



TRICAP 2009

ThRee-way methods In Chemistry And Psychology



Vall de Núria – Spain
June 14 - 19, 2009



MINISTERIO
DE CIENCIA
E INNOVACIÓN

WELCOME TO TRICAP 2009

We are happy to welcome you all to the 6th edition of the TriCAP conference, to be held in one of the most beautiful valleys of the Catalan Pyrenees, Vall de Núria. The choice of the venue is not a coincidence, a secluded valley and an inspiring mountain landscape looked as a perfect environment to foster relaxed and fruitful scientific discussions among the three-way community.

We would like to thank all the people coming to this conference: those who are eager to learn from this meeting and, in a particular way, those who will share their scientific expertise through their communications. All conferees have helped to configure a very attractive scientific program, where theoretical and applied aspects combine, where very diverse multi-way analysis prospects are represented and where experience and youth can find their place. All your contributions have resulted in a very balanced program that we hope you can enjoy.

We also thank our sponsors, the research institutions and societies (CSIC, Universitat de Barcelona and Societat Catalana de Química, IEC) and the Catalan and Spanish governments for trusting this scientific initiative and providing financial support to develop it.

It is a pleasure for us hosting this 6th edition of TriCAP and we sincerely hope that the combination of science and nature can meet your personal and scientific expectations.

Benvinguts a la Vall de Núria!

Welcome to TriCAP 2009!

Romà Tauler
Age Smilde
Anna de Juan
Joaquim Jaumot

INDEX

	Pag.
- General program	4
- Technical program	8
- ABSTRACTS	

MULTIWAY DATA ANALYSIS, THEORETICAL ASPECTS

<u>Comm. 1</u> - Brett W. Bader - Advances in Data Analysis using PARAFAC2 and Three-way DEDICOM	12
<u>Comm. 2</u> - Lieven de Lathauwer - Best multilinear rank approximation: new results	13
<u>Comm. 3</u> - Paolo Giordani - Three-way component models for imprecise data	14
<u>Comm. 4</u> - Boris Khoromskij - Tensor-structured numerical methods for solving multidimensional equations	15
<u>Comm. 5</u> - Morten Morup - Bayesian and Non-linear Multi-way modelling	16
<u>Comm. 6</u> - Dimitri Nion - Selecting the basic parameters of a decomposition in rank-(L,L,1) terms	17
<u>Comm. 7</u> - Jorge Tendeiro - Simplicity transformations for three-way arrays with symmetric slices	18
<u>Comm. 8</u> - Marina Cocchi - A classification tool for N-way array based on SIMCA methodology	19

MULTIWAY DATA ANALYSIS, COMPUTATIONAL ASPECTS

<u>Comm. 9</u> - Nikos Sidiropoulos - Adaptive Algorithms to Track the PARAFAC Decomposition of a Third-Order Tensor	20
<u>Comm. 10</u> - Alwin Stegeman - The Candecomp/Parafac decomposition – diverging components and how to avoid them	21
<u>Comm. 11</u> - Evrim Acar - An Optimization Approach for Fitting a CANDECAMP/ PARAFAC Model with Applications in Social Network Analysis	22
<u>Comm. 12</u> - Mohamed Hanafi -New computational properties for	

Hierarchical Principal Component Analysis (HPCA) and its relation to PARAFAC Model	23
<u>Comm. 13</u> - Giorgio Tomasi - QR PARAFAC	24
<u>Comm. 14</u> - Hai-Long Wu - A comparison of several second-order calibration algorithms	25
<u>Comm. 15</u> - Alberto Ferrer and José Manuel Prats - 3-way methods: a practitioner perspective	26
<u>Comm. 16</u> - Michel Tenenhaus - A criterion based PLS approach to structural equation modeling	27

MULTISET DATA ANALYSIS

<u>Comm. 17</u> - Age Smilde - Multilevel Multiway Analysis	28
<u>Comm. 18</u> - Katrijn Van Deun - A unifying framework for simultaneous component methods	29
<u>Comm. 19</u> - Iven Van Mechelen - A generic model for data fusion	30
<u>Comm. 20</u> - Eva Ceulemans - CLASSI modeling of sequential processes and individual differences therein	31

MULTIWAY AND MULTISET DATA ANALYSIS, APPLICATIONS

<u>Comm. 21</u> - Sungjin Hong - Multilinear modeling of brain imaging data	32
<u>Comm. 22</u> - Nathaniel E. Helwig - Parallel Factor Analysis of Gait Data	33
<u>Comm. 23</u> - Ilgils Ibragimov and Vladislav Y. Orekhov - Application of Multi-Dimensional Decomposition in BioMolecular NMR Spectroscopy	34
<u>Comm. 24</u> - Mark Van Benthem - Three-Way Factor Analysis of Large-Microscopic Hyperspectral Images: Compression and Analysis of Very Large, Small Images	35
<u>Comm. 25</u> - Sarah Rutan - Analysis of Multi-way 2D-Liquid Chromatography Diode Array Data for Metabolomic Studies	36
<u>Comm. 26</u> - Federico Marini - Two and three way methods for the resolution of overlapping peaks in the determination of polyphenols in olive oil by HPLC-DAD	37
<u>Comm. 27</u> - José Manuel Amigo - Improvements on GC-MS aroma profile of IIDRØD PIGEON (Malus domestica) Apples exposed to different length of ripening time by using PARAFAC2	38
<u>Comm. 28</u> - Ricardo Leardi - Three-way Principal Component Analysis applied to Noodles Sensory Data Analysis	39

GENERAL PROGRAM

SUNDAY 14/06

El Prat Airport, meeting point Terminal A 13.00-14:00

Romà Tauler mobile phone number: (+34) 667.826.757

14.00 h - Bus service airport to Ribes de Freser (service included)

17.40 h - Rack railway to Núria (service included)

18.40 h - Arrival to Núria

Check in hotel

20.00 h - Dinner

21.00 h - Registration. Collecting information (Hall Auditorium)

21.30 h - Auditorium. Video presentation about Vall de Núria

Bar until 24.00 h (services not included)

MONDAY 15/06

8.00 h – Breakfast

Morning working sessions chair: Age Smilde

9.00 h - Working session 1 - Communication 1 (9.00 - 9.40h)

Discussion (9.40 - 9.55h)

Communication 2 (9.55 - 10.35h)

Discussion (10.35 - 10.50h)

10.50 h - Coffee break

11.10 h - Working session 2 - Communication 3 (11.10 - 11.50h)

Discussion (11.50 - 12.05h)

Communication 4 (12.05 - 12.45h)

Discussion (12.45 - 13.00h)

13.00 h - Lunch

Afternoon working sessions chair: Nikos Sidiropoulos

15.00 h - Presentation of the special issue of Journal of Chemometrics in memorial of Richard Harshman

15:15 - Working session 3 - Communication 5 (15.15 - 15.55h)

Discussion (15.55 - 16.10h)

Communication 6 (16.10 - 16.50h)

Discussion (16.50 - 17.05h)
17.05 h - Coffee break
17.25 h - Working session 4 - Communication 7 (17.25 - 18.05h)
Discussion (18.05 - 18.20h)
Communication 8 (18.20 - 19.00h)
Discussion (19.00 - 19.15h)
20.00 h - Dinner
21.30 h – Evening concert. Works of Catalan composers and other piano cello repertoire.
Carolina Gispert (piano), Mireia Quintana (cello).
Bar until 24.00 h (services not included)

TUESDAY 16/06

8.00 h – Breakfast

Morning working sessions chair: Lieven de Lathauwer

9.00 h - Working session 5 - Communication 9 (9.00 - 9.40h)
Discussion (9.40 - 9.55h)
Communication 10 (9.55 - 10.35h)
Discussion (10.35 - 10.50h)
10.50 h - Coffee break
11.10 h - Working session 6 - Communication 11 (11.10 - 11.50h)
Discussion (11.50 - 12.05h)
Communication 12 (12.05 - 12.45h)
Discussion (12.45 - 13.00)
13.00 h - Lunch
14.00 h – Non-scientific activity: Rowing boats on the lake (included)

Afternoon working session chair: Eva Ceulemans

15.00 h - Working session 7 - Communication 13 (15.00 - 15.40h)
Discussion (15.40 - 15.55h)
Communication 14 (15.55 - 16.35h)
Discussion (16.35 - 16.50h)
16.50 h - Coffee break
17.10 h - Working session 8 - Communication 15 (17.10 - 17.50h)
Discussion (17.50 - 18.05h)
Communication 16 (18.05 - 18.45h)
Discussion (18.45 - 19.00h)
20.00 h - Dinner
21.30 h – Non-scientific activity: Walk promenade and stars watching

Bar until 24.00 h (services not included)

WEDNESDAY 17/06

8.00 h - Breakfast

9.00 h - Hiking activity.

