VOC Newsletter 64

November 2024

Vereniging voor Ordinatie en Classificatie / Dutch-Flemish Classification Society

Chairman: Mark de Rooij, Universiteit Leiden, Faculteit Sociale Wetenschappen, Instituut Psychologie, Methodologie & Statistiek, Postbus 9555, 2300 RB Leiden, Nederland (rooijm@fsw.leidenuniv.nl)

Secretary: Kathrin Gruber, Erasmus Universiteit Rotterdam, Erasmus School of Economics, Econometric Institute, Postbus 1738, 3000 DR Rotterdam, Nederland (gruber@ese.eur.nl)

Treasurer: Tom Wilderjans, Universiteit Leiden, Faculteit Sociale Wetenschappen, Instituut Psychologie, Methodologie & Statistiek, Postbus 9555, 2300 RB Leiden, Nederland (t.f.wilderjans@fsw.leidenuniv.nl)

Editor: Pieter Schoonees, Erasmus Universiteit Rotterdam, Erasmus School of Economics, Econometric Institute, Postbus 1738, 3000 DR Rotterdam, Nederland (schoonees@ese.eur.nl)

VOC website: http://www.voc.ac
Postbankrekening (IBAN) NL86 INGB 0000 161723 t.n.v. Vereniging voor Ordinatie en Classificatie.

Program in short

VOC Jubilee Conference

21 - 22 November 2024, ISVW Leusden

Dodeweg 8 3832 RD Leusden

Thursday, 21 November 2024

11:00 - 11:30	*Arrival and Welcome*
11:30 - 12:15	Fred van Eeuwijk
12:15 - 13:30	*Lunch*
13:30 - 15:00	Marieke Timmerman
	Eyke Hüllermeier
15:00 - 15:45	*Coffee Break*
15:45 - 18:00	Sanwouly Marlene Yao
	Bert van der Veen
	Katrijn van Deun
18:00 -	*Drinks and Dinner*

Friday, 22 November 2024

07:00 - 09:00	*Breakfast*
09:00 - 10:30	Marina Cocchi
	Iven van Mechelen
10:30 - 11:15	*Coffee Break*
11:15 - 12:45	Andreas Alfons
	Alberto Ferrer
12:45 - 14:00	*Lunch*
14:00 - 15:30	Yves Rosseel
	Jeanine Houwing-Duistermaat
15:30 -	*Closing*

From the President

Dear VOC colleagues,

This year, the VOC celebrates its 35th anniversary. As we do with every 5th anniversary, we organized a two-day conference with invited speakers. The program of our meeting can be found in this newsletter, and as you can see, we accomplished a very diverse program with distinguished speakers from different disciplines. All speakers have a passion for data and what we can learn from data, but the field of application widely differs. Some speakers have a background in the social, economic, or behavioral sciences, others in chemistry, artificial intelligence, biology, or medicine. Such diversity is characteristic for our society. I am really looking forward to the meeting, I hope you do too!

Mark de Rooij

Conference Announcement

Joint VOC/CLADAG Meeting
8 - 10 September 2025
Naples, Italy

The VOC is jointly organizing an international joint meeting with the Italian Classification Society (CLADAG) in Naples, Italy, in September 2025. Anyone interested in the conference is welcome to register and submit a short abstract being presented at the conference.

Participants have in addition the opportunity to submit a paper (minimum 8, maximum 12 pages) to the forthcoming Springer volume "Advances in Supervised and Unsupervised Statistical Data Analysis", focusing on developments in supervised and unsupervised statistical methods and models.

Areas of interest include, but are not limited to clustering, pattern recognition, fuzzy methods, proximity structure analysis, mixture models, decision trees, data analysis, model selection, and textual classification. Proposals for critical and innovative applications of these types of models to real data, highlighting the fundamental contribution of statistical science to modern data analysis, are particularly welcome.

Registrations will open in January 2025. The deadline for the submission of the paper in the Springer volume is March 1, 2025. Information about the conference can be found at https://cladag2025.unina.it/.

Sponsors

The VOC Jubilee Conference is sponsored by the Erasmus Research Institute of Management (ERIM), LUXs Data Science and Biometris. Thank you for the support!

