From R. Bro: Multi-way Analysis in the Food Industry. Ph.D. thesis, Amsterdam, 1998

A new approach called the *core consistency diagnostic* has been suggested for determining the proper number of components for multi-way models (Bro & Kiers, J. Chemom, 1998). It applies especially to the PARAFAC model, but also any other model, that can be considered a restricted Tucker3 model. First the principle behind the method will be given and it will then be shown that it can indeed be an effective way of judging model complexity.

Consider a three-way PARAFAC model. Normally the structural model is stated

$$\mathbf{X} = \mathbf{A}(\mathbf{C} \otimes \mathbf{B})^T, \tag{2}$$

but it may equivalently be stated as a restricted Tucker3 model

$$\mathbf{X} = \mathbf{A}\mathbf{T}^{(F \times FF)}(\mathbf{C} \otimes \mathbf{B})^T, \tag{3}$$

where the core array $\underline{\mathbf{T}}$ is a binary array with zeros in all places except for the superdiagonal which contains only ones. After having fitted the PARAFAC model (**A**, **B**, and **C**), verification that the trilinear structure is appropriate can be obtained by calculating the least-squares Tucker3 core given **A**, **B**, and **C**, according to

$$\mathbf{X} - \mathbf{A}\mathbf{G}(\mathbf{C} \otimes \mathbf{B})^T \tag{4}$$

A regression model can be constructed to solve the problem based on the loss function

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

If the PARAFAC model *is* valid then $\underline{\mathbf{G}}$ should resemble $\underline{\mathbf{T}}$. If the data can not approximately be described by a trilinear model or too many components are used then the core, $\underline{\mathbf{G}}$, will differ from $\underline{\mathbf{T}}$. To explain this, assume that an *F*-component trilinear model is an adequate model for the systematic part of the data. An additional component (or rather the *F*+1-component model as a whole) will not only describe the systematic variation but also the random part which is distributed more evenly in the variable space. Hence the extra component, even though it is forced to be trilinear, will be descriptive not only of trilinear variation, but also variation distributed all over the array. This follows because, if an extra component could be found that is not descriptive of evenly (or rather non-trilinear) distributed variation, then naturally a component could be found descriptive of trilinear variation. In that case an extra component *would* be appropriate. Hence, per definition, a least squares model with an extra component will not only describe trilinear variation. Therefore the core array of equation 4 will provide a better-fitting model by deviating from $\underline{\mathbf{T}}$ (see Bro & Kiers 98).