Option 1. Hike to Queralbs (Trail by “Camí de Núria” down to Queralbs (3 hours) and going back by rack railway) (*easy but leg breaking*)

Option 2. Go and return to Queralbs by rack railway (*not trained, very easy*)

Option 3. Hike to one of the surrounding mountains (*only trained*)

14.30 h - Lunch

Afternoon working session chair: Anna de Juan

16.00 h - Working session 9 - Communication 17 (16.00 - 16.40h)

Discussion (16.40 - 16.55h)

Communication 18 (16.55 - 17.35h)

Discussion (17.35 - 17.50h)

17:50 h - Coffee break

18.10 h - Working session 10 - Communication 19 (18.10 - 18.50h)

Discussion (18.50 - 19.05h)

Communication 20 (19.05 - 19.45h)

Discussion (19.45 - 20.00h)

21.00 h - Dinner

Bar until 24.00 h (services not included)

THURSDAY 18/06

8.00 h – Breakfast

Morning working session chair: Brett W.Bader

9.00 h - Working session 11 - Communication 21 (9.00 - 9.40h)

Discussion (9.40 - 9.55h)

Communication 22 (9.55 - 10.35h)

Discussion (10.35 - 10.50h)

10.50 h - Coffee break

11.10 h - Working session 12 - Communication 23 (11.10 - 11.50h)

Discussion (11.50 - 12.05h)

Communication 24 (12.05 - 12.45h)

Discussion (12.45 - 13.00)

13.00 h - Lunch

Afternoon working session chair: Romà Tauler

15.00 h - Working session 13 - Communication 25 (15.00 - 15.40h)
Discussion (15.40 - 15.55h)
Communication 26 (15.55 - 16.35h)
Discussion (16.35 - 16.50h)

16.50 h – Coffee break

17.10 h - Working session 14 - Communication 27 (17.10 - 17.50h)
Discussion (17.50 - 18.05h)
Communication 28 (18.05 - 18.45h)
Discussion (18.45 - 19.00h)

19.00 h – Closing scientific part of the meeting

20.00 h - Conference dinner (“Cabanya dels pastors”, included)
Snack and drinks on the terrace
Dinner (approx. at 21.00h)
Drinks and closing conference (until late night)

FRIDAY 19/06

8.00 h - Breakfast
Checking out hotel

10.10 h - Rack railway back to Ribes del Freser (service included)

11.00 h - Bus service airport from Ribes to El Prat Airport (service included)

13.30 h - Arrival to El Prat Airport

TECHNICAL PROGRAM

Monday 14th June

MULTIWAY DATA ANALYSIS, THEORETICAL ASPECTS

Working session 1

Communication 1 (9.00h) - Brett W. Bader

Advances in Data Analysis using PARAFAC2 and Three-way DEDICOM

Communication 2 (9.55h) - Lieven de Lathauwer

Best multilinear rank approximation: new results

Working session 2

Communication 3 (11.10h) - Paolo Giordani

Three-way component models for imprecise data

Communication 4 (12.05h) - Boris Khoromskij

Tensor-structured numerical methods for solving multidimensional equations

Working session 3

Communication 5 (15.15h) - Morten Morup

Bayesian and Non-linear Multi-way modelling

Communication 6 (16.10h) - Dimitri Nion

Selecting the basic parameters of a decomposition in rank-(L,L,1) terms

Working session 4

Communication 7 (17.25h) - Jorge Tendeiro

Simplicity transformations for three-way arrays with symmetric slices

Communication 8 (18.20h) - Marina Cocchi

A classification tool for N-way array based on SIMCA methodology

Tuesday 15th June

MULTIWAY DATA ANALYSIS, COMPUTATIONAL ASPECTS

Working session 5

Communication 9 (9.00h) - Nikos Sidiropoulos

Adaptive Algorithms to Track the PARAFAC Decomposition of a Third-Order Tensor

Communication 10 (9.55h) - Alwin Stegeman

The Candecomp/Parafac decomposition – diverging components and how to avoid them

Working session 6

Communication 11 (11.10h) - Evrim Acar

An Optimization Approach for Fitting a CANDECOMP/ PARAFAC Model with Applications in Social Network Analysis

Communication 12 (12.05h) - Mohamed Hanafi

New computational properties for Hierarchical Principal Component Analysis (HPCA) and its relation to PARAFAC Model

Working session 7

Communication 13 (15.00h) - Giorgio Tomasi

QR PARAFAC

Communication 14 (15.55h) - Hai-Long Wu

A comparison of several second-order calibration algorithms

Working session 8

Communication 15 (17.10h) - Alberto Ferrer and José Manuel Prats

3-way methods: a practitioner perspective

Communication 16 (18.05h) - Michel Tenenhaus

A criterion based PLS approach to structural equation modeling

Wednesday 16th June

MULTISET DATA ANALYSIS

Working session 9

Communication 17 (16.00h) - Age Smilde

Multilevel Multiway Analysis

Communication 18 (16.55h) - Katrijn Van Deun

A unifying framework for simultaneous component methods

Working session 10

Communication 19 (18.10h) - Iven Van Mechelen

A generic model for data fusion

Communication 20 (19.05h) - Eva Ceulemans

CLASSI modeling of sequential processes and individual differences therein

Thursday 18th June

MULTIWAY AND MULTISET DATA ANALYSIS, APPLICATIONS

Working session 11

Communication 21 (9.00h) - Sungjin Hong

Multilinear modeling of brain imaging data

Communication 22 (9.55h) - Nathaniel E. Helwig

Parallel Factor Analysis of Gait Data

Working session 12

Communication 23 (11.10h) - Ilgils Ibragimov and Vladislav Y. Orekhov

Application of Multi-Dimensional Decomposition in BioMolecular NMR Spectroscopy

Communication 24 (12.05h) - Mark Van Benthem

Three-Way Factor Analysis of Large-Microscopic Hyperspectral Images: Compression and Analysis of Very Large, Small Images

Working session 13

Communication 25 (15.00h) - Sarah Rutan

Analysis of Multi-way 2D-Liquid Chromatography Diode Array Data for Metabolomic Studies

Communication 26 (15.55h) - Federico Marini

Two and three way methods for the resolution of overlapping peaks in the determination of polyphenols in olive oil by HPLC-DAD

Working session 14

Communication 27 (17.10h) - José Manuel Amigo

Improvements on GC-MS aroma profile of I IDRØD PIGEON (Malus domestica) Apples exposed to different length of ripening time by using PARAFAC2

Communication 28 (18.05h) - Riccardo Leardi

Three-way Principal Component Analysis applied to Noodles Sensory Data Analysis

ABSTRACTS

Communication 1

Advances in Data Analysis using PARAFAC2 and Three-way DEDICOM

B. W. BADER¹

¹ Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-1318, {*bwbader@sandia.gov*}

PARAFAC2 is a model for analyzing three-way data consisting of symmetric frontal slices, usually of cross-product matrices (e.g., covariance matrices or scalar product matrices). Three-way DEDICOM is a related model that fits three-way data with asymmetric frontal slices and has been used to discover patterns in international import/export data and in social networks in the Enron email data set¹. The traditional approach for fitting these models to data is with an alternating least-squares procedure, where the factor matrices are updated one at a time.

In this presentation, algorithmic alternatives for fitting PARAFAC2 and DEDICOM to large data sets are developed. The new techniques use gradient-based optimization methods, which perform an all-at-once update of the factor matrices. These techniques are suitable for large-scale data, and several illustrative examples of analyzing large data sets from bibliometrics and cyber traffic are shown.

Acknowledgement

This work was funded by Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

We thank Danny Dunlavy, Evrim Acar and Tamara Kolda for the use of their Poblano Toolbox in MATLAB.

References

1. B. Bader, R. Harshman, T. Kolda, Proceedings of the Seventh IEEE International Conference on Data Mining (2007), 33-42.

