JOURNAL OF CHEMOMETRICS
J. Chemometrics 2002; **16**: 444–450
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.749

PARAFAC with splines: a case study†

Marlon M. Reis and Márcia M. C. Ferreira*

Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, PO Box 6154, CEP 13083-970 Campinas, SP, Brazil

Received 30 September 2001; Revised 10 May 2002; Accepted 24 May 2002

The PARAFAC model has been used in several applications in chemistry, e.g. for overlapped spectra resolution and second-order calibration. In general, the PARAFAC method uses a vector space approach by considering the matrices resulting from the decomposition as a collection of vectors. This paper presents a PARAFAC application where the factors resulting from the decomposition are considered as functions. The functional objects used for this are spline functions. The methodology used performs the Spline-PARAFAC decomposition based on the Bro–Sidiropoulos approach for the unimodality constraint. One of the advantages of using splines is the possibility of achieving a controlled degree of smoothing on the decomposed components. The amount of smoothing applied on the components in the presented methodology is controlled by a penalty parameter or by the number of basis functions. Thus Spline-PARAFAC requires the calculation of the parameter λ and the number of basis functions, which were determined in this work by using ordinary cross-validation (OCV). Spline-PARAFAC was applied to a carbon monoxide data set comprising concentrations measured every hour during the years 1997 and 1999 in the city of São Paulo, Brazil. Each data set was arranged in a three-way array of dimension (24 hours $\times$ 5 days $\times$ 52 weeks). Spline-PARAFAC showed a good performance, producing smoothed profiles describing the daily variations in emitted gas and the seasonal effects during the year. **Copyright © 2002 John Wiley & Sons, Ltd.**

KEYWORDS: PARAFAC; smoothing splines; carbon monoxide

1. INTRODUCTION

Multiway methods, which first appeared in psychometrics [1,2], have received more and more attention in chemometrics in the last two decades. This family of methods is appropriate for the analysis of large structured data sets, which have become common in chemistry owing to instrumental developments such as hyphenated instruments (e.g. LC–UV, GC–MS, MS–MS).

In general, the chemometrics approach is based on the decomposition of data matrices into latent variables, as for example in the principal component analysis (PCA) method, where a matrix $\mathbf{X}$ is decomposed into the product of the score and loading matrices (i.e. $\mathbf{X} = \mathbf{T}\mathbf{P}^{T}$, where the superscript denotes the transposed matrix). In the same way, three-way methods can be considered as an extension of PCA-like methods to multiway data, since it is possible to perform a similar kind of data analysis.

The multiway methods developed in psychometrics, especially the PARAFAC and Tucker models, have been used for second-order calibration, curve resolution and other chemical applications, after some refinements that are needed owing to unpredictable variation in the experiments. In curve resolution, for example, when applied on overlapped spectra and time profiles, the aim of multiway methods is to fit each of the overlapped spectra and time profiles. In this case a non-negative constraint has been shown to be useful, since it is known *a priori* that the components to be fitted are non-negative. Smoothness, which can be achieved by several methods, may be required to avoid rapid variation in the decomposed profiles. In the multiway analysis context, Bro [3] applied PARAFAC with a smoothing constraint based on a penalty approach for curve resolution of fluorescence data. Timmerman and Kiers [4] have also used a smoothing spline approach for three-way component analysis.

The spline, which historically originated in engineering to draw a smooth curve between specified points, has become a mathematical term, consisting in the solution of a constrained optimization problem. Splines are purely interpolatory in nature. Although interpolating splines are useful for non-noisy data, which have limited use in experimental data analysis, there is a type of smoothing spline that makes it possible to describe the data but not be constrained to

*Correspondence to: M. M. C. Ferreira, Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, PO Box 6154, CEP 13083-970 Campinas, SP, Brazil.
E-mail: marcia@iqm.unicamp.br
†Paper presented at the 7th Scandinavian Symposium on Chemometrics, Copenhagen, Denmark, 19–23 August 2001.
Contract/grant sponsor: FAPESP; Contract/grant number: 97/13046-4; Contract/grant number: 99/09643-2.

Copyright © 2002 John Wiley & Sons, Ltd.

interpolating exactly [5]. Ramsay and Silverman [6], Besse and Ramsay [7] and Silverman [8] have shown the usefulness of functional analysis applied on principal component-like methods by means of smoothing spline-like methods. In chemometrics, splines have been used for curve fitting [9], data compression [10] and linearization of non-linear regression problems [11].

