MODELLING PEAK INTENSITIES FROM MASS SPECTROMETRY PROTEOMIC PROFILES USING ZERO-MODIFIED DISTRIBUTIONS

J.H. Barrett, D.A. Cairns, R.E. Banks

University of Leeds, Leeds, UK

Email: *j.h.barrett@leeds.ac.uk*

Intensities for an individual peak from mass spectrometry (MS) proteomic profiling of a group of samples are usually modelled as following a normal distribution, possibly after transformation, or alternatively reduced to a binary variable. There are generally a number of samples where the peak is absent (intensity zero), and it is common practice to include imputed values for these observations, with little apparent justification. We propose that the data can be more accurately described by a mixture of a (log)-normal distribution and a block of zeroes, from samples where the peak is absent (the associated protein or protein fragment is not expressed). We discuss the statistical properties of the zero-modified (log)-normal distribution and demonstrate goodness-of-fit to peak intensities from a study of renal cancer based on surface-enhanced laser desorption/ionization (SELDI) MS profiling of serum samples. Within this framework groups of samples (e.g. cases and controls) can be compared using likelihood ratio tests, either assuming homogeneity of variance or allowing for differences in variance between groups. In this data set more peaks are identified as being differentially expressed using the new approach than using two-sample t-tests, and the method allows one to identify the nature of the difference between sample groups. We use simulation to further evaluate performance in terms of power and false positive rate in comparison with standard approaches.