Statistical analysis of proteomic mass spectrometry data for the identification of biomarkers and disease diagnosis

Tyman Stanford

Thesis submitted for the degree of
Doctor of Philosophy in Statistics at
The University of Adelaide

October 30, 2015

Discipline of Statistics
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

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: ....

Contents

Signed statement i

Acknowledgements xii

Abstract xiii

Acronyms and abbreviations xv

1 Disease, proteins and mass spectrometry **1**

1.1 Motivation . . . 2

1.2 Biological background . . . 4

1.2.1 Amino acids and proteins . . . 4

1.2.2 Proteins via genes . . . 5

1.2.3 Gene and protein expression . . . 7

1.2.4 Protein function, sera and cancer . . . 7

1.2.5 Proteomics . . . 8

1.2.6 Biomarker discovery . . . 11

1.2.7 Linear TOF-MS . . . 13

1.2.8 Ionisation of proteins . . . 14

1.3 MS Data . . . 17

1.3.1 Synthetic data . . . 18

1.3.2 Proteomic MS for cancer classification . . . 19

1.3.3 Adam et al. (2002) . . . 20

1.3.4 de Noo et al. (2006) . . . 22

1.3.5 Asthma studies . . . 23

1.3.6 Gastric cancer mice study . . . 26

1.4 Summary of data . . . 29

2 Methods of intra-spectra pre-processing **33**

2.1 Pre-processing step I: signal smoothing . . . 36

2.1.1 Savitzky-Golay smoothing . . . 37

2.1.2 Comparison to moving-average filters . . . 38

2.1.3 Further considerations . . . 39

2.2 Pre-processing step II: baseline correction 40
2.2.1 Morphological image analysis and theory 41
2.2.2 Implementation 44
2.2.3 A more efficient implementation 46
2.2.4 Towards automated baseline correction 49
2.2.5 Morphological analysis for unequally spaced values 51
2.2.6 The novel continuous line segment algorithm 52
2.2.7 Comparison of piecewise and continuous baseline subtraction . 58
2.3 Recommendations 62

3 Methods of inter-spectra pre-processing 63
3.1 Pre-processing step III: spectra normalisation 64
3.1.1 Total ion current normalisation 66
3.1.2 Empirical quantile normalisation 66
3.1.3 Cyclic LOESS normalisation 67
3.1.4 Evaluating methods of normalisation 70
3.2 Pre-processing step IV(a): peak detection 74
3.3 Pre-processing step IV(b): peak alignment 78
3.3.1 Dynamic programming 81
3.3.2 Modifications required for MALDI/SELDI TOF-MS 85
3.3.3 Implementation and calibration 91
3.3.4 Comparison to standard peak alignment techniques 94
3.4 Recommendations 96

4 Data visualisation, intermediate analysis and biomarker discovery 98
4.1 Unsupervised exploratory analysis 99
4.1.1 Principal component analysis 99
4.1.2 Missing values 100
4.2 Using linear models to identify potential biomarkers 108
4.2.1 Linear modelling of the GC mice data 108
4.2.2 Regression on the other peak expression data 120
4.3 GC mice: the effect of missingness on statistical inference 125
4.3.1 Missing peak expression as an outcome 127
4.3.2 Linear models with imputed data 130
4.4 Unknown and unwanted variation 134
4.4.1 Surrogate variable analysis 136
4.4.2 Remove unwanted variation 140
4.5 GC mice biomarker candidate summary 145

5 Statistical and computational methods of classification 148
5.1 Supervised learning in the context of diagnostics 149
5.2 Unbiased error prediction 150
5.3 Statistical classification 153

5.3.1 Linear discriminant analysis . . . 153
5.3.2 Quadratic discriminant analysis . . . 155
5.4 Extending statistical classification for SnLp problems . . . 156
5.4.1 Pairwise fusion discriminant analysis . . . 156
5.4.2 Regularised discriminant analysis . . . 158
5.5 Computational classification . . . 158
5.5.1 RandomForests . . . 159
5.5.2 Support Vector Machines . . . 162
5.6 Feature Selection . . . 166
5.6.1 Fisher Score . . . 167
5.6.2 Classification with in-built feature selection . . . 168
5.6.3 RandomForests . . . 168
5.6.4 Pareto Fronts . . . 168
5.7 Summary . . . 174