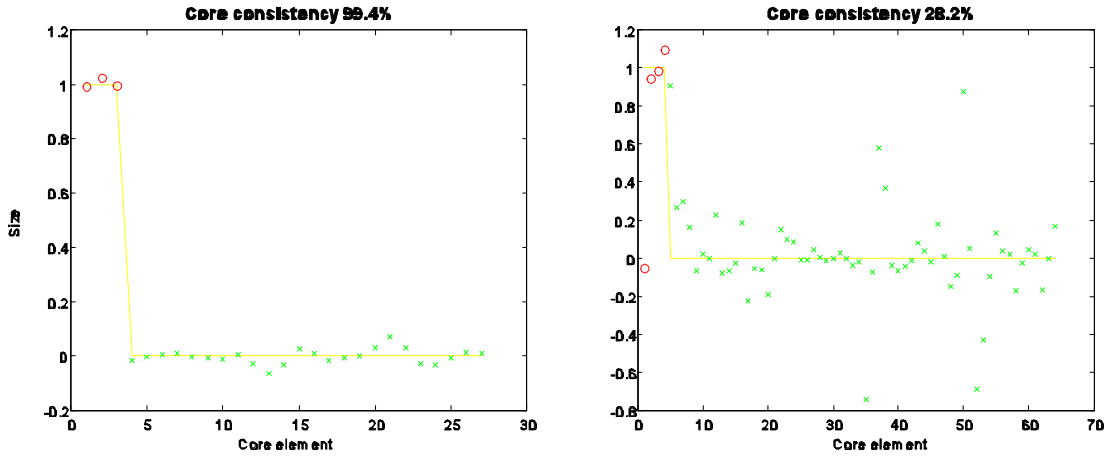

Figure 1. Core consistency plot of a three-component PARAFAC model (left) and a four-component model (right) of five samples of fluorescence excitation-emission (a 5 × 201 × 61 array). Each sample contains different amounts of tryptophan, tyrosine, and phenylalanine, and should theoretically be modeled by a three-component model. In the plots the circles are the superdiagonal elements of ***G*** *hence these should ideally be one. The crosses are the off-superdiagonal elements which should ideally be zero. The line segment is made from the elements of* ***T*** *and is hence the target that* ***G*** *should resemble.*

A simple way to assess if the model structure is reasonable is therefore to monitor the distribution of superdiagonal and off-superdiagonal elements of **G**. If the superdiagonal elements are all close to one, and the off-superdiagonal elements are close to zero the model is not overfitting. 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 disturb the model. It is possible to calculate the superidentity of **G** to obtain a single parameter for the model quality. The superidentity or *core consistency* is defined here as

$$\text{Core Consistency} = 100\left(\frac{1 - \sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}(g_{def} - t_{def})^2}{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}t_{def}^2}\right) \quad (6)$$

i.e. percentage of the variation in **G** consistent the variation in **T**. It is called the *core consistency* as it reflects how well the Tucker3 core fits to the assumptions of the model. The difference between using superdiagonality and superidentity is generally small. The superidentity, though, is directly reflecting the sought consistency.

The core consistency diagnostic may at first seem less strong than other approaches for determining dimensionality, but it is actually extremely powerful. This will be exemplified here by showing the results for some of the PARAFAC models discussed in this thesis (for more examples see Bro & Kiers 98). In the following figures the distribution of the core elements are shown in so-called *core consistency plots* for models of the same complexity as used in the actual applications and models using one more component. Note the following important points

- For *all* models the appropriate complexity was determined by other means than core consistency.
- The data vary from simple laboratory data with almost perfect trilinear structure (amino acids) over more complicated though probably quite trilinear data (sugar) to very noisy data with no *a priori* knowledge of reasonable structure (bread).

In Figure 1 the core consistency plot is shown for two competing models of a simple fluorescence data set. It is easy to see that the four-component model is strongly overfitting the data, as several off-superdiagonal elements of **G** are larger or of similar size as the superdiagonal elements. In this case there is thus no doubt that the three-component solution is preferable to the four-component solution.

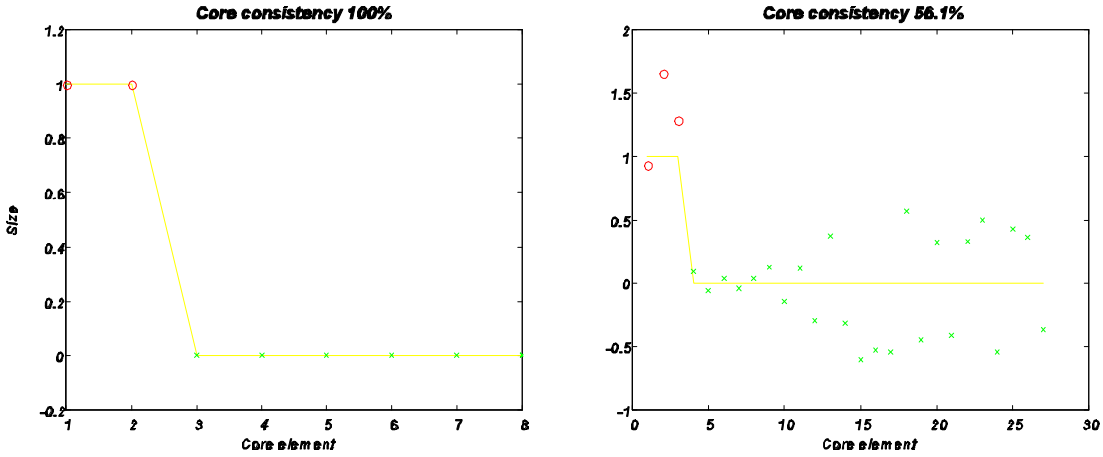

Figure 2. Core consistency plots of a two- and a three-component PARAFAC model of the bread data.

