Multivariate Bayesian variable selection for NIR spectra applied to biscuit dough pieces

P. J. Brown*, M. Vannucci† and T. Fearn‡

SUMMARY

We are concerned with predicting the composition of biscuit dough pieces in the baking industry. The eventual aim is to implement rapid on-line quality control of biscuit production using Near Infrared spectroscopy to measure the composition of the dough pieces, i.e. the unbaked biscuits. This measurement makes it possible to detect problems with the metering and mixing of ingredients at early stage. The data arise from an experiment to test the feasibility of the approach. We have available 300 Near Infrared reflectance measurements on each sample of dough. There are 40 samples in the training data, and a further 40 available for validation. In this test situation in both sets we know the actual composition by weight of the ingredients used in the formulation.

Our application uses an extension of the decision theory approach to regressor variable selection of Lindley(1968) to both a multivariate response and informative prior distributions. This involves two terms, a prediction mean squares error and an additive cost penalty for the number of wavelength variables included. It is necessary to carefully specify the cost per variable added as well as the broad prior structure of the data. The application is also multivariate, involving four constituents of the dough pieces (Fat, Sucrose, Dry Flour, Water). Rather than focus on one constituent for prediction at the expense of others we choose a relative weighting of the responses and seek to optimise for the worst predicted response.

Because of the large number of variables involved (2^{300} possible subsets) simulated annealing is used to maximise the expected utility. Our implementation involved a steadly reduction of "temperature" and reheating at the first optimum encountered. The stochastic search used in the application involves deleting, adding and swapping moves. Great computational savings are achieved by employing fast updating algorithms.

*University of Kent at Canterbury, Institute of Mathematics and Statistics, Cornwallis Building, Canterbury, CT2 7NF, UK

†University of Kent at Canterbury

‡University College London, UK

The small set of selected variables leads to commendable accuracy in terms of prediction for each of the four responses. Looking at the Bayes coefficients of the 11 chosen wavelengths, it can be seen that they divide up into subsets which are close to wavelengths traditionally chosen to identify the separate ingredients.