6 Classification results 176
6.1 Towards optimal discrimination . . . 177
6.1.1 A standardised approach . . . 178
6.2 Optimising discrimination of the GC mice dataset . . . 181
6.2.1 Averaging over peak expression vectors . . . 181
6.2.2 Threshold of replicates to achieve successful discrimination . . . 183
6.2.3 Individual mice and their classification . . . 187
6.3 Comparison of data handling . . . 190
6.3.1 Change in basis of data . . . 190
6.3.2 Treatment of missing values . . . 191
6.3.3 Peak expression quantification . . . 195
6.4 Comparison of feature selection and classification methods . . . 196
6.4.1 Feature selection . . . 197
6.4.2 Classification model . . . 199
6.5 The effects of spectra normalisation on classification . . . 200
6.6 Comparison of proteomic data results . . . 202
6.7 Summary . . . 206

7 Concluding remarks 209

A R code 211

B Continuous line segment algorithm proof 214

C Biomarker investigation supplementary information 218

D Probabilistic LDA classification of GC mice 232

Bibliography 235

List of Figures

1.1 Elemental structure of an amino acid. 5

1.2 The components of blood. 9

1.3 The context of biomarker discovery. 10

1.4 An example of a mass spectrum (m/z vs intensity): a serum derived SELDI TOF-MS raw spectrum from the Adam et al. (2002) study. . 12

1.5 An illustration of the basic construct of TOF-MS. 13

1.6 A schematic of a MALDI TOF-MS system. 15

1.7 An example spectrum from the Morris et al. (2005) generated data. . 19

1.8 Schematic of location and run order of samples on the MALDI Bruker Daltronics Anchorchip for the asthma2 data. 26

1.9 Scatter plot of the median absolute deviation of spectra against the natural logarithm of total signal detected for each spectrum in the asthma2 dataset. 27

1.10 Experimental design for the GC mice data. 29

2.1 Sources of false signal in MALDI/SELDI TOF-MS data. 34

2.2 Pipeline of MALDI/SELDI TOF-MS pre-processing and analysis. . . 35

2.3 A randomly selected raw spectrum from the GC mice data on a subset of the m/z-axis with overlays of Savitzky-Golay and moving-average smoothing techniques. 39

2.4 An example of ϵ_B, ω_B and τ_B on a set f. 44

2.5 A spectrum from the asthma1 dataset demonstrating the baseline estimates in the signal using the top-hat operator, spline and LOESS methods. 46

2.6 A random selection of five spectra from the GC mice dataset and their respective peak widths of detected peaks. 50

2.7 A random selection of ten spectra represented by different colours from the GC mice dataset and their respective peak widths of detected peaks. 51

2.8 An example of data where the x values are unevenly spaced points in $\mathbb{R}$ for a SE of size $k = 3$ and the point $x = 2.44$. 52

2.9 An example of data where $\theta_{w_i^{\triangledown}} = \theta_{w_i^{\vartriangle}+1}$ and the computation required for the continuous line segment algorithm. 55

2.10 An example of data where $\theta_{w_i^{\triangledown}-1} = \theta_{w_i^{\vartriangle}}$ and the computation required for the continuous line segment algorithm. 56
2.11 Computation time of calculating the morphological opening for randomly generated data in R. 58
2.12 Baseline correction on a randomly selected spectrum at the low end of m/z-values from the GC mice dataset using piecewise top-hat and transformed TOF-values with constant SE width and the continuous definition of morphological operators. 59
2.13 Baseline correction on a randomly selected spectrum at an intermediary section of m/z-values using piecewise top-hat and transformed TOF-values with constant SE width and the continuous definition of morphological operators. 60
2.14 Baseline correction on a randomly selected spectrum at the high end of m/z-values using piecewise top-hat and transformed TOF-values with constant SE width and the continuous definition of morphological operators. 61