In Figure 2 the core consistency plot of a two-component model of the bread data is perfect. The core elements on the superdiagonal (to the left in the figure) are close to one and all off-diagonal elements are zero. Further the superidentity is 100%. Hence there is no doubt that the two-component model is suitable from this point of view. For a three-component model the picture changes. This model is clearly inferior to the two-component model. Note that in this case the data consist of different assessors' judgement of different breads with respect to different attributes. These data are noisy and there is no theory stating that the structure of the data should be trilinear. A two-component model seems appropriate, but the somewhat intermediate core consistency of the three-compponent solution indicates that some additional systematic variation is present there though. Further analysis is necessary in this case to elucidate the appropriateness of the three-component model.

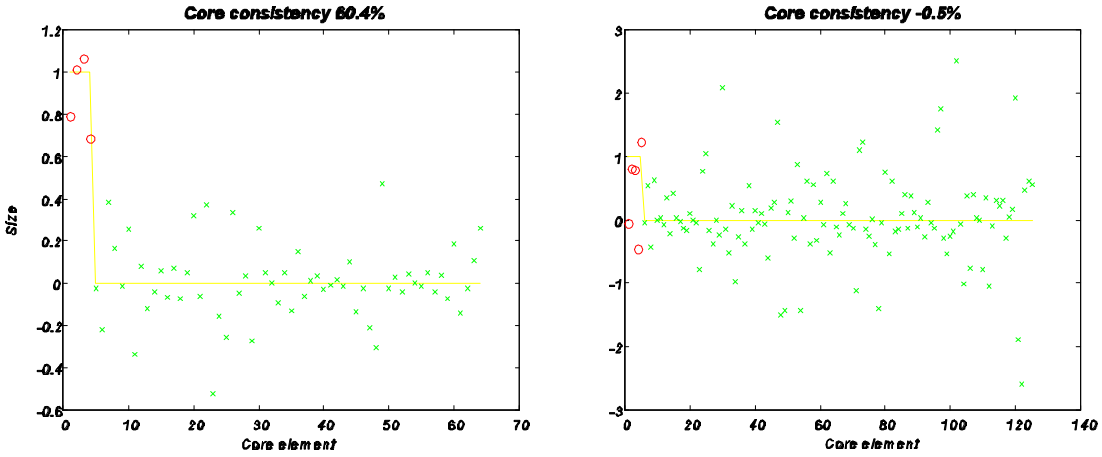

Figure 3. Core consistency plots of a four- and a five-component PARAFAC model of the sugar data.

In Figure 3 the core consistency plots are not as clear-cut as for the other examples. Even though the five-component PARAFAC solution is inappropriate as judged from the core consistency plot, the four-component model is not perfect either. However, as the diagonality is high and the data *are* difficult to model adequately, the four-component model may still be the best choice. Further, the deviation from perfect superidentity of the four-component solution simply points to the problems also described in detail in chapter seven.

If the core consistency diagnostic is to be used for other types of models, e.g., the restricted PARATUCK2 model the principle is basically the same. For a given model the loading matrices of the model are used for calculating a Tucker3 core. To be able to compare with a Tucker3 core, the corresponding restricted Tucker3 core of the model must be used, just like the superidentity array $\underline{\mathbf{T}}$ is used for the PARAFAC model. For a restricted PARATUCK2 model this array can be identified by restating the model as a restricted Tucker3 model. Suppose the restricted PARATUCK2 model has first mode model dimension R and second and third mode model dimension S. Then

$$\mathbf{X} = \mathbf{AH}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}} \Rightarrow$$