An important topic to be considered when analysing a gas concentration in the atmosphere is the correlation between the gas emission sources and the atmospheric conditions. An interesting example mentioned by Comrie and Diem [12] refers to a taxi strike in the city of New York when carbon monoxide (CO) emission was reduced by 34% but, owing to coincident low wind speeds, its concentration in the atmosphere was reduced by only a few per cent. The data set studied in this work is built from measurements of CO concentration collected every hour for a year. We consider the data as having a three-way structure with the following modes: hours of the day × days of the week × weeks of the year (HD × DW × WY). The first mode, HD, represents the emission during the day; the second mode, DW, refers to the contribution of the days of the week to the CO emission; and the last mode, WY, represents the contribution from seasonal effects during the year. Additionally, the profiles for the HD mode are assumed to be systematic and with a gradual variation (i.e. without random changes) within their elements. Thus the profiles for the HD mode would be related to systematic sources of CO emission and the profiles for the WY mode would represent changes in climatic conditions due to the different seasons during the year. These characteristics, i.e. systematic profiles and smooth variation, are imposed on the model by considering the profiles for each mode as functions, which is done here by means of splines.

Paatero and Juntto [13] have carried out a data analysis of carbon monoxide data using their method to perform a PARAFAC-like decomposition.

The aim of the present work is to fit the systematic variation for the HD mode and verify the presence of seasonal effects in the WY mode. Additionally, the usefulness of combining PARAFAC and functional analysis, by using Bro's approach to applying the unimodality constraint [3], is evaluated.

2. THE DATA

The data set comprises measurements of carbon monoxide (CO) concentration in the city of São Paulo, Brazil during the years 1997 and 1999 collected every hour, every day.

3. METHODS

PARAFAC is based on the decomposition of a multiway data set into a linear combination of multilinear components [14,15]. This decomposition can be constrained if the data set demands it, e.g. a non-negative constraint can be used for curve resolution of time profiles and spectra. The approach described in this work considers the PARAFAC factors as functions.

Before starting the description of the functional objects, it is necessary to describe the PARAFAC method represented by Equation (1) for a three-way array:

$$\mathbf{X} = \mathbf{A}\mathbf{I}_{\mathrm{DS}}(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) + \mathbf{E} \tag{1}$$

where $\mathbf{X}$ ($M \times [N{\cdot}R]$) and $\mathbf{I}_{\mathrm{DS}}$ ($F \times [F{\cdot}F]$) denote the matrix representation of the data set and superdiagonal three-way arrays respectively. The matrix $\mathbf{X}$ is built by juxtaposing horizontally R matrices of dimension ($M \times N$), which are called slices, e.g. 52 matrices of dimension (24 hours × 5 weekdays). F denotes the number of trilinear components fitted. The matrix form of the superdiagonal three-way array is built in the same way as $\mathbf{X}$, where each slice is a square ($F \times F$) matrix which has only one element different from zero and equal to one, i.e. the (f, f) element of its diagonal, where f is the slice number. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are the component matrices of dimensions ($M \times F$), ($N \times F$) and ($R \times F$) respectively [16]. $\mathbf{E}$ ($M \times [N{\cdot}R]$) corresponds to the part of $\mathbf{X}$ that cannot be accommodated in the trilinear model.

The estimates for parameters (i.e. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$) of the PARAFAC model described in Equation (1) can be determined by an alternating least square (ALS) algorithm where the component matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are found one at each step. This formulation is called global by Bro [3]. Another approach suggested to find $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ is the columnwise formulation [3]. Here Equation (1) is first rewritten as Equation (2) [3,16]:

$$\begin{aligned}\mathbf{X}_{\mathrm{A}}^{\mathrm{T}} = \mathbf{Z}_{\mathrm{A}}\mathbf{A}^{\mathrm{T}} = \mathbf{z}_{\mathrm{A},1}\mathbf{a}_1^{\mathrm{T}} + \mathbf{z}_{\mathrm{A},2}\mathbf{a}_2^{\mathrm{T}} + \cdots \\ + \mathbf{z}_{\mathrm{A},f}\mathbf{a}_f^{\mathrm{T}} + \cdots + \mathbf{z}_{\mathrm{A},F}\mathbf{a}_F^{\mathrm{T}} + \mathbf{E}\end{aligned} \tag{2}$$

where

$$\mathbf{Z}_{\mathrm{A}} = [\mathbf{I}_{\mathrm{DS}}(\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}})]^{\mathrm{T}} = (\mathbf{z}_{\mathrm{A},1}|\mathbf{z}_{\mathrm{A},2}|\cdots|\mathbf{z}_{\mathrm{A},F}) \tag{3}$$

and $\mathbf{X}_{\mathrm{A}}$ is the same as matrix $\mathbf{X}$, but where the subscript A is used to denote that the component matrix $\mathbf{A}$ is not participating directly in the product ($\cdots\otimes\cdots$).

