Chromatographic profiling: From samples to information

Peters, S.

Citation for published version (APA):
Peters, S. (2013). Chromatographic profiling: From samples to information

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 5

A new method for the automated selection of the number of components for deconvolving overlapping chromatographic peaks

This chapter has been accepted for publication in Analytica Chimica Acta.
Summary

Mathematical deconvolution methods can separate co-eluting peaks in samples for which (chromatographic) separation fail. However, these methods often heavily rely on manual user-input and interpretation. This is not only time-consuming but also error-prone and automation is needed if such methods are to be applied in a routine manner.

One major hurdle when automating deconvolution methods is the selection of the correct number of components used for building the model. We propose a new method for the automatic determination of the optimum number of components when applying multivariate curve resolution (MCR) to comprehensive two-dimensional gas chromatography-mass spectrometry (GC×GC-MS) data. It is based on a two-fold cross-validation scheme. The obtained overall cross-validation error decreases when adding components and increases again once over-fitting of the data starts to occur. The turning point indicates that the optimum number of components has been reached. Overall, the method is at least as good as and sometimes superior to the inspection of the eigenvalues when performing singular-value decomposition. However, its strong point is that it can be fully automated and it is thus more efficient and less prone to subjective interpretation. The developed method has been applied to two different-sized regions in a GC×GC-MS chromatogram. In both regions, the cross-validation scheme resulted in selecting the correct number of components for applying MCR. The pure concentration and mass spectral profiles obtained can then be used for identification and/or quantification of the compounds. While the method has been developed for applying MCR to GC×GC-MS data, a transfer to other deconvolution methods and other analytical systems should only require minor modifications.
5.1 Introduction

In the field of separation sciences, the development of two-dimensional chromatographic techniques such as comprehensive two-dimensional gas chromatography (GC×GC) or liquid chromatography (LC×LC) is the most important improvement made in recent years. In these methods, the peak capacity is increased by around one order of magnitude in comparison to one-dimensional systems. This increased separation power allows the analysis of mixtures to a detail not possible in the past. For complex samples, however, in spite of the high peak capacities, there is still a significant chance that not all compounds will be separated from each other, as has been demonstrated by Davis [1]. In such situations, the use of mass spectrometry as a detection system may help isolating the profiles of co-eluting compounds as it provides another separation dimension. Nevertheless, co-elution still represents a major hurdle in the identification of unknowns in complex mixtures.

If a true (chemical) separation is not possible the mathematical deconvolution of the profiles represents an alternative to deal with partial co-elution. There are several chemometric techniques that deal with the problem of deconvolution for tri-linear data (such as e.g. GC×GC-MS data). The most important ones are the generalized rank annihilation method (GRAM) [2], parallel factor analysis (PARAFAC) [3] and multivariate curve resolution (MCR) [4]. All of them have their own advantages and also suffer from their own limitations. In our case, we have selected the MCR-alternating least squares (ALS) algorithm. However, the proposed cross-validation approach could be adapted to the other methods.

In one-dimensional chromatography, automated deconvolution is not new. The ‘Automated Mass Spectral Deconvolution and Identification System’ (AMDIS) [5] from the National Institute of Standards and Technology is a widely-used deconvolution software that is often already embedded in instrument software. However, to the authors knowledge, no transfer to two-dimensional
chromatography has yet been made. The major hurdle when automating deconvolution methods is that most methods require the input of the number of chemical compounds expected in the mixture. This number corresponds to the number of mathematical components used for creating a deconvolution model of the data. The solution often is to gradually increase the number of components and manually check the results for their model’s validity. For target-compound analysis, this is rather easy as the obtained spectra using a number of components increasing from one to $n$ can be compared to existing library spectra of the known compounds. This method has been used, for example, by Parastar et al. [4] or Hoggard et al. [6]. For profiling applications however, the decision on when to stop adding components is less straight-forward. One attempt has been made by the latter-mentioned group [7]. In that paper, Hoggard et al. compare the obtained pure spectra using one to $n$ components in a non-target region. The optimum number of components is detected when going from $n$ to $n+1$ components results in the same spectra. In this case, an added component will just describe departures from bi-linearity (due to, for example, matrix effects). Other methods for determining the optimum number of components in a (semi-) automated way, although not applied to GC×GC-MS data, have been presented by Vivó-Truyols et al. [8] and Jellema et al. [9]. In the former, the structure of the residuals is investigated, looking to situations in which non-structured noise is obtained (which is a sign that the correct number of components has been detected). The latter method is based on singular-value decomposition, a technique that has also been frequently applied by other groups [e.g. 10, 11].

