ELSEVIER

Chemometrics and Intelligent Laboratory Systems 64 (2002) 123–135

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

A multi-way analysis of starch cassava properties

Marlon M. Reis [a], Márcia M.C. Ferreira [a,*], Silene B.S. Sarmento [b]

[a] *Chemistry Institute-UNICAMP, Cidade Universitária Zeferino Vaz, s/n, PO Box 6154, 13083-970 Campinas, São Paulo, Brazil*
[b] *Department of Agroindustry, Food and Nutrition, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Brazil*

Received 31 August 2001; received in revised form 18 June 2002; accepted 26 June 2002

Abstract

The original methods proposed by Ledyard R. Tucker during the 1960s present the rotational freedom problem, making the interpretation of their results rather difficult to be carried out. Aiming to make the multi-way data analysis more acceptable, this work suggests a methodology for extracting meaningful information from the data set. This methodology is based on the decomposition of data set in three-way blocks by using Tucker models. With the aim of keeping in one block similar information about the data properties, a decomposition based on a constrained Tucker model was used, where the core array has some of its elements fixed to zero. This methodology is successfully applied to a data set formed by physical and physicochemical properties of starches of four cassava cultivars, harvested at different ages during the period usually taken for harvest of industrial uses.

Keywords: Constrained Tucker model; Cassava starch properties; Multi-way analysis

1. Introduction

The Tucker models introduced by Tucker [1] during the 1960s for the interpretation of psychological data have been applied to three-way environmental [2] and chemical data for exploratory analysis [3], identification of compounds and second order calibration [4], among others. The original methods proposed by Tucker present the rotational freedom problem, making the interpretation of its results rather difficult to be carried out [5]. A procedure presented to avoid such problem is the rotation of the component matrices of the fitted model. An alternative approach is based on constraining certain parameters in the Tucker model to zero. In such instances, fitting the constrained Tucker model to data set is not meant for assessing which model should be used to describe the data, but for determining which parameter values govern certain aspects of the data [5].

Starch composition and its properties vary mainly with the plant source from which it is derived [6,7]. It is the main constituent of cassava roots. The industrial uses of starch are primarily determined by its physical–chemical properties. This work deals with a data set formed by the properties of four cassava cultivars, harvested at different ages during the period usually taken for harvest of industrial uses [6].

* Corresponding author. Tel.: +55-19-3788-3102; fax: +55-19-3788-3023.
E-mail address: marcia@iqm.unicamp.br (M.M.C. Ferreira).

0169-7439/02/$ - see front matter
PII: S0169-7439(02)00053-9

The complexity of the data set, age and genetic factors makes the data analysis rather complex. The multivariate data analysis on a subset of this data by principal component analysis [8] confirms the importance of using chemometric methods in such kind of study. This data complexity demands the application of more sophisticated methods such as three-way methods.

The methodological aspect of this work lies on the evaluation of the possibility of extracting meaningful information from the data set. It is based on the decomposition of the data set in three-way blocks by using Tucker models in such a way to keep in one block the similar information of starch properties among the cultivars. Three decompositions were tested: the first being a Tucker model without constraints, the second is originated by rotation of the first model and the last one is a constrained Tucker model inspired on the models presented by Kiers et al. [9], where the core array had part of its elements fixed to zero.

The goal of this work is to provide information about carrying out an exploratory data analysis by three-way methods suggesting some tools such as inertia functions for the evaluation of the fitted models. These abilities were verified by the meaningful information extracted from the data set used.

2. Notation

See Appendix A.

3. Data description

The data set consists of the results from experimental measurements of physicochemical and functional properties of starch extracted from four cassava cultivars (i.e. *Manihot esculenta* Crantz: SRT 59-Branca de Santa Catarina, SRT 1287-Fibra, SRT 1105 Mico and IAC 12-829), growing in Campinas, São Paulo State (Brazil) and harvested at eight different ages, 10, 12, 14, 16, 18, 20, 22 and 24 months after being planted (i.e. the first harvest July and the last one in September of the following year). As the harvest period is long, the roots are subjected to well-defined climatic conditions in this region: a rainy season with high temperatures and rainfall, and a drought season with low temperatures and rainfall. The variables studied were: amylose content (AC); granule absolute density (GAD); swelling power (SP) and solubilities (%S) at 60, 75 and 90 °C; water-binding capacity (WBC); enzymatic susceptibility (Glu) expressed in amount of reducing sugars produced after 3, 6 and 24 h; granular specific surface area (GSSA) and pasting properties: peak of viscosity (PV) and retrogradation (R)[1]. These experimental results can be arranged in a three-way data array with modes (age × properproperties × cultivars).

4. Preprocessing

The multi-way array slices (i.e. cultivars matrices (age × properties)) were juxtaposed vertically and the resulting matrix was autoscaled (i.e. the mean value and standard deviation of each column were calculated and each column element was subtracted from its mean and divided by its standard deviation). The preprocessing is illustrated in Scheme 1.

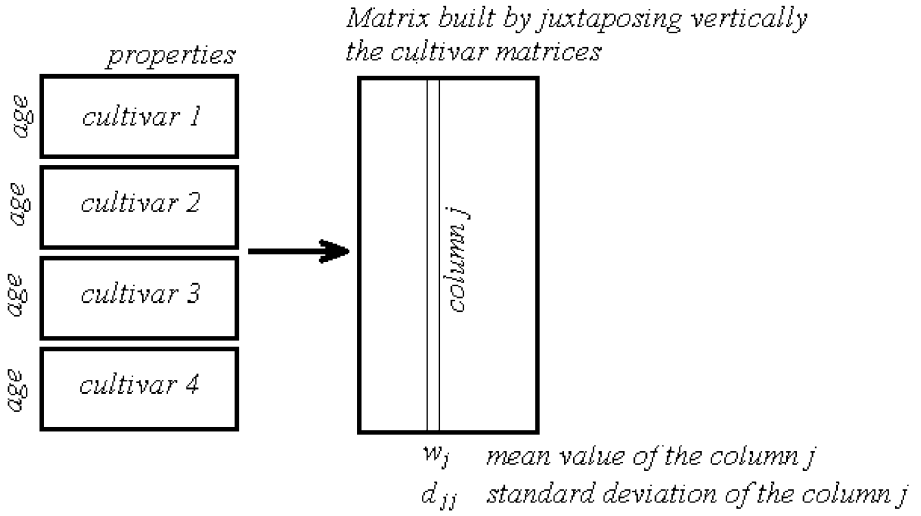
w is a column vector where each element corresponds to the mean value of the columns of a matrix ($[IK] \times J$) and **D** is a diagonal square matrix of order J where the diagonal elements correspond to the inverse of the standard deviations of the columns of vertically juxtaposed cultivar matrices. The unfolded autoscaled three-way array **X** ($I \times [JK]$) can be represented by the following equation

$$\mathbf{X} = [\mathbf{X}_{\text{org}} - \mathbf{1}_I(\mathbf{1}_K^T \otimes \mathbf{w}^T)](\mathbf{I}_K \otimes \mathbf{D}) \quad (1)$$

Here, $\mathbf{X}_{\text{org}}$ ($I \times [JK]$) corresponds to the original unfolded three-way array, $\mathbf{1}_I$ and $\mathbf{1}_K$ are column vectors of ones with I and K elements, respectively, and $\mathbf{I}_K$ is an identity matrix of order K. The vector **w** and matrix **D** were obtained according to Scheme 1 described above. $\mathbf{X}_{\text{org}}$ and **X** are built by juxtaposing horizontally the cultivar matrices, which have as rows the ages mode and as columns the properties modes.