The global PARAFAC decomposition is obtained by an ALS optimization where the function described by Equation (4) is minimized:

$$\begin{aligned}l(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_F) = \|\mathbf{X}_{\mathrm{A}}^{\mathrm{T}} - (\mathbf{z}_{\mathrm{A},1}\mathbf{a}_1^{\mathrm{T}} + \mathbf{z}_{\mathrm{A},2}\mathbf{a}_2^{\mathrm{T}} + \cdots \\ + \mathbf{z}_{\mathrm{A},f}\mathbf{a}_f^{\mathrm{T}} + \cdots + \mathbf{z}_{\mathrm{A},F}\mathbf{a}_F^{\mathrm{T}})\|^2\end{aligned} \tag{4}$$

The columnwise formulation is found by rewriting Equation (4) as Equations (5) and (6), where the fth multilinear component fitting is represented as shown in Equation (7) for mode A:

$$\begin{aligned}l_f(\mathbf{a}_f) = \|(\mathbf{X}_{\mathrm{A}}^{\mathrm{T}} - \mathbf{z}_{\mathrm{A},1}\mathbf{a}_1^{\mathrm{T}} - \mathbf{z}_{\mathrm{A},2}\mathbf{a}_2^{\mathrm{T}} - \cdots \\ - \mathbf{z}_{\mathrm{A},F}\mathbf{a}_F^{\mathrm{T}}) - \mathbf{z}_{\mathrm{A},f}\mathbf{a}_f^{\mathrm{T}}\|^2\end{aligned} \tag{5}$$

$$^{(-f)}\mathbf{Y} = \mathbf{X}_{\mathrm{A}}^{\mathrm{T}} - \mathbf{z}_{\mathrm{A},1}\mathbf{a}_1^{\mathrm{T}} - \mathbf{z}_{\mathrm{A},2}\mathbf{a}_2^{\mathrm{T}} - \cdots - \mathbf{z}_{\mathrm{A},F}\mathbf{a}_F^{\mathrm{T}} \tag{6}$$

$$l_f(\mathbf{a}_f) = \|^{(-f)}\mathbf{Y} - \mathbf{z}_{\mathrm{A},f}\mathbf{a}_f^{\mathrm{T}}\|^2 \tag{7}$$

Thus, for every fitting of the factor f, the minimum of Equation (7) is sought.

The solution for the optimization problem of minimizing the function described by Equation (7) when $\mathbf{a}_f$ is under constraint is equivalent to the problem [3,16]

$$\min_{\boldsymbol{\alpha}} \|\mathbf{a}_f - \boldsymbol{\alpha}_f\|^2 \tag{8}$$

Copyright © 2002 John Wiley & Sons, Ltd.

where $\boldsymbol{\alpha}_f$ is subjected to a constraint and $\mathbf{a}_f$ is the unconstrained least square solution for the problem described by Equation (7).

3.1 The functional constraint

The unconstrained least square solution for the problem expressed in terms of Equation (7) can incorporate some information in the PARAFAC factors which is not directly related to the phenomenon studied. Therefore it would be interesting to incorporate some extra information, based on characteristics of the data set, to constrain the PARAFAC fitting. In this work the factors for one mode are required to describe a systematic variation, whereas the variation within the elements of each factor must be smooth. Thus splines are used to impose these characteristics on the PARAFAC factors. A special problem in applying functional methods (i.e. splines) to experimental data is due to sudden high local variation. The goal is to obtain a spline representation of the data that reflects some natural behaviour without being affected by random changes (rapid local variation). Producing a good fit and avoiding too much rapid local variation can be achieved by incorporating some regularization into the fitting of the PARAFAC components. One simple method to obtain this regularization is to represent the components by a linear combination of basis functions, where the degree of regularization is controlled by the number of basis functions. Another method is to measure such rapid local variation by a roughness penalty parameter [6,17]. In this work the roughness penalty parameter used to smooth the fth vector of a component matrix is represented by the integrated squared second derivative added to the minimized expression in Equation (8), as shown in the following:

$$l_{f\lambda}(\alpha_f) = \sum_t (a_{f,t} - \alpha_{f,t})^2 + \lambda \int (g_f'')^2 \mathrm{d}t \tag{9}$$

where

$$\mathbf{a}_f^{\mathrm{T}} = (\mathbf{z}_f^{\mathrm{T}}\mathbf{z}_f)^{-1}\mathbf{z}_f^{\mathrm{T}(-f)}\mathbf{Y} \tag{10}$$

$$\boldsymbol{\alpha}_f = \left(g_f(t_1)g_f(t_2)\cdots g_f(t_M)\right)^{\mathrm{T}} \tag{11}$$

and $\boldsymbol{\alpha}_f$ denotes the fth column vector of a component matrix having its values calculated by the function $g_f(t)$ on the t points in a given interval. The term $\int (g_f'')^2 \mathrm{d}t$ in Equation (9) is responsible for the curvature of the function $g_f(t)$ (or the rate of exchange between residual error and local variation). In this way, changing the value of λ causes the $\int (g_f'')^2 \mathrm{d}t$ value to be changed and consequently the curvature of $g_f(t)$ to be adjusted [17], where the double-prime superscript denotes the second derivative of $g_f(t)$.

The algorithm for the solution of Equation (9) is not discussed here but can be found elsewhere [6,18–21].

