University of Groningen

A review of tensor-based methods and their application to hospital care data

Giordani, Paolo; Kiers, Henk A. L.

Published in:
Statistics in Medicine

DOI:
[10.1002/sim.7514](https://doi.org/10.1002/sim.7514)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Giordani, P., & Kiers, H. A. L. (2018). A review of tensor-based methods and their application to hospital care data. *Statistics in Medicine*, 37(1), 137-156. <https://doi.org/10.1002/sim.7514>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

RESEARCH ARTICLE

A review of tensor-based methods and their application to hospital care data

Paolo Giordani¹ | Henk A.L. Kiers²¹ Department of Statistical Sciences,
Sapienza University of Rome, Rome, Italy² Department of Psychology, University of
Groningen, Groningen, The Netherlands**Correspondence**Paolo Giordani, Department of Statistical
Sciences, Sapienza University of Rome,
Rome, Italy.

Email: paolo.giordani@uniroma1.it

In many situations, a researcher is interested in the analysis of the scores of a set of observation units on a set of variables. However, in medicine, it is very frequent that the information is replicated at different occasions. The occasions can be time-varying or refer to different conditions. In such cases, the data can be stored in a 3-way array or tensor. The Candecomp/Parafac and Tucker3 methods represent the most common methods for analyzing 3-way tensors. In this work, a review of these methods is provided, and then this class of methods is applied to a 3-way data set concerning hospital care data for a hospital in Rome (Italy) during 15 years distinguished in 3 groups of consecutive years (1892–1896, 1940–1944, 1968–1972). The analysis reveals some peculiar aspects about the use of health services and its evolution along the time.

KEYWORDS

Candecomp/Parafac, hospital care data, tensor-based methods, 3-way data, Tucker3

1 | INTRODUCTION

In many situations, a researcher is interested in the analysis of the scores of a set of observation units on a set of variables. Examples include gene expression data, the vital signs observed for some patients in a hospital department, the characteristics of some children measured by a pediatrician, the symptoms observed on a group of patients suffering from panic disorder by a psychiatrist, and the incidence of hospitalizations in a certain area distinguished by classes of age. In such cases, the available information can be stored in a standard matrix, say $\mathbf{X}$, the generic element of which, x_{ij} , expresses the score of the i -th observation ($i = 1, \dots, I$) with respect to the j -th variable ($j = 1, \dots, J$). More in detail, $\mathbf{X}$ is a 2-way 2-mode matrix where the 2 modes are the 2 sets of entities (observation units, variables) and the 2 ways are the 2 sets of indices ($i = 1, \dots, I$ and $j = 1, \dots, J$). The matrix $\mathbf{X}$ can equivalently be denoted as a tensor in $\mathfrak{R}^{I \times J}$ (a vector of length I is a tensor in $\mathfrak{R}^I$).

In medicine, it is very frequent that the information is replicated at different occasions. The occasions can be time-varying (eg, the vital signs of some patients recorded daily or the characteristics of some children collected monthly) or refer to different conditions (eg, the symptoms observed on a group of patients by more than 1 psychiatrist or the incidence of hospitalizations distinguished by not only classes of age but also hospital departments). The above examples are 3-way 3-mode. In fact, there is an additional mode (occasions), and the data are stored in a 3-way array or (3-way) tensor $\mathbf{X}$ in $\mathfrak{R}^{I \times J \times K}$ with generic element x_{ijk} expressing the score of observation unit i for variable j at occasion k ($k = 1, \dots, K$).

In the last years, the analysis of tensors has received a growing attention in the statistical literature.^{1–4} However, research on tensor-based methods has a long history. In fact, the first works on tensor decompositions were due to Hitchcock^{5,6} and Cattell.⁷ These findings were reconsidered leading to the so-called Candecomp/Parafac (CP)^{8,9} and Tucker3 (T3)¹⁰ methods. The CP and T3 methods represent the most common tensor-based methods. When analyzing 3-way tensors (or briefly “tensors” unless otherwise specified), the research interest is usually to summarize the huge amount of data by representing the relevant information by components. CP and T3 achieve this goal from an exploratory point

of view and, therefore, represent 3-way extensions of classical Principal Component Analysis (PCA). Differently from PCA, they allow to discover the 3-way interactions among the modes. This is done by detecting a limited number of components for every mode and investigating the existing interactions. Even though ordinary PCA can also be used to analyze a tensor by either performing separate PCA's (1 for every occasion) or collapsing the data (for instance, computing the average values over the occasions), the triple interactions then remain hidden and cannot be discovered.

This paper offers a review of tensor-based methods, with particular reference to the CP and T3 methods. The methods are illustrated by their application to a 3-way data set concerning hospital care. In particular, the data are about the health services observed in the hospital "Pio Istituto di Santo Spirito ed Ospedali Riuniti di Roma" in Rome, Italy. The data refer to 15 years distinguished in 3 groups of 5 years. The first era concerns the end of the 19th century (the years 1892–1896) and provides a picture of the Roman health status a few decades after the birth of the Italian nation (1861). The next era (1940–1944) is during the Fascism regime and the Second World War. Finally, the last era (1968–1972) corresponds to the end of the Italian economic boom. The available information allows us to appreciate how the health features evolved during the 20th century: some diseases characterizing the first era vanished or, at least, remarkably reduced their importance and some others emerged in the more recent years. Although the data refer to 1 hospital, they offer insight into the health status in Italy. In fact, as we shall see, some results can be better interpreted by taking into account what happened in Italy and not just in the city of Rome.

The paper is organized as follows. The next section contains a review of CP and T3. The results of the application of the methods to the previously described data set by using a freely available package for R are given in Section 3. Final remarks in Section 4 conclude the paper.

2 | TENSOR-BASED METHODS

In this section, the T3 and CP methods will be reviewed. These tools represent extensions of the classical PCA for higher order tensors. For convenience, PCA is briefly recalled. PCA fits the model

$$x_{ij} = \sum_{s=1}^S a_{is}b_{js} + e_{ij}, i = 1, \dots, I, j = 1, \dots, J, \quad (1)$$

where a_{is} and b_{js} are the component score of observation unit i (eg, a type of disease) for component s ($i = 1, \dots, I, s = 1, \dots, S$) and the component loading of variable j (eg, the mortality rate) for component s ($j = 1, \dots, J, s = 1, \dots, S$), respectively, S denoting the number of components, and e_{ij} is the error term. In matrix form (1) can be rewritten as

$$\mathbf{X} = \mathbf{AB}' + \mathbf{E}, \quad (2)$$

where $\mathbf{A}$ ($I \times S$) with generic element a_{is} and $\mathbf{B}$ ($J \times S$) with generic element b_{js} denote the component score and component loading matrices, respectively, and $\mathbf{E}$ ($I \times J$) is the error matrix. The data matrix $\mathbf{X}$ often contains standardized values. This is usually done by centering across the observations (rows) so that every variable has zero mean, and normalizing within the variables (columns) so that every variable has unit variance. It is desirable to find such estimates for $\mathbf{A}$ and $\mathbf{B}$ in (2) in such a way that as little information as possible is missed. To this end, PCA fits model (2) by minimizing

$$\|\mathbf{E}\|^2 = \|\mathbf{X} - \mathbf{AB}'\|^2 \quad (3)$$

with respect to $\mathbf{A}$ and $\mathbf{B}$, where the symbol $\|\cdot\|$ denotes the Frobenius norm. The optimal $\mathbf{A}$ and $\mathbf{B}$ are thus obtained in a least squares sense. As is well known, the minimum of (3) can be found by means of the singular value decomposition (SVD) of $\mathbf{X}$. Let $\mathbf{X} = \mathbf{PDQ}'$ with $\mathbf{P}$ and $\mathbf{Q}$ the columnwise orthonormal matrices holding in their columns, respectively, the left and right singular vectors of $\mathbf{X}$ and $\mathbf{D}$ the diagonal matrix containing the singular values of $\mathbf{X}$ in the main diagonal. We denote by $\mathbf{P}_S$ and $\mathbf{Q}_S$ the matrices the columns of which contain the first S columns of $\mathbf{P}$ and $\mathbf{Q}$, respectively, and by $\mathbf{D}_S$ the diagonal matrix with the S highest singular values. Then, the optimal component matrices are

$$\mathbf{A} = \mathbf{P}_S \mathbf{D}_S, \mathbf{B} = \mathbf{Q}_S. \quad (4)$$

The solution in (4) provides the best rank S approximation of $\mathbf{X}$. It is not unique because equally fitting solutions can be found by post-multiplying $\mathbf{A}$ by an arbitrary (non-singular) matrix $\mathbf{T}$ ($S \times S$) (new component score matrix $\mathbf{A}^R = \mathbf{AT}$) and post-multiplying $\mathbf{B}$ by $(\mathbf{T}')^{-1}$ (new component loading matrix $\mathbf{B}^R = \mathbf{B}(\mathbf{T}')^{-1}$). In fact, it is easy to see that $\mathbf{AB}' = \mathbf{A}^R \mathbf{B}^{R'}$. The transformation matrices are often chosen to be orthonormal and are then called rotation matrices.

