

IMPERIAL COLLEGE LONDON

DOCTORAL THESIS

Statistical Approaches for Metabolomics and Omics Data Integration

Author:

Takoua JENDOUBI
BEDHIAFI

Supervisors:

Prof. Korbinian
STRIMMER
Dr. Timothy M.D.
EBBELS
Prof. Robert C. GLEN
Dr. Marc-Emmanuel
DUMAS

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy*

in the

Department of Epidemiology and Biostatistics
School of Public Health

November 2, 2018

I herewith certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed:

Date:

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

I herewith certify that all material in this dissertation which is not my own work has been properly acknowledged.

Takoua Jendoubi

Abstract

Biological processes are the result of multiple interactions between various omic entities and are inherently complex. Metabolomics profiling plays a key role into deciphering mechanisms of biological functions in living organisms and is hence gaining popularity. In the last twenty years, the parallel acquisition of high-throughput datasets from the genome, metabolome, proteome, and transcriptome has seen a tremendous boost. The integrative analysis of these datasets is promising to enhance the understanding of biological functions and uncover their underlying mechanisms.

The main objectives of this thesis consist in i) developing and investigating novel statistical models for integrative analysis of metabolomics data with other omics technologies, ii) enriching the offer of probabilistic models tailored to metabolomics data and iii) providing enhanced interpretability of results. This thesis is mainly concerned with designing models that can be introduced in different steps of a typical analysis pipeline of metabolomics data. Chapters 2 and 3 motivate the importance of data integration and review popular statistical techniques used in metabolomics. Chapter 4 simultaneously covers two steps of the analysis pipeline, by building a single integrative Bayesian model that is able to perform both cross-omics biomarker discovery and infer potential perturbed pathways. Chapter 5 focuses solely on integrative statistical analysis by uncovering hidden associations between multi-omics data. Finally, in Chapter 6 we investigate the incorporation of pathway information into a Bayesian nonparametric clustering model and its potential to help metabolite annotation.

Where possible, simulation studies are used to get a better understanding of our methods and test their applicability. These simulations are always followed by analysis of real data and comparison to competing methods. In most instances, our methods have resulted in plausible biological findings when applied to real data, and represent, to our knowledge, one of the first applications of such probabilistic models in integrative analysis of metabolomics data.

Acknowledgements

I would like exceptionally to thank Dr Korbinian Strimmer and Dr Timothy Ebbels for their valuable expert guidance and patience during my PhD research and all my supervisors for their continuous support.

I would like to acknowledge Panagiotis Vorkas for providing the time-course metabolomics data used in Chapter 4 and his expert advice on MS, Ibrahim Karaman and Goncalo Correia for their patient help in understanding NMR and MS data.

My financial support is acknowledged from a Wellcome Trust ISSF PhD studentship.

I am very grateful to my Stratigrad friends with whom I spent great escapades from the UK to Japan and many other places. I am particularly grateful to Nazanin for sharing my random nonsense during those escapades, for showing me support as well as motivating me.

I am unbelievably grateful to my family, in particular my mother Hedia, grandmother Janette and grandfather Ali, without whom I never would have got this far, for raising me to be the person I am now. I am appreciative of my husband's love and understanding throughout this PhD journey. Lastly, but by no means least, I am incredibly grateful to Jana, my daughter for inspiring my curiosity with her imagination, for enlightening my life with her shining smile and for showing me my strengths.

Dedicated to my beloved Jana, Hedia and Janette

Contents

Abstract	5
Acknowledgements	6
1 Introduction	27
1.1 Overview and motivation	27
1.2 Summary of contributions	32
2 Biological background and omics technologies	35
2.1 Basics of molecular biology	35
2.2 The omics cascade and omics technologies	39
2.3 Metabolomics	41
2.3.1 ¹ H Nuclear Magnetic Resonance Spectroscopy	43
2.3.2 Liquid Chromatography Mass Spectrometry	45
2.3.3 Data preprocessing	46
2.3.4 Metabolite annotation	49
2.3.5 Metabolic pathways	52
2.4 Summary and discussion	54
3 Statistical multi-omics data integration	57
3.1 Multi-omics data integration and challenges	57
3.2 Statistical background and gaps	65
3.2.1 Bayesian inference	66
3.2.2 Integrative regression	67
3.2.3 Latent variable dimension reduction methods for data in- tegration	69
3.2.4 Integrative clustering	72
3.3 Summary and discussion	74
4 Integrative Longitudinal Analysis and Metabolic Biomarker Dis- covery	79

4.1	Background	79
4.2	Model Framework	84
4.2.1	Metabolite dependencies	84
4.2.2	Integrative analysis	89
4.2.3	Experimental design	91
4.3	Results	93
4.3.1	Simulation study	94
4.3.2	Case study	99
4.4	Implementation	109
4.5	Discussion	112
5	A new approach for canonical correlation analysis and correlation-based network model	115
5.1	Background	115
5.2	A whitening approach for probabilistic canonical correlation analysis	116
5.2.1	Classical CCA	117
5.2.2	Whitening transformations and CCA	119
5.2.3	CCA and latent variable multivariate regression	123
5.3	Entropy-based criteria for multivariate association	127
5.3.1	The RV coefficient	127
5.3.2	Approach	128
5.4	Results	132
5.4.1	Synthetic data	132
5.4.2	Nutrimouse data	134
5.4.3	NCI-60 microRNA data	140
5.4.4	The Cancer Genome Atlas LUSC data	144
5.5	Software	147
5.5.1	Computational aspects	147
5.5.2	Implementation in R	150
5.6	Conclusions	150
6	Investigating infinite functional clustering for metabolite annotation	153
6.1	Introduction	153
6.2	Background	155
6.3	Infinite functional clustering model for metabolite annotation	160

6.3.1	The HDP mixture model	160
6.3.2	Modeling pathway information	161
6.4	Inference	164
6.5	Results	165
6.5.1	NCI-60 metabolomics data	167
6.5.2	Nutrimouse data	171
6.6	Discussion	180
7	Conclusion	185
A	Appendix: iCARH model	191
A.1	Inferring missing values in iCARH	191
A.2	Sensitivity analysis	192
B	Appendix: CCA whitening approach	195
B.1	Induction of correlation by a shared latent variable	195
B.2	Synthetic data : Assessment of estimated canonical correlations	196
B.3	Analysis of nutrimouse data	198
C	Appendix: Investigating infinite functional clustering for metabo- lite annotation	203
C.1	Finite Gaussian mixture models	203
C.2	Latent class model and the Dirichlet process	205
C.3	Hierarchical Dirichlet process and collapsed Gibbs sampling . . .	208
	Bibliography	211