3.1 Randomly selected spectra from the Adam et al. (2002) dataset. . . . 65
3.2 An example empirical quantile normalisation on two spectra with intensities at three m/z-values. 67
3.3 Identical MA plots of two spectra from the Adam et al. (2002) data. 68
3.4 The effect of MA-LOESS adjustment on two spectra. 70
3.5 Density and histogram plots of total ion current for asthma2 spectra over the 4000-7000 subsection of the m/z-axis with constant x-axes. . 72
3.6 Heat map plots of MALDI chip location total ion current for asthma2 spectra over the 4000-7000 subsection of the m/z-axis using (a) TCN, (b) EQN and (c) CLN. 75
3.7 Simple S2N ratio peak detection on the Adam et al. (2002) data (patient 20) over a subset of the m/z-axis. 76
3.8 An illustration of peak width against peak location using TOF-values and m/z-values; peaks are generally wider for larger molecular masses. 77
3.9 An estimated non-linear regression Gaussian curve fitted to the intensities satisfying $f(x_i) \geq \frac{1}{2}a_f$. 78
3.10 W matrix for the amalgamated alignment of two 4-alignments of the de Noo et al. (2006) dataset. 84
3.11 A calibrated T matrix for two spectra from the Adam et al. (2002) data. The four heatmaps below are S matrices using different metrics. . . . 87
3.12 Alignment amalgamation de Noo et al. (2006) spectra. 89
3.13 Peak alignment example for two spectra along a subset of the m/z-axis. 90
3.14 A subset of the GC mice spectra over the 8788-9907Da subinterval of the m/z domain; a potential alignment. 93
3.15 Alignment heatmaps of two datasets and two alignment methods for comparison. 95

4.1 PCA plots of GC mice peak expression data with disease classification group ($k = 1, 2, \ldots, 5$) and chip ($j = 1, 2, 3$) labelled. 102
4.2 PCA plots of GC mice peak expression data with labels for chip only ($j = 1, 2, 3$). 104
4.3 PCA plots of GC mice peak expression from chip 1 only with run-order information. The 360 PCA points are plotted in a random order irrespective of chip run-order to avoid a visual bias from plotting points in run-order. 105
4.4 PCA plots of peak expression intensities of the Adam et al. (2002) dataset. 106
4.5 PCA plots of peak expression intensities of the de Noo et al. (2006) dataset. 107
4.6 Experimental design of GC mice dataset with linear mixed effect model annotation. 109
4.7 GC mice peak expression data as a heatmap. 110
4.8 Theoretical and empirical pairwise correlation structure of peak expression for each set of 27 spectra for each mouse in the GC mice dataset. 112
4.9 Proportional random effect contributions to the variance for each of the 159 LME peak models for (a) the four-level model of Equation (4.1) and (b) the three-level model of Equation (4.6). 117
4.10 Volcano plot of phenotype group differences for each peak of the GC mice expression data; adjusted p-value vs. fold change on the $\log_2$-scale. 119
4.11 Parallel plot of the GC mice peak expression data for peaks identified in Figure 4.10. 120
4.12 Volcano plot for the de Noo et al. (2006) dataset for group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. . 122
4.13 Volcano plot for the Adam et al. (2002) dataset for group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. . 123
4.14 Parallel plots of peak expression on the $\log_2$-scale relative to the model-estimated control group effect for each peak identified in the volcano plot in Figures 4.12 and 4.13. 124
4.15 Scatter plot of the mean $\log_2$ expression against proportionate missingness for each peak in the GC mice dataset. 126
4.16 Parameter estimates for identified potential biomarkers, relative to the WT group, for the LME models when missing values are ignored and when the missing values are kNN imputed. 132
4.17 Volcano plot of phenotype group differences for each peak of the GC mice expression data with kNN imputed data. 133
4.18 Peak expression data as a heatmap for the 12 peaks satisfying a fold change of 1.5 and significant group effect when modelled (a) with no imputation and (b) using kNN imputed data. 135

4.19 Plot of estimated number of surrogate variables, h, found by the permutation p-values of eigenvalues method for the peak expression datasets. 139
4.20 Plots of GC mice peak expression data for increasing peak location on the y-axis. 142
4.21 PCA plot of RUV-rep adjusted GC mice peak expression data with disease group ($k = 1, 2, \ldots, 5$) and chip ($j = 1, 2, 3$) labelled for principal components one and three. 144

5.1 The two-class supervised learning problem. 149
5.2 Error prediction process using GFCV. 151
5.3 An example of a binary classification tree for two-class data, successively splitting data towards terminal nodes. 159
5.4 An illustration of the construction of a classification tree for two-class data in a RandomForest. 161
5.5 The expression for two different features in a two-class problem. . . . 167
5.6 Feature importance calculation example for a single RandomForest tree. 169
5.7 An example with two criteria ξ_1, ξ_2 to be maximised and the first four Pareto Fronts. 171
5.8 A receiver operator curve showing the sensitivity and specificity of the feature ranking methods in selecting the 25 differentially expressed features and the cumulative proportion of truly differentiated features selected by the feature ranking methods. 174