Classical PCA is inadequate to summarize tensors because it cannot handle interactions between the 3 modes jointly in a proper way. The CP and T3 methods fill this gap.

2.1 | Candecomp/Parafac (CP)

The CP model was independently proposed by Carroll and Chang⁸ and Harshman.⁹ In particular, it was named Candecomp (from Canonical Decomposition)⁸ or Parafac (from Parallel Factors).⁹ Hereinafter, we refer to it as CP to acknowledge both papers. Although the latter is probably the most widely used name, some authors refer to it as the Polyadic decomposition to emphasize the role of Hitchcock.^{5,6} The intuition behind CP is straightforward. Bearing in mind (1), the component scores a_{is} express the relation between the observation units (from now on simply called “units”) and the components and the component loadings b_{js} express the relation between the variables and the components. In the CP model, new loadings, denoted by c_{ks} ($k = 1, \dots, K, s = 1, \dots, S$), linking occasions and components are added. In scalar form, the CP model can then be formulated as

$$x_{ijk} = \sum_{s=1}^S a_{is} b_{js} c_{ks} + e_{ijk}, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K, \quad (5)$$

where e_{ijk} is the generic error term belonging to the error term tensor $\underline{\mathbf{E}}$. The CP model in (5) exploits the principle of the parallel proportional profiles,⁷ ie, the same underlying components are used on all occasions, but they have different scalings depending on the occasions. In other words, the patterns linking the variables to the components remain the same (the loadings b_{js} do not differ with respect to the various occasions) but are proportional across the various occasions (with proportionality factors c_{ks}). The CP model could also be expressed in tensor formulation. However, it appears to be more useful to give the CP model in matrix formulation. Two equivalent formulations are:

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C} \bullet \mathbf{B})' + \mathbf{E}_A = \mathbf{A} \mathbf{I}_A (\mathbf{C} \otimes \mathbf{B})' + \mathbf{E}_A, \quad (6)$$

where $\mathbf{X}_A$, $\mathbf{I}_A$, and $\mathbf{E}_A$ represent the unit mode matricizations (see Figure 1) of $\underline{\mathbf{X}}$, $\underline{\mathbf{I}}$, and $\underline{\mathbf{E}}$, respectively, where $\underline{\mathbf{I}}$ denotes the identity array of order S , with the generic element $i_{s's''s'''}$ equal to 1 when $s' = s'' = s'''$ and 0 otherwise. For instance, $\mathbf{X}_A$ is the matrix of order $(I \times JK)$ obtained juxtaposing next to each other the units-by-variables matrices $\mathbf{X}_k$ (the slices of $\underline{\mathbf{X}}$) at occasion k ($k = 1, \dots, K$):

$$\mathbf{X}_A = [\mathbf{X}_1 \ \dots \ \mathbf{X}_k \ \dots \ \mathbf{X}_K]. \quad (7)$$

Moreover, $\mathbf{A}$ ($I \times S$), $\mathbf{B}$ ($J \times S$), and $\mathbf{C}$ ($K \times S$) are the component matrices for the units, variables, and occasions, respectively, and $\bullet$ denotes the so-called Khatri-Rao product (if $\mathbf{Y}$ and $\mathbf{Z}$ are 2 matrices with V columns, respectively, $\mathbf{y}_v$ and $\mathbf{z}_v$, $v = 1, \dots, V$, $\mathbf{Y} \bullet \mathbf{Z} = [\mathbf{y}_1 \otimes \mathbf{z}_1 \ \dots \ \mathbf{y}_V \otimes \mathbf{z}_V]$, where $\otimes$ denotes the Kronecker product of matrices). The CP method consists of fitting the CP model to the data by minimizing

$$\|\mathbf{E}_A\|^2 = \left\| \mathbf{X}_A - \mathbf{A}(\mathbf{C} \bullet \mathbf{B})' \right\|^2 \quad (8)$$

with respect to $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. For this purpose, several alternating least squares (ALS) algorithms have been proposed in the literature (for an overview, refer to Tomasi and Bro¹¹). The standard ALS algorithm consists of iteratively updating the 3 component matrices by solving ordinary regression problems.^{8,9} The expression in (8) is usually used to evaluate the fit percentage of the CP solution:

$$[1 - (\|\mathbf{E}_A\|^2 / \|\mathbf{X}_A\|^2)] 100. \quad (9)$$

The closer to 100, the better the fit of the CP model. The fit percentage in (9) is used to select the optimal number of components. As for PCA, it is usually desirable to search for a good balance between parsimony and fit of the solution (ie, applying “Occam’s razor”¹²). For increasing values of S , the fit percentage increases. Therefore, S should be such that the fit of CP with S components is noticeably higher than the 1 with $S - 1$ components and slightly lower than the 1 with $S + 1$ components. Graphically, the classical scree test plot¹³ can be applied to choose S in connection with the first elbow.

A nice property of the CP solution is its uniqueness up to scaling and permutation of the columns of the component matrices under mild conditions (see Kruskal,¹⁴ Domanov and De Lathauver,¹⁵ and Domanov and De Lathauver¹⁶ and

FIGURE 1 Three-way array or tensor and unit mode (eg, types of disease), variable mode (eg, number of admissions and mortality rates), and occasion mode (eg, years) matricizations [Colour figure can be viewed at wileyonlinelibrary.com]

references therein). Similarly to PCA, the CP solution with S components provides the best approximation of $\underline{\mathbf{X}}$ of tensorial rank S .¹³ The concept of tensorial rank represents the multi-way extension of that of (matrix) rank. $\underline{\mathbf{X}}$ has tensorial rank R if it is perfectly fitted by CP with R components. Although this definition is similar to that of (matrix) rank (the rank of a matrix is the number of components R such that $\mathbf{X}$ is perfectly fitted by PCA), their properties are different. A detailed comparison is out of the scope of this paper; interested readers may refer to Kruskal¹⁴ and references therein. Nonetheless, it is relevant to observe that the (tensorial) rank of a tensor may not exist. This has a practical implication, namely, with CP there always is the risk of obtaining degenerate solutions.¹⁷ The phenomenon of degeneracy refers to the case in which the obtained solution has no practical meaning because the components are highly col-linear and diverging and thus uninterpretable. Remedies are available for solving the degeneracy problem. See for instance Giordani and Rocci,¹⁸ Giordani and Rocci,¹⁹ and Stegeman²⁰ and references therein.

2.2 | Tucker3 (T3)

In the CP model, the same number of components S is used for the unit, variable, and occasion modes, and the underlying components are constructed according to Cattell's principle of parallel proportional profiles.⁷ The CP model is therefore based on quite strict assumptions preventing its applications in several cases. This occurs when the complexity of the 3 modes differs and different numbers of components for the 3 modes would be desirable or when Cattell's principle is violated. A less restricted tensor-based model is the T3 model.¹⁰ It can be formulated as

$$x_{ijk} = \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}, i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K, \quad (10)$$

in scalar notation, and

$$\mathbf{X}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C} \otimes \mathbf{B})' + \mathbf{E}_A, \quad (11)$$

in the matrix one. By inspecting (11), as for CP, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the component matrices for the 3 modes, but now their order is $(I \times P)$, $(J \times Q)$, and $(K \times R)$, respectively, where P , Q , and R denote the numbers of components for the units, the variables and the occasions, respectively. Furthermore, $\mathbf{G}_A$ is the unit mode matricization of the so-called core tensor $\underline{\mathbf{G}}$ of order $(P \times Q \times R)$ with generic element g_{pqr} , which expresses the triple interaction among component p of the unit mode, component q of the variable mode, and component r of the occasion mode. In practice, a high value of g_{pqr} in absolute sense highlights a strong relation among such components. The comparison between the third term of (6) and (11) allows us to discover the existing difference between CP and T3. In fact, if $\mathbf{G}_A = \mathbf{I}_A$ and, obviously, $P = Q = R = S$, then (6) and (11) coincide. This means that the CP model with S components can be seen as a constrained version of T3 with S components for all the modes where 1 component for the units is related to only 1 component for the variables and to only 1 component for the occasions. In general, the T3 model allows for all the triple interactions among the components of the 3 modes and, therefore, represents a more flexible tool for handling tensors. The T3 method consists of minimizing the sum of the squared errors