List of Figures

2.1	Simplistic representation of the central dogma in molecular biology. When a cell divides DNA is replicated such that genetic information is maintained in the resulting cell. Transcription concerns transmission of genetic information from DNA to RNA whilst translation is the process of protein synthesis based on RNA coding. RNA can also be replicated (RNA replication) or used to create DNA by reverse transcription.	36
2.2	The omics cascade starting with the level of the genes to the last level of the metabolites triggering a particular phenotype (adapted from Haukaas et al. (2017)). For example, type 2 diabetes mellitus is a multi-factorial disease resulting from the breakdown of the insulin secretion pathway and the insulin sensitivity pathway along with some genetic (and environmental) factors.	40
2.3	Representative analysis pipeline of metabolomics data	43
2.4	Example of a typical NMR spectrum of a rat urine sample with some annotated metabolites (Beckonert et al., 2007). All peaks are referenced to the resonance of TSP (sodium salt of 3-trimethylsilylpropionic acid) at 0 ppm. Sharp peaks indicate presence of low weight molecular compounds whilst broad ones indicate presence of larger molecules. Some aromatic signals in the 6-9 ppm region are low but they can be resolved. The asterisk refers to a suppressed water signal.	44

- 2.5 Example of a LC-MS spectrum of a mix of 7 proteins obtained from Sashimi data repository (file 7MIX_STD_110802_1 from <http://sashimi.sourceforge.net/repository.html>). The x-axis of a mass spectrum represents the m/z values denoting the mass-to-charge ratio formed by dividing the mass of an ion by the unified atomic mass unit and by its charge number (positive absolute value). The y-axis represents retention times of the ions. The z-axis represents signal intensities of the ions. 45
- 2.6 Transformation of phosphoenol-pyruvate into pyruvate (irreversible reaction) and pyruvate into L-lactate (reversible reaction). A substrate is the chemical substance that generates a product. For example, in the irreversible reaction, phosphoenol-pyruvate is the substrate whilst pyruvate is the product. Each reaction is catalyzed by an enzyme encoded here by its KEGG sequence 2.7.1.40 and 1.1.1.27 respectively. If the enzyme is inhibited, the reaction rate is low and the substrate tends to accumulate. 52
- 3.1 Examples of a multi-stage integrative analysis approach. Example (a) illustrates a three-step framework where genomic and metabolomic datasets are concurrently tested for association with the phenotype resulting in smaller datasets. These datasets are then investigated to infer linked variables. Example (b) illustrates a typical scenario where genomic variables are tested for association with transcripts which are in turn associated to metabolotypes. These metabolotypes might for instance explain the expression of phenotype. 59
- 3.2 Different data integration strategies. Figure (a) illustrates early integration where data is combined into a single data matrix before modeling. Figure (b) depicts the intermediate data integration level where data matrices are transformed or mapped into a common meaningful representation before modeling. In figure (c), each data model is generated separately and is then combined with models based on other data types to generate the integrative or high-level model. 61

4.1	Figure from Walther et al. (2010). The figure shows that pairwise correlation levels (y axis) decrease when the shortest pathway distance between metabolites increase for different temperature settings.	85
4.2	Prior $p(\kappa_{mk} \tau, \sigma_{\beta_m})$ on κ_{mk} for different values of σ_{β_m} and τ . The prior distribution skews towards 1 if τ increases or σ_{β_m} decreases.	90
4.3	Left : Boxplots of Area Under the Curve (AUC) for pathway perturbation inference with uniform prior on ϕ compared to a beta like prior on ϕ for 10 simulated datasets. We infer perturbations based on the posterior distribution of $\phi^{\text{controls}} - \phi^{\text{cases}}$. Right : Distribution of $\phi^{\text{controls}} - \phi^{\text{cases}}$ under the uniform prior (white) and the beta like prior (grey). True perturbed pathways are printed in bold on the x axis.	94
4.4	Average Receiver Operating Characteristic (ROC) curves for pathway perturbation inference across 10 datasets for different fractions of “wrongly” assigned metabolites.	95
4.5	Left: Average Receiver Operating Characteristic (ROC) curves reflecting the effect of pathway size across 10 datasets for 18 % of “wrongly” assigned metabolites. Pathway size is the average of the number of metabolites per pathway for each simulation for a total of 11 pathways. Right: Average Receiver Operating Characteristic (ROC) curves reflecting the effect of the number of pathways across 10 datasets for 18 % of “wrongly” assigned metabolites.	97
4.6	Assessing the horseshoe like prior behavior : boxplots for estimated β coefficients (black) compared to true simulated coefficients (red). The horseshoe like prior performs well in terms of shrinking the noise terms to zero ($\kappa_{km} \approx 1$) and preserving the high terms ($\kappa_{km} \approx 0$).	98

4.7	Metformin data: Figure shows boxplots of correlations between metabolites that are part of the same pathway (intra-pathway correlations) compared to correlations between metabolites that belong to different pathways (inter-pathway correlations). Although correlations are not very close to 1, there is an empirically observed correlation signature related to whether metabolites are in the same pathway or not. In fact average correlations of intra-pathway metabolites are higher than average correlations of inter-pathway metabolites.	101
4.8	Metformin data: Posterior predictive checks for mean absolute deviation (MAD) compared to DPPCA for different numbers of metabolites used for inference. The MAD decreases as the number of metabolites increases. Our model performs clearly better than the DPPCA model.	102
4.9	Metformin data: Absolute mean error of estimated metabolic profiles using our model (red) and the SME model (green) for each metabolite (right) and on average (left). Our model performs in a similar way to SME which is indicative of a good fit.	102
4.10	Metformin data: Right and left panels show model fit assessment for controls and cases data. Left : quantile-quantile normal plot of $\Psi_{\text{cases}}^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$. Right : quantile-quantile normal plot of $\Psi_{\text{controls}}^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$	103
4.11	Estimates of effects of treatment on metabolite profiles are captured by β_m^α . Only part of the data is plotted as we are mainly interested in the feature “LP-Neg_0295.2270_00.96” which is a hydroxy octadecadienoic acid.	104
4.12	Metformin data: Posterior distributions of ϕ^e for each pathway for each treatment group. Posterior distributions of ϕ^e are very similar for both controls and cases which is indicative of no -or mild- pathway alterations. These pathways correspond as well to the top 11 pathways inferred by MetaboAnalyst (Chong et al., 2018). The p-values are the p-values returned by MetaboAnalyst for each pathway.	105