In this work we demonstrate a new method for the automatic determination of the optimum number of components used in the MCR-ALS algorithm. We propose a cross-validation strategy in which the obtained overall error of cross-validation can then be used to determine the optimum number of components. The aim is to construct a simple and reliable method that can be the base of a curve-resolution
strategy. The new method is illustrated on a smaller and a larger region of a GC×GC-MS chromatogram.

5.2 Theory

5.2.1 Unfolding the data

A GC×GC-MS apparatus is a third-order instrument [12]. This means that the data obtained can be organised as a third-order tensor (i.e. a “cube of data”), with the rows being the second-dimension retention time (2D-tR) points, the columns being the first-dimension retention time (1D-tR) points and the “slices” the mass-to-charge ratios (m/z). In order to apply MCR to the data, 2nd-order (and not 3rd-order) data should be produced. There are several ways to unfold the data and we decided to unfold the data as shown in Figure 5.1.

![Figure 5.1](image)

**Figure 5.1** Selected way of unfolding the three-way GC×GC-MS data. K corresponds to the first chromatographic dimension, L to the second dimension and M to the m/z-dimension.

The first-dimension retention time scans are hereby set on top of each other, leaving the mass-dimension (M) as the common space in the augmented matrix. The obtained matrix, X, is then used for the cross-validation routine described below.

5.2.2 Cross-validation routine

The proposed method to find the optimum number of components is based on a two-fold cross-validation scheme [13], and is illustrated in Figure 5.2.
In a first step, the (unfolded) X-data matrix (see Section 5.2.1 for a description of the data augmentation) was divided into two parts, \( X_{\text{even}} \) and \( X_{\text{odd}} \), both used as calibration and as validation set. Each part contains half of the rows of the X-matrix by selecting every second point in the new retention time axis for both \( X_{\text{even}} \) and \( X_{\text{odd}} \). Knowing that a fast-scanning mass spectrometer has been used in this study it can be assumed that in this way the data in either set still embodies the chromatographic nature of the signal (\( i.e. \) its continuity). As it will be illustrated later, this continuity is used to construct a predicted data set from the calibration set. Note that both \( X_{\text{even}} \) and \( X_{\text{odd}} \) contain all m/z-channels.

Two independent analyses were then performed per half data set (\( i.e. \) for \( X_{\text{even}} \) and \( X_{\text{odd}} \), as illustrated in Fig. 5.2. Each analysis consists of applying SIMPLISMA (“Simple-to-use, interactive, self-modelling, mixture analysis”) [14] to the calibration set, submitting the results of SIMPLISMA to MCR-ALS, and interpolating the results from MCR-ALS to construct the modeled X-matrix of the other half of the data (\( i.e. \) the validation set). These three operations
(SIMPLISMA, MCR-ALS and interpolation) were applied considering a different number of components (from 1 to nc), see Fig. 5.2. Firstly, SIMPLISMA is used to obtain an initial guess for the pure spectra that were then used as input for the MCR-ALS- routine. Both routines used were based on the work as presented by Jaumot et al. [15]. While these authors have developed an interactive tool for applying MCR-ALS we have used their routines in an automated manner. For example, SIMPLISMA was applied automatically finding those columns in the initial data matrix that offer the highest selectivity for each compound in the mixture. We have applied SIMPLISMA column-wise since it is easier to find selectivity in the m/z direction than in the chromatographic direction (i.e. finding m/z-channels that respond only to one compound is easier than finding retention times free of overlap). The output of the SIMPLISMA algorithm is a (collection of) initial guesses of pure spectra that are then the input of the MCR-ALS algorithm. Although other initial (pure) spectra may have been selected if one was to manually select them, we skipped this manual optimization step and accepted the automated answer from SIMPLISMA as the valid one. For a more detailed discussion of the different ways of finding initial guesses in MCR, refer to Vivó-Truyols et al. [16].