Communication 2

Best multilinear rank approximation: new results

L. DE LATHAUWER^{1,2}, M. ISHTEVA², P.-A. ABSIL³ AND S. VAN HUFFEL²

¹ K.U.Leuven Campus Kortrijk, Group Science, Engineering and Technology,
E. Sabbelaan 53, 8500 Kortrijk, Belgium, *Lieven.DeLathauwer@kuleuven-kortrijk.be*

² K.U.Leuven, E.E. Dept. (ESAT), Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

³ Université catholique de Louvain, Dept. of Mathematical Engineering, Av. G. Lemaître,
B-1348 Louvain-la-Neuve, Belgium

We will present three new algorithms for the computation of the best multilinear rank approximation of a given higher-order tensor: (i) a Newton-type algorithm, (ii) a trust-region method and (iii) a conjugate gradient-type method. These methods exploit differential geometric properties of the manifolds on which they operate. We will discuss the issue of local optima. We will introduce the concept of hierarchical Tucker compression. Time permitting, we will also discuss some applications.

References

1. M. Ishteva, L. De Lathauwer, P.-A. Absil, S. Van Huffel, *Int. Journal of Pure and Applied Mathematics* **42** (2008), 337--343.
2. M. Ishteva, L. De Lathauwer, P.-A. Absil, S. Van Huffel, *Numerical Algorithms* (2009), to appear.

Communication 3

Three-way component models for imprecise data

P. GIORDANI

Department of Statistics, Probability and Applied Statistics, Sapienza University of Rome, P.le Aldo Moro, 5, Rome, 00185, Italy, *paolo.giordani@uniroma1.it*

The data under investigation are often uncertain. Uncertainty is usually referred to as randomness. Nonetheless, other sources of uncertainty may occur (see Klir, 2006). In particular, the empirical information may be imprecise because of the use of linguistic labels, partial ignorance, the calibration of the measurement devices or the vagueness associated to the definition of concepts. Generally speaking, imprecision derives from the uncertainty concerning the placement of an outcome in a given class. A possible way to manage imprecision is to transform the available information into interval valued or fuzzy data.

Here, we address the problem of summarizing a set of three-way interval valued or fuzzy data. In fact, also in these cases, it can be fruitful to analyze the underlying structure of the data. It is important to note that two-way component models for imprecise data are available in the literature. See, for instance, Cazes et al. (1997), Coppi et al. (2006), D'Urso, Giordani (2005), Giordani, Kiers (2004). The corresponding three-way extensions are based on suitable generalizations of the well-known Tucker3 (Tucker, 1966) and CANDECOMP/PARAFAC (Carroll, Chang, 1970; Harshman, 1970) models. When the (three-way imprecise) data are a random sample drawn from a reference population, the data at hand are affected not only by imprecision but also by randomness. Although the Tucker3 and CANDECOMP/PARAFAC models, as well as their generalizations for imprecise data, are usually performed in an exploratory context, it may be desirable to provide a measure of the statistical validity of the extracted components. In this respect, on the basis of the findings in Kiers (2004), a non-parametric bootstrap procedure for computing the standard errors of the model parameters can be adopted.

The results of a simulation experiment and some applications to real data are reported in order to show how these component models work.

References

1. P. Cazes, A. Chouakria, E. Diday, Y. Schektman, *Revue de Statistique Appliquée* **45** (1997), 5–24.
2. J.D. Carroll, J.J. Chang, *Psychometrika* **35** (1970), 283–319.
3. R. Coppi, P. Giordani, P. D'Urso, *Psychometrika* **71** (2006), 733–761.
4. P. D'Urso, P. Giordani, *Fuzzy Sets and Systems* **150** (2005), 285–305.
5. P. Giordani, H.A.L. Kiers, *Computational Statistics and Data Analysis* **45** (2004), 519–548.
6. R.A. Harshman, *UCLA Working Papers in Phonetics* **16** (1970), 1–84.
7. H.A.L. Kiers, *Journal of Chemometrics* **18** (2004), 22–36.
8. G.J. Klir, *Uncertainty and information: foundations of generalized information theory*; John Wiley: New York (2006).
9. L.R Tucker, *Psychometrika* **31** (1966), 279–311.

Communication 4

Tensor-structured numerical methods for solving multidimensional equations

B. N. KHOROMSKIJ

Max-Planck-Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany

E-mail address: *bokh@mis.mpg.de*

Numerical methods of tensor-product approximation allow an efficient data-structured representation of the operators and functions in higher dimensions (cf. [1] – [7]). Our main tool is based on the multigrid accelerated mixed Tucker-canonical algebraic decomposition, which is free of the “curse of dimensionality” (linear scaling in the dimension parameter d). We discuss tensor-structured iterative methods for solving the integral/differential equations in \mathbb{R}^d . In particular, we focus on the problems arising in electronic structure calculation (3D Hartree-Fock equation) and in stochastic PDEs. Numerical illustrations will demonstrate the perspectives of tensor methods in various multidimensional applications.

References

1. I. P. Gavriluk, W. Hackbusch and B. N. Khoromskij: Tensor-Product Approximation to Elliptic and Parabolic Solution Operators in Higher Dimensions. *Computing* 74 (2005), 131-157.
2. B.N. Khoromskij: An Introduction to Structured Tensor-Product Representation of Discrete Nonlocal Operators. *Lecture Notes 27, MPI MIS, Leipzig 2005.*
3. W. Hackbusch and B.N. Khoromskij: Low-rank Kronecker product approximation to multi-dimensional nonlocal operators. Parts I/II. *Computing* 76 (2006), 177-202/203-225.
4. B.N. Khoromskij and V. Khoromskaia: Multigrid accelerated tensor approximation of function related multidimensional arrays. Preprint 40/2008, MPI MIS, Leipzig 2008 (*SIAM J. Sci. Comp.*, submitted).
5. W. Hackbusch, B.N. Khoromskij, S. Sauter, and E. Tyrtshnikov: Use of tensor formats in elliptic eigenvalue problems. Preprint 78/2008, MPI MIS Leipzig, 2008 (*SINUM*, submitted).
6. B.N. Khoromskij and Ch. Schwab: Tensor approximation of multi-parametric elliptic problems in stochastic PDEs. *ETH Zuerich, 2008/2009 (in preparation).*
7. B.N. Khoromskij: Tensor-structured Preconditioners and Approximate Inverse of Elliptic Operators in \mathbb{R}^d . Preprint 82/2008, MPI MIS Leipzig, 2008 (*Constructive Approximation*, submitted).

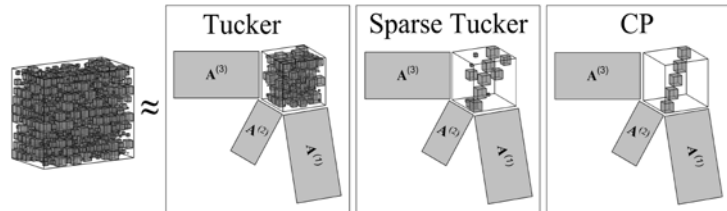
Communication 5

Bayesian and Non-linear Multi-way modelling

M. MØRUP¹, L. KAI HANSEN¹

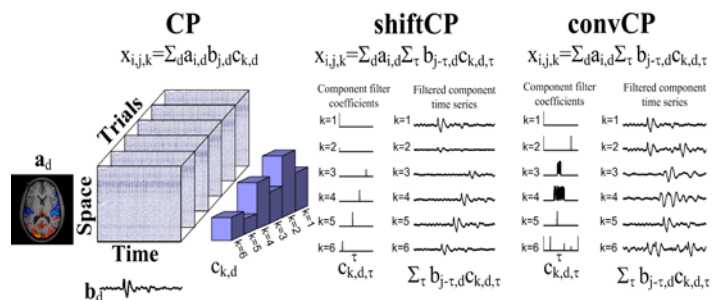
¹ DTU Informatics, Richard Petersens Plads bld. 321/118, Kgs. Lyngby, 2800, Denmark, {mm,lkh}@imm.dtu.dk

Estimating the adequate number of components is an important yet difficult problem in multi-way modeling. We demonstrate how a Bayesian framework for model selection



based on automatic relevance determination (ARD) can be adapted to the Tucker and CandeComp/PARAFAC (CP) models. By assigning priors for the model parameters and learning the hyper-parameters of these priors the method is able to turn off excess components and simplifying the core structure at a computational cost of fitting the conventional Tucker/CP model (i.e., forming the Sparse Tucker representation in the figure above interpolating between a Tucker (full core) and CP (diagonal core) representation).

We further present an algorithm for multi-linear decomposition that allows for arbitrary shifts along one modality forming the shiftCP model. The method is applied to neural activity arranged in the three modalities space, time, and trial. Thus, the algorithm models neural activity as a linear superposition of components with a fixed time course that may vary across



either trials or space in its overall intensity and latency. Its utility is demonstrated on simulated data as well as actual EEG, and fMRI data where the algorithm successfully cope with variable latencies and avoid the CP-degenerate solutions that occur when analyzing the data by instantaneous multi-linear decompositions. We finally extend the shiftCP model to a model that can both accommodate an arbitrary number of component delays as well as shape variability using a convolutional representation forming the convCP model (see figure above). Imposing sparseness on the convolutional filters it becomes possible to interpolate between an arbitrary number of component delays within each trial and the single delay shiftCP model.