4.13	Metformin data: Pathway inference using sparse group PLS. For each pathway, bars indicate whether or not the pathway was selected in the first (blue) or second (red) component. Non selected pathways indicate that all related genes correspond to zero weights in a given component. “Glycerophospholipid metabolism” and “GPI-anchor biosynthesis” were not selected in both components whilst other pathways were selected at least once.	106
4.14	Metformin data: Estimates of effects of bacteria on metabolite profiles in iCARH are captured by β_m	108
4.15	iCARH’s simple pipeline and functions. The package also provides the <code>iCARH.simulate</code> function used to simulate data presented in Section 4.3. This function allows for different setting parameters.	109
5.1	Probabilistic CCA as a two layer latent variable generative model. The middle layer contains the CCA-whitened variables $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$, and the top layer the uncorrelated generative latent variables $\mathbf{Z}^{\mathbf{X}}$, $\mathbf{Z}^{\mathbf{Y}}$, and $\mathbf{Z}^{\text{shared}}$	126
5.2	Top: Percentage of estimated canonical correlations with correctly identified signs in dependence of the sample size and the strength of the true canonical correlation. Bottom: Percentage of estimated positive and negative canonical correlations respectively (from left to right) in dependence of the sample size and the strength of the true canonical correlation. Performance in recovering the signs separately for positive and negative canonical correlations is very similar for all sample sizes.	133
5.3	Visualization of the correlations within and between gene expression levels ($\mathbf{X}$) and lipid concentrations ($\mathbf{Y}$) from the Nutrimouse study. The graphs show strong correlation patterns (≥ 0.5) within the gene expression data compared to the lipids data. The cross-correlation matrix indicates mild correlation patterns between the genes and the lipids which will be further exploited by CCA.	135
5.4	Plot of the estimated canonical correlations for the Nutrimouse data. The majority of the correlations indicate a negative association between the corresponding canonical variables.	136

5.5	Squared correlations loadings between the first 5 components of the canonical covariates $\tilde{\mathbf{X}}^{\text{CCA}}$ and $\tilde{\mathbf{Y}}^{\text{CCA}}$ and the corresponding observed variables $\mathbf{X}$ and $\mathbf{Y}$ for the Nutrimouse data.	137
5.6	Scatter plots between corresponding pairs of canonical covariates for the Nutrimouse data.	138
5.7	Scatter plots between first and second components within each canonical covariate for the Nutrimouse data.	138
5.8	Nutrimouse data. Top: Distribution of the K matrix coefficients for our CCA whitening method and sparse CCA. These are mostly centered around 0 with a higher concentration for our CCA whitening method. Bottom plot shows a sample network centered around metabolite “C18.2n.6” induced respectively by CCA whitening (upper plot) and sparse CCA (lower plot). The network shows a sparser structure for our CCA whitening method compared to sparse CCA.	139
5.9	Visualization of the correlations within and between the two investigated variables (microRNA expression and growth inhibitory response) from the NCI-60 data (Liu et al., 2010). The graphs show strong correlation patterns (≥ 0.5) within $\mathbf{Y}$. The cross-correlation matrix indicates mild correlation patterns (≤ 0.5) between $\mathbf{Y}$ and $\mathbf{X}$ which will be further exploited by CCA.	141
5.10	Plot of the estimated canonical correlations for the NCI-60 data.	142
5.11	Scatter plots between corresponding pairs of canonical covariates for the NCI-60 data with different CCA models applied. Samples are colored according to their cancer type.	142
5.12	NCI-60 data. Top: Distribution of the K matrix coefficients for our CCA whitening method and sparse CCA. These are mostly centered around 0 with a higher concentration for our CCA whitening method. Bottom plot shows a sample network centered around drug “Amonafide” induced respectively by CCA whitening (upper plot) and sparse CCA (lower plot). The network shows a sparser structure for our CCA whitening method compared to sparse CCA.	143

5.13	Visualization of the correlations of the gene expression ($\mathbf{X}$) and methylation levels ($\mathbf{Y}$) for the TCGA LUSC data. The graphs show strong correlation patterns (≥ 0.5) within $\mathbf{Y}$. The cross-correlation matrix indicates mild correlation patterns (≤ 0.5) between $\mathbf{Y}$ and $\mathbf{X}$ which will be further exploited by CCA.	144
5.14	Plot of the estimated canonical correlations for the TCGA LUSC data. The graphs show some strong correlation patterns (≥ 0.5) within $\mathbf{X}$. The cross-correlation matrix indicates few strong correlations between $\mathbf{Y}$ and $\mathbf{X}$. As most of the correlations are very mild in this example (green), CCA will be mainly used for exploratory analysis.	145
5.15	Squared correlations loadings between the first 10 components of the canonical covariates $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ and the corresponding observed variables $\mathbf{X}$ and $\mathbf{Y}$ for the TCGA LUSC data. These indicate that around 4 components are needed for $\mathbf{X}$ and around 2 components are needed for $\mathbf{Y}$	146
5.16	Scatter plots between first component of $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ (left) and within the first two components of $\widetilde{\mathbf{X}}^{\text{CCA}}$ (right) for the TCGA LUSC data.	147
5.17	Plot of the survival probabilities for male and female patients for the TCGA LUSC data.	148
6.1	Graphical representation of infinite functional clustering for metabolite annotation model. The cluster parameters θ_k are deliberately removed from the figure for clarity. The model consists in three components: the integrative level Dirichlet process, the data-specific Dirichlet processes and the metabolite annotation level. Fixed parameters are represented by squares and random variables are represented by circles. Depending on the score function used, metabolite annotations are conditioned on cluster assignments z_{dj} or cluster assignments z_{dj} and data points $\mathbf{x}_{dj}$	162
6.2	Similarity matrix for HDP clustering of the NCI-60 data. The matrix shows a diagonal pattern suggesting small size clusters.	168
6.3	Histogram of the number of clusters for NCI-60 data following MCMC inference.	169

6.4	Left: Boxplot of intra-pathway and inter-pathway correlations for the NCI-60 data. Intra-pathway correlation are correlation between metabolites in the same pathway by contrast to inter-pathway correlations designating metabolites in different pathways. Right: The histogram of the proportion of "true" pathway connections in each cluster shows that 0 is the dominant value suggesting that the clustering is not in accordance with metabolic pathways.	169
6.5	Mean rank (out of 128) of true recovered metabolite identities for NCI-60 data based on decreasing score functions. The plot shows a poor performance of the model for all score functions. .	170
6.6	Boxplot of the number of clusters for different values of γ across different configurations of (α_1, α_2) for the nutrimouse data. . . .	172
6.7	Figure shows different validity measures used to assess clustering of nutrimouse data and compare it with mclust model. Values obtained for mclust are represented by a dashed line. Top left panel shows -WAIC values on a positive scale. Best performance is obtained for highest values. Top left: -WAIC values are at maximum when γ is at lowest values. Top right and bottom left : Fukuyama-Sugeno index and Xie-Beni index plots show that all HDP models perform similarly but remarkably better than mclust. Bottom right : Partition coefficient values show less uncertainty in the clustering induced by mclust (partition coefficient equal to 1) compared to the HDP model (partition coefficient less than 0.4) . Best model is given by $\gamma = 3$ and $(\alpha_1, \alpha_2) = (12, 14)$	173
6.8	Similarity matrix for HDP clustering of the nutrimouse data for $\gamma = 3$ and $(\alpha_1, \alpha_2) = (12, 14)$. Darker colors indicate higher posterior probabilities that two omics variables are in the same cluster. Matrix rows and columns are arranged by cluster labels. Horizontal side bar encodes data types while the vertical side bar encodes the different cluster labels.	175

