Investigating singular value decomposition as a tool for data management in tourism

Kekez, Ivan; Garbin Praničević, Daniela

Proceedings SOR

Rupnik V. and L. Bogataj (Editors): The 1st Symposium on Operational Research, SOR'93. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1993, 310 pp.

Rupnik V. and M. Bogataj (Editors): The 2nd International Symposium on Operational Research in Slovenia, SOR'94. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1994, 275 pp.

Rupnik V. and M. Bogataj (Editors): The 3rd International Symposium on Operational Research in Slovenia, SOR'95. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1995, 175 pp.

Rupnik V., L. Zadnik Stirn and S. Drobne (Editors.): The 4th International Symposium on Operational Research in Slovenia, SOR'97. Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1997, 366 pp. ISBN 961-6165-05-4.

Rupnik V., L. Zadnik Stirn and S. Drobne (Editors.): The 5th International Symposium on Operational Research SOR '99, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 1999, 300 pp. ISBN 961-6165-08-9.

Lenart L., L. Zadnik Stirn and S. Drobne (Editors.): The 6th International Symposium on Operational Research SOR '01, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2001, 403 pp. ISBN 961-6165-12-7.

Zadnik Stirn L., M. Basti and S. Drobne (Editors): The 7th International Symposium on Operational Research SOR'03, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2003, 424 pp. ISBN 961-6165-15-1.

Zadnik Stirn L. and S. Drobne (Editors): The 8th International Symposium on Operational Research SOR'05, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2005, 426 pp. ISBN 961-6165-20-8.

Zadnik Stirn L. and S. Drobne (Editors): The 9th International Symposium on Operational Research SOR'07, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2007, 460 pp. ISBN 978-961-6165-25-9.

Zadnik Stirn L., J. Žerovnik, S. Drobne and A. Lisec (Editors): The 10th International Symposium on Operational Research SOR'09, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2009, 604 pp. ISBN 978-961-6165-30-3.

Zadnik Stirn L., J. Žerovnik, J. Povh, S. Drobne and A. Lisec (Editors): The 11th International Symposium on Operational Research SOR'11, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2011, 358 pp. ISBN 978-961-6165-35-8.

Zadnik Stirn L., J. Žerovnik, J. Povh, S. Drobne and A. Lisec (Editors): The 12th International Symposium on Operational Research SOR'13, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2013, 390 pp. ISBN 978-961-6165-40-2.

Zadnik Stirn L., J. Žerovnik, M. Kljaji Borštnar, S. Drobne (Editors): The 13th International Symposium on Operational Research SOR'15, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2015, 559 pp. ISBN 978-961-6165-45-7.

Zadnik Stirn L., J. Žerovnik, M. Kljaji Borštnar, S. Drobne (Editors): The 14th International Symposium on Operational Research SOR'17, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2017, 567 pp. ISBN 978-961-6165-50-1.

Zadnik Stirn L., M. Kljajl Borštnar, J. Žerovnik, S. Drobne, J. Povh (Editors): The 15th International Symposium on Operational Research SOR'19, Proceedings. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2019, 618 pp. ISBN 978-961-6165-55-6.

ISBN 978-961-6165-57-0

9 789616 165570

Proceedings SOR '21

Proceedings of the 16th International Symposium on OPERATIONAL RESEARCH in Slovenia

SOR '21

September 22-24, 2021

Edited by:

S. Drobne • L. Zadnik Stirn • M. Kljajić Borštnar • J. Povh • J. Žerovnik

SOR '21 Proceedings

The 16th International Symposium on Operational Research in Slovenia

September 22 - 24, 2021, Online

Edited by:

S. Drobne, L. Zadnik Stirn, M. Kljajić Borštar, J. Povh and J. Žerovnik

Slovenian Society INFORMATIKA (SDI)
Section for Operational Research (SOR)

Proceedings of the 16th International Symposium on Operational Research in Slovenia, SOR'21 in Slovenia, September 22 - 24, 2021, Online.

Organiser: Slovenian Society INFORMATIKA – Section for Operational Research, SI-1000 Ljubljana, Litostrojska cesta 54, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

Co-organiser: University of Maribor, Faculty of Organizational Sciences, SI-4000 Kranj, Kidričeva cesta 55a, Slovenia (http://www.fov.um.si/)

Co-organiser: University of Ljubljana, Faculty of Mechanical Engineering, SI-1000 Ljubljana, Aškerčeva cesta 6, Slovenia (https://www.fs.uni-lj.si/)

First published in Slovenia in 2021 by Slovenian Society INFORMATIKA – Section for Operational Research, SI 1000 Ljubljana, Litostrojska cesta 54, Slovenia (www.drustvo-informatika.si/sekcije/sor/)

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

519.8(082)
519.8:005.745(082)
519.81:519.233.3/.5(082)