All the above approaches can be considered non-linear generalizations of the multi-linear CP and Tucker models.

Acknowledgement

This research was supported by the European Commission through the EU FP6 NEST Pathfinder grant PERCEPT (043261).

References

1. M. Mørup, L. K. Hansen Automatic Relevance Determination for Multi-way Models, Journal of Chemometrics (2009), published online.
2. M. Mørup, L. K. Hansen, S. M. Arnfred, L.-H. Lim, K. H. Shift Invariant Multi-linear Decomposition of Neuroimaging Data, NeuroImage (2008) **42(4)**, 1439-1450.

Communication 6

Selecting the basic parameters of a decomposition in rank-(L,L,1) terms

D. NION¹ AND L. DE LATHAUWER¹

¹ K.U. Leuven, Campus Kortrijk, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
Dimitri.Nion@kuleuven-kortrijk.be , *Lieven.DeLathauwer@kuleuven-kortrijk.be*

The recently introduced decompositions of a higher-order tensor in block terms unify the Tucker and PARAFAC decompositions. In this talk, we focus on one of these decompositions, namely, the decomposition in multilinear rank-(L,L,1) terms. We investigate how L and the number of terms R can be determined. To this end, we present several criteria, among which a block term generalization of the Core Consistency Diagnostic (CORCONDIA). We show that the joint analysis of the proposed criteria allows one to select values of R and L that yield a reasonable model.

Communication 7

Simplicity transformations for three-way arrays with symmetric slices

J.N. TENDEIRO¹, J.M.F. TEN BERGE AND H.A.L. KIERS

¹University of Groningen-Heijmans Institute of Psychological Research, Grote Kruisstraat 2/1, Groningen, 9712 TS, The Netherlands, *j.n.tendeiro@rug.nl*

Tucker three-way PCA and Candecomp/Parafac are two well-known methods of generalizing principal component analysis to three way data. Candecomp/Parafac yields solutions that are typically unique up to jointly permuting and rescaling the components. Tucker-3 analysis, on the other hand, has full transformational freedom. That is, the fit does not change when the component matrices are postmultiplied by nonsingular transformation matrices, provided that the inverse transformations are applied to the so-called core array \mathbf{G} . This freedom of transformation can be used to create a simple structure in the component matrices, and/or in \mathbf{G} . In our research we deal with the latter possibility exclusively. It revolves around the question of how a core array, or, in fact, any three-way array can be transformed to have a maximum number of zero elements. Direct applications are in Tucker-3 analysis, where simplicity of the core may facilitate the interpretation of a Tucker-3 solution, and in constrained Tucker-3 analysis, where hypotheses involving sparse cores are taken into account. So far, a number of simplicity results have been attained, pertaining to arrays sampled randomly from continuous distributions. These results do not apply to three-way arrays with symmetric slices in one direction. We propose a number of simplicity results for arrays with symmetric slices. The issues of typical rank and maximal simplicity of the targets to be presented will be addressed, either by formal proofs or by relying on simulation results.

Communication 8

A classification tool for N-way array based on SIMCA methodology

M. COCCHI¹, C. DURANTE¹ AND R. BRO²

¹University of Modena and Reggio Emilia, Chemistry Department, Via Campi 183, Modena, 41100, Italy, *marina.cocchi@unimore.it*

²University of Copenhagen, Faculty of Life Sciences, Dept. of Food Science, Rolighedsvej 30, DK-1958 Frederiksberg C DK

In the literature there are few papers concerning classification methods suitable for multi-way array [1-5] while the most common procedure is to unfold the three-way data array and then to apply the traditional bi-dimensional tools for classification. In particular, discriminant multi-way partial least squares regression has been used as supervised recognition method [4]. Another approach has been to investigate the classification capabilities of PARAFAC together with Fisher's LDA, applying the last one to the scores obtained by the multi-way decomposition method [3] or SIMCA [6]. A first attempt to use TUCKER3 in conjunction with SIMCA classification has been reported [5] dealing with a particular data set and without trying any generalization. Thus, a true multi-way classification algorithm is still missing.

Aim of this work is to extend the SIMCA method to three and higher order arrays and developing a suitable code. In analogy with the two-way SIMCA, a classification model is separately built for each multi-way data set for each class, using a PCA extension to higher order arrays such as PARAFAC or TUCKER3. The choice of the best dimensionality, i.e. number of latent factors, is chosen according to a cross-validation criterion. In order to estimate the class limits for each class model, both leverage and D and Q statistics, are tested. Using the distributions of these, confidence limits for the two parameters are obtained analogously to the two-way case. Classification performance using different definition of class boundaries and classification rules, including the use of cross-validated residuals, are compared.

The proposed N-SIMCA classification algorithm, besides simulated data sets of varying dimensionality, has been tested on two case studies, concerning food authentication tasks for PDO products.

References

1. M. Dyrby M. Petersen, A.K. Whittaker, L. Lambert, L. Nørgaard, R. Bro, S.B. Engelsen, *Anal. Chim. Acta*, 531 (2005) 209-216.
2. F. Guimet , J. Ferré, R. Boqué. *Anal. Chim. Acta*, 544 (2005) 143-152.
3. Guimet F., J. Ferré, R. Boqué, *Chemom. Intell. Lab. Syst.*, 81 (2006) 94-106
4. D. Ballabio, V. Consonni, R. Todeschini, *Anal. Chim. Acta*, 605 (2007) 134-146.
5. G. R. Flåten, B. Grung, O. M. Kvalheim. *Journal of Chemometrics*, 18 (2004) 173 – 182.
6. G.J. Hall, J.E. Kenny, *Anal. Chim. Acta*, 581 (2007) 118-124.

Communication 9

Adaptive Algorithms to Track the PARAFAC Decomposition of a Third-Order Tensor

D. NION¹, N. D. SIDIROPOULOS²

¹ KU Leuven, Kortrijk Campus, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium, e-mail Dimitri.Nion@kuleuven-kortrijk.be

² Department of ECE, TU Crete, Kounoupidiana, chania – Crete 73100, Greece, e-mail nikos@telecom.tuc.gr

The PARAFAC decomposition of a higher-order tensor is a powerful multilinear algebra tool that becomes more and more popular in a number of disciplines. Existing PARAFAC algorithms are computationally demanding and operate in batch mode - both serious drawbacks for on-line applications. When the data are serially acquired, or the underlying model changes with time, adaptive PARAFAC algorithms that can track the sought decomposition at low complexity would be highly desirable. This is a challenging task that has not been addressed in the literature, and the topic of this paper. Given an estimate of the PARAFAC decomposition of a tensor at instant t , we propose two adaptive algorithms to update the decomposition at instant $t+1$, the new tensor being obtained from the old one after appending a new slice in the 'time' dimension. The proposed algorithms can yield estimation performance that is very close to that obtained via repeated application of state-of-art batch algorithms, at orders of magnitude lower complexity. The effectiveness of the proposed algorithms is illustrated using a MIMO radar application (tracking of directions of arrival and directions of departure) as an example.

Acknowledgement

D. Nion was supported by a post-doctoral grant from the Délégation Générale pour l'Armement (DGA) via ETIS Lab., UMR 8051 (ENSEA, CNRS, Univ. Cergy-Pontoise), France.

References

1. D. Nion and N.D. Sidiropoulos, "Adaptive Algorithms to Track the PARAFAC Decomposition of a Third-Order Tensor," *IEEE Trans. on Signal Processing*, accepted subject to minor revision.

Communication 10

The Candecomp/Parafac decomposition – diverging components and how to avoid them

A. STEGEMAN¹

¹ Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, The Netherlands, a.w.stegeman@rug.nl, www.gmw.rug.nl/~stegeman

Candecomp/Parafac (CP) is the most well-known decomposition of multi-way arrays or tensors. For a given array, CP yields a best rank- R approximation in terms of R rank-1 components, where the number of components R is prespecified. As such, CP can be seen as a multi-way generalization of Principal Component Analysis (PCA) or the (truncated) Singular Value Decomposition (SVD) for matrices. An attractive feature of CP is that it yields rotationally unique components under relatively mild conditions. This is not the case for PCA. A disadvantage of CP is that an optimal CP solution may not exist. That is, unlike a matrix, a multi-way array may not have a best rank- R approximation (reference 5). If this is the case, then a CP algorithm will terminate with diverging components, which is also referred to as “degeneracy” (reference 3). Diverging components can be avoided by imposing additional restrictions in CP, such as orthogonality or nonnegativity. However, they are a significant problem in the practical application of unrestricted CP (references 1, 2, and 4). In this talk, the mathematical background of diverging CP components will be discussed (references 1, 2, and 4). Moreover, for 3-way arrays with two slices, i.e. $p \times q \times 2$ arrays, a method will be presented that avoids the problems of diverging CP components (reference 6). Using this method, we are able to obtain the nondiverging CP components separately. Also, we can obtain a Tucker3 decomposition of the limit point of the diverging CP components.

