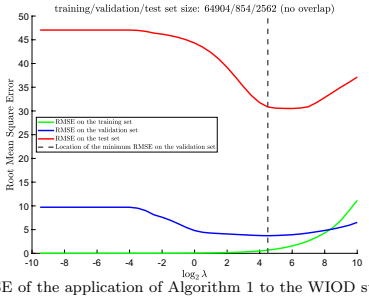
I/O, year			
DEU/ITA, 2010	IND/ITA, 2010	MLT/ITA, 2010	SVN/ITA, 2010
DEU/ITA, 2011	IND/ITA, 2011	MLT/ITA, 2011	SVN/ITA, 2011
DEU/ITA, 2012	IND/ITA, 2012	MLT/ITA, 2012	SVN/ITA, 2012
DEU/ITA, 2013	IND/ITA, 2013	MLT/ITA, 2013	SVN/ITA, 2013
DEU/ITA, 2014	IND/ITA, 2014	MLT/ITA, 2014	SVN/ITA, 2014

Similar comments as in Table 4 apply

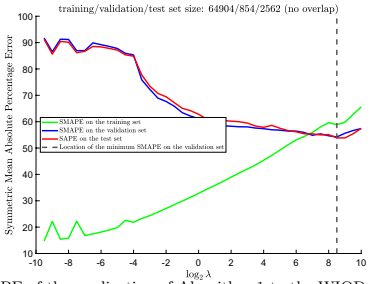
Italian intermediate/final outputs and sectors from the two countries c_1 and c_2 in input ($\mathbf{T}^{c_1,Italy}$ and $\mathbf{T}^{c_2,Italy}$). In other words, the dissimilarity of the two countries c_1 and c_2 in their Italian export patterns is evaluated in the first case, whereas their dissimilarity in the respective Italian import patterns is evaluated in the second case. Both the hierarchical clustering analyses are repeated taking as inputs stacked I/O blocks associated with several years (2010, 2011, 2012 and 2013), and using complete linkage to perform clustering. Figures 4 and 5 report the dendrograms obtained, where c_1 and c_2 are, respectively, both output countries (Fig. 4), and both input countries (Fig. 5).

In this way, it is possible to extract from Fig. 4 two groups of 4 output countries (see Tables 5 and 6) that are, respectively, in the same cluster, and in 4 different clusters.

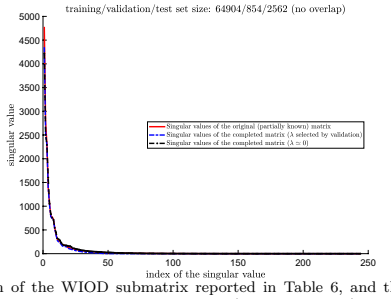
Table 5 refers to a WIOD submatrix whose blocks have Italian sectors in input and intermediate/final outputs associated with the first group of extracted countries (specifically, Austria, Belgium, Germany and the Netherlands). In contrast, in Table 6, the intermediate/final outputs refer to the second group of extracted countries (specifically, Australia, Belgium, Japan and Malta). For predictive/MC



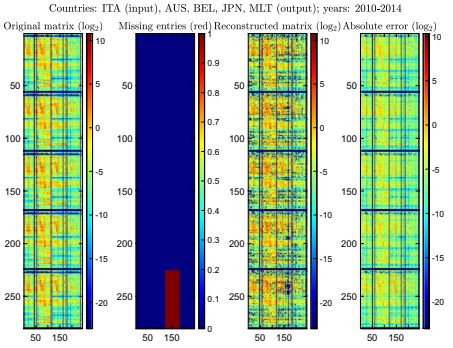
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6.



(c) Singular values' distribution of the WIOD submatrix reported in Table 6, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion).



(d) Colored visualization of the elements of the WIOD submatrix reported in Table 6, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error.

Fig. 8 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7. **c** Singular values' distribution of the WIOD submatrix reported in Table 7, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion). **d** Colored visualization of the elements of the WIOD submatrix reported in Table 7, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (color figure online)

purposes, the tables contain also data related to the year 2014.[27] Then, the MC Algorithm 1 is applied to both submatrices, after obscuring all the elements of their last block (highlighted in bold in Tables 5 and 6), which refers to a specific output country in 2014.

Figures 6a and 7a report the results of the application of the MC Algorithm 1 to the two WIOD submatrices whose structures are described in Tables 5 and 6, respectively. As expected, the results show a better performance of the MC algorithm, measured in terms of the percentage of reduction of the RMSE on the validation set from $\lambda \simeq 0$ to the optimal choice of λ, in the case of the first submatrix, whose intermediate/final outputs are associated with more similar countries.

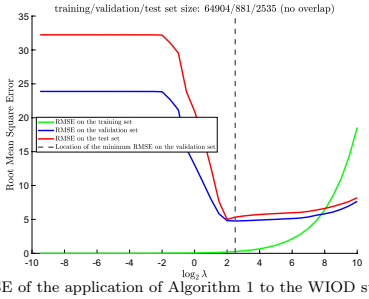
It is worth observing that quite similar results have been obtained if a different 2014 block corresponding to another country in the group of 4 countries is obscured in each of the two WIOD submatrices, or when the analysis has been repeated by considering Italy in output and 4 similar/dissimilar countries in input (see the dendrogram reported in Fig. 5). In this second analysis, the selected subset of 4 similar countries in input is made by Belgium, Germany, Spain and France (see Table 7), whereas the selected subset of 4 dissimilar countries in input is made by Germany, India, Malta and Slovenia (see Table 8). Corresponding results of the MC analysis are reported in Figs. 8a and 9a. Again, similar comments as before apply: when more similar input countries are considered and the RMSE criterion is considered, the performance of MC improves.

Moreover, a comparison of Figs. 6b, 7b, 8b and 9b show that, also when the SMAPE performance measure is used, MC applied to similar countries produces better results (in terms of relative improvement with the respect to the baseline case) than MC applied to dissimilar countries.

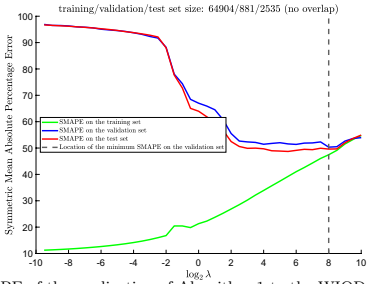
As shown later in Sect. 3.7, qualitatively similar results as in this subsection have been obtained by varying the random choices of the validation and test sets.

