A THREE-WAY "INDIRECT" REDUNDANCY ANALYSIS

Laura Marcis [1], Maria Chiara Pagliarella[1] and Renato Salvatore [1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: laura.marcis@unicas.it, mc.pagliarella@unicas.it, rsalvatore@unicas.it)

ABSTRACT: This work introduces a composite Three-Way application of the High Order Singular Value Decomposition. Two of the three component data matrices are processed by a standard Redundancy Analysis. The remaining "external" data matrix is related to the others in a heterogeneous system of relations, that can be well suited to tensor analysis. The external data are set to be linked with the first matrix, while with the second matrix the relations are explained only through multivariate linear regression. An application introduces the method, based on the official data from the Italian Equitable and Sustainable Well-being indicators.

KEYWORDS: Tucker decomposition, high order singular value decomposition, redundancy analysis.

1 Introduction and background

Tensor decomposition (Kolda & Bader, 2009) has the main objective of reducing complex information detected by higher dimensional arrays of data. From a pure statistical perspective, there are two important exploitations of the tensor analysis: the Candecomp/Parafac decomposition and the Tucker decomposition. They play the role of the extension to tensor objects of the principal component analysis (PCa), recognized as an explorative way to approach multidimensional information (Kroonenberg, 2008). In the literature, the most popular tensor decompositions are "Canonical Decomposition" and the "High Order SVD" (HOSVD, De Lathauwer *et al.*, 2000). The HOSVD decomposes an N-mode tensor, as a multidimensional array, in a core reduced-order tensor, multiplied by component matrices alongside each of the N modes. Three-way PCa was the first extension of the PCa to a three-way data set, giving the first useful employment of tensor analysis to explorative statistical analysis. In standard PCa, the components that come from the SVD that summarize individuals are uniquely related to the components that summarize variables. In a three-way PCa the components that summarize entities in each of the modes are related with the remaining two. Redundancy Analysis (RDA, Legendre and Legendre, 2012) was originally introduced in order to capture the effect onto a reduced space $\widehat{\mathbf{Y}}_X = \mathbf{X}\widehat{\mathbf{B}}$ of the linear dependence by a set of criterion variables $\mathbf{Y}$ from a set of predictors $\mathbf{X}$, where $\widehat{\mathbf{B}}$ is the matrix of the ordinary least squares multivariate regression estimates. RDA provides a constrained analysis of the whole linear relations between the two sets of variables, and an unconstrained analysis given by the set of multivariate regression residuals. It can be considered as an extension of multivariate regression because models the effects of the explanatory variables on a response matrix. Partial RDA (pRDA) explores the effects of the predictors in $\mathbf{X}$ on the $\mathbf{Y}$ variables, given the covariates of some additional exploratory variables in a matrix $\mathbf{Z}$. It is a standard RDA performed taking into account the $\mathbf{X}$ variables as predictors on $\mathbf{Y} - \widehat{\mathbf{Y}}_Z$, with the "effect" by $\mathbf{Z}$ removed. Nevertheless, the relations between the variables $\mathbf{Y}$ and $\mathbf{Z}$ may be quite several. While remaining the same the role of the predictors $\mathbf{X}$ on $\mathbf{Y}$, a third set of variables $\mathbf{Z}$ may be related and depend on $\mathbf{Y}$, by an existing but not well defined dependence. Thus, applying multivariate regression may result hardly appropriate. Variables in $\mathbf{Z}$ in some cases can not be modeled on $\mathbf{Y}$ as predictors in a multivariate regression, while $\mathbf{X}$ predict $\mathbf{Y}$ and, indirectly through $\mathbf{Y}$, the variables in $\mathbf{Z}$. Residuals $\mathbf{Y} - \widehat{\mathbf{Y}}_X$ may take in account the role of $\mathbf{X}$ in the "indirect" explanation of $\mathbf{Z}$. This is somewhat different from pRDA, because $\mathbf{Y}$ is not regressed on $\mathbf{Z}$, as the external set of covariates from which we remove the effect on $\mathbf{Y}$, and also $\mathbf{Z}$ is not related with $\mathbf{Y}$ through linear regression. Given a 3rd-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, the Tucker decomposition through the HOSVD decomposes the tensor $\mathcal{X}$ into a core tensor $\mathcal{G}$ and factor matrices along each mode, as follows:

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$

with the correspondent elementwise expression $x_{ijk} = \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{t=1}^{T} g_{rst} a_{ir} b_{js} c_{kt}$, with $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$. The factor matrices are columnwise orthonormal, $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_R], \mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_S], \mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_T]$, with $r = 1, ..., R, s = 1, ..., S, t = 1, ..., T$. The matricized forms, one per mode, of the 3-way tensor $\mathcal{X}$ are:

$$\begin{aligned}
\mathbf{X}_{(1)} &\approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})' = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})', \\
\mathbf{X}_{(2)} &\approx \mathbf{B}(\mathbf{C} \odot \mathbf{A})' = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})', \\
\mathbf{X}_{(3)} &\approx \mathbf{C}(\mathbf{B} \odot \mathbf{A})' = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})',
\end{aligned}$$

with the symbols $\odot$ and $\otimes$ that are the Khatri-Rao and Kronecker products, respectively. If $r_R(\mathcal{X})$ is the rank of the tensor $\mathcal{X}$ alongside one of the modes,

Table 1. *Description of the variables used for the application*

Variables	Description
S8	Age-standardised mortality rate for dementia and nervous system diseases
IF3	People having completed tertiary education (30-34 years old)
L12	Share of employed persons who feel satisfied with their work
REL4	Social participation
POL5	Trust in other institutions like the police and the fire brigade
SIC1	Homicide rate
BS3	Positive judgement for future perspectives
PATR9	Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest
AMB9	Satisfaction for the environment - air, water, noise
INN1	Percentage of R&D expenditure on GDP
Q2	Children who benefited of early childhood services
BE1	Per capita adjusted disposable income
LBE1	Logarithm of Per capita adjusted disposable income

the HOSVD may uses Alternating Least Squares, in order to find:

$$\min_{\mathcal{G},\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathcal{X} - \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \right\| .$$

Making the substitutions $\mathbf{A} = \mathbf{Y}$, $\mathbf{B} = \mathbf{Y} - \widehat{\mathbf{Y}}_X$, $\mathbf{C} = \mathbf{Z}$, with $I = J = K = n$, $R = S = r(\mathbf{Y}) = r(\mathbf{Y} - \widehat{\mathbf{Y}}_X)$, and $T = r(\mathbf{Z})$, we achieve the desired result, by finding a Three-Way version of the "indirect" RDA, with the proper data matrices. Like in the standard RDA, the data in $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ have to be preprocessed by centering and standardazing their column vectors. This is requested before the application of the RDA of $\mathbf{Y}$ on $\mathbf{X}$.

2 Application study

The Equitable and Sustainable Well-being indicators (BES) are designed to define the economic policies which largely act on some fundamental aspects of the quality of life. Table 2 reports the description of these indicators. We use the latter as the predictor variable in the RDA that gives the constrained analysis in the subspace of $\widehat{\mathbf{Y}}_X$. Table 2 reports the correlation matrix between the column vectors of $\mathbf{Y}$, $\mathbf{Y}^*$, and $\mathbf{Z}$. Correlations in bold are significant. It is interesting to remark that in some cases the variables in $\mathbf{Z}$ are correlated with the columns of $\mathbf{Y}$, while they are generally poorly related with the RDA residuals vectors (given by the unconstrained RDA). In particular, the evidence is that even if $\mathbf{Z}$ may be regressed on $\mathbf{Y}$, for some variables the regression on $\mathbf{X}$ results inappropriate. One of the important cases is shown by the variable AMB9. This variable (Satisfaction for the environment - air, water, noise) is permanently correlated with the variable BS3 (Positive judgement for future