Acknowledgement

Research supported by the Dutch Organisation for Scientific Research (NWO), VENI grant 451-04-102, and VIDI grant 452-08-001.

References

1. A. Stegeman, *Psychometrika*, **71** (2006), 483-501.
2. A. Stegeman, *Psychometrika*, **72** (2007), 601-619.
3. W.P. Krijnen, T.K. Dijkstra, A. Stegeman, *Psychometrika*, **73** (2008), 431-439.
4. A. Stegeman, *SIAM Journal on Matrix Analysis and Applications*, **30** (2008), 988-1007.
5. V. De Silva, L.-H. Lim, *SIAM Journal on Matrix Analysis and Applications*, **30** (2008), 1084-1127.
6. A. Stegeman, L. De Lathauwer, *SIAM Journal on Matrix Analysis and Applications*, **30** (2009), 1614-1638.

Communication 11

An Optimization Approach for Fitting a CANDECAMP/PARAFAC Model with Applications in Social Network Analysis

E. ACAR¹, T. G. KOLDA¹ AND D. M. DUNLAVY²

¹ Computational Science and Mathematics Science Research Department, Sandia National Laboratories, Livermore, CA 94551-9159
ecarat@sandia.gov, tgkolda@sandia.gov

² Computer Science and Informatics Department, Sandia National Laboratories, Albuquerque, NM 87123-1318, *dmdunla@sandia.gov*

The task of fitting the CANDECAMP/PARAFAC (CP) model to large scale datasets is known to be computationally challenging. The traditional approach for fitting a CP model is based on alternating least squares (ALS) optimization. The ALS approach is generally fast but is not robust to over-factoring (i.e., ALS does not recover the true components when more than that number is specified in the model). Previously, nonlinear least squares (NLS) methods have also been recommended and are robust to over-factoring (i.e., the true components are recovered and the extra components are set to zero); however, these methods do not scale to large problem sizes.

To address these issues, we propose the use of gradient-based optimization (OPT) methods for fitting the CP model. We discuss the mathematical calculation of the derivatives and show that they can be computed efficiently, at the same cost as one iteration of ALS. Numerical experiments are conducted on both simulated and real datasets. We demonstrate that gradient-based optimization methods are more accurate than ALS and much faster than NLS.

To demonstrate scalability, we use the proposed OPT algorithm for the analysis of large social network data. We arrange the time-evolving data as a third-order tensor where time is the third dimension. Our goal is to predict links in future years. We model the tensor using CP. The factors extracted by the CP model are used to forecast future relationships between objects and give promising performance in terms of link prediction.

Acknowledgement

This work was fully supported by Sandia's Laboratory Directed Research & Development (LDRD) program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Communication 12

New computational properties for Hierarchical Principal Component Analysis (HPCA) and its relation to PARAFAC Model

M. HANAFI

Unité de Sensométrie et de Chimiométrie, ENITIAA, BP 82225, 44322 NANTES CEDEX 3, FRANCE

Mohamed.hanafi@enitiaa-nantes.fr

Abstract

Several methods have been proposed in the literature to reveal covariant patterns between and within different multivariate data sets. Among these methods, the present talk focuses on Hierarchical Principal Component analysis (HPCA)^[1]. HPCA has been introduced as a generalisation to several data sets of the well known NIPALS algorithm for Principal Component Analysis.

The computation of the parameters of HPCA (so-called Block scores and global scores) is based on an iterative procedure without much knowledge about its properties^[2]. The main result^[3] presented in this talk discloses that the HPCA iterative procedure increases a criterion. As a consequence the monotony convergence of HPCA is proved. Thus, it turns out that HPCA is tightly related to a method called Common Components and specific weight analysis (CCSWA)^[4,5,6]. In addition clarifications concerning the link between these two methods (HPCA, CCSWA) and the well know PARAFAC model is discussed.

References.

1. Wold S, Kettaneh N, Tjessem K. *J. Chemometrics*, (1996), 463-482
2. Smilde AK, Westerhuis JA, De Jong S. *J. Chemometrics*. (2003), 323-337
3. Hanafi M.. *Computational Statistics*. (2009) (Submitted)
4. Qannari E. M., Wakeling I., Courcoux, Ph., MacFie, M.F. *Food Quality and Preference*. (2000), 151-154.
5. Hanafi, M. Qannari, EM. *Journal de la Société Française de Statistique*.(2008), 75-97
6. Hanafi, M., Mazerolles, G., Dufour, E., Qannari, E. M. *J. Chemometrics*. (2006), 172-183

Communication 13

QR PARAFAC

G. TOMASI

Dept. Basic Science and Environment, Faculty of Life Sciences,
University of Copenhagen, Denmark, *gt@kvl.dk*

Many algorithms have been proposed for the fitting of the PARAFAC model. The Gauss-Newton algorithm in particular appears to be particularly promising for difficult problems. However, it is not applicable in the current versions to relatively large data sets in which the number of estimated parameters is in the order of several thousands.

Compression, which reduces the size of the problem for an N-way array to F^N , where F is the rank, is not entirely satisfactory because it cannot be straightforwardly extended to the bounded or constrained case.

Here a GN algorithm is presented based on iterative QR decompositions which compresses the problem to $F^{(N-1)} \times (I + J + K + \dots)$ and beside being faster than standard GM, is in principle extendable to non-negative decompositions.

Communication 14

A comparison of several second-order calibration algorithms

YONG-JIE YU, HAI-LONG WU, JIN-FANG NIE, SHU-RONG ZHANG, SHU-FANG LI, YUAN-NA LI, SHAO-HUA ZHU and RU-QIN YU

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China. E-mails: *hlwu@hnu.cn* (Hai-Long Wu) and *rqyu@hnu.cn* (Ru-Qin Yu)

A comprehensive and systemic strategy for evaluating the performances of second-order calibration methods is presented in this paper, in particular with a view of practical applications. Second-order calibration methods such as PARAFAC, ATLD, SWATLD and APTLD, which have the so-called “second-order advantage” and are gaining widespread acceptance in the field of chemometrics, were compared. Based on different input parameters including noise level, initial value, number of estimated components and collinearity in simulated and real data, the performances of these methods were evaluated in terms of recovery, consistency of resolved and real profiles, fitness obtained by selected components and speed of convergence. The obtained results give a reevaluation of the position and role of these second-order calibration methods in chemometrics and provide a guidance in practical applications for solving analytical chemistry problems. It is useful and helpful to choose, for example, which algorithm would be more suitable for predicting the concentration of the analyte of interest even in the presence of many unknown interferents in complex systems.

Acknowledgement

The authors would like to acknowledge the National Natural Science Foundation of China (Grant Nos. 20775025 and 20435010) and the National Basic Research Program (No. 2007CB216404) for financial supports.

References

1. G.M. Escandar, N.K.M. Faber, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, R.J. Poppi, *Trends Anal. Chem.* 26 (2007) 752-765.
2. A.C. Olivieri, *Anal. Chem.* 80 (2008) 5713–5720.
3. V. Gómez, M. P. Callao, *Anal. Chim. Acta* 627 (2008) 169-183.
4. A.C. Olivieri, H.L. Wu, R.Q. Yu, *Chemom. Intell. Lab. Syst.* 96 (2009) 246-251.
5. K.S. Booksh, B.R. Kowalski, *Anal. Chem.* 66 (1994) A782-A790.
6. H.L. Wu, M. Shibukawa, K. Oguma, *J. Chemom.* 12 (1998) 1-26.
7. R. Bro, *Chemom. Intell. Lab. Syst.* 38 (1997) 149-171.
8. Z.P. Chen, H.L. Wu, J.H. Jiang, Y. Li, R.Q. Yu, *Chemom. Intell. Lab. Syst.* 52 (2000) 75-86.
9. A.L. Xia, H.L. Wu, D.M. Fang, Y.J. Ding, L.Q. Hu, R.Q. Yu, *J. Chemom.* 19 (2005) 65-76.
10. N.K.M. Faber, R. Bro, P.K. Hopke, *Chemom. Intell. Lab. Syst.* 65 (2003) 119-137.
11. G. Tomasi, R. Bro, *Comput. Stat. Data Anal.* 50 (2006) 1700-1734.
12. T.G. Kolda, B.W. Bader, *Tensor decompositions and applications*, *SIAM Review*. To appear (accepted June 2008).