[27] It is worth observing that, in Tables 5 and 6, and in the successive Tables 7 and 8, the blocks associated with a specific year are concatenated horizontally according to a given order, and such order does not change when considering the blocks associated with different years. The choice of a specific order is actually irrelevant for the application of MC, as it follows by combining the two following observations:

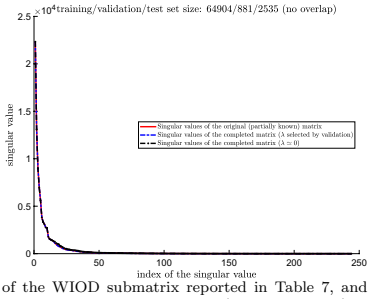
- The singular values of a rectangular matrix are invariant with respect to permutations of rows and/or columns of that matrix;
- If the Soft Impute algorithm is applied to two different rectangular matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ that differ only for permutations of rows and/or columns, and if the same permutations are applied to the matrix that represents the locations of the observed/unobserved elements (i.e., such matrix contains 1 for every observed entry, and 0 for every unobserved one), then the two matrices $\hat{\mathbf{M}}_1$ and $\hat{\mathbf{M}}_2$ obtained as output of MC in the two cases differ only for the same permutations.



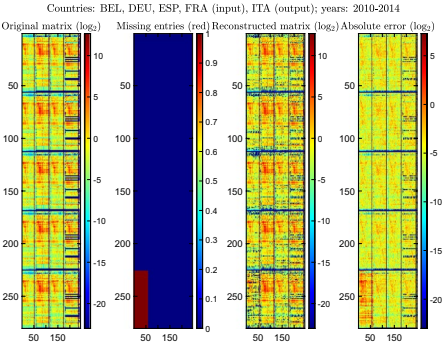
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7.



(c) Singular values' distribution of the WIOD submatrix reported in Table 7, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion).



(d) Colored visualization of the elements of the WIOD submatrix reported in Table 7, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error.

Fig. 9 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8. **c** Singular values' distribution of the WIOD submatrix reported in Table 8, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion). **d** Colored visualization of the elements of the WIOD submatrix reported in Table 8, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (color figure online) ▶

3.6 Performance of matrix completion on simulated matrices

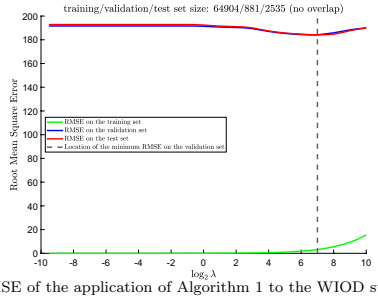
In this subsection, we show that the application of MC on the simulated data of Sect. 3.3 produces similar results as its application to the original data (see Sect. 3.5). In the following, for illustrative purposes, we focus just on one of the simulated matrices considered in Sect. 3.3 (in the next figures, the synthetic countries are still named as the original countries, since their respective data are obtained by perturbations of the ones of the associated original countries).

Figures 10 and 11 show the results of the hierarchical clustering, obtained, respectively, with Italy in input and in output.

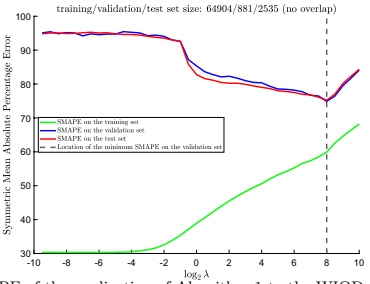
Then, based on the dendrogram shown in Figs. 10 and 12 compares the MC performance (for short, limiting to the RMSE criterion) for the cases—analogous to the ones considered in Sect. 3.5—in which the 4 selected synthetic countries belong, respectively, to the same cluster (similar synthetic countries: ESP, FRA, GBR, USA, obscured one in the last year: ESP) and to different clusters (dissimilar synthetic countries: CYP, ESP, IDN, MEX, obscured one in the last year: CYP). Analogously, based on the dendrogram shown in Figs. 11 and 13 compares the MC performance (again, limiting to the RMSE criterion) for the cases—analogous to the ones considered in Sect. 3.5—in which the 4 selected synthetic countries belong, respectively, to the same cluster (similar synthetic countries: BEL, DEU, ESP, FRA, obscured one in the last year: BEL) and to different clusters (dissimilar synthetic countries: CZE, DEU, EST, IND, obscured one in the last year: DEU). The results are qualitatively similar to the ones reported in for the original data, and demonstrate the robustness of the proposed approach of analysis, which combines hierarchical clustering and MC. Similar results, not reported here, are obtained when the SMAPE criterion is used to compare the performance of MC for similar and dissimilar countries.

3.7 Results for different choices of the validation and test sets

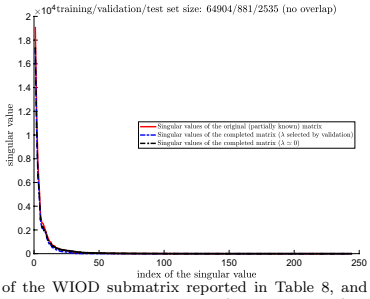
In order to investigate how the results obtained in Sect. 3.5 may depend on the random choices of the validation and test sets inside the obscured blocks, we report in Fig. 14a–d some variations of Figs. 6a, 7, 8 and 9a, achieved by considering, for illustrative purposes, 5 such random choices. Similarly, in Fig. 15a–d, we do the same to produce variations of Figs. 6b, 7, 8 and 9a. Figures 14a–d and 15a–d show that qualitatively similar results as in Sect. 3.5 are obtained in this way. Moreover, especially in the case of the RMSE criterion, they further justify focusing on the relative improvement achieved by MC, as there is some variability in the validation and test set RMSEs obtained for $\lambda \simeq 0$. It is also worth noticing that the variability



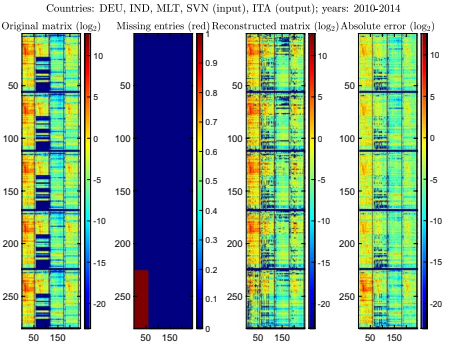
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8.



(c) Singular values' distribution of the WIOD submatrix reported in Table 8, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion).



(d) Colored visualization of the elements of the WIOD submatrix reported in Table 8, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error.

