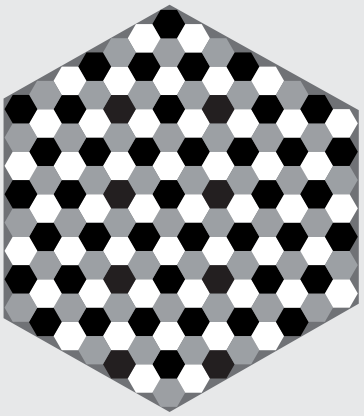


TRICAP - 2015

31 May - 5 June 2015, Pecol di Zoldo (Belluno) - Italy

PROGRAM AND ABSTRACTS

Program

Sunday 31st May

18.00 *Welcome aperitif*
20.00 *Dinner*

Monday 1st June

8.45 - 9.00 **Marina Cocchi – Federico Marini opening**

Chairman **Eva Ceulemans**
9.00 - 10.00 **Arthur Tenenhaus**
Regularized Generalized Canonical Correlation analysis for Multiway data

10.00 - 11.00 **Katrijn Van Deun**
Sparse Common and Distinctive Components

11.00 -11.20 *Coffe Break*

Chairman **Beata Walczak**
11.20 - 12.20 **Romà Tauler**
Multilinear constraints in Multivariate Curve Resolution of multiway and multiset data

12.20 - 13.20 **Mohamed Hanafi**
An algebraic framework for multiblock data analysis.

13.20 - 14.30 *Lunch*

Chairman **Pieter Kroonenberg**
14.30 - 15.30 **Mark Van Benthem**
The Amazing Flexibility of Fluorescence Spectroscopy for Trilinear Modeling

15.30 - 16.30 **Sugnet Lubbe**
Visualising three-way arrays: An outerproduct rank perspective

16.30 - 16.50 *Coffe Break*

16.50 - 17.50 **Tom Wilderjans**
Clusterwise three-way models to account for heterogeneity in three-way data

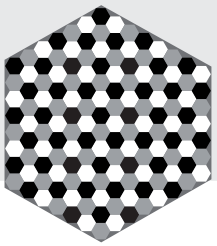
20.15 *Dinner*
after Dinner with live music by Fede

Tuesday 2nd June

Chairman **Rasmus Bro**
9.15 - 10.15 **Paolo Giordani**
An SVD penalized approach to Candecomp/Parafac estimation

10.15 - 11.15 **Nikos Sidiropoulos**
Two takes on constrained CP decomposition using the alternating direction method of multipliers

11.15 - 11.35 *Coffee Break*



11.35 - 12.35 **Laurent Albera - Xavier Luciani**
Joint Eigen-Value Decomposition (JEVD) of matrices, Canonical Polyadic (CP) and non-linear decompositions of tensors: algorithms and application to Independent Component Analysis (ICA) and fluorescence spectroscopy.

12.35 - 13.35 *Lunch*

Chairman **Nikos Siridopolous**
13.35 - 14.35 **Andre De Almeida**
Distributed large-scale tensor decompositions in collaborative networks.

14.35 - 15.35 **Rasmus Bro**
Speeding up and automating multi-way models

15.35 -15.55 *Coffee Break*

15.55 - 16.55 **Casper Albers**
A contribution to the visualisation of three-way arrays

17.15 - 18.30 Meeting of TRICAP Committee

19.30 *Dinner*

21.00 *Traditional "Ladina" music*

Wednesday 3rd June

Chairman **Romà Tauler**
9.00 - 10.00 **Sarah Rutan**
Constraints in Multiway and Multiset Analysis - Applications in Chromatography

10.00 - 11.00 **Martin Haardt**
A semi-algebraic framework for approximate CP decompositions via Simultaneous Matrix Diagonalizations (SECSI)

11.00 - 11.20 *Coffee Break*

11.20 - 12.20 **Orly Alter**
Discovery of Principles of Nature from Matrix and Tensor Modeling of Large-Scale Molecular Biological Data

12.20 - 13.20 *quick Lunch*

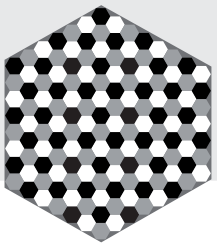
13.20 - on *Hiking Tour & Social Dinner at Rifugio Passo Giau*

Thursday 4th June

Chairman **Paolo Giordani**
9.15 - 10.15 **Nico Vervliet**
A randomized block sampling approach for the CP decomposition of large-scale tensors

10.15 - 11.15 **Alberto Ferrer - Raffaele Vitale**
A novel proposal for the identification of common and distinctive sources of variability in multi-set data

11.15 - 11.35 *Coffee Break*



11.35 - 12.35 **Pieter Kroonenberg**
Assessing factorial invariance of two-way rating designs using three-way methods

12.35 - 13.35 *Lunch*

Chairman **Marieke Timmermann**
13.40 - 14.40 **José Montalbán**
Variable selection in N-PLS

14.40 - 15.40 **Otto De Bals**
About tensorization for blind signal separation: concept and Illustrations

15.40 -16.00 *Coffee Break*

16.00 - 17.00 **Age Smilde**
Separating common and distinct variation in multi-set data: theory and some examples

17.00 - 18.00 Scientific "Jam" Session
by **Henk Kiers**
Chairmans **Marina Cocchi Federico Marini**

20.15 *Dinner*
after Dinner with live music by Fede

Friday 5th June

Chairman **Age Smilde**
9.00 -10.00 **Marieke E. Timmerman**
Sparse modelling of Multiset data

10.00 - 11.00 **Eva Ceulemans**
Two-mode K-Spectral Centroid analysis for studying multivariate dynamical processes

11.00 - 12.00 Last but not least: overall discussion, remarks, future hints

12.15 *Lunch or pocket lunch depending on departure time*

Scientific Committee

Rasmus Bro (University of Copenhagen)
Eva Ceulemans (University of Leuven)
Marina Cocchi (University of modena and Reggio Emilia)
Paolo Giordani (University of Rome "Sapienza")
Pieter Kroonenberg (Leiden University)
Lieven de Lathauwer (University of Leuven)
Federico Marini (University of Rome "Sapienza")
Iven Van Mecheln (University of Leuven)
Bill Rayens (University of Kentucky)
Nikos Sidiropoulos (University of Minnesota)
Age Smilde (University of Amsterdam)
Romà Tauler (IDAEA-CSIC)
Marieke Timmerman (University of Groningen)
Barry M. Wise (Eigenvector Research)