6.9	Left: Boxplot of intra-pathway and inter-pathway correlations for the nutrimouse dataset. Intra-pathway correlation are correlation between metabolites in the same pathway by contrast to inter-pathway correlations designating metabolites in different pathways. Right: The histogram of the proportion of “true” pathway connections in each cluster shows that 1 is the dominant value suggesting that the clustering is to some extent consistent with metabolic pathways.	176
6.10	Mean rank (out of 30) of true recovered metabolite identities for nutrimouse data based on decreasing score functions. All score functions seem to give similar results with the homogeneity score slightly underperforming compared to the two other score functions.	178
6.11	KEGG pathway: Biosynthesis of unsaturated fatty acids. Highlighted metabolites (red and blue) and genes (yellow) are those who correspond to detected variables in the nutrimouse dataset. The pathway comprises two clusters: A homogeneous cluster represented by metabolites in blue and a heterogeneous cluster represented by the ACOT2 gene in yellow (EC 3.1.2.2) and palmitic acid in red.	179
A.1	Metformin data: Posterior distributions and means of missing values in acylcarnitine for the last two timepoints. Posterior means in green seem to provide reasonable estimates of missing values. Orange violin plots show the posterior distributions and span a wide range.	192
A.2	Metformin data: Boxplots of prediction error for 4 observed time points to 8 observed time points. iCARH performs similarly in predicting future timepoints with a prediction error always centered around 0.	193
B.1	Estimated canonical correlations with correctly identified signs in dependence of the sample size and the strength of the true canonical correlation for both positive (left) and negative (right) signs.	196

B.2	Estimated canonical correlations from fast regularized CCA in dependence of the sample size and the strength of the true canonical correlation for both positive (left) and negative (right) signs.	197
B.3	Scatter plots between corresponding pairs of canonical covariates for the nutrimouse data. Samples are colored according to wild type (WT) and PPAR labels. Plots show that our CCA model outperforms classical and sparse CCA for nutrimouse data for exploratory analysis purposes by exhibiting better correlation and better discrimination across labels.	198
B.4	PCA explained variance for gene expression data (left) and lipids data (right).	199
B.5	Scatter plots between different pairs of components for the nutrimouse data. Samples are colored according to mice genotypes and diet labels. Plots show that the first two PCA components are able to discriminate between the genotype of the mice whilst components (3,4) and (4,5) are able to identify COC and SUN diets respectively.	200
B.6	Scatter plots between different pairs of components for the nutrimouse data. Samples are colored according to mice genotypes and diet labels. Plots show that the first two PCA components are able to discriminate between the different diets whilst components 2 and 3 show some information relative to the genotype of the mice. Further components potentially identify some diet types.	201
C.1	Figure from Orbanz (2012). The figure shows realizations of Dirichlet processes with different concentration parameters γ . The x-axis shows atoms locations and y-axis shows atom weights. Left: For $\gamma = 1$ the density is concentrated on few atoms who have large weights compared to others. Thus most of the observations will likely belong to small number of clusters. Right : For $\gamma = 10$ the weights are more even and a higher number of clusters is expected.	207

List of Tables

3.1	Classification of different data integration approaches. The examples list is by no means exhaustive.	64
4.1	iCARH model summary	93
6.1	Model summary for infinite functional clustering for metabolite annotation.	164
6.2	NCI-60 dataset: This table shows different measures of interest comparing our model with <code>mclust</code> . Validity measures show that they perform similarly.	168
6.3	Total clusters refer to the total number of clusters. Average number of heterogeneous clusters refer to the number of clusters containing variables from both data averaged over the different values of α_d . As expected, the clusters sizes are highest when γ is low, however, the average number of heterogeneous clusters is highest when $\gamma \in \{3, 5\}$. The exception here being $\gamma = 1$ and $\gamma \in \{7, 14\}$ since these values do not favor integrative analysis as explained in Section 6.3.1.	172
7.1	Summary of different features of the result chapters in this thesis	186

Acronyms

NMR	Nuclear Magnetic Spectroscopy
MS	Mass Spectrometry
PCA	Principal Component Analysis
PLS	Partial Least Squares
O2PLS	Two-way Orthogonal Partial Least Squares
CCA	Canonical Correlation analysis
CAR	Conditional Auto-Regressive model
iCARH	integrative Conditional Auto-Regressive Horseshoe model
DPPCA	Dynamic Probabilistic PCA
SME	smoothing Splines Mixed Effects model
GWAS	Genome Wide Association Studies
MWAS	Metabolome Wide Association Studies
mGWAS	metabolome Genome Wide Association Studies
DP	Dirichlet Process
HDP	Hierarchical Dirichlet Process
DPMM	Dirichlet Process Mixture Models
PSDF	Patient Specific Data Fusion
FSI	Fukuyama-Sugeno Index
XBI	Xie-Beni Index
PC	Partition Coefficient

Notation

Unless otherwise stated, we will use upper case letters to indicate random variables or matrices and lower case letters for specific observations. Vector quantities are written in bold type. Matrices are written in bold upper case letters. For example, $\mathbf{X}$ can be used to represent

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

However, in order to distinguish matrices from random vectors, we specify in the corresponding context if the quantity is a matrix or a random vector. If i and j are indexes that runs through rows and columns of $\mathbf{X}$ respectively. The notation x_{ij} represents the i th observation of the j th variable in $\mathbf{X}$, $\mathbf{x}_j$ represents the j th column vector of $\mathbf{X}$ and $\mathbf{x}_i$ represents the i th row vector of $\mathbf{X}$. A superscript T is used to indicate matrix or vector transpose. $\mathbf{I}_p$ is used to indicate the identity matrix of order p .

