JOURNAL OF CHEMOMETRICS
*J. Chemometrics* 2003; **17**: 274–286
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.801

# A new efficient method for determining the number of components in PARAFAC models

**Rasmus Bro[1]* and Henk A. L. Kiers[2]**

[1]Chemometrics Group, Department of Food and Dairy Science, Royal Veterinary and Agricultural University, DK-1958 Frederiksberg C, Denmark
[2]Heymans Institute (DPMG), University of Groningen, Groningen, The Netherlands

Received 1 December 1998; Revised 2 December 2002; Accepted 28 February 2003

**A new diagnostic called the *core consistency diagnostic* (CORCONDIA) is suggested for determining the proper number of components for multiway models. It applies especially to the parallel factor analysis (PARAFAC) model, but also to other models that can be considered as restricted Tucker3 models. It is based on scrutinizing the 'appropriateness' of the structural model based on the data and the estimated parameters of gradually augmented models. A PARAFAC model (employing dimension-wise combinations of components for all modes) is called appropriate if adding other combinations of the same components does not improve the fit considerably. It is proposed to choose the largest model that is still sufficiently appropriate. Using examples from a range of different types of data, it is shown that the core consistency diagnostic is an effective tool for determining the appropriate number of components in e.g. PARAFAC models. However, it is also shown, using simulated data, that the theoretical understanding of CORCONDIA is not yet complete. Copyright © 2003 John Wiley & Sons, Ltd.**

**KEYWORDS:** PARAFAC; validation; number of components; loss function; scree plots; cross-validation

## 1. INTRODUCTION

The parallel factor analysis (PARAFAC) model [1–4] is gaining widespread use e.g. in the field of chemometrics, especially owing to its highly attractive uniqueness property. In certain applications of curve resolution [5,6] and second-order calibration [7,8], this uniqueness of the PARAFAC model is essential for solving the problems. A major practical obstacle in the use of the PARAFAC model is how to determine the appropriate number of components. Fitting a single model can be time-consuming, especially because refitting from different starting points is usually essential for assuring convergence to the global minimum. The use of resampling techniques such as cross-validation is often either unfeasible or unattractive because of the heavy computations involved and owing to the non-sequential fitting of the PARAFAC model. Therefore a diagnostic based on *single* analyses for all relevant dimensionalities seems to be called for. Simply comparing fit values (scree-like plots) and searching for a point where the increase becomes too small is often not easy, because the increases typically tend to decrease gradually, without a clear jump. In the present paper a diagnostic is suggested that is also based on single analyses but, in contrast to fit values, usually gives clear differences for different models.

First the relation between the PARAFAC model and the Tucker3 model will be described. By expressing the PARAFAC model as a special case of the Tucker3 model, it will be shown how to use a certain Tucker3 core for assessing the appropriateness of a PARAFAC model using the *core consistency diagnostic* (CORCONDIA). This diagnostic was originally presented in a less developed form by Bro [9] and implemented in the *N*-way Toolbox for Matlab [10]. It has subsequently been used in several research papers [11–13]. However, as of yet, there has been no formal scientific description of the method and its merits. This paper formally introduces the diagnostic and evaluates how it performs in different settings.

Only three-way models will be considered here, but the method extends straightforwardly to higher-order models as well. Recommendations will be given for how to use and interpret CORCONDIA and also for handling some special cases. Finally it will be shown, using simulated data and a large body of real data sets, that CORCONDIA can indeed be an effective additional tool for judging model complexity.

*Correspondence to: R. Bro, Chemometrics Group, Department of Food and Dairy Science, Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
E-mail: rb@kvl.dk
Contract/grant sponsor: LMC (Centre for Advanced Food Studies).
Contract/grant sponsor: NWayQual; Contract/grant number: GRD1-1999-10377.
Contract/grant sponsor: STVF.

## 2. THEORY

### 2.1. Background

Consider a three-way $F$-component PARAFAC model. The structural model can be stated as

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} + \mathbf{E} \tag{1}$$

The matrix $\mathbf{X}$ is an $I \times J \times K$ data array $\underline{\mathbf{X}}$ matricized (rearranged) to an $I \times JK$ matrix, the first $J$ columns being the first horizontal slab of $\underline{\mathbf{X}}$, etc. The matrices $\mathbf{A}$ ($I \times F$), $\mathbf{B}$ ($J \times F$) and $\mathbf{C}$ ($K \times F$) hold the parameters of the model, and $\mathbf{E}$ ($I \times JK$) holds the residual variation not explained by the model. The operator $\odot$ is the so-called Khatri-Rao product [14], which is equivalent to a column-wise Kronecker product [15]. The PARAFAC model may equivalently be stated in the form of a restricted Tucker3 model [8,9,16,17] as

$$\mathbf{X} = \mathbf{A}\mathbf{T}^{(F \times FF)}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}} \tag{2}$$

where the core array $\underline{\mathbf{T}}$) is a binary array with zeros in all places except for the superdiagonal, which contains only ones; that is, $t_{def} = 1$ for $d = e = f$, else $t_{def} = 0$. The two-way matrix $\mathbf{T}^{(F \times FF)}$ is the $F \times FF$ matricized three-way core. It is assumed that the components are scaled such that every loading vector within a component has the same sum of squares. Thus it holds for the columns that $\|\mathbf{a}_f\| = \|\mathbf{b}_f\| = \|\mathbf{c}_f\|$ for every $f = 1,\ldots,F$. In practice, this reparametrization of the model may be built into the algorithm performing CORCONDIA and thus does not have to be considered by the practitioner.

Assume that a PARAFAC model has been fitted. The model is given by the parameter matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ and the core array $\underline{\mathbf{T}}$. An *appropriate* PARAFAC model is a model where the components primarily reflect (low-rank) trilinear variation in the data. We will define what this means more explicitly below. Note that even though the model is appropriate it may not be *the* preferred model. Other alternative structural models or PARAFAC models using other preprocessing may provide better representations of the data. In this paper the aim is to develop a tool for finding the most appropriate complexity of a PARAFAC model, not paying attention to whether the PARAFAC model as such is the best structural model. Additionally, several PARAFAC models with different numbers of components may be appropriate and, in fact, this is mostly the case. The most reasonable model, though, is the one that comes last in the sequence of models (starting from the one-component model) and that is still appropriate. This follows naturally, because if two models with different numbers of components are both appropriate, the one with fewer components fails to describe some of the low-rank trilinear variation. Having a tool for determining appropriateness is therefore the only thing needed for determining the necessary number of components, as the appropriate model with the highest number of components must be the sought model.

A PARAFAC model is called appropriate if the component matrices defined by the model only span variation in the data that is component-wise trilinear. In other words, a PARAFAC model is considered appropriate if adding interactions between the components from the different modes does not improve the PARAFAC model. If adding interactions *will* improve the fit relatively much, this may be caused either by the fact that the components used are unsystematic (because unsystematic components will contribute more or less equally in *all* combinations in which they are used, and not only in the combination implied in the PARAFAC model) or that the data should actually be modelled by a Tucker3 model rather than PARAFAC. In both cases the PARAFAC solution is indeed 'inappropriate'. The term 'degree of appropriateness' will be quantitatively defined after first introducing an essential part of the assessment of appropriateness, namely the calculation and properties of a Tucker3 core for given PARAFAC component matrices.

Let the loading matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ contain the PARAFAC components for the three different modes. To see if these components *alone* describe the data almost as well as does the model involving interactions of these components (as modelled in Tucker3), we fit the full Tucker3 model to the data, but using the components found by PARAFAC, by minimizing

$$\sigma(\mathbf{G}) = \left\|\mathbf{X} - \mathbf{A}\mathbf{G}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \tag{3}$$

over $\mathbf{G}$ (which is the matricized version of the $F \times F \times F$ core array $\underline{\mathbf{G}}$). This idea also underlies the so-called PFCORE procedure [18], but PFCORE is used mainly to *understand* why a PARAFAC model does not fit as well as expected. Here the core array $\underline{\mathbf{G}}$ that minimizes $\sigma(\mathbf{G})$ is used to study quantitatively *to what extent* interactions play a role in describing the data. The optimal $\mathbf{G}$ minimizing $\sigma(\mathbf{G})$ can be obtained upon writing Equation (3) as

$$\sigma(\mathbf{G}) = \|\mathrm{vec}\mathbf{X} - (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\mathrm{vec}\mathbf{G}\|_{\mathrm{F}}^2 \tag{4}$$

which comes down to a straightforward regression problem. Thus the optimal $\underline{\mathbf{G}}$ can be determined as

$$\mathrm{vec}\mathbf{G} = (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})^{+}\mathrm{vec}\mathbf{X} \tag{5}$$

where the superscript '+' indicates a pseudoinverse. To understand CORCONDIA, the following simple lemma is of importance.