Organising Committee

Marina Cocchi (University of Modena and Reggio Emilia)
Federico Marini (University of Rome "Sapienza")

Joint Eigen-Value Decomposition (JEVD) of matrices, Canonical Polyadic (CP) and non-linear decompositions of tensors: algorithms and application to Independent Component Analysis (ICA) and fluorescence spectroscopy

Laurent Albera - Xavier Luciani

Université de Rennes 1, Rennes, France

laurent.albera@univ-rennes1.fr

The first part of this talk will address the CP decomposition of symmetric even higher order arrays. Numerical methods based on orthogonal and non-orthogonal JEVD of an appropriate set of matrices will be presented by considering the toy example of fourth order cumulant arrays. In addition, a fast and efficient Jacobi-like approach, named JET (Joint Eigenvalue decomposition based on Triangular matrices), will be described for the non-orthogonal JEVD of a set of non-defective matrices based on the LU factorization of the matrix of eigenvectors. In addition, we will illustrate in the ICA context the interest of being able to compute efficiently a symmetric CP decomposition using the JET technique. More particularly, we will show that, when the noise covariance matrix is unknown and the source kurtoses have different signs, solving the non-orthogonal JEVD problem can lead to a more robust ICA method.
The second part will address the CP decomposition of any multi-way arrays by i) generalizing the approach presented in the first part of the talk to the case of unsymmetric arrays and ii) leading to a novel direct solution, named DIAG (DIrect AlGorithm for canonical polyadic decomposition). At this occasion, another JEVD algorithm called JDTM will be briefly described. We will then focus on CPD of fluorescence tensors built from Fluorescence Excitation-Emission Matrices (FEEM). We will notably show that direct algorithms such as DIAG can help to solve the over-factoring problem that occurs with traditional iterative algorithms when the "chemical rank" of the fluorescence tensor, i.e. the number of fluorophores, is unknown. Eventually, the end of the talk will be dedicated to the non-linear model of fluorescence taking into account inner filter effects. This model can be seen as an original non-linear tensor decomposition called NLFD (Non Linear Fluorecence Decomposition) and several decomposition methods will be (very) briefly presented.

More details can be founded in the following references:

[1] L. ALBERA, A. FERREOL, P. COMON and P. CHEVALIER, "Blind identification of Overcomplete MixturEs of sources (BIOME)," in Elsevier Linear Algebra and its Applications, vol. 391C, pp. 3-30, November 2004.

[2] L. ALBERA, A. FERREOL, P. CHEVALIER and P. COMON, "ICAR, a tool for blind source separation using fourth order statistics only", in IEEE Transactions On Signal Processing, vol. 53, no. 10, pp. 3633-3643, October 2005.

[3] X. LUCIANI and L. ALBERA, "Semi-algebraic canonical decomposition of multi-way arrays and joint eigenvalue decomposition,"in ICASSP 11, 2011 IEEE International Conference on Acoustics Speech and Signal Processing, Prague, Czech Republic, May 22-27 2011, pp. 4104-4107.

[4] X. LUCIANI and L. ALBERA, "Joint eigenvalue decomposition of non-defective matrices based on the LU factorization with application to ICA", submitted to IEEE Transactions On Signal Processing.

[5] X. LUCIANI and L. ALBERA, "Canonical Polyadic Decomposition based on joint eigenvalue decomposition," in Elsevier Chemometrics and Intelligent Laboratory Systems, vol. 132, pp. 152-167, March 2014.

[6] X. LUCIANI, S. MOUNIER, R. REDON and A. BOIS, "A simple correction method of inner filter effects affecting FEEM and its application to the PARAFAC decomposition", Chemometrics and Intelligent Laboratory Systems, Volume 96, Issue 2, 15 April 2009, pp. 227-238.

[7] X. LUCIANI, S. MOUNIER and R. REDON, "How to correct inner filter effects altering 3D fluorescence spectra by using a mirrored cell", Chemometrics and Intelligent Laboratory Systems, Volume 126, July 2013, pp. 91-99

[8] J. COHEN and P. COMON, "On almost sure identifiability of non multilinear tensor decomposition", in EUSIPCO'13, 2013 Proceedings of the 22nd European Signal Processing Conference, pp. 2245-2249.

A contribution to the visualisation of three-way arrays

Casper Albers[1], **John Gower**[2]

[1] University of Groningen, Grote Kruisstraat 2/1 , 9712 TS Groningen
[2] Department of Mathematics and Statistics , The Open University (UK)

c.j.albers@rug.nl

My presentation is concerned with the visualisation of three-way arrays, and the visualisation of interactions in three-way arrays in particular. I'll review and extend some standard ways of visualising biadditive modelling. Three-way interaction is modelled by the Individual Scaling method as applied to interaction arrays that have main effects and biadditive terms removed. I will discuss methods to visualise rank-two and rank-three interactions accurately in three dimensions, using polyhedra, and in two dimensions. The three-dimensional visualisations are appealing from a geometric point of view, yet are not adequate for practical purposes. The two-dimensional visualisation is based on what we term the triplot, an extension of the biplot. For larger arrays, interactive visualisation and/or clutter-removing methods are required in order to achieve visualisation with practical applicability.

Discovery of Principles of Nature from Matrix and Tensor Modeling of Large-Scale Molecular Biological Data

Orly Alter

University of Utah (USA)

orly@sci.utah.edu

I will describe the use of matrix and tensor decompositions in the simultaneous modeling of different types of large-scale molecular biological data, from different studies of cell division and cancer [1] and from different organisms, to computationally predict previously unknown physical, cellular and evolutionary mechanisms that govern the activity of DNA and RNA. I will present novel multi-matrix and multi-tensor generalizations of the singular value decomposition as well as experimental verification and validation of some of the computational predictions. These models bring physicians a step closer to one day being able to predict and control the progression of cell division and cancer as readily as NASA engineers plot the trajectories of spacecraft today.

