Clustering and data reduction models for three-way preference data

Maurizio Vichi

University of Rome "La Sapienza", Dep. Statistics, Probability and Applied Statistics,
P.le A. Moro 5, I-00185, Rome, Italy

Three-way data are the result of the observation of a *multivariate-multioccasion* phenomenon, characterized by a set **X** of *nkr* values of *k* variables, measured on a set of *n* units, in *r* different occasions. The elements of the *three-way data set* **X** are arranged into an array with respect to units, variables and occasions. Frequently, **X** refers to the same set of units and variables, but different occasions. In other cases, **X** is a set of classifications, i.e. distance matrices with order constraints on the triplets, observed in different occasions on the same set of units. In many sensometrics studies, preference data are collected in one of the two forms of **X**. For example, in the last three Sensometrics meetings, such arrays related to different steamed potatoes or grape/raspberry beverages or Riesling wines (units), by regarding a set of sensory variables, estimated by a set of trained assessors (occasions) have been statistically analyzed. Other experiments require to directly classify products according to assessed preferences. Open questions in the last three Sensometrics meetings were: what are the sensory differences among products? Are consumers uniform in their liking and disliking of products? What is the relation between sensory and instrumental data observed in addition to **X**? Appropriate answers to these questions are acquired only by using methodologies that take into account the multi-way structure of the data, mainly by using clustering and data reduction models for **X**. Differences between products (units) or consumers (occasions) can be evaluated by fitting clustering models (e.g., partitions, hierarchies, fuzzy classifications) to **X**. Note that a proper clustering of **X** is a more complex problem than the classical multivariate case. In fact, each assessor actually induces a personal classification of products according to the sensory variables and a global classification of products must represent a consensus classification of those induced by each assessor. Similarly, each product induces a different classification of the assessors that needs to be synthesized by a consensus. Methodologies have been proposed that fit a clustering model to **X** using a least-square approach [1, 2].

Relations between variables are examined by three-way factorial models. Here the aim is to find virtual sensory variables (factors associated to observed variables) and virtual assessors (factors associated to occasions) to synthesize the relations between sensory variables and between occasions. An inferential approach that simultaneously classifies units and reduces variables and occasions is proposed. It is assumed that the data to be clustered are sampled from a finite mixture of Gaussians, i.e, each observation is a realization of a mixture density, where the components correspond to underlying clusters. The clusters-constrained covariance matrix depends on occasions and variables following a direct product model. This allows to decompose the within group variability into the parts due to variables and occasions, respectively. The mean vectors of the clusters are constrained to lie in a reduced subspace according to a Tucker model. In this way the latent factors for variables and occasions that best explain the between variability of the data are identified. [3].

References

[1] Gordon, A.D., Vichi, M. (1998). Partitions of Partitions, *J. of Classification*, Springer-Verlag, New York, 15, 265-285.

[2] Vichi, M. (1999). One mode classification of a three-way data set, *J. of Classification*, Springer-Verlag, New York, 16, 27- 44.
[3] Rocci, R. Vichi, M (2001). Mixture Models for Simultaneous Reduction and Classification. Submitted.