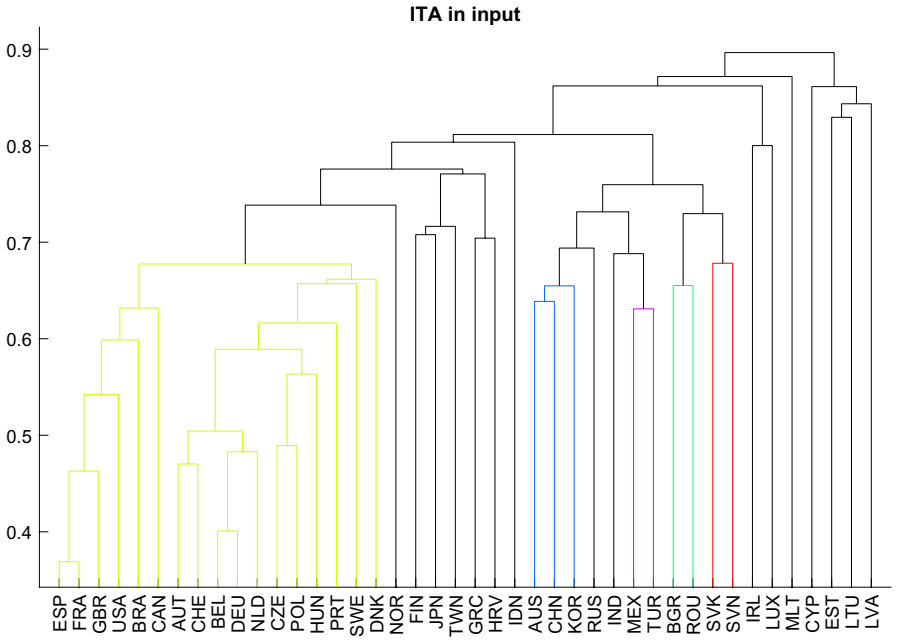


Fig. 10 Dendrograms of synthetic output countries with Italian sectors in input, based on synthetic WIOD tables (stacked over the years 2010–2013). Hierarchical clustering performed with the AACD dissimilarity measure (y-axis) and complete linkage. 21 desired groups (compare with Table 3). Countries in the same cluster are depicted with the same color. Countries in singleton clusters are highlighted in black (color figure online)

of the curves looks larger for the cases of Fig. 14b and d, which refer to situations in which MC is applied to I/O subtables made of dissimilar countries. This is in agreement with our intuition that MC performs better when it is applied to I/O subtables made of similar countries (see Fig. 14a and c).

Qualitatively similar results, not reported here, have been obtained by changing randomly the validation and test sets related to Figs. 12 and 13 in Sect. 3.6.

3.8 Application of matrix completion to a WIOD submatrix containing both intra-country and inter-country blocks

In Table 9, we consider the following variation of Table 4, in which we take into account also the domestic block associated with Italy, evaluated in different years.

For what concerns the application of MC, due to the different orders of magnitude of the elements contained in the domestic blocks compared to the ones belonging to the other blocks, the range of values for the regularization parameter has been increased for this specific example, by setting $\lambda_k = 2^{k/2-20}$, for $k = 1, \ldots, 80$. As

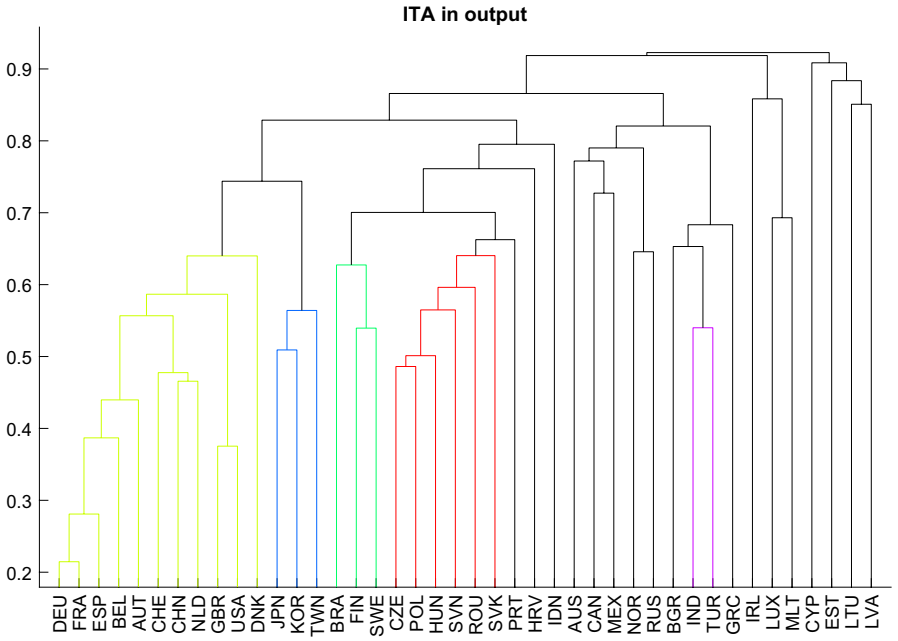


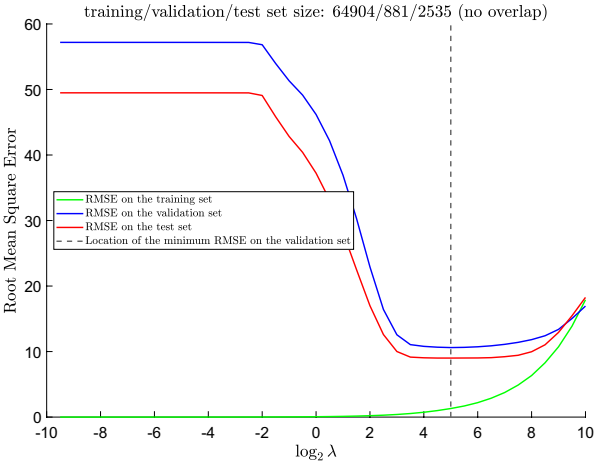
Fig. 11 Dendrograms of synthetic input countries with Italian sectors in output, based on synthetic WIOD tables (stacked over the years 2010–2013). Hierarchical clustering performed with the AACD dissimilarity measure (y-axis) and complete linkage. 22 desired groups (compare with Table 3). Countries in the same cluster are depicted with the same color. Countries in singleton clusters are highlighted in black (color figure online)

highlighted by Fig. 16, in this case, the performance of MC is quite bad, likely due to the highly different orders of magnitude of the elements in the various blocks.