6.1 Discrimination results for GC mice dataset using different levels of averaging of the peak expression vectors. 182
6.2 Discrimination results for the GC mice dataset for varying replication in the training data (C8 averaged). 184
6.3 Modified sensitivity and specificity plots for replicates on GC mice data. 186
6.4 Heatmap of the SVM predictive error for GC mice dataset for different peak expression vector averaging schemes: (a) none, (b) C8, (c) aliquot and (d) mouse. 188
6.5 Relationship between SVM prediction certainty and prediction outcome for the C8 averaged data. 190
6.6 Discrimination results for GC mice and Adam et al. (2002) datasets using varying data transformation methods. 191
6.7 Discrimination results for GC mice and Adam et al. (2002) datasets using different imputation methods. 193
6.8 Modified sensitivity and specificity plots for different missing value imputation methods on the GC mice data. 194

6.9 Discrimination results for GC mice and Adam et al. (2002) datasets using varying peak quantification methods. 196
6.10 Discrimination results for GC mice and Adam et al. (2002) datasets using varying feature selection methods. 197
6.11 Number of features selected for the final predictive model in the 100 iterations. 198
6.12 Discrimination results for GC mice and Adam et al. (2002) datasets using different classification methods. 200
6.13 Discrimination results for GC mice and Adam et al. (2002) datasets using different spectra normalisation methods. 201
6.14 The final discrimination results using SVM and LDA classification. . 202
6.15 Confusion matrices as heatmaps of the average test predictive error for the GC and Adam et al. (2002) datasets. 205

C.1 PCA plots of GC mice peak expression data by chip 2 run-order. . . . 219
C.2 PCA plots of GC mice peak expression data by chip 3 run-order. . . . 220
C.3 PCA plots of peak expression intensities of the asthma1 dataset by group labels. 221
C.4 PCA plots of peak expression intensities of the asthma2 dataset by group labels. 222
C.5 Statistics to assess different random effects model structures. 223
C.6 Schematic of the three-level LME model structure for the GC mice data of the form $\mathbf{Y}_p = \mathbf{X}\boldsymbol{\beta}_p + \mathbf{Z}\mathbf{B}_p + \boldsymbol{\epsilon}_p$. 224
C.7 Parameter estimates, relative to the WT group, for the LME models when missing values are ignored and when the missing values are kNN imputed. 227
C.8 Volcano plots for asthma1 dataset relating to group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. 230
C.9 Volcano plots for asthma2 dataset relating to group differences peak expressions; adjusted p-value vs. fold change on the $\log_2$-scale. 231

D.1 Heatmap of the LDA predictive error for GC mice dataset for different peak expression vector averaging schemes: (a) none, (b) C8, (c) aliquot and (d) mouse. 233
D.2 Relationship between prediction certainty and prediction outcome for the C8 averaged GC mice data using LDA classification. 234

List of Tables

1.1 Summary of subjects in the Adam et al. (2002) study. 20
1.2 Demographic information for the de Noo et al. (2006) study. 23
1.3 Classification results for the de Noo et al. (2006) study. 23
1.4 Groups in the GC mice data. 28
1.5 Summary of experimental groups for the datasets. 30
1.6 Summary of important features of the MS datasets. 32

2.1 Computation time of top-hat operator using different code on generated data. 48

3.1 CV_r value using different normalisation methods and data. 73
3.2 Mean misclassification proportion (bootstrapped 95% confidence interval of the mean) using the normalisation methods TCN and EQN. 73
3.3 Dynamic programming parameters. 91
3.4 Harmonic means of pairwise dynamic programming and vote alignment on randomly sampled Adam et al. (2002) spectra. 94

4.1 Top five peaks contributing to the first three principal components of the GC mice peak expression data. 103
4.2 GEE GLM modelling of the binary outcome of missingness in the GC mice peak expression dataset using the predictors of disease group and chip. 129
4.3 Pipeline for using the RUV-rep method on the GC mice data. 143
4.4 GEE GLM modelling of the binary outcome of missingness in the GC mice peak expression dataset using the predictors of disease group and chip. 146

5.1 Probability of an individual observation being sampled for a tree in a RandomForest. 160
5.2 Summary of attributes of discrimination methods used. 175