All calculations were performed in an IBM compatible PC using MATLAB® running in Windows®

[1] Figure 1b.



Scheme 1.

system. The Multi-Way toolbox (version 1.08, Oct. 1998) was downloaded from Internet [10].

5. Theory

5.1. Tucker model

Tucker model is the most general model for three-way data. For a three-way array, it is given by Eq. (2).

$$x_{ijk} = \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} \right) + e_{ijk} \tag{2}$$

where a_{ip}, b_{jq} and c_{kr} denote elements of the component matrices **A** (for A mode), **B** (for B mode) and **C** (for C mode) of order $I \times P$, $J \times Q$ and $K \times R$, respectively. Furthermore, g_{pqr} denotes the element (p,q,r) of the $P \times Q \times R$ core array **G**. Finally, e_{ijk} denotes the error term for element x_{ijk} when approximated by a Tucker model and is an element of the $I \times J \times K$ array **E**.

The matrix formulation for the Tucker model is given by Eq. (3).

$$\mathbf{X} = \mathbf{AG}(\mathbf{C}^T \otimes \mathbf{B}^T) + \mathbf{E} \tag{3}$$

where **A**, **B** and **C** denote the component matrices of order $I \times P$, $J \times Q$ and $K \times R$, respectively, **G** denotes the unfolded core array of order $(P \times [QR])$ and finally, **E** denotes the error term by approximating **X** by the Tucker model.

In this work, it is considered the decomposition of the Tucker model in R blocks as demonstrated by Eq. (4), where each term "$\mathbf{AG}_r(\mathbf{c}_r{}^T \otimes \mathbf{B}^T)$" corresponds to one block.

$$\mathbf{X} = \mathbf{AG}_1(\mathbf{c}_1^T \otimes \mathbf{B}^T) + \mathbf{AG}_2(\mathbf{c}_2^T \otimes \mathbf{B}^T) + \cdots + \mathbf{AG}_R(\mathbf{c}_R^T \otimes \mathbf{B}^T) + \mathbf{E} \tag{4}$$

5.2. Inertia functions

Tucker models are fitted by the alternating least squares (ALS) algorithm [11], where the loss function, Eq. (5), is minimized.

$$\text{loss}(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{G}) = \| \mathbf{X} - \mathbf{AG}(\mathbf{C}^T \otimes \mathbf{B}^T) \|^2 \tag{5}$$

The goodness of the ALS algorithm fitting is usually checked by the amount of explained inertia [3]. In this work, the Tucker models were evaluated by means of two kinds of "inertia function", which informs how good is the fitting. The first, "total inertia function", given by Eq. (6), is based on the

loss function which is minimized by the ALS used to fit the Tucker models.

$$f = \left(1 - \frac{\|\mathbf{X} - \mathbf{AG}(\mathbf{C}^T \otimes \mathbf{B}^T)\|^2}{\|\mathbf{X}\|^2}\right) \times 100\% \quad (6)$$

The other function, "partial inertia function", Eq. (7), is used to verify how the core slices describe each slice of the data array.

5.3. Partial inertia function

$$f(\mathbf{G}_k, c_{qk}, \mathbf{X}_q) = \left(1 - \frac{\|\mathrm{vec}\mathbf{X}_q - [(\mathbf{A} \otimes \mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}]\|^2}{\mathrm{vec}\mathbf{X}_q^T \mathrm{vec}\mathbf{X}_\mathrm{q}}\right) \times 100\% \quad (7)$$

where $\mathrm{vec}\mathbf{X}_q$ is the data array slice reshaped as a column vector by the operator vec, $\mathrm{vec}\mathbf{G}_k$ is the core slice reshaped as a column vector and c_{qk} is a C-mode component matrix element.

5.4. Rotated Tucker model

The rotated Tucker model is generated by rotating the core array using an Orthomax [12] rotation for the three modes as follows:

$$\mathbf{X} = \mathbf{ASS}^T\mathbf{G}(\mathbf{U} \otimes \mathbf{T})(\mathbf{U}^T \otimes \mathbf{T}^T)(\mathbf{C}^T \otimes \mathbf{B}^T) \quad (8)$$

since

$$\mathbf{SS}^T = \mathbf{S}^T\mathbf{S} = \mathbf{I},\ \mathbf{U}^T\mathbf{U} = \mathbf{UU}^T = \mathbf{I},\ \mathbf{T}^T\mathbf{T} = \mathbf{TT}^T = \mathbf{I} \quad (9)$$

then

$$\mathbf{X} = \widetilde{\mathbf{A}}\widetilde{\mathbf{G}}(\widetilde{\mathbf{C}}^T \otimes \widetilde{\mathbf{B}}^T) \quad (10)$$

taking

$$\widetilde{\mathbf{A}} = \mathbf{AS},\ \widetilde{\mathbf{G}} = \mathbf{S}^T\mathbf{G}(\mathbf{U} \otimes \mathbf{T}),\ \widetilde{\mathbf{C}} = \mathbf{CU},\ \widetilde{\mathbf{B}} = \mathbf{BT} \quad (11)$$

where **S**, **T** and **U** are the rotation matrices for the A, B and C modes, respectively, and $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{C}}$ corresponds to the rotated component matrices.

5.5. Constrained Tucker model

The constrained Tucker model fits a Tucker model subject to certain constraints on the core. In this case, some elements of the constrained core are fixed to zero. It is also fitted by the ALS algorithm, having as difference the update of the core elements [5,9].