Meeting Location: ISVW

The VOC Jubilee Meeting takes place at the International School for Philosophy (Internationale School voor Wijsbegeerte, ISVW) in Leusden, the Netherlands. The ISVW, founded in 1916, is a non-profit institution and center for learning, practicing and developing philosophy located in the heart of the Netherlands on the vast ISVW estate in the woods of Leusden. The school shares its site with a hotel, conference center and publishing house.

More information about the ISVW is available from its website.

The ISVW is reachable by public transportation, but **the options are limited and time restricted**. Therefore, **please plan ahead** using https://www.9292.nl/en (or the corresponding phone app) to bus stop ISVW, Leusden, if you are travelling with public transport.

Detailed information about reaching the ISVW is available (in Dutch only) at https://isvw.nl/over-ons/contact/. Below is a summary.

Bus

The ISVW is reachable by Syntus Bus 19 (Direction ISVW via Oud-Leusden) from Amersfoort Centraal train station. Get off at the final stop. The bus runs only on weekdays twice per hour from 07:04 to 18:46. It is possible to pay by OV chipkaart or debit/credit card for your journey of approximately 15 minutes.

Alternatives to the bus

The ISVW lists alternative, more expensive, options to the bus on their website and they offer a shuttle service, see this website.

Bicycle

ISVW is reachable by bicycle from Amersfoort Centraal in under 20 minutes (4.6 km). An public OV bicycle can be rented at Amersfoort Centraal station for less that 4.55 euro per 24 hours (maximum rental period of 72 hours). The ISVW has a recommended route indicated at https://isvw.nl/over-ons/contact/.

Car
There is ample parking available at ISVW. Please consult https://isvw.nl/over-ons/contact/ or your preferred navigation app for information on reachability.

Programme: VOC Jubilee Conference

Leusden, 21 - 22 November 2024

Internationale School voor Wijsbegeerte (ISVW) Dodeweg 8 3832 RD Leusden

Thursday, 21 November 2024

11:00 – 11:30	Arrival (11:00) and Welcome (11:25)
11:30 – 12:15	Fred van Eeuwijk (WUR) – Modelling of Genotype by Environment by Management Interactions
12:15 – 13:30	Lunch
13:30 – 15:00	Marieke Timmerman (RUG) – Statistical Modelling Approaches to Derive Norms for a Psychological Test
	Eyke Hüllermeier (LMU Munich) – Uncertainty Quantification in Machine Learning: From Aleatoric to Epistemic
15:00 – 15:45	Coffee Break
15:45 – 18:00	Sanwouly Marlene Yao (RU) – Handheld Spectroscopy for Quality Assessment in the Chemical industry *(Student Presentation Award Winner)*
	Bert van der Veen (NTNU) – From Model-based to Hierarchical Ordination in Community Ecology
	Katrijn van Deun (TIU) – Easy Analysis of Complex Data: Regularized Multigroup Approximate Exploratory Factor Analysis
18:00 –	Drinks (18:00) and Dinner (19:00)

Friday, 22 November 2024

07:00 – 09:00	Breakfast
09:00 – 10:30	Marina Cocchi (Modena) – Developing Locally Weighted Multiblock Approaches for Real Time Quality Prediction in Latent Variable Multivariate-based Process Monitoring
	Iven van Mechelen (KU Leuven) – Redeeming at Last the Classipedia Promise: Onset of a Conceptual Outline Map to Get a Hold on the Jungle of Cluster Analysis
10:30 – 11:15	Coffee Break
11:15 – 12:45	Andreas Alfons (EUR) – Robust Correlation Estimation with Discrete Rating-scale Data
	Alberto Ferrer (València) – Multivariate Six Sigma for Industry 4.0
12:45 – 14:00	Lunch
14:00 – 15:30	Yves Rosseel (Ghent) – The Structural-after-measurement (SAM) Approach to SEM
	Jeanine Houwing-Duistermaat (RU) – Statistical Integration of Cross-sectional and Longitudinal Omics Datasets
15:30	Closing

Vereniging voor Ordinatie en Classificatie
Dutch/Flemish Classification Society

VOC

VOC Jubilee Conference
21-22 November 2024

Internationale School voor Wijsbegeerte (ISVW), Leusden

Dodeweg 8 3832 RD Leusden

Book of Abstracts

Scope

The Dutch/Flemish Classification Society, VOC, aims at communicating scientific principles, methods, and applications of ordination and classification. The VOC is a member of the International Federation of Classification Societies (IFCS).

