Autumn Meeting of the VOC

Data Fusion

November 26, 2010

Leiden University, FSW, Pieter de la Court Building (SB11)

Program:

10.00	Registration and Coffee
10.15	Iven Van Mechelen: A generic linked-mode decomposition model for data fusion
11.15	Michel van de Velden: Generalized canonical correlation analysis with missing values
12.00	Lunch
13.00	Peyman Zarrineh: Module-based comparative gene expression analysis: evolutionary conserved coexpression in *Bacillus subtilis* and *Escherichia coli*
13.45	Pascal Van Hattum: The proof of the pudding is in the eating. Data fusion: An application in marketing
14.30	Tea
15.00	Hans Kiesl: Uncertainty in data fusion
16.00	Drinks

Registration details for the Autumn Meeting:

Those who would like to participate are welcome and are kindly requested to register by sending an e-mail to meeting@voc.ac with subject 'Registration Autumn Meeting 2010' and including your name and affiliation in the body of the e-mail. Participation is free, lunch is available for 10 Euros and must be requested upon registration. Registration deadline: November 19th.

Abstracts for the Autumn Meeting

Iven Van Mechelen (University of Leuven): A generic linked-mode decomposition model for data fusion

As a consequence of our information society, not only more and larger data sets become available, but also data sets that include multiple sorts of information regarding the same system. Such data sets can be denoted by the terms coupled, linked, or multiset data, and the associated data analysis can be denoted by the term data fusion. In this talk, I will first give a formal description of coupled data, which allows the data-analyst to typify the structure of a coupled data set at hand. Second, I will list two meta-questions and a series of complicating factors that may be useful to focus the initial content-driven research questions that go with coupled data, and to choose a suitable method of data fusion. Third, I will propose a generic framework for a family of decomposition-based models pertaining to an important subset of data fusion problems. I will conclude the talk with a long list of research challenges that go with the proposed generic modeling approach. Throughout the talk I will illustrate with examples from the domain of systems biology.

Reference:
Van Mechelen, I., & Smilde, A.K. (in press). A generic linked-mode decomposition model for data fusion. *Chemometrics and Intelligent Laboratory Systems*. doi:10.1016/j.chemolab.2010.04.012

Iven Van Mechelen *is Professor of Quantitative Psychology at the University of Leuven. He received a Master of Mathematics degree from the University of Antwerp (1980) and a PhD degree in Psychology from the University of Leuven (1989). His research work includes the development and taxonomic organization of two-mode clustering methods, methods for multiway data, and custom-made data-analytic methods for a contextualized study of individual differences in personality, emotions, and affective dynamics. He has published in a broad range of journals, including Psychometrika, Journal of Classification, Computational Statistics and Data Analysis, Bioinformatics, and various methodological and substantive psychological journals. He has been on the board of VOC from 1998 till 2004. At present he is President-Elect of the International Federation of Classification Societies (IFCS).*
(see also http://ppw.kuleuven.be/okp/people/Iven_Van_Mechelen/)

Michel van de Velden (Erasmus University Rotterdam): Generalized canonical correlation analysis with missing values

Generalized canonical correlation analysis is a versatile technique that allows the joint analysis of several sets of data matrices. The generalized canonical correlation analysis solution can be obtained through an eigenequation and distributional assumptions are not required. When dealing with multiple set data, the situation frequently occurs that some values are missing. In this paper, two new methods for dealing with missing values in generalized canonical correlation analysis are introduced. The first approach, which does not require iterations, is a generalization of the Test Equating method available for principal component analysis. In the second approach, missing values are imputed in such a way that the generalized canonical correlation analysis objective function does not increase in subsequent steps. Convergence is achieved when the value of the objective function remains constant. By means of a simulation study, we assess the performance of the new methods. We compare the results with those of two available methods; the missing-data passive method, introduced in Gifi's homogeneity analysis framework, and the GENCOM algorithm developed by Green and Carroll. An application using world bank data is used to illustrate the proposed methods.

