

Application of approaches to assess the value of the evidence of multivariate data

Sonja Menges

Four recently discussed methods which have been developed to assess the value of the evidence for multivariate continuous data are applied to existing data. Two of them make use of statistic significance tests, the multiple performance of the univariate t-test on the one hand and the application of Hotelling's T^2 -test on the other hand. The other two methods are based on the evaluation of a likelihood ratio. The underlying probabilistic model takes two types of variation into account, that within sources and that between sources. The between source variability is both modelled by a multivariate normal distribution and a multivariate kernel density estimate. In order to compare the performances of these methods, they are applied to glass data. The BKA provides favourable conditions since several databases are available here. They consist of measurements of element concentrations of float glasses which result from laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) analysis.

Two datasets are used as population databases. One contains element concentrations of 63 float glasses which had been collected from different manufacturers whereas the other one consists of real case data. A third database contains the element concentrations of 33 samples that are taken from different parts of one window and are hence known to come from the same source. For each sample of the databases, six replicate measurements are made.

The above-mentioned methods are applied to sets of multivariate data consisting of different numbers of components. The element concentrations taken into consideration are chosen from 19 measured elements. For each selection, the performance of the approaches is compared and assessed in the following way. On the one hand, pairs of samples that are known to come from the same source are compared. For this purpose, the six replicate measurements are split up in two groups of three members each and for each sample of the underlying population database, the corresponding groups of triple measurements are compared. Furthermore, the methods are applied to pairs of samples taken from the window database. In this case, the other two databases act as reference populations when performing the likelihood ratio approach. On the other hand, all pairs of samples belonging to the underlying population database are compared in order to get results for groups of data that are known to come from different sources.

Besides the comparison of the different approaches, it is also a future aim to study the relationship between the number of components and the size of the database that is used to estimate the distribution of the characteristics in a reference population within the likelihood ratio approach. This seems to be an important question if it is possible – as in the case of glass elemental analysis evidence – to select from a high number of components. It is planned to base this study on computer simulations.

Apart from applications to glass data, we are trying to find concepts for using the multivariate likelihood ratio approach in the field of forensic speaker recognition with respect to long-term distribution data. Instead of applying it to the fundamental frequency F0, the formants F2 and F3 for which data of 70 speakers exist will be the focus of attention, though, it is not clear yet how to determine the parameters.