### Lemma

For a perfect-fitting PARAFAC model the Tucker3 core $\underline{\mathbf{G}}$ based on the PARAFAC loading matrices will be identical to the superdiagonal array of ones, $\underline{\mathbf{I}}$, if the loading matrices have full column rank.

*Proof*

Let the PARAFAC model

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} = \mathbf{A}\mathbf{T}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}} \tag{6}$$

be a perfect-fitting PARAFAC model of $\underline{\mathbf{X}}$. Let $\underline{\mathbf{G}}$ be the least squares Tucker3 core given the loading matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$; that is, the minimum of Equation (4). As the intrinsic core $\underline{\mathbf{T}}$ provides a perfect-fitting model, we have

$$\sigma(\mathbf{T}) = \|\mathrm{vec}\mathbf{X} - (\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A})\mathrm{vec}\mathbf{T}\|_{\mathrm{F}}^2 = 0 \tag{7}$$

which clearly gives the absolute minimum of Equation (4). As the problem of minimizing Equation (4) has a unique solution (provided that $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ have full column rank, as

is assumed here but which will be relaxed later), the Tucker3 core $\underline{\mathbf{G}}$ will be identical to the superdiagonal binary array $\underline{\mathbf{T}}$ if the PARAFAC model fits perfectly. QED

The situation where the loading matrices do not have full column rank is important e.g. in second-order calibration. This situation, however, is covered by means of restating the model as a kind of restricted Tucker3 model as will be shown in Appendix I.

## 2.2 The core consistency diagnostic

The Tucker3 core array can be considered as the regression of the three-way array onto the subspaces defined by the component matrices or, alternatively, the co-ordinates of the part of the three-way array within these subspaces. *The Tucker3 core is thus per definition the optimal representation of the data array with respect to the subspaces defined by the components*, whereas the PARAFAC model specifically disregards any variation associated with off-superdiagonal core elements. If the data, for example, consist of purely random variation, the regression onto *any* subspace will provide a core array with regression coefficients/elements all of similar magnitude. For very collinear component matrices the regression coefficients will be correlated but nevertheless span variation of similar size, because all directions contain variation of similar size. Regression coefficients of an array holding systematic low-rank trilinear variation, on the other hand, will be large only for the superdiagonal core elements of the Tucker3 core array if the loading matrices reflect this trilinear variation. Stating that a specific PARAFAC model is valid implies that the dimension-wise trilinear combinations of components are the only entities needed for describing the data and that interactions between the components do not contribute considerably. Then an unconstrained Tucker3 core array must be mainly superdiagonal.

The principle in CORCONDIA is to assess the degree of superdiagonality or, in fact, the similarity between the implicitly imposed superdiagonal array $\underline{\mathbf{T}}$ of ones and the least squares-fitted $\underline{\mathbf{G}}$ for a series of models with a gradually increasing number of components, starting from the one-component model. The last model in this series for which the Tucker3 core array is still similar to $\underline{\mathbf{T}}$ will be the adequate number of components to use provided that the data can be modelled by PARAFAC at all.

A simple way to assess if $\underline{\mathbf{G}}$ and $\underline{\mathbf{T}}$ are similar is to monitor the distribution of superdiagonal and off-superdiagonal elements of $\underline{\mathbf{G}}$. If the superdiagonal elements are all close to the corresponding elements of $\underline{\mathbf{T}}$ and the off-superdiagonal elements are close to zero, the model is appropriate. If, on the other hand, this is not the case, then either too many components have been extracted, the model is mis-specified or gross outliers and/or noise disturb the model. One may quantify the similarity between $\underline{\mathbf{G}}$ and $\underline{\mathbf{T}}$ as

$$\text{core consistency} = 100\left(1 - \frac{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}\left(g_{def} - t_{def}\right)^2}{F}\right) \quad (8)$$

i.e. the percentage of the alikeness expressed by one minus the sum of squares of the elements of $\underline{\mathbf{G}} - \underline{\mathbf{T}}$ divided by the sum of squares of the elements of $\underline{\mathbf{T}}$, which equals $F$. This is called the *core consistency* as it reflects how well the Tucker3 core fits to the assumptions of the model.

It is important to realize that the principle behind CORCONDIA is quite different from using residual-based methods such as e.g. in the scree test. Residual methods are almost entirely based on assessing the magnitude of the residuals as compared with the magnitude of the variation described by the (components of the) model. Thus no real assessment of the appropriateness of the structural model is investigated. CORCONDIA, on the other hand, does not focus on the magnitude of the noise as such. CORCONDIA utilizes the fact that a valid PARAFAC model only spans a very limited part of the available space. An $I \times J \times K$ array can be described as a point in an $IJK$-dimensional space, and an $F$-component PARAFAC model is situated within an $F^3$-dimensional subspace defined by the space spanned by $\mathbf{C} \otimes \mathbf{B} \otimes \mathbf{A}$. However, owing to the definition of the core array being superdiagonal in PARAFAC, the model is actually constrained or assumed to be within an $F$-dimensional subspace only (corresponding to the superdiagonal part of the core). Hence the PARAFAC model is a very constrained model even within the subspaces of which it is defined. It is this restriction that CORCONDIA specifically investigates the appropriateness of, by using the fact that non-PARAFAC structure will distribute more or less evenly across the entire $F^3$-dimensional subspace.

The core consistency is always less than or equal to 100% and may also be negative. A core consistency close to 100% implies an appropriate model. The closeness to 100% is to be understood relative to the changes compared with models with fewer components. As a rule of thumb, a core consistency above 90% can be interpreted as 'very trilinear', whereas a core consistency in the neighbourhood of 50% would mean a problematic model with signs of both trilinear variation and variation which is not trilinear. A core consistency close to zero or even negative implies an invalid model, because the space covered by the component matrices is then not primarily describing trilinear variation. A negative core consistency implies a very inappropriate model, as more variation is associated with the off-superdiagonal core elements than with the superdiagonal core elements. Many examples on the interpretation of CORCONDIA will be given in the application part of the paper.

For a one-component model the core consistency is always 100%, because there are no off-superdiagonal elements in a $1 \times 1 \times 1$ Tucker3 core array. Thus from this point of view a one-component PARAFAC model is always appropriate.

If a data set is modelled by PARAFAC models with an increasing number of components, the core consistency will typically decrease more or less monotonically and slowly with the number of components, because the influence of noise and other non-trilinear variation increases with increasing number of components. Note, though, that it is not intrinsic that the core consistency decreases monotonically, because there is no direct mathematical link between estimating the component parameters and calculating the Tucker3 core. Once the maximal number of appropriate

 

components is exceeded, the core consistency will decrease more dramatically, because some directions in the model subspace will mainly be descriptive of noise or other variation, leading to high off-superdiagonal core values.

The number of elements in the data must exceed the number of elements in the Tucker3 core significantly, otherwise the elements in the Tucker3 core can become unstable owing to underdeterminacy. Thus, for *very* small arrays, CORCONDIA may not provide useful results.

In Appendix I it is shown how CORCONDIA may be used for assessing models other than PARAFAC as well as PARAFAC models with rank-deficient loading matrices, appearing e.g. in second-order calibration.

The core consistency test is quite powerful for determining the number of components to use for real data. This will be exemplified by analysing different PARAFAC models of several different types of data from the literature. After having shown results from the use of CORCONDIA on real data, several studies on simulated data will be used for illustrating some details of the properties that enable the use of CORCONDIA. It is also emphasized that CORCONDIA is an *ad hoc* tool and its results should always be assessed in combination with other diagnostics.