References:

[1] P. Sankaranarayanan, T. E. Schomay, K. A. Aiello and O. Alter, "Tensor GSVD of Patient- and Platform-Matched Tumor and Normal DNA Copy-Number Profiles Uncovers Chromosome Arm-Wide Patterns of Tumor-Exclusive Platform-Consistent Alterations Encoding for Cell Transformation and Predicting Ovarian Cancer Survival", *PLoS One* 10, article e0121396 (2015); doi: 10.1371/journal.pone.0121396.

Speeding up and automating multi-way models

Rasmus Bro

Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frb. C

rb@food.ku.dk

In this talk automated tensor modeling will be discussed. Tensor models have great potential in e.g. analysis of chromatographic data but the usefulness is hampered because the users that need this tool are 1) unaware that they need it (hence, maybe they don't!) and 2) are not skilled in data analysis and in selecting meta-parameters etc. Hence, the gap between the theoretically fancy methods and the actual real world users is way too big. Here, we will discuss how we aim to fix this and especially talk about what still needs to be fixed.

Two-mode K-Spectral Centroid analysis for studying multivariate dynamical processes

Eva Ceulemans

University of Leuven, BE

eva.ceulemans@ppw.kuleuven.be

Researchers that study dynamic processes, often collect multivariate time profiles, mapping the evolution of a set of variables over time, for multiple subjects. For instance, many clinical studies focus on the differential effect of an intervention on different symptoms, by repeatedly measuring symptom severity. To parsimoniously describe the huge information in such data and to pursue an insightful overview on how time profiles vary as a function of both subjects and variables, we propose two-mode K-Spectral Centroid (2M-KSC) analysis. This method, that combines the key ideas of multi-mode partitioning and one-mode K-Spectral Centroid analysis, simultaneously reduces subjects to subject clusters and variables to variable clusters. This clustering is based on the shape of the time profiles under study, implying that time profiles that correspond to a specific combination of a person cluster and variable cluster are modeled with one specific reference profile, reflecting the typical evolution over time. Furthermore, each time profile receives an amplitude score, indicating its overall intensity relative to its corresponding reference profile. We apply the new 2M-KSC method to time profiles reflecting the intensity of depression symptoms during citalopram treatment.

Distributed Large-Scale Tensor Decompositions in Collaborative Networks

André L. F. de Almeida

Department of Teleinformatics Engineering - Federal University of Ceará (UFC), Fortaleza, Brazil.

andre@gtel.ufc.br

In many disciplines, data inherently has more than two axes of variation and can be arranged as tensors (i.e. multi-way arrays). Computing tensor decompositions of multi-way datasets is particularly useful to extract hidden patterns and structure in data analytics problems. The traditional approach to compute tensor decomposition assumes a single machine with enough memory and processing power to process the whole data tensor. However, in large-scale (big data) applications where tensors can have billions of entries, tensor factorizations imply huge storage and processing capabilities that call for distributed algorithms affording some degree of parallel processing. The PARAFAC decomposition, also known as canonical polyadic (CP) decomposition, is a very popular tool for multi-way analysis, which has found innumerous applications in several domains including data mining, signal processing, and machine learning. In this talk, we discuss a fully distributed framework to compute the PARAFAC decomposition of a large-scale three-way data tensor across a network of collaborative machines with limited storage and computation resources. The big tensor is partitioned into several smaller sub-tensors that are allocated to different machines. Each machine operates on its sub-tensor to partially compute the factor matrices. The distributed computation of the PARAFAC decomposition is based on a "multi-way collaboration architecture": collaborative sub-networks of machines are established for each slicing direction, or mode, of the data tensor to estimate the factor matrices in a distributed and parallel fashion using alternating least squares. Such a multi-way network allows a meaningful reconstruction of the global factor matrices as for traditional (centralized) PARAFAC-ALS algorithms at the expense of some communication cost between machines. We also discuss how the communication cost scales with the number of machines. Some insights on extensions and applications of this approach will be given.

About tensorization for blind signal separation: concept and Illustrations

Otto Debals, **Lieven De Lathauwer**

KU Leuven, Kasteelpark, Arenberg 10, 3001 Leuven, Belgium;

otto.debals@esat.kuleuven.be

Given a mixture of some source signals, the Blind Signal Separation (BSS) problem consists of the identification of both the mixing matrix and the original sources. Specifically, one tries to determine M and S in X=MS with only M known. The source signals, mixing vectors and sensor signals are contained in S, M and X, respectively. One can see that the problem consists of a matrix factorization of X. The solution is not unique (except for a single source), and scaling and permutation of the mixing part and source signals are two insurmountable indeterminacies.

The eigenvalue decomposition, the QR/LQ decomposition and the singular value decomposition are the most well known matrix factorization techniques. However, the constraints defining these (such as triangularity and/or orthogonality) are too restrictive for BSS. More suitable techniques have been introduced in recent history, e.g., Non-negative Matrix Factorization (NMF), Sparse Component Analysis (SCA) and Independent Component Analysis (ICA).

The possibility of obtaining a single unique solution is a desirable quality of a BSS-technique. Proving this is a hard task. For example, NMF needs additional constraints (such as sparsity) to enforce uniqueness of the solution. Tensors (as higher-order generalizations of vectors and matrices) provide a solution for a lot of BSS techniques. Decompositions of tensors, such as the Canonical Polyadic Decomposition (CPD) and Block Term Decomposition (BTD), are very powerful as they are unique under only mild conditions. They can be used for BSS after the observation matrix is mapped to a tensor through tensorization, translating the assumptions made on the sources and/or mixing vectors. The decomposition afterwards ensures the uniqueness.

Different tensorization techniques have recently appeared but in a disparate manner. Also, the link to tensor decompositions has been omitted in many existing techniques for BSS despite it being inherently present. The multilinear setting however provides a comprehensive framework. We will present tensorization as an important concept by itself.