THURSDAY, 21 NOVEMBER 2024

Modelling of Genotype by Environment by Management Interactions

Fred van Eeuwijk
Biometris Wageningen University

A main objective of plant breeding is the development of new plant varieties with improved genetic properties related to yield, quality, and sustainability. Societal and consumer demands ask for high yielding varieties with good taste and shelf life at low inputs and minimum ecological imprints. To achieve this objective large numbers of genotypes (varieties) are evaluated under different management and environmental conditions in series of controlled (management) and uncontrolled (environment) trials. The latter trials consist of evaluations in field trials at a number of locations over various years. When differences between genotypes depend on environment and management, genotype by environment by management interactions (GxExM) occur. Such interactions complicate the life of plant breeders and farmers because genotypic superiority and variety recommendation will be specific to management and environment and no unconditional recommendations can be made. For the analysis of GxExM, a wide range of statistical methods have been proposed, many of them firmly based in quantitative genetic theory. The most popular methods for GxExM are members of the classes of linear models, linear mixed models (with advanced variance-covariance models to model genetic and environmental variances and correlations) , and bilinear models (including biplots). Over the last decade, satellites, drones and sensors have added lots of longitudinal phenomic and enviromic information to the huge amount of genomic information of the decade before. The challenge has become to predict yield for any plant genotype as a function of genomic, phenomic and enviromic information. Because of the at first sight large amounts of data, the GxExM problem has also attracted the attention of AI researchers. Simultaneously, plant physiologists have become more interested in plant breeding because phenomic and enviromic data appear to facilitate predictive models with more biology than statistical and AI models. In the presentation, I will describe traditional and new approaches to model GxExM. For VOC members, GxExM data are just an example of three-way data.

Statistical Modelling Approaches to Derive Norms for a Psychological Test

Marieke E. Timmerman
University of Groningen

A norm-referenced score on a psychological test, like an IQ-score, expresses the position of an individual test taker in the reference population, like the Dutch population of the same age as the test taker. Norm-referenced scores are derived from test scores obtained from a normative sample. This sample is drawn from all reference populations involved, like the Dutch population in the age range of the intelligence test. In this presentation, I will provide an overview of available statistical modelling approaches, and their properties for deriving precise and unbiased norms.

Current leading norming approaches involve some form of regression modelling of the raw test scores (e.g., sum scores). I will motivate the usefulness of generalized additive models for location, scale, and shape (GAMLSS) for norming, and discuss model selection, transformation to normed scores and visualization of results. I will also discuss how to adjust for non-representativeness of the normative sample using multilevel regression and poststratification (MRP). This approach is then compared to other available adjustment methods: weighted regression and weighted cNORM. The results of our simulation study indicate that MRP is more efficient than weighted regression and weighted cNORM. I will illustrate the use of GAMLSS and MRP with normative data from a language comprehension test.

Further, I will outline how item responses, rather than raw test scores, can be modelled to estimate norms. I will introduce 2PL-norm, involving a Bayesian two-parameter logistic IRT model with age-dependent mean and variance of the latent trait distribution. 2PL-norm is then compared to the raw test score based methods GAMLSS and cNORM. The results of our simulation study indicate that 2PL-norm has the best overall performance of the three, but that 2PL-norm is less effective at the tails of the latent trait. Moreover, the credible intervals from 2PL-norm, expressing error due to measurement and sampling variability, have a much better coverage than the confidence intervals of cNORM and GAMLSS. I will illustrate the use of 2PL-norm with normative data from an intelligence test.
I will conclude by recommendations for norming practice, and a wish list for future developments.

Co-authors: Casper J. Albers, Hannah Heister, Lieke Voncken, and Klazien de Vries.