$$\|\mathbf{E}_A\|^2 = \|\mathbf{X}_A - \mathbf{A}\mathbf{G}_A(\mathbf{C} \otimes \mathbf{B})'\|^2 \quad (12)$$

with respect to the component matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ and the core array $\underline{\mathbf{G}}$ and the fit proportion of the T3 model can be evaluated by means of (9) where $\|\mathbf{E}_A\|^2$ is defined according to (12). The best known ALS algorithm for minimizing (12) is proposed by Kroonenberg and de Leeuw.²¹ In brief, the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are updated sequentially, while the core is updated implicitly. Note that the updates of the component matrices are carried out by constraining such matrices to be columnwise orthonormal. This can be done without loss of fit because the T3 solution is not unique. In fact, if we post-multiply $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ by transformation matrices of appropriate order denoted by $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$, we obtain the new component matrices $\mathbf{A}^R = \mathbf{A}\mathbf{S}$, $\mathbf{B}^R = \mathbf{B}\mathbf{T}$, and $\mathbf{C}^R = \mathbf{C}\mathbf{U}$. In order to get an equally fitting solution, $\mathbf{G}_A$ must be replaced by $\mathbf{G}_A^R = \mathbf{S}^{-1}\mathbf{G}_A(\mathbf{U}^{-1'} \otimes \mathbf{T}^{-1'})$. This guarantees that the transformations of the component matrices are compensated in the core. Specifically, we have

$$\mathbf{A}^R \mathbf{G}_A^R (\mathbf{C}^R \otimes \mathbf{B}^R)' = \mathbf{A} \mathbf{S} \mathbf{S}^{-1} \mathbf{G}_A (\mathbf{U}^{-1'} \otimes \mathbf{T}^{-1'}) (\mathbf{C} \mathbf{U} \otimes \mathbf{B} \mathbf{T})' = \mathbf{A} \mathbf{G}_A (\mathbf{C} \otimes \mathbf{B})'. \quad (13)$$

Such an indeterminacy of the T3 model can be exploited in order to search for simplicity of the component matrices or of the core tensor or of both. The concept of simplicity is expressed as a solution where a few component scores have high values in the absolute sense and the remaining ones have near-zero values. In this way, only subsets of entities of the various modes will influence components. Similarly, a limited number of core elements far from zero will be simple to interpret because only a few triple interactions capture the information hidden in the data tensor. In the literature, particular attention has been devoted to the simplicity of the core tensor attempting to set to zero some elements (see, eg, Kiers,²² ten Berge and Kiers,²³ and Rocci and ten Berge²⁴). This can be done without loss of fit by using the rotational freedom (and compensating the transformation in the component matrices) or allowing a certain loss of fit. Some general procedures for the joint simplicity of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and of $\underline{\mathbf{G}}$ are available.^{25,26} This goal is achieved by extending the standard varimax rotation²⁷ and the more general orthomax rotation.^{28,29} By exploiting the T3 rotational freedom, this consists of maximizing

$$f_{OR-T3}(\mathbf{S}, \mathbf{T}, \mathbf{U}) = f_{OR}^T(\mathbf{G}^R) + w_A f_{OR}(\mathbf{A}^R, \gamma_A) + w_B f_{OR}(\mathbf{B}^R, \gamma_B) + w_C f_{OR}(\mathbf{C}^R, \gamma_C) \quad (14)$$

with respect to orthogonal rotation matrices $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$. The function in (14) is a weighted sum of the orthomax criteria (the functions f_{OR}) applied to $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{G}$ (when $\gamma_A = \gamma_B = \gamma_C = 1$, the varimax criteria are used) and w_A , w_B , and w_C are non-negative weights balancing the simplicity for the pieces of the T3 solution. The selection of the optimal number of components for T3 is much more complex than CP because one has to choose the optimal number of components for every mode and a solution with a lower total number of components can have a better fit than a solution with a larger total number of components. This depends on the different level of complexity of the 3 modes. In the literature, several tools are available.³⁰⁻³³ The so-called Convex Hull (CHull) procedure³³ is a widely used numerical model selection heuristic where the fits of different T3 models using different numbers of components are analyzed, and the parsimony is evaluated by means of the number of free parameters.^{34,35} Let f_m and fp_m be the fit and the number of free parameters for solution m , respectively. The CHull is based on the following statistic

$$st_m = \frac{f_m - f_{m-1}}{fp_m - fp_{m-1}} \bigg/ \frac{f_{m+1} - f_m}{fp_{m+1} - fp_m}, \quad (15)$$

where $m - 1$ denotes the solution with the closest smaller number of free parameters. Solutions characterized by high st values represent a good compromise between fit and parsimony. To speed up the procedure, approximate solutions can be used very well³¹ (also see the supplementary material to this paper).

In the literature, the T3 model is sometimes referred to as Higher Order Singular Value Decomposition (HO-SVD).³⁶ In HO-SVD, the component matrices and the core tensor are explicitly constrained to be orthogonal. The HO-SVD solution is based on computing the eigendecompositions of $\mathbf{X}_A \mathbf{X}_A'$, $\mathbf{X}_B \mathbf{X}_B'$, and $\mathbf{X}_C \mathbf{X}_C'$ consistently with the original idea of Tucker.¹⁰ Note that such a strategy does not minimize the loss function in (12). Despite its name, there is no connection between HO-SVD and the tensorial rank of a tensor. For the sake of completeness, we mention 2 variants of T3. In T3, by choosing $P < I$, $Q < J$, and $R < K$, one summarizes all the 3 modes in terms of a limited number of components for every mode. It may occur that the researcher is interested in summarizing only 2 modes or 1 mode. In the former, case T3 reduces to the Tucker2 (T2) model. For instance, assuming that the unit mode is not reduced, the T2 model looks like

$$\mathbf{X}_A = \mathbf{G}_A(\mathbf{C} \otimes \mathbf{B})' + \mathbf{E}_A, \quad (16)$$

where, taking into account (11), the component matrix for the units $\mathbf{A}$ disappears ($\mathbf{A} = \mathbf{I}$) and the core array has order $(I \times Q \times R)$ with elements describing scores of the observed units on combinations of components of the variable mode and components of the occasion mode. If only 1 mode is summarized, then T3 (and T2) reduces to the Tucker1 (T1) model. Suppose that the interest is to extract components only for the unit mode, the T1 model can be formulated as

$$\mathbf{X}_A = \mathbf{A} \mathbf{G}_A + \mathbf{E}_A, \quad (17)$$

where, taking into account (11), $\mathbf{B}$ and $\mathbf{C}$ vanish and the core array has order $(P \times J \times K)$. By comparing (2) and (17), it is straightforward to conclude that PCA and T1 coincide.

2.3 | Software and related references

Readers interested in a deeper insight into 3-way methods can refer to monographs³⁷⁻⁴⁰ and to surveys.⁴¹⁻⁴³ Interesting applications can also be found.⁴⁴⁻⁴⁶ Three-way methods have been implemented in different software. The most important ones are the N-way toolbox⁴⁷ in Matlab, the Three-way m-files (see, eg, Kiers and Van Mechelen⁴⁵) in Matlab, the Tensor toolbox⁴⁸ in Matlab, the program 3WayPack,⁴⁹ and the R package ThreeWay.⁵⁰ In particular, the R package ThreeWay will be adopted for obtaining the results reported in the next section.

3 | AN APPLICATION OF TENSOR-BASED METHODS TO HOSPITAL DATA

The analysis of a 3-way dataset concerning the hospital care for a hospital complex in Rome (Italy) during 15 years distinguished in 3 groups of consecutive years was carried out. The data referred to the admissions and the mortality rate of the hospital complex named “Pio Istituto di Santo Spirito ed Ospedali Riuniti di Roma”. The hospital “Pio Istituto di

Santo Spirito” was one of the oldest ones in Rome and, even, in the world: it was founded by Pope Innocent III at the end of the XII century. Along the centuries, it was the most important hospital in Rome and included some other Roman hospitals becoming one of the largest hospital complexes in Europe. By means of the recorded admissions to the hospital along the years, we aimed at studying how the use of health care services evolved across time. Because the hospital was a reference in Rome, it was reasonable to think that the data offered a valuable description of the Roman health care system.

