B140 Validation of Computer-Assisted Classification of Trace Evidence Data: Lies, Damned Lies, and Multivariate Statistics

Stephen L. Morgan, PhD, Department of Chemistry & Biochemistry, University of South Carolina, Columbia, SC 29208; and Edward G. Bartick, PhD, FBI Laboratory, Counterterrorism & Forensic Science Research Unit, FBI Academy, Quantico, VA 22135*

After attending this presentation, attendees will gain a basic understanding of multivariate statistical methods such as principal component analysis and linear discriminant analysis; the application of these techniques to the discrimination of forensic data; an understanding of when to use (and not use) these methods; issues in the validation of computer-assisted data analysis applied to forensic problems.

This presentation will impact the forensic community and/or humanity by increasing understanding of the potential benefits and misuses of multivariate statistics in the interpretation of trace evidence.

Identification of patterns in analytical chemical data and interpretation of observed differences is a frequent task for forensic chemists. A fiber examiner might perform UV/visible microspectrophotometry on known and questioned fibers to evaluate possible associations between source and location. An arson investigator might compare headspace gas chromatograms taken from questioned debris at the scene of a suspicious fire with chromatograms of known accelerants. Most forensic scientists are familiar with statistics associated with a single type of measurement variable, for example, the refractive index of questioned glass fragments might be measured for a comparison with known window glass fragments from a crime scene. The resulting set of sample data values have only one variable (refractive index) measured for a number of different objects (different glass fragments). Statistics for summarizing such *univariate* measurements, including the sample mean and the sample standard deviation, and hypothesis testing, such as the two-sample *t*-test for means, is commonly used. However, spectroscopic, chromatographic, or mass spectrometric instruments routinely produce megabytes of data of high dimensionality from a single analytical run on a single forensic sample. Scientists are less familiar with *multivariate* statistics based on measurements of two or more variables for each sample. Real differences in multivariate data might not show up as the presence or absence of a single peak, or as inflation or reduction of single peaks. A *combination* of small changes could be the key to recognizing a significant difference. Trace evidence found at a crime scene may also involve mixtures of several chemical components. How should the similarity of complex patterns be evaluated? Are the patterns similar enough to be considered to have originated from similar source materials? Can differences be explained simply by sampling variability from the original source or by experimental variability in the laboratory? How can outliers be recognized? Increases in computing power have made computationally intensive data analysis feasible, and the increased availability of software for multivariate statistics has made these techniques accessible.

The scope of this presentation includes the techniques of principal component analysis (PCA) and linear discriminant analysis (LDA). Examples of forensic applications will include arson investigation, analysis of polymer trace evidence, and forensic fiber examination. PCA and LDA can be employed as exploratory tools to detect and to visualize patterns and as predictive tools to classify and discriminate among data from different analytical samples. The statistical significance of differences and similarities can be assessed, and the reliability of discrimination among groups of samples can be evaluated using multivariate statistics. A summary of practical guidelines for use and interpretation will be presented along with examples of misuse and steps that should be taken to validate such computer assisted data analysis.

Trace Evidence, Statistics, Validation

 ** Presenting Author*