5.6. Methodology

The methodology described in this work for the multi-way data analysis is based in a multi-way decomposition in which the information about the starch properties, common among the cultivars, is captured in a block ($\mathbf{AG}_r(\mathbf{c}_r^T \otimes \mathbf{B}^T)$ in Eq. (4)). For that, the first assumption used in the analysis is that at least one vector of the component matrix, which represents the cultivars mode, must be nonnegative (or to have all the elements with same sign). This assumption suggests that this(ese) vector(s) represents the same kind of information in all cultivars but with different weights. In this case, if any element of vector $\mathbf{c}_r$, for block $\mathbf{AG}_r(\mathbf{c}_r^T \otimes \mathbf{B}^T)$, has a different sign from the others, it means that the information described by $\mathbf{AG}_r\mathbf{B}^T$ for this element will be the opposite of the other elements of $\mathbf{c}_r$ (i.e. each element represents a slice of the C mode, in this work, a cultivar matrix). To make this assumption more clear, consider as an example the total luminescence experiment based on the excitation of a sample in a range of wavelengths resulting in a spectrum of emission intensities for each exciting wavelength [13]. As a result, for each sample there is an emission intensity surface (i.e. excitation × emission). If there is more than one sample, the data set corresponds to a three-way array (i.e. excitation × emission × sample). In this case, if an absorbing specie is present in all samples but in different concentrations, the emission intensity surface for this species corresponds to the "same kind of information" (the emission intensity surface is unique for each species at unity concentration) present in all samples but in different "weights" (concentrations). In the present work, only one vector can be nonnegative (or negative) due to the orthogonality constraint applied in the component matrices (i.e. **A**, **B**, and **C**) for the three modes. Therefore, the core slice associated with this component vector should represent the similar relationship between the plants age

Core slices

Rotated Tucker model Constrained Tucker model

$$\widetilde{\mathbf{G}}_1 = \begin{pmatrix} -9.1 & 0.2 & 0.01 \\ 3 & -1 & -0.1 \\ -0.2 & 0.02 & 0.35 \end{pmatrix} \quad \mathbf{G}_{C-1} = \begin{pmatrix} g_{111-C} & 0 & 0 \\ g_{211-C} & g_{221-C} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$ First built slice

$$\widetilde{\mathbf{G}}_2 = \begin{pmatrix} 1.1 & -0.45 & -0.1 \\ 0.01 & 0.5 & -2.1 \\ -1.2 & 2.2 & 0.35 \end{pmatrix} \quad \mathbf{G}_{C-2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & g_{232-C} \\ g_{312-C} & g_{322-C} & 0 \end{pmatrix}$$ Second built slice

$$\widetilde{\mathbf{G}}_3 = \begin{pmatrix} -0.1 & 0.15 & 0.34 \\ 1.01 & -0.23 & 1.1 \\ 0.4 & -1.1 & 0.15 \end{pmatrix} \quad \mathbf{G}_{C-3} = \begin{pmatrix} 0 & g_{123-C} & g_{133-C} \\ 0 & 0 & 0 \\ 0 & 0 & g_{333-C} \end{pmatrix}$$ Third built slice

Scheme 2.

and their properties for all cultivars. In this way, the aim of this multi-way data analysis is to extract the information about the starch properties during the studied period, which is similar for all cultivars.

The constrained Tucker model has as fundamental problem in designing the core array, in other words, how to choose the core elements that should be fixed to zero. In some applications, the core design is based on the data characteristics [5]. In the exploratory analysis context, the core design should have a structure that makes the data analysis easily carried out. Additionally, the information in one block should be independent from the other blocks. Thus, the elements to be fixed to zero in the constrained core are chosen in this work, at first, considering the orthogonality of the vectors in VEC($\mathbf{G}_C$) (the orthogonality is imposed by the restriction that the nonfixed elements in one slice must be fixed as zero in the other slices), second, by considering the results of the rotated Tucker model and, finally, the values of the partial inertia function which must be nonnegative. In this way, the first constrained slice of the core array to be built is that one associated with the nonnegative vector of the C-mode component matrix found in rotated Tucker model. This is done by choosing the elements, to be fixed to zero, as those having small values on the respective core slice fitted by rotated Tucker model. The second slice is built considering the values magnitude for the respective core elements of rotated Tucker model and the restriction that the nonfixed elements in one slice must be fixed to zero in the other slices. The last parameter to be considered is the function $f(\mathbf{G}_{G-qk})$, the partial inertia function, which must be nonnegative for all core slices and data slices. In this way, a constrained Tucker model is fitted and $f(\mathbf{G}_{G-qk})$ is verified for negative values. If there is one or more negative values, a new constrained model is built and fitted again. A simple example on building the core slices is given in Scheme 2.

6. Results

The first analysis was carried out using the variables amylose content (AC); granule absolute density

Table 1
Singular values for cultivar mode matrices (age × properties), horizontally and vertically juxtaposed

	Singular values					
	BSC[a]	Mico[a]	Fibra[a]	IAC[a]	X1[b]	X2[c]
1	7.9349	4.6923	5.0551	5.9045	10.8104	10.0823
2	4.2190	3.9153	4.1699	3.9585	8.8409	8.3358
3	3.1204	3.4352	3.8742	3.5939	6.5597	6.5593
4	2.3468	2.8444	3.4254	2.8041	6.3111	6.0054
5	2.0159	2.2793	2.9542	2.4515	4.6355	5.1407
6	1.4671	1.8580	2.2703	1.6155	4.3996	4.8358
7	1.0417	1.5234	1.6500	1.2270	3.7283	4.1320
8	0.4981	0.4314	0.7402	0.8817	2.8932	3.0891

[a] Cultivars matrices.
[b] Vertically juxtaposed cultivar matrices.
[c] Horizontally juxtaposed cultivar matrices.

Table 2
Loadings of nonnegative vector of the cultivar mode component matrices

	Loadings		
	TM	RTM	CTM
BSC	0.2141	0.4738	0.4718
Mico	0.4636	0.4886	0.4981
Fibra	0.6163	0.4607	0.5107
IAC	0.5995	0.5697	0.5182

TM—Tucker model.
RTM—rotated Tucker model.
CTM—constrained Tucker model.

(GAD); swelling power (SP) and solubilities (%S) at 60 and 90 °C; water-binding capacity (WBC); enzymatic susceptibility (Glu) expressed in amount of reducing sugars produced after 3, 6 and 24 h; granular specific surface area (GSSA). These data form a three-way data array with modes (age × properties × cultivars) having the dimension ($8 \times 11 \times 4$).

The singular values calculated for scaled data slices and of juxtaposed matrices presented in Table 1 suggest that the rank of the matrices is 8 for the data slices. Thus, Tucker models were fitted with dimension for the A, B and C modes as ($8 \times 8 \times 4$), respectively. The Tucker model fitted 95.80% as measured by the inertia function given in Eq. (6) and constrained Tucker model fitted 85.81%.