INTERNATIONAL Symposium on Operational Research in Slovenia (16 ; 2021 ; online)
SOR '21 proceedings : the 16th International Symposium on Operational Research in Slovenia : September 22 - 24, 2021, online / [organiser] Slovenian Society Informatika (SDI), Section for Operational Research (SOR), [co-organiser University of Maribor, Faculty of Organizational Sciences [and] University of Ljubljana, Faculty of Mechanical Engineering] ; edited by S. Drobne ... [et al.]. - Ljubljana : Slovenian Society Informatika, Section for Operational Research, 2021

ISBN 978-961-6165-57-0
COBISS.SI-ID 75727107

Proceedings of the 16th International Symposium on Operational Research in Slovenia (SOR'21) is cited in: ISI (Index to Scientific & Technical Proceedings on CD-ROM and ISI/ISTP&B online database), Current Mathematical Publications, Mathematical Review, MathSci, Zentralblatt für Mathematic / Mathematics Abstracts, MATH on STN International, CompactMath, INSPEC, Journal of Economic Literature

Technical editor: Samo Drobne
Designed by: Samo Drobne
Printed by: BISTISK d.o.o., Ljubljana, Slovenia
Number of copies printed: 160

Reviewers:

Wellington Alves, Business School, Instituto Politécnico de Viana do Castelo; ALGORITMI Research Centre, Universidade do Minho, Portugal

Josip Arnerić, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia

Marija Bogataj, University of Ljubljana, Faculty of Economics, and INRISK, Slovenia

Marko Bohanec, Jožef Stefan Institute, Department of Knowledge Technologies, Ljubljana, Slovenia

D. Bokal, University of Maribor, Faculty of Natural Sciences and Mathematics, and DataBitLab d.o.o., Maribor, Slovenia

Katarína Cechlárová, P. J. Šafarik University, Faculty of Science, Košice, Slovakia

Vesna Čančer, University of Maribor, Faculty of Economics and Business

Ćukušić Maja, University of Split, Faculty of Business, Economics and Tourism, Split, Croatia

Samo Drobne, University of Ljubljana, Faculty of Civil and Geodetic Engineering, Ljubljana, Slovenia

Ksenija Dumičić, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia

Eisenberg-Nagy Marianna, Corvinus University of Budapest, Corvinus Centre for Operations Research

Helena Gaspars-Wieloch, Poznan University of Economics and Buinsess, Poznan, Poland

Eloy Hontoria, Technical University of Cartagena, Business Management Department, Cartagena, Spain

Tibor Illés, Corvinus University of Budapest, Corvinus Centre for Operations Research

Mario Jadrić, University of Split, Faculty of Business, Economics and Tourism, Split, Croatia

Janja Jerebic, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia

Veljko Jeremić, University of Belgrade, Faculty of Organizational Sciences, Belgrade, Serbia

Sandi Klavžar, University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia

Mirjana Kljajić Borštnar, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia

Andrej Košir, University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

Vasja Leban, University of Ljubljana, Biotechnical Faculty, Ljubljana, Slovenia

Zrinka Lukač, University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia

Karmen Pažek, University of Maribor, Faculty of Agriculture and Life Sciences

Mirjana Pejić Bach, Faculty of Economics & Business, University of Zagreb, Croatia

Tunjo Perić, University of Zagreb, Faculty Economics and Business, Department of Informatics, Zagreb, Croatia

Snježana Pivac, University of Split, Faculty of Economics, Split, Croatia

Janez Povh, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia

Uroš Rajkovič, University of Maribor, Facutly of Organizational Sciences, Kranj, Slovenia

Vladislav Rajkovič, University of Maribor, Facutly of Organizational Sciences, Kranj, Slovenia

Helena Sofia Rodrigues, Business School, Instituto Politécnico de Viana do Castelo; Center for Research and Development in Mathematics and Applications (CIDMA), Universidade de Aveiro

Črtomir Rozman, University of Maribor, Faculty of Agriculture and Life Sciences

Darja Rupnik, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia

Ângela Silva, Business School, Instituto Politécnico de Viana do Castelo; COMEGI, Universidade Lusíada, Portugal

Jurij Šilc, Jožef Stefan Institute, Computer Systems Department, Ljubljana, Slovenia

Tadeusz Trzaskalik, University of Economics, Department of Operational Research Katowice, Poland

Lidija Zadnik Stirn, University of Ljubljana, Biotechnical Faculty, Ljubljana, Slovenia

Anja Žnidaršič, University of Maribor, Faculty of Organizational Sciences, Kranj, Slovenia

Janez Žerovnik, University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia

Chairs:

Preface

This volume, Proceedings of the 16th International Symposium on Operational Research, called SOR'21, contains papers presented at SOR'21 (https://sor.fov.um.si/), organised by Slovenian Society INFORMATIKA (SDI), Section for Operational Research (SOR), University of Maribor, Faculty of Organisational Sciences, Kranj, Slovenia (FOV), and University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia (UL FS). The SOR'21 symposium, held 22-24 September 2021, was originally planned to take place in Bled, Slovenia, but was moved online due to the situation of COVID-19 in Slovenia and beyond. The volume contains blind peer-reviewed papers or abstracts of papers presented at the symposium.