3.1 | Data

The admissions of the hospital complex distinguished by gender and age groups and the mortality rates yearly registered were studied. The information came from the annual bulletins of the statistical office of the hospital complex. The year 1892 was the first one in which the information was available. Unfortunately, for various reasons, the bulletins were either not published in certain years or lost. For instance, the information was missing for a lot of consecutive years at the beginning of the XX century. The bulletins were surely published in some years during the forties. After the Second World War, no information was available for approximately 20 years, whilst the bulletins from the mid of the sixties to 1972 were at disposal (since 1972, the bulletin was probably no longer published). In practice, the information about 3 consecutive groups of years of different length was available. This allowed for obtaining 3 different groups of pictures of the health conditions in Rome in 3 different eras. To avoid that groups would unduly have different influences on the solution, we decided to select 5 consecutive years for each era. In particular, we chose the earliest years, 1892 to 1896, some years during the Second World War, 1940 to 1944, and the most recent ones, 1968 to 1972 ($K = 5 \times 3 = 15$). For every year, the admissions and the mortality rates for $I = 15$ main types of disease (see Table 1) were recorded. Such a classification scheme was used in the bulletins for the last 2 groups of years. In particular, in these groups of years, the bulletins reported the number of admissions distinguished by types of disease. This classification was not used in the bulletins for the years 1892 to 1896, which recorded the number of admissions for specific diseases. Thus, data have been preprocessed by aggregating them over specific diseases belonging to the same type. The bulletins in 1968 to 1972 also contained the type of disease labelled “mental disorder”. Such an information was not considered in the analysis because it was not available in the first 2 groups of years when these patients were admitted to psychiatric institutions.

For every year, the available information about admissions was gender specific and age specific. In particular, age was divided into 8 classes of length equal to 10 years (less than 10 years, from 10 to 20, from 20 to 30, from 30 to 40, from 40 to 50, from 50 to 60, from 60 to 70, more than 70). The age of some patients was unknown, and, thus, their admissions belonged to the age group labelled “unknown age”. Such a scheme was inherited from the first era and was followed in

TABLE 1 $I = 15$ types of disease

Code	Type of disease
01	Infectious and parasitic diseases
02	Neoplasms
03	Constitutional diseases
04	Diseases of the blood
05	Diseases of the nervous system
06	Diseases of the circulatory system
07	Diseases of the respiratory system
08	Diseases of the digestive system
09	Diseases of the genitourinary system
10	Pregnancy and childbirth disorders
11	Diseases of the skin and subcutaneous tissue
12	Diseases of the musculoskeletal system and connective tissue
13	Congenital malformations, deformations, and chromosomal abnormalities
14	External causes of morbidity and mortality
15	Other diseases

the remaining years with a few exceptions. Namely, in 1940 to 1944, the age group “less than 10 years” was split in 3 subgroups, and, hence, such data were aggregated. In 1968 to 1972, data concerning the age group “more than 70” were aggregated because 2 subgroups were present. Moreover, for every year, the gender-specific mortality rates were registered. The male and female admissions for all of the above-described 9 age groups and the mortality rate defined the “variable” mode of the 3-way data set (see Table 2), hence $J = 20$ ($2 \times (9 + 1)$, ie, gender $\times$ (age groups + mortality rate)). Such data were stored in a 3-way array $\underline{\mathbf{X}}$ of order ($I = 15 \times J = 20 \times K = 15$), the generic element of which was x_{ijk} expressing, for instance, the number of admissions of a certain combination of age group and gender (“variable” j) for disease i in year k ($i = 1, \dots, 15$; $j = 1, \dots, 20$; $k = 1, \dots, 15$).

Further preprocessing was necessary because, for the years 1968 to 1972, the admissions distinguished by gender and age groups referred only to the alive patients. For this purpose, given the total number of male or female deaths for a certain year, we imputed its distribution with respect to the age groups according to 3 scenarios and then applied 3-way methods to all the scenarios to evaluate the stability of the results. In Scenario 1, we imputed the number of deaths for every age group proportional to the corresponding one of the alive. In Scenario 2, we assumed that the risk of death was increasing for older patients. This was done by imputing the values for each age group proportional to g^2 , where g is the order of the age group ($g = 1, \dots, 8$). In Scenario 3, we split the total number of deaths uniformly among the age groups. Of course, there were many other alternatives for imputing the distribution of the non-alive patients and, for instance, Scenario 3 appeared to be quite unrealistic. Nonetheless, we aimed at evaluating how the solutions differed with respect to the various scenarios, including extreme scenarios such as Scenario 3. As we shall see, the obtained results were remarkably similar.

3.2 | Descriptive statistics

Some descriptive statistics could help for a preliminary analysis of the data. The number of patients considerably increased during the 3 eras passing from approximately 25 000 per year on average in 1892 to 1896 to approximately 75 000 in 1940 to 1944 and almost 200 000 per year on average in 1968 to 1972. Such an increase was approximately proportional to that of the population of Rome. In fact, the census population estimates were approximately 0.4 and 1.1 million in 1901 and 1936, respectively, and more than 2.7 million in 1971, emerging that, in the first 2 eras, the trends were similar and the increase in the number of patients was slightly higher than the one of population. Remarkably, the number of male patients was twice the one of female in 1892 to 1896, whilst the number of females was slightly higher than that of male in the other 2 eras.

Some relevant distinctions among years and gender could be highlighted by analyzing the mortality rates reported in Figure 2. In the first era, the mortality rates for the female patients were higher than those for the male. Taking into account the number of female patients, one might conclude that women were admitted to the hospital in case of more severe diseases with respect to men and, to a limited extent as we shall see, for childbirth. In fact, especially in the past, childbirth was a dangerous event. In the years 1940 to 1944, the mortality rates for males were uniformly higher than the corresponding ones for females and increasing trends, especially for the males, could be observed. This is likely related to

TABLE 2 $J = 20$ “variables”

Code	“Variable”	Code	“Variable”
1	Less than 10—M	11	Less than 10—F
2	From 10 to 20—M	12	From 10 to 20—F
3	From 20 to 30—M	13	From 20 to 30—F
4	From 30 to 40—M	14	From 30 to 40—F
5	From 40 to 50—M	15	From 40 to 50—F
6	From 50 to 60—M	16	From 50 to 60—F
7	From 60 to 70—M	17	From 60 to 70—F
8	More than 70—M	18	More than 70—F
9	Unknown—M	19	Unknown—F
10	Mortality rate—F	20	Mortality rate—F

Note: The entities refer to age classes and gender. For instance, the entities labelled “from 10 to 20—F” refers to admissions of women from 10 to 20 years old.

FIGURE 2 Mortality rates (number of deaths expressed as a percentage of the total number of patients per year per gender category) [Colour figure can be viewed at wileyonlinelibrary.com]

the Second World War. The last era was characterized by the lowest mortality rates. The rates were almost constant during the 5 years and the ones for women were noticeably lower than the ones for men.

In Figure 3, we represented the average number of admissions for every era distinguished by type of disease using the codes given in Table 1. By inspecting the figure it was clear that the number of admissions increased with respect to time and that every era was characterized by a specific distribution of admissions. During the years 1892 to 1896, the most relevant reason for admissions was due to Infectious and parasitic diseases (code 01), which is one of the less relevant types of disease in 1968 to 1972. The admissions in the years 1940 to 1944 were distributed rather uniformly with respect to the types of disease. In such years, the mode was given by Diseases of the digestive system (08). In the last era, health conditions were definitively better than in the past and hospital admissions mainly depended on External causes of morbidity and mortality (14). This kind of disease described approximately one sixth of the admissions in the years 1968 to 1972.

These preliminary statistics allowed us to discover that the phenomenon under investigation noticeably varied in the 3 eras. The distribution of the admissions by type of disease changed across time, thus highlighting the existence of 3-way interactions among the modes. Therefore, it was reasonable to expect that a full analysis of the data would benefit from applying 3-way methods. This was not the case if, for instance, the distributions of the admissions remained stable across time (no 3-way interactions). If so, to summarize the relevant information, standard PCA on data averaged across the years would be enough to discover the existing 2-way (diseases $\times$ “variables”) interactions because the data compression would not produce any loss of information.