## 3. RESULTS ON REAL DATA SETS

Several real data sets are analysed in this section. The results of prior analyses given in the existing literature are briefly described for each data set. It is interesting to note that for many of the data sets the conclusion from prior analyses required extensive experience with three-way analysis as well as deep insight into the problem and the use of time-consuming resampling techniques. After presenting the results of the original analyses, the resulting core consistencies and fit values fitting PARAFAC with one to six components are shown. The algorithm for fitting the PARAFAC model is based on alternating least squares with a line search [9,10]. For every number of components, five randomly started runs of the algorithm are used and the best-fitting is taken as the solution.

The following results are shown for each number of components.

1. LOSS: the loss function value (the sum of squared errors). In all tables the loss function value will be given relative to the loss function value of the one-component model.
2. RELFIT: the percentage of explained variation given as

$$\text{RELFIT} = 100\left(1 - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk} - m_{ijk}\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}{}^2}\right)$$

3. CORCO: the core consistency—Equation (8).

A fair number of alternative methods could be envisioned for shedding light on the appropriate number of components. However, the loss and fit values are the ones most commonly used and are also based on single-fit analysis. It therefore seems adequate to compare to these.

The real data vary from simple laboratory data with almost perfect low-rank trilinear structure (amino acids) over more complicated though probably quite trilinear data (sugar) to noisy data with no *a priori* knowledge of reasonable structure (bread).

### 3.1. Cars and stars data

Harshman and De Sarbo [19] have given a very thorough analysis of the so-called cars and stars data. These data consist of ratings of different brands of cars and different potential commercial spokesmen (stars) for these cars. By rating both cars and spokesmen on the same scale, it is hoped to find a structure in which it is possible to select which spokesmen would reinforce appropriately a given car. The data are arranged in a $39 \times 25 \times 34$ array. The first mode is the set of 39 different adjectives used for assessing the cars and stars. The second mode is the set of cars and stars and the third mode is the 34 different persons evaluating the cars and stars.

Harshman and De Sarbo argue that centring the data across the first and second modes and scaling within the first and third modes is feasible. In their analysis an unconstrained PARAFAC model of the preprocessed data brings about three valid components, but there are certain indications that more components are present. An additional analysis with orthogonality constraints on the rating mode loadings gives additional valid components, and through split-half analysis and visual interpretation it is argued that at least four components are valid and that more but slightly unstable components are also present. By reanalysing the data, we will study if CORCONDIA supports the conclusions of the paper.

From Table I it is easily seen that a three-component PARAFAC model is appropriate in this case. A three-component model has a core consistency of 94%, while a four-component model has a core consistency of only 17%. Thus it is easy to assess the correct number of components even though the data are certainly not perfectly approximated by the low-rank trilinear model. It is difficult to conclude anything similarly clear-cut from the fit-related values. In fact, it requires substantive insight and thorough analysis to reach the same conclusion by traditional analysis, as evidenced in the original publication.

For the orthogonality-constrained model, Harshman and

**Table I.** Loss function values relative to the value for a one-component model, fit values in percentages and core consistencies versus the number of components in an unconstrained PARAFAC model of the preprocessed cars and stars data

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 11.73 | 100.0 |
| 2 | 0.915 | 19.21 | 100.0 |
| 3 | 0.843 | 25.50 | 93.5 |
| 4 | 0.797 | 29.62 | 16.6 |
| 5 | 0.760 | 32.95 | −0.1 |
| 6 | 0.726 | 35.88 | 5.8 |

**Table II.** Loss function values relative to the value for a one-component model, fit values in percentages and core consistencies versus the number of components in a PARAFAC model of the cars and stars data with the scale mode constrained to have an orthogonal loading matrix

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 11.73 | 100.0 |
| 2 | 0.916 | 19.19 | 99.4 |
| 3 | 0.844 | 25.49 | 94.4 |
| 4 | 0.802 | 29.18 | 87.4 |
| 5 | 0.771 | 31.94 | 73.0 |
| 6 | 0.750 | 33.77 | 57.8 |

**Table III.** Loss function values, percentage of variation explained and core consistency for a one- to a six-component three-way PARAFAC model of the chromatographic data using non-negativity constraints on all modes

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 73.65 | 100.0 |
| 2 | 0.280 | 92.63 | 100.0 |
| 3 | 0.087 | 97.70 | 86.2 |
| 4 | 0.037 | 99.02 | 93.5 |
| 5 | 0.023 | 99.41 | 21.6 |
| 6 | 0.016 | 99.57 | 12.2 |

De Sarbo find that four components can now be fitted and that there might even be more components in the data. This somewhat subjective statement is, in fact, precisely stated in the core consistency values in Table II. Notice that, instead of a sharp change in the consistency, the values change gradually from 100% to 58%. That is, there is no clear-cut correct number of components, simply because models with an increasing number of components only lose their appropriateness gradually. Most likely, the reason is that these data are not that well approximated by a low-rank orthogonality-constrained trilinear model. In any case the core consistency confirms the subjective notions experienced through elaborate analysis in the original work [19], namely that a four-component model is good but further components are also relevant.

## 3.2. Chromatographic data

Fifteen samples of thick juice from different sugar factories were introduced into a Sephadex G25 low-pressure chromatographic system using 0.02 M $NH_4Cl/NH_3$ buffer (pH 9.00) as carrier. In this way the high-molecular-weight reaction products between reducing sugars and amino acids/phenols are separated from the low-molecular-weight free amino acids and phenols. The high-molecular-weight substances elute first, followed by the low-molecular-weight species. Aromatic components are retarded the most. The sample size was 300 μl and a flow rate of 0.4 ml $min^{-1}$ was used. Twenty-eight discrete fractions of 1.2 ml were sampled and measured spectrofluorometrically on a Perkin Elmer LS50B spectrofluorometer. The column was a 20 cm long glass cylinder with an inner radius of 10 mm packed with Sephadex G25 fine gel. The water used was doubly ion exchanged and millipore filtrated upon degassing. The excitation–emission matrices were collected using a standard 10 mm by 10 mm quartz cuvette, scanning at 1500 nm $min^{-1}$ with 10 nm slit widths in both excitation and emission monochromators (250–440 nm excitation, 10 nm intervals; 250–560 nm emission, 4 nm intervals).

The size of the four-way data set is 28 (fractions) × 20 (excitation) × 78 (emission) × 15 (samples). For each sample, 28 excitation–emission matrices are measured, one for each fraction collected. Owing to retention time shifts, the elution profiles shift from experiment to experiment, as discussed at length in the original publication [20]. However, by rearranging the data into a three-way structure where one mode is a combined mode consisting of the original sample and elution modes, a unique PARAFAC model may be fitted which is not affected by possible retention time shifts. On fitting three-way PARAFAC models to these data, the results in Table III are obtained. CORCONDIA suggests that four components are appropriate. This may also perhaps be deduced from the fit values, but with less confidence. In the original publication [20] it is shown that four components are indeed appropriate, as judged from split-half analysis [19] and from analysis using slightly different models.

## 3.3. Sugar data

Sugar was sampled continuously every eighth hour during 3 months of operation of a sugar plant in Scandinavia to give a total of 268 samples, of which three were discarded as extreme outliers. The sugar was sampled directly from the final unit operation (centrifuge) of the process. The sugar was dissolved in unbuffered millipore-filtrated water (2.25 g per 15 ml) and the solution was measured spectrofluorometrically in a 10 mm by 10 mm cuvette on a Perkin Elmer LS50B spectrofluorometer. Raw non-smoothed data were output from the fluorometer. For every sample the emission spectra from 275 to 560 nm were measured in 0.5 nm intervals (571 wavelengths) at seven excitation wavelengths (230, 240, 255, 290, 305, 325 and 340 nm). The array to be modelled is thus 265 × 571 × 7. Bro [21] argues that four components provide a good model. One of the components, however, is found to be partly governed by random and Rayleigh scatter-related variation in the low-wavelength area. By means of split-half analysis it is concluded that a non-negativity-constrained model with additional unimodality constraints on the emission mode loadings circumvents the problems and provides a valid four-component model. The same model is fitted here.