We shall discuss the following tensorization techniques: 1) Higher-Order Statistics, useful for ICA; 2) the stacking of matrices depending on a single (or multiple) parameters. Stacked covariance matrices are an example also suitable for ICA; 3) `Hankelization', useful for a separation into exponential polynomials; 4) `Löwnerization', useful for a separation into rational functions; 5) `Segmentation', useful for big data with compact representations through low-rank approximations.

The first item is a stochastic tensorization technique, as well as the example in the second item. The latter three are purely deterministic techniques. They work well for a low number of samples. There are other techniques which we omit for brevity. We will provide some separation examples with real-life datasets, especially for the Löwnerization technique.

A novel proposal for the identification of common and distinctive sources of variability in multi-set data

Raffaele Vitale[1], Onno E. de Noord[2], Alberto Ferrer[1]

[1] Dept. de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
[2] Shell Global Solutions International B.V., Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN Amsterdam, The Netherlands

aferrer@eio.upv.es

Differentiating common and distinctive sources of variability in multi-set data has been recently found to be an interesting and challenging task in many research and practical fields [1]. For instance, in an industrial scenario, one might be interested in identifying the underlying events driving the same production process evolving in reacting units of different scale and/or monitored by means of different sensors. In the last decades, many techniques have been proposed to this end, but lately much work and attention has been devoted to Joint-Y PLS (JY-PLS), a novel regression method, which is able to model the latent variable structure shared by two or more sets of data [2]. Until now, this approach has only been used for product transfer or monitoring model transfer between manufacturing sites or plants. Nevertheless, due to its algorithmic structure, it seems to be perfectly suitable for recognising phenomena or mechanisms described by either part or all of the data blocks under study. The main aim of this work is to evaluate the potential of JY-PLS for such applications and compare its performance to that of well-established methods, commonly resorted to for this purpose.

References:

[1] Van Deun, K., Smilde, A.K., Thorrez, L., Kiers, H.A.L. and Van Mechelen, I., *Chemometr. Intell. Lab.*, 2013 (129), 40-51
[2] García-Muñoz, S., MacGregor, J.F. and Kourti, T., *Chemometr. Intell. Lab.*, 2005 (79), 101-114

Acknowledgements:

This research work was partially supported by the Spanish Ministry of Economy and Competitiveness under the project DPI2011-28112-C04-02 and Shell Global Solutions International B.V. (Amsterdam, The Netherlands).

An SVD penalized approach to Candecomp/Parafac estimation

Paolo Giordani[1] and Roberto Rocci[2]

[1] Department of Statistical Sciences, Sapienza University of Rome
[2] Department of Business Government Philosophy Studies, University of Rome "Tor Vergata"

paolo.giordani@uniroma1.it

The Candecomp/Parafac (CP) model is a widely-used tool for synthetizing three-way arrays through a limited number of components. Unfortunately, its applicability may be prevented by the risk of obtaining degenerate solutions. In such cases, the CP algorithm takes a long time to converge. Upon convergence, solutions with diverging uninterpretable component matrices are observed. In the literature, CP degeneracy has been deeply investigated by several authors. A recognized remedy is to impose orthogonality constraints to one of the component matrices, denoted by A. Although this strategy solves the problem from a theoretical point of view, it fails from a practical point of view. In particular, it is unclear why degeneracy should be interpreted as an indication of the presence of orthogonal components underlying the data. For this purpose, we recently proposed some estimation procedures for avoiding CP degeneracy. These are the CP-Lasso (Giordani and Rocci, 2013a) and the CP-Ridge (Giordani and Rocci, 2013b), based on the Least absolute shrinkage and selection operator (Lasso) and on the ridge regularization term, respectively. The intuition behind both the proposals is to bound the condition number (the ratio of the largest singular value of a matrix to the smallest singular value) of A. In fact, it is well-known that degeneracy occurs when the condition number of A tends to infinity (Krijnen et al., 2008). This suggests us to develop a different estimation procedure for CP taking into account explicitly the singular values of A. We refer to this proposal as the CP-SVD. In CP-SVD we rewrite A in terms of its Singular Value Decomposition (SVD) and penalize directly the singular values far from 1. This is done by minimizing the CP loss function with a penalization term based on the singular values of A. Such a penalization term is weighted by a tuning parameter denoted by λ. A strategy to select the value of λ is proposed. The results of the application of CP-SVD to simulated and real life data show its effectiveness.

References:

Giordani, P., and Rocci, R. (2013a). Candecomp/Parafac via the Lasso., Psychometrika, 78, 669-684.
Giordani, P., and Rocci, R. (2013b). Candecomp/Parafac with ridge regularization. Chemometrics and Intelligent Laboratory Systems, 129, 3-9.
Krijnen, W.P., Dijkstra, T.K., and Stegeman, A. (2008). On the non-existence of optimal solutions and the occurrence of "degeneracy" in the CANDECOMP/PARAFAC model. Psychometrika, 73, 431-439.

A semi-algebraic framework for approximate CP decompositions via Simultaneous Matrix Diagonalizations (SECSI)

Martin Haardt

Ilmenau University of Technology, Communications Research Laboratory
P.O. Box 10 05 65, D-98684 Ilmenau, Germany

Martin.Haardt@tu-ilmenau.de

The Canonical Polyadic (CP) decomposition of R-way arrays is a powerful tool in multilinear algebra. Algorithms to compute an approximate CP decomposition from noisy observations are often based on Alternating Least Squares (ALS) which may require a large number of iterations to converge. To avoid this drawback we investigate semi-algebraic approaches that algebraically reformulate the CP decomposition into a set of simultaneous matrix diagonalization (SMD) problems.

To this end, we propose a *SE*mi-algebraic framework that allows the computation of approximate *C*P decompositions via *SI*multaneous Matrix Diagonalizations (SECSI). In contrast to previous Simultaneous Matrix Diagonalization (SMD)-based approaches, we use the tensor structure to construct not only one but the full set of possible SMDs. Solving all SMDs, we obtain multiple estimates of the factor matrices and present strategies to choose the best estimate in a subsequent step. This SECSI framework retains the option to choose the number of SMDs to solve and to adopt various strategies for the selection of the final solution out of the multiple estimates. A best matching scheme based on an exhaustive search as well as heuristic selection schemes with a reduced computational complexity are devised to flexibly adapt to specific applications. Several example algorithms with different accuracy-complexity trade-off points are compared to state-of-the-art algorithms. We obtain more reliable estimates and a reduced computational complexity.

