Institute for Advanced Studies
in Basic Sciences
Gava Zang, Zanjan, Iran

# Relationship between Duality Concept and Complementarity-Coupling Theorems in Case of Reducing the Rotational Ambiguity

Ph.D. Thesis

**Elnaz Tavakkoli**

**Supervisors: Prof. H. Abdollahi**

**Prof. R. Rajkó**

# Abstract

Self Modeling Curve Resolution (SMCR) is a class of techniques concerned with estimating pure profiles underlying a set of measurements on chemical systems. In general, the estimated profiles are ambiguous (non-unique) except if some special conditions fulfilled. Implementing the adequate information can reduce the so-called rotational ambiguity effectively, and in the most desirable cases lead to the unique solution. Therefore, studies on circumstances resulting in unique solution are of particular importance. In bilinear chemical (e.g., spectroscopic) data matrix, there is a natural duality between its row and column vector spaces using minimal constraints (non-negativity of concentrations and absorbances). The correspondence between the points in the considered space and the hyperplanes in the dual space can be applied to extract information in SMCR analysis. Infact, the conditions for achieving a unique solution can be investigated based on the duality concept as a general principle

## I. Duality based investigations in SMCR

### A. Clarification and visualization of the duality concept of curve resolution using three-component non-negative bilinear chemical data

In this work, it has been intended to clarify the duality concept and implement it in data structure visualization. Accordingly, duality principle is used to visualize the relationship between data points in each space and non-negativity borders in its dual space. Duality concept state that one point in one data space is corresponding with one hyperplane in the other space. Therefore, by implementing this concept and using data points, it has been visualize that the intersection of the all defined hyperplanes generates the outer boundaries in the dual mode. In order to reach a better understanding of this concept and its application in SMCR, hypothetical chromatographic system of three components is visualized in PC spaces.

### B. Duality based direct resolution of unique profiles using zero concentration region information

The duality relation can also be the proposed approach to exploit more information of the chemical data matrix if additional knowledge of the system is available. In this section, by applying the reliable information about the concentration window of components, the conditions of unique resolution has been explored. Therefore, the conditions of the unique solution according to duality concept and using zero concentration region information is intended to show using simulated and experimental datasets. It gives an easy way for finding how to use zero concentration region information based on duality in special cases to obtain unique solutions in SMCR methods. Additionally, the knowledge of a pure profile (e.g. spectrum) is used to discovery of information in SMCR method. In this way, graphical tools of PCA are utilized for illustrating the data structure.

## II .Development on generalized rank annihilation problem by implementing duality

The analytical chemist is frequently confronted with the problem of analyzing complex mixtures in the presence of any component in the sample that is not included in the calibration model. In these cases, it is desirable to be able to obtain quantitative information for a particular component without concern for the rest of the components in the sample. The property of quantitation of an analyte in the presence of unknown constituents is called second-order advantage. Additionally, in recent studies it has been shown that when the solution is unique or information is available for obtaining the unique solution, the duality concept,is a useful approach to extract information in SMCR. In a bilinear data matrix, there is a natural duality between its row and column spaces and it is possible to transform the coordinates in row space to the coordinates in column space and vice versa without using any constraint. In this section, a novel algorithm to achieve “second order advantage” is introduced based on duality. Moreover, informative geometrical visualizations beside mathematical formulas are presented for the proposed method. In this way, one of the most common approaches for visualization of data, Principle Component Analysis (PCA), is used. It is shown in order to determine λ, the value related to the contribution of the constituent of interest in an unknown mixture, the subspace of the all interfering compounds should be defined properly. Therefore, a matrix of λ values will be obtained in a systematic way and it is possible to calculate more precise estimation of λ.

## III. A rank reduction based normalization and visualization of rank annihilation

### A. A rank reduction based normalization

In this work a special normalization for a particular component in the system by using pure component spectra is derived. By implementing the general rank annihilation Wedderburn formula, it has been shown that the introduced normalization specifies the intensity of the same component in the dual mode. The interesting point is that the derived normalization is particularly applied for the analyte of interest without considering the remain part of the abstract space. By the help of the figures and plots, the relation between "known profile" and "normalization" is visualized. The introduced normalization result in the rank reduction for the known component.