In Table IV the results are given. In the original work, elaborate and computationally expensive analyses were necessary for establishing the right dimensionality. Indeed, it is difficult to decide on the number of components from the fit values. They point to three or perhaps four components. The core consistency gives results that are in line with the original conclusions [21], namely that four components are valid though somewhat unstable.

## 3.4. Bread data

Five different breads were baked in duplicate to give a total of 10 samples. Eight different judges assessed the breads with respect to 11 different attributes in a fixed vocabulary

**Table IV.** Loss function values, fit values in percentages and core consistencies versus the number of components in a non-negativity-constrained PARAFAC model of the sugar fluorescence data. Additionally, unimodality constraints are imposed in the emission model loadings

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 94.14 | 100.0 |
| 2 | 0.249 | 98.54 | 99.9 |
| 3 | 0.080 | 99.53 | 81.9 |
| 4 | 0.023 | 99.87 | 60.4 |
| 5 | 0.012 | 99.93 | −0.5 |
| 6 | 0.009 | 99.95 | 3.5 |

profiling analysis [9]. The data are arranged in a three-way array (10 breads × 11 attributes × 8 judges). Although a PARAFAC model can be seen as a reasonable approximate model incorporating saliencies for each assessor, there is no fundamental theory stating the nature of the data, and hence a significant model error is anticipated. The data are also noisier than e.g. the spectral data treated above. The data are centred across the sample mode before fitting the model. This centring removes assessor-specific offsets on the attribute scales. Using a PARAFAC model, the approximation is that the judges use the same latent variables, only in different proportions. Bro [9] assessed that two components were sufficient for providing an adequate description of the data.

From the fit values in Table V it is almost impossible to assess the right complexity of the model. It is easily seen from the core consistency that the two-component model is suitable. For a three-component model the picture changes. This model is clearly inferior to the two-component model, though it does represent some systematic trilinear variation. Hence a two-component model seems appropriate, but a weak third component might be present. In this case, additional analysis would be required to verify if the third component is valid. However, regardless of the result, the core consistency suggests that the component is somewhat unstable.

## 3.5. Tongue data

Harshman *et al.* [22] performed an interesting analysis of data on tongue shapes with respect to a defined grid when uttering different English vowels. The data were generated from X-rays taken of five different speakers during their pronunciation of the vowels. Details of the experimental conditions may be found in the literature. The data array has the size 5 (speakers) × 10 (vowels) × 13 (position in cm on grid). The data were centred across vowels (R. A. Harshman and M. E. Lundy, personal communication) and it was found that two components were definitely present. A possible third component could not be substantiated, partly because apparently it was difficult to fit three components reliably. This in itself is an indication that too many components are extracted from the data.

**Table V.** Loss function values, fit values in percentages and core consistencies versus the number of components in an unconstrained PARAFAC model of the sensory profiling data of bread samples centred across the sample mode

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 35.25 | 100.0 |
| 2 | 0.784 | 49.22 | 100.0 |
| 3 | 0.658 | 57.40 | 56.1 |
| 4 | 0.576 | 62.17 | 1.8 |
| 5 | 0.507 | 67.17 | −0.9 |
| 6 | 0.456 | 70.47 | 0.5 |

**Table VI.** Loss function values, fit values in percentages and core consistencies versus the number of components in an unconstrained PARAFAC model of the tongue data centred across vowels

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 79.30 | 100.0 |
| 2 | 0.357 | 92.62 | 100.0 |
| 3 | 0.280 | 94.21 | −11.2 |
| 4 | 0.214 | 95.58 | 24.4 |
| 5 | 0.165 | 96.58 | 0.0 |
| 6 | 0.116 | 97.61 | 2.6 |

As is readily seen from Table VI, the core consistency verifies that two components are adequate. This may also be deduced from the loss function value, but with considerably less confidence.

## 3.6. Amino acids

This data set consists of five simple laboratory-made samples [1]. Each sample contains different amounts of tyrosine, tryptophan and phenylalanine dissolved in phosphate-buffered water. The samples were measured by fluorescence (excitation 250–300 nm, emission 250–450 nm, 1 nm intervals) on a Perkin Elmer LS50B spectrofluorometer with an excitation slit width of 2.5 nm, an emission slit width of 10 nm and a scan speed of 1500 nm $s^{-1}$. The array to be decomposed is hence $5 \times 51 \times 201$. In Figure 1, measurements of one of the samples are shown. Ideally these data should be describable with three PARAFAC components [1]. This is so because each individual amino acid gives a rank-one contribution to the data.

Non-negativity-constrained PARAFAC models were fitted with one to six components. Scrutinizing Table VII, it is seen that both fit-related measures (loss value and relative fit) point to three components being adequate. This is in accordance with the theory, as there are three flourescent analytes in the samples. However, CORCONDIA points clearly and surprisingly to four components being correct.

In Figures 2 and 3 the normalized loadings of a three- and a four-component model respectively are shown. It is readily seen that the three loadings of the three-component model are also found in the four-component model. These three loadings resemble the pure spectra of tryptophan, tyrosine and phenylalanine. The fourth component does not resemble any of the analytes and, in fact, does not seem to be reflecting chemical information. The reason for the presence of this fourth and quite distinct component must be that non-

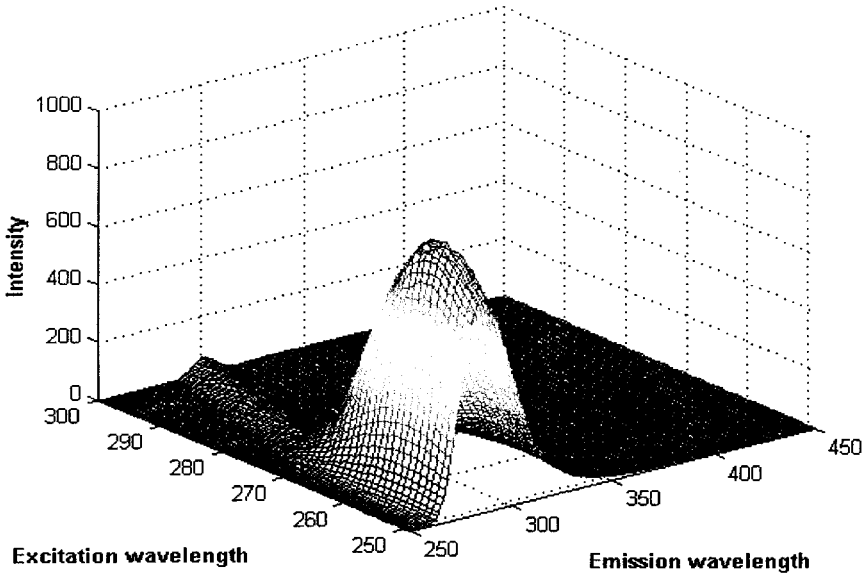


**Figure 1.** Fluorescence landscape of a sample containing only phenylalanine.

**Table VII.** Amino acid fluorescence data. Results from fitting one- to six-component non-negativity-constrained PARAFAC models to the amino acid data without missing elements

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 66.16 | 100.0 |
| 2 | 0.540 | 88.17 | 99.7 |
| 3 | 0.002 | 99.94 | 99.8 |
| 4 | 0.002 | 99.95 | 93.1 |
| 5 | 0.001 | 99.97 | 1.7 |
| 6 | 0.001 | 99.98 | 55.7 |

linearities or scatter effects cause some additional systematic variation.