The opening address at SOR'21 was given by Prof. Dr. Lidija Zadnik Stirn, President of SOR, Mr. Niko Schlamberger, President of SDI, representatives of FOV and UL FS, Prof. Dr. Mario Jadrić, President of Croatian Operational Research Society (CRORS), Dr Sarah Fores, manager of The Association of European Operational Research Societies (EURO), and presidents/representatives of some others Operational Research Societies from abroad.

SOR'21 is the scientific event in the field of Operational Research, another in the traditional series of biennial international OR conferences organised in Slovenia by SDI-SOR. It is the continuation of fifteen previous symposia. The main objective of SOR'21 is to promote knowledge, interest and education in the field of OR in Slovenia, Europe and worldwide in order to build the intellectual and social capital essential for maintaining the identity of OR, especially at a time when interdisciplinary cooperation is proclaimed as particularly important for solving problems in today's challenging times. By joining IFORS and EURO, the SDI-SOR has also agreed to collaborate with different disciplines, i.e., to balance the depth of theoretical knowledge in OR and the understanding of theory, methods, and problems in other fields within and outside OR. We believe that SOR'21 creates the advantage of these goals, contributes to the quality and reputation of OR by presenting and sharing new developments, opinions and experiences in the theory and practise of OR.

SOR'21 was highlighted by five distinguished keynote speakers. The first part of Proceedings SOR'21 contains invited abstracts, presented by five outstanding scientists: Assist. Prof. Nikolina Ban, University of Innsbruck (UIBK), Department of Atmospheric and Cryospheric Sciences, Innsbruck, Austria, Assist. Prof. Vedran Kojić, University of Zagreb, Faculty of Economics & Business, Zagreb, Croatia, Prof. Panos Patrinos, KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Leuven, Belgium, Prof. Suresh P. Sethi, Eugene McDermott Chair Professor of Operations Management, Director, Center of Intelligent Supply Networks, Naveen Jindal School of Management, The University of Texas at Dallas, Dallas, USA, and Prof. Jerneja Žganec Gros, Alpineon Ltd, Ljubljana, Slovenia.

The Proceedings includes 118 papers or abstracts by 240 authors. Most of the authors of the contributed papers came from Slovenia (82), then Croatia (52), Hungary (23), Portugal (23), Serbia (17), Poland (9), Czech Republic (8), Slovak Republic (7), Spain (6), Netherlands (4), Bosnia and Herzegovina (2), Austria (1), Belgium (1), France (1), Germany (1), Romania (1), Ukraine (1), United Kingdom (1), and United States of Amerika (1). The papers published in the Proceedings are divided into Plenary Lectures (5 abstracts), eleven special sessions: Application of Operational Research in Smart Cities (6 papers), Computational Mathematical Optimization (7 papers and 6 abstracts), Data Science – Methodologies and Case Studies (10 papers), Graph Theory and Algorithms (2 papers), *High-Performance Computing and Big Data (3 papers), Industry & Society 5.0: Optimization in Industrial and Human Environments (6 papers), International Projects in Operations Research (2 papers), Lessons Learned from the COVID-19 Pandemic: Applications of Statistical and OR Methods (8 papers), Logistics and Sustainability (9 papers), Operational Research in Ageing Studies and Social Innovations (5 papers), Operations Research in Agricultural Economics and Farm Management (5 papers), and eight sessions: Econometric Models and Statistics (6 papers), Environment and Social Issues (5 papers), Finance and Investments (6 papers), Location and Transport, Graphs and their Applications (5 papers), Mathematical Programming and Optimization (5 papers and 1 abstract), Multi-Criteria Decision-Making (10 papers), Theory of Games (3 papers), and Problems Approaching OR (3 papers).*

Proceedings of the previous fifteen International Symposia on Operational Research organised by the Slovenian Section on Operational Research, listed at https://www.drustvo-informatika.si/sekcije/sor/sor-publikacijepublications/, are indexed in the following secondary and tertiary publications: Current Mathematical Publications, Mathematical Review, Zentralblatt fuer Mathematik/ Mathematics Abstracts, MATH on STN International and CompactMath, INSPEC. It is expected that Proceedings SOR'21 will be covered by the same bibliographic databases.