### B. Visualization of rank annihilation

Although rank annihilation is a crucial concept in chemometrics, but it has not been investigated considering the dimensionality of the data. In this work, illustrative visualization of rank annihilation procedure has been provided. It has been illustrated that the reduction of the dimension of the residual, when standard matrix is subtracted from the mixture, can be monitored in rank annihilation problem. Therefore, in this section graphical visualizations are provided to depict the relation of "rank" and "dimensionality" when the standard component is annihilated from the data.

## IV. Soft-trilinear constraints for improved quantitation in Multivariate Curve Resolution

When a set of samples is to be analyzed with one data matrix per sample, the data is often presumed to have "trilinear" structure if the profile for each compound does not change shape or position from one sample to the other. By applying this information as a trilinearity constraint in SMCR methods, overlapping peaks related to the pure compounds of interest can be resolved in a unique way. In practice, many systems have non-trilinear behavior due to deviation from ideal response, for example a sample matrix effect, or changes in instrumental response (e.g., shifts and/or changes in the shape of chromatographic peaks). In such cases, the trilinear model is not valid because every analyte does not have the same peak shape or position in every sample. In such cases, the unique profiles obtained by strictly enforced trilinearity constraints will not necessarily produce true profiles because the data set does not follow the assumed trilinear behavior.

In this section, we introduce "soft-trilinearity constraints" and a new MATLAB program to permit peak profiles of the components of interest to have small deviations in their shape and position from sample to sample. In order to visualize the results, soft-trilinearity constraints were incorporated into a systematic grid search algorithm for the case of a three-component system [1]. This algorithm is general and can be applied to any MCR method. Results are provided for simulated noisy data with non-trilinear behavior and one experimental data set. The results show that implementing soft-trilinearity constraints reduces the range of feasible solutions considerably compared to the application of simpler non-negativity constraints. The results of this approach are compared to other methods including PARAFAC2 and MCR-ALS with hard-trilinearity constraints. It is shown that the methods employing hard-trilinearity constraints lead to incorrect solutions, or produce solutions outside the range feasible solutions.

## Contents

Chapter 1 ........ 1
General Introduction ........ 1
1.1 Chemometrics ........ 1
1.2 Data Mining ........ 2
1.2.1 Data ........ 2
1.3 Factor Analysis ........ 4
1.3.1 Principal Component Analysis ........ 4
1.4 Curve Resolution Method ........ 5
1.4.1 Soft Modeling Methods ........ 6
1.4.1.1 Ambiguities ........ 6
1.4.1.2 Constraints, Definition ........ 7
1.4.1.2.1 Non-negativity ........ 8
1.4.1.2.2 Unimodality ........ 8
1.4.1.2.3 Equality, Known Profiles ........ 8
1.4.1.2.4 Closure ........ 9
1.4.1.2.5 Trilinearity ........ 9
1.4.1.2.6 Selectivity, Zero concentration windows, Local-Rank constraints ........ 10
1.4.1.2.6.1 Unique solution: Manne's resolution theorems ........ 11
1.4.1.2.6.2 Window Factor Analysis (WFA) ........ 12
1.4.1.2.6.2 Sub Window Factor Analysis (SWFA) ........ 13
1.5 Duality ........ 14
1.6 The Extent of Rotational Ambiguity ........ 15
1.6.1 Two-component system: Lawton-Sylvestre Method ........ 15
1.6.2 Resolution of Rotational Ambiguity in Three-Component Systems ........ 17
1.6.3 Borgen Method ........ 20
1.6.4 The FACPACK-toolbox for AFS-computations ........ 24
1.7 Normalization ........ 25
Chapter 2 ........ 27
2. Historical background ........ 27
2.1 Self Modeling Curve Resolution Methods (SMCR) ........ 27
2.2 PCA and duality ........ 27
2.3 Unique Solution ........ 31
2.4. Duality based second order advantage ........ 31