MCR-ALS is now applied in a second step to determine the pure chromatographic profiles and m/z-channels of all the compounds present in the data. MCR is a matrix-decomposition method that decomposes the matrix into so-called scores \( C \) and loadings \( S^T \) that are capable of modeling the original matrix excluding an error \( E \) (Eq. 1), where \( X \) in this case is \( X_{\text{even}} \) or \( X_{\text{odd}} \).

\[
X = C \cdot S^T + E \quad \text{Eq. 1}
\]

Generally, the solution of MCR is refined by alternating least squares, applying constraints like non-negativity, unimodality or known selectivity. At the end of the process, MCR returns the pure concentration profiles (C) and spectra (S) for \( nc \) components considered for both input matrices \( X_{\text{even}} \) and \( X_{\text{odd}} \) (in our case, for \( X_{\text{odd}}, C_{\text{odd}} \) and \( S_{\text{odd}} \) are obtained and for \( X_{\text{even}}, C_{\text{even}} \) and \( S_{\text{even}} \) are obtained). Note
that $C_{\text{odd}}$ and $C_{\text{even}}$ are still under-sampled chromatographic profiles, whereas $S_{\text{odd}}$ and $S_{\text{even}}$ are not under-sampled (all the m/z channels were considered).

In a third step, the modeled concentration profiles ($C_{\text{odd}}$ and $C_{\text{even}}$) are used to interpolate an estimation of the profiles of the other half of the data set that was not used for modeling. This is done by linear interpolation. One should remember that $X_{\text{odd}}$ and $X_{\text{even}}$ were selected in a way that the continuity of the chromatographic nature of the data set is more or less sustained. Linear interpolation is therefore very suitable to model the other half of the data set. If the number of scans per second-dimension peak is very low (e.g. if a slow-scanning mass spectrometer has been used or if very narrow peaks elute from the second-dimension column) other interpolation methods should be selected.

In this way, $C_{\text{odd}}$ is used to obtain $(C_{\text{even}})^{\text{int}}$, and $C_{\text{even}}$ is used to obtain $(C_{\text{odd}})^{\text{int}}$ (where the superscript “int” indicates that the values were interpolated). Next, the least squares estimation for $S$ (i.e. the modeled $\hat{S}$) is obtained from the interpolated values of $C$. For simplification, Eq. 2 shows the modification of Eq. 1 to solve for $S$ when $X_{\text{odd}}$ is used as the original $X$-matrix:

$$\hat{S}_{\text{even}} = ((C_{\text{even}})^{\text{int}}, (C_{\text{even}})^{\text{int}})^{-1} \cdot (C_{\text{even}})^{\text{int}} \cdot X_{\text{even}}$$

Eq. 2

In a next step, the modeled matrix $\hat{X}_{\text{even}}$ is obtained by the application of Eq. 1:

$$\hat{X}_{\text{even}} = (C_{\text{even}})^{\text{int}} \cdot \hat{S}_{\text{even}}$$

Eq. 3

Of course, $\hat{S}_{\text{odd}}$ and $\hat{X}_{\text{odd}}$ are obtained in the same way, using the interpolated concentration profiles for $C_{\text{odd}} ((C_{\text{odd}})^{\text{int}})$ and then obtaining $\hat{X}_{\text{odd}}$.

The overall cross-validation error is the sum of the squared differences between $X_{\text{even}}$ and $\hat{X}_{\text{even}}$ and between $X_{\text{odd}}$ and $\hat{X}_{\text{odd}}$. These steps are repeated for $nc$ number of components and the error is monitored. In case limited computational power is an issue it is advisable to restrict the maximum number of components to use (i.e. how often the loop is to be repeated) to the maximum number of chemical compounds expected. In our case, as we are investigating smaller regions of the chromatogram, it was set to 10 components as we did not expect more than 10
peaks to appear in any small region of the GC×GC-MS chromatogram. Note that this number only effects the computational time and is not critical to the developed algorithm.

5.3 Experimental

5.3.1 Sample background and instrumentation
The sample used to exemplify our method is an algae sample obtained from Wetsus and NIOO (Nederlands Instituut voor Onderzoek Oppervlaktewater). Raw dried algae cells were mixed with a 0.25 M aqueous trimethylsulfonium hydroxide (TMSH) solution. A very small aliquot of this mixture was put manually into a micro-cup which was placed into a liner, sealed and inserted into an Optic 3 PTV injector (ATAS GL) at room temperature. The injector was then first heated to 120 °C at 5 °C/s (1.5 min.) and then up to 250 °C (30 °C/s) where it was kept until the end of the run. The GC instrument was a HP7890 (Agilent, Amstelveen, The Netherlands) equipped with a LECO (Mönchengladbach, Germany) dual-stage, quad-jet thermal modulator. For detection, a LECO Pegasus III TOF-MS system was used. The acquisition rate of the TOF-MS was 100 spectra / second at a mass range of 50 to 500 Da.