Table 2 presents the C-mode nonnegative vectors for the three models. In this table, it is possible to notice that vectors for a rotated Tucker model and a constrained Tucker model present more similar values, a good aspect since the block corresponding to this vector, should describe the starch properties, which is supposed to be similar among plant varieties. Although the nonnegative vectors for the fitted models have presented meaningful variation among their values, the partial inertia function presented negative values for the rotated Tucker model as shown in Table 3. In spite of these negative values being small, they suggest which of the corresponding core slices kept more information than it should have. In other words, the distribution of values in the nonnegative vectors has no physical meaning for the rotated Tucker model, even being correctly calculated as shown in Appendix B.

Table 4 shows the partial inertia function values associated with the nonnegative vectors of the three models; it can be noticed that the vector elements for the constrained Tucker model are more similarly distributed.

The partial inertia function values were important in order to determine the best-fixed to zero positions in the constrained Tucker model core array since the best combination was that one which provided nonnegative values for partial inertia function. The partial inertia function values used as parameter yield information on how the constrained Tucker model describes the data. Negative values for partial inertia function indicates that the nonmodeled part of the data has large influence as shown in Appendix B.

Table 5 shows the core slice for the constrained Tucker model corresponding to the nonnegative vector in the C mode. The element in column 6 and row 7 is the third most important of all core elements. This element corresponds to vector 7 of A-mode component matrix, $\mathbf{a}_7$, and vector 6 of B-mode component matrix, $\mathbf{b}_6$. $\mathbf{a}_7$ gives information about the seasonal effects on starch characteristics. Fig. 1a shows the $\mathbf{a}_7$

Table 3
Partial inertia function values

Cultivars	Partial inertia function (%)							
	RTM				CTM			
	G_1[a]	G_2[a]	G_3[a]	G_4[a]	G_1[a]	G_2[a]	G_3[a]	G_4[a]
BSC	60.3390	22.7970	0.8486	1.8170	26.2985	62.3342	0.1074	0.4401
Mico	− 0.4535	52.9419	− 1.3969	42.6410	50.6217	0.0115	3.4313	26.7182
Fibra	35.8150	47.6668	26.5168	− 0.3346	45.6891	25.7686	7.2344	7.9505
IAC	20.8181	57.8973	7.4225	9.3023	49.4570	16.2763	17.0759	0.7268

RTM—rotated Tucker model.
CTM—constrained Tucker model.
[a] Core slices.

Table 4
Partial inertia function values corresponding to the nonnegative vector of component matrix for the cultivar mode

	Partial inertia function (%)		
	TM	RTM	CTM
BSC	6.7886	22.7973	26.2985
Mico	48.1736	52.9427	50.6217
Fibra	67.0748	47.6670	45.6891
IAC	67.0252	57.8967	49.4570

TM—Tucker model.
RTM—rotated Tucker model.
CTM—constrained Tucker model.

vector, where there are two distinct groups: the first is formed by samples corresponding to the ages of 10, 22 and 24 months (cold and dry season) and the second to the ages of 12, 14, 16, 18 and 20 months (hot and rainy season).

The first group corresponds to the drought period when the plants do not produce or produce only minor amounts of starch or consume this component to provide plant physiology needs, since there was the loss of leaves, followed by the renewal of the plant top. The second group represents the rainy season, stage when there is intense root formation and starch accumulation. It is interesting to notice the correlation between values of the age of the plants and the vector values in the second group. The information represented by the vectors $\mathbf{a}_7$ and $\mathbf{b}_6$ is uniquely determined since only one core slice of constrained Tucker model has the column 6 and row 7 with only one element not fixed to zero. The results presented by the $\mathbf{a}_7$ vector confirm the ability of the constrained Tucker model to keep in one block the common information about the starch structure.

Starch granules are primarily composed of amylose and amylopectin. Amylose is essentially a linear polymer consisting of (1–4)-linked α-D-glucopyranosyl units. Amylopectin is a highly branched polymer of α-D-glucopyranosyl units, primarily linked by (1–4) bonds, with branches resulting from (1–6) linkages, as illustrated by Fig. 2. These two biopolymers can appear in the granules with different conformations, chain length and in different ratios.

Table 5
Core slice for the constrained model corresponding to the nonnegative vector of the component matrix for the cultivar mode

0.04	0	0	0	0	0	0	0
0	−7.10	0	0	0	0	0	0
0	3.93	0	0	0	0	0	1.38
0	0	0	−0.94	0	0	0	0
0	0	2.76	0	0	0	0	0
0	0	0	0	4.03	0	0	0
0	0	0	0	0	−6.04	0	0
0	0	0	0	0	0	−3.64	0

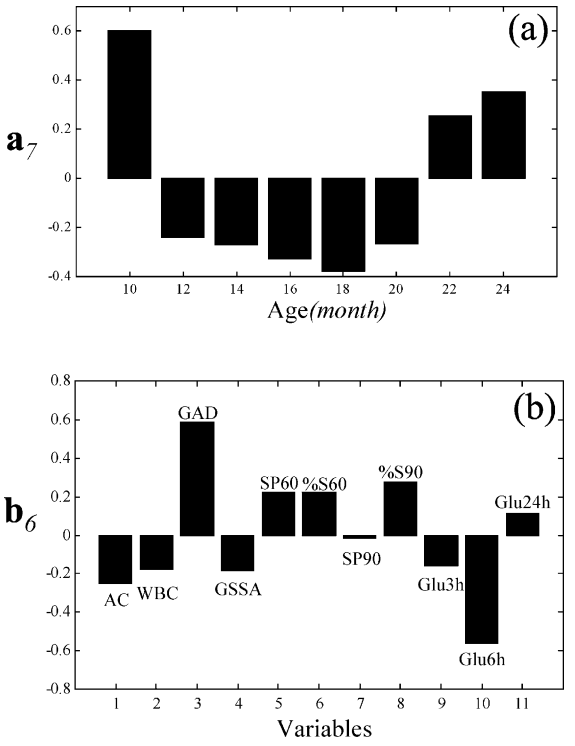


Fig. 1. First analysis. (a) Loadings values for vector 7 in the age mode; (b) loadings values for vector 6 in properties mode (for legend, see data description in the text).

By heating a starch suspension in water, the granules slowly start to absorb water and swell. Over a relatively narrow temperature range, all the granules swell irreversibly and are said to have undergone gelatinization. As the temperature of a starch aqueous suspension is raised above the gelatinization range, intramolecular hydrogen bonds originated from hydroxyl groups continue to be disrupted, water molecules interact with the liberated hydroxyl groups and the granules continue to swell. As a direct result of granule swelling, there is a parallel increase in starch solubility.