2.5 Soft trilinearity ........ 34
Chapter 3 ........ 38
3. Numerical Experiments ........ 38
3.1 Beer-Lambert’s Law ........ 38
3.2 Gaussian Curves ........ 40
3.3 Generation of Simulated Mixtures ........ 40
3.3.1 Three-component mixtures with overlapping Gaussian chromatography peaks ........ 40
3.3.2 Two-component multi set data ........ 41
3.3.3 Three-component multi set data ........ 42
3.3.4 Non-trilinear, simulated data: ........ 43
3.4 Real data sets ........ 45
3.4.1 Three component amino acids ........ 45
3.4.2 JODA excitation-emission data[105] ........ 45
3.4.3 Non-trilinear HPLC-DAD data[106, 107] ........ 47
3.5 Software ........ 47
Chapter 4 ........ 48
4. Results and Discussion ........ 48
4.1. General Introduction on SMCR problem ........ 48
4.1.1 Clarification and visualization of the duality concept of curve resolution using three-component non-negative bilinear chemical data ........ 49
4.1.2 Duality based direct resolution of unique profiles using zero concentration region information ........ 54
4.1.2.1Concentration profile ........ 55
4.1.2.2 Spectral profile ........ 57
4.1.2.3 Known pure profile ........ 61
4.1.2.4 Real example ........ 66
4.1.2.5 Conclusion ........ 69
4.2 Development on generalized rank annihilation problem by implementing duality ........ 70
4.2.1 Geometrical Visualization ........ 72
4.2.2 Numerical experiment ........ 77
4.2.2.1 Simulated data ........ 77
4.2.2.1.1 Example1 ........ 78
4.2.2.1.2. Example 2 ........ 79
4.2.2.1.3 Example 3-5 ........ 80

4.2.2.1.4 Conclusion ........................................................................................... 81
4.3 A rank reduction based normalization and visualization of rank annihilation .................. 82
4.3.1 Introduction to generalized rank reduction problem...................................................... 83
4.3.1.1 Wedderburn’s formula ........................................................................................... 83
4.3.1.2 Rank annihilation based normalization.......................................................................... 85
4.3.2 Visualization of Rank annihilation ................................................................... 88
4.3.2.1 Dimension reduction ........................................................................................... 89
4.3.3 A Look to the Future Work:................................................................................ 92
4.3.3.1 Rank Annihilation and Trilinearity Constraint ............................................................ 92
4.3.4 Conclusion ........................................................................................................ 94
4.4 Soft-trilinear constraints for improved quantitation in Multivariate Curve Resolution ..... 96
4.4.1 Soft Trilinarity constraint in Grid search ......................................................... 96
4.4.1.2. Soft-trilinearity constraints: .......................................................................... 98
4.4.1.3 The AFS under soft non-negativity constraints: ......................................................... 100
4.4.1.4 The reduction of AFS under soft-trilinearity constraints:........................................... 101
4.4.1.4.1 Example 1: ...................................................................................................... 101
4.4.1.5 Hard constraints .............................................................................................. 105
4.4.1.5.1 Example 1: ...................................................................................................... 105
4.4.2 Example 2: ......................................................................................................... 107
4.4.3 RSM ..................................................................................................................... 108
4.4.4. Conclusion .......................................................................................................... 110

# Table of Figures:

Figure 1.1 The relationship and distinction among the concepts of "data order" and "data way". This figure is borrowed from the reference.[14] ..... 3

Figure 1.2 Implementation of the trilinear constraint in the MCR-ALS algorithm. This figure is borrowed from the reference [48] ..... 10

Figure 3.1 Diagram demonstrating Beer-Lambert absorption of a beam of light as it travels through a cuvette of size l containing solution of concentration c and molar absorptivity ε..... 38