For tensors with $R > 3$ dimensions, it is beneficial to combine the SECSI framework with the concept of generalized unfoldings (SECSI-GU) in order to enhance their identifiability. These generalized unfoldings are known from the "Semi-Algebraic Tensor Decomposition" (SALT) algorithm. The resulting SECSI-GU framework offers a large number of degrees of freedom to flexibly adapt the performance-complexity trade-off. As we show in numerical simulations, it outperforms SECSI and SALT for tensors with $R > 3$ dimensions.

An algebraic framework for multiblock data analysis

Mohamed Hanafi

Unité de Sensométrie et de Chimiométrie, ONIRIS, site de la Géraudière, Nantes, France

mohamed.hanafi@oniris-nantes.fr

A challenging problem in multivariate statistics is the study of relationships between several blocks of data, also known as "multi-block data analysis".
After twenty years of involvement around this methodology, and nearly a century of literature in this field, I have arrived to the conclusion that : the framework provided by matrix algebra is not enough to describe the whole richness and the complexity of block matrices, their manipulation, or the factorization strategies behind multi-block methods.
The aims of the present talk can be summarized as follows:

(i) To clarify definitions and vocabulary for block matrices and block tensors. What is a block matrix and what it is not?
(ii) To introduce the block dimension concept for block matrices and block tensors (as an extension of the usual dimension of matrices and tensors). How to take into account the block structure?
(iii) To provide (16) products for block matrices. How to extend the well known products for matrix to block matrix?
(iv) To discuss how to handle block matrix or block tensors on scientific softwares like R or matlab.

The main consequence of these contributions is twofold:

(i) An algebraic framework which extends algebra for matrix to block matrix is obtained. This new framework opens the possibility to study various block matrix factorization strategies as a natural extension of usual matrix factorizations.
(ii) To provide operational solutions for fast prototyping procedures, in particular for multi-block data analysis methods. Experimental packages are presented under R or matlab.

References

Chessel, D., Hanafi,M. (1996). Revue de Statistique Appliquée.(XLIV),2,35-60.
Hanafi, M., Ten Berge, J. (2003). Psychometrika,vol.68, No.1,97–103.
Hanafi, M., Kiers , H.A.L. (2006)..Computational Statistics and Data Analysis.(51),3,1491-1508.
Hanafi, M. (2007). Computational Statistics.22, 275-292.
Hanafi, M., Kohler, A., Qannari,E.M.(2010). Journal of Chemometrics, 24,703–709.

Assessing factorial invariance of two-way rating designs using three-way methods

P. M. Kroonenberg

Institute of Education and Child Studies, Leiden University,
Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

kroonenb@fsw.leidenuniv.nl

Assessing the factorial invariance of two-way rating designs such as ratings of concepts on several scales by different groups can be carried out with three-way models such as the Parafac and Tucker models. By their definitions these models are double-metric factorially invariant. The differences between these models lie in their handling of the links between the concept and scale spaces. These links may consist of unrestricted linking (Tucker2 model), invariant component covariances but variable variances per group and per component (Parafac model), zero covariances and variances different per group but not per component (Replicated Tucker3 model) and strict invariance (Component analysis on the average matrix). This hierarchy of invariant models, and the procedures by which to evaluate the models against each other, is illustrated in some detail with an international data set from attachment theory.

Visualising three-way arrays: An outerproduct rank perspective

Sugnet Lubbe and Darryn Williams

Department of Statistical Sciences, University of Cape Town, South Africa

sugnet.lubbe@uct.ac.za

Graphical displays of data can often provide a rudimentary understanding of the relationships inherent in the data set under investigation. In the context of two-way data, the most popular graphical tool is the PCA biplot.
This paper addresses some of the complexities of rank in a three-way context, particularly the fact that the two popularly used decomposition techniques, Tucker3 and PARAFAC, conceive of rank in different ways, namely multilinear rank and outerproduct rank respectively. The concept of orthogonal rank is also addressed. In the literature the multilinear rank decomposition receives a great deal of attention as a starting point for visualisation techniques. The outerproduct rank decomposition, however, receives less attention for the purpose of three-way visualisation.
A suitable orthogonal outerproduct rank decomposition is justified for the construction of a plot. The use of an orthogonal outerproduct rank decomposition allows the interpretation of the plot to be closely aligned with that of the PCA biplot, making it intuitive to use. Linear axes are fitted with markers that facilitate reading off the original data values, similar to the PCA biplot suggested by Gower and Hand (1996).
The methodology is applied to a dataset related to a study of blue crabs shell disease (Gemperline et. al, 1992) and the inferences made from the plot are compared to their conclusions. This includes a comparison of the proposed plot to the popularly used joint plot in order to show that it yields similar results and a discussion as to why this triplot can arguably be considered slightly more intuitive than the joint plot.

References

Gemperline, P. J., Miller, K. H., West, T. L., Weinstein, J. E., Hamilton, J. C., and Bray, J.T. 1992. Principal component analysis, trace elements, and blue crab shell disease. *Analytical Chemistry*, 64, A523-A531.
Gower, J.C. and Hand, D.J. 1996. *Biplots*. Chapman & Hall.

Variable selection in N-PLS

David Hervás[1], Abel Folch-Fortuny[2], Agustín Lahoz[3], Alberto Ferrer[2] and José Manuel Prats-Montalbán[2]

[1] Unidad de Bioestadística, Instituto de Investigación Sanitaria - Fundación Hospital La Fe, Valencia, Spain
[2] Grupo de Ingeniería Estadística Multivariante. Departamento de Estadística e IO Aplicadas y Calidad. Universitat Politècnica de València, Spain.
[3] Unidad de Hepatología Experimental, Instituto de Investigación Sanitaria - Fundación Hospital La Fe, Valencia, Spain.

jopramon@eio.upv.es

Variable selection appears as one paramount objective in different fields, such as *–omics* or image analysis. This dissertation presents two case studies: the first related to the inclusion of the LASSO method into the NPLS algorithm, in order to be able to shrink any of the modes related to the variables; the second one related to the reduction of the number of wavelengths in a hyperspectral camera in order to use 3 to 5 filters in a modified commercial camera.