The function $g_f(t)$ used to represent the PARAFAC components is a linear combination of basis functions, as mentioned before. Two of the most common basis functions used to represent a data set are B-spline functions and Fourier series. B-splines have the computational advantage of being represented by a basis function matrix (to be described next) which is a banded matrix (i.e. one whose elements are zero everywhere except over a finite interval). For those cases where the data are periodic, Fourier series are indicated. As an advantage, Fourier series can produce an orthogonal basis function matrix when the data points are equally spaced. The function $g_f(t)$ represented by B-spline functions is given by Equation (12) (see References [5,9,18] for an introduction and Reference [19] for technical details):

$$g_f(t) = \sum_{j=1}^{n_{\mathrm{basis}}} \boldsymbol{\tau}_j Q_j(t) = \mathbf{Q}\boldsymbol{\tau} \tag{12}$$

where $Q_j(t)$ is the jth column vector of the basis function matrix $\mathbf{Q}$, τ is the vector of τ_j coefficients and n_{basis} is the number of B-spline basis functions. The knot sequence is ξ_σ, $\sigma = 1,2,\ldots,n_{\mathrm{knots}}$ plus the boundary knots ξ_0 and $\xi_{\sigma+1}$, where n_{knots} is the number of knots.

The description of the $g_f(t)$ function in terms of a Fourier series is given in Equation (13):

$$g_f(t) = c_0 + c_1 \sin(\widetilde{\omega}t) + c_2 \cos(\widetilde{\omega}t) + c_3 \sin(2\widetilde{\omega}t) + c_4 \cos(2\widetilde{\omega}t) + \cdots \tag{13}$$

which can be described by

$$g_f(t) = \boldsymbol{\phi}\boldsymbol{\gamma} \tag{14}$$

where the columns of the basis function matrix $\boldsymbol{\phi}$ are $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin(r\widetilde{\omega}t)$ and $\phi_{2r}(t) = \cos(r\widetilde{\omega}t)$. The parameter $\widetilde{\omega}$ determines the period $2\pi/\widetilde{\omega}$, which is equal to the length of the interval T, and $g_f(t)$ is periodic [6]. The parameter r defines the number of cycles (each cycle is a full period). For instance, in a matrix for a basis with five functions the first column is a vector of ones, the second and third columns correspond to $r = 1$ and the fourth and fifth columns correspond to $r = 2$. In this work the number of basis functions is taken as an odd number in order to produce a complete basis set. The vector $\boldsymbol{\gamma}$ is the vector of coefficients.

3.2 Choosing the smoothing parameter λ and the number of basis functions

The smoothing, as mentioned before, is accomplished here by two methods: one using a relatively small number of basis functions and the other using a penalty parameter. The first method is based on achieving the regularization of the profiles only by the penalized least square approach (for an odd number of data points the the number of basis functions is the same, otherwise it is one less). The second uses the penalized least square approach for one mode and the smoothing control by the number of basis functions for another mode. A similar approach was used by Timmerman and Kiers [4].

The CO data may present a systematic variation with a period of 24 hours and a seasonal effect during the year due to changes in the climatic conditions. Thus PARAFAC was tested with two different combinations of splines. Fourier series were chosen as basis functions to represent the hours of the day (HD), because this mode is periodic and one of the characteristics required for the factors of this mode is to present a systematic, or periodic, variation among their elements. The fitting of this mode was regularized by the penalized least square approach, and the weeks of the year (WY) mode was smoothed by Fourier series or B-spline basis

Copyright © 2002 John Wiley & Sons, Ltd.

Table I. Outline of methodology

Mode	Smoothing method		Method name	
	Basis functions	Regularization control	Method A	Method B
Hours of the day (HD)	Fourier series	Penalty parameter (λ) (penalized least square)	×	×
Weeks of the year (WY)	Fourier series	Penalty parameter (λ) (penalized least square)	×	
	Cubic B-splines	Number of basis functions		×
Days of the week (DW)	Considered as constant	–	×	×

functions. In the first case the WY mode profile was regularized by the penalized least square approach and in the second case the amount of smoothing was controlled by the number of B-spline basis functions. The contribution of the days of the week (DW) mode was considered as constant, since the contribution of the weekdays is expected to be the same (see Section 3.3). Table I summarizes the methodologies described above, and the two combinations of smoothing methods are termed method A and method B.

When the PARAFAC fitting is constrained by method A, the smoothing depends on one parameter, i.e. the penalty parameter. For method B, which uses B-splines for the WY mode, there are several parameters to be adjusted (i.e. number of knots, knot positions, penalty parameter, spline order). In this case the PARAFAC fitting is prohibitive because of the large number of parameters to be optimized. Thus the number of parameters used for the B-splines was reduced by using equally spaced knots, with the number of knots given by $n_{\text{knots}} = n_{\text{basis}} - (ord_{\text{pol}} + 1) + 2$, where ord_{pol} is the polynomial order, $ord_{\text{pol}} + 1$ is the number of knots needed to span the spline space, and the value 2 corresponds to the boundary knots. We also used cubic B-splines (i.e. $ord_{\text{pol}} = 3$) as basis functions.