The success of the scientific events at SOR'21 and of the present conference proceedings should be seen because of joint efforts. On behalf of the organisers, we would like to express our sincere gratitude to all those who assisted us in the preparation of the event. Without the dedicated and advice of the active members of the Slovenian Operations Research Section, we would not have been able to attract so many top-class speakers from all over the world. Many thanks to them. In addition, we would like to express our deepest gratitude to the prominent keynote speakers, the members of the Programme and Organising Committees, the reviewers who improved the quality of SOR'21 with their useful suggestions, the section chairs and all the numerous people - far too many to list individually here - who helped in organizing of the 16th International Symposium on Operational Research SOR'21 and compiling this proceedings. Finally, we thank the authors for their efforts in preparing and presenting the papers that made the 16th Symposium on Operational Research SOR'21 a success.

We would like to give special thanks to the Partnership for Advanced Computing in Europe (PRACE) for their financial support.

Ljubljana and Kranj, September 22, 2021

Samo Drobne
Lidija Zadnik Stirn
Mirjana Kljajić Borštnar
Janez Povh
Janez Žerovnik
(Editors)

INVESTIGATING SINGULAR VALUE DECOMPOSITION AS A TOOL FOR DATA MANAGEMENT IN TOURISM

Ivan Kekez and **Daniela Garbin Praničević**
University of Split, Faculty of Economics, Business and Tourism
CviteFiskovica 5, 21000 Split, Croatia
E-mail: ivank3kez@gmail.com, daniela@efst.hr

Abstract: This paper contains a brief description of a singular value decomposition method as a tool for data management and performance improvement in the context of tourism activities – online hotel ratings. Throughout the paper, the authors introduced elementary linear theory background and SVD mathematical algorithm in a simplified way in order to express its contribution to the analytical value of data. Demonstrated algorithm and achieved results indicate two decisions. To perform high compression despite potential analytical and misinterpretation risks due to the details loss or keeping the data volume, only with minimal reduction for a largely dependent, false, and outlier data.

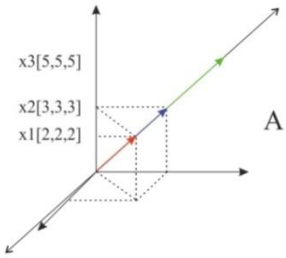
Keywords: singular value decomposition, tourism, data management, dimensionality reduction

1 INTRODUCTION

The continuous rise of digital technologies in everyday activities has led to an exponential volume increase of digital data [29], drawing attention to big data terminology [13]. Generally, big data can be defined across three dimensions: Volume, Velocity, Variety [24], with volume as its main characteristic [9]. Along with an increased volume, various data types started to appear, making it challenging to analyze. Some studies pointed out key challenges such as analytical complexities, data storage, and data management [6][33]. Over the years, the impact of big data also grew on tourism due to the influential role of customers' online activities [10]. Hotels and other tourism and hospitality companies become closer to the big data, considering helping them generate more revenue and deliver a better experience for customers. Moreover, the ability to manage a large amount of data means an option for value creation from big data and its supporting technologies, which justifies considering hotel reviews in a smart tourism context. It is often used for predictive and behavioral analytics to assist companies in recognizing the patterns related to their business practices [17]. But occasionally, defining relations between data, information, and knowledge is not always fully transparent. Knowledge is often seen as an organized data structure that creates information uniquely observed by specific users due to its experiences and practices [2]. Therefore, depending on the data source, type, and user perception, data receive different traits that define its meaning [1]. The accepted view of data is a fact that becomes valuable as it is formed into a structure that creates high-quality information [30]. The quoted definition is accepted regardless of differences between scientific and practical disciplines, but at the same time doesn't answer questions about what data structure is and how it is manageable. For that matter, it is essential to add a suitable numerical representation, a collection of individual numerical values that accurately describes the data [31]. Correspondingly, this paper aims at presenting the simplified mathematical view on data management as the numerical presentation of elemental pieces of information, in this case, the hotel reviews.