Uncertainty Quantification in Machine Learning: From Aleatoric to Epistemic

Eyke Hüllermeier
LMU Munich

Due to the steadily increasing relevance of machine learning for practical applications, many of which are coming with safety requirements, the notion of uncertainty has received increasing attention in machine learning research in the recent past. This talk will address questions regarding the representation and adequate handling of (predictive) uncertainty in (supervised) machine learning. A particular focus will be put on the distinction between two important types of uncertainty, often referred to as aleatoric and epistemic, and how to quantify these uncertainties in terms of appropriate numerical measures. Roughly speaking, while aleatoric uncertainty is due to the randomness inherent in the data generating process, epistemic uncertainty is caused by the learner's ignorance of the true underlying model.

Handheld Spectroscopy for Quality Assessment in the Chemical Industry

Sanwouly Marlene Yao
Radboud University

In industry, it is important to do a quick quality assessment to steer the process in terms of cost and sustainability, however, routine measurements in the lab are time and effort consuming. Handheld spectroscopy offers a promising solution for quick, non-invasive quality assessment in the chemical industry. Here we explore the use of handheld spectroscopic devices to predict physical properties of methylene diphenyl isocyanate (MDI), indicators of product quality that influence its performance and suitability for industrial applications. Various spectrometers were evaluated under different experimental conditions to determine their effectiveness. One of the NIR spectrometers emerged as the most suitable device for both colour and viscosity predictions. The optimal condiions for these measurements were found to depend on the modelling results as well as the operator input, with the recommended protocol being to scan in a dark environment using a round jar. Future work will focus on validating this protocol in real-world conditions, developing robust predictive models for field use, and training operators to ensure accurate implementation.

This project is co-funded by TKI-E&I with the supplementary grant 'TKI- Toeslag' for Topconsortia for Knowledge and Innovation (TKI's) of the Ministry of Economic Affairs and Climate Policy.

From Model-based to Hierarchical Ordination in Community Ecology

Bert van der Veen
Norwegian University of Science and Technology (NTNU)

The field of community ecology frequently deals with sparse multivariate datasets. These are sparse because organisms, such as plants or insects, tend to favor particular environmental conditions, leading to their occurrence in only a few locations. For the past 70 years, ordination methods have been the method of choice for analyzing such data.
In pursuit of an ideal method to construct a low-dimensional visualizations of community data, researchers have argued for or against particular methods, though these discussions slowed at the turn of the millennium. For the last three decades a consensus has remained in favour of certain methods. Recently, the development of new ordination techniques has diverged on multiple paths. There are machine learning methods on the one side, and model-based ordination methods on the other. This talk will focus on recent advances and ongoing challenges in developing model-based ordination methods that operate within a fully probabilistic framework, such as the Generalised Linear (Mixed) Models to which they are related.

Multivariate analysis with such statistical models holds great potential, not only for community ecology but also for many other fields. These models offer various advantages, including flexible tools for inference, the ability to account for data properties through changes in response distributions or by incorporating random effects, the use of residual diagnostics for validation, and straightforward forecasting. Perhaps most importantly, many different ordination methods exist within this same class. Nonetheless, many challenges persist in the development and application of these new methods, not in the least their slow-paced adoption.

Easy Analysis of Complex Data: Regularized Multigroup Approximate Exploratory Factor Analysis

Katrijn van Deun
Tilburg University

Exploring multi-group data for similarities and differences in the measurement model forms a substantial part of the research conducted in the behavioral and social sciences: Examples include studying measurement invariance of psychological scales over age or ethnic groups and comparing symptom correlations between different psychological disorders. Multigroup exploratory factor analysis is often the method of choice. Yet, currently available methods are restrictive in their use. First, these methods cannot handle complex data with small sample sizes relative to the number of variables while high-dimension low sample size data are more and more used as a result of digitalization (e.g., omics data or word counts obtained by text mining of tweets). Second, the use of existing software is often arduous.