In fact, these data have been investigated at several times and always using three-component PARAFAC models. Even for second-order calibration the use of three components has given satisfactory results. This is so because the fourth component has a very low variance. The variance of this fourth component is only 0.03%, as compared with 50.7%, 25.5% and 16.2% for the three 'chemical' components. Therefore the bulk variation is not affected significantly by the fourth component, and this is also the reason why traditional tools based on magnitude of residual variation have difficulties in detecting this fourth component. It is also

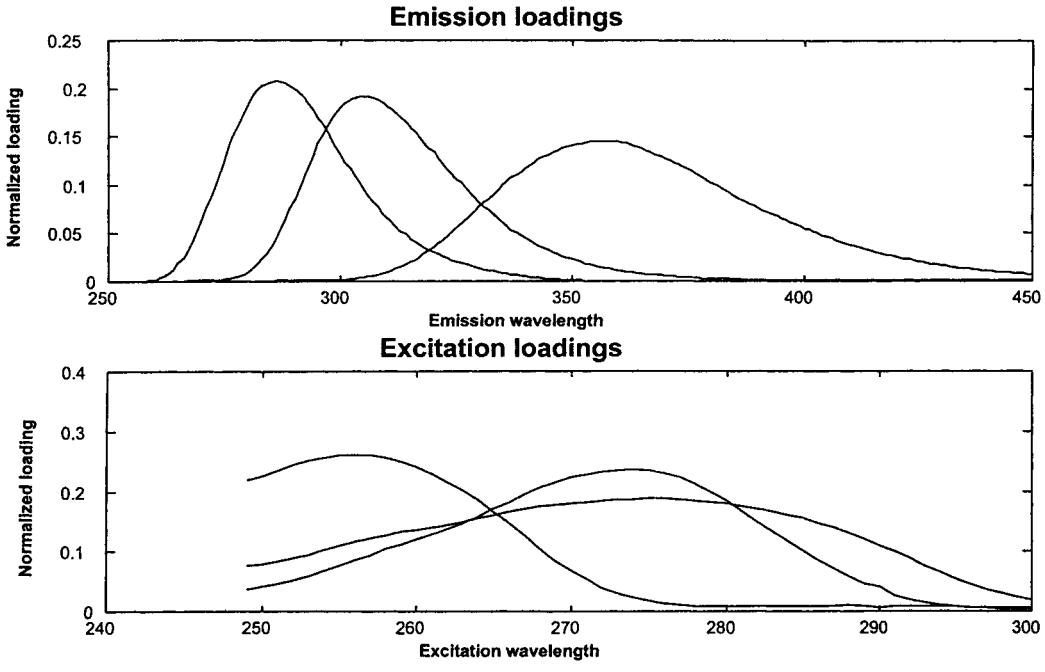


**Figure 2.** Loading vectors resulting from fitting a three-component PARAFAC model to the amino acid data.

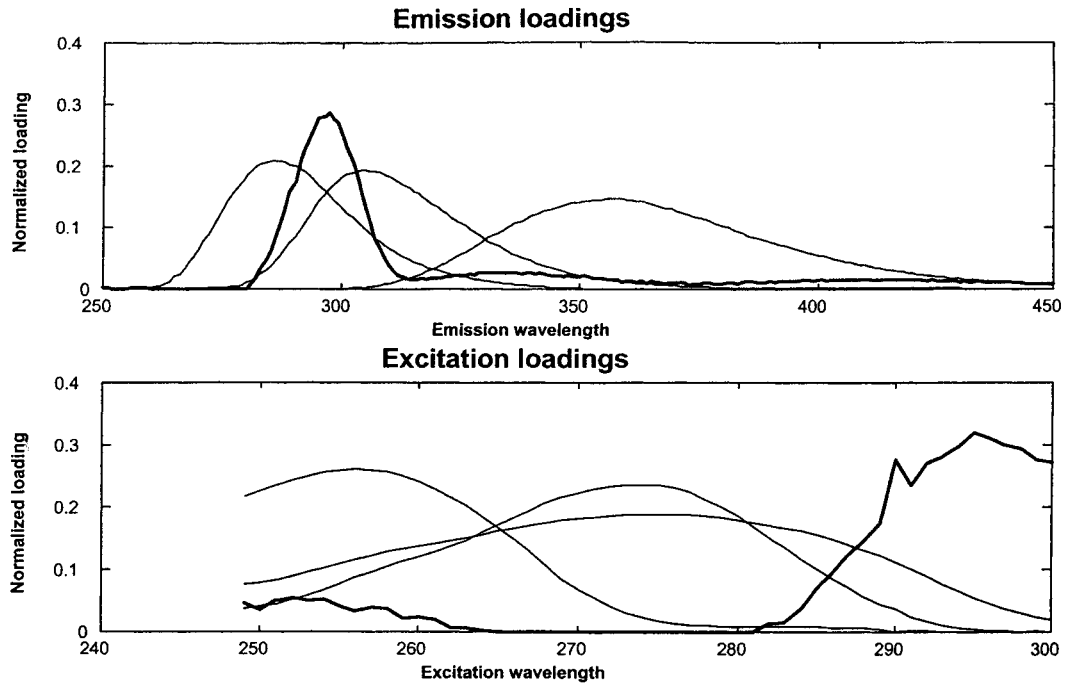


**Figure 3.** Loading vectors resulting from fitting a four-component PARAFAC model to the amino acid data. The fourth suspicious component is shown with a thicker line.

not likely that such a component can be detected e.g. using split-half analysis, because it is an idiosyncratic manifestation of non-trilinear variation in the five specific samples. Using other samples could possibly lead to a different but valid fourth component.

As an explanation for our finding, it is important to notice in Figure 1 the Rayleigh scatter in the left part, which is not multilinear in its nature [23]. It is situated around a diagonal of corresponding emission and excitation wavelengths. Additionally to Rayleigh scatter, the emission below the excitation wavelength does not vary according to the multilinear model, since the emission intensity is zero (up to the noise) regardless of excitation. In fact, the emission mode loading of the fourth 'spurious' component resembles the Rayleigh scatter. To avoid such spurious results, the lower part of the data (emission below excitation wavelength) as well as the part corresponding to Rayleigh scatter should not be used in fitting the model. Rather these elements must be set to missing values in the three-way array in order not to bias the model.

When all appropriate elements of the array have been set to missing, the values in Table VIII are obtained. Clearly, CORCONDIA now correctly identifies that there are three trilinear components in the data.

In Table VIII it is seen that a six-component solution has a quite high core consistency, but this is preceded by two very low values. Hence the choice here clearly is to take three components.

## 4. ON THE PROPERTIES OF CORCONDIA

Any diagnostic for assessing the number of components is based on certain implicit or explicit assumptions. Hence any diagnostic can also be misleading in certain situations. It is therefore of importance to have an idea about when CORCONDIA can be expected to provide reliable results and especially when it cannot. This is particularly important in the case of a diagnostic such as CORCONDIA, which is an *ad hoc* diagnostic and therefore should be used only as an *additional* indication together with other diagnostics.

It is intuitively tempting to assume that the interesting behaviour of CORCONDIA when too many components are computed is due to there being no support for more systematic variation in the data, hence that pure i.i.d. noise is being fitted by the additional components. However, it is interesting that simulations indicate that this is not what is actually causing the behaviour. This is important because it points to properties of trilinear analysis that have not yet been alluded to in the literature and may be beneficial to investigate in more detail in the future.

To illustrate the behaviour of CORCONDIA, some results on ideal trilinear data with i.i.d. noise will first be described. Data according to the PARAFAC model were made with

**Table VIII.** Amino acid fluorescence data. Results from fitting one- to six-component PARAFAC models to the amino acid data with missing values

| Components | LOSS | RELFIT | CORCO |
|---|---|---|---|
| 1 | 1.000 | 68.57 | 100.0 |
| 2 | 0.293 | 90.81 | 99.9 |
| 3 | 0.001 | 99.96 | 99.4 |
| 4 | 0.001 | 99.98 | 28.2 |
| 5 | 0.001 | 99.99 | 13.7 |
| 6 | 0.000 | 99.99 | 62.8 |

four components and an array of size $10 \times 8 \times 9$, and increasing amounts of Gaussian i.i.d. noise (10 different data sets for each of three noise cases) ranging between small (10%), medium (25%) and large (50%) amounts of noise. Similar results to those reported here were obtained also for smaller noise levels. All component matrices were chosen to be orthogonal, hence $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, $\mathbf{B}^T\mathbf{B} = \mathbf{I}$ and $\mathbf{C}^T\mathbf{C} = \mathbf{I}$. Thirty data sets were constructed and models were fitted by picking the best-fitting of five models using an alternating least squares algorithm with a convergence criterion of $10^{-8}$. The results in terms of core consistency are shown in Appendix II (Table X). From the CORCONDIA values the correct dimensionality will always be chosen based on choosing the number of components before the first abrupt decrease from close to 100%. Note, though, that some models with five components get very high core consistencies (>90%) compared with what was seen in the analyses of real data. Although the relative change leaves no doubt that five components are too many, it is of concern that invalid models get such high core consistencies. This is especially interesting because this phenomenon has not yet been observed for real data.