Turning to specific points of notation, $\mathbf{X}$ and $\mathbf{Y}$ will typically designate two omics datasets. Unless otherwise specified, p and q will designate the number of omics variables in the data matrices $\mathbf{X}$ and $\mathbf{Y}$ respectively. n will designate the number of observations, d the data index and D the total number of datasets. $\mathbf{X}_{-i}$ or $\mathbf{x}_{-i}$ refers to the subset of datapoints in $\mathbf{X}$ or $\mathbf{x}$ where item i has been removed. $\mathbf{1}_{\cdot}$ is the indicator function equal to 1 when condition $\cdot$ holds and 0 otherwise.

Additional notation specific to each context will be detailed in the corresponding chapter.

Software packages, programming languages and programming language keywords will be denoted in typewriter font.

Chapter 1

Introduction

1.1 Overview and motivation

Biological processes and functions are the result of multiple interactions between tens of thousands of omics variables and are inherently complex. In the last 30 years, the parallel acquisition of high-throughput datasets from the genome, metabolome, proteome, and transcriptome has seen a tremendous boost. The genome, metabolome, proteome, and transcriptome form different layers of the so-called “omics cascade”, each of which characterizes a biosystem or an organism at different biomolecular levels (Haukaas et al., 2017). The ensemble of a number of omics layers is usually designated by the term *multi-omics* or *cross-omics*, reflecting the complexity and heterogeneity of the system they attempt to measure. Omics entities or variables belonging to distinct omics layers are said to be *heterogeneous* by contrast to variables belonging to the same layer which are said to be *homogeneous* (Ritchie et al., 2015; Tseng, Ghosh, and Zhou, 2015).

Metabolomics deals with multiple and complex chemical reactions within living organisms and how these are influenced by external or internal perturbations. Data acquisition in metabolomics is generally achieved via analytical platforms such as Nuclear Magnetic Resonance spectroscopy (Nicholson, Lindon, and Holmes, 1999) or Mass Spectrometry (Dettmer, Aronov, and Hammock, 2007) ¹ by profiling adequate biological samples to generate a set of measurements. We call the set of measurements for each sample interchangeably by *metabolic profiles* or *metabotypes*. The terms *variables* or *metabolic variables* are used to designate metabolites. Unless otherwise stated, the terms *datapoints* and *observations* are used as a more general terminology that might refer to either samples or variables depending on the context.

¹Mass Spectrometry is also used in proteomics.

Metabolomics lies at the heart of omics profiling technologies not only as the underlying biochemical layer that reflects information expressed by the genome, the transcriptome and the proteome, but also as the closest layer to the phenome (Nicholson and Wilson, 2003; Coughlin, 2014; Holmes et al., 2008). Therefore, metabolic profiles or metabotypes can be indicative of changes in genetic profiles (genotypes) or proteomic profiles (proteotypes) and also markers of phenotypic traits (phenotype). As such, metabolomics offers a great potential to identify both metabotype-genotype interactions and metabotype-phenotype associations (Ritchie et al., 2015). Inference of significant associations between phenotypic traits and metabolic profiles is termed as *biomarker discovery* whilst inference of significant associations in multi-omics data is often termed *cross-omics biomarker discovery* (Robinette, Lindon, and Nicholson, 2013; Winning, 2009; Chadeau-Hyam et al., 2010).

The most frequently invoked omics approach is genomics defined as the quantitative study of the genome, gene mapping and interaction (Mayer, 2011). High-throughput technologies in genomics offer the ability to interrogate tens of thousands of genes simultaneously. Data acquisition is usually performed by means of microarray techniques or next-generation sequencing techniques. This rapidly increasing sequence information has been also successfully applied to e.g. transcriptomics to gain gene expression information at high throughput level. Sequencing techniques have also seen a boost in epigenomics as a powerful tool to decipher the regulation of gene expression resulting from e.g. external or environmental factors (Bird, 2007; Vineis et al., 2013; Niedzwiecki et al., 2019). Epigenomics is the study of changes to gene activity and regulation that do not alter the genetic sequence itself (Holliday, 1990). Differences in gene expression, activity or regulation are responsible for multiple phenotypic differences as well as indicative of cellular responses to external perturbations or environmental stimuli.

The integrative analysis of metabolomics and other omics datasets such as genomics, transcriptomics, proteomics and epigenomics principally deals with the association, correlation, combination, coupling and linking of data from multiple sources in a *meaningful* way to provide a more comprehensive view of the biosystem of interest. It is a promising area of research to enhance current understanding of biological functions, elucidate their underlying mechanisms and uncover hidden associations between multiple variables (Chadeau-Hyam et al., 2013; Yugi et al., 2016; Yuan, Savage, and Markowitz, 2011; Walther et al.,

2010; Tang et al., 2014; Bottolo et al., 2013; Schmidt et al., 2007).

When dealing with metabolomics data for integrative analysis, multiple challenges arise and these are in some cases shared with the other omics technologies. On a first instance, omics are not fully characterized. Profiling technologies in metabolomics are subject to an incredibly complex chemical heterogeneity where metabolites are typically not easily identified. The metabolome, in fact, is characterized by a high diversity comprising thousands to hundreds of thousands of chemicals ² (Devlin, 2011). As a consequence, unknown entities might be entities that are not identified but might also represent chemicals that have not been reported in the literature. On the other hand, genomic variables are not usually fully characterized by the profiling technology and require an annotation step. Gene annotation is subject to two major bottlenecks: Identifying elements on the genome (Gene finding) and adding biological information to these elements (Gene function). Uncovering the function of genes is critical to understanding its biological roles and corresponding cellular mechanisms. This challenge in the characterization of variable identities is not only likely to induce biases in interpretability but can also lead to uninterpretable results (Sumner et al., 2007).

On a second instance, distinct omics datasets have their own limitations and require complex analysis pipelines prior to performing data integration. For instance, analysis of methylation data is complicated by the uneven distribution of methylation target sequences across the genome requiring specific normalization and scaling strategies (Holliday, 1990). Each omics platform faces unique challenges such as experimental and inherent biological noise, differences among experimental platforms and detection bias (Richelle, Joshi, and Lewis, 2019). In a similar vein to processing genomic data, a supplementary step is critical to ensure interpretability in metabolomics data: metabolite identification. In an agnostic approach, where metabolites are putatively annotated, integrative analysis can be performed regardless of the metabolite identification step. However in a more specific approach, integrative analysis needs to be performed with regards to whether or not metabolite identification has been realized beforehand. For example, if metabolites have not been identified, data integration would be rather limited to almost purely statistical analysis i.e. classification purposes, prediction purposes or inference of significant variables whereas when metabolite identities are known enrichment analysis methods can be applied.

²Details in Chapter 2

Additional challenges arise since there is often not a one to one relationship between genes and metabolites.

On a third instance, the metabolome is intrinsically different between individual samples due to its rich chemical diversity and hence some observed patterns in metabolic profiles might not be the result of perturbations in the biosystem or phenotype of interest. Therefore, integrative analysis ideally needs robust models to account for individual variations.