Table 2. *Correlations - Matrices* $\mathbf{Y}$, $\mathbf{Y}^{\star}$, *and* $\mathbf{Z}$

Variable	$\mathrm{Y1}_{BS3}$	$\mathrm{Y2}_{INN1}$	$\mathrm{Y3}_{IF3}$	$\mathrm{Y4}_{Q2}$	$\mathrm{Y5}_{L12}$	$\mathrm{Y6}_{S8}$
$Z1_{AMB9}$	**0,4029**	−0,0239	**0,4570**	**0,6852**	**0,8090**	**0,6926**
$Z2_{POL5}$	0,1906	**0,3629**	**0,2594**	**0,6395**	**0,6330**	**0,5973**
$Z3_{PATR9}$	0,1800	**0,3759**	0,0426	0,0353	0,0146	**0,2420**
$Z4_{REL4}$	**0,5133**	**0,2601**	**0,4413**	**0,7026**	**0,8380**	**0,6507**
$Z5_{SIC1}$	**−0,2215**	−0,1150	**−0,4665**	**−0,5397**	**−0,5925**	**−0,6343**
Variable	$\mathrm{Y1}^{\star}_{BS3}$	$\mathrm{Y2}^{\star}_{INN1}$	$\mathrm{Y3}^{\star}_{IF3}$	$\mathrm{Y4}^{\star}_{Q2}$	$\mathrm{Y5}^{\star}_{L12}$	$\mathrm{Y6}^{\star}_{S8}$
$Z1_{AMB9}$	**0,4605**	−0,1075	**0,2848**	0,1294	0,0423	−0,0119
$Z2_{POL5}$	0,0042	−0,1972	−0,0523	0,0662	−0,0624	−0,0755
$Z3_{PATR9}$	−0,1311	0,2081	**−0,2749**	**0,2794**	0,0053	0,1774
$Z4_{REL4}$	**0,3595**	−0,0025	−0,0056	0,0993	−0,1227	−0,1229
$Z5_{SIC1}$	−0,2029	−0,0184	**−0,3021**	−0,1787	−0,0291	−0,0234

perspectives), whatever is $\mathbf{y}$ or $\mathbf{y}^{*} = \mathbf{y} - \widehat{\mathbf{y}}_X$ (with $corr(y, y^{*}) = 0.7293$). We have a moderate correlation between the variable BS3 and the correspondent RDA residuals, and a moderate explanation of this variable is given by the BE1 (Per capita adjusted disposable income). Then, a tentative conclusion is that the "Satisfaction for the environment" (a $\mathbf{Z}$ variable) does not depend on the "Disposable income" (the RDA predictor $\mathbf{X}$). An opposite case occurs when we try to assess the same AMB9 variable, versus L12 (Share of employed persons who feel satisfied with their work). Even we have that $corr(y, y^{*}) = -0.2395$, AMB9 has the greatest correlation with the observed L12 (y), which reduces to be not significant in terms of L12 RDA residuals (y^{*}). Thus, even the "Share of employed persons who feel satisfied with their work" depends on the "Disposable income", and the "Satisfaction for the environment" can be explained by the relation with "People that feel satisfied with their work", the "Satisfaction for the environment" depends on the "Disposable income" through its relation with the "People that feel satisfied with their work".

References

De Lathauwer, Lieven, De Moor, Bart, & Vandewalle, Joos. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, **21**(4), 1253–1278.

Kolda, Tamara G, & Bader, Brett W. 2009. Tensor decompositions and applications. *SIAM review*, **51**(3), 455–500.

Kroonenberg, Pieter M. 2008. *Applied multiway data analysis*. John Wiley & Sons.

Legendre, Pierre, & Legendre, Louis. 2012. *Numerical ecology*. Elsevier.