5.3.2 Chromatographic analysis
The GC temperature programme of the first-dimension column started at 60 °C for 4.5 min, followed by a ramp of 5 °C / min to 220 °C followed by 10 °C / min to 325 °C with a final hold of 4 min. The second-dimension column followed the same programme with a temperature offset of 5 °C. The modulation time was 6 s. The GC×GC separations were performed using a 30 m x 0.25 mm I.D. TC-5MS column (5% phenyl-methyl siloxane) with a film thickness of 0.25 μm (GL Sciences) and as the second-dimension column an IntertCap 17 (50% phenyl-methylsiloxane) of 2 m x 0.1 mm I.D. with a film thickness of 0.1 μm (GL
Sciences). The 2D-GC data was exported as cdf-file and read into Matlab© for all further processing.

5.3.3 Data pre-processing and analysis

Figure 5.3 shows the entire GC×GC-MS chromatogram used as example chromatogram in this work (see Experimental for further sample information). Two regions (annotated A and B in Fig. 5.3) were selected from this chromatogram to exemplify our method. All masses below 80 Da were excluded as they do not reflect specific chemical information. This is not necessary for the developed algorithm but it simplifies building good models by MCR-ALS.

![Figure 5.3 Example GC×GC chromatogram. See experimental section for details. Two regions are indicated in which the cross-validation routine was applied.](image)

The method was applied to each region individually. In a first step, the data in the respective region was unfolded and the cross-validation routine was applied to the
unfolded data as described in Section 5.2. Multivariate curve resolution-alternate least squares (MCR-ALS) was performed for each first-dimension retention time “slice” using a non-negativity constraint in all dimensions, a maximum of 50 iterations and a 0.1 % convergence criterion. The algorithm applied was adapted from [15]. We decided to limit the constraints to non-negativity. The reason is that if we applied more constraints, the model was not flexible enough to model small departures from bi-linearity (originated from the compounds themselves) and the number of components obtained was incorrect. Generally, the more constraints applied the less flexible a model is. By using only one set of parameters (rather than manually optimising them for every application as done in [15]) less perfect models were accepted for reaching the greater goal, automation.

5.4 Results and discussion
The cross-validation strategy outlined in Section 5.2 has been developed in order to automatically determine the optimum number of components to be used for deconvolution (in our case by MCR). This can be done because the strategy allows the detection of situations of over-fitting: the overall cross-validation error should decrease with the number of components until over-fitting starts to occur (i.e. when the number of components is larger than the number of chemical compounds present in the mixture). This can be explained by the interpolation step. Interpolation only works on smooth chromatographic profiles, meaning that the obtained interpolated concentration profiles only make sense when the number of components is still lower than or equal to the number of chemical compounds present. Once the number of components is higher than the number of chemical compounds, noise is being modeled. However, as noise is a random (non-smooth) component, the interpolation step (which supposes that the profiles are smooth) will increase the error of the model. The result is that the previously decreasing
error of cross-validation will now increase - something that can be easily detected in an automated manner.

Figure 5.4 Selected region in GC×GC chromatogram (A) and the obtained errors of cross-validation for one to ten components (B).

The newly-developed cross-validation scheme is illustrated for the peak region depicted in Figure 5.4A, which corresponds to region A in Fig 5.3. From visual inspection two peaks can be seen in this figure (as annotated by the arrows). One would therefore expect an optimum model using two components. Note that the (mathematical) model just tries to describe the variation in the data and thus could not distinguish between a real, chemical compound and other (unwanted) chromatographic effects as for example column bleed that may appear in the chromatogram. Fig. 5.4B shows the plot of the overall error of cross-validation versus the number of components. As expected, the error decreases first when adding components and then increases again once the model starts to overfit the data (at four components, see encircled). Therefore, the optimum number of components for the region depicted in Figure 5.4A is not two but three. Note that even though only a slight increase can be observed it is enough to be detected in an automated manner. That the correct number of components indeed is three can be seen when looking at the obtained pure concentration profiles (Fig. 5.5A) and spectra (5.5B).
Figure 5.5 Obtained pure concentration profiles for component one (solid line), two (dotted line) and three (dashed line) (A) and the corresponding pure mass spectra for component one to three (B).

Clearly, there are three peaks present, which is what the cross-validation procedure was finding and what was not visual in the total-ion chromatogram in Fig. 5.4A. If a fourth component is added to build the model, the obtained pure chromatogram and spectrum of the fourth component are the same as the ones of the second (data not shown). As mentioned previously by Hoggard [8], the appearance of very similar spectra with additional components is a sign that the optimum number of components had been reached.