The choice of the penalty parameter or the number of basis functions must produce a realistic curve that expresses some characteristics of the phenomenon represented by the data set. Additionally, every factor fitting must preserve the convergence of the ALS algorithm. In this work the penalty parameter λ in Equation (9) and the number of basis functions for the B-splines were found by using an ordinary cross-validation (OCV) method [17,20]. This method is based on the principle of leaving the data points out one at a time and choosing that value for the desired parameter when the missing data points are best predicted by the remainder of the data set. The best value for the parameter is the one which minimizes the cross-validation scores given by Equation (15):

$$\text{OCV}(\lambda) = \sum_{i=1}^{l} (a_{f,i} - {}^{\lambda,i}\alpha_{f,i})^2 \tag{15}$$

where ${}^{\lambda,i}\alpha(t)$ is the resulting function from the approximation due to Equation (9) when fitted without the point i, and ${}^{\lambda,i}\alpha_{f,i}$ is the ith value calculated by the function ${}^{\lambda,i}\alpha(t)$ for the parameter λ.

For method B the variable in the function OCV is the number of basis functions, and the term $\int (g''_f)^2 \, dt$ in Equation (9) vanishes since it was not penalized.

OCV is chosen to determine the penalty parameter because it minimizes the sum of squares given in Equation (8), as pointed out by Hastie and Tibshirani [18], and consequently preserves the ALS algorithm convergence. The OCV method as described by Equation (15) is time-consuming, but an efficient calculation of the cross-validation score can be used [21].

All the calculations were performed on an IBM-compatible PC using MATLAB® running under Windows®. The spline toolbox used on this work is described in the book by Ramsay and Silverman [6]. The toolbox was obtained through the internet [22].

3.3 PARAFAC model for CO data

As mentioned in Section 1, the gas concentration in the atmosphere depends on, among other factors, the gas emission and the climatic conditions. In short, if the climatic conditions are not favourable for gas dispersion, a sample of the gas concentration in the atmosphere is measured, otherwise only a fraction of it is measured. In this way the gas data analysis would be performed by a model that discriminates between the gas source contribution and the contribution to the gas dispersion due to changes in the climatic conditions. In this work the concentration of carbon monoxide (CO) in the atmosphere is analysed. In this case the CO emission is assumed to be periodic, with a period of 24 hours. Considering that this periodic characteristic of the CO emission can be used to identify the emission source and that the climatic conditions are not equal during the year [23], acting differently on the CO dispersion, a PARAFAC model is suggested in Equation (16) for the data analysis:

$$\hat{x}_{mnr} = \sum_{f=1}^{F} a_{mf} b_{nf} c_{rf} \tag{16}$$

where $\hat{x}_{mnr}$ is the estimated CO concentration at hour m of day n in week r, a_{mf} is proportional to the emission of source f at hour m, b_{nf} is the contribution of day n of the week for source f, c_{rf} is proportional to the seasonal effect for week r on source f, and F is the number of factors (sources).

For the data set dealt with in this work, it is known *a priori* that automobile traffic is the main source of carbon monoxide [23] at the measurement site. Thus the amount of gas collected at the site depends mainly on the number of automobiles and on the climatic conditions. It is assumed that the automobile traffic is the same for the 5 weekdays during the whole year (the weekend days would present a different daily periodic variation during the year), with a 24 hour periodic variation in the number of automobiles. It is also assumed that there are no significant changes in the

Copyright © 2002 John Wiley & Sons, Ltd.

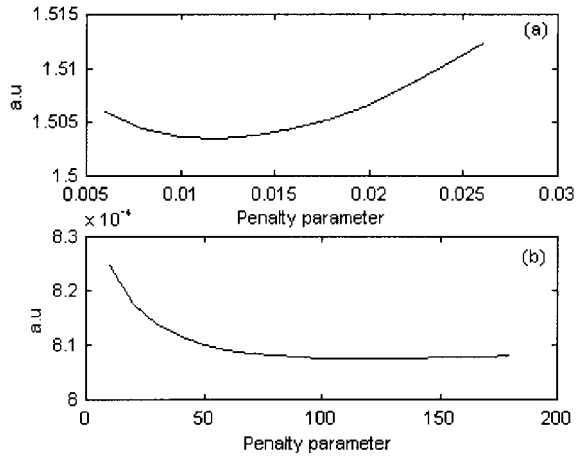

Figure 1. Ordinary cross-validation functions (for penalty parameter using method A) for (a) hours of day mode and (b) weeks of year mode for 1999 (a.u., arbitrary unit).