Several other simulation studies were carried out, including studies where the component matrices were not orthogonal, studies where the noise was not i.i.d., but was correlated and/or systematic (e.g. bilinear noise), and studies where the component matrices had higher correlations. The results from these studies were similar to the ones presented here. These results strongly indicate that none of the real data have characteristics close to the simulated data and that results obtained on data simulated in this way cannot be taken to be indicative for PARAFAC results on real data.

A first indication of the reason for this can be obtained by hypothesizing the following characteristics of real data. Assume that a data set is of approximately trilinear structure. The data consist of a trilinear part plus some random, possibly i.i.d., noise. The crucial part, however, is that, in addition to these two parts, the data also consist of minor deviations—model error—from the trilinear model. For example, if one component vector in a spectral data set corresponds to the pure absorbance spectrum of a chemical analyte, it is very reasonable to assume that the spectrum may be subject to minor differences when observed in different samples or on different occasions. Similar deviations can be anticipated for almost any kind of data, whether the data are bilinear, trilinear or have a different structure. Another characteristic that may be anticipated for real data is that the variations in the different underlying phenomena will likely not be completely independent of each other. A minor dependence between real phenomena can occur owing to the properties of the system but also on a mathematical level, because failing to describe one phenomenon perfectly affects the residual variation, which again affects the estimation of other components.

How can such phenomena be incorporated into simulation studies? One simple way of adding errors approximating the above is the following. Let the initial data follow the model

$$\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{E} \tag{9}$$

where $\mathbf{E}$ holds the actual random noise (e.g. i.i.d.). In order to simulate the above-mentioned interactions and model errors, let the total data set have the form

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{AG}(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E} \tag{10}$$

where $\mathbf{G}$ ($F \times FF$) is a core array with small random numbers. This core array adds different random combinations of the 'true' trilinear component vectors, which is a reasonable way to simulate that minor model errors appear. It is crucial to be aware that this by no means can be expected to be a perfect model of model error. On the contrary, it is well known that many types of model errors are not just simple linear combinations of the true parameters. However, the proposed model of model errors does capture a reasonable aspect of model errors as described above, namely that there are often dependences between the phenomena observed.

It is almost explicit in the term model error that it cannot be described well. The hope with the above data structure is that the added variance can be a reasonable first approximation of how model errors influence model fitting. If results of analysing data with such errors can explain the differences between results on real data and simulated trilinear data, then this hypothesized model of model error can be considered useful for getting a first indication of why the differences occur. Indeed, it will be shown that this model does provide results that are similar to the ones observed for real data; furthermore, it will be shown that the adequacy of the model of model errors can be empirically verified from the real data.

By letting $\|\mathbf{G}\|_F$ be small compared with $\|\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T\|_F$, it is ensured that the primary variation in the data follows the PARAFAC model. Adding the model error term may seem to imply that a Tucker3 rather than a PARAFAC model should be used to model the data. Indeed, if this structure of the data is a reasonable model of real data, then in theory a Tucker3 model will be more reasonable. However, first of all, it is not expected that this model of model error is perfect. It is merely a reasonable approximation which is useful to illustrate some points in PARAFAC modelling which have not yet been considered. Furthermore, for sufficiently small numbers in $\mathbf{G}$ the variation in the Tucker part of the model will be too small to be reliably modelled. In any case the interesting aspect here is not to develop a theoretically correct model. *Any* model will have deficiencies in modelling real data. The interesting question is therefore whether empirically validated PARAFAC models can be assumed to lead to the observed CORCONDIA results due to the structure defined in Equation (10). In order to verify this, a study was conducted using the amino acid data as a starting point. A three-component model was fitted to the raw data, yielding components $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. A synthetic data set was created from these estimated components as

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T + \mathbf{AG}(\mathbf{C} \otimes \mathbf{B})^T + \mathbf{E} \tag{11}$$

where $\|\mathbf{AG}(\mathbf{C} \otimes \mathbf{B})^T\|^2$ and $\|\mathbf{E}\|^2$ are describing the combined effect of model error and random noise. The Tucker3 core was assigned random Gaussian numbers. The combined magnitude of noise was set to 5% of the variation in the

model part, and two different proportions of noise were investigated. Seventy realizations were tested where 90% of the noise was due the Tucker3 part, and 70 realizations were tested where 90% of the noise was due to the i.i.d. error. A five-component model was fitted to the 140 data sets in order to assess the effect of fitting too many components.

For each fitted model the core consistency was calculated. Ideally the core consistency should be low for all models, and indeed this is the case for the data sets with substantial model error. All 70 models have core consistencies below zero. However, for data with little model error the core consistency is higher than 50% for 12 models. The consistency is even close to 100% in several cases. This is clearly not feasible but is in line with the differences observed between modelling real and simulated data.

Thus the Tucker3-based model error seems to be useful for assessing the current discrepancy between ideal and real data. In order to understand the implication of structured model error, it is useful to look at some examples of the estimated parameters in the different models of the simulated data (Figure 4).

As can be seen in Figure 4, there is a significant difference between what is obtained in situations with and without model error. When relatively large model errors are present, the excessive loadings have a tendency to model the three real components in such a way that all five fitted components are linear combinations of the three real components. As a result, these components are all of similar size (Figure 4, right). On the other hand, when there is little or no model error, the model is able to distinguish clearly between real trilinear phenomena and noise. The two excessive components only describe the noise and hence are of small magnitude. What is the importance of this observation? First of all, it is evident that CORCONDIA cannot be expected to work well in situations when there is no model error. This may seem problematic, but given the applications on widely different types of data, it is realized that this may not be a severe problem, as most real data seem to have significant model error such that the core consistency correctly guides the component selection. Furthermore, in the case of no model error it is typically easy to find the appropriate number of components, e.g. by simply inspecting the size of the loadings, whereas for the (more realistic) case of having model error it is more difficult to assess the appropriate number of components on the basis of the size of the loadings. Fortunately, in these cases, CORCONDIA seems to work well.

As stated initially, it is not expected that the model of model errors given by the core array in Equation (10) is a perfect model. However, the extent to which it provides a reasonable approximation of model errors for the real data can be evaluated. If the model of model error is perfect, then it holds that, on analysing a real data set with one component more than adequate, the obtained component matrices must be linear combinations of the component matrices obtained with the right number of components. Conversely, for perfect trilinear data (with added i.i.d. noise), this should not be the case.

For the real data above, this implied property was investigated by fitting models with the right number of components (as defined in Section 3) and with one more component. Furthermore, the same was done for simulated data sets based on the same estimated components (with the right number of components) plus 1% added Gaussian i.i.d. noise. For the real and simulated data and for each mode (**A**, **B** and **C**) the prediction of the component matrix (e.g. $\mathbf{A}_{F+1}$ signifying the first-mode component matrix with one more component than needed) from $\mathbf{A}_F$ was obtained from multiple linear regression. The quality of the prediction was given in terms of $R^2$ in percentages using normalized component matrices.