When large amounts of data are available, introducing too few relevant parameters may produce some bias in the estimations, whereas the opposite may increment the variance unnecessarily. One way of dealing with these data, when N-way arrays are available and some variable is to be predicted, is applying N-PLS, which segregates the parameters producing a more parsimonious model. In order to further select the relevant variables and obtain simpler models for improving the interpretability and/or the prediction capabilities, different selection techniques can be applied. In this work, we present the introduction of LASSO in N-PLS. LASSO shrinks the coefficients of the model, causing some of them to be exactly zero and thus performing variable selection at the same time. We compare the results provided by LASSO-N-PLS with the creation of random null distributions of VIP's and weights, with posterior calculation of the statistical significance. The method showed good ability for the reliable selection of those important variables comprised in large *-omic* data sets. The second case deals with the detection of rotten oranges in the Valencian fruit export. These exports reached €2.7 billion in 2014. Therefore, it is of paramount importance to control the absence of rotten oranges in the warehouses, since they can infect other fruits and spoil hundreds of kg. Here we use hyperspectral images to discriminate between rotten and healthy oranges. A NPLS model is built based on a set of LCTF hyperspectral images taken from oranges from different varieties. Some oranges were infected by a fungus and others were infiltrated just with water. For each variety and wavelength a set of features are extracted, defining thus a three way X data set per variety: orange sample times features times wavelengths. After splitting the data in calibration and external test sets, a N-way Partial Least Squares Discriminant Analysis (NPLS-DA) model is applied for each variety, selecting also the very few wavelengths offering the best correct classification rates in the validation set.

Constraints in Multiway and Multiset Analysis - Applications in Chromatography

Sarah C. Rutan[1], Daniel W. Cook[1] and Robert C. Allen[2]

[1] Dept. of Chemistry, Virginia Commonwealth University, Richmond, VA 23284-2006
[2] SABIC Innovative Plastics, Selkirk, NY 12158

srutan@vcu.edu

Multiway and multiset methods are particularly useful for the analysis of liquid chromatographic data. Ideally these data provide trilinear and quadrilinear data structures, but in some cases small deviations from multilinear structure can limit the performance of the PARAFAC algorithm. In these situations, multivariate curve resolution-alternating least squares (MCR-ALS) can often be used to obtain chemically valid results. In this case, it can be important to consider a broad range of constraints to allow the fitting algorithm to converge to a chemically valid solution. Some of these constraints include nonnegatively, unimodality, selectivity or local rank, linear regression/calibration, peak separation, peak detection, peak shifting, normalization and closure. In this presentation, the implementation and relative influence of these constraints will be discussed and strategies for effective utilization of these constraints will be described.

Two takes on constrained CP decomposition using the alternating direction method of multipliers

Nikos Sidiropoulos, Athanasios Liavas, Kejun Huang

Dept. of Electrical and Computer Engineering, University of Minnesota
200 Union St. SE, Minneapolis, MN 55455

nikos@umn.edu

With few recent exceptions, all tensor factorization algorithms were originally developed for centralized, in-memory computation on a single machine; and the few that break away from this mold do not easily incorporate practically important constraints, such as non-negativity. In this talk, I will first introduce the alternating direction method of multipliers (ADMoM). Then, I will show how ADMoM can be used for constrained CP decomposition in ways that maintain simplicity similar to unconstrained CP, and naturally yield distributed algorithms suitable for parallel implementation in high performance computing architectures. Two families of algorithms will be presented: pure ADMoM, and one that uses ADMoM as a building block.

Separating common and distinct variation in multi-set data: theory and some examples

Age K. Smilde[1,3], Johan Westerhuis[1], Frans van der Kloet[1], Ingrid Måge[2], Tormod Naes[2,3] and Rasmus Bro[3]

[1] Biosystems Data Analysis, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.
[2] Nofima AS, Osloveien 1, NO-1430, s, Norway
[3] Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.

a.k.smilde@uva.nl

One of the active research topics in multi-set data analysis or data fusion is to separate common from distinct variation in multiple blocks of data. This goes under different names such as shared/unshared variation, joint/ individual variation, unique variation and specific variation. We will present a mathematical and geometrical framework for common and distinct variation and give definitions to structure the field. Some methods will be placed in that framework and worked out in an example of fusing mRNA and miRNA data.

Regularized Generalized Canonical Correlation analysis for Multiway data

Arthur Tenenhaus and Laurent Le Brusquet

Laboratoire des Signaux et Systèmes, CentraleSupélec, Gif-sur-Yvette, France

arthur.tenenhaus@supelec.fr

Several examples of either three-way data or multiblock data can be found in a variety of domains including chemometrics, bioinformatics to name but a few. Nowadays, it frequently occurs to encounter data combining multiway and multiblock data. Regularized Generalized Canonical Correlation Analysis (RGCCA) is currently geared for the analysis of two-way data matrices. In this work, RGCCA is extended to the multiway data configuration (Multiway RGCCA - MGCCA) by applying appropriate kronecker constraints on the RGCCA outer weight vector. The main aim of MGCCA is to study the relationships between a collection of multi-way data table.

Sparse Common and Distinctive Components

Katrijn Van Deun & Zhengguo Gu

Methodology & Statistics, Tilburg University

Katrijn.VanDeun@psy.kuleuven.be

Multiset data consist of several heterogeneous data sets that are obtained for the same set of objects. For example, both liquid chromatograph and gas chromatography mass spectrometry data obtained for the same set of fermentation batches. Often it is of interest to find both the common sources of variation underlying such multiset data as well as the sources of variation that are specific for a particular dataset. Several component based approaches have been proposed for this purpose, yet most of these approaches are dense in the sense that all variables load on the components. Mainly from an interpretational point of view, having sparse components based on a few non-zero loadings only is most desirable. Here, we propose a novel approach that yields sparse common and distinctive components making use of structured penalties like the sparse group lasso and the elitist lasso. A highly efficient algorithm is proposed and a stability selection procedure is used to tune the variable selection.