The swelling power of the starch, determined by heating a weighed dry starch sample in excess water, is

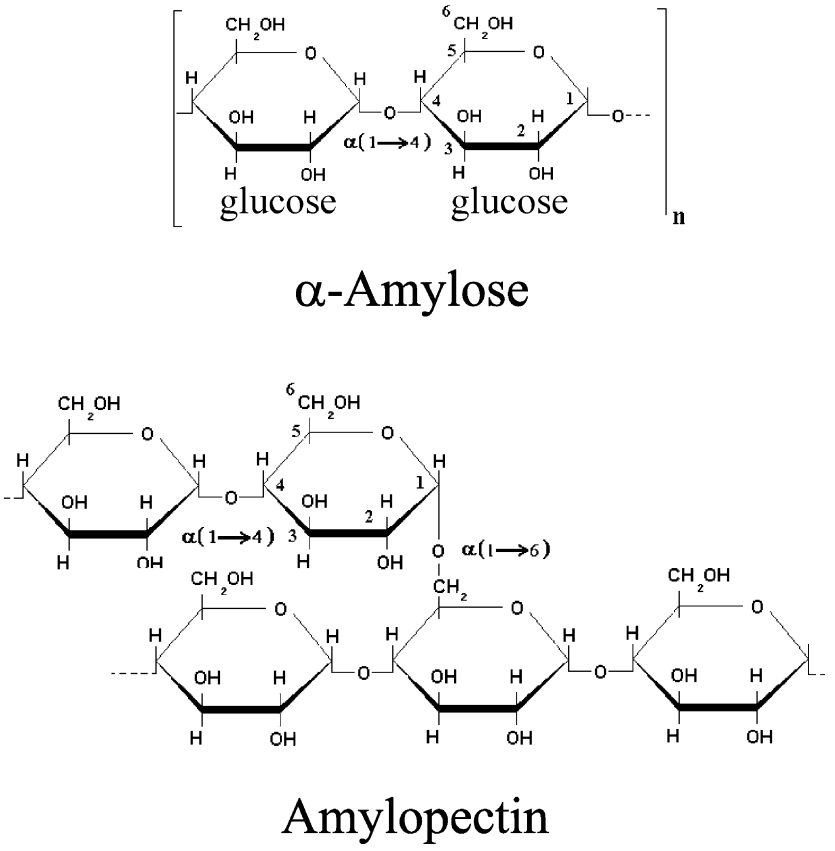


Fig. 2. Chemical structure of α-amylose and amylopectin macromolecules.

defined as the swollen sediment weight per gram of dry starch. The solubilities can also be determined on the same solution [6,7]. According to Leach [14], the strength and character of micellar network within the granule are the major factors controlling swelling behavior of starch. At molecular level, many factors may influence the degree and kind of association between the polymers (i.e. amylose and amylopectin). These factors include the ratio of amylose to amylopectin, the characteristics of each fraction in terms of molecular weight/distribution, degree/length of branching and conformation of these polymers. The presence of the naturally occurring noncarbohydrates such as lipids is also an important factor. The formation of amylose–lipid complexes can restrict swelling and solubilization. The difference in the solubility behavior of the root may be attributed to differences in the conformation of the amylose components in native starch granules. In root starches, lipids are almost absent and the amylose occurs in amorphous state and is converted by heat treatment into a less soluble helical form [6,7]. The loadings in vector $\mathbf{b}_6$ for swelling power and solubilities (see Fig. 1b: SP-60, %S60, %S90, swelling power at 60 °C and percentage of solubilities at 60 and 90 °C) presented a positive correlation with the first group (i.e. samples at age of 10, 22 and 24 months) and a negative correlation with the second group (i.e. samples 12, 14, 16, 18 and 20 months old). The signs of these loadings are opposite to amylose content loading. The values of loadings for solubilities suggest that the difference between those two sample groups is due to amylose conformation and not due to the amount of this macromolecule.

Water molecules, which interact with the macromolecules, are termed ‘bound water’, and reflect the ability of a molecular surface to form hydrogen bonds

with water. The molecular surface available for such 'binding' of water molecules are reduced in the regions with extensive intramolecular hydrogen bonds. The amount of 'bound water' associated with starch granules influences the swelling characteristics of the granules. High 'water-binding' is attributed to the loss of association ability of the starch polymers in native granule. The water-binding sites are considered to be the hydroxyl groups and their interglucose oxygen atoms. During the gelatinization process, the water-binding sites are increased as the heat starts to disrupt the intragranular bonds [6,7]. The loadings in vector $\mathbf{b}_6$ for water-binding capacity (see Fig. 1b, WBC) presented positive correlation with amylose and the second group (i.e. samples at the age of 12, 14, 16, 18 and 20 months). It presented also negative correlation with the first group (i.e. samples at the age of 10, 22 and 24 months), swelling power and solubilities. These differences between those two sample groups are due to the amount of amylose and to the existence of a different number of water-binding sites, which suggests different proportions of amorphous/crystalline areas.

The loadings for granular specific surface area (GSSA) and granule absolute density (GAD) are opposite in signs. The granule absolute density is positively correlated to the first group and the granular specific surface area is positively correlated to the second group. The specific granule surface area is inversely proportional to granule size. This indicates that the first group contains granules bigger in size and more compact than the second group.

The susceptibility of starch granules to digestion by glucoamylase was evaluated by measuring the amount of glucose units (i.e. reducing sugars) produced after a period of enzymatic attack. The glucoamylase cleaves the α(1–4) and α(1–6)-linkages of starch chain producing glucose units. The amorphous part in starch granule is more susceptible to enzymatic attack, making possible the use of enzymatic degradation to study the ratio between the amorphous and the crystalline parts in starch granule. The loadings related to the amount of reducing sugars after 3 and 6 h are positively correlated with amylose content, water-binding capacity and granular specific surface area (granules smaller in size) and with the second group. Amount of reducing sugar produced after 24 h is positively correlated to swelling power (60) and solubilities (60,90) and to the first group.

Two new analyses were done. Firstly, variables about the pasting properties were included. For the second analysis, variables about the pasting properties and swelling power (%SP) and solubilities at 75 °C were included, while the variable of the amount of reducing sugar produced after 24 h was excluded. This was done to verify if the constrained Tucker model would extract the same kind of information in both instances: by including or excluding new variables. The results are shown in Figs. 3 and 4.

The pasting properties showed higher viscosity peaks for starches of the first group and these peaks were negatively correlated to amylose content, water-binding capacity, enzymatic susceptibility and granule specific surface area (or lower size granules). Retrogradation was negatively correlated with viscosity peaks and positively correlated with amylose amounts, the last one being recognized as the main factor influencing this property.