Table II. Results for ordinary cross-validation fitting of spline parameters

		Year	
Method	Regularization control	1997	1999
A	Penalty parameter λ(HD)	0.149	0.012
	Penalty parameter λ(WY)	180.000	120.000
B	Penalty parameter λ(HD)	0.151	0.012
	Number of basis functions (WY)	5	5

climatic conditions during each week, i.e. sudden changes are considered to be random. Based on these assumptions, the CO data sets for the years 1997 and 1999 were arranged as three-way arrays of dimension $25 \times 5 \times 52$, where the 25 in the HD mode corresponds to 24 hours plus 1 hour to complete a full cycle (i.e. a period from zero to 2π).

4. RESULTS

A one-factor PARAFAC model was fitted to the CO data sets to verify the possibility of describing the automobile traffic contribution and how it is affected by the climatic changes during the year. The resulting PARAFAC components are the following: **a** is the systematic profile (the automobile contribution), which represents the HD mode; **b** is fixed, having unit elements for all five days of the DW mode; and **c** is used to describe the seasonal effects during the year in the WY mode.

Two typical cross-validation curves for the year 1999 fitted by method A (i.e. these curves correspond to the last fit) are shown in Figure 1 and the results for the years 1997 and 1999 are summarized in Table II.

Figure 2 shows the PARAFAC components for the HD mode for the two years. These components describe the daily systematic variation, showing a high correlation with the traffic flow during the day. The positions of peaks and valleys are in agreement with the automobile traffic as described by a report from the traffic department for April of these two years [24].

The PARAFAC components for the WY mode are shown in Figures 3 and 4 for the two years. The two methods described in Table I differ in the fitting of the WY mode. From the results shown in Figures 3 and 4, it is possible to verify small differences between the profiles fitted by methods A and B. Comparing these differences with the variation within the elements of the profiles found by unconstrained PARAFAC (Figure 3) and the differences between the smoothed profiles by the same method for the two years (Figure 4), it is reasonable to suggest that the differences due to methods A and B are not significant with respect to the different kinds of variation, i.e. rapid local variation and variation due to different years.

Seasonally, the CO concentrations are highest in the dry season during stagnant conditions. Among the factors responsible for these stagnant conditions, a high frequency of atmospheric temperature inversion and a low wind speed are important factors for less propitious conditions for CO

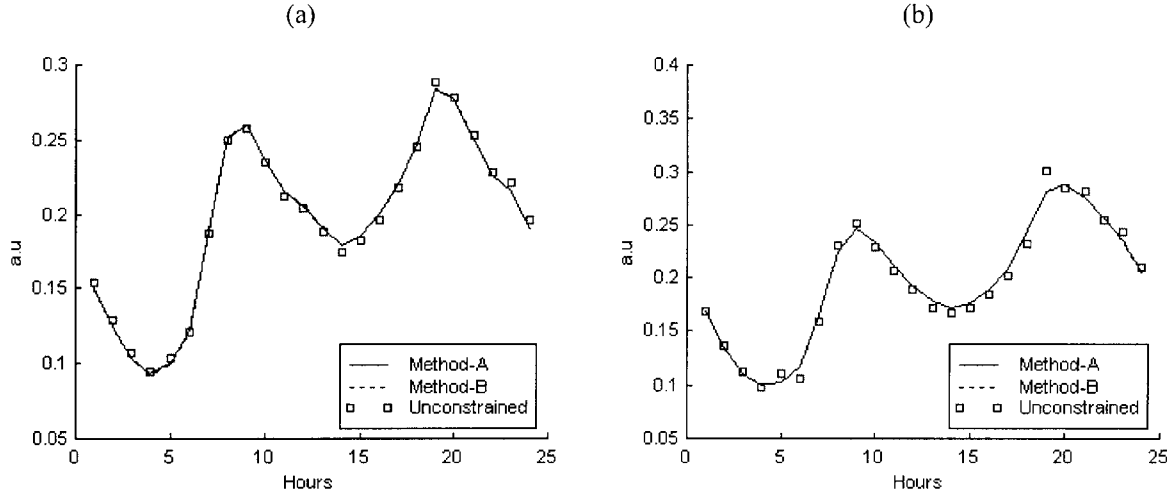

Figure 2. PARAFAC loadings for hours of day mode for (a) 1999 and (b) 1997 (a.u., arbitrary unit).

Copyright © 2002 John Wiley & Sons, Ltd.