2 MATRIX STORAGE AND DIMENSIONALITY REDUCTION

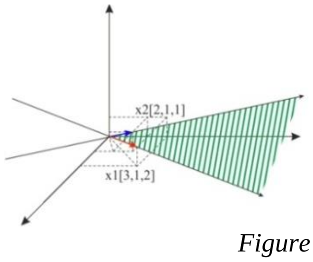
Every numerical analysis starts with a single number, a numerical representation of data with the corresponding value called a scalar [12]. An array of scalars creates a vector where each value is in order according to the corresponding index. As the scalar index in an array defines the vector, the corresponding index of vectors describes the matrix structure ordered in m rows and *n* columns of unit vectors [31]. A more practical way to think about matrix is to view it as a data set, where columns and rows represent particular data attribute and each attribute make its dimension. Dimensionality reduction within stored data aims at reducing the number of data attributes for future analytical examination [8]. But before reviewing data reduction, it is necessary to analyze the term dimension. Generally speaking, term dimension can be associated with mathematical space and related objects positioned in correspondence with spatial coordinates needed to specify its location. From a data point of view, a dimensional „feature“ of a scalar is determined by its data attributes that create dimension [22]. The dimensionality of a data set filled with scalar values is the minimum number of attributes needed for data representation without significant information loss [3]. With additional added dimensions, data is stored in multidimensional tensor-based structures [32]. From the tensor perspective, a vector with two scalar values is a principal component of the first-rank tensor in a two-dimensional space [25]. A vector with three scalar values requires a third space dimension - a second-rank tensor. An example of a typical additional dimension for a large proportion of stored data within numerous databases is time series [4]. Depending on the analytical approach, the structural rank of a matrix can be viewed as the maximum achievable rank among all matrices taken into an analytical procedure with the same structure but different scalar values [23], as a number of its positive singular values [20] or as a number of linearly independent column vectors of the matrix [28]. The simplest forms of the data structures are one-dimensional matrices, where each of the scalar values is positioned on the component line in space (*Figure 1*).



$$A = \begin{bmatrix} 2 & 3 & 5 \\ 2 & 3 & 5 \\ 2 & 3 & 5 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = x1 \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} + x2 \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} + x3 \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix}$$

Figure 1: Matrix A rank 1

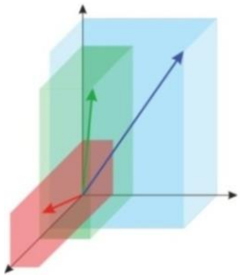
For each matrix transformation from an array of scalars *x1*, *x2*, *x3*, newly calculated values will be on the hypothetical line – rank one. Just by looking at the presented example, the simplicity of the formed matrix can't be unnoticed. However, most of the time, connections between values within data sets aren't noticeable and acquire more data aggregation. In order to show that the maximum number of column vectors of the matrix does not always define its structural rank and that linear dependence can be more challenging to see, a simple matrix A *rank 2* (*mxn* = 3x3) is presented in the example below.



$$A = \begin{bmatrix} 3 & 2 & 5 \\ 1 & 1 & 2 \\ 2 & 1 & 3 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = x1 \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} + x2 \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} + x3 \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix}$$

Figure 2: Matrix A rank 2

If we examine matrix A (*Figure 2*), at first glance, it could look like the scalar values within the matrix are randomized. If that is true, it would mean that the structural rank of the matrix A equals 3. But in this case, this is not true because there is linear dependence of the third column vector as a sum of the scalar values of first and second vectors. Thus, a presence of linear dependence among column vectors suggests that stored values can lower their dimensionality level [31]. For each matrix transformation from an array of numbers *x1*, *x2*, *x3*, newly calculated values will be on the hypothetical two-dimensional plane because the third matrix column is linearly dependent as a sum of the first two column vectors. In statistical terms, there is a multicollinearity problem of the third column vector, so its values are predictable as a combination of the first and second column vector - the column has no analytical value [7]. If none of the vectors are a linear combination of other vectors, the matrix is linear independent [12].



$$A = \begin{bmatrix} 1 & 3 & 1 \\ 2 & 5 & 3 \\ 5 & 2 & 2 \end{bmatrix} \begin{bmatrix} x1 \\ x2 \\ x3 \end{bmatrix} = x1 \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix} + x2 \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix} + x3 \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

Figure 3: Matrix A rank 3

In the examples above, the intention was to present data management in two-dimensional matrices. Although two-dimensional structures are essential for data management, it is necessary to exclude the possibility that all data is: (1) stored in square matrix where $m=n$; (2) low dimensional order; (3) relations are easily manageable. In addition to that, the authors introduced a standard singular value decomposition (SVD) model as one of the most universal and fundamental mathematical methods for dimensionality reduction and data management. Although some noted studies point out limitations for a particular field [21][5][16], authors consider the SVD model essential for understanding data management and manipulation, especially in this study format, as a straightforward introduction to raw data management.

3 SINGULAR VALUE DECOMPOSITION

Theoretical analysis and data experimentation suggested that the theory of matrix decomposition is one of the most beneficial ideas for data management. Therefore, it is not surprising that matrix decomposition is a fundamental discipline of linear algebra for various analytical problems within collected and stored data [27]. Dimensionality reduction aims to lower dimensionality space without significant loss of information [19]. With dimensionality reduction, practicing the matrix decomposition algorithm reduce the time complexity, memory usage and potentially increase accuracy of data analysis [16]. Although SVD is a well-known technique in various fields across computer science, machine learning, and more, social scientists are often very familiar with the value of observed results [15]. The equation for SVD of a given $m \times n$ matrix A with scalar values for structural rank r in $\boldsymbol{m \geq n}$ matrix structure is a factorization of three matrices [11]:

$$A = U \Sigma V^{t} \tag{1}$$

where diagonal $r \times r$ matrix $\sum$ contains singular values $\partial 1, \partial 2, \ldots, \partial n \geq 0$ valued from highest to lowest, matrix U contains left singular vectors in m dimension, and matrix V *transpose* contains right singular vectors in n dimension. If the observed matrix is a square, its factorized matrices will be orthogonal. An orthogonal matrix is a square matrix whose columns and rows values are mutually orthonormal containing eigenvectors for $A^t A$ and AA^t [26]. Calculated eigenvalues on diagonal of orthogonal matrix $\sum$ are singular values of matrix factorization for $\boldsymbol{m \neq n}$**,** and their index presents the structural rank of matrix A.

4 RESULTS AND DISCUSSION

In this section, the authors examined an available dataset to demonstrate matrix factorization on simple data matrix examples [14]. Out of available data attributes, the focus was on object categorization: *hotels*, *name of the object*, *username*, and *hotel rating*.

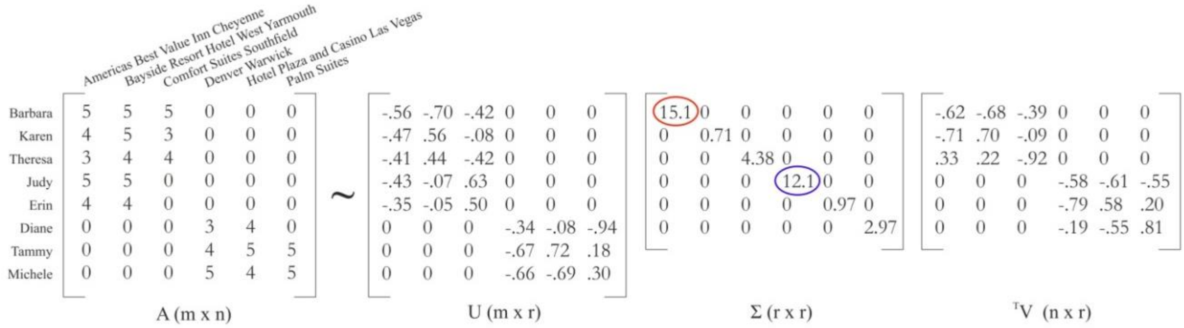


Figure 4: Singular Value Decomposition of matrix A

Important to state is that all null (*empty*) and false values in matrix A are assumed to be zero (0) valued. Zero is not an acceptable user rating for data attributes in the range 1 - 5, and so it displays replacement value. The obtained values on the diagonal of the $\sum$ matrix show the strength of the relationship between values of the matrix A attributes - singular vectors. Vector on coordinates $u_1\partial_1 v_1 t$ contains the highest singular value equal to 15.11. The vector is called *the First Principal Component* [34] and indicates the line that best defines data observations of the matrix A (values have the lowest measured distance projection on component line) – contains the highest analytical value. The structural rank of the matrix represented with only one vector equals one (*see Figure 1*). In the example above (*Figure 4*), there are six positive singular values and question how many dimensions are possible to discard (remove singular values) without decreasing the analytical value of the matrix. A lower percentage of data retained by a combination of singular vectors is at least 90% of the sum of all squared singular values [18]. In our example, a sum of all squared singular values equals $15{,}11^2+12{,}13^2+4{,}38^2+2{,}97^2+0{,}97^2+0{,}71^2=$**405**. Removing singular values from the equation in ascending order results in: **404.50; 394.73; 375.59**. By keeping principal component lines written on coordinates $u_1\partial_1 v_1 t$ and $u_4\partial_4 v_4 t$ results in 92,74% of the sum of squared singular values (375,59/405*100), which is larger than the proposed minimum of 90%. Therefore, reducing the dimensionality of matrix A on rank two (2) (*see Figure 2*) caused a loss of 7,26% stored data – approximation of matrix A (*matrix A'*).

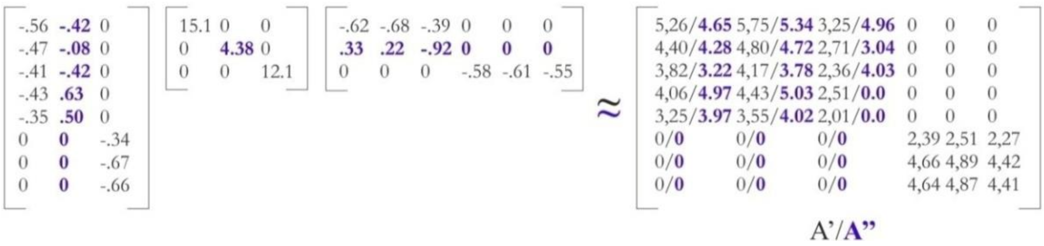


Figure 5: Approximation of matrix A (A' & A")

However, if we do an inverse calculation in the approximate values of matrix A, we can notice deviations in the data values of matrix A' compared to matrix A (*Figure 5*). Although the vectors $u_1\partial_1 v_1 t$ and $u_4\partial_4 v_4 t$ represent a large part of the values of matrix A, the vector $u_3\partial_3 v_3 t$ is significantly related to its coordinated values. The reason is that the selected attributes from the sample are randomized, and although it may seem that there is a linear dependence between the elements, that cannot be viewed as a rule for all stored data. In other words, hotel rating values alone are not sufficient to describe possible attribute connections.