The loadings in vector $\mathbf{b}_6$ show positive correlation among water-binding capacity (more water-bind-

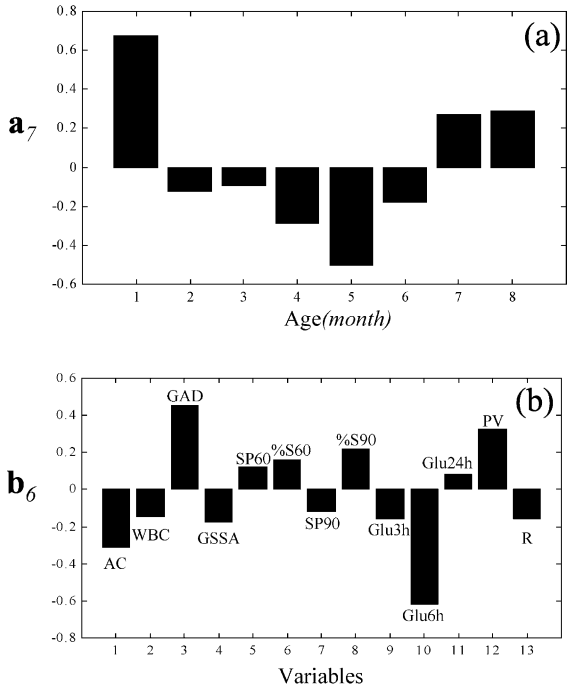


Fig. 3. Second analysis. (a) Loadings values for vector 7 in the age mode; (b) loadings values for vector 6 in properties mode (for legend, see data description in the text).

ing sites), production of reducing sugars in short periods (more susceptible regions for enzymatic attack), specific granule surface area (smaller granule size) and amylose content. These properties and composition are negatively correlated to granule density, swelling power and solubilities and production of reducing sugars in long period. These correlations indicate that the granules of samples in the first group (for rainy season, i.e. 10, 22 and 24 months old, see Figs. 1, 2 and 3a) are more compact with less amorphous regions. Whereas the second group of samples (drought conditions, i.e. samples 12, 14, 16, 18 and 20 months old, see Figs. 1, 3 and 4a) present less compact granules having probably more amorphous regions which are associated with the amylose content.

It should be noticed that the $\mathbf{a}_7$ vector presents an aging gradient. In short, it is positive for the age of 10 months, then its values decrease gradually until the age of 18, and finally its values increase gradually until the age of 24 months. The average monthly rainfall [6] presented in Fig. 5 shows a similar variation of its values for the studied period. In other words, there is a correlation between the average monthly rainfall and the $\mathbf{a}_7$ vector, confirming the seasonal effects on the starch properties.

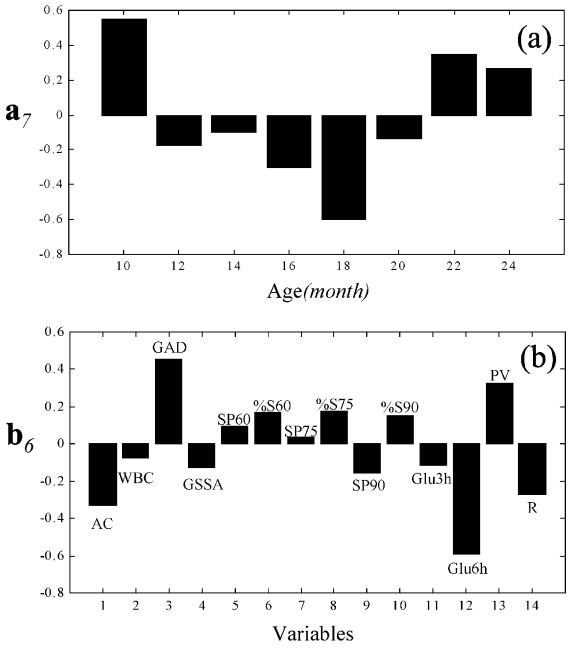


Fig. 4. Third analysis. (a) Loadings values for vector 7 in the age mode; (b) loadings values for vector 6 in properties mode (for legend, see data description in the text).

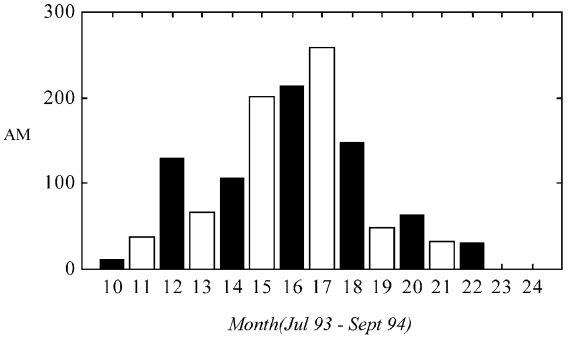


Fig. 5. Average monthly (AM) rainfall for the harvest period, July 93–September 94. The black bars correspond to the month before the harvest.

7. Conclusion

The proposed methodology for the data analysis of the cassava starch properties showed that these starch properties are susceptible to variations in a long-period harvest, these variations being more propitious to happen due to changes in the climatic conditions (i.e. temperatures and rainfall) or due to physiological stage of the plant than to the age of the plant. The used methodology still shows how the changes in the climatic conditions or the physiological stage of the plant act on the starch properties.

The correlation between starch properties, the starch structure, age stage and different cultivars is a complicated puzzle. On the other hand, this data set (periods × properties × cultivars) can be arranged as a three-way data set and analyzed as such. The singular values of slices and unfolded data show its complexity since the rank in different unfolded directions is not reduced; thus, one would argue that the three-way analysis is not worthy. On the other hand, the three-way analysis carried out to extract a target information was able to provide useful information about the data set. The information about starch structure, which is similar among the four cultivars, presents a certain structure in the three-way data set, named as block.

This three-way structure, block, was best approximated by constrained Tucker model. The formulation of the constrained Tucker model was done by considering the independence among blocks, which was formulated theoretically. The inertia function, which gave information about how the data slices are described by the core slice, was fundamental for the final adjustment of the constrained Tucker model. This methodology is interesting since the vectors in the A mode and B mode which show the correlation between properties and age are directly related in one block, making its analysis quite easy.

The information presented by the constrained Tucker model is helpful for choosing the best period to harvest cassava plants since these properties are important for food and industrial use of its starch.

Acknowledgements

The authors acknowledge the financial support from FAPESP (grant number 97/13046-4) for carrying out this work and also Prof. Dr H.A.L. Kiers for sharing the Orthomax codes.