On a fourth instance, metabolomic datasets are characterized by high correlation structures in that many peaks can arise from the same metabolites and metabolites operate within networks of chemical reactions. Hence, two highly correlated metabolites might not be directly related but high correlation can be the result of complex interactions with other metabolites in common. Finally, additional fundamental challenges are similar to typical challenges that usually arise in data integration frameworks, for example, incompleteness of each data type (i.e. missing values), high dimensionality and heterogeneity of data obtained from multiple sources. It is noteworthy to acknowledge that all data analysis steps need to be performed in account of the technical and experimental limitations in metabolomics. Metabolomics requires relatively high-cost instrumentation, complex data analysis and still suffers from issues of variable sensitivity, high volatility and sample-to-sample variability (Riekeberg and Powers, 2017). As a result, reproducibility is one of the significant hurdles in metabolomics (Keun et al., 2002). In fact, the delicate stability of metabolites implies that biological samples need to be handled quickly and uniformly. Minor changes in the experimental conditions or procedure (e.g. different laboratories, external conditions ...) can cause critical changes in the observed metabolome which might impact results.

A part from these challenges, a preliminary examination of the literature at the time of research shed light on two substantial shortcomings. Most of the current integrative analysis approaches are conducted separately from the main stream of the analysis, i.e. as a supplementary step. These two-step integrative approaches are very informative to prioritize data signals, nevertheless, they are not optimal. The heterogeneity of biosystems suggests that interrelationships between the various omics entities is key to exhibiting specific phenotype implying that data integration plays an important role into deciphering mechanisms of biological functions in living organisms (Joyce and Palsson, 2006). As a consequence, integrative analysis should be part of the main analysis pipeline. On

top of that, a close survey of the literature reveals that applications of probabilistic models for integrative analysis in metabolomics are very scarce. This is mostly ascribable to the limited number of available software on probabilistic models in the field which restricted their popularity.

From the outset, our research is primarily motivated by the need of enhanced interpretability in omics data integration. We are interested to answer to three different questions that will be addressed in each chapter of this thesis while simultaneously attempting to provide user-friendly software for the analysis. The leading questions we want to answer are specifically:

1. Which variables (or group of variables) of the first data type are significantly associated with the second data type?
2. Which cellular processes are perturbed as a result of treatment or disease?
3. How can a different type of omics data deepen our understanding of metabolomics, in particular, can a different type of omics data help metabolite identification?

The first question illustrates one of the most popular topics in integrative analysis from a statistical perspective. Multiple questions that are concerned with statistical issues examine predictive ability of integrated datasets compared to that of a single dataset (Shen et al., 2012) or seek to examine the significance of association between heterogeneous variables (or specific combinations of heterogeneous variables) that are indicative of a phenotype. On the other hand, biological issues are also at the core of integrative analysis (Ritchie et al., 2015). In this case, researchers might be interested in determining whether findings of one data are consistent with the other data, whether they are contradictory or whether they complement each other. Biological issues also include interrogations on the regulatory processes and cellular mechanisms leading to the expression of phenotype. The last two questions we endeavor to answer comprise this biological dimension. Despite the apparent disconnection, the statistical dimension and the biological dimension are closely related. For instance, biological mechanisms can be investigated by coupling questions 1 and 2. Metabolomics allows to detect subtle alterations in concentrations of metabolites related to cellular activity. By examining association between metabolite concentrations and the phenotype of interest and integrating this information into e.g. functional approaches, the analysis can move from biomarker discovery towards cellular mechanisms.

The third question motivates integrative analyses where different types of omics data are used to inform metabolomics data. However, we note that multi-omics data integration can go far beyond informing metabolomics only and that metabolomics can in turn inform other omics. In fact, metabolic reactions are far-reaching as they can be involved in the biosynthesis of cell building blocks (Vineis et al., 2013). By means of example, metabolites can help identify genes³ that perturb metabolism and enrich mechanistic insight by reflecting back to the function of the gene (Suhre et al., 2011b; Shin et al., 2014). In a similar fashion, metabolites also play an important role in strengthening the annotation of gene functions by uncovering downstream targets from a given gene (Ji et al., 2011; Rattray et al., 2018; Vineis et al., 2013).

All three questions raised in this work illustrate distinct steps of a typical pipeline for metabolomics data analysis. For instance, statistical analysis of significant variables with respect to an outcome of interest is performed in the statistical analysis step. Metabolite identification can be carried out on these significant variables to put forward plausible biological hypotheses. It is optionally followed by examination of perturbed biological processes, re-examination of data, orthogonal biological analysis⁴ and further experiment to provide potential targets for further mechanistic studies.

1.2 Summary of contributions

In this thesis, we seek to address the issues discussed in the previous section by exploring different research aspects relevant to the metabolomics field. In particular, the main objectives of this thesis consist in:

1. Developing and investigating novel statistical models for integrative analysis of metabolomics data with other omics technologies
2. Enriching the offer of probabilistic models tailored to metabolomics data
3. Providing enhanced interpretability of results

The third objective comes spontaneously as a result of the second objective. This is explained by the fact that probabilistic and Bayesian frameworks in general make explicit assumptions about parameters of interest via prior

³More specifically alleles which are variant forms or copies of a gene. When the copies of a given gene are different, they are known as alleles.

⁴orthogonal refers to dissimilar or different biological analysis.

distributions or generative models⁵. This allows to model uncertainty in parameters and facilitates a common sense interpretation of statistical conclusions by contrast to frequentist statistics where usually p-values can be a source of confusion. Thanks to its flexibility, the Bayesian framework can also boost interpretation if the parameters are appropriately modeled in a consistent way with available knowledge about the underlying problem. Moreover, the flexibility of the Bayesian setting allows for complex multi-layered model specifications and, hence, seemingless incorporation of integrative analysis components⁶.

In addition to the above aspects, this thesis is mainly concerned with designing models that can be introduced in different steps of a typical analysis pipeline of metabolomics data. Typical steps include *statistical analysis*, *metabolite identification* and *pathway analysis*. First of all, Chapter 2 motivates the importance of data integration from a biological perspective and describes the processing and analysis pipeline of metabolomics data whereas Chapter 3 covers some common statistical techniques used in metabolomics and that will be further studied throughout the thesis. Chapter 4⁷ covers the *statistical analysis* and *pathway analysis* steps of the analysis pipeline, by building a single comprehensive integrative Bayesian model that is able to perform both cross-omics biomarker discovery and infer potential perturbed pathways. Using simulations and real data, we demonstrate that our model compete with state-of-the-art models while performing steps of the analysis pipeline simultaneously. Since variations in the data are modeled explicitly, the model effectively benefits from ease of interpretation of the results.