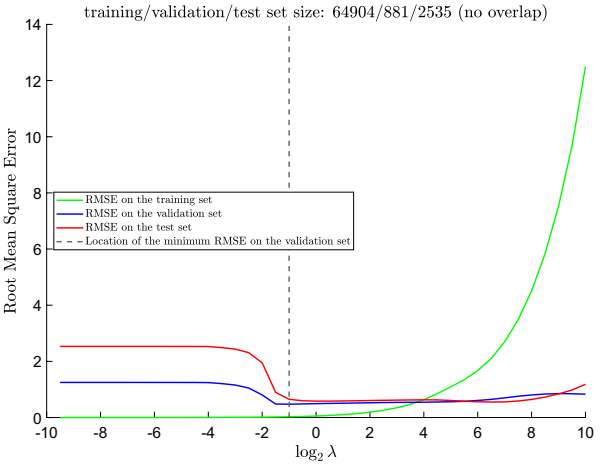
4 Future research and concluding remarks

This work represents the first attempt to adopt a matrix completion (MC) algorithm, combined with a hierarchical clustering pre-preprocessing step, to predict missing entries in submatrices of I/O tables in the context of a panel data analysis.

The particular structure of I/O tables, reported in the article, makes the data reconstruction and prediction problems not trivial. Hence, in the pre-processing phase, we have employed the dissimilarity pattern of countries to define low-rank I/O subtables with few dominant singular values. A panel matrix completion with nuclear norm penalty has been tested on those low-rank subtables. The effectiveness of the proposed method according to historical data available from previous years has been

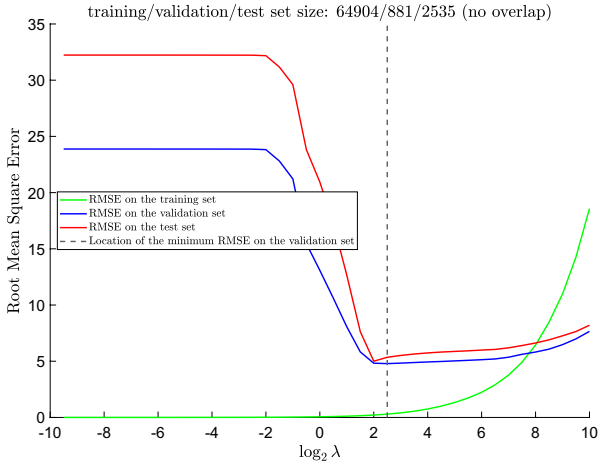


(a) Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 similar synthetic countries, with Italy in input.

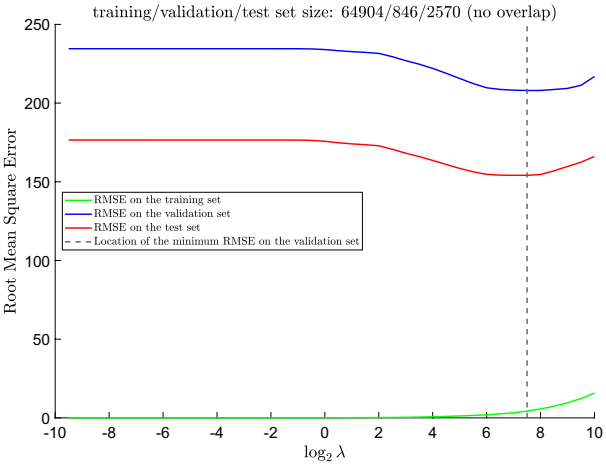


(b) Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 dissimilar synthetic countries, with Italy in input.

Fig. 12 **a** Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 similar synthetic countries, with Italy in input. **b** Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 dissimilar synthetic countries, with Italy in input



(a) Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 similar synthetic countries, with Italy in output.



(b) Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 dissimilar synthetic countries, with Italy in output.

Fig. 13 **a** Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 similar synthetic countries, with Italy in output. **b** Results in terms of RMSE of the application of Algorithm 1 to one synthetic WIOD submatrix made of 4 dissimilar synthetic countries, with Italy in output

Fig. 14 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5. **b** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6. **c** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7. **d** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8

demonstrated when the considered I/O subtables are obtained by selecting similar countries.

A first possible extension of the analysis concerns comparing matrix reconstruction of I/O tables (in one year, based on current and previous years) based on the repeated application of matrix completion to several subtables[28] of the original I/O table, instead of a single more computationally expensive and (presumably) less effective application to the whole table (possibly after removing domestic blocks, likewise in this article). For what concerns the possible dependence of the results on the cluster size (in the case of I/O subtables associated with countries coming from the same cluster), it is worth noticing that the results reported in the present work refer to clusters having slightly different sizes. So, the proposed approach has the potential to work well (compared to the alternative selection of countries from distinct clusters) with different cluster sizes. A more extensive analysis (based either on artificial data or on real-world data, possibly with various selections of the given country in input or output) would be needed to further check this. This is left as a future development, since it would require a much larger number of (computationally intensive) repeated applications of MC.

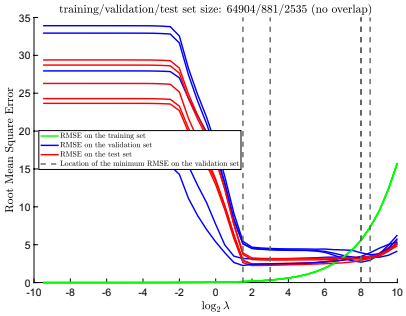
The proposed methodology is expected to be applicable, with similar results, also to other I/O tables (either industry-by-industry and product-by-product ones), because their structure is often similar to the one of WIOD tables, as highlighted in this work.[29]

As a second possible extension, algorithms for clustering and matrix completion different from those employed in the present article could be used. Moreover, matrix completion itself could be compared with other imputation methods for missing entries in panel data models. A comparison with alternative methods such as the one suggested by Rueda-Cantuche et al. (2018) is left for future research.

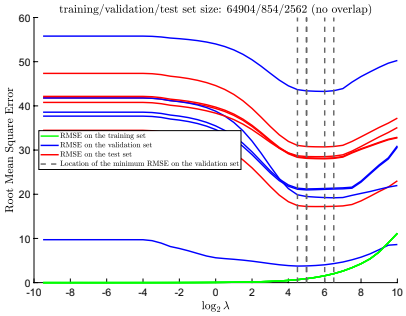
A third possible extension consists in applying the MC algorithm not to the original I/O subtable, but to its suitable pre-processed version, obtained by removing from that subtable its prediction provided by another method. This "ensemble learning" approach would combine the two methods, with the aim of possibly obtaining better predictions.