Appendix A

Notation

$\mathbf{X}_q = ({}_q\mathbf{x}_1\ {}_q\mathbf{x}_2 \cdots {}_q\mathbf{x}_M)$ — The q slice matrix of a three-way array; ${}_q\mathbf{x}_{1\ldots M}$ are column vectors

$\mathbf{X} = (\mathbf{X}_1 | \ \mathbf{X}_2 | \ \cdots | \ \mathbf{X}_O)$, $\mathbf{G} = (\mathbf{G}_1 | \ \mathbf{G}_2 | \ \cdots | \ \mathbf{G}_R)$ — Unfolded three-way arrays ($N \times [MO]$) and ($P \times [QR]$)

$\mathbf{A} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_P)$, $\mathbf{B} = (\mathbf{b}_1 \ \mathbf{b}_2 \ \cdots \ \mathbf{b}_Q)$, $\mathbf{C} = (\mathbf{c}_1 \ \mathbf{c}_2 \ \cdots \ \mathbf{c}_R)$ — Loading matrices (component matrices)

$\mathbf{a}_1 = (a_{11} \ a_{21} \ \cdots \ a_{N1})^T$, $\mathbf{b}_1 = (b_{11} \ b_{21} \ \cdots \ b_{M1})^T$, $\mathbf{c}_1 = (c_{11} \ c_{12} \ \cdots \ c_{O1})^T$ — Column vectors

${}^1\mathbf{c} = (c_{11} \ c_{12} \ \cdots \ c_{1R})$ — Row vector

$$\mathrm{vec}\mathbf{X}_q = \begin{pmatrix} {}_q\mathbf{x}_1 \\ {}_q\mathbf{x}_2 \\ \vdots \\ {}_q\mathbf{x}_M \end{pmatrix}$$ — Vectorized form of slice q

$\mathrm{VEC}(\mathbf{X}) = (\mathrm{vec}\boldsymbol{X}_1 \ | \mathrm{vec}\boldsymbol{X}_2 | \cdots | \mathrm{vec}\boldsymbol{X}_O)$, $\mathrm{VEC}(\mathbf{G}) = (\mathrm{vec}\boldsymbol{G}_1 \ | \mathrm{vec}\boldsymbol{G}_2 | \cdots | \mathrm{vec}\boldsymbol{G}_R)$ — VEC(·) denotes an operator used to vectorize each slice, of the unfolded three-way arrays

$$(\mathbf{C}^T \otimes \mathbf{B}^T) = \begin{pmatrix} c_{1,1}\mathbf{B}^T & c_{2,1}\mathbf{B}^T & \cdots & c_{O,1}\mathbf{B}^T \\ \vdots & \vdots & \ddots & \vdots \\ c_{1,R}\mathbf{B}^T & c_{2,R}\mathbf{B}^T & \cdots & c_{O,R}\mathbf{B}^T \end{pmatrix}$$ — Tensorial Kronecker Product

Appendix B

This appendix shows how the function $f(G_k, c_{qk}, \mathbf{X}_q)$ may take negative values or not. For that, the function $f(\mathbf{G}_{G-qk})$ is studied since for the cases where the inequality $f(\mathbf{G}_{G-qk}) \le \mathrm{vec}\mathbf{X}_q{}^T\mathrm{vec}\mathbf{X}_q$ is valid, the $f(G_k, c_{qk}, \mathbf{X}_q)$ is nonnegative.

$$f(G_k, c_{qk}, \mathbf{X}_q) = \left(1 - \frac{\| \mathrm{vec}\mathbf{X}_q - [(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}] \|^2}{\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q}\right) \times 100\% = \left(1 - \frac{f(\mathbf{G}_{G-qk})}{\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q}\right) \times 100\% \quad (12)$$

where $\mathrm{vec}\mathbf{G}_k$ is the column k of the matrix VEC($\mathbf{G}$) and c_{qk} is the kth element of row q of the component matrix $\mathbf{C}$. VEC($\cdot$) denotes an operator used to vectorize each slice, of order $(P \times Q)$, of the unfolded core array $\mathbf{G}$ of order $(P \times [QR])$, the result of which is a matrix of order $([PQ] \times R)$.

Consider the evaluation of the $f(\mathbf{G}_{G-qk})$ used to describe the information of q slice in the data array by the slice k of the core array.

$$f(\mathbf{G}_{G-qk}) = (\{\mathrm{vec}\mathbf{X}_q - [(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}]\}^T \times \{\mathrm{vec}\mathbf{X}_q - [(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}]\}) \quad (13)$$

$$\mathbf{W}_1 = \mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q - \mathrm{vec}\mathbf{X}_q^T(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk} - [(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}]^T\mathrm{vec}\mathbf{X}_q \quad (14)$$

$$\mathbf{W}_2 = [(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}]^T[(\mathbf{A}\otimes\mathbf{B})\mathrm{vec}\mathbf{G}_k c_{qk}] \quad (15)$$

$$f(\mathbf{G}_{G-qk}) = \mathbf{W}_1 + \mathbf{W}_2 \quad (16)$$

Considering that the q slice corresponds to:

$$\mathrm{vec}\mathbf{X}_q = (\mathbf{A}\otimes\mathbf{B})\mathrm{VEC}(\mathbf{G})({}^q\mathbf{c})^T + \mathbf{e} \quad (17)$$

using

$$n = c_{11}\mathrm{vec}\mathbf{G}_1^T(\mathbf{A}^T\otimes\mathbf{B}^T)\mathbf{e} \quad (18)$$

where $({}^q\mathbf{c})$ denotes the qth row of the $\mathbf{C}$ component matrix and $\mathbf{e}$ denotes the nonmodeled part of $\mathrm{vec}\mathbf{X}_q$.

The function $f(\mathbf{G}_{G-q1})$ can be rewritten as follows:

$$\mathbf{W}_1 = (\mathrm{vec}\boldsymbol{X}_q^T\mathrm{vec}\boldsymbol{X}_q) - 2\lfloor({}^q\mathbf{c})\mathrm{vec}\underline{\mathbf{G}}^T\mathrm{vec}\mathbf{G}_k c_{qk}\rfloor - 2n \quad (19)$$

considering the orthogonality of the loadings matrices, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$:

$$(\mathbf{A}^T\otimes\mathbf{B}^T)(\mathbf{A}\otimes\mathbf{B}) = \mathbf{I} \quad (20)$$

and

$$[\mathrm{VEC}(\mathbf{G})]^T\mathrm{VEC}(\mathbf{G}) = \mathbf{\Lambda} \quad (21)$$

where $\mathbf{\Lambda}$ is a positive diagonal matrix since:

$$\mathrm{VEC}(\mathbf{X}) = (\mathbf{A}\otimes\mathbf{B})\mathrm{VEC}(\mathbf{G})\mathbf{C}^T \quad (22)$$

$$(\mathbf{A}^T\otimes\mathbf{B}^T)\mathrm{VEC}(\mathbf{X}) = \mathrm{VEC}(\mathbf{G})\mathbf{C}^T \quad (23)$$

where the columns of VEC($\mathbf{G}$) are orthogonal due to orthogonality among the columns of $\mathbf{C}$ (see Magnus in one mode component analysis). For the constrained model, the $\mathbf{\Lambda}$ is a positive diagonal matrix due to the constrain imposed on the core array (for description of Eq. (22), see Ref. [3], p. 453).