FIGURE 3 Average number of admissions for every era distinguished by type of disease. For each graph, starting from the center, the first radial line denotes 5000 admissions, the other ones +10 000 admissions; for instance, the third radial line denotes 25 000 admissions [Colour figure can be viewed at wileyonlinelibrary.com]

3.3 | Three-way analysis

The results of the 3-way analysis were obtained by applying the R package `ThreeWay`.⁵⁰ Let `hospital` be the R object holding the data (the data set is available as supplementary material). It is a `data.frame` corresponding to the disease mode unfolding of the tensor data according to Scenario 2. We started by running a T3 analysis.

```
R> library(ThreeWay)
R> T3.res <- T3(Hospital, Diseases, Variables, Years)
R> Specify the number of A-mode entities
R> 1: 15
R> Specify the number of B-mode entities
R> 1: 20
R> Specify the number of C-mode entities
R> 1: 15
```

where `Diseases`, `Variables`, and `Years` are optional vectors containing the labels of the 3 modes (available as supplementary material). The function `T3` performs an interactive T3 analysis in such a way that the user inspects some preliminary output and decides how to proceed accordingly. These steps are discussed later. For a thorough explanation, the interested reader is invited to refer to the appendix (see supplementary material).

The first decision to be made concerned how to preprocess the data (the functions `cent3` and `norm3` were automatically called). By centering, the aim is to remove possible off-set terms so that resulting scores can be seen as measurements on proportional scales. Because a meaningful zero point exists in our case (ie, if $x_{ijk} = 0$, then there are either no admissions for the entities involved or a mortality rate equal to zero), we decided not to center the data. The scaling step was performed by normalizing within the “variable” mode. This allowed us to guarantee that the results of the analysis were not affected by the magnitudes of admissions for the age groups, thus, avoiding that the older age groups played the most relevant role in extracting the components.

```
R> How do you want to normalize your array?
R> 0 = none (default)
R> 1 = within A-mode
R> 2 = within B-mode
R> 3 = within C-mode
R> 1: 2
R> Read 1 item
R> Data have been normalized within B-mode
```

The next step was the selection of the optimal numbers of components by applying the C-Hull procedure³³ to approximate T3 solutions (by calling the functions `DimSelector` and `T3dimensionalityplot`).

```
R> If you want to do the PCASUPs for choosing your dimensionality, specify '1':
R> 1: 1
R> For the generalized scree test it is needed to indicate the maximum number of dimensions
for each mode you want to study
R> Up to how many A-mode components do you want to use?
R> 1: 6
R> Up to how many B-mode components do you want to use?
R> 1: 6
R> Up to how many C-mode components do you want to use?
R> 1: 6
```

We got the results reported in Figure 4 and summarized in Table 3. Figure 4 displays the plot of the fit percentages versus the number of free parameters for all approximate T3 solutions starting from the simplest one with 1 component for every mode ($P = Q = R = 1$) to the most complex one (we decided to extract not more than 6 components for every mode,

FIGURE 4 Plot of the obtained approximate T3 solutions for different values of P , Q , and R ; x-axis: Number of effective parameters; y-axis: Fit percentages. The dashed lines represent the upper boundary of the convex hull, and the arrows identify the most valuable T3 solutions (reported in Table 3)

TABLE 3 Most valuable T3 solutions according to the CHull procedure: Numbers of components (P , Q , R), total number of components $P + Q + R$, approximate goodness-of-fit values (ie, based on the eigendecompositions of $\mathbf{X}_A\mathbf{X}_A'$, $\mathbf{X}_B\mathbf{X}_B'$, and $\mathbf{X}_C\mathbf{X}_C'$,³³ using the function `pcasup3`) rounded to the nearest integer and st values

(P , Q , R)	$P + Q + R$	Fit	st
(1, 1, 1)	3	57	—
(2, 2, 1)	5	67	1.09
(3, 3, 1)	7	75	1.62
(4, 4, 2)	10	85	2.12
(5, 5, 2)	12	88	1.61
(5, 6, 3)	14	91	1.79
(6, 6, 4)	16	93	3.95
(6, 6, 5)	17	93	2.22
(6, 6, 6)	18	93	—

thus $P = Q = R = 6$). The most valuable solutions balancing fit and parsimony are located along the higher boundary of the convex hull. Solutions below such a boundary should be discarded because there exists at least 1 solution with higher fit and same level of parsimony. The CHull procedure numerically assesses the higher boundary of the convex hull and selects the best solutions. In particular, among the solutions belonging to the higher boundary, the interest is in solutions associated with a relatively big increase (see arrows), as expressed in their st values. This subset of solutions with the corresponding st values was given in Table 3. The maximum st value (3.95) was registered for solution (6, 6, 4), whilst the second highest st value (2.22) pertained to solution (6, 6, 5). Both solutions appeared extremely complex, having 16 and 17 components, respectively. A more reasonable solution was the one with $P = 4$, $Q = 4$, and $R = 2$, having the third highest st value with a fit of 85%. Taking into account the total number of components, 10, we decided to investigate this solution because it represented the best compromise in terms of fit and parsimony. In fact, the solutions with higher st values were much more complex (6 or 7 additional components) and accounted for a limited increase of fit (approximately +8%). By inspecting Table 3, an interesting conclusion was drawn. It was clear that the complexity of the occasion mode was lower than that of the remaining modes. In other words, the number of components needed for summarizing the occasion mode should be lower than those for the other modes ($R < \min\{P, Q\}$). This finding was interpreted as a hint that the CP structure was not adequate to fit the data reasonably well. Nonetheless, prior to focusing our attention on the

T3 model, we ran different CP analyses setting $S = 2, 3$, and 4 . The function CP was used:

```
R> library(ThreeWay)
R> CP.res <- CP(Hospital, Diseases, Variables, Years)
```

and essentially the same steps previously described were made (the code is not reported). The registered fit percentages were 68%, 77%, and 86%. By comparing such values with the T3 ones reported in Table 3, it was clear that T3 should be preferred. More specifically, the $S = 2$ solution (total number of components = 6) should not be preferred to the T3 (2, 2, 1) one (approximately same fit, 1 component less). A similar reasoning allowed us to discard the CP solutions with $S = 3$ and $S = 4$ components, taking into account the T3 ones in cases (3, 3, 1) and (4, 4, 2). In the latter case, the more parsimonious T3 model (10 components) had a higher fit than the CP one with $S = 4$ (12 components). Apart from the fit values, we found that the CP solutions for varying S were not informative. Moreover, when $S = 4$, the typical features of degeneracy, ie, uninterpretable, highly collinear, and diverging components, were encountered.

3.3.1 | T3 with $P = 4$, $Q = 4$ and $R = 2$

The T3 model with $P = 4$, $Q = 4$, and $R = 2$ components has been fitted to the data.

```
R> How many A-mode components do you want to use?
R> 1: 4
R> How many B-mode components do you want to use?
R> 1: 4
R> How many C-mode components do you want to use?
R> 1: 2
```

The ALS algorithm was run using a convergence criterion equal to 10^{-6} , and 5 random starts were considered to limit the risk of attaining local optima. The fit of the model was 85.47%.

In order to interpret the solution, we exploited the rotational indeterminacy trying to reach the maximal simplicity for the core tensor and the 3 component matrices by applying the general procedures in Crawford and Ferguson²⁸ and Jennrich²⁹ (this was automatically done in T3 calling the function `varimcoco`). We observed that the component matrix for the years (**C**) had a simple structure even in cases where its simplicity obtained low weight (eg, see Table 7 below), and therefore we concentrated our attention on the simplicity of **A** (component matrix for the types of diseases), **B** (component matrix for the “variables”), and **G** (core tensor). For this purpose, we maximized (14) in the varimax case (ie, $\gamma_A = \gamma_B = \gamma_C = 1$), setting $w_C = 0$ and choosing different values in the set $\{1, 2, 5, 10\}$ for w_A and w_B . The results are reported in Table 4.