The results are shown in Table IX. The results are not

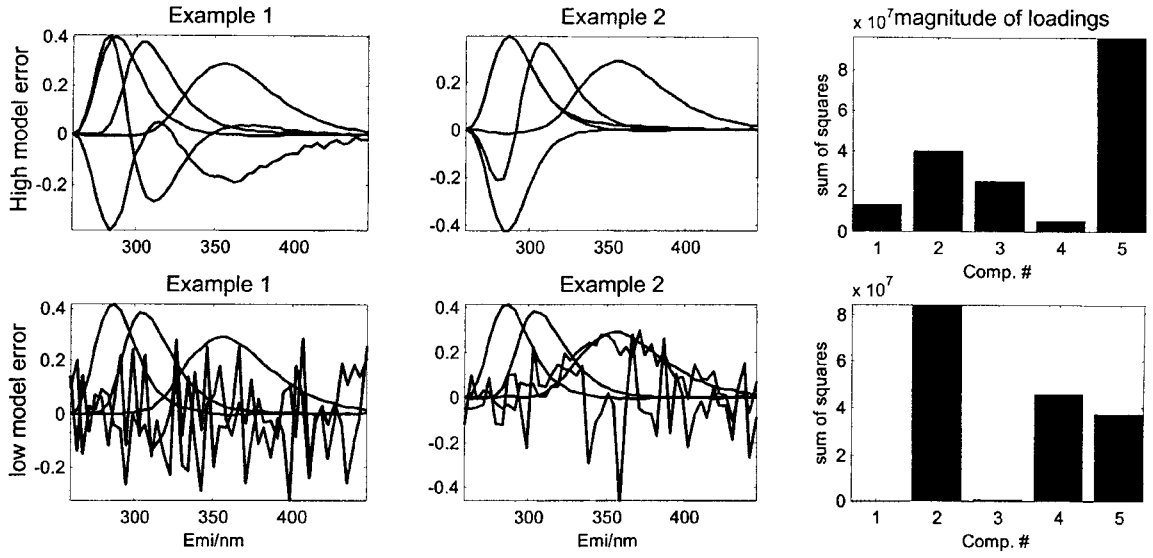


**Figure 4.** Typical examples of emission loadings from five-component models. The top row shows two examples of estimated emission loadings from data sets with high amounts of model error, while the lower plots show results for low model errors. The right-most column shows the magnitude of the individual components (the sum of squares of the elements) for the first data set of each type.

**Table IX.** $R^2$ values predicting a component matrix with one additional component from a component matrix with the correct number of components. Values are in percentages

| | Real data | | | Simulated data | | |
|---|---|---|---|---|---|---|
| Data | **A** | **B** | **C** | **A** | **B** | **C** |
| Amino acids | 97 | 91 | 91 | 70 | 28 | 49 |
| Cars and stars | 74 | 50 | 93 | 60 | 55 | 68 |
| Chromatographic | 83 | 97 | 99 | 56 | 52 | 61 |
| Sugar | 100 | 100 | 99 | 37 | 24 | 75 |
| Tongue | 94 | 78 | 96 | 66 | 60 | 87 |
| Bread | 75 | 76 | 95 | 79 | 79 | 82 |

perfectly consistent. This is expected, because we do not purport to have a perfect model of the model errors. However, looking through the table, it is evident that the general trend is that, for real data, additional components tend to be linear combinations of the 'true' components, whereas this is less pronounced for the simulated data. Hence these empirical observations support the adequacy of the approximate model of model errors and hence that the explanation for why the core consistency works so well lies in the model error. Further work is needed, though, for elaborating the models of model errors for specific data types to gain further insight into this. For example, it is very likely that the appearance of two-factor degeneracies when overfactoring the model also plays a role in the behaviour of the core consistency. The core consistency can sometimes vary dramatically over a few iterations for ill-posed models. This is not in itself a problem, because these models are always highly unstable and all have a very low core consistency. The dramatic changes typically occur because the regression problem solved when determining the least squares core is ill-posed, a situation that is expected to arise e.g. in the situation where there are two-factor degeneracies.

## 5. CONCLUSION AND DISCUSSION

The core consistency diagnostic aids in choosing the proper model complexity of PARAFAC models. The core consistency diagnostic provides stronger and often clearer answers than e.g. inspecting fit values. Further, no *a priori* assumptions regarding residuals are required and no distributionally related degrees of freedom are required. Additionally, by assessing the size of the core consistency, the diagnostic also helps in pointing out if the model is strongly stable or slightly unstable. It has been shown throughout the examples that it is possible to quantify subjective results very clearly.

The results shown here for data of quite different natures indicate that CORCONDIA has versatile applicability, and it is suggested that it is always used to *supplement* other methods for determining dimensionality. Core consistency, though, should never form the sole basis upon which the choice of the number of components is made. Other types of validation (exploratory as well as confirmatory) have to be used as well.

Sometimes the core consistency value was seen to increase again after having been quite low. This by itself has no implications for the number of components to choose, as this is to be based on the last in a succession of sufficiently high core consistency values. Nevertheless, the phenomenon is worth noting and may at first seem a bit surprising. Such a phenomenon may point to the existence of weak structural components that were simply missed by the preceding analysis. Indeed, if the components are too weak, they may simply be undetectable given the presence of noise. An alternative explanation for the phenomenon could be as follows. Every data set can be fitted perfectly by a PARAFAC model provided that a sufficient number of components are used (usually very many components). For such perfect-fitting models the core consistency value will by definition be perfect (see the lemma in Section 2.1). Hence for any data set the core consistency value starts at 100% for the one-component model and will then start decreasing. However, if one continues increasing the number of components, at a certain point the value of the core consistency will always *increase* again. It is expected that such an *increase* in the core consistency value will go gradually and may in fact start with a number of components much smaller than the actual number of components needed for a perfect PARAFAC fit. In fact, one may expect the core consistency value to be high as soon as the fit of the data approximates that of a perfect PARAFAC fit.

## 6. COMPUTATIONS

The algorithms have been implemented in MATLAB for Windows v5.3 (MathWorks, Inc.) and can be obtained from the Internet at http://www.models.kvl.dk as part of the *N*-way Toolbox for MATLAB [10]. All calculations were performed on a 266 MHz Pentium II Dell Latitude PC with 128 MB RAM.

## Acknowledgements

R. Bro gratefully acknowledges support provided by LMC (Center for Advanced Food Studies), NWayQual (GRD1-1999-10377) and STVF project 'Multi-way Analysis'. Claus A. Andersson is thanked for allowing us to use the amino acid fluorescence data. Magni Martens is thanked for allowing the use of the sensory data. Richard A. Harshman is thanked for giving us permission to use the cars and stars data as well as the tongue data. An anonymous reviewer is gratefully thanked for a very thorough and inspiring review helping us in improving the manuscript substantially.

## APPENDIX I

CORCONDIA is also useful for assessing the appropriateness of types of models other than the PARAFAC model. Bro [9,24] suggested the restricted PARATUCK2 model for modelling certain rank-deficient data. This model was later generalized to the so-called PARALIND family of models [25]. One specific type of PARALIND model is defined as

$$\mathbf{X} = \mathbf{AH}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} \tag{12}$$

which distinguishes itself from the PARAFAC model by the

introduction of an interaction matrix **H**. This interaction matrix makes it possible to have a different column dimension in **A** from that in **B** and **C** as well as providing a structural approach to modelling linear dependences in **A** (so-called closure). The PARALIND model with the first-mode model dimension $R$ and second- and third-mode model dimensions $S$ is equivalent to the restricted Tucker3 model

$$\mathbf{X} = \mathbf{A}\mathbf{T}^{(R\times SS)}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}} \tag{13}$$

where the core array $\underline{\mathbf{T}}$ (R × S × S) has a specific structure. The element

$$t_{rss} = h_{rs} \tag{14}$$

and all other elements are zero.

The principle for testing the PARALIND model with CORCONDIA is basically the same as for the PARAFAC model. For a given model the component matrices of the model are used for estimating a Tucker3 core. To be able to compare with a Tucker3 core, the corresponding restricted Tucker3 core, the elements of which are given in Equation (14), must be used, just like the superdiagonal array $\underline{\mathbf{T}}$ is used for the PARAFAC model. The unconstrained Tucker3 $\underline{\mathbf{G}}$ is then compared with the expected core $\underline{\mathbf{T}}$, again using Equation (8).

