Chemometrics and Intelligent Laboratory Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Visualising three-way arrays

Darryn Williams*, Sugnet Gardner-Lubbe

Department of Statistical Sciences, University of Cape Town, South Africa

ARTICLE INFO

Article history:
Received 14 February 2016
Received in revised form 31 July 2016
Accepted 7 August 2016
Available online xxxx

Keywords:
Triplot
Tensor Singular Value decomposition
Principal Component Analysis biplot

ABSTRACT

This paper considers the use of the Tensor Singular Value decomposition approximation as a basis for visualising three-way data, particularly in the spirit of the PCA biplot, and proposes a plot based on the orthogonal rank decomposition of a three-way array. It is shown how the decomposition can be used to partition terms as a product of one mode and a combination of the remaining two modes. This allows for a two dimensional representation with triplot axes, similar to PCA biplots.

1. Introduction

Graphical displays of data can often provide a rudimentary understanding of the relationships inherent in the data set under investigation. In the context of two-way data, one such graphical tool is the PCA biplot which [13, p. 492] eloquently describes as “allow[ing] for the analysis of two-way interaction in a table of *n* objects and *p* variables such that systematic patterns between rows, between columns and between columns and rows can readily be assessed and evaluated”. Gower and Hand [10] provide a comprehensive discussion on the theoretical foundations of the construction of this plot. It is relevant that the construction relies on determining the optimal low rank approximation of the data table as well as on the concept of orthogonality. This paper sets out to define a visualisation of three-way data that, in the spirit of Kroonenberg’s definition, could be described as affording the means to analyse three-way interaction such that the systematic patterns between objects, between variables, between conditions, between any combination of these as well as between all three of these collectively can be readily assessed.

This paper provides the theoretical framework for the construction and interpretation of the proposed plot. These ideas are based on those put forward by Araújo [3] but improves on his proposition in three ways. Firstly, it allows the plot to be applied more generally as an exploratory plot than the PARAFAC decomposition, which is what Araújo used, by alleviating the potential problem of degeneracy. Secondly, the interpretation of the proposed plot is closely aligned with that of the PCA biplot, making it more intuitive. Finally, where the original plot as conceived by Araújo did not allow for the reading off of data values, this paper includes linear axes fitted with markers that facilitate this, thus enhancing the original plot. Araújo [3] christened this the triplot while Albers and Gower [1] is, to the best of our knowledge, the first reference in English for this type of plot. Analogous to the term biplot for representing both the rows and columns of a data matrix, here triplot refers to the simultaneous representation of the three modes of a three-way data array. This is however not the only use of the term ‘triplot’ as it is also used for a triangle-shaped plot (see for example [15]) and Gardner-Lubbe [8] uses the term for a plot simultaneously representing the samples, variables and classes in a multiclass classification problem. For ease of reference this name will be used henceforth to refer to the plot discussed in the paper.

The work in this paper lies very closely with the work done by Gower and Albers [1,2]; it is distinct in that the approaches adopted in constructing the visualisation technique are rather different. This paper relies on a geometric approach, whilst Gower and Albers adopted a linear algebraic approach. Reading this paper together with Albers and Gower will provide much additional insight with respect to visualising three-way arrays.

Finally, the methodology is applied to a dataset related to a study of blue crabs shell disease [9] and the inferences made from the plot are compared to their conclusions. This includes a comparison of the proposed plot to the popularly used joint plot in order to show that it yields similar results and a discussion as to why the triplot can arguably be considered slightly more intuitive than the joint plot.

* Corresponding author.
E-mail address: Darryn.williams@gmail.com (D. Williams).

http://dx.doi.org/10.1016/j.chemolab.2016.08.003
0169-7439/© 2016 Published by Elsevier B.V.