Here, we propose a regularized approximate exploratory factor analysis method that addresses these issues by building on a strong computational framework: The resulting method yields solutions that are constrained to show simple structure and similarity of the loadings over groups when supported by the data. The minimal required input is restricted to the data and number of factors. In a simulation study we show the method considerably outperforms existing methods, also in the low-dimensional setting. How to use the method both in the low and high-dimensional setting is illustrated with empirical data.

Co-author: Trà Lê.

FRIDAY, 22 NOVEMBER 2024

Developing Locally Weighted Multiblock Approaches for Real Time Quality Prediction in Latent Variable Multivariate-based Process Monitoring

Marina Cocchi
Università degli Studi di Modena e Reggio Emilia

The presentation will illustrate the challenges of handling multiblock data for on-line quality prediction in a full-scale plant scenario, which can be extremely complex considering the data volume, diversity, as well as the additional sources of variance introduced because different products can be manufactured in the same production line at different times, by changing operational conditions and formulations without interrupting production. This is an ideal scenario to evaluate the use of multiblock data analysis methods, which can enhance data interpretation, visualization, and predictive performances.

In particular, recent work of our research group, concerning development of locally weighted multiblock methods (LW-MB), such as LW-MB Partial Least Squares and LW-Response-Oriented Sequential Alternation (ROSA), including also a robust version of LW-MB-PLS will be presented, paving attention to interpretation of local model outcome. Furthermore, evaluation of the incremental addition of data blocks to capture early estimation of product quality will be discussed.

Redeeming at Last the Classipedia Promise: Onset of a Conceptual Outline Map to Get a Hold on the Jungle of Cluster Analysis

Iven van Mechelen
KU Leuven

In my 2013 IFCS Presidential Address, I argued that, whereas clustering is alive and kicking as a research domain, the available clustering models, algorithms, and data-analytic techniques in their entirety form an inconvenient and intricate jungle. I further argued that this is a most problematic obstacle for developers of methods, for students who want to familiarize themselves with the domain, and for applied researchers. As a way to overcome this obstacle, I proposed to work out within IFCS a road map for the clustering domain, called Classipedia.

To redeem at last in part the Classipedia promise, most recently Christian Hennig, Henk Kiers, and I published a paper on an outline map for the clustering domain, based on an overarching conceptual framework and a common language (Van Mechelen, Hennig, & Kiers, 2024). With this outline map, my co-authors and I wanted to contribute to structuring the domain, to characterizing methods that have often been developed and studied in quite different contexts, to identifying links between methods, and to introducing a frame of reference for optimally setting up cluster analyses in data-analytic practice. In this talk, I will briefly introduce the overall scheme, comprising six interrelated angles, that acts as backbone of the outline map. Subsequently, I will highlight a few taxonomic distinctions within some of the angles, to give a better sense of the nature of our proposed approach. I will conclude by showcasing how the outline map can be used in practice with two illustrative examples.

Van Mechelen, I., Hennig, C., & Kiers, H. A. L. (2024). Onset of a conceptual outline map to get a hold on the jungle of cluster analysis. *WIREs Data Mining and Knowledge Discovery*, 14, e1547, 1-37. https://doi.org/10.1002/widm.1547

Robust Correlation Estimation with Discrete Rating-scale Data

Andreas Alfons
Erasmus University Rotterdam

Correlation estimation is often an important building block in the analysis of rating data, particularly for structural equation models. In this context, polychoric correlation is a frequently used model: it assumes latent normal variables such that thresholds for discretization are estimated together with the correlation of the latent variables. However, the commonly employed maximum likelihood (ML) estimator is highly susceptible to misspecification of the polychoric correlation model, for instance through violations of latent normality assumptions. We propose a novel estimator that is designed to be robust to partial misspecification of the polychoric model, that is, the model is only misspecified for an unknown fraction of observations, for instance (but not limited to) careless respondents. In contrast to existing literature, our estimator makes no assumption on the type or degree of model misspecification. It furthermore generalizes ML estimation and is consistent as well as asymptotically normally distributed. We demonstrate the robustness and practical usefulness of our estimator in simulation studies and an empirical application from a survey on the Big Five personality traits. In the latter, the polychoric correlation estimates of our estimator and ML differ substantially, which, after further inspection, is likely due to the presence of careless respondents that our estimator helps identify.