$$\mathbf{X} = \mathbf{AT}^{(\mathrm{R} \times \mathrm{SS})}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}}, \tag{7}$$

where the core array $\underline{\mathbf{T}}$ has a specific structure. The element

$$t_{\mathrm{rss}} \equiv h_{\mathrm{rs}}, \tag{8}$$

and all other elements are zero. It is easily verified that the structural model thus posed is identical to the restricted PARATUCK2 model. It is of no concern if the interaction matrix is fixed or estimated. In both cases $\underline{\mathbf{G}}$ is compared to the expected core $\underline{\mathbf{T}}$.

When some loading matrices in a PARAFAC model do not have full column rank the core consistency diagnostic does not work because the rank of the problem defining the core is deficient. However, this problem can be easily circumvented. Assume that the first mode loadings, $\mathbf{A}$, of a PARAFAC model has dimensions 2×4, i.e., there are only two levels in the first mode but four components. Such a situation occurs frequently in second-order calibration. Let $\mathbf{A}_1$ be two columns of $\mathbf{A}$ that are not collinear and let $\mathbf{A}_2$ be the remaining columns. Define

a 2 × 4 matrix **H** as

$$\mathbf{H} = [\mathbf{I}\ \mathbf{A}_1^{+}\mathbf{A}_2] \tag{9}$$

where **I** is the two by two identity matrix. It then holds that the PARAFAC model

$$\mathbf{X}^{(I\times JK)} = \mathbf{A}(\mathbf{C}|\otimes|\mathbf{B})^T = \mathbf{A}_1\mathbf{H}(\mathbf{C}|\otimes|\mathbf{B})^T, \tag{10}$$

i.e., the PARAFAC model can be posed as a restricted PARATUCK2 model. Since all loading matrices of this restricted PARATUCK2 model are of full rank the model can thus be tested as an ordinary restricted PARATUCK2 model. In practice, a QR decomposition can be used for rearranging the model.

The core consistency diagnostic may also be used for judging other models. It is not necessary that the model to test is unique, but it is essential that the model used to test the parameters against is less restricted than the model being judged. It is not feasible, for example, to test a two-way PCA model ($\mathbf{TP}^T$) against the 'core' of a bilinear model where the core is not restricted to be diagonal ($\mathbf{AGB}^T$). The posed problem here is to estimate

$$||\mathbf{X} - \mathbf{AGB}^T|| \tag{11}$$

where **A** and **B** are set equal to **T** and **P** from the PCA model. The calculated core **G** will be equal to the identity matrix. This is so because the model $\mathbf{AGB}^T$ is mathematically equivalent to the model $\mathbf{TP}^T$. Having **G** equal to the identity matrix is implicitly given even in the PCA model and as the model $\mathbf{AGB}^T$ does not offer any increased modeling power no other setting of **G** can give a better-fitting model.

The core consistency diagnostic often gives a clear-cut answer to whether a model is appropriate or not. It does not, however, tell if the model is *the* correct model. For a data set that can be modeled by, say, a three-component PARAFAC model, one will find that a one- and two-component PARAFAC model is also valid. The core consistency will consequently show that all these models are valid in the sense that they do not overfit. By assuming that noise is not trilinear however, it follows that the valid model with the highest number of components must be the one to choose. Also, though, it must be considered that another model structure or a model using other constraints or preprocessing may be more appropriate.

If a data set is modeled by, e.g., PARAFAC models of increasing number of components, the superidentity will typically decrease monotonically with the number of components. After the maximal number of appropriate components the superidentity will decrease, though, much more dramatically, and often more clearly than if using a scree-plot or similar.

The core consistency diagnostic helps in choosing the proper model complexity. Further, no a priori assumptions regarding residuals are required, since it is the deterministic and systematic rather than the probabilistic part of the data that is being used for assessing the model. The results shown here for data of quite different nature indicate that it has a versatile applicability and it is suggested that it is used to supplement other methods for determining dimensionality.