6.1 The median and 25th percentile sensitivity, specificity and area under the curve for the GC mice data using varying numbers of peak expression vector replicates in the training data (model creation) for test dataset prediction. 187
6.2 The interquartile range of the number of features used in the test data prediction model for the Adam et al. (2002) and GC mice datasets. . 199
6.3 Proportional representation of peaks in the 100 final classification models used to classify the GC mice test data, restricted to peaks occurring in 80 or more models. 204
6.4 The step-wise approach of Chapter 6 to 'optimise' the discrimination of classification groups. 208

C.1 Peaks with significant group difference and fold changes of at least $\frac{3}{2}$ in the GC mice peak expression dataset. 225
C.2 Missingness observed for peak expressions for peaks 8337, 8607 and $12161 m/z$ for GC mice dataset. 226
C.3 Peaks with significant group difference and fold changes of at least 2 in the de Noo et al. (2006) peak expression dataset. 228
C.4 Peaks with significant group difference and fold changes of at least $\frac{3}{2}$ in the Adam et al. (2002) peak expression dataset. 229
C.5 Peaks with fold changes of at least $\frac{3}{2}$ in the asthma1 peak expression dataset. 230
C.6 Peaks with fold changes of at least $\frac{3}{2}$ in the asthma2 peak expression dataset. 231

Acknowledgements

I must emphasise my gratitude to my supervisors, Professor Patty Solomon and Dr Chris Bagley. Patty, I have tremendous admiration for your statistical knowledge and thank you for your fantastic insights, guidance and wisdom along the way. Chris, your incredibly sharp eye combined with your patience when answering my questions is thoroughly appreciated.

Mum, Dad, Mel and Liana, I hope we can do away with the “are you done yet?” question. Thank you to you all for your encouragement from near and far, recently and formerly. I am under no illusions that without your combined support I would not be writing this now. I will be a better son, brother and partner now I promise. I extended this to friends who have been very understanding of my absenteeism. Thank you for your support as well.

Thank you to Chris Davies also for his initial work in his honours thesis that allowed me to hit the ground running. Many thanks to The University of Adelaide, specifically to the School of Mathematical Sciences and those I have had the most contact with at the Adelaide Proteomics Centre: Megan Penno, Vicki Clifton and Peter Hoffmann.

Since I have the floor, there are some more general sentiments I would like to make. I am grateful to exist in the time and location I do, and to be able to do what I love. My exclamation of “what a time to be alive!” is rarely sarcastic, albeit a poor attempt at humour. It would be remiss of me not to reference ‘standing on the shoulders of giants’ (but to complete the metaphor, in my case, rather than standing I might be sitting or even sliding off). I also wish to thank others that I have not met, those who get insufficient acknowledgement for what is a tremendous service to society: people that create and maintain publicly available software. Particularly the authors and contributors of R and TeX/LaTeX, software I have used extensively in this thesis.

Abstract

Proteomic spectra obtained from matrix-assisted laser desorption ionisation (MALDI) time-of-flight mass spectrometry (TOF-MS) are generated from the proteins and peptides present in serum obtained from blood. By ionising the proteins and resolving them in the mass spectrometer, data on the expression of proteins can be obtained, realised from the amplitude of signal for different mass to charge ratios. Of primary interest is the biological signal, in particular, the expression of proteins related to disease. In common with many 'omic' technologies, the raw spectra suffer from systematic errors due to technological artefacts and batch-effects, in addition to sample and biological variability. To negate these effects, novel application of genetic microarray pre-processing and analysis methods to proteomic TOF-MS data are presented. However, there are important differences between microarray and TOF-MS data which require consideration and non-trivial modifications to be successfully applied. One important difference between MALDI TOF-MS data and other high-throughput data, seldom addressed, is the high proportion of missing values.

The pre-processing of raw proteomic TOF-MS data needs to be undertaken prior to analysis and remains a mathematical and statistical challenge. Performed in distinct steps, pre-processing consists of signal smoothing, baseline correction, spectra normalisation, peak detection and peak alignment. An argument as to why the order of these steps is highly important is presented. Standard and novel data pre-processing methods are investigated and compared to optimise the process. Each step is given due consideration since the cumulative effects of substandard pre-processing can render subsequent statistical analysis highly unreliable.