[28] In view of this possible future development, it looks preferable to identify a small subset of similar countries (as done in our work, through our choice of the threshold used in hierarchical clustering aimed to find the optimal number of clusters). The choice of a small subset of similar countries is supported not only by the fact that Algorithm 1 is slower for larger matrices (which would arise by selecting a larger subset of countries in the analysis), but also by the fact that larger cluster sizes may worsen the performance of MC, as the risk of having less similar countries in the same cluster would increase. In other words, it looks preferable to obtain clusters with as much as possible small within-cluster sum of squares.

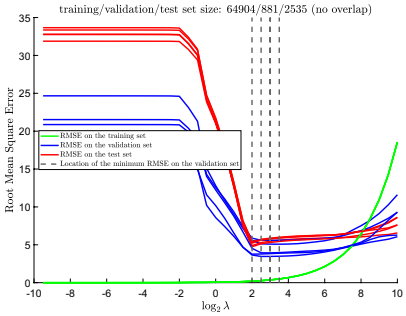
[29] In this regard, it is worth mentioning that WIOD is currently available only till 2014, and it is unlikely that there will be important updates of it in the future, whereas OECD and FIGARO I/O tables are expected to be the main official sources for the future.



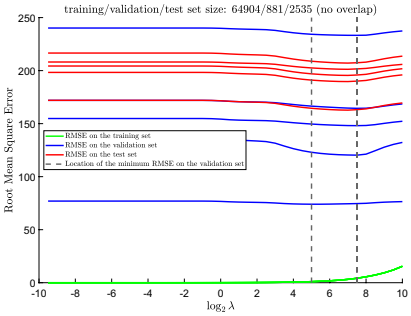
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5.



(b) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6.



(c) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7.



(d) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8.

Fig. 15 **a** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6. **c** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7. **d** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8 ▶

Table 9 Structure of the WIOD submatrix used for the example reported in Fig. 16

I/O, year		
FRA/ITA, 2010	ITA/FRA, 2010	ITA/ITA, 2010
FRA/ITA, 2011	ITA/FRA, 2011	ITA/ITA, 2011
FRA/ITA, 2012	ITA/FRA, 2012	ITA/ITA, 2012
FRA/ITA, 2013	ITA/FRA, 2013	ITA/ITA, 2013
FRA/ITA, 2014	ITA/FRA, 2014	**ITA/ITA, 2014**

Similar comments as in Table 4 apply

Finally, as another possible extension, the approach adopted in the paper could be applied to generate counterfactuals of I/O submatrices: e.g., by predicting how the entries of a suitably specified input–output submatrix related to Japan would have changed, in case the March 2011 earthquake and tsunami and the successive Fukushima Daiichi nuclear disaster (Yonemoto 2016) would have not occurred. To do this, one would preliminary need to identify sectors of the economy that were not affected by such events (i.e., untreated sectors), then obscure (and reconstruct) the entries of that submatrix related to other sectors that were affected (i.e., treated sectors).

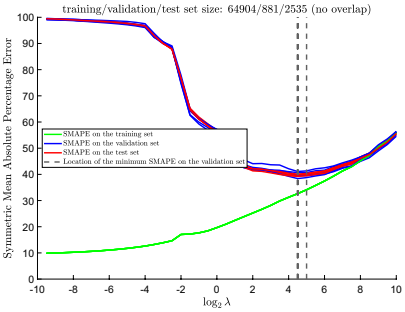
Appendix

Relationship between performance of matrix completion and singular value decomposition of the matrix to be completed

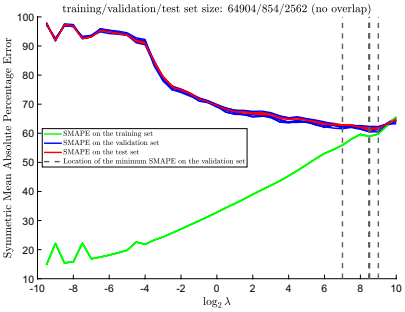
Let k indicate any non-negative integer smaller than the rank of a rectangular matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ (denoted as $\text{rank}(\mathbf{M})$). Moreover, let σ_i (for $i = 1, \ldots, \text{rank}(\mathbf{M})$) be the singular values of $\mathbf{M}$, ordered nonincreasingly. Then, by Eckart–Young theorem (Moitra 2018, Theorem 2.1.2), the best rank-k approximation $\mathbf{M}_k$ of $\mathbf{M}$ according to the Frobenius norm is provided by a truncated singular value decomposition of $\mathbf{M}$, in which one keeps only its k-largest singular values, and zeroes all the others. Its Root Mean Square Error (RMSE) of reconstruction is equal to

$$RMSE_k := \frac{1}{\sqrt{mn}} \|\mathbf{M} - \mathbf{M}_k\|_F = \frac{1}{\sqrt{mn}} \sum_{i=k+1}^{\text{rank}(\mathbf{M})} \sigma_i^2 . \tag{A1}$$

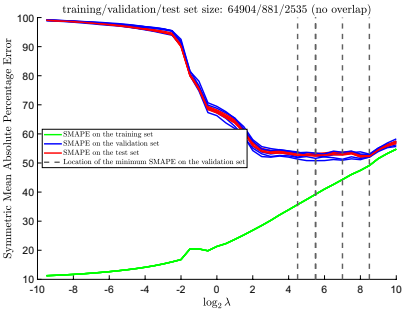
Equation (A1) shows that either a low rank of the matrix to be completed or a fast decay to 0 of its singular values' distribution are important for a successful



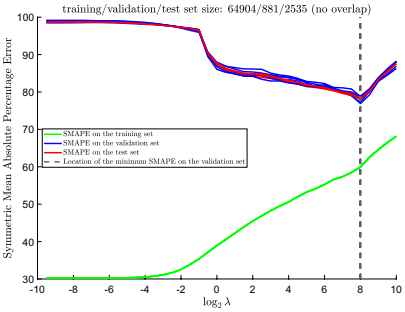
(a) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 5.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 6.



(c) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 7.



(d) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 8.

Fig. 16 **a** Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported ▶ in Table 9. **b** Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 9. **c** Singular values' distribution of the WIOD submatrix reported in Table 9, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion). **d** Colored visualization of the elements of the WIOD submatrix reported in Table 9, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error (color figure online)

application of MC to that matrix, as they imply a small right-hand side of the equation. It has to be remarked, however, that they provide only a necessary condition for such a successful application since, in the context of MC, finding exactly the singular value decomposition of the matrix $\mathbf{M}$ is infeasible, being $\mathbf{M}$ only partially observed. The reader is referred to Nguyen et al. (2019) for more details about properties of a matrix and on the distribution of its sampled elements which allow a successful application of some forms of MC to that matrix. It is worth remarking that some of such properties refer to the possibility of an exact (or "perfect") reconstruction of the matrix, which is not really needed neither feasible in our specific application of MC to I/O matrices.