Communication 15

3-way methods: a practitioner perspective

A. FERRER¹ AND J.M. PRATS-MONTALBÁN²

Universidad Politécnica de Valencia; Dep. of Applied Statistics, Operations Research and Quality; Multivariate Statistical Engineering Group³; Camino de Vera s/n, Edificio 7A, 46022, Valencia (Spain),

¹ *aferrer@eio.upv.es*

² *jopramon@eio.upv.es*

³ *mseg.webs.upv.es*

In this talk we discuss on the potentials and limitations of 3-way methods from a practical perspective. Several examples from different fields (environmental analysis, bioinformatics and multivariate process monitoring) are going to be used to illustrate the issues under discussion.

A descriptive environmental analysis is presented, showing the benefits of Tucker3 vs PCA models for interpretation aims. A second example deals with bioinformatics data analysed by using Tucker3 as well as N-PLS when input-output dataset structure is considered. Finally, the multivariate monitoring scheme of a wastewater batch process is introduced in the third example. The straightforward analysis of the 3-way data structure via a Tucker3 model is compared to bilinear approaches using unfold-PCA. Diagnostic issues by using contribution plots are discussed in order to compare the performance of the different models.

Acknowledgements

This research was supported by the Spanish Government (MICINN) and the European Union (RDE funds) under grant DPI2008-06880-C03-03.

References

1. J.C. García-Díaz, J.M. Prats-Montalbán, *Chemometrics and Intelligent Laboratory Systems* **76** (2005) 15-24.
2. R. Henrion, *Chemometrics and Intelligent Laboratory Systems* **25** (1994) 1-23.
3. A. Smilde, R. Bro, P. Geladi, *Multi-way Analysis, Application in the Chemical Sciences*, John Wiley & Sons: England (2003)

Communication 16

A criterion based PLS approach to structural equation modelling

M. TENENHAUS

HEC PARIS (GREGHEC), 1 rue de la Libération, Jouy-en-Josas, France, tenenhaus@hec.fr

Abstract

For more than two blocks, the properties of the Wold's PLS algorithm for SEM have long remained unknown. We propose in this paper a more general approach based on a continuum going from a new mode A to the usual mode B. When a structural equation model based on J blocks X_1, \dots, X_J of manifest variables measured on the same set of individuals is considered, we propose the following optimization problem:

Maximize the criterion $\sum_{i=1, j=1, i \neq j}^J c_{ij} g(\text{Cov}(X_i a_i, X_j a_j))$ with respect to the weights a_1, \dots, a_J ,
subject to the constraints $\tau_i \|a_i\|^2 + (1 - \tau_i) \text{Var}(X_i a_i) = 1, i = 1, \dots, J$.

where:

- The function g can be the identity (*Horst scheme*), the absolute value (*centroid scheme*) or the squared value (*factorial scheme*),
- $c_{ij} = 1$ for two connected blocks, and 0 otherwise,
- $0 \leq \tau_i \leq 1$.

For $\tau_i = 0$, mode B is found again, for $\tau_i = 1$, a new mode A is defined. This new mode A yields exactly PLS regression for a two blocks situation.

Pragmatic solutions to these maximization problems are found by using the following procedure:

- (1) Construct the Lagrangian function related to the maximization problem.
- (2) Define stationary equations by cancelling the derivatives of the Lagrangian function.
- (3) Find a solution of the stationary equations by using an iterative procedure.

The convergence of the algorithm is proven when the Wold's iterative procedure is used. That means that the bounded criterion to be maximized is increasing at each step of the procedure. When g is the absolute or squared value and $\tau_i = 0$ for all i , PLS mode B + centroid or factorial scheme is found again. Our approach generalizes results of Hanafi (2007) on PLS mode B and of Krämer (2007) on PLS mode A.

References

1. Hanafi M. (2007): PLS Path modelling: computation of latent variables with the estimation mode B, *Computational Statistics*, 22, 275-292.
2. Krämer N. (2007): Analysis of high-dimensional data with partial least squares and boosting. Doctoral dissertation, Technischen Universität Berlin.

Communication 17

Multilevel Multiway Analysis

A. K. SMILDE

Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV, Amsterdam, The Netherlands
A.K.Smilde@uva.nl

In systems biology, data are becoming more and more complex. For answering biological questions usually multiple data sets have to be analyzed simultaneously. Moreover, the data can have extra structure such as multiple levels, time-profiles and/or underlying experimental designs. The presentation will start with a short overview of multiway problems in systems biology. Then the focus will be on multilevel multiway problems. Methods to tackle these problems will be discussed and these will be illustrated with examples from the field of human metabolomics and biotechnology.

Communication 18

A unifying framework for simultaneous component methods

K. VAN DEUN¹, I. VAN MECHELEN¹, A. SMILDE², H. KIERS³, AND M. VAN DER WERF⁴

¹ Department of Psychology, University of Leuven, Tiensestraat 102, Leuven, 3000, Belgium, *katrijn.vandeun@psy.kuleuven.be*

² Biosystems data analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam, 1018 WV, The Netherlands

³Heymans Institute, University of Groningen, Groningen, Grote Kruisstraat 2/1, 9712 TS, The Netherlands

⁴TNO, Quality of Life, Utrechtseweg 48, 3704 HE, The Netherlands

Nowadays frequently different pieces of information are gathered on the same set of entities or variables with the different pieces stemming, for example, from different conditions or measurement techniques. This implies that more and more data appear that consist of two or more data arrays that have a shared mode. A broad range of methods can be used to analyze such data, an important class of them originating from the component analysis domain, called simultaneous component methods (e.g., SUM-PCA, unrestricted PCovR, MFA, STATIS, and SCA-P). Yet, different simultaneous component methods may lead to quite different results. Moreover, the methods are not easy to compare as they stem from different research domains in which different terminologies are being used. In this presentation we offer a general framework that encompasses all simultaneous component methods and that highlights both the common core of the methods and the specific elements with regard to which they differ. An overview of principles is given that may guide the data analyst in choosing an appropriate simultaneous components method. Several theoretical and practical issues are illustrated with an empirical example on metabolomics data as obtained with different separation methods.

Acknowledgement

This research was supported by the Research Council of KU Leuven (EF/05/007 and grant GOA/2005/04) and the Flemish Government, IWT (SBO-BioFrame).

Communication 19

A generic model for data fusion

I. VAN MECHELEN

Department of Psychology, University of Leuven, Tiensestraat 102, Leuven, 3000, Belgium, *iven.vanmechelen@psy.kuleuven.ac.be*

Abstract:

In many research contexts, data show up that take the form of multiple linked data blocks. As an example, one may think of several batches of information with regard to a same set of entities as stemming from different sources of information. Data sets that consist of multiple linked data blocks imply novel challenges for the data analyst. More in particular, the data analyst may wish to go for a simultaneous modeling of the different linked blocks, and this for several possible reasons, including: (a) to arrive at more reliable inferences, (b) to grasp in an effective way common as well as distinct pieces of structural information as included in the different data blocks, and (c) to get a better understanding of the linkage relations between the different data blocks.

In this paper, we will outline a conceptual framework for the simultaneous modeling of linked data blocks. Subsequently, we will introduce a generic model for this problem, which subsumes a broad range of specific models (existing as well as to be developed) as special cases.

Communication 20

CLASSI modeling of sequential processes and individual differences therein

E. CEULEMANS

Katholieke Universiteit Leuven, Centre for Methodology of Educational Research,
Andreas Vesaliusstraat 2 box 3762, B-3000 Leuven, Belgium,
eva.ceulemans@ped.kuleuven.be

Abstract:

In psychological research, one often aims at explaining individual differences in S-R profiles, that is, individual differences in the responses (R) with which people react to specific stimuli (S). To this end, researchers often postulate an underlying sequential process, which boils down to the specification of a set of mediating variables (M) and the processes that link these mediating variables to the stimuli and responses under study. Obviously, a crucial task is to chart how the individual differences in the S-R profiles are caused by individual differences in the S-M link and/or by individual differences in the M-R link. Ceulemans & Van Mechelen (2008) proposed a new model, called CLASSI, which was explicitly designed for this task. In particular, the key principle of CLASSI consists of reducing the S, M, and R nodes of a sequential process to a few mutually exclusive types and inducing an S-M and an M-R person typology from the data, with the S-M person types being characterized in terms of if S type then M type rules and the M-R person types in terms of if M type then R type rules. As such, the S-M and M-R person types and their associated if-then rules represent the important individual differences in the S-M and M-R links of the sequential process under study. The CLASSI model suffers from two important restrictions, however: Firstly, it can only handle binary data. Secondly, the CLASSI model requires the persons and stimuli to be fully crossed, implying that each person has to rate the same set of stimuli. In many research domains, this is a major restriction since not all stimuli are equally relevant for every person. Therefore, in this paper, we discuss the extension of the CLASSI model to real-valued and nested data, the latter implying that the set of rated stimuli differs across persons.