We believed that a good compromise for the simple structure of **A**, **B**, and **G** was found when $w_A = 2$ and $w_B = 5$. To motivate this choice, in Figure 5, we displayed the percentages of improvement from the chosen set of weights of the simplicity of **A**, **B**, **C**, and **G**, ie, the differences in percentage between the values of the functions f_{OR} for w_A and w_B in $\{1, 2, 5, 10\}$ with respect to those for $w_A = 2$ and $w_B = 5$.

```
R> Find useful simple structure rotation of core and components
R> You can now carry out SIMPLE STRUCTURE rotations with varying weights.
R>
R> Specify (range of) relative weight(s) for A (default=0):
R> 1: 2
R> 2:
R> Read 1 item
R> Specify (range of) relative weight(s) for B (default=0):
R> 1: 5
R> 2:
R> Read 1 item
R> Specify (range of) relative weight(s) for C (default=0):
R> 1: 0
R> 2:
R> Read 1 item
```

TABLE 4 Results of the orthomax procedure for different combinations of weights

w_A	w_B	w_C	$f_{OR}^T(\underline{G}^R)$ [core]	$f_{OR}(A^R, 1)$ [diseases]	$f_{OR}(B^R, 1)$ ["variables"]	$f_{OR}(C^R, 1)$ [years]
1	1	0	13.02	6.78	2.93	0.93
1	2	0	12.95	6.79	2.97	0.94
1	5	0	12.68	6.82	3.04	0.96
1	10	0	8.68	7.33	3.46	1.03
2	1	0	12.91	6.84	2.94	0.94
2	2	0	12.84	6.85	2.98	0.95
2	5	0	12.55	6.88	3.05	0.97
2	10	0	5.16	8.54	3.65	1.06
5	1	0	6.92	8.40	3.17	1.03
5	2	0	5.78	8.54	3.48	1.03
5	5	0	5.08	8.59	3.63	1.06
5	10	0	4.98	8.60	3.65	1.07
10	1	0	5.42	8.60	3.45	1.02
10	2	0	5.18	8.60	3.58	1.05
10	5	0	5.02	8.60	3.64	1.07
10	10	0	4.93	8.60	3.65	1.07

FIGURE 5 Results of the orthomax procedure for different combinations of weights. Percentages of improvement from the chosen set of weights of the simplicity of **A**, **B**, **C**, and **G** [Colour figure can be viewed at wileyonlinelibrary.com]

Higher weights implied a core tensor difficult to interpret, whilst lower values of w_A and w_B affected the simplicity of the corresponding component matrices. The resulting component matrices were reported in Tables 5-7. In order to interpret the components of every mode, we inspected the 3 matrices separately. As for standard PCA, high values in absolute sense reflect a relevant role of the corresponding entity.

In Table 5, we see that Component 1 for the type of disease mode mainly depended on diseases 14 (with negative sign), 06, and 02. This component is associated with some of the most frequent causes of admission in the last era (see right side of Figure 3). In particular, this component highlights the contrast, ie, a few diseases with positive scores and some others with negative ones, between, on the 1 side, the 2 “big killers” Neoplasms (02) and Diseases of the circulatory systems (06) and, on the other side, External causes of morbidity and mortality (14). Component 2 is strongly

TABLE 5 Component matrix A

Code	Type of disease	Lifestyle diseases	Pregnancy and childbirth disorders	Subset of internal diseases	Infectious and parasitic diseases
01	Infectious and parasitic diseases	0.00	−0.01	0.08	0.92
02	Neoplasms	0.38	0.00	0.29	0.00
03	Constitutional diseases	0.09	0.01	0.10	−0.01
04	Diseases of the blood	−0.02	−0.03	0.06	0.14
05	Diseases of the nervous system	0.18	−0.03	0.23	0.09
06	Diseases of the circulatory system	0.55	−0.10	0.41	−0.24
07	Diseases of the respiratory system	−0.02	−0.10	0.28	0.07
08	Diseases of the digestive system	0.04	0.05	0.40	0.16
09	Dis. of the genitourinary system	0.14	0.19	0.17	0.02
10	Pregnancy and childbirth disorders	0.00	0.97	0.05	−0.01
11	Dis. of the skin and subcut. tissue	0.01	0.00	0.03	0.11
12	Dis. of the musc. syst. and conn. tiss.	0.03	0.00	0.08	0.06
13	Cong. malf., deform. and chrom. abn.	−0.06	−0.02	0.04	0.03
14	External causes of morb. and mort.	−0.68	−0.05	0.56	−0.14
15	Other diseases	−0.14	0.01	0.29	−0.05

Note: component scores higher than 0.20 in absolute value are in bold style.

TABLE 6 Component matrix B

“Variable”	Men and older women	Older vs younger age groups	Mortality rates	Women in the fertility period
Less than 10—M	0.22	−0.29	−0.07	−0.08
From 10 to 20—M	0.21	−0.38	−0.04	−0.07
From 20 to 30—M	0.20	−0.34	0.09	−0.05
From 30 to 40—M	0.22	−0.24	0.13	−0.04
From 40 to 50—M	0.27	−0.06	0.08	−0.04
From 50 to 60—M	0.29	0.14	0.06	−0.05
From 60 to 70—M	0.29	0.29	0.00	−0.06
More than 70—M	0.28	0.35	−0.01	−0.06
Unknown—M	0.24	−0.22	−0.29	−0.08
Mortality rate—M	0.07	−0.00	0.58	0.01
Less than 10—F	0.21	−0.28	−0.01	−0.08
From 10 to 20—F	0.17	−0.23	0.10	0.23
From 20 to 30—F	0.03	−0.01	0.00	0.65
From 30 to 40—F	0.07	0.03	0.00	0.62
From 40 to 50—F	0.24	0.13	0.03	0.20
From 50 to 60—F	0.29	0.15	−0.02	−0.03
From 60 to 70—F	0.30	0.24	−0.06	−0.06
More than 70—F	0.28	0.30	−0.06	−0.08
Unknown—F	0.22	−0.09	−0.28	0.22
Mortality rate—F	0.08	−0.02	0.66	0.00

Note: Component scores higher than 0.20 in absolute value are in bold style.

TABLE 7 Component matrix C

Year	Early period	Late period
1892	0.33	0.03
1893	0.32	0.03
1894	0.31	0.03
1895	0.34	0.03
1896	0.33	0.03
1940	0.26	0.14
1941	0.27	0.13
1942	0.28	0.13
1943	0.28	0.12
1944	0.33	0.11
1968	−0.05	0.42
1969	−0.08	0.45
1970	−0.09	0.38
1971	−0.12	0.43
1972	−0.19	0.46

Note: component scores higher than 0.20 in absolute value are in bold style.

related to Pregnancy and childbirth disorders (10) and thus reflects such a specific disease. Component 3 is related to a subset of types of disease: in order of importance, External causes of morbidity and mortality (14), Diseases of the circulatory systems (06), Diseases of the digestive systems (08) and, to a lesser extent, Other diseases (15), Neoplasms (02), Diseases of the respiratory systems (07), and Diseases of the nervous systems (05). These diseases shared similar profiles with respect to admissions and mortality rates. For a more comprehensible description of this phenomenon, the components for the other modes must be considered, and the interactions should be analysed through the core elements. Finally, Component 4 essentially reflects Infectious and parasitic diseases (01) and, to a limited extent, Diseases of the circulatory systems (06) with negative sign.

Table 6 shows that Component 1 for the “variable” mode is related to several entities, namely, to all the age classes for the men and to the older ones for the women as well as, to a lesser extent, to the female age class “less than 10”. Component 2 also depends on a large number of age classes. In particular, it is positively connected with the older age groups and negatively with the younger. Moreover, distinctions by gender do not emerge. More “specific” interpretations pertain to Components 3 and 4. Specifically, Component 3 strongly reflects the mortality rates (both for female and male) and, to a limited extent with negative sign, the age group labelled “unknown”. Finally, Component 4 is gender oriented. It is mainly associated with women in the fertility period. In fact, note the highest component scores for the age groups from 20 to 40 years old and the quite high ones for the “unknown” group and for the age class from 10 to 20 years (thus highlighting the presence of women close to 20).

As previously mentioned, the component matrix for the years (reported in Table 7) is very simple to interpret. Components 1 and 2 discriminate 2 groups of years. The 2 older eras (1892–1896 and 1940–1944) influence the former component, whilst the more recent era (1968–1972) plays a relevant role in characterizing the latter one. This allows us to discover that the distribution of the admissions was very different in the last era in comparison with the first two.

So far, the components have been interpreted separately for each mode. A full and deeper comprehension of the data was achieved by integrating the previously described results with the core tensor. In fact, its elements show the interactions among the components of the various modes. The matricized version of the core is reported in Table 8.

The higher in absolute value an element of the core, the stronger is the interaction among the components involved. The highest score (= 50) refers to the triple of components (“Subset of internal diseases”, “Men and older women”, “Late period”). Taking into account the component interpretations, we conclude that, during the years 1968 to 1972, there were relatively many admissions with the subset of internal diseases related to Component 3 for the diseases. To corroborate this claim, we inspected the data tensor and noted that, in the last 5 years, for each pair of age group and gender

TABLE 8 Core matrix G_A

	Early period				Late period			
	Men and older women	Older vs younger age groups	Mortality rates	Women in the fertility period	Men and older women	Older vs younger age groups	Mortality rates	Women in the fertility period
Lifestyle diseases	1	−3	4	1	−1	18	7	−0
Pregnancy and childbirth disorders	1	−0	−1	2	2	−1	−1	21
Subset of internal diseases	0	−3	14	1	50	−2	3	4
Infectious and parasitic diseases	10	−7	5	1	3	−4	3	1

Note: scores higher than 10.00 in absolute value are in bold style. Values are rounded to the nearest integer.

related to the component labelled “Men and older women”, large numbers of admission were registered for the diseases associated with the component “Subset of internal diseases”. In particular, the higher the component score, the larger the number of admissions was. We also observed that, except for a few cases, the first 5 high ranked diseases in terms of number of admissions for the previously described pairs of age group and gender came from the “Subset of internal diseases”. These findings can be explained by noting that the years of the “Late period” coincided with the end of the so-called Italian economic miracle. Starting from the end of the Second World War, a strong economic growth in Italy and, indeed, in Rome, was registered. During this phase, the country became a major industrial power, and a high percentage of people were employed in the manufacturing sector. Obviously, this had an impact on the Italian society: the quality of life remarkably improved for the great majority of the population, thus enhancing health condition and increasing life expectancy. This produced the growth of diseases connected with a more stressing life style and with old age, in particular those related to Component 3 for the diseases.