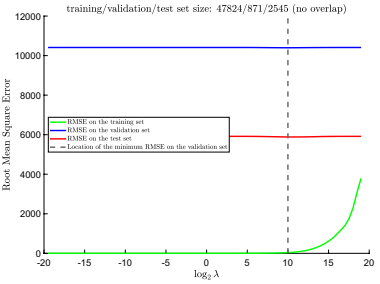
Alternative displacement of the blocks

It is worth observing that the several blocks of the WIOD submatrix reported in Table 4 have been displaced in such a way to make the application of MC possible. To clarify this issue, an alternative displacement of the same blocks is reported in Table 10.

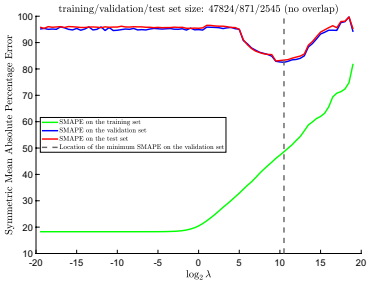
In the situation described by Table 10, differently from the one illustrated by Table 4, some rows of the submatrix are obscured as a whole, and the optimal solution to the optimization problem (11) is trivial, in the sense that the obscured rows are reconstructed by setting all their elements to 0. This can be explained as follows. Given a partially observed matrix $\mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{M})$ and setting, in Algorithm 1, its tolerance parameter tol to 0 and its upper bound N^{it} on the number of iterations to ∞, it is well known from the convergence of that algorithm (Mazumder et al. 2010, Lemma 5) that the sequence of solutions $\hat{\mathbf{M}}^{\text{new}}$ it generates[30] tends to an optimal solution $\hat{\mathbf{M}}_{\lambda^\circ}$ of the optimization problem (11). In case the i-th row of $\mathbf{M}$ is not observed, one can show[31] that all the elements in the same i-th row of each matrix $\hat{\mathbf{M}}^{\text{new}}$ generated by the algorithm (which is renamed as $\hat{\mathbf{M}}^{\text{old}}$ at the end of the corresponding

[30] The choices $tol = 0$ and $N^{\text{it}} = \infty$ prevent early termination of the algorithm, i.e., its termination before convergence has been achieved.

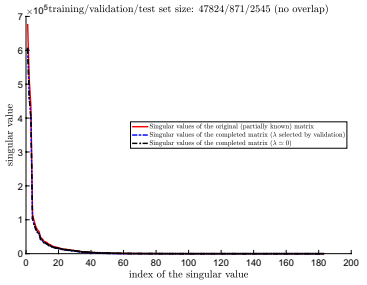
[31] Detailed arguments are reported in this footnote. In the situation considered here, if at a generic iteration of Algorithm 1 the i-th row of the matrix $\hat{\mathbf{M}}^{\text{old}}$ is made of all zeros (and this surely holds at the initialization of the algorithm), then also the i-th row of the matrix $\mathbf{N}^{\text{old}} = \mathbf{P}_{\Omega^{\text{tr}}}(\mathbf{M}) + \mathbf{P}^{\perp}_{\Omega^{\text{tr}}}(\hat{\mathbf{M}}^{\text{old}})$ is made of all zeros. Let $\mathbf{N}^{\text{old}} = \mathbf{U}^{\text{old}}\mathbf{\Sigma}^{\text{old}}\left(\mathbf{V}^{\text{old}}\right)^T$ be the singular value decomposition of such matrix. Then, when the index j is associated with a singular value of $\mathbf{N}^{\text{old}}$, the element in position (i, j) of $\mathbf{U}^{\text{old}}$ is equal to 0 (this is shown by finding the singular value decomposition of the matrix obtained from $\mathbf{N}^{\text{old}}$ by removing its i-th row, then obtaining the singular value decomposition of $\mathbf{N}^{\text{old}}$ by extending with zero components the resulting left-singular vectors). Since the matrix $\hat{\mathbf{M}}^{\text{new}} = \mathbf{S}_{\lambda}\left(\mathbf{N}^{\text{old}}\right)$ has the same left-singular vectors as $\mathbf{N}^{\text{old}}$, also the i-th row of $\hat{\mathbf{M}}^{\text{new}}$ is made by all zeros. Finally, since such matrix becomes $\hat{\mathbf{M}}^{\text{old}}$ at the successive iteration of Algorithm 1, this property holds for all the matrices $\hat{\mathbf{M}}^{\text{new}}$ generated by that algorithm.



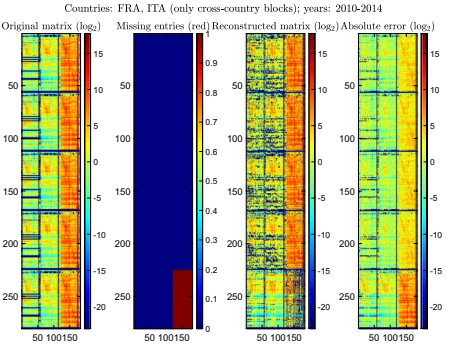
(a) Results in terms of RMSE of the application of Algorithm 1 to the WIOD submatrix reported in Table 9.



(b) Results in terms of SMAPE of the application of Algorithm 1 to the WIOD submatrix reported in Table 9.



(c) Singular values' distribution of the WIOD submatrix reported in Table 9, and the one of the completed submatrix produced by Algorithm 1 for the optimal regularization constant (RMSE criterion).



(d) Colored visualization of the elements of the WIOD submatrix reported in Table 9, positions of the missing entries, reconstructed submatrix obtained for the optimal regularization constant (RMSE criterion), and element-wise absolute value of the reconstruction error.

Table 10 Alternative displacement of the blocks for the WIOD submatrix reported in Table 4

I/O, year
FRA/ITA, 2010
ITA/FRA, 2010
FRA/ITA, 2011
ITA/FRA, 2011
FRA/ITA, 2012
ITA/FRA, 2012
FRA/ITA, 2013
ITA/FRA, 2013
FRA/ITA, 2014
ITA/FRA, 2014

All the entries contained in the block highlighted in bold are obscured. In this case, the optimization problem (11) is trivial

iteration) are equal to 0. Then, by induction and convergence, also all the elements in the i-th row of $\hat{\mathbf{M}}_{\lambda^\circ}$ are equal to 0. More generally, the problem addressed by low-rank MC is either trivial (if a regularization term is present) or ill-posed (if there is no regularization term) when some rows or columns of the matrix to be completed are obscured as a whole. This depends on the fact that the other rows/columns provide no information on the obscured ones (Candès and Tao 2010). In both cases, the application of MC becomes meaningless.