Additionally, adding the third singular value - 4.38 in an inverse calculation (matrix A"), combined with the previous singular values on the $\sum$ matrix diagonal, covers 97.46% of the sum of squared singular values (new rank of the matrix is equal to three 3; see *Figure 3*). By using the formula for the data compression ratio [26] and new structural rank of matrices A 'and A", it is possible to calculate the percentage of data reduction and approximately assume the analytical data value:

$$CR = \frac{m * n}{(m + n + 1) * r} \tag{2}$$

where: m – rows; n – columns; r – singular values of new matrix. For matrix A' expression is: $CR(A') = (8 * 6)/((8 + 6 + 1) * 2) = 6.4{:}\,1 \approx 15.63\%$ is a volume of compressed data which means that 92,74% of data can be retained by keeping the principal component lines written on coordinates $u_1\partial_1 v_1 t$ i $u_4\partial_4 v_4 t$ – rank 2. For matrix A" expression is: $CR(A'') = (8 * 6)/((8 + 6 + 1) * 3) = 1.067{:}\,1 = 93.72\%$ is volume of compressed data which means that 97,46% of data can be retained by keeping the principal component lines written on coordinates $u_1\partial_1 v_1 t$, $u_3\partial_3 v_3 t$ i $u_4\partial_4 v_4 t$ - rank 3. Due to the size of matrix A" rank 3, there is no significant contribution in data compression of matrix A. On the other hand, there is a significant data reduction of matrix A - rank 2, but also it brings analytical risks and possible misinterpretation due to the data loss on $u_3\partial_3 v_3 t$ coordinates.

5 CONCLUSION

Practical implications of data management address many issues such as predicting tourist demand, supporting demand level anticipation, enabling better decision making, managing knowledge flows with customers, and providing the best service more efficiently and effectively. The presented SVD method is one of the fundamental mathematical algorithms essential for analyzing large amounts of data. Although pointed benefits of data management using SVD arise from this and other studies, there are also several limitations. Firstly, it is vital to have the highest possible transparency within data attributes due to the potential clusters within data values with similar qualitative features. Hotel ratings in their numerical form don't reveal enough information to provide definite conclusions about existing connections within numerical data. That is visible from dimensionality reduction, where data on third component line have significant analytical value no matter its low volume. Additional descriptive and cluster analysis for hotel objects and users would contribute more insights about connections within ratings, which would prove or disapprove its significance to remaining data projected on first and second component lines. Additionally, with a higher number of attributes and observations, the risk of efficiency failure rises, especially if there is low data transparency. Finally, SVD is a reliable tool for dimensionality reduction and graph analysis, but on the other hand, its complex algorithm can be overwhelming. Thus, the paper aimed to demonstrate a few practical SVD possibilities and enhance understanding of its value throughout simplistic form for the application in the smart tourism domain as a data management tool. For that purpose, an open data source served well in the demonstrative purpose of the study.

References

[1] Ancona, D. 2012. Sensemaking: Framing and Acting in the Unknown. In Snook, S, Nohria, N, Khurana, R. (Eds.). The Handbook for Teaching Leadership: Knowing, Doing, and Being (pp. 3-19). London: SAGE Publication, Inc.

[2] Bhatt, G.D. 2001. Knowledge management in organizations: examing the interaction between technologies, techniques, and people. Journal of Knowledge Management, 5(1): 68-75.

[3] Camastra, F. 2003. Data dimensionality estimation methods: a survey. Pattern Recognition, 36(12): 2945-2954.

[4] Chakrabarti, K., Keogh, E., Mehrotra, S., Pazzani, M. 2002. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. ACM Transactions on Database Systems, 27(2): 188-228.

[5] Cherry, S. 1996. Singular Value Decomposition Analysis and Canonical Correlation Analysis. Journal of Climate, 9(9): 2003-2009.

[6] Espinosa, J.A., Kaisler, S., Armour, F., Money, W.H. 2019. Big Data Redux: New Issues and Challenges Moving Forward. Hawaii International Conference on System Sciences.

[7] Farrar, D.E., Glauber, R.R. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics, 49(1): 92-107.