Communication 21

Multilinear modeling of brain imaging data

S. HONG

University of Illinois, 603 E Daniel St, Urbana-Champaign, US, hongsj@uiuc.edu

Independent component analysis has been frequently used for decomposition of brain imaging data. Beckmann recently proposed tensorial probabilistic ICA (tensorial PICA) for constrained Parafac analysis of brain imaging data, which combines the fast ICA decomposition with a further decomposition of the component weights into two other modes. In tensorial PICA, the components are constrained to be statistically independent in the spatial mode. However, given the uniqueness property of the Parafac model, tensorial PICA does not produce independent components as presumed. Instead, it may be considered as an alternative algorithm to fit the Parafac model. Some results will be shown to compare the fitting performance of tensorial PICA and the ALS Parafac. In addition, a case will be shown where the Parafac2 model is useful for decomposition of seemingly incomparable brain images arising from some typical experimental designs. A practical challenge in component analysis of brain imaging data is excessive size of the data. A typical three-mode dataset (e.g., voxels \times time points \times subjects) could easily become too large to apply any component analysis on a personal computer. A speedup procedure will be discussed.

Communication 22

Parallel Factor Analysis of Gait Data

N. E. HELWIG AND S. HONG

University of Illinois at Urbana-Champaign, 603 E Daniel St, US, nhelwig2@illinois.edu

Gait analysis refers to the systematic study of human (or animal) walking behavior. One method in gait analysis is to study the 2D or 3D shape patterns made by the joint centers in the lower limbs (e.g., ankle or knee) during ambulation. Another possibility is the analysis of various 1D data trajectories (e.g., the angle between joint centers, the acceleration at a joint center, etc.) extracted from the kinematic data. This study will show how the Parafac model can reveal effects of mobility constraints (e.g., injuries) on the lower limbs via differently formulated data of the same source. First, normal (i.e., healthy) 2D joint center shape patterns are aligned by Generalized Procrustes Analysis (GPA), and then deviations from the GPA consensus for both normal and mobility-constrained data are simultaneously decomposed by Parafac. The second approach applies a warping technique to temporally align 1D joint angle trajectories. Then, the resulting temporal and intensity deviations from a normal trajectory (for both the normal and mobility-constrained data) are simultaneously modeled by Parafac. Results from these alternatively formulated data on essentially the same information will be compared, and possible diagnostic implications will be discussed along with the Parafac results.

Application of Multi-Dimensional Decomposition in BioMolecular NMR Spectroscopy

V. OREKHOV^{1,2}, I. IBRAGHIMOV³, S. HILLER², G. WAGNER², V. JARAVINE¹, A. ZHURAVLEVA¹, P. PERMI⁴, L. SAMUELSSON⁵, J. LARSSON⁵

¹Swedish NMR Centre at University of Gothenburg, Gothenburg, Sweden;

²Department Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, USA;

³Elegant Mathematics Ltd, Ottweiler, Germany;

⁴Inst. Biotechnology, Helsinki, Finland;

⁵Inst. Neurosciences & Physiol. University of Gothenburg, Gothenburg, Sweden

Multi-Dimensional Decomposition (MDD) is a powerful signal-processing tool in Nuclear Magnetic Resonance (NMR) spectroscopy of large biological molecules (10^3 - 10^5 atoms), such as proteins and nuclear acids. Contemporary NMR routinely deals with large data sets of three and more dimension. Raw signal, which is often collected in sparse mode, comes from spectrometer in the time domain as a superposition of multidimensional sine/cosine waves. Spectrum in the frequency domain is typically obtained by Fourier transform.

Multidimensional spectrum of a protein is mostly empty with several hundreds, resolved peaks sticking out from flat noise baseline. Moreover, NMR theory tells that with rare exceptions the signals are completely defined by one-dimensional line shapes along all spectral dimensions. These properties make MDD and related methods very useful for the analysis and allow using of sparse (non-uniform) sampling for optimizing spectra resolution and sensitivity. We apply MDD for processing of individual spectra and co-processing of combination of experiments [1-4]. The approach has been successfully demonstrated for several essentially different situations: processing of four-dimensional NOESY spectra for the de novo structure determination of the integral human membrane protein VDAC-1 in detergent micelles with effective molecular weight of 70–90 kDa; rapid real-time data collection of 9-dimensional hyper-spectrum for automated backbone assignments of 13 kDa naturally disordered cytoplasmic part of the T-cell receptor; and analysis of 50 two-dimensional spectra from a metabolomics study.

References

1. V. Tugarinov, W.Y. Choy, V.Y. Orekhov, and L.E. Kay, Proc. Natl. Acad. Sci. USA. **102** (2005) p. 622-627.
2. V. Jaravine, I. Ibraghimov, and V.Y. Orekhov, Nature Methods **3** (2006) p. 605-607.
3. S. Hiller, R.G. Garces, T.J. Malia, V.Y. Orekhov, M. Colombini, and G. Wagner, Science **321** (2008) p. 1206-1210.
4. V. Jaravine, A. Zhuravleva, P. Permi, I. Ibraghimov, and V.Y. Orekhov, J. Am. Chem. Soc. **130** (2008) p. 3927-3936.

Communication 24

Three-Way Factor Analysis of Large-Microscopic Hyperspectral Images: Compression and Analysis of Very Large, Small Images

M. H. VAN BENTHEM¹ AND M. R. KEENAN²

¹ Sandia National Laboratories Albuquerque, NM 87185-0886

² 8346 Roney Rd., Wolcott, NY 14590

Hyperspectral imaging microscopy is an extremely valuable tool for chemical and biological research. (Colarusso et al., 1998; Schultz et al., 2001) In this technique one collects a full spectrum, which can represent, inter alia, energy or wavelength (as in x-ray or visible or IR spectroscopy) or mass (as in mass spectroscopy), for each pixel in an image. Under the proper experimental conditions, one can even generate three-way and higher hyperspectral images. Since hyperspectral imaging is a data-rich technique, it produces enormous quantities of data for a single chemical or biological sample. (Kotula et al., 2003) It is this “richness” of data that gives such power to the technique and facilitates the understanding of the material or phenomenon under scrutiny. Unfortunately, these datasets can be quite sizable and unwieldy, comprising hundreds of megabytes in raw form. (Van Benthem and Keenan, 2008) Multivariate analysis, three-way decomposition in particular, permits one to reduce the dimensionality of the data to reveal the underlying information: the species giving rise to the spectra, their behavior under varying experimental conditions, and where they exist and relative abundances in the image. Unfortunately, the size of these data typically requires special handling. For example, Sandia National Laboratories’ hyperspectral imaging microscope can produce images of biological samples with dimensions of 200 × 200 image pixels by 512 wavelength elements by 18 photobleaching increments. Given the extent of these data, careful processing and efficient analysis algorithms are paramount. Accordingly, we have implemented novel fast algorithms and compression techniques that can quickly perform factor analysis. In this presentation, we will describe the efficient methods we have used to process and analyze these very large data sets.

Acknowledgement

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

- Colarusso, P., et al., 1998. Infrared spectroscopic imaging: From planetary to cellular systems. *Applied Spectroscopy*. 52, 106A-120A.
- Kotula, P. G., et al., 2003. Automated analysis of SEM X-ray spectral images: a powerful new microanalysis tool. *Microscopy and Microanalysis*. 9, 1-17.
- Schultz, R. A., et al., 2001. Hyperspectral imaging: A novel approach for microscopic analysis. *Cytometry*. 43, 239-247.
- Van Benthem, M. H., Keenan, M. R., 2008. Tucker1 model algorithms for fast solutions to large PARAFAC problems. *Journal of Chemometrics*. 22, 345-354.