Ultimately, the aim of proteomic MS is to analyse the protein profiles. Two different but related approaches to the analysis are undertaken. The first approach is to identify biological markers (biomarkers) that exhibit differential expression between disease groups. Identifying potential biomarkers for further research requires appropriate exploratory, visual and statistical modelling which is addressed in detail here. The second approach is to perform statistical discrimination between groups, a classical supervised learning problem. The ability of mathematical models to predict disease groups using differential biological signal provides insight into the plausibility of diagnostic tests. Methodologically, supervised learning is a multifaceted problem given that feature selection, model parameter optimisation, and the handling of the training and test data all contribute to the inference that can be made from the results. Empirical appraisal of the methods applied to the proteomic data are provided with the outcome of discrimination error as a quantitative benchmark.

A number of proteomic TOF-MS datasets with differing characteristics are used throughout this thesis to assess the validity of the methods presented. The detailed analysis of a murine model MALDI TOF-MS dataset has facilitated the discovery of potential biomarkers for gastric cancer. Correct classification of spectra to their respective disease group (gastric cancer or control mice) as high as 97.4% was achieved using supervised learning. The thorough treatment of all the differently behaved datasets contained in this thesis, starting from the raw data pre-processing steps through to the challenging process of identifying potential biomarkers, provides a comprehensive and best-practice pipeline to analyse real-world proteomic MS data.

Acronyms and abbreviations

For simplicity, many abbreviations will be used throughout this thesis. The abbreviation/acronym will appear in parentheses at the first occurrence of the phrase but the table below provides a comprehensive list for quick reference.

Abbreviation	Meaning
APC	Adelaide Proteomics Centre
C	The portable and compiled programming language
C8 beads	Alkyl group beads used in proteomic sample fractionation
CLSA	Continuous line segment algorithm
CLN	Cyclic LOESS normalisation
CRC	Colorectal cancer
CV	Coefficient of variation
(k)Da	(kilo)Daltons; $^1/_{12}{}^{\text{th}}$ of a carbon-12 atom's mass ($\sim 1.7 \times 10^{-27}$kg)
DNA	Deoxyribonucleic acid
DP	Dynamic programming
EQN	Empirical quantile normalisation
FDR	False discovery rate
FS	Fisher score
FWHM	Full-width at half-maximum
GC	Gastric cancer
GC-MS	Gas chromatography-mass spectrometry
GEE	Generalised estimating equation
GFCV	G-fold cross-validation, traditionally denoted k-fold
GLM	Generalized linear model
HM	Harmonic mean
IMAC-Cu	Immobilised metal affinity chromatography - copper
kNN	k-nearest neighbours
LC-MS	Liquid chromatography?mass spectrometry
LDA	Linear discriminant analysis
LME	Linear mixed effects
LOESS	Locally weighted scatterplot smoothing (local regression)
LSA	Line segment algorithm

Abbreviation	Meaning
MA	A transformation of paired minus vs. average log intensities
MAR	Missing at random
MALDI	Matrix-assisted laser desorption/ionisation
MCAR	Missing completely at random
MS	Mass spectrometry
m/z	Mass divided by charge: the x-axis of TOF-MS
μm	Micrometre (10^{-6} metres)
Nd:YAG	Neodymium-yttrium aluminium garnet (laser)
n_k	The number of patients/subjects in $k = 1, \dots, K$ groups
nm	Nanometre (10^{-9} metres)
NW	Needleman and Wunsch (algorithm)
OOB	Out-of-bag
OLS	Ordinary linear least-squares (regression)
PC	Prostate cancer
PCA	Principal component analysis
PF	Pareto Front
PFDA	Pairwise fusion discriminant analysis
PLS	Penalised least squares (regression)
pH	Acidity/akalinity scale; hydrogen ion concentration metric
pmol/μL	Molecular concentration/microlitre; pmol $\approx 6 \times 10^{11}$ molecules
QDA	Quadratic discriminant analysis
R	The statistical programming environment
RDA	Regularised discriminant analysis
REML	Restricted maximum likelihood
RF	RandomForest
RNA	Ribonucleic acid
RUV	Remove unwanted variation
S2N	Signal to noise (ratio)
SAX	Strong anion exchange
SE	Structuring element
SELDI	Surface-enhanced laser desorption/ionisation
S-G	Savitzky-Golay
SnLp	Small-n Large-p (problem)
SVA	Surrogate variable analysis
SVD	Singular value decomposition
SVM	Support vector machine
SW	Smith and Waterman (algorithm)
TCN	TIC normalisation
TIC	Total ion current
TOF	Time-of-flight
$\mathrm{T_x}$	Treatment
UV	Ultraviolet
WCX	Weak cation exchange