In Chapter 5⁸ we focus solely on *statistical analysis* by uncovering hidden associations between multi-omics data via canonical correlation analysis (CCA). We revisit CCA from the perspective of statistical whitening of random variables and propose a simple yet flexible probabilistic model for CCA in the form of a two-layer latent variable generative model. The advantages of this variant of probabilistic CCA include non-ambiguity of the latent variables, provisions for negative canonical correlations as well as ease of interpretation on all levels

⁵Generative models encode a joint distribution over all variables including observations and parameters and allow and specify how to generate data that fit such distributions. For further details see Barber (2012) and Murphy (2012)

⁶Another appealing feature with respect to metabolomics data is that the Bayesian framework can naturally handle missing values in accordance with the generative model underlying the data

⁷Material in this chapter is mainly based on our paper presented at Advances in Modeling and Learning Interactions from Complex Data workshop NIPS 2017 available online as a preprint (Jendoubi and Ebbels, 2018).

⁸Material from this chapter is based on our preprint Jendoubi and Strimmer (2018)

of the model. In addition, we show that it lends itself to computationally efficient estimation in high-dimensional settings using regularized inference. It also allows for straightforward derivation of dependency measures between groups of variables based on an entropy criteria. We test our approach to CCA analysis in simulations and apply it to different omics data sets illustrating the integration of gene expression data, lipid concentrations and methylation levels. We also show that the proposed entropy-based coefficient is also able to uncover underlying network structure.

Finally, in Chapter 6 we address challenges relative to *metabolite identification*. We investigate the incorporation of additional omics data and pathway information into a Bayesian nonparametric clustering model and its potential to help metabolite identification. Clustering for functional analysis is widely used in omics technologies and proven to provide more understanding of the forces underlying cellular processes and an unbiased method for researchers to identify related functional modules. In the Bayesian setting more flexible models based on hierarchical layers were evaluated. Hierarchical models offer more degrees of freedom than one-level models and thus allows to define for each data its own parameters that might (or might not) be shared. To the best of our knowledge, infinite functional clustering has not been investigated for metabolomics data. In Chapter 6 we combine network information from pathway databases, genomic data and infinite clustering via hierarchical Dirichlet processes and we report on the potential of this model to improve metabolite identification, that is to say whether or not genomic data contributes to enriching our knowledge of metabolomics data.

When possible, simulation studies are used to get a better understanding of our methods and test their applicability. These simulations are always followed by analysis of real data and comparison to competing methods. In most instances, our methods have resulted in plausible biological findings when applied to real data, and represent, to our knowledge, one of the first applications of such probabilistic models in integrative analysis involving metabolomics data.

Chapter 2

Biological background and omics technologies

In order to explain the inherent need for data integration, this chapter provides the reader with basic understanding of molecular biology and the omics cascade pinpointing interactions between the genomic, transcriptomic and metabolomic layers. A particular emphasis is given to metabolomic studies. Data acquisition technologies such as Nuclear Magnetic Resonance (NMR) and Mass Spectrometry (MS) are discussed with a brief introduction to corresponding preprocessing steps. Finally, the “metabolite annotation” and “metabolic pathway analysis” steps are highlighted separately.

2.1 Basics of molecular biology

The cell is the main building block of all living organisms from the simplest to the most complex. Living cells constitute single independent systems such as bacteria and multicellular organisms such as humans with over 100 trillion cells. Molecular biology focuses on studying the molecular composition inside and outside of the cell and interrelationship of these molecules. Three classes of macromolecules play important roles in the central dogma of molecular biology : the deoxyribonucleic acid (DNA), the ribonucleic acid (RNA) and proteins. The central dogma of molecular biology is concerned with explaining how information is encoded, transmitted copied and maintained within the cell. It is schematically represented in figure 2.1. DNA replication occurs when a cell divides so that a copy of the DNA is conserved in the new cell. Most importantly, the central dogma of molecular biology states that genetic information is transmitted from DNA to RNA by “transcription” and translated from RNA to proteins by “translation”. Nowadays we know that each stage feeds back

to previous ones. For instance, RNA can be used to create DNA by reverse transcription.

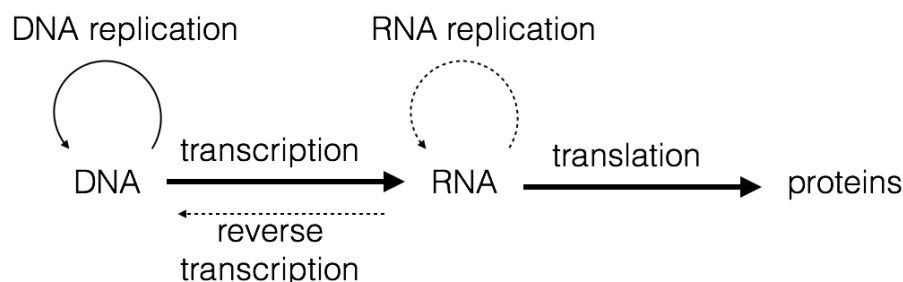


FIGURE 2.1: Simplistic representation of the central dogma in molecular biology. When a cell divides DNA is replicated such that genetic information is maintained in the resulting cell. Transcription concerns transmission of genetic information from DNA to RNA whilst translation is the process of protein synthesis based on RNA coding. RNA can also be replicated (RNA replication) or used to create DNA by reverse transcription.

Nucleic acids (DNA and RNA) are the primary units used for genetic information storage in the cell. DNA and RNA are able to encode vast amounts of information using a simple four-letter code. The genetic information inside the cell is hence composed by the nucleobases that differ from one unit to another. Three major bases are commonly found in nucleic acids: Guanine (G), Adenine (A), Cytosine (C). However, Thymine (T) is generally found only in DNA and Uracil (U) is generally found only in RNA. Due to their chemical properties, T and A only bind to each other and C and G only bind to each other as well. RNA encodes the same information as the transcribed DNA chunk as it uses the same coding, only replacing T by U.

There are different types of RNA. These include, for example, messenger RNA (mRNA), ribosomal RNA (rRNA) and micro RNA (miRNA). The mRNA acts as a messenger from the DNA with the required information to synthesize protein molecules. The rRNA is the main constituent of protein factories i.e. ribosomes. It is essential for protein synthesis but is not translated into proteins. Finally, the miRNA participates in post-transcriptional regulation. Specifically, the miRNA molecule is a small fragment of RNA that has a matching sequence in its “target” mRNA i.e the nucleobases from the miRNA have their matching nucleobases in the target mRNA. By binding to their target mRNA, the miRNA molecule is able to inhibit or down regulate its function. This process enables the cell to regulate the amounts of active mRNA molecules and thus their protein products.