Figure 3.2 Diagram demonstrates decomposition of spectra R into corresponding concentration (C) and bsorptivity (S) matrices. The matrix E is representative of the experimental noise associated with any real measurement. ..... 39

Figure 3.3, Plots of (a) concentration profiles, (b) spectra profiles, (c) simulated data matrix ..... 41

Figure 3.4, Plots of (a) concentration profiles, (b) spectra profiles, (c) simulated data matrix ..... 41

Figure 3.5. Simulated two-component data: (a) elution profiles, (b) spectra profiles, (c) simulated standard of one component, containing red component, (d) simulated standard of two component (e) simulated mixture data ..... 42

Figure 3.6. Simulated three-component data: (a) elution profiles, (b) spectra profiles, (c) simulated standard data, $R_s$ , containing red component (d) simulated mixture data, R. ..... 43

Figure 3.7. (a) Simulated concentration profile(s), (b) simulated spectra profiles, (c) $R_s$, standard matrix of analyte with $R_{ts}=50$ and $W_s=23.55$; (d) $R_1$, first mixture with $R_{t1}=48$ and $W=20.02$ for analyte and (e) $R_2$, second mixture data with $R_{t2}=53$ and $W=27.09$ for analyte (f), $R_{aug}$, augmented data with added noise amount of 0.5% maximum of the data set. ..... 44

Figure 3.8, (a) Augmented fluorescence dataset, including sample 2, 4 and 5. (b) Augmented fluorescence dataset, including sample 1, 3 and 4. ..... 45

Figure 3.9. Experimental fluorescence data, test samples in Example 3 (a), 4 (b), 5 (c) and standard (d). ..... 46

Figure 3.10, Non-trilinear HPLC-DAD data ..... 47

Figure 4.1. Illustration of the inner and outer pyramids for a hypothetical three component system. (a) and (b) show the plot of inner pyramid in V- and U-space, respectively. The data point of the first row is highlighted as a star in (c) and (d) shows its dual plane in U-space. The data point of the first column is highlighted as a star in (e) and (f) shows its dual plane in V-space, (g) and (h) are a depiction of inner-outer pyramids in V- and U-space, respectively. ..... 53

Figure 4.2 (a) Depiction of concentration window of components. The solid line rectangle shows the concentration window of the blue component (analyte) and the dashed line rectangle shows its absence window. (b) Score plot of data matrix (pink dots) and the spectral subspace of the green and red components (interferents shown with blue stars). (c) Calculated concentration profile for the blue component. ..... 57

Figure 4.3 (a) Depiction of concentration window of components. The solid line rectangle shows the concentration window of the blue (interferent) and green (analyte) components, and the dashed line rectangle shows the concentration window of green (analyte) and red (interferent) components. (b) The spectral subspace of the blue and green components on the score plot. (c) The spectral subspace of the green and red components on the score plot. (d) Concentration subspace of the blue and red components (interferents). (e) The analyte is at the intersection of spectral subspaces in (a) and (b). (f) Calculated spectral profile for the green component. ..... 60

