Contents

Acknowledgments ii

Abstract v

Contents xvii

I Introduction and literature review 1

1 Introduction, contents and aims of the thesis 3

1.1 Introduction 3
1.2 Characteristics of 'omic' data 5

2 Common strategies for the analysis of metabolomic data 9

2.1 Classical bivariate tests 9
2.2 Linear modelling of omic data 12
2.3 Corrections for multiple comparisons 13
2.4 Principal Component analysis 16

3 Modern techniques for the analysis of metabolomic data 19

3.1 Projection based methods 19
3.2 Penalization methods 22
3.3 Other methods 29

4 Challenges for analyzing metabolomics data. Strategies and aims of the thesis 31

4.1 Challenges for analyzing metabolomic data. 31
4.2 Strategies and aims of the thesis. 32

II Materials and methods 35

5 Multiway data, methods for its analysis 37

5.1 From 2D matrices to 3D arrays 37
5.2 Exploration and description of N-way arrays. 38
5.3 Regression methods for N-way arrays, N-PLS. 41
5.4 Software for analyzing three-way data 43

6 Variable selection techniques 45

6.1 VIP scores 45
6.2 Selectivity Ratio 48
6.3 Variable selection through L1 penalization 50

III Results 55

7 Sparse N-PLS, a method for variable selection in multiway data sets 57

7.1 Sparse N-PLS, integration of L1 penalization in the N-PLS algorithm 57
7.2 Hyperparameters of the sparse N-PLS algorithm. 58

8 sNPLS package, a comprehensive software for N-PLS and sparse N-PLS analysis 63

8.1 Implementation of the sNPLS algorithm in R 63

8.2 The sNPLS package 70

9 Validation of the sparse *N*-PLS method and sNPLS package 85

9.1 Random synthetic data sets 85

9.2 Data driven synthetic data sets 88

9.3 Real data sets 90

9.4 Comparison with other methods 96

IV Conclussions and future work 107

10 Conclusions 109

10.1 Developed topics and contributions 109

10.2 Future work and research lines 110

Bibliography 115