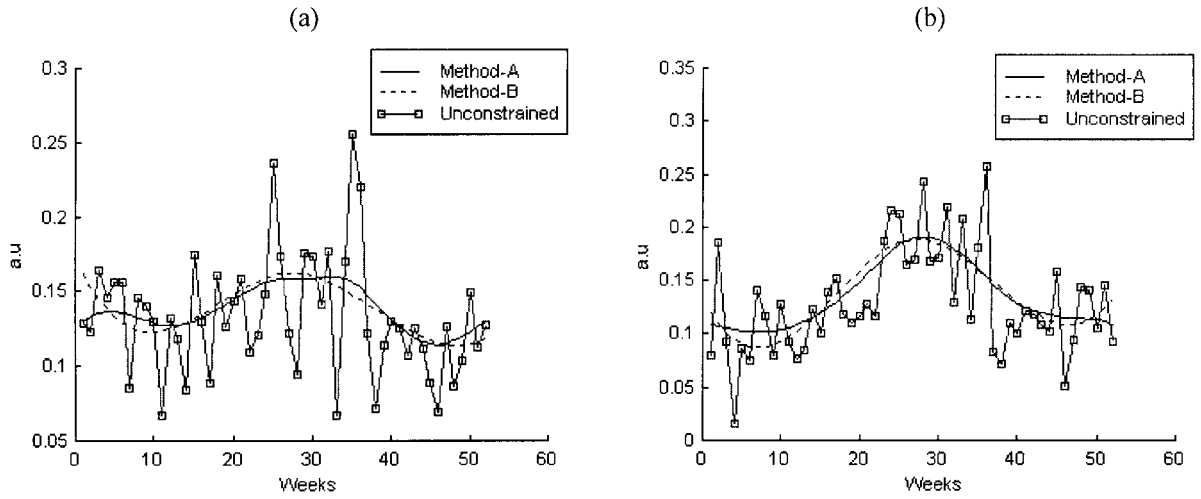

Figure 3. PARAFAC loadings for weeks of year mode for (a) 1999 and (b) 1997 (a.u., arbitrary unit).

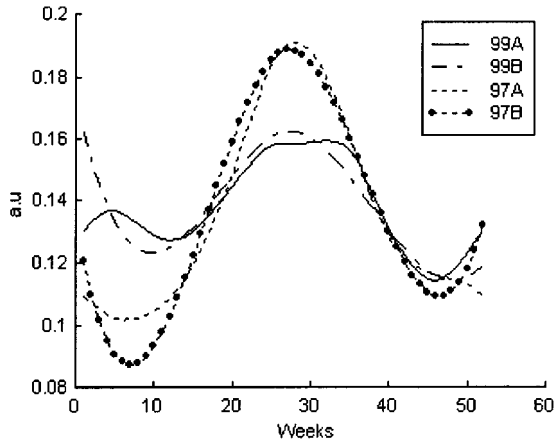

Figure 4. PARAFAC loadings for weeks of year mode for 1999 and 1997 (a.u., arbitrary unit).

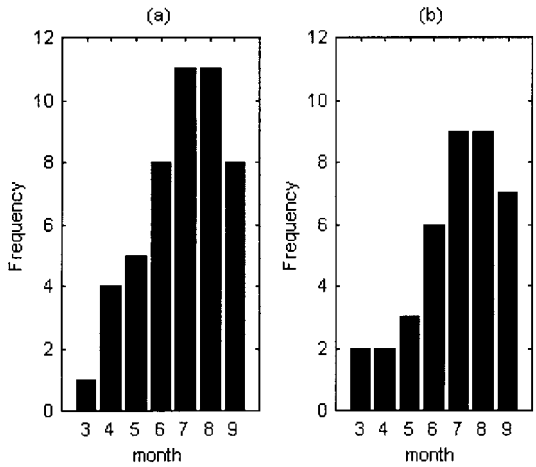

Figure 5. Frequencies of temperature inversion for (a) 1997 and (b) 1999.

dispersion [25]. Figure 5 shows the frequencies of temperature inversion [23] during the dry season for these two years, where it is possible to verify that in 1997 there are higher frequencies than in 1999. Since the temperature inversion limits the vertical ventilating capacity of the atmosphere [25], the higher frequency of atmospheric temperature inversion for 1997 is assumed to have contributed to the relatively higher carbon monoxide concentration compared to 1999. Table III shows the number of days per month with conditions not propitious for CO dispersion. Figure 5 and Table III describe the climatic conditions during the dry season (weeks 19–39, months 5–8), which suggest that 1997 presented less propitious conditions for CO dispersion compared to 1999. In terms of PARAFAC trilinear components the daily systematic variation is better captured in a week with many days not propitious for CO dispersion, resulting in a high value for the element in the WY loading. On the other hand, a week with good conditions for CO dispersion presents a small value for its element in the WY

Table III. Number of days not propitious for CO dispersion

	Number of days not propitious for CO dispersion						
	January	May	June	July	August	September	November
Year	(weeks 1–4)	(weeks 19–21)	(weeks 23–25)	(weeks 27–30)	(weeks 32–34)	(weeks 36–39)	(weeks 45–47)
1997	0	4	3	10	14	5	1
1999	1	0	1	4	10	8	0

Copyright © 2002 John Wiley & Sons, Ltd.

loading. In short, these data, which reflect the climatic conditions in the dry period (Figure 5 and Table III), show that the WY PARAFAC component does describe the seasonal effects for these two different years.