2. Theoretical framework

The biplot, as introduced by Gabriel [7], depends on the factorisation of a rank R two-way array matrix $\boldsymbol{X} \in \Re^{d_1 \times d_2}$ as $\boldsymbol{X} = \boldsymbol{G}\boldsymbol{H}^{'}$ where $\boldsymbol{G} \in \Re^{d_1 \times R}$ and $\boldsymbol{H} \in \Re^{d_1 \times R}$. To construct a biplot in r dimensions, for $R > r$, the best rank r approximation of $\boldsymbol{X}$, $\hat{\boldsymbol{X}}$, is obtained from the Eckart and Young [17] theorem, using the singular value decomposition of the matrix $\boldsymbol{X}$. The factorisation of the rank r matrix can be written as

$$\hat{\boldsymbol{X}} = \sum_{i=1}^{r} \sigma_r \boldsymbol{g}_i \circ \boldsymbol{h}_i, \tag{1}$$

where $\boldsymbol{g}_i$ is the i^{th} column of $\boldsymbol{G}$ and $\boldsymbol{h}_i$ is the i^{th} column of $\boldsymbol{H}$ and $\circ$ is the outerproduct operator. For practical purposes, $r = 2$ or $r = 3$ is used to construct the biplot in two or three dimensions respectively.

In the context of three-way tensors defined as $\mathcal{X} := [\![x_{ijk}]\!] \in \Re^{d_1 \times d_2 \times d_3}$, the concept of the rank of a tensor is more complex than what is encountered in the context of matrices (see for example [12]). For the purposes of this paper, a detailed exposition on tensor rank is not necessary; it suffices to provide a brief description of a tensor decomposition that will allow the construction of a triplot in the spirit of the biplot as envisaged by Gabriel [7]. To this end, the definition of the Tensor Singular Value decomposition is defined as specified in [4].

Definition 2.1. A tensor $\mathcal{X} \in \Re^{d_1 \times d_2 \times d_3}$ admits a tensor SVD if it can be written in the form

$$\mathcal{X} = \sum_{r=1}^{R} \sigma_r \boldsymbol{u}_r^{(1)} \circ \boldsymbol{u}_r^{(2)} \circ \boldsymbol{u}_r^{(3)}, \tag{2}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_R > 0$ and $< \boldsymbol{u}_{\boldsymbol{j}}^{(\boldsymbol{n})}, \boldsymbol{u}_{\boldsymbol{k}}^{(\boldsymbol{n})} >= \delta_{ij}$ for $n = 1, 2, 3$. δ_{ij} is the Kronecker delta, σ_r's are the singular values and $\boldsymbol{u}_r^{(n)}$ for $r = 1, 2, \ldots, R$ are the n-mode singular vectors.

An equivalent representation, taken from Chen and Saad, is given by

$$\mathcal{X} = \mathcal{D} \times_1 \boldsymbol{U}^{(1)} \times_2 \boldsymbol{U}^{(2)} \times_3 \boldsymbol{U}^{(3)}, \tag{3}$$

where $\mathcal{D} \in \Re^{R \times R \times R}$ is the diagonal core tensor with $\mathcal{D}_{ii\ldots i} = \sigma_i$, $\times_i$ the matrix tensor multiplication operator as defined in [5] and

$$\boldsymbol{U}^{(n)} = \left(\boldsymbol{u}_1^{(n)}, \boldsymbol{u}_2^{(n)}, \ldots, \boldsymbol{u}_R^{(n)}\right) \in \Re^{d_n \times R}, \tag{4}$$

are orthogonal matrices for $n = 1, 2, 3$. A tensor $\mathcal{X}$ will admit a TSVD if and only if the core tensor arising from a Higher Order Singular Value Decomposition (HOSVD) is diagonalisable but in general this cannot be done [4]. The TSVD definition refers to a tensor of orthogonal rank R being expressed as in Eq. (3).

The importance of this assertion is that a TSVD might not exist to fully decompose a tensor into the sum of R outerproducts where R is the orthogonal rank of the tensor but this does not make a statement regarding the ability to use TSVD to find a lower rank approximation to the tensor $\mathcal{X}$. Here the problem of interest is to minimise

$$\parallel \mathcal{X} - \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i^{(1)} \circ \boldsymbol{u}_i^{(2)} \circ \boldsymbol{u}_i^{(3)} \parallel^2, \tag{5}$$

subject to the constraint that $< \boldsymbol{u}_j^{(n)}, \boldsymbol{u}_k^{(n)} >= \delta_{ij}$ for $n = 1, 2, 3$. It has already been established that this problem might not have a solution if $r = R$, the rank of the tensor $\mathcal{X}$. The contribution of Chen and Saad [4] was to show that the minimisation problem will always have a solution for any $\mathcal{X} \in \Re^{d_1 \times d_2 \times d_3}$ and any $r \leq \min\{d_1, d_2, \ldots, d_n\}$. This decomposition technique does not suffer from the problem of degeneracy that can show itself when PARAFAC decompositions are used. For a particular low rank r, a degenerate solution implies that a low rank-r approximation of the tensor does not exist [6]. Although there is a restriction on r in Chen and Saad; the triplot is two dimensional, making this restriction negligible. Ultimately, using this decomposition technique makes the triplot more generally applicable as an exploratory plot. The low orthogonal rank decomposition is thus the conception of rank that is the basis for constructing the triplot.