The Amazing Flexibility of Fluorescence Spectroscopy for Trilinear Modeling

Mark H. Van Benthem[1], Benjamin Grant[2], Kurt C. Peterson[2], and Gregory D. Gillispie[2]

[1] Sandia National Laboratories, Albuquerque, NM 87185-0886
[2] Fluorescence Innovations, Inc., Minneapolis, MN 55414

mhvanbe@sandia.gov

The inherent flexibility of molecular fluorescence spectroscopy, given the several independent data acquisition modalities, makes it ideally suited to trilinear methods, such as PARAFAC. There is a variety of ways to generate trilinear fluorescence data where one of the modes is concentration, say for ligand binding. Traditionally the other two modes are chosen from excitation wavelength, emission wavelength, and fluorescence decay. We are exploring an alternative mode in which the shape of the excitation pulse is systematically varied. We have compared various combinations of modes satisfying the trilinear criterion for their ability to resolves compound mixtures. The work involves extensive simulations with realistic representations of noise, followed by experimental studies to confirm the predictive ability of the simulations. The feasibility of routinely applying trilinear analysis in high-throughput assays, e.g., for drug discovery data acquired with a microplate reader, is assessed.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

A randomized block sampling approach for the CP decomposition of large-scale tensors

Nico Vervliet[1] **and Lieven De Lathauwer**[1,2]

[1] Dept. of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven;
iMinds Medical IT, Leuven, BE.
[2] Group of Science, Engineering and Technology, KU Leuven Kulak, E. Sabbelaan 53, B-8500 Kortrijk, BE.

nico.vervliet@esat.kuleuven.be

To make sense out of large-scale datasets, simple structures are often assumed. In the case of tensors such structures are, for example, the low multilinear rank approximation, the block term decomposition or tensor trains. In this talk we focus on the decomposition in rank-1 terms, also called the (canonical) polyadic decomposition (CPD), PARAFAC or CANDECOMP. Over the years many algorithms have been developed to compute this decomposition, ranging from direct methods over alternating least squares algorithms to all-at-once optimization routines. In all these algorithms the size of the tensors that can be handled is limited by the so called curse of dimensionality. This curse encompasses the computational and memory difficulties arising from the exponential growth of the data for higher order tensors. New algorithms exploiting sparsity, incompleteness and parallelization have emerged recently, as have some algorithms that work on blocks of the tensor instead of the full tensor. A few examples are grid PARAFAC, ParCube, GigaTensor and PARACOMP.

In this talk we will demonstrate another strategy. Inspired by the success of randomization and stochastic optimization, we propose the randomized block sampling CPD algorithm. This new algorithm samples a small, random subblock from a tensor in every iteration, computes an update for the variables based on this block and then moves on to the next block. Exploiting the locality property of a CPD, the computations can be performed efficiently. This process is repeated up to convergence. As a stopping criterion, we present a new measure based on the Cramér-Rao bound which combines information from noise estimates and how much the variables change over a few iterations, rendering it more informative than criteria based on the function value or the step size alone. For the randomized block sampling CPD algorithm, we experimentally show how carefully choosing the block size can decrease the computational cost significantly by sampling only a fraction of the data. Using a step restriction schedule, the resulting decompositions are almost as accurate as the results from state-of-the-art decomposition algorithms for full tensors. How to select this schedule is illustrated with some experiments. The scalability of our method is illustrated by the decomposition of an 8TB tensor in a few minutes on a standard laptop.

Multilinearity in Multivariate Curve Resolution of multiway and multiset data

Romà Tauler

IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, SPAIN

Roma.Tauler@idaea.csic.es

Multivariate Curve Resolution (MCR) has evolved in the recent years as a powerful tool to investigate complex chemical systems measured by multivariate instrumental analytical methods (e.g. spectroscopy, chromatography, ...). The basic assumption of MCR methods is that the experimental data can be arranged in a 2D data table or in a two-way (two-mode) data matrix which can be appropriately described using a low-rank bilinear model factor decomposition, and that the sought factor solutions fulfill a number of natural constraints like non-negativity. Typically this MCR data decomposition can be easily achieved by means of a constrained alternating least squares (MCR-ALS) algorithm. This analysis can be easily extended to multiset and multiway data by means of matrix augmentation strategies of the different data sets concatenated in the direction of the different data modes and with the extension of the bilinear model data decomposition to them. It has been already shown (1) that this extension can also include the fulfillment of the trilinear model in the analysis of three-way data by means of the implementation of a trilinear constraint (2). Results obtained by the application of this constraint in the analysis of three-way data are equivalent to those obtained by other three-way methods like PARAFAC and/or its extensions (3). Furthermore the implementation of this constraint can report some advantages like the possibility to implement the trilinear model only to some of the resolved components (mixed bilinear and trilinear models). Recently, the same idea behind the implementation of the trilinear constraint has been extended to the quadrilinear (4) model for the analysis of four-way data, and it is proposed to be generalized to any type of multilinear model, including also possible component interactions (5). In the presentation of this work, the extension of the multilinear constraint for MCR analysis of different examples of multiway data will be shown and the results discussed.

References

1 R. Tauler, A.K. Smilde, B.R. Kowalski. J. Chemometr., 1995, 9, 3.; 40. R. Tauler. Chemom. Intell. Lab. Sys., 1995, 30, 133.

2 R. Tauler, I. Marqués, E. Casassas. J. Chemometr., 1998, 12, 55.

3. A. de Juan, R. Tauler. J. Chemom. 2001, 15, 749.

4. A. Malik, R. Tauler. Anal. Chim. Acta, 2013, 794, 20; 52. A. Malik, R. Tauler. Chemom. Intell. Lab. Sys. 2014, 135, 223–234.

5. E. Peré-Trepat, A. Ginebreda, R. Tauler. Chemom. Intell. Lab. Sys. 2007, 88, 69; M. Alier, M. Felipe, I. Hernández, R. Tauler. Anal. Bioanal. Chem. 2011, 399, 2015.