Communication 25

Analysis of Multi-way 2D-Liquid Chromatography Diode Array Data for Metabolomic Studies

H. P. BAILEY¹, R. ALLEN¹, S. C. RUTAN¹ AND P. W. CARR²

¹Virginia Commonwealth University, Box 842006, Department of Chemistry, Richmond, VA, 23284-2006, USA, *srutan@vcu.edu*

²University of Minnesota, 207 Pleasant St. SE, Minneapolis, MN, 55455-0431, USA

Advances in analytical instrumentation in recent years have resulted in techniques that easily allow for the acquisition of multiple order data. In particular, two-dimensional liquid chromatography (2D-LC) coupled with diode array detection (DAD), when used to characterize multiple samples, gives rise to four-way data sets comprised of absorbance measurements as a function of 1st dimension elution time, 2nd dimension elution time, wavelength, and sample number.¹ While these data sets should in theory result in quadrilinear data, in practice, experimental errors such as retention time shifts lead to significant deviations in quadrilinearity. Our work focuses on the development of pretreatment methods that improve the quadrilinear nature of the data and of the application of data analysis methods such as PARAFAC, iterative key set factor analysis and multivariate curve resolution-alternating least squares (MCR-ALS) for the analysis of 2D-LC DAD data. The primary pretreatment issue is the alignment of the chromatograms in the first and second dimensions. While methods have been reported in the literature for 2D gas chromatography data,² the more highly overlapped nature of LC signals has prevented the direct extension of these methods to 2D-LC DAD data. Utilization of various interpolation methods such as cubic polynomial splines, especially for the 1st dimension chromatograms which, by necessity, are undersampled, is a key component of the alignment approaches developed in this work. For the analysis of complex data sets, we have found that MCR-ALS methods which have relaxed multilinearity requirements, to be the most useful of the chemometric tools currently at our disposal.³ These studies will provide important advances for metabolomic analyses, where changes in the concentrations of small molecule metabolites in response to a biological change can be used as biomarkers for disease.

Acknowledgement

This work was funded by NIH Grant GM 054585.

References

1. S.E.G. Porter, D.R. Stoll, S.C. Rutan, P.W. Carr, J.D. Cohen, *Anal. Chem.* **78** (2006) 5559–5569.
2. K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, *Anal. Chem.* **77** (2005) 7735–7743.
3. E. Bezemer, S.C. Rutan, *Chemom. Intell. Lab. Syst.* **81** (2006) 82–93.

Communication 26

Two and three way methods for the resolution of overlapping peaks in the determination of polyphenols in olive oil by HPLC-DAD

F. MARINI, A. D'ALOISE, R. BUCCI, A. D. MAGRI, A. L. MAGRI

Dept. Chemistry, Sapienza University of Rome, I-00185, Rome, Italy
fmmonet@hotmail.com

Polyphenols are important components of virgin olive oils, due to their antioxidant properties. Indeed, their beneficial health effects are known since many years and many studies report their role in the prevention of various diseases. Among the various analytical methods proposed for their determination, HPLC play a dominant role, both coupled to UV-spectrophotometric and MS detection of the eluted component. However, the optimal time required for the full separation of all the components can be quite long. In this study, we investigated the possibility of reducing the analytical time of HPLC-DAD analysis of polyphenols by allowing the overlapping and coelution of some peaks, whose signals were separated off-column by two- and three-way resolution methods. In particular, we focused our attention on a peak where four phenolic acids (syringic acid, vanillic acid, p-hydroxybenzoic acid, and caffeic acid) coeluted. The results of applying PARAFAC, PARAFAC2, N-PLS and MCR-ALS on this data set will be discussed and compared.

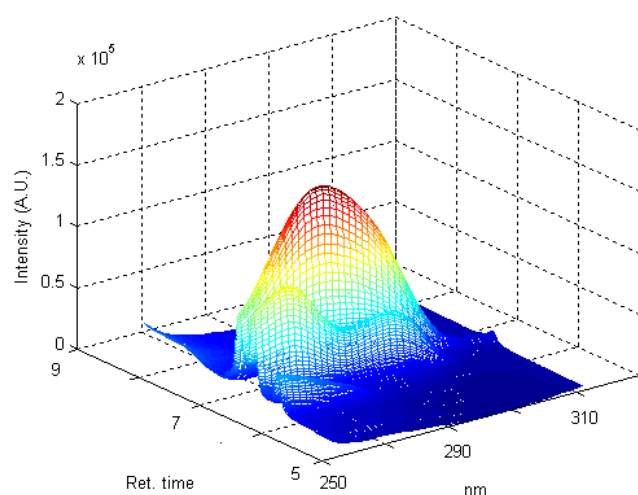


Figure 1 – A typical 3D HPLC-DAD landscape, resulting from the analysis of the 4 coeluting peaks.

1. R.W. Owen, A. Giocosa, *Lancet Oncology*, **1** (2000) 102-112.
2. A. Smilde, R. Bro, P. Geladi, *Multi-way analysis: Applications in the chemical sciences*, Wiley, New York, 2004.
3. R. Tauler, A. De Juan, *Crit. Rev. Anal. Chem.*, **36** (2006) 163-176.

Communication 27

Improvements on GC-MS aroma profile of IIDRØD PIGEON (*Malus domestica*) apples exposed to different length of ripening time by using PARAFAC2

J. M. AMIGO^A, M. J. POPIELARZ^A, R. M. CALLEJÓN^B, M. L. MORALES^B, A. M. TRONCOSO^B, T. B. TOLDAM-ANDERSEN^C, M. A. PETERSEN^A

a Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark, b Departamento Bioquímica, Bromatología, Toxicología y Medicina Legal. Facultad de Farmacia, Universidad de Sevilla. C/ P. Garcia González, 2. 41012 Sevilla, Spain, c Department of Agriculture and Ecology/Crop Science, Faculty of Life Sciences, University of Copenhagen, Højbakkegård Allé 21, 2630 Taastrup, Denmark
e-mail: jmar@life.ku.dk

The establishment of an appropriate length of room temperature exposure (ripening time) after cold storage to achieve full aroma profile of IP apples is still a great importance issue.

This variety of apples is enormously demanded because of their qualities in traditional-Christmas cooking in Denmark. Therefore, the product has to be sold in the markets in the proper time with the proper quality.

One of the alternatives to study the aroma profile behavior is to storage a representative amount of apples at room temperature and, afterwards to analyze each sample by Gas Chromatography – Mass Spectrometry (GC-MS) according to the method of Petersen et al. (1).

The most straightforward method to analyze the obtained GC-MS dataset is to integrate the major peaks (the ones that can be easily identified by their MS profile) and to perform a Principal Component Analysis (PCA). This alternative has several drawbacks: baseline drifts are scarcely considered, the need of an internal standard and, what is most important, the integration boundaries are not always well defined (long tails, overlapped peaks, etc.)

To improve this working methodology, this work proposes the modeling of the raw dataset by using PARAFAC2 algorithm in selected areas of the GC profile and using the obtained well-resolved profiles to develop a further PCA model. With this working method, not only the problems arising from instrumental artifacts are overcome, but the classification of the different ripening times and detection of new analytes is achieved.

References

1. Petersen, M.A., Poll, L., Toldam-Andersen, T.B. (2007). In: Recent highlights in flavor chemistry & biology - Proceedings of the 8th Wartburg Symposium, (Hofmann, T., Meyerhof, W., Schieberle, P., eds.); Deutsche Forschungsanstalt für Lebensmittelchemie, pp. 345-34

Three-way Principal Component Analysis applied to Noodles Sensory Data Analysis

R.LEARDI¹ AND C.B.Y.CORDELLA²

¹ Department of Pharmaceutical and Food Chemistry and Technology, via Brigata Salerno ponte), 16147 Genova, Italy, *riclea@dictfa.unige.it*

² Institut National de Recherche Agronomique, UMR214 Ingénierie Analytique pour la Qualité des Aliments, INRA/INA-PG, 16 rue Claude Bernard, 75005 Paris, France, *Christophe.Cordella@paris.inra.fr*

The results presented in this paper are issued from the study and the interpretation of a 3-way data matrix built on sensory data analysis of wheat noodles.

The aim of this work was to provide a better understanding of internal relationships existing between the chemical composition of the studied noodles and their specific sensory attributes such as colour, surface smoothness, elasticity or chewiness.

The application of the Tucker3 algorithm (the three modes being the noodles, the sensory attributes and the assessors) allows the detection and the interpretation of the differences among the types of noodles, together with the estimation of the effect of different sources of variability on the sensory evaluation.

A joint interpretation of the first and of the second mode (noodles and sensory attributes) allows to link appearance and texture attributes with the composition of the noodles.