Co-authors: Max Welz and Patrick Mair.

Multivariate Six Sigma for Industry 4.0

Alberto Ferrer
Universitat Politècnica de València

In the digitalized industrial age not only has the amount of registered data increased dramatically, but also there has been a significant change in their nature. These data are mostly collected from daily production (i.e., happenstance data) and often exhibit high correlation, rank deficiency, low signal-to-noise ratio, and missing values.

Data Science projects without a scientific method approach and a proved problem – solving strategy are doomed to failure in Industry 4.0 when process understanding is needed for troubleshooting and process optimization.

Six Sigma has proven to be a successful problem-solving methodology for process improvement in the last decades. However, the traditional Six Sigma statistical toolkit, mainly composed of classical statistical techniques (e.g., scatter plots, correlation coefficients, hypothesis testing, and linear regression models from experimental designs), must be revisited to face the new challenges derived from Industry 4.0.

The upgrade of the Six Sigma toolkit with latent variables-based multivariate statistical techniques such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), widely used in process chemometrics, and machine learning tools, is essential for addressing the complex data characteristics in Industry 4.0, giving rise to the so-called Multivariate Six Sigma (Ferrer 2021).

In this talk, we will discuss several hot topics such as data quantity vs data quality, correlation vs causality, what type of models are useful in Industry 4.0 (predictive vs explanatory models), and illustrate the benefits of the Multivariate Six Sigma approach through several industrial real case studies.

Alberto Ferrer (2021) “Multivariate six sigma: A key improvement strategy in industry 4.0,” *Quality Engineering*, 33(4):758–763, DOI: 10.1080/08982112.2021.1957481

The Structural-after-measurement (SAM) Approach to SEM

Yves Rosseel
Ghent University

In structural equation modeling (SEM), the measurement and structural parts of the model are usually estimated simultaneously. But already since the birth of SEM in the '70s, various authors have advocated that we should first estimate the measurement part, and then estimate the structural part. We call this the Structural-After-Measurement (SAM) approach. In the first part of the presentation, I will give a brief historical overview of various SAM approaches, and discuss their advantages and disadvantages. Next, I will describe the so-called `local' SAM method where the mean vector and variance–covariance matrix of the latent variables are expressed as a function of the observed summary statistics and the parameters of the measurement model. The method includes two-step corrected standard errors and local fit measures. In the second part of the presentation, I will discuss several recent developments that are based on the SAM approach, including the inclusion of latent quadratic and interaction terms, the use of non-iterative estimators for the measurement part of the model, small-sample corrections, and various approaches to study measurement invariance in the setting where the number of groups is very large. Finally, I will discuss a software implementation of the SAM approach that is available in the R package lavaan.

Statistical Integration of Cross-sectional and Longitudinal Omics Datasets

Jeanine Houwing-Duistermaat
Radboud University

Multivariate methods such as O2PLS and P2PLS can be used for omics data integration. These methods decompose the outcome space into joint, data-specific and residual components. In this talk, we are interested in the joint components for two problems. Firstly, when integrating genetic markers and an omics dataset, the joint genetic component represents the genetic part of the omics features. Such an approach is an alternative for univariate approaches such as polygenic scores, which do not address the randomness and the heterogeneity in the genetic data. Secondly, for two longitudinally measured omics datasets, we propose to let the joint components capture the change over time.

To evaluate the methods, we performed simulations. The considered designs reflect the multivariate and the univariate model, and normally distributed and categorical variables. The software Cosi2 is used to obtain realistic genetic data. For the longitudinal modelling, we compare the longitudinal with a cross-sectional model where the latter ignores the changes over time. Finally, we apply the methods to several datasets. We estimate the genetic part from metabolomics data from the epidemiological study ORCADES. To identify dynamic metabolic components, we fit the longitudinal model to two metabolomics datasets measured at three timepoints from TwinsUK study.

It appears that multivariate statistical integration approaches are good alternatives for omics based polygenic scores and that the parameters of the longitudinal models are well estimated.

Co-authors: He Li and Said el Bouhaddani.