Michel van de Velden *is assistant professor at the Econometric Institute of the Erasmus University Rotterdam. His research interests concern development and application of visualization methods for multivariate data. His academic work has been published in journals from several disciplines, such as mathematics, statistics, psychometrics, sensometrics and archaeology. Since 2005, Michel has been the treasurer for the International Association for Statistical Computing (IASC). He is also a board member for the Dutch/Flemish classification society (VOC) and the Economic section of the Netherlands society of statistics and operations research (VVS). More information about his past and current academic activities can be found on his personal website.*

Peyman Zarrineh (University of Leuven): Module-based comparative gene expression analysis: evolutionary conserved coexpression in *Bacillus subtilis* and *Escherichia coli*

Increasingly large scale expression compendia for different species are becoming available. By exploiting the modularity of the coexpression network, these compendia can be used to identify biological processes for which the expression behavior is conserved over different species. However, comparing module networks across species is not trivial. The definition of a biologically meaningful module is not a fixed one and changing the distance threshold that defines the

degree of coexpression give rise to different modules. As a result when comparing modules across species, many different partially overlapping conserved module pairs across species exist and deciding which pair is most relevant is hard.

Therefore we developed a method referred to as COMODO (COnserved MODules across Organisms) that uses an objective selection criterium to identify conserved expression modules between two species. The method uses as input microarray data and a gene homology map and provides as output pairs of conserved modules and searches for the pair of modules for which the number of sharing homologs is statistically most significant relative to the size of the linked modules. To demonstrate its principle, we applied COMODO to study coexpression conservation between the two well studied bacteria *Escherichia coli* K12 and *Bacillus subtilis*.

Peyman Zarrineh *obtained his Bachelor and Master of Science in Computer Science and in Bioinformatics at Tehran University (Iran) and an additional master in bioinformatics from Chalmers university of technology. Currently he is a PhD student at the University of Leuven, Department of Electrical Engineering and Department of Microbial and Molecular Systems, where he prepares a PhD on developing coclustering methods for the comparison of coexpression behavior across species.*

Pascal van Hattum (The SmartAgent Company and University of Utrecht): The Proof of the Pudding is in the eating. Data Fusion: An Application in Marketing

Data fusion, or combining multiple data sets in one data set, is not a new concept. However, due to the increasing desire of differentiated direct marketing strategies, it is getting more popular in marketing. This paper shows how marketing information can be fused to a company's customer database. Using real marketing applications, two traditional data fusion methods, that are, polytomous logistic regression and nearest neighbor algorithms, are compared with two model based clustering approaches. Finally, the results are evaluated using internal and external criteria.

Pascal van Hattum *studied Business Mathematics and Computer Science at the VU University Amsterdam (2001). In cooperation with The SmartAgent Company he worked part-time on his Ph.D project 'Market Segmentation Using Bayesian Model Based Clustering' at the Department Methodology and Statistics at the University of Utrecht (2009). Currently he is manager Data Intelligence at The SmartAgent Company and combines this function with further research at the Department of Methodology and Statistics at the University of Utrecht.*

Hans Kiesl (Regensburg University): Uncertainty in data fusion

Data fusion (also called statistical matching) tries to combine information from different data sets by matching on those variables that are common to both files. Algorithms like nearest neighbour or Mahalanobis distance matching are routinely applied, but it is well known that they implicitly assume conditional independence of those variables that have not been jointly observed (called specific variables). In this presentation, we discuss how to quantify the amount of uncertainty in the matching process by calculating bounds on the feasible correlations of the specific variables. Since data fusion might be viewed as a missing data problem, we propose a multiple imputation algorithm that creates different matched data sets with different feasible correlation matrices. Since several recent studies have used propensity score matching for combining different data sets, we will also discuss why propensity score matching is appropriate for the estimation of average treatment effects in the context of Rubin's causal model (where we have to deal with a different conditional independence assumption) but should not be applied in the data fusion setting.

Hans Kiesl *is a professor of Statistics at Regensburg University of Applied Sciences, Germany. He received his PhD in 2002 (Bamberg University, Germany) with a thesis on measures of ordinal dispersion. He worked as a statistician and survey methodologist at the German Federal Statistical Office (Destatis) and at the German Institute for Employment Research (IAB). His primary research interests include sampling theory, missing data methods, variance estimation and statistical matching.*