Clustering of Proteomics Imaging Mass Spectrometry Data

Annie Conway

Thesis submitted for the degree of

Master of Philosophy

in

Statistics

at

The University of Adelaide

Faculty of Mathematical and Computer Sciences

School of Mathematical Sciences

January 12, 2016

Contents

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

SIGNED: .... DATE: ....

Acknowledgements

I give my foremost thanks to my supervisor Inge Koch for her helpful advice and teaching, as well as her unwavering support and encouragement.

I would also like to thank Lyron Winderbaum and the Adelaide Proteomics Centre for providing data, assistance and information. I acknowledge The University of Adelaide for my scholarship and The School of Mathematical Sciences for giving additional support as well as providing me with a place to work.

Finally, I would like to thank my friends and family for their support and encouragement. Many people were a source of insightful conversation which was valuable for both my research and understanding of mathematics.

Abstract

This thesis presents a toolbox for the exploratory analysis of multivariate data, in particular proteomics imaging mass spectrometry data. Typically such data consist of 15000 - 20000 spectra with a spatial component, and for each spectrum ion intensities are recorded at specific masses. Clustering is a focus of this thesis, with discussion of k-means clustering and clustering with principal component analysis (PCA). Theoretical results relating PCA and clustering are given based on Ding and He (2004), and detailed and corrected proofs of the authors' results are presented. The benefits of transformations prior to clustering of the data are explored. Transformations include normalisation, peak intensity correction (PIC), binary and log transformations. A number of techniques for comparing different clustering results are also discussed and these include set based comparisons with the Jaccard distance, an information based criterion (variation of information), point-pair comparisons (Rand index) and a modified version of the prediction strength of Tibshirani and Walther (2005).

These exploratory analyses are applied to imaging mass spectrometry data taken from patients with ovarian cancer. The data are taken from slices of cancerous tissue. The analyses in this thesis are primarily focused on data from one patient, with some techniques demonstrated on other patients for comparison.

Notation Index

$\mathbb{R}$	the real numbers.
p	number of variables.
n	number of observations.
x	(observation of) a functional random variable.
$\mathbf{x}$	(observation of) a multivariate random variable, $p \times 1$.
$\mathbb{X}$	data matrix, $p \times n$.
$\boldsymbol{\mu}$	the mean of $\mathbf{x}$, $p \times 1$.
Σ	the covariance matrix of $\mathbf{x}$, $p \times p$.
$\bar{\mathbf{x}}$	the sample mean of $\mathbb{X}$, $p \times 1$.
S	the sample covariance matrix of $\mathbb{X}$, $p \times p$.
$d(\mathbf{x}_1, \mathbf{x}_2)$	distance between two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.
$\delta(C_1, C_2)$	distance between two sets C_1 and C_2.
k	number of clusters.
$\mathcal{C} = \{C_1, \ldots, C_k\}$	a k-cluster arrangement.
$\mathcal{P}(\mathbb{X})$	the power set of $\mathbb{X}$, i.e. the set of all subsets of $\mathbb{X}$.
m/z	mass-on-charge.