[8] Fu, X., Wang, L. 2003. Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, 33(3): 399-409.

[9] Gandomi, A., Haider, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2): 137-144.

[10] Gajdošik, T. 2019. Big Data Analytics in Smart Tourism Destinations. A New Tool for Destination Management Organizations?. In. Katsoni, V, Segarra-Ona, M, (Eds.). Smart Tourism as a Driver for Culture and Sustainability (pp. 15-33). Springer, Cham.

[11] Golub, G.H., Reinsch, C. 1970. Singular value decomposition and least squares solutions. Numerische Mathematik, 14(5): 403-420.

[12] Goodfellow, I., Bengio, Y., Courville, A. 2016. Deep Learning. Cambridge: The MIT Press.

[13] Jun, S.-P., Yoo, H. S., Choi, S. 2018. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change, 130: 69-87.

[14] Hotel Reviews. 2019. Kaggle. https://www.kaggle.com/datafiniti/hotel-reviews [30/04/2021].

[15] Kosinski, M., Wang, Y., Lakkaraju, H., Leskovac, J. 2016. Mining Big Data to Extract Patterns and Predict Real-Life Outcomes. Psychol Methods, 21(4): 493-506.

[16] Kuang, L., Hao, F., Yang, L.T., Lin, M., Luo, C., Min, G. 2014. A Tensor-Based Approach for Big Data Representation and Dimensionality Reduction. IEEE (TETC), 2(3): 280-291.

[17] Li, J., Xu, L., Tang, L., Wang, S, Li, L. 2018. Big data in tourism research: A literature review. Tourism Management, 68: 301-323.

[18] Leskovec, J., Rajaraman, A., Ullman, J.D. 2014. Mining of Massive Datasets. Cambridge University Press.

[19] L'Heureux, A., Grolinger, K., Elymany, H.F., Capretz, M. 2018. Machine Learning with Big Data: Challenges and Approaches. IEEE Access, 5: 7776-7797.

[20] Ma, S., Goldfarb, D., Chen, L. 2009. Fixed point and Bregman iterative methods for matrix rank minimization. Mathematical Programming, 128: 321-353 (2011).

[21] Newman, M., Sardeshmukh, P.D. 1995. A Caveat Concerning Singular Value Decomposition. Journal of Climate, 8(2): 352-360.

[22] Parsons, L., Haque, E., Liu, H. 2004. Subspace Clustering for High Dimensional Data: A Review. Acm siggkdd explorations newsletter, 6(1):90-105.

[23] Pinar, A., Chow, E., Pothen, A. 2006. Combinatorial Algorithms for Computing Column Space Bases That Have Sparse Inverses. Electronic Transactions on Numerical Analysis, 22: 122-145.

[24] Singh, S., Singh, N. 2012. Big Data Analytics. 2012 International Conference on Communication, Information & Computing Technology (ICCICT), 1-4.

[25] Spiegel, M.R. 1959. Vector Analysis and an introduction to tensor analysis. NY: McGraw Hill.

[26] Stacchini de Souza, J.C., Lessa Assis, T.M., Pal, B.C. 2017. Data Compression in Smart Distribution Systems via Singular Value Decomposition. IEEE Transactions on Smart Grid, 8(1): 275-284.

[27] Steward, G.W. 1993. On the Early History of the Singular Value Decomposition. SIREV, 35(4): 551-566.

[28] Strang, G. 2019. Linear algebra and Learning from data. MA Wellesley: Cambridge Press.

[29] Tanwar S., Ramani, T., Tyagi, S. 2018. Dimensionality Reduction Using PCA and SVD in Big Data: A Comparative Case Study. In Patel, Z, Gupta, S. (Eds.). ICFITT 2017: Future Internet Technologies and Trends (pp. 116-125). Springer, Cham.

[30] Tuomi, I. 1999. Data is More Than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. Journal of Management Informational Systems, 16(3): 107-121.

[31] Zhang, A., Lipton, Z.C., Li, M., Smola, A.J. 2020. Dive into Deep Learning. https://d2l.ai [30/04/2021].

[32] Zhang, Y., Chen, M., Mao, S., Hu, L., Leung, V.C.M. 2014. CAP: Community Activity Prediction Based on Big Data Analysis. IEEE Network, 28(4): 52-57.

[33] Wani, M.A., Jabin S. 2018. Big Data Issues, Challenges, and Techniques in Business Intelligence. In Aggarwal, V, Bhatnagar, V, Mishra, D. (Eds.). Big Data Analytics. Advances in Intelligent Systems and Computing (pp. 613-628). Singapore: Springer.

[34] Wold, S., Esbensen, K., Geladi, P. 1987. Principal Component Analysis. Chemometrics and Intelligent Laboraty Systems, 2: 37-52.