When some loading matrices in a PARAFAC model do not have full column rank, the core consistency test does not work, because the rank of the matrix of independent variables defining the core regression problem in Equation (4) is deficient. This is of practical importance when one mode has a dimension lower than the number of components. However, this problem can be easily circumvented. Assume that the first-mode component matrix **A** of a PARAFAC model has dimension 2 × 5, i.e. there are only two levels in the first mode but five components. Such a situation occurs frequently in second-order calibration. Let $\mathbf{A}_1$ contain two columns of **A** that are not collinear and let $\mathbf{A}_2$ contain the remaining columns. Define a 2 × 5 matrix **H** as

$$\mathbf{H} = [\mathbf{I} \qquad \mathbf{A}_1^{-1} \qquad \mathbf{A}_2] \tag{15}$$

where **I** is the 2 × 2 identity matrix*. It then holds that the PARAFAC model

$$\mathbf{X} = \mathbf{A}(\mathbf{C}\odot\mathbf{B})^{\mathrm{T}} = \mathbf{A}_1\mathbf{H}(\mathbf{C}\odot\mathbf{B})^{\mathrm{T}} \tag{16}$$

i.e. the PARAFAC model can be posed as a PARALIND model. As all component matrices of this PARALIND model are of full rank, the model can be tested as an ordinary PARALIND model. A practical problem is that the core consistency of the PARAFAC model will depend on how the columns of **A** are used for calculating **H**. However, the matrix $\mathbf{A}_1$ defining the first–mode model subspace will span the same space regardless, so the differences in core consistency only reflect the variations focusing on different non-trilinear variation. Hence, if the model is appropriate, it will produce a high core consistency regardless of how **H** is computed. Note also that, although fitting an unconstrained Tucker3 core to the original component matrices has an unidentified solution, this will not be the case for the transformed problem. Although there are several ways of transforming the model into the form of Equation (16), once this model has been chosen, the fitting of the Tucker3 core array to the transformed problem is an identified problem owing to the full rank of the involved component matrices.

*One may alternatively use a QR decomposition so as to reparametrize **A** by **Q** and use **R** instead of **H**. This makes it possible to automate the search for columns even in situations where some columns are exactly identical, as is sometimes the case in second-order calibration.

## APPENDIX II

**Table X.** Core consistency of models using one to five components for 30 different data sets with four-dimensional PARAFAC structure with orthogonal components

| | Components | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Low noise | 100 | 100 | 99.89 | 99.60 | 87.08 |
| | 100 | 100 | 99.89 | 99.72 | 87.26 |
| | 100 | 100 | 99.84 | 99.65 | 72.25 |
| | 100 | 100 | 99.95 | 99.73 | 85.23 |
| | 100 | 100 | 99.69 | 99.64 | 96.11 |
| | 100 | 100 | 99.84 | 99.54 | 92.48 |
| | 100 | 100 | 99.82 | 99.68 | 88.38 |
| | 100 | 100 | 99.78 | 99.70 | 85.04 |
| | 100 | 100 | 99.90 | 99.74 | <0 |
| | 100 | 100 | 99.60 | 99.31 | <0 |
| Medium noise | 100 | 100 | 99.45 | 99.17 | 94.82 |
| | 100 | 100 | 98.96 | 98.58 | <0 |
| | 100 | 100 | 99.44 | 98.29 | <0 |
| | 100 | 100 | 99.27 | 98.28 | 90.46 |
| | 100 | 100 | 99.14 | 98.25 | <0 |
| | 100 | 100 | 99.29 | 96.99 | <0 |
| | 100 | 100 | 99.26 | 98.19 | <0 |
| | 100 | 100 | 99.58 | 99.04 | <0 |
| | 100 | 100 | 99.04 | 98.44 | 57.02 |
| | 100 | 100 | 99.53 | 98.82 | 89.83 |
| High noise | 100 | 100 | 97.91 | 90.71 | <0 |
| | 100 | 100 | 98.22 | 94.78 | 74.27 |
| | 100 | 100 | 98.99 | 92.45 | <0 |
| | 100 | 100 | 98.33 | 94.41 | 81.86 |
| | 100 | 100 | 93.70 | 91.24 | 50.72 |
| | 100 | 100 | 97.89 | 94.95 | <0 |
| | 100 | 100 | 96.93 | 93.45 | 9.55 |
| | 100 | 100 | 98.39 | 92.60 | 30.91 |
| | 100 | 100 | 98.37 | 92.22 | 55.44 |
| | 100 | 100 | 98.75 | 94.88 | <0 |

## REFERENCES


1. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
2. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multimodal factor analysis. *UCLA Working Papers Phonet.* 1970; **16**: 1–84.
3. Kruskal JB. Multilinear methods. *Proc. Symp. Appl. Math.* 1983; **28**: 75–104.
4. Snyder CW, Walsh WD and Pamment PR. Three-mode PARAFAC factor analysis in applied research. *J. Appl. Psychol.* 1983; **68**: 572–583.
5. Mitchell BC and Burdick DS. An empirical comparison of

resolution methods for three-way arrays. *Chemometrics Intell. Lab. Syst.* 1993; **20**: 149–161.
6. Nørgaard L. Spectral resolution and prediction of slit widths in fluorescence spectroscopy by two and three way methods. *J. Chemometrics* 1996; **10**: 615–630.
7. Kiers HAL and Smilde AK. Some theoretical results on second-order calibration methods for data with and without rank overlap. *J. Chemometrics* 1995; **9**: 179–195.
8. Smilde AK, Tauler R, Henshaw JM, Burgess LW and Kowalski BR, Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 3. Medium-rank second-order calibration with restricted Tucker models. *Anal. Chem.* 1994; **66**: 3345–3351.
9. Bro R. Multi-way analysis in the food industry. Models, algorithms, and applications. *PhD Thesis*, University of Amsterdam, 1998 (http://www.mli.kvl.dk/staff/foodtech/brothesis.pdf).
10. Andersson CA and Bro R. The *N*-way Toolbox for MATLAB. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 1–4.
11. Estienne F, Matthijs N, Massart DL, Ricoux P and Leibovici D. Multi-way modelling of high-dimensionality electroencephalographic data. *Chemometrics Intell. Lab. Syst.* 2001; **58**: 59–72.
12. Fernandez FM, Tudino MB and Troccoli OE. Multicomponent kinetic determination of Cu, Zn, Co, Ni and Fe at trace levels by first and second order multivariate calibration. *Anal. Chim. Acta* 2001; **433**: 119–134.
13. Moberg L, Robertsson G and Karlberg B. Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis. *Talanta* 2001; **54**: 161–170.
14. Rao CR and Mitra S. *Generalized Inverse of Matrices and Its Applications*. Wiley: New York, 1971.
15. MacDonald RP. A simple comprehensive model for the analysis of covariance structures: some remarks on applications. *Br. J. Math. Statist. Psychol.* 1980; **33**: 161–183.
16. Harshman RA and Lundy ME. The PARAFAC model for three-way factor analysis and multidimensional scaling. In *Research Methods for Multimode Data Analysis*, Law HG, Snyder CW, Hattie JA, McDonald RP (eds). Praeger: New York, 1984; 122–215.
17. Smilde AK, Wang Y and Kowalski BR. Theory of medium-rank second-order calibration with restricted Tucker models. *J. Chemometrics* 1994; **8**: 21–36.
18. Lundy ME, Harshman RA and Kruskal JB. A two-stage procedure incorporating features of both trilinear and quadrilinear model. In *Multiway Data Analysis*, Coppi R, Bolasco S (eds). Elsevier: Amsterdam, 1989; 123–130.
19. Harshman RA and De Sarbo WS. An application of PARAFAC to small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In *Research Methods for Multimode Data Analysis*, Law HG, Snyder CW, Hattie JA, McDonald RP (eds). Praeger: New York, 1984; 602–642.
20. Bro R, Andersson CA and Kiers HAL. PARAFAC2—Part II. Modeling chromatographic data with retention time shifts. *J. Chemometrics* 1999; **13**: 295–309.
21. Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics Intell. Lab. Syst.* 1999; **46**: 133–147.
22. Harshman RA, Ladefoged P and Goldstein L. Factor analysis of tongue shapes. *J. Acoust. Soc. Am.* 1977; **62**: 693–707.
23. Ewing GW. *Instrumental Methods of Chemical Analysis*. McGraw-Hill: New York, 1985.
24. Smilde AK, Tauler R, Saurina J and Bro R. Calibration methods for complex second-order data. *Anal. Chim. Acta* 1999; **398**: 237–251.
25. Bro R, Harshman R and Sidiropoulos ND. New rank-deficient models for multi-way data. *J. Chemometrics*; submitted.

 