



### Chapter 1 Introduction

- 1.0 Background 1
- 1.1 Problems/Challenges 16
- 1.2 Reproducibility Issues 31
- 1.3 Data Collection Outline 49

### Chapter 2 Scaling

- 2.0 Introduction 57
- 2.1 Mean Centering 60
- 2.2 Variance Scaling / Autoscaling 62
- 2.3 Scale Row Area / Integral Normalization 63
- 2.4 Pareto Scaling 64
- 2.5 Range Scaling 66
- 2.6 Level Scaling 67
- 2.7 Log Transformation 68
- 2.8 Power Transformation - Square Root 69
- 2.9 Discretization 70
- 2.10 Generalized Log Transform 73
- 2.11 Probabilistic Quotient Normalization (PQN) 75
- 2.12 Variable Stability Scaling (VAST) 78
- 2.13 Bucketing/Binning 79
- 2.14 Histogram Matching (HM) 83
- 2.15 Orthogonal Signal Correction (OSC) 86
- 2.16 Digitization 89
- 2.17 Multiplicative Scatter Correction (MSC) 90
- 2.18 Standard Normal Variates (SNV) 92

### Chapter 3 Preprocessing

- 3.0 Introduction 95
- 3.1 Zero Filling (NMR) 96
- 3.2 Windowing/Filtering (NMR) 97
- 3.3 Signal-to-Noise Ratio (SNR) 99
- 3.4 Noise reduction and differentiation 101
- 3.5 Savitzky-Golay 101
- 3.6 Differentiation 103
- 3.7 Smoothing 106

The best preprocessing methods will be the ones that ultimately produce a robust model with the most accurate predictive ability. Unfortunately, there are no particularly straightforward rules to guide investigators to the best selection of preprocessing options; the subsequent trial and error optimization process may be quite time consuming and confusing. However, spending little or no time investigating preprocessing options is likely to result in less than optimal results.

The primary objective of this book is to present a relatively focused outline of the major options available for data analysis, with an emphasis on the advantages and disadvantages of the techniques discussed.

Shipping weight : 650g  
414 pages

# Data Preprocessing For Chemometric and Metabonomic Analysis

ISBN 978-1-926825-61-8

**MRi Consulting**

8 Wilmot Street  
Kingston, Ontario CANADA K7L 4V1  
info@chemometrics-analysis.com

**MRi Consulting**

Tel: (+1) 613-541-0609

**To purchase go to:**

**www.chemometrics-analysis.com**

- 3.8 Moving Averages 106
- 3.9 Heteroscedastic / Homoscedastic Noise 109
- 3.10 Phasing 112
- 3.11 Baseline 112
- 3.12 Linear Regression Baseline Fitting 115
- 3.13 Two Point Linear Baseline 116
- 3.14 Baseline correction, function fit baseline 116
- 3.15 Asymmetric least squares baseline fit 118
- 3.16 airPLS 119
- 3.17 Baseline Offset 120
- 3.18 Mass Spectrometry 122
- 3.19 Wavelets 132
- 3.19.1 Continuous Wavelet Transform (CWT) 133
- 3.19.2 Scaling 135
- 3.19.3 Shifting 136
- 3.19.4 Discrete Wavelet Transform (DWT) 136
- 3.19.5 Approximations and Details 137
- 3.19.6 Multiple-Level Decomposition 138
- 3.19.7 Dimension Reduction 141
- 3.19.8 Baseline/Background Correction 144
- 3.19.9 Denoising 146
- 3.19.10 FT Denoising 146
- 3.19.11 Wavelet Denoising 147
- 3.20 Peak Alignment 154
- 3.20.1 Warping 157
- 3.20.2 Peak alignment by Fast Fourier Transform 158
- 3.20.3 Recursive alignment Fast Fourier Transform 159
- 3.20.4 Recursive Segment-wise Peak Alignment 160
- 3.20.5 Generalized Fuzzy Hough transform (GFHT) 161
- 3.20.6 Dynamic Time Warping (DTW) 165
- 3.20.7 Parametric Time Warping (PTW) 166
- 3.20.8 Correlation Optimized Warping (COW) 168
- 3.20.9 Peak Alignment - Reduced Set Mapping 169
- 3.20.10 Partial Linear Fit (PLF) 170
- 3.20.11 Peak Alignment by a Genetic Algorithm 171
- 3.20.12 Beam Search 173
- 3.20.13 iCOSHIFT 176
- 3.20.14 FFT Cross Correlation 178
- 3.20.15 Progressive Consensus Alignment NMR 179

#### Chapter 4 Sample Subset Selection 181

- 4.0 Introduction 181
- 4.1 Sample size recommendations 182
- 4.2 Representativity 184
- 4.3 Median Absolute Deviation (MAD) 187
- 4.4 Dixon's Test 188
- 4.5 Grubbs Test 189
- 4.6 Cochran test 190
- 4.7 Constituent Value Range 190
- 4.8 General Notes 191
- 4.9 Overview of Sample Subset Selection Options 191
- 4.10 Random Subsampling 193

- 4.11 Bootstrapping 193
- 4.12 Cross Validation 195
- 4.13 Mahalanobis distance (MD) 196
- 4.14 Kennard-Stone (KS) 201
- 4.15 Duplex Method 203
- 4.16 Sample Set Partitioning X–Y distances (SPXY) 204
- 4.17 Rank Select 206
- 4.18 Kohonen Neural Networks 206