In summary, the two methods used for regularization in the fitting of PARAFAC for the WY mode, one controlling the amount of smoothing by a penalty parameter (method A) and the other by the number of basis functions (method B), have resulted in profiles which are in agreement with the climatic conditions for the period under consideration. It is worth mentioning that method B provides a more parsimonious model, since the component for the WY mode uses a small number of B-spline basis functions without being penalized by the integrated squared second derivative, making it computationally more efficient.

5. CONCLUSIONS

The focus of the present work is the description of PARAFAC components in terms of functional objects (i.e. splines). Splines are useful mathematical tools, though their use is not a simple task owing to the number of parameters to be optimized. For the experimental data used in this work, the unpredictable variation (or rapid local variation) results in an additional difficulty, which is to produce a curve with physical meaning, representing some natural behaviour of the data set, without being affected by the rapid local variation. However, one of the most important points for applying splines on such experimental data is how to control the amount of rapid local variation. Additionally, the description of the PARAFAC components by splines must preserve the convergence of the alternating least square algorithm. Thus the complexity of the approach described in this work is mainly determined by the spline fitting. In this context, two methodologies are discussed here in terms of the kind of basis functions as well as ways of controlling the amount of rapid local variation, which were tested on real data. As a result, Spline-PARAFAC produced components which are in agreement with the natural characteristics of the data set, showing a large influence of rapid local variation. In conclusion, the results described here suggest that splines can be useful in PARAFAC fitting of data sets that require functional characteristics as well smoothing variation within the components.

Acknowledgements

The authors acknowledge the financial support from FAPESP (grants 97/13046-4, and 99/09643-2) for carrying out this work, and CETESB for kindly supplying the data set. We also thank one of the anonymous referees for drawing our attention to the independent work by Timmerman and Kiers which seems to use the same approach as ours.

REFERENCES

1. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J. Chemometrics* 2000; **14**: 105–122.
2. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279–311.
3. Bro R. Multi-way analysis in the food industry, models, algorithms and applications. *PhD Thesis*, University of Amsterdam, 1998.
4. Timmerman ME and Kiers HAL. Three-way component analysis with smoothness constraints. *Comput. Statist. Data Anal.* in press.
5. Wegman EJ and Wright IW. Splines in statistics. *J. Am. Statist. Assoc.* 1983; **78**: 351–365.
6. Ramsay JO and Silverman BW. *Functional Data Analysis*. Springer: New York, 1997.
7. Besse P and Ramsay JO. Principal component analysis of sampled functions. *Psychometrika* 1986; **51**: 285–311.
8. Silverman BW. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* 1996; **24**: 1–24.
9. Wold S. Spline functions in data analysis. *Technometrics* 1974; **16**: 1–11.
10. Alsberg BK and Kvalheim OM. Compression of ***n***th-order data arrays by B-splines. Part 1: Theory. *J. Chemometrics* 1993; **7**: 61–73.
11. Ferreira MMC, Ferreira Jr WC and Kowalski BR. Rank determination and analysis of non-linear processes by global linearizing transformation. *J. Chemometrics* 1995; **10**: 11–30.
12. Comrie AC and Diem JE. Climatology and forecast modeling of ambient carbon monoxide in Phoenix, Arizona. *Atmos. Environ.* 1999; **33**: 5023–5036.
13. Paatero P and Juntto S. Determination of underlying components of cyclical time series by means of two-way and three-way factor analytic techniques. *J. Chemometrics* 2000; **14**: 241–259.
14. Harshman RA and Lundy ME. PARAFAC: parallel factor analysis. *Comput. Statist. Data Anal.* 1994; **18**: 39–72.
15. Bro R. PARAFAC: tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
16. Bro R and Sidiropoulos ND. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics* 1998; **12**: 223–247.
17. Silverman BW. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. R. Statist. Soc. B* 1985; **47**: 1–52.
18. Hastie TJ and Tibshirani RJ. *Generalized Additive Models*. Chapman and Hall: London, 1997.
19. De Boor C. *Practical Guide to Splines*. Springer: New York, 1987.
20. Wahba G and Wold S. A completely automatic French curve: fitting spline functions by cross validation. *Commun. Statist.* 1975; **4**: 1–17.
21. Green PJ and Silverman BW. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall: London, 1994.
22. [Online]. Available: http://www.psych.mcgill.ca/faculty/ramsay/software.html [30 July 2001].
23. CETESB. *Report of Air Quality in São Paulo for 1999*.
24. [Online]. Available: http://200.19.93.5/internew/informativo/balanco/relad.html [31 July 2000].
25. Colucci JM and Begeman CR. Carbon monoxide in Detroit, New York and Los Angeles air. *Environ. Sci. Technol.* 1969; **4**: 3–39.

Copyright © 2002 John Wiley & Sons, Ltd.