Sparse modelling of Multiset data

Marieke E. Timmerman[1], Eva Ceulemans[2], Henk A.L. Kiers[1]

[1] Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen
[2] KU Leuven, Belgium

m.e.timmerman@rug.nl

Multiset modelling aims at expressing the major sources of variance within the sets, in an interpretable way. To this end, various multi-set models are available. In empirical practice, the challenge is to select a well-fitting model that is easy to interpret. Typically, the interpretation is facilitated when the model has a certain degree of sparseness. To achieve this one could consider three main approaches, namely 1. exploiting the transformational freedom in a model; 2. soft sparseness: steering the model estimates towards sparseness by imposing a penalty on non-sparseness; 3. hard sparseness: imposing sparseness on some of the model estimates. We will present a new flexible transformational procedure that explicitly aims at retrieving a sparse solution. A nice feature is the option to specify subsets of variables for which an equal degree of sparseness is desired. This new transformational approach will be contrasted to the soft and the hard sparseness approaches to multi-set modelling, including illustrations by empirical examples.

Clusterwise three-way models to account for heterogeneity in three-way data

Tom Wilderjans

KU Leuven, BE

tom.wilderjans@ppw.kuleuven.be

Nowadays, in many fields of science, challenging research questions often call for the analysis of three-way data sets (e.g., conventional sensory profiling, multi-subject EEG, fluorescence spectroscopy and nuclear magnetic resonance data). To reveal the structure underlying such data, often three-way component methods (e.g., Parafac, Tucker3) have been used. These methods postulate the underlying components being the same for all elements of the three modes, an assumption which may be violated in many situations (e.g., groups of raters adopting different sensory dimensions). To account for this type of heterogeneity, we propose the class of clusterwise three-way component models. Members of this class have in common that the elements of one of the modes (e.g., subjects) are clustered and, simultaneously, a three-way component model is fitted to the data within each cluster. As such, qualitative differences (i.e., heterogeneity) in underlying component structure across clusters can be disclosed. The goal of this presentation is to introduce this model class by highlighting some of its representative members (e.g., Clusterwise Parafac) and illustrating them with empirical data. During the talk, we will also discuss models in which the number of components is allowed to vary across clusters, resulting in a challenging model selection problem.

First name	Last name	MAIL	Address
Laurent	Albera	laurent.albera@univ-rennes1.fr	Université de Rennes 1, Rennes, France
Casper	Albers	c.j.albers@rug.nl	University of Groningen, Grote Kruisstraat 2/1 , 9712 TS Groningen, NL
Orly	Alter	orly@sci.utah.edu	University of Utah (USA)
Rasmus	Bro	rb@life.ku.dk	Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frb. C
Eva	Ceulemans	eva.ceulemans@ppw.kuleuven.be	University of Leuven, BE
Marina	Cocchi	marina.cocchi@unimore.it	Dept. of Chemical and Geolgical Sciences, University of Modena and Reggio Emilia, Via Campi 103, Modena, Italy
André	De Almeida	andre@gtel.ufc.br	Department of Teleinformatics Engineering - Federal University of Ceará (UFC), Fortaleza, Brazil.
Otto	De Bals	otto.debals@esat.kuleuven.be	KU Leuven, Kasteelpark, Arenberg 10, 3001 Leuven, Belgium;
Alberto	Ferrer	aferrer@eio.upv.es	Dept. de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
Paolo	Giordani	Paolo.Giordani@uniroma1.it	Department of Statistical Sciences, Sapienza University of Rome, P.le A. Moro, Rome, Italy
Martin	Haardt	martin.haardt@tu-ilmenau.de	Ilmenau University of Technology, Communications Research Laboratory, P.O. Box 10 05 65, D-98684 Ilmenau,Germany
Mohamed	Hanafi	mohamed.hanafi@oniris-nantes.fr	Unité de Sensométrie et de Chimiométrie, ONIRIS, site de la Géraudière, Nantes, France
Henk A. L.	Kiers	H.A.L.Kiers@rug.nl	Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen NL
Pieter	Kroonenberg	kroonenb@fsw.LeidenUniv.nl	Institute of Education and Child Studies, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, NL
Sugnet	Lubbe	sugnet.lubbe@uct.ac.za	Department of Statistical Sciences, University of Cape Town, South Africa
Xavier	Luciani	lucianix@gmail.com	Université de Rennes 1, Rennes, France
Federico	Marini	Federico.Marini@uniroma1.it	Department of Chemistry, Sapienza University, Rome, P.le A. Moro, Rome, Italy
Jose Manuel	Prats Montalban	jopramon@eio.upv.es	Dept. de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
Sarah	Rutan	srutan@vcu.edu	Virginia Commonwealth University, Richmond (USA)
Nicos	Sidiropoulos	nikos@umn.edu	Dept. of Electrical and Computer Engineering, University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455
Age	Smilde	A.K.Smilde@uva.nl	Biosystems Data Analysis, University of Amsterdam, Science Park 904, 1098 XH, Amsterdam NL
Romà	Tauler	rtaqam@iiqab.csic.es	IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, SPAIN
Arthur	Tenenhaus	Arthur.Tenenhaus@supelec.fr	Supelec GIF SUR YVETTE CEDEX
Marieke	Timmerman	m.e.timmerman@rug.nl	Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen, NL
Giorgio	Tomasi	giorgio.tomasi@gmail.com	University of Copenhagen, (DK)
Katrijn	Van Deun	Katrijn.VanDeun@psy.kuleuven.be	Methodology & Statistics, Tilburg University
Mark	Van Benthem	mhvanbe@sandia.gov	Sandia National Laboratory Livermore CA (USA)
Nico	Vervliet	nico.vervliet@esat.kuleuven.be	Dept. of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven; iMinds Medical IT, Leuven, BE.
Thomaz	Veloso da Silva	thomazveloso@gmail.com	Faculty of Science, University of Copenhagen, Rolighedsvej 30, 1958 Frb. C
Raffaele	Vitale	rvitale86@gmail.com	Dept. de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain
Beata	Walczak	beata.walczak@us.edu.pl	Silesian University, Katowice (PL)
Tom	Wilderjans	tom.wilderjans@ppw.kuleuven.be	KU, Leuven (Belgium)