It has

$$[\mathrm{VEC}(\mathbf{G})]^T\mathrm{vec}\mathbf{G}_k = \begin{pmatrix} 0 \\ \vdots \\ \mathrm{vec}\mathbf{G}_k^T\mathrm{vec}\mathbf{G}_k \\ \vdots \\ 0 \end{pmatrix} \quad (24)$$

$$\mathbf{W}_1 = (\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q) - 2c_{qk}^2\mathrm{vec}\mathbf{G}_k^T\mathrm{vec}\mathbf{G}_k - 2n \quad (25)$$

$$\mathbf{W}_2 = c_{qk}^2\mathrm{vec}\mathbf{G}_k^T\mathrm{vec}\mathbf{G}_k \quad (26)$$

$$f(\mathbf{G}_{G-qk}) = (\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q) - c_{qk}^2\mathrm{vec}\mathbf{G}_k^T\mathrm{vec}\mathbf{G}_k - 2n \quad (27)$$

Thus, for $f(\mathbf{G}_{G-qk}) \le \mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q$ to be valid, the scalar $2n$, if negative, must be smaller than or equal to the positive term $c_{qk}^2\mathrm{vec}\,\mathbf{G}_k^T\,\mathrm{vec}\mathbf{G}_k$, which means that

the nonmodeled part of the slice q has small importance if compared to that one described by the slice k of the core array. In those cases where nonmodeled part of the slice one has large importance, $f(\mathbf{G}_{G-qk}) \leq \mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q$ is not valid resulting in negative values for $f(G_k, c_{qk}, \mathbf{X}_q)$. This fact is important on building the constrained Tucker model.

For the rotated model, $f(\mathbf{G}_{G-qk})$ becomes $f(\widetilde{\mathrm{G}}_{\widetilde{G}\mathbf{U}-qk})$ where the hat " ~ " refers to the rotated model. In the same way done for $f(\mathbf{G}_{G-qk})$, it has

$$\widetilde{\mathbf{W}}_1 = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) - 2\left[\left({}^q\widetilde{\mathbf{c}}\right)\mathrm{vec}\widetilde{\mathbf{G}}_q^T\mathrm{vec}\widetilde{\mathbf{G}}_q\widetilde{c}_{qk}\right] - 2\widetilde{n} \tag{28}$$

$$\widetilde{\mathbf{W}}_1 = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{U}\mathbf{U}^T[\mathrm{VEC}(\mathbf{G})]^T \times \mathrm{VEC}(\mathbf{G})\mathbf{u}_k\left({}^q\mathbf{c}\right)\mathbf{u}_k\right] - 2\widetilde{n} \tag{29}$$

see Eq. (11)

$$\widetilde{\mathbf{W}}_1 = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{\Lambda}\mathbf{u}_k\left({}^q\mathbf{c}\right)\mathbf{u}_k\right] - 2\widetilde{n} \tag{30}$$

$$\widetilde{\mathbf{W}}_1 = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right)\right] - 2\widetilde{n} \tag{31}$$

$$\widetilde{\mathbf{W}}_2 = \left({}^q\mathbf{c}\right)\mathbf{u}_k\mathbf{u}_k^T\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) \tag{32}$$

$$f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right) = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right)\right] - \mathbf{2}\widetilde{n} + \left({}^q\mathbf{c}\right)\mathbf{u}_k\mathbf{u}_k^T\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) \tag{33}$$

$$f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right) = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) + \left[\left({}^q\mathbf{c}\right)\mathbf{\Omega}\mathbf{\Lambda}\mathrm{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right)\right] - \mathbf{2}\widetilde{n} + \left({}^q\mathbf{c}\right)\mathbf{u}_k\mathbf{u}_k^T\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) \tag{34}$$

where

$$\mathbf{\Omega} = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

$$f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right) = \left(\mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q\right) + \left({}^q\mathbf{c}\right) \times \left[\mathbf{\Omega} + \mathbf{u}_k\mathbf{u}_k^T\right]\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) - 2\widetilde{n} \tag{35}$$

For $f(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}) \leq \mathrm{vec}\mathbf{X}_q^T\mathrm{vec}\mathbf{X}_q$ to be valid depends on how the model was rotated not only as good as the slice q is described by the model.

References

[1] L.R. Tucker, Psychometrika 31 (3) (1966) 279–311.
[2] E.S. Barcellos, M.M. Reis, M.M.C. Ferreira, submitted for publication.
[3] H.A.L. Kiers, Psychometrika 56 (3) (1991) 449–470.
[4] A.K. Smilde, R. Tauler, J. Saurina, R. Bro, Analytica Chimica Acta 398 (1999) 237–251.
[5] H.A.L. Kiers, A.K. Smilde, Journal of Chemometrics 12 (1998) 125–147.
[6] S.B.S. Sarmento, Caracterização da Fécula de Mandioca (*Manihot esculenta* C.) no Período de Colheita de Cultivares de uso Industrial, PhD thesis, Universidade de São Paulo-Faculdade de Ciências Farmacêuticas-Departamento de Alimentos e Nutrição-1997.
[7] J.E. Rickard, M. Asaoka, J.M.V. Blanshard, Tropical Science 31 (1991) 189–207.
[8] S.B.S. Sarmento, M.M. Reis, M.M.C. Ferreira, M.P. Cereda, M.V.C. Penteado, C.B. Anjos, Brazilian Journal of Food Technology 2 (1–2) (1999) 131–137.
[9] H.A.L. Kiers, J.M.F. ten Berge, R. Rocci, Psychometrika 62 (3) (1997) 349–374.
[10] http://www.models.kvl.dk/source/ (actual site, August 2001).
[11] H.A.L. Kiers, P.M. Kroonenberg, J.M.F. ten Berge, Psychometrika 57 (3) (1992) 415–422.
[12] H.A.L. Kiers, Psychometrika 62 (4) (1997) 579–598.
[13] M.M. Reis, D.N. Bilioti, M.M.C. Ferreira, F.B.T. Pessine, Applied Spectroscopy 55 (7) (2001) 847–851.
[14] H.W. Leach, Gelatinization of starch, in: R.L. Whistler, E.F. Paschall (Eds.), Starch Chemistry and Technology, vol. I, Academic Press, New York, 1965, pp. 289–307.