WORKSHOP: PLS regression and related component based methods in sensory science

Lead by Tormod Næs

Tormod Næs will give an introduction on component based methods, then there will be a presentation by the three participants before a discussion.

Participants: Marieke E. Timmerman, Harald Martens, John Castura

A taxonomy of linear models, and its use for empirical analysis - on the relationships between structural equation modeling, regression analysis, component analysis etc.

Marieke E. Timmerman

Psychometrics and Statistics, Heymans Institute for Psychological Research,
University of Groningen, the Netherlands; m.e.timmerman@rug.nl

The analysis of empirical data is both easier and more difficult than ever. The broad availability of useful models and accompanying software allows for any type of analysis needed. However, there are so many models around that it can be a difficult task to select a suitable model for the empirical data at hand. This task is greatly facilitated when one has proper insight into the characteristics of various models – and their mutual relationships. In this talk I will discuss a taxonomy of linear models, which encompasses regression analysis, component analysis, common factor analysis, structural equation modeling and mixture modeling. I will discuss key similarities and differences, and their mutual relationships by explaining them in terms of observed and latent variables. I will devote attention to modeling dependent variables of different natures (continuous, ordinal, categorical), and to rotational freedom within exploratory models. Further, I will highlight how this taxonomy is of use to select promising candidate models for an empirical data set at hand.

Interpretable machine learning with an eye for the physics:

Discovery, quantification and error detection in subspace models of everlasting, high-dimensional streams of Big Data

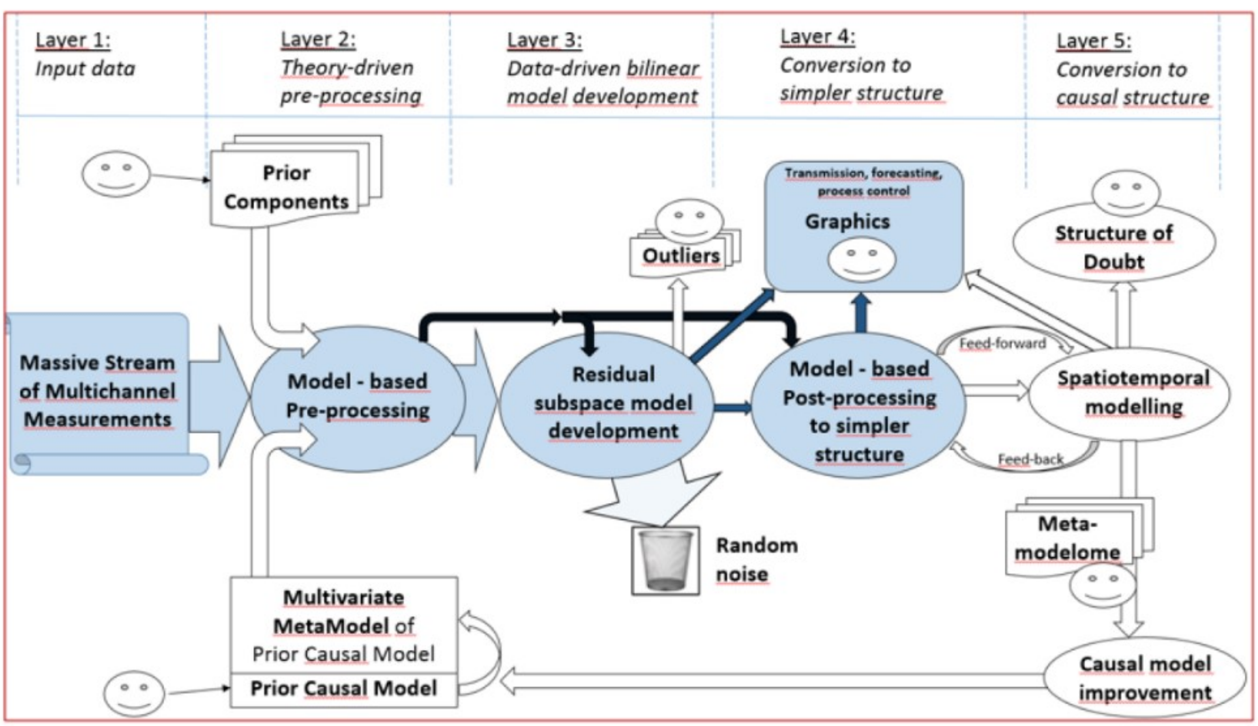
Harald Martens[1,2,3]