In our specific application of MC to I/O subtables, the arrangement of blocks reported in Table 4 makes this issue of triviality/ill-posedness of the problem addressed by MC disappear (while such issue is present in the case of the arrangement of blocks reported in Table 10). This can be justified as follows. In the situation described by Table 4, differently from the one illustrated by Table 10, the particular arrangement of the blocks makes it possible for MC to "learn," if present, a relationship between the observed blocks "FRA/ITA, x" and "ITA/FRA, x" in the same year x, where $x = 2010, 2011, 2012, 2013$. Then, when the block "FRA/ITA, 2014" is observed, such a relationship makes it possible to estimate (although with some error) the block "ITA/FRA, 2014," which is not observed. In the case of Table 10, the different arrangement of the same blocks makes impossible for MC to "learn" the relationship above,[32] so the MC estimate of the block "ITA/FRA, 2014" is now arbitrary.

[32] Loosely speaking, this conclusion comes from the fact that, if one applies the Soft Impute algorithm to two partially observed matrices that differ only by a permutation of their rows and if also the matrix representing the locations of the observed/unobserved entries is subject to the same permutation, then the two respective completed matrices found by the algorithm differ only by the same permutation. In the case of Table 10, this means that MC fails to distinguish if the missing block "ITA/FRA, 2014" has indeed the stated form, or if it is instead another block of the form "FRA/ITA, x" with $x = 2010, 2011, 2012, 2013, 2014$, or one of the form "ITA/FRA, y," with $y \neq 2014$.

It is worth mentioning that the above-mentioned issue of triviality/ill-posedness of the problem addressed by MC does not actually arise from possibly having either a row or a column made of all zeros in the matrix to be completed, but from the fact of observing no element in such row or column. For simplicity, in the following we illustrate this issue for the case in which one looks for a completed matrix $\hat{\mathbf{M}}$ having zero error on the set of observed entries.[33] So, suppose that one knows that the matrix $\mathbf{M}$ to be completed has rank 1, and let the symbol $*$ be used to denote each of its unobserved entries. If $\mathbf{M} = \begin{bmatrix} 1 & 2 \\ * & 0 \\ 3 & * \end{bmatrix}$, then its unique rank-1 reconstruction (with no error in the observed entries) is $\hat{\mathbf{M}} = \begin{bmatrix} 1 & 2 \\ 0 & 0 \\ 3 & 6 \end{bmatrix}$. Instead, if one has $\mathbf{M} = \begin{bmatrix} 1 & 2 \\ 0 & 0 \\ * & * \end{bmatrix}$, then there is no unique rank-1 reconstruction of that matrix. On the other hand, taking an arbitrary such reconstruction would be useless.

Author Contributions (CRediT author statement) RM contributed to methodology, software, formal analysis, investigation, data curation and writing—original draft; GG contributed to validation, methodology, software, formal analysis, investigation, data curation and writing—original draft; FB contributed to formal analysis, investigation and writing—original draft; MR contributed to conceptualization, formal analysis, methodology, resources, writing—review and editing, supervision and project administration. All the authors analyzed the results and reviewed the manuscript.

Funding Open access funding provided by Università degli Studi di Salerno within the CRUI-CARE Agreement.

Availability of data and materials Data are freely available online at www.wiod.org.

Code availability Codes are available upon request.

Declarations

Conflict of interest No conflicts of interest/competing interest.

[33] This is not exactly the case considered in the paper, in which a nonzero error is actually allowed, but the two settings are clearly related, as the first one arises as a limiting case of the optimization problem (11) when λ tends to 0^+.

References

Aggarwal, C.C., Reddy, C.K. (eds.): Data Clustering: Algorithms and Applications. Chapman & Hall/CRC Press, London (2014)

Aldstadt, J.: Spatial clustering. In: Fischer, M.M., Getis, A. (eds.) Handbook of Applied Spatial Analysis, pp. 279–300. Springer, Berlin, Heidelberg (2010)

Arto, I., Dietzenbacher, E., Rueda-Cantuche, J.M.: Measuring bilateral trade in terms of value added. European Commission Joint Research Centre (JRC) Technical report. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC116694/jrc116694_manuscript_2019-05-28_online.pdf (2019)

Athey, S., Imbens, G.W.: Machine learning methods that economists should know about. Ann. Rev. Econ. **11**, 685–725 (2019)

Bhattacharya, T., Bhandari, B., Bairagya, I.: Where are the jobs? Estimating skill-based employment linkages across sectors for the Indian economy: an input–output analysis. Struct. Change Econ. Dyn. **53**, 292–308 (2020)

Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)

Carvalho, V.: Aggregate fluctuations and the network structure of intersectoral trade. Working Papers-Universitat Pompeu Fabra. Departamento de Economía y Empresa, 1206(1) (2009)

Cerina, F., Zhu, Z., Chessa, A., Riccaboni, M.: World input-output network. PLoS ONE **10**(7), e0134025 (2015)

Chen, G.Q., Wu, X.D., Guo, J., Meng, J., Li, C.: Global overview for energy use of the world economy: household-consumption-based accounting based on the world input-output database (WIOD). Energy Econ. **81**, 835–847 (2019)

Dietzenbacher, E., Los, B., Stehrer, R., Timmer, M., De Vries, G.: The construction of world input–output tables in the WIOD project. Econ. Syst. Res. **25**(1), 71–98 (2013)

Fagiolo, G., Reyes, J., Schiavo, S.: On the topological properties of the world trade web: a weighted network analysis. Physica A **387**(15), 3868–3873 (2008)

Fernandez-Vazquez, E.: A generalized cross entropy formulation for matrix balancing with both positive and negative entries (2016, unpublished)

Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, New York (2015)

Iori, G., De Masi, G., Precup, O.V., Gabbi, G., Caldarelli, G.: A network analysis of the Italian overnight money market. J. Econ. Dyn. Control **32**(1), 259–278 (2008)

Lance, G.N., Williams, W.T.: A general theory of classificatory sorting strategies. 1. Hierarchical systems. Comput. J. **9**(4), 373–380 (1967)