Figure 4.4 The representation of the relation between known profile in spectral space and the deduced plane (line, with normalization) in concentration space when the blue component is known. 63
Figure 4.5. The representation of the relation between known profile in spectral space and the deduced plane (line, with normalization) in concentration space when the red component is known. 64
Figure 4.6. The representation of the relation between known profile(s) in spectral space and the deduced plane (line, with normalization) in concentration space when the red and blue components are known. 65
Figure 4.7. (a) Concentration window of tyrosine, phenylalanine, and tryptophan is shown with horizontal blue, green and red lines respectively. (b) resolved excitation profile of tyrosine. 67
Figure 4.8. (a) Concentration window of tyrosine, phenylalanine, and tryptophan is shown with horizontal blue, green and red lines in turn. (b) resolved emission profile of tyrosine. 68
Figure 4.9. Duality relation between standard, red line, and interference subspace, blue plane, in (a) column and (b) row space. 74
Figure 4.10. Linear combination of standard and interfering compound to produce data points in (a) column and, (b) row space of the data. 75
Figure 4.11. **nc** is the vector orthogonal to interference subspace. Defining **nc** based on duality, leads to determine λ. 76
Figure 4.12. Illustration of data points (black dots), inner (dashed black lines), outer (green lines) boundaries in a two-component system. (a) $s_1$ is the pure spectrum of the first component, shown with black circle. (b) Depiction of normalization line which contains pure concentration of the first component, $c_1$. 86
Figure 4.13. (a) coordinate of the fixed point in V-space is changed as shown with red arrow.(b) the normalized subspace related to $c_1$ will change as shown with arrow as well. 87
Figure 4.14. The normalization subspace contains $c_1$ while the duality line provides a unique solution for $c_2$. 88
Figure 4.15 depiction of data (black), standard (red) and residual (blue) points in V-space of the data. 89
Figure 4.16 (a) the *30th* row of data matrix R, $R_s$ and E are shown with pink lines. (b) By multiplying with λ, the standard and residual point change to new positions, shown with red and blue circles on solid pink lines. 90
Figure 4.17 (a) *1-dimensional* residual subspace by multiplying the standard points in $\lambda_1$=0.8571 (red dashed line) and (b) *1-dimensional* residual subspace by multiplying the standard points in $\lambda_2$=2 (red dashed line).The original standard data points are shown with red dot-line. 91
Figure 4.18 the linear combination of the standard and residual points to produce *31th* and *22th* row of the data in pink and green respectively. 92
Figure 4.19 resolved profiles and λ values for the case of two component system. 94
Figure 4.20. Implementation of trilinearity soft constraints. The concentration profiles in the first column of the resolved concentration matrix [$c_{11}$;$c_{12}$;$c_{13}$] are folded to give the matrix $C_1$. By applying PCA to $C_1$, eigenvalues can be obtained and then *k* is calculated. This procedure is applicable for any component which is desired to follow trilinearity constraints. 99
Figure 4.21. (a) Area of Feasible Solutions (AFS) for analyte (component B) by non-negativity and trilinearity soft constraints using different alpha levels [$\alpha$ =1.60, 1.54, 1.52, 1.51, 1.49], y1 and y2 are the coordinates (scores) of the projected column vectors in the concentration space and we added a description on figure caption, (b-g) translated concentration profiles. 103

Figure 4.22 Calculated concentration profiles with soft trilinearity constraint (α =1.54), PARAFAC2 (green) and MCR-ALS for Example 1 (black)..................................................... 104
Figure 4.23 (a) calculated AFS for fenitrothion with soft nonnegativity (blue) and soft trilinearity constraint (red), (b) translated concentration profiles, (c) result of PARAFAC2 (green) and MCR-ALS (black). ............................................................................................................ 107
Figure 4.24. Obtained Response Surface for different non-trilinear behaviors due to changes in retention time(X), peak width (Y) and noise level. .................................................................. 110

## List of Schematics

Scheme 1.1 Algorithm of grid search algorithm in three-component system[1] 19

# List of tables

Table 4.1 Calculated concentration of tyrosine in the mixture with Relative Error (RE) ............ 68
Table 4.2. The result obtained by the proposed algorithm for simulated two component system. ........................................................................................................................................ 78
Table 4.3. The result obtained by the proposed algorithm for simulated three component system. ........................................................................................................................................ 79
Table 4.4. The result obtained by the proposed algorithm for real data sets. ............................... 80
Table 4.5 Calculated concentration values associated with the application of soft-trilinear constraints using different levels of $\alpha$ and methods using hard-trilinear constraints, PARAFAC2 and MCR-ALS for analyte (B) in simulated HPLC data set. ..................................................... 104
Table 4.6 Calculated concentration values associated with the application of soft-trilinear constraint using $\alpha$=3 and methods using hard-trilinear constraints, PARAFAC2 and MCR-ALS for fenitrothion in experimental HPLC data set.......................................................................... 108
Table 4.7 Three level fractional factorial design. ....................................................................... 109