Bull. Aust. Math. Soc. (First published online 2025), page 1 of 2*
doi:10.1017/S0004972725100294
*Provisional—final page numbers to be inserted when paper edition is published

VARIABLE SELECTION IN MULTIVARIATE DATA BY ANALYSIS OF DATA PATTERN

LUCY CONRAN

(Received 2 June 2025)

2020 *Mathematics subject classification*: 62H99.

Keywords and phrases: multivariate analysis, variable selection, data pattern, density based clustering, canonical correlation analysis.

In this thesis, we develop statistical methods that are focused on isolating which variables influence a prediction, cause change or discriminate between groups. We examine variables of multivariate data and split our work into two parts based on the structure of the data under analysis.

Part I. Variable selection for data with more observations than variables

We worked with flow cytometry experts to develop methods that isolate sets of interesting variables in flow cytometry data. Our variables are biomarkers and our aim is to identify which variables discriminate between two or more classes. For example, how does the expression of a biomarker change with infection by a specific virus? The data to which we apply our methods are collected from human blood samples by a flow cytometer and have many more observations than variables. Our approach is to simultaneously create density based cluster patterns in two different but related datasets, and estimate the distance between the clusters and modes as a proxy for 'distance' between the underlying datasets. We expand our methods with our new method, Multi-SOPHE, to isolate the variables that differentiate most, that is, which subset of variables exhibits most change between the clusters of the datasets under consideration.

Part II. Variable selection for data with more variables than observations

Proteomics is the large-scale study of proteins, peptides and ions. We worked with proteomics experts to expand and develop methods that isolate which ions and proteins

Thesis submitted to the University of Western Australia (UWA) in December 2024; degree approved on 24 March 2025; supervisors Berwin Turlach (UWA) and Inge Koch (RMIT).

https://doi.org/10.1017/S0004972725100294 Published online by Cambridge University Press

provide accurate classification of tissue samples. For example, into tumour or normal, or into whether the tumour has metastasised or not. Our data are collected by matrix assisted laser desorption ionisation mass spectrometry imaging and have more than 171 000 variables (our variables are ion masses). Our aim is to find a subset of these variables that is capable of predicting the class of the underlying tissue sample. Our approach is to use canonical correlation analysis to select a small subset of variables that correlate with a change in class label. We use this reduced set of variables to classify tissue samples. In this way, we isolate ion masses that are the best predictors of tissue class and therefore pivotal in cancer growth.

LUCY CONRAN, School of Physics, Mathematics and Computing,
University of Western Australia, Crawley, Western Australia 6009, Australia
e-mail: lucyconran@gmail.com

https://doi.org/10.1017/S0004972725100294 Published online by Cambridge University Press