[1]Idletechs AS Trondheim Norway (www.idletechs.com, harald.martens@idletechs.com), [2]Dept. Engineering Cybernetics, Norwegian U. of Science and Technology NTNU, Trondheim, [3]School of Business, Macau U. of Science and Technology MUST, MACAU

Data is not information. The more data you have, the less information, unless you have adequate tools to interpret the data, in light of their own context and your own knowledge. The "standard Artificial Intelligence model" - based on unexplainable "black box" machine learning from Big Data - now receives massive criticism from many fields - especially concerning application in critical stages in medicine, industry, transport etc.

But Big Data themselves - at least wisely selected Quantitative Big Data from modern measuring devices- can be very informative: Modern multichannel sensors (thermal video cameras, hyperspectral cameras, microphones, chemical sensor arrays etc) simultaneously deliver tens of thousands of different variables (pixels, wavelength channels, vibration frequencies, chemical compounds), many times per second. How to interpret and utilize such "ever-lasting", overwhelming streams of high-dimensional Big Data in practice? Traditional bilinear subspace modelling methods, like PCA from psychometrics and PLSR from chemometrics, have spawned many modern extensions, in sensometrics, engineering cybernetics etc. These have been implemented in software tools that people can now use in practice, to quantify known phenomena from data, discover unexpected patterns, gain system overview and get early warnings for dangerous developments. Bilinear subspace modelling of such Big Data streams gives efficient summaries of the main information content, with locally simple structure suitable for human interpretation and for discovery and modelling of causality chains, feed-back mechanisms and process dynamics.

This lecture outlines one practical implementation of such a tool that helps ordinary people to interpret and use Big Data in e.g. industry and shipping, without having to take an education in mathematics, statistics or computer science first. A software system combining thermal monitoring, subspace data modelling, multivariate computer graphics based on cognitive science allows operators to explore, understand and control a relatively complex industrial system wrt its expected and unexpected types of variation. A five-layered framework will then be proposed for interpretable «deep learning» - dynamic, open-ended but with respect for both the laws of physics and people's practical experience:



Person, product, context, choice: Connecting consumers' motivations and perceptions with their preferences

J. C. Castura

Compusense Inc., Guelph, Ontario, Canada

One of the secret service agents in John Le Carré's *Tinker, Tailor, Soldier, Spy* makes the following assertion: "The more identities a man has, the more they express the person they conceal." The same can be said of the data that arises from the ordered, labelled categorical scales that are commonly used in sensory evaluation. Data obtained from these scales (nine-point hedonic scale, the five-point purchase intent scale, etc.) can be thought of as having multiple identities. Although technically incorrect, they are often treated as interval data, summarized by a mean and a standard deviation. If results are thought of as having a discrete categorical or a binary outcome, they can be summarized by crosstabulations: Top Box data, Top 2 Box data, etc. Cumulative crosstabulations treat the data as ordinal. Scale data can also be considered to be ranking data, or as another type of binary outcome, e.g., ranked first or ranked last.

Standardizing nine-point hedonic scale data weights each consumer equally. Patterns in the original data are preserved, but standardized data are disconnected from the original scale anchors; for example, a standardized preference of -1 might for one consumer refer to a sample that is liked moderately, but for another consumer refer to a sample that is disliked moderately. Discretizing data into a Top-2 box response focuses on which products elicit a strongly positive liking response (1) vs. any other response (0) regardless of whether that other response is one of disliking or a relatively tepid positive response. Each treatments of the data gives a different perspective.

How do relationships between consumers' motivations, preferences, and choice behaviours results differ when we consider ordered, labelled categorical data in different ways? Data arising from two large-scale consumer studies to PLSR along with other approaches to explore this question. Collectively these analyses provide perspectives that express what the data conceals about the relationship between person, product, context, choice.