#### Chapter 5 Variable Subset Selection 217

- 5.0 Introduction 217
- 5.1 Missing Values 217
- 5.2 Imputation Methods 220
- 5.3 Multiple Imputation 222
- 5.4 Why variable selection 223
- 5.5 Chance 225
- 5.6 Generalizability 226
- 5.7 Bias 227
- 5.8 Filter Methods 230
- 5.9 Wrapper Methods 231
- 5.10 Embedded Methods 232
- 5.11 Exhaustive Methods 233
- 5.12 Information Leak 233
- 5.13 Cross Validation 235
- 5.14 Variable Selection by Stepwise Algorithms 238
- 5.15 Sequential Forward Floating Selection (SFFS) 238
- 5.16 Variable Selection Stepwise Regression 239
- 5.17 F-Test 240
- 5.18 t-Test 242
- 5.19 Fisher Index / Coomans Index 244
- 5.20  $\chi^2$ -Test 244
- 5.21 Kolmogorov–Smirnov Test 246
- 5.22 Wilcoxon Rank Sum Test 248
- 5.23 Analysis of Variance (ANOVA) 252
- 5.24 SELECT 256
- 5.25 Simple Variable Reduction 257
- 5.26 Selection of Univariate Tests 259
- 5.27 Simple Pairwise Correlation Method 260
- 5.28 KIF Index Method 260
- 5.29 VIF Method 261
- 5.30 B2 And B4 Methods 262
- 5.31 First Eigenvector Method 263
- 5.32 Overlap Density Heatmap (ODH) 264
- 5.33 Successive Projections Algorithm (SPA) 269
- 5.34 Linear Discriminant Analysis (LDA) 272
- 5.35 Uncorrelated Linear Discriminant Analysis (ULDA) 276
- 5.36 Principal Component Analysis (PCA) 276
- 5.37 Partial Least Squares (PLS) 282
- 5.38 Interval PCA (iPCA) and PLS (iPLS) 286
- 5.39 Interval PLS (iPLS) 287
- 5.40 Moving Window PLS (mwPLS) 289
- 5.41 Backward Interval PLS (biPLS) 290
- 5.42 Synergy Interval PLS (siPLS) 292

- 5.43 Variable Importance in Projection (VIP) 294
- 5.44 Non-Orthogonalized PLS1 (IFRNOPLS) 295
- 5.45 Outer Product Analysis PLS DA 296
- 5.46 PLS Uninformative Variable Elimination 299
- 5.47 Partial Least Squares Genetic Algorithm 301
- 5.48 Linear Discriminant Analysis-Genetic Algorithm 309
- 5.49 Mutual Information (MI) 312
- 5.50 Particle Swarm Optimization (PSO) 313
- 5.51 Relief 316
- 5.52 Decision Trees 318
- 5.53 CART 319
- 5.54 Recursive Feature Elimination (RFE) 322
- 5.55 Naïve Bayesian Belief Network (BBN) 323
- 5.56 Support Vector Machines (SVM) 328
- 5.57 Ant Colony Optimization (ACO) 334
- 5.58 Minimum Redundancy–Maximum Relevance 336
- 5.59 Neural Networks (NN) 340
- 5.60 Back-Propagation Neural Network (BP-NN) 341
- 5.61 Probabilistic Neural Networks (PNN) 347
- 5.62 Random Forest (RF) 351
- 5.63 Independent Component Analysis (ICA) 355
- 5.64 Random Permutations 362
- 5.65 Correlation-Based Feature Selection (CFS) 363
- 5.66 Fast Correlation Based Feature Selection (FCBF) 364
- 5.67 Simulated Annealing (SA) 366
- 5.68 Multidimensional Scaling (MDS) 368
- 5.69 Stochastic Proximity Embedding (SPE) 368
- 5.70 Isomap 369
- 5.71 Fast Maximum Variance Unfolding (FastMVU) 370
- 5.72 Kernel PCA (KPCA) / kernel PLS 370
- 5.73 Generalized Discriminant Analysis (GDA) 372
- 5.74 Diffusion Maps (DM) 373
- 5.75 Stochastic Neighbor Embedding (SNE) 374
- 5.76 Local Linear Embedding (LLE) 374
- 5.77 Laplacian Eigenmaps (LE) 375
- 5.78 Hessian LLE (HLLE) 375
- 5.79 Local Tangent Space Analysis (LTSA) 375
- 5.80 Conformal Eigenmaps (CCA) 376
- 5.81 Maximum Variance Unfolding (MVU) 376
- 5.82 Linearity Preserving Projection (LPP) 377
- 5.83 Neighborhood Preserving Embedding (NPE) 377
- 5.84 Locally Linear Coordination (LLC) 377
- 5.85 Manifold Charting (MC) 378
- 5.86 Coordinated Factor Analysis (CFA) 378
- 5.87 Stochastic Neighbor Embedding (SNE) 378
- 5.88 Evaluation Criteria 382

- List of Abbreviations 392
- Appendix 1 Test Data 395
- Appendix 2 Software 1 399
- Appendix 3 Software 2 413