Lenzen, M., Kanemoto, K., Moran, D., Geschke, A.: Mapping the structure of the world economy. Environ. Sci. Technol. **46**(15), 8374–8381 (2012)

Leontief, W. (ed.): Input–Output Economics. Oxford University Press, New York (1986)

Li, C., Zhou, H.S.: Singular value thresholding in MATLAB. J. Stat. Softw. (2017). https://doi.org/10.18637/jss.v081.c02

Liang, S., Qi, Z., Qu, S., Zhu, J., Chiu, A.S., Jia, X., Xu, M.: Scaling of global input–output networks. Physica A **452**, 311–319 (2016)

MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Sympos. Math. Stat. Probab. **1**(14), 281–297 (1967)

Mazumder, R., Hastie, T., Tibshirani, R.: Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. **11**, 2287–2322 (2010)

McNerney, J., Fath, B.D., Silverberg, G.: Network structure of inter-industry flows. Physica A **392**(24), 6427–6441 (2013)

Metulini, R., Riccaboni, M., Sgrignoli, P., Zhu, Z.: The indirect effects of foreign direct investment on trade: a network perspective. World Econ. **40**(10), 2193–2225 (2017)

Moitra, A.: Algorithmic Aspects of Machine Learning. Cambridge University Press, Cambridge (2018)

Murtagh, F., Legendre, P.: Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? J. Classif. **31**, 274–295 (2014)

Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: optimal bounds with noise. J. Mach. Learn. Res. **13**(1), 1665–1697 (2012)

Nguyen, L.T., Kim, J., Kim, S., Shim, B.: Localization of IoT networks via low-rank matrix completion. IEEE Trans. Commun. **67**(8), 5833–5847 (2019)

Nguyen, L.T., Kim, J., Shim, B.: Low-rank matrix completion: a contemporary survey. IEEE Access **7**, 94215–94237 (2019)
OECD, Input–Output Tables. For download at http://oe.cd/i-o. Organisation for Economic Co-operation and Development, Paris (2018)
Oliva, G., Setola, R., Panzieri, S.: Critical clusters in interdependent economic sectors: a data-driven spectral clustering analysis. Eur. Phys. J. Spec. Top. **225**, 1929–1944 (2016)
Pavia, J.M., Cabrer, B., Sala, R.: Updating input–output matrices: assessing alternatives through simulation. J. Stat. Comput. Simul. **79**(12), 1467–1482 (2009)
Percoco, M., Hewings, G., Senn, L.: Structural change decomposition through a global sensitivity analysis of input–output models. Econ. Syst. Res. **18**(2), 115–131 (2006)
Rémond-Tiedrez, I., Rueda-Cantuche, J.M.: Full international and global accounts for research in input–output analysis (FIGARO). Eurostat (2019)
Revelle, W.: Hierarchical cluster analysis and the internal structure of tests. Multivar. Behav. Res. **14**(1), 57–74 (1979)
Riccaboni, M., Wang, X., Zhu, Z.: Firm performance in networks: the interplay between firm centrality and corporate group size. J. Bus. Res. **129**(C), 641–653 (2021)
Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
Rueda-Cantuche, J.M., Amores, A.F., Beutel, J., Remond-Tiedrez, I.: Assessment of European use tables at basic prices and valuation matrices in the absence of official data. Econ. Syst. Res. **30**(2), 252–270 (2018)
Sgrignoli, P., Metulini, R., Schiavo, S., Riccaboni, M.: The relation between global migration and trade networks. Physica A **417**, 245–260 (2015)
Thorndike, R.L.: Who belongs in the family? Psychometrika **18**(4), 267–276 (1953)
Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Methodol.) **63**(2), 411–423 (2001)
Timmer, M.P., Dietzenbacher, E., Los, B., Stehrer, R., De Vries, G.J.: An illustrated user guide to the world input–output database: the case of global automotive production. Rev. Int. Econ. **23**(3), 575–605 (2015)
Timmer, M.P., Los, B., Stehrer, R., De Vries, G.J.: An anatomy of the global trade slowdown based on the WIOD 2016 release. GGDC research memorandum number 162, University of Groningen, available online at the following hyperlink. https://www.rug.nl/ggdc/html_publications/memorandum/gd162.pdf (2016)
Tukker, A., De Koning, A., Wood, R., Hawkins, T., Lutter, S., Acosta, J., Rueda Cantuche, M., Bouwmeester, M., Oosterhaven, J., Drosdowski, T., Kuenen, J.: EXIOPOL-development and illustrative analyses of a detailed global MR EE SUT/IOT. Econ. Syst. Res. **25**(1), 50–70 (2013)
United Nations, Handbook on supply and use tables and input–output tables with extensions and applications (2018)
Valderas-Jaramillo, J.M., Rueda-Cantuche, J.M., Olmedo, E., Beutel, J.: Projecting supply and use tables: new variants and fair comparisons. Econ. Syst. Res. **31**(3), 423–444 (2019)
Valderas-Jaramillo, J.M., Rueda-Cantuche, J.M., Beutel, J.: The Euro and SUT-RAS methods: some further considerations. Econ. Syst. Res. **33**(2), 276–286 (2021)
Wang, H., Wang, C., Zheng, H., Feng, H., Guan, R., Long, W.: Updating input–output tables with benchmark table series. Econ. Syst. Res. **27**(3), 287–305 (2015)
Wang, H., Ding, L., Guan, R., Xian, Y.: Effects of advancing internet technology on Chinese employment: a spatial study of inter-industry spillovers. Technol. Forecast. Soc. Chang. **161**, 120259 (2020)
Wen, S., Xu, F., Wen, Z.: Robust linear optimization under matrix completion. Sci. China Math. **57**(4), 699–710 (2014)
Xu, M., Liang, S.: Input-output networks offer new insights of economic structure. Physica A **527**, 121178 (2019)
Xu, F., Lin, C., He, G., Wen, Z.: Nonnegative matrix completion for life-cycle assessment and input-output analysis. CER working paper, available online at the following hyperlink: http://www.cer.sdu.edu.cn/info/1093/3013.htm (2014)
Yonemoto, K.: Changes in the input–output structures of the six regions of Fukushima, Japan: 3 years after the disaster. J. Econ. Struct. **5**(1), 2 (2016)

Zhu, Z., Puliga, M., Cerina, F., Chessa, A., Riccaboni, M.: Global value trees. PLOS ONE **10**(5), e0126699 (2015)
Zhu, Z., Morrison, G., Puliga, M., Chessa, A., Riccaboni, M.: The similarity of global value chains: a network-based measure. Netw. Sci. **6**(4), 607–632 (2018)