In literature, a manual evaluation of the obtained residuals by a given model is also often used to select the correct number of components: the number of components is increased gradually and the obtained residuals are inspected (e.g. [8]). If a peak is still present in the error matrix (i.e. that part of the matrix that has not been modelled; the component E in Eq. 1), the residuals will still show some structure and not just random variation as would be expected when only noise was left. This structure should disappear once the correct number of components has been reached. Figure 5.6A shows the residuals along the retention time axis for one, two and three components for the same example region shown in Fig. 5.4 (see
Figure caption). To produce this plot, the m/z direction is summed for each residual in the same way that the TIC represents the sum over all m/z channels in a chromatogram. From the figure, it becomes clear that no conclusion can be drawn from the structure of the residuals, something we have seen in all of our test regions (data not shown).

Another method often applied to determine the number of components is to inspect the accumulated sum of eigenvalues by performing singular-value decomposition (SVD) on the data matrix (e.g. [9]). Fig. 5.6B shows the explained variance obtained for one to ten components. One should note that, as the percentage of variance explained by the SVD is calculated with the same data used to perform the decomposition (i.e. it is not cross-validated), the percentage of variance explained will always decrease when adding more components. In this method the optimum number of components is usually determined by finding a change in the way the explained variance is decreasing (from a certain number of components onwards it only decreases slowly). Automatically detecting this slower decrease in a decreasing “curve” is intrinsically more difficult than...
detecting an increase (compare Figs. 5.4B with 5.6B) and in this case, one would visually struggle to determine the number of components to use.

**Application to larger regions**

Figure 5.7A shows region B of Fig. 5.3, together with the plot of the error of cross-validation (5.7B) obtained by our method and the variance explained when performing singular-value decomposition (5.7C).

![Figure 5.7](image)

**Figure 5.7** (A) Part of the GCxGC-TIC chromatogram shown in Fig. 5.3, (B) the error of cross-validation and (C) the variance explained by a SVD.

In this example, the error of cross-validation does not increase when overfitting has occurred but continues a gradual decrease. This might be due to the presence of structured noise that is unfortunately surprisingly common in analytical instrumentation [17]. Structured noise makes the addition of an extra component during the interpolation step (Section 5.2.2) beneficial from the residuals point of view. Or, in other words, sometimes noise can be interpolated as it shows a smooth trend. In situations like this, a good rule-of-thumb is to stop adding components when the decrease in the cross-validated error is less than 10%. The resulting optimum number of components, when applying this rule of thumb for this example would then be four. This is the same number as one would probably
select from the plot obtained from SVD (Fig. 5.7C). The results from the MCR-ALS routine using four components can be seen in Fig. 5.8.

![Figure 5.8](image_url)

**Figure 5.8** Obtained pure chromatograms (A) and spectra (B) from the MCR-ALS routine when using four components for the chromatogram shown in Fig. 5.7.

Figs. 5.8A show the obtained pure chromatograms, re-folded as two-dimensional chromatograms and Figs. 5.8B show the obtained pure spectra for these four components. From these plots, it can be deduced that indeed four peaks are present and a fifth component will only describe noise (data not shown). Clearly, the method can also be used for larger regions though there will be a practical limit, especially when an automatic solution is desired (e.g. the 10%- rule-of-thumb may fail or other stopping points may be necessary). Also, larger regions require higher computational power, something that still represents a huge problem when common personal computers are used for the calculations.
5.5 Conclusions

A new method for the automatic determination of the optimum number of components for applying MCR to GC×GC-MS data has been presented. It is based on a two-fold cross-validation scheme. The obtained overall cross-validation error decreases when adding components and increases again once over-fitting of the data starts to occur, *i.e.* when too many components have been used to build the model. In practice, however, the error sometimes follows a slow decrease only. This has been seen when the region analysed via MCR is relatively large. For these cases other rules on selecting the optimum number of components from the cross-validation plot have been developed, although they are not universally-applicable yet and still require some user-intervention. Overall, the method is at least as good as and sometimes superior to the inspection of the SVD eigenvalues. Its strong point lies in the fact that it is fully automatic and in that way making it efficient and less subject to subjective manual interpretation.

The developed method has been applied to smaller and larger regions in a GC×GC-MS chromatogram. In both cases the cross-validation scheme resulted in selecting the correct number of components for applying MCR-ALS in the respective region.

References