The interaction (core element = 21) among “Pregnancy and childbirth disorders”, “Women in the fertility period” and “Late period” highlights a specific phenomenon. In the years 1968 to 1972, several admissions concerned women in the fertility period for Pregnancy and childbirth disorders. Such a strong interaction between Component 2 for the diseases and Component 4 for the “variables” was not discovered in connection with the early period. This might be explained by the fact that the category “Pregnancy and childbirth disorders” also include admissions of problem free childbirth, and the latter hardly ever was a reason for going to hospital in the earlier 2 eras, because then childbirth typically occurred at home.

A high interaction (= 18) is observed for the components labelled “Lifestyle diseases”, “Older vs younger age groups”, and “Late period”. It should be noted that in the first 2 components the contrast between subsets of entities noticeably emerged. In detail, such components were explained by large scores in absolute sense with opposite sign. The diseases with positive component scores were related to the age groups with positive component scores (and vice-versa). Hence, we concluded that, in the years 1968 to 1972, relatively younger people were admitted due to External causes of morbidity and mortality, whilst the admissions of older people were caused more by Neoplasms and Diseases of the circulatory systems.

Two other core elements merit our attention. Both refer to the first 2 time eras. The interaction (= 14) among “Subset of internal diseases”, “Mortality rates”, and “Early period” highlighted that, in the years 1892 to 1896 and 1940 to 1944, the subset of internal diseases was the most problematic one, being associated with a higher risk of death in comparison with the other diseases.

An interesting discovery concerns the element g_{411} (= 10), expressing the triple interaction among “Infectious and parasitic diseases”, “Men and older women”, and “Early period”. The analysis highlighted a peculiar occurrence: Infectious and parasitic diseases implied a relevant number of admissions for almost all the age groups in the years 1892 to 1896 and 1940 to 1944, whilst it had a negligible impact in 1968 to 1972. Malaria and other parasitic diseases were present in Italy at the end of the 19th century. During these years, in order to defeat malaria, the Italian government stimulated the low-cost administration of quinine. This was an open health problem in Rome, due to the Pontine Marshes, located in the surroundings of Rome. The area was reclaimed in 1920 to 1930, but infectious and parasitic diseases were present all around Italy and, indeed, in Rome during years 1940 to 1944 due to the Second World War which contributed to the

deterioration of the extremely poor Italian health condition. This was a consequence of the weak economic condition after the First World War and the extended post-war economic depression.

3.3.2 | Stability analysis

The stability of the obtained solution was checked with respect to 2 points of view. In order to assess the stability of the solution due to the relatively small data size, we decided to perform a split-half analysis (see, eg, Harshman and De Sarbo⁵¹). In splitting, the 2 halves would need to be as similar as possible. For this reason, we studied the stability of the solution across the years. In particular, we adopted an odd-even split where, for each group of 5 years, 3 non-consecutive years belonged to Split 1 and the remaining ones to Split 2. Therefore, Split 1 contained the years 1890, 1892, 1894, 1940, 1942, 1944, 1968, 1970, and 1972, resulting in a tensor of order $(15 \times 20 \times 9)$ and Split 2 produced a tensor of order $(15 \times 20 \times 6)$. The split-half analysis was automatically implemented in the function `T3` (calling the function `splithalfT3`). However, to implement the previously described splitting, some data manipulations were needed and the specific function `splithalfT3` was run.

For each split, we applied `T3` with $P = 4$, $Q = 4$, and $R = 2$ to the data after performing the same preprocessing step chosen for the full data tensor. Because the numbers of units and variables remained the same, for each split we obtained a pair of component matrices **A** and **B** of the same order and fully comparable with those obtained using the full data (reported in Tables 5 and 6). Once the solutions are comparable, the Tucker congruence coefficient φ ⁵² can be adopted (the Tucker congruence coefficient is implemented in the function `phi`). It measures the proportionality between corresponding columns and takes values in the interval $[-1, 1]$. Values of φ higher than 0.85 reflect stable solutions. The component matrices were very similar with values of φ very often higher than 0.99. With respect to **C**, which had order (9×2) and (6×2) for Split 1 and Split 2, respectively, we combined the component matrices from the 2 splits and compared it with the one from the full data. The observed congruence was high (>0.99). Finally, we checked how much the core arrays differed. In this case, the core tensors had the same size, but their magnitudes needed to be corrected by taking into account the numbers of years in the 2 splits.* We found that the cores from the 2 splits well resembled the one from the full data. The most relevant triple interactions were discovered by considering the splits and the maximum difference between corresponding core elements was equal to approximately 2. All in all, the solutions were fully comparable, and thus the obtained `T3` solution appeared to be stable.

Furthermore, a sensitivity analysis was carried out in order to evaluate whether the solutions were comparable with respect to the 3 scenarios considered for handling the distribution of the deaths. For this purpose, we ran the function `T3` using the same preprocessing and extracting the same number of components on the data tensor obtained under scenarios 1 and 3. We found that the components were interpreted in the same way. To assess in a more objective way, how much the component matrices differed, we computed the Tucker congruence coefficient between pairs of corresponding columns. The results were very satisfactory, the coefficients were always higher than 0.99, thus denoting an almost perfect congruence. The 3 core arrays discovered the same strong interactions among triples of components. The magnitudes of the cores were fully comparable admitting a comparison. On average, just 12.5% of core elements differed more than 0.50 with respect to the corresponding one obtained according to a different scenario (4.17% more than 1.00). The maximum difference we found was 2.06.

4 | FINAL DISCUSSION

In this paper, we have reviewed the most common tensor-based methods, CP and `T3`. They represent 3-way extensions of the classical PCA. The primary scope of CP and `T3` is to detect the underlying structure of the data by extracting a limited number of components for all the modes characterizing the tensor. `T3` is more general than CP by allowing for different numbers of components for the modes and synthesizing the triple interactions among the components in the so-called core tensor. In the application, we considered the `T3` method, which appeared more fruitful in order to analyse the hospital care data. In particular, our study on health condition was carried out by means of `T3` with 4 components for the types of disease, 4 components for the admissions distinguished by gender and age groups and the mortality rates and 2

*Because the matrix **C** is always normalized such that $\mathbf{C}'\mathbf{C} = \mathbf{I}$, in the first split (where 9 out of 15 occasions are used), its elements will be artificially increased by a factor of $\sqrt{15/9}$; as a consequence, the elements in the core will be *decreased* by that same factor; likewise, in the second split (where 6 out of 15 occasions are used), its elements will be artificially increased by a factor of $\sqrt{15/6}$, and the elements in the core will be *decreased* by that same factor.

components for the years. This solution was motivated by the fact that the complexity of the underlying structures of the modes differed and, therefore, considering a different number of components for each mode was the most sensible choice. In particular, the occasion mode appeared to be the simplest one requiring only 2 components despite the fact that the data referred to 3 different historical eras. One component was related to the first 2 eras and another component to the last one. The T3 model with one more component for the years appeared not to be interesting because the increase of fit was negligible. This highlighted that the differences in admissions and mortality rates between the first 2 eras were not relevant enough to justify the need for an additional component discriminating such 2 eras. The solution we found was stable and easy to interpret. The latter point was done by exploiting the rotational freedom of T3, and several components were interpreted in connection with historical events in Italy.

The obtained results were not possible if we analysed the data by means of PCA. To properly deal with the tensor-structure of the data, ad-hoc methods, such as CP and T3, should be used. In fact, differently from 2-way analyses, 3-way models simultaneously take into account all the modes and their interactions offering a full description of the tensor.

ACKNOWLEDGEMENTS

We are grateful to the AE and the reviewers for their insightful comments. We also thank Elena Rodriguez for providing the data analysed in this paper.