2.1. Constructing and interpreting the triplot

2.1.1. Construction of triplot

Note that since the triplot is two-dimensional, it implies that $r = 2$. The TSVD decomposition can be represented as

$$\hat{x}_{ijk} = \sum_{n=1}^{2} \sigma_n \boldsymbol{u}_{in}^{(1)} \boldsymbol{u}_{jn}^{(2)} \boldsymbol{u}_{kn}^{(3)}. \tag{6}$$

As an example, consider the expression for $\hat{x}_{111}$ which is given by

$$\begin{aligned}\hat{x}_{111} &= (\sigma_1 \boldsymbol{u}_{11}^{(1)}) \boldsymbol{u}_{11}^{(2)} \boldsymbol{u}_{11}^{(3)} + (\sigma_2 \boldsymbol{u}_{12}^{(1)}) \boldsymbol{u}_{12}^{(2)} \boldsymbol{u}_{12}^{(3)} \\ &= y_1 y_2 y_3 + z_1 z_2 z_3.\end{aligned} \tag{7}$$

For the purpose of explaining the construction process, the σ_i terms have been grouped with the elements of the first matrix $\boldsymbol{U}^{(1)}$.

With the aid of Fig. 1 it is possible to interpret this representation in a graphical sense. S_1 represents the first row of the matrix $\boldsymbol{U}^{(1)}$ scaled by the singular values plotted relative to Cartesian axes, so labeled because the first mode often refers to subjects. V_1 and T_1 represent the rows of $\boldsymbol{U}^{(2)}$ and $\boldsymbol{U}^{(3)}$ plotted relative to Cartesian axes, labeled as they are due to the fact that variables and time often comprise the second and third modes respectively. V_1T_1 is the result of taking the product of corresponding elements of the vectors $\overline{OV}_1$ and $\overline{OT}_1$. Consider a somewhat different representation of each of the elements comprising the expansion in Eq. (7).

$$y_1 = \overline{OS}_1 \cos(\beta_1 + \theta_1) \qquad z_1 = \overline{OS}_1 \sin(\beta_1 + \theta_1) \tag{8}$$

$$y_2 = \overline{OV}_1 \cos(\alpha_1 + \alpha_2) \qquad z_2 = \overline{OV}_1 \sin(\alpha_1 + \alpha_2) \tag{9}$$

$$y_3 = \overline{OT}_1 \cos(\alpha_1) \qquad z_3 = \overline{OT}_1 \cos(\alpha_1) \tag{10}$$

$$y_2 y_3 = c_1 = \overline{OV_1T_1} \cos(\theta_1) \qquad z_2 z_3 = d_1 = \overline{OV_1T_1} \sin(\theta_1). \tag{11}$$

With this notation Eq. (7) can be written as

$$\begin{aligned}\hat{x}_{111} &= y_1 c_1 + z_1 d_1 \\ &= \overline{OS}_1 \cos(\beta_1 + \theta_1) \overline{OV_1T_1} \cos(\theta_1) + \overline{OS}_1 \sin(\beta_1 + \theta_1) \overline{OV_1T_1} \sin(\theta_1)) \\ &= \overline{OS}_1 \overline{OV_1T_1} \cos(\beta_1 + \theta_1 - \theta_1) \\ &= \overline{OS}_1 \overline{OV_1T_1} \cos(\beta_1) \\ &= \overline{OP}_{S_1} \overline{OV_1T_1},\end{aligned} \tag{12}$$

Please cite this article as: D. Williams, S. Gardner-Lubbe, Visualising three-way arrays, Chemometrics and Intelligent Laboratory Systems (2016), http://dx.doi.org/10.1016/j.chemolab.2016.08.003