ORCID

Paolo Giordani <http://orcid.org/0000-0003-4091-3165>

REFERENCES

1. Bhattacharya A, Dunson DB. Simplex factor models for multivariate unordered categorical data. *J Am Stat Assoc.* 2012;107:362-377.
2. Zhou J, Bhattacharya A, Herring AH, Dunson DB. Bayesian factorizations of big sparse tensors. *J Am Stat Assoc.* 2015;110:1562-1576.
3. Hoff PD. Equivariant and scale-free Tucker decomposition models. *Bayesian Anal.* 2016;11:627-648.
4. Johndrow J, Bhattacharya A, Dunson DB. Tensor decompositions and sparse log-linear models. *Ann Stat.* 2017;45:1-38.
5. Hitchcock FL. The expression of a tensor or a polyadic as a sum of products. *J Math Phys.* 1927;6:164-189.
6. Hitchcock FL. Multiple invariants and generalized rank of a p -way matrix or tensor. *J Math Phys.* 1927;7:39-79.
7. Cattell RB. Parallel proportional profiles and other principles for determining the choice of factors by rotation. *Psychometrika.* 1944;9:267-283.
8. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an n -way generalization of Eckart-Young decomposition. *Psychometrika.* 1970;35:283-319.
9. Harshman RA. Foundations of the Parafac procedure: models and conditions for an 'explanatory' multimodal factor analysis. *UCLA Work Pap Phon.* 1970;16:1-84.
10. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika.* 1966;31:279-311.
11. Tomasi G, Bro R. A comparison of algorithms for fitting the PARAFAC model. *Comput Stat Data Anal.* 2006;50:1700-1734.
12. Vandekerckhove J, Matzke D, Wagenmakers E-J. Model comparison and the principle of parsimony. In: Townsend JT, Busemeyer JR, eds. *The Oxford Handbook of Computational and Mathematical Psychology.* Oxford: Oxford University Press; 2015.
13. Cattell RB. The scree test for the number of factors. *Multivariate Behav Res.* 1966;1:245-276.
14. Kruskal JB. Three-way arrays: rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. *Linear Algebra Appl.* 1977;18:95-138.
15. Domanov I, De Lathauver L. On the uniqueness of the canonical polyadic decomposition of third-order tensors—part I: basic results and uniqueness of one factor matrix. *SIAM J Matrix Anal Appl.* 2013;34:855-875.
16. Domanov I, De Lathauver L. On the uniqueness of the canonical polyadic decomposition of third-order tensors—part II: uniqueness of the overall decomposition. *SIAM J Matrix Anal Appl.* 2013;34:876-903.
17. Kruskal JB, Harshman RA, Lundy ME. How 3-MFA data can cause degenerate PARAFAC solutions, among other relationships. In: Coppi R, Bolasco S, eds. *Multiway Data Analysis.* Amsterdam: Elsevier; 1989:115-122.
18. Giordani P, Rocci R. Candecomp/Parafac via the lasso. *Psychometrika.* 2013;78:669-684.

19. Giordani P, Rocci R. Some clarifications of remedies for Candecomp/Parafac degeneracy by means of an SVD-penalized approach. *Chemometr Intell Lab Syst.* 2017;162:172-181.
20. Stegeman A. Finding the limit of diverging components in three-way Candecomp/Parafac—a demonstration of its practical merits. *Comput Stat Data Anal.* 2014;75:203-216.
21. Kroonenberg PM, De Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika.* 1980;45:69-97.
22. Kiers HAL. TUCKALS core rotations and constrained TUCKALS modelling. *Stat Appl.* 1992;4:659-667.
23. ten Berge JMF, Kiers HAL. Simplicity of core arrays in three-way principal component analysis and the typical rank of $P \times Q \times 2$ arrays. *Linear Algebra Appl.* 1999;294:169-179.
24. Rocci R, ten Berge JMF. Transforming three-way arrays to maximal simplicity. *Psychometrika.* 2002;67:351-365.
25. Kiers HAL. Three-way SIMPLIMAX for oblique rotation of the three-mode factor analysis core to simple structure. *Comput Stat Data Anal.* 1998;28:307-324.
26. Kiers HAL. Three-mode orthomax rotation. *Psychometrika.* 1997;62:579-598.
27. Kaiser HF. The varimax criterion for analytic rotation in factor analysis. *Psychometrika.* 1958;23:187-200.
28. Crawford CB, Ferguson GA. A general rotation criterion and its use in orthogonal rotation. *Psychometrika.* 1970;35:321-332.
29. Jennrich RI. Orthogonal rotation algorithms. *Psychometrika.* 1970;35:229-235.
30. Timmerman ME, Kiers HAL. Three-mode principal components analysis: choosing the numbers of components and sensitivity to local minima. *Br J Math Stat Psychol.* 2000;53:1-16.
31. Kiers HAL, der Kinderen A. A fast method for choosing the numbers of components in Tucker3 analysis. *Br J Math Stat Psychol.* 2003;56:119-125.
32. Kiers HAL, Bro R. A new efficient method for determining the number of components in PARAFAC models. *J Chemometr.* 2003;17:274-286.
33. Ceulemans E, Kiers HAL. Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. *Br J Math Stat Psychol.* 2006;59:133-150.
34. Weesie J, Van Houwelingen H. *GEPCAM Users' Manual (First Draft)*. Utrecht: Institute of Mathematical Statistics, State University of Utrecht; 1983.
35. Wansbeek T, Verhees J. Models for multidimensional matrices in econometrics and psychometrics. In: Coppi R, Bolasco S, eds. *Multiway Data Analysis*. Amsterdam: Elsevier; 1989:543-552.
36. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. *SIAM J Matrix Anal Appl.* 2000;21:1253-1278.
37. Kroonenberg PM. *Three-Mode Principal Component Analysis: Theory and Applications*. Leiden: DSWO Press; 1983 Available at: http://www.leidenuniv.nl/fsw/three-mode/bibliogr/kroonenbergpm_thesis/kroonenberg1983.pdf.
38. Bro R: Multi-way analysis in the food industry, models algorithms and applications. Ph.D. Thesis. Amsterdam: University of Amsterdam, 1998. Available at: <http://www.models.life.ku.dk/~rasmus/thesis/brothesis.pdf>
39. Smilde A, Bro R, Geladi P. *Multi-way Analysis: Applications in the Chemical Sciences*. Chichester: Wiley; 2004.
40. Kroonenberg PM. *Applied Multiway Data Analysis*. Hoboken: Wiley; 2008.
41. Acar E, Yener B. Unsupervised multiway data analysis: a literature survey. *IEEE Trans Knowl Data Eng.* 2009;21:6-20.
42. Kolda TG, Bader BW. Tensor decompositions and applications. *SIAM Rev.* 2009;51:455-500.
43. Mørup M. Applications of tensor (multi-way array) factorizations and decompositions in data mining. *WIREs Data Min Knowl Discov.* 2011;1:24-40.
44. Andersen AH, Rayens WS. Structure-seeking multilinear methods for the analysis of fMRI data. *Neuroimage.* 2004;22:728-739.
45. Kiers HAL, Van Mechelen I. Three-way component analysis: principles and illustrative application. *Psychol Methods.* 2001;6:84-110.
46. Kroonenberg PM, Lammers CJ, Stoop I. Three-mode principal component analysis of multivariate longitudinal organizational data. *Sociol Methods Res.* 1985;14:99-136.
47. Andersson CA, Bro R. The N -way toolbox for MATLAB. *Chemometr Intell Lab Syst.* 2000;52:1-4.
48. Bader BW, Kolda TG, Sun J, et al. MATLAB tensor toolbox version 2.6, 2015. Available at: <http://www.sandia.gov/~tgkolda/TensorToolbox/>
49. Kroonenberg PM. 3WAYPACK user's manual, 1996. Available at: <http://three-mode.leidenuniv.nl/document/programs.htm#3WayPack>
50. Giordani P, Kiers HAL, Del Ferraro MA. Three-way component analysis using the R package ThreeWay. *J Stat Softw.* 2014;57(7):1-23.
51. Harshman RA, De Sarbo WS. An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In: Law HG, Snyder CW Jr, Hattie JA, McDonald RP, eds. *Research Methods for Multimode Data Analysis*. New York: Praeger; 1984:602-642.

52. Tucker LR. *A Method for Synthesis of Factor Analysis Studies. Personnel Research Section Report No. 984*. Washington: Department of the Army; 1951.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Giordani P, Kiers HAL. A review of tensor-based methods and their application to hospital care data. *Statistics in Medicine*. 2018;37:137–156. <https://doi.org/10.1002/sim.7514>