A gene is a fragment of the DNA which is transcribed or copied into a functional RNA molecule. Genes generally contain the necessary information to synthesize proteins responsible for various vital biological processes such as cell growth and reproduction. The third type of macromolecules involved in the central dogma of molecular biology is proteins. Proteins are macromolecules that are formed by amino acid chains e.g. lysine, valine, leucine which represent 3 of the 20 molecules of amino acids in living organisms. Proteins perform a striking number of dynamic functions from catalysis of chemical reactions and metabolic control to oxygen transport and also play a protective role against viral or bacterial infection. A special type of proteins that is particularly interesting for metabolic reactions are “enzymes”. Enzymes act as catalysts for most biochemical reactions in biological systems to ensure that these reactions achieve a sufficient level of energy to form the product. As a result, the molecular composition of a cell is mainly determined by its enzyme composition which also enhances a particular phenotype. For example, altered levels of enzyme production may favor a specific genetic disease, a pathological condition or a metabolic imbalance (Baumeier et al., 2017; Linda et al., 2009).

Apart from proteins, metabolites also serve a complex, regulated system of chemical reactions known as metabolism. Metabolites are molecules with a mass under 1500 daltons (Da) and are also responsible for diverse life sustaining chemical processes including energy producing processes and processes involved in transcription as well as translation. Metabolites include amino acids, vitamins, fatty acids and carbohydrates amongst others. The complete set of metabolites in a biological organism forms the “metabolome”. Metabolites might be produced via two types of reactions: catabolic reactions which define “catabolism” and anabolic reactions which define “anabolism”. Catabolism represents the ensemble of reactions for degradation of large molecules into energy and smaller ones. In contrast to catabolism, anabolism consists in energy-utilizing reactions responsible for biosynthesis of large molecules as well as cell building blocks i.e. protein synthesis or DNA synthesis.

To sum up, most of the functions in living organisms are a result of a complex interrelationship between various types of molecules: The DNA encoding information for protein synthesis, is transcribed into RNA. The RNA is translated into proteins. Proteins along with metabolites characterize the particular biochemistry of the system (Devlin, 2011). They represent the actual functional phenotype i.e. what the cell “can do”, and offer the advantage, when correctly

measured, to show an accurate snapshot of the system's physiological state. For example, in metabolomics the analysis of volatile compounds contributing to e.g. food aromas, acids and hydrocarbons is challenging. In fact, MS conventional sample preparation techniques are not applicable as a sample might be subject to losses or incomplete extraction of analytes (Dunn and Ellis, 2005).

Gene expression is very rich and encode important information about the cell state and activity of genes (Duensing and Munger, 2004; Morozova and Marra, 2008; Cushman and Bohnert, 2000). In the context of medicine and human health, the knowledge gained from these types of measurements can help determine the causes and consequences of disease or exogenous exposure such as smoking (Guida et al., 2015; Fasanelli et al., 2015; Chatziioannou et al., 2017) or environmental impact (Mostafavi et al., 2017; Stringhini et al., 2015). In fact a wide range of GWAS (Genome Wide Association Study) studies have helped elucidating potential enzymatic mutations that can narrow down the DNA search space of variables linked to exposures (Ji et al., 2011; Rattray et al., 2018; Vineis et al., 2013; Suhre et al., 2011a; Bottolo et al., 2013). Epigenomics is widely adopted for this purpose, it provides stability and diversity to the phenotype (as a product of the genotype in juxtaposition with the environment) via local preserved transcriptions as a result of chromatin marks (Knaap and Verrijzer, 2016; Heard and Martienssen, 2014; Bird, 2007; Holliday, 1990; Niedzwiecki et al., 2019).

Unlike the genome and epigenome, the transcriptome is very dynamic and responds rapidly to perturbations (Joyce and Palsson, 2006; Liu et al., 2010). For most genes, changes in mRNA abundance are related to changes in protein abundance (Vogel and Marcotte, 2012). Proteomics, the study of proteins, is an additional discovery tool and may provide insights previously unappreciated. However, the analysis of proteins has major roadblocks because of the huge number of proteins, variability in size and post-translational modifications. Proteomics research leads to the handling of complex data sets, which include hundreds of proteins; it requires specialized data analysis that involves statistics and bioinformatics techniques (Gevaert and Vandekerckhove, 2000; Kopczynski, Sickmann, and Ahrends, 2017). Metabolites and proteins both form the biochemical state of the cell; hence, metabolome changes are closely associated with the enzymatic functions of the perturbed genes and corresponding proteins (Verhoeckx et al., 2004; Rantalainen et al., 2006).

Altogether, these different molecules and macromolecules form the entities

of interest in systems biology and their differential expression is key to capture acute and long term effects (Niedzwiecki et al., 2019; Robinette, Lindon, and Nicholson, 2013; Nicholson and Wilson, 2003; Nicholson, Lindon, and Holmes, 1999; Lindon, Nicholson, and Holmes, 2011). In order to enhance understanding of the integrity of the biological system, systems biology surpasses the study of each molecular level such as DNA, RNA, proteins and metabolites in isolation and considers the system as an integrated interacting network. The need for data integration in a way that is coherent with the numerous biological parameters is hence inherent to a systems biology approach. In the next section we will discuss typical data types in omics and present the omics cascade concept in a systems biology approach.

2.2 The omics cascade and omics technologies

Over the past years, there has been a significant development in high-throughput omics technologies e.g. genomics, transcriptomics, proteomics and metabolomics along with a growing interest into joint modeling of multi-omic data (Joyce and Palsson, 2006; Ebrahim et al., 2016; Yugi et al., 2016). Each omics discipline involves the study of a specific building block of the cell namely DNA, RNA, proteins and metabolites and is referred to as “omic layer”. The set of omics layers and their biochemical interrelationship constitutes the omics cascade (Figure 2.2). Metabolomics is a relatively new field and represents the last layer in the omics cascade after genomics, transcriptomics and proteomics. Omics are closely related to phenotypic changes of an organism which why they have seen a tremendous boost over the last years. The overall set of expressed phenotypes is called phenome.

High-throughput omics technologies offer the ability to measure different biological parameters at an increasing rate in a highly parallel framework. Microarrays are the most widely adopted high-throughput technique. DNA microarrays are able to measure the levels of gene expression i.e number of RNA transcripts transcribed from it. DNA microarray chips are formed of tens to hundreds of thousands of probes. Each probe is designed such that it tends to bind to a specific target molecule. The molecules in the sample are labeled with fluorescent labels, as a consequence the amount of fluorescence of each probe determine the amount of molecules that are bound to the probe and hence relative abundance of the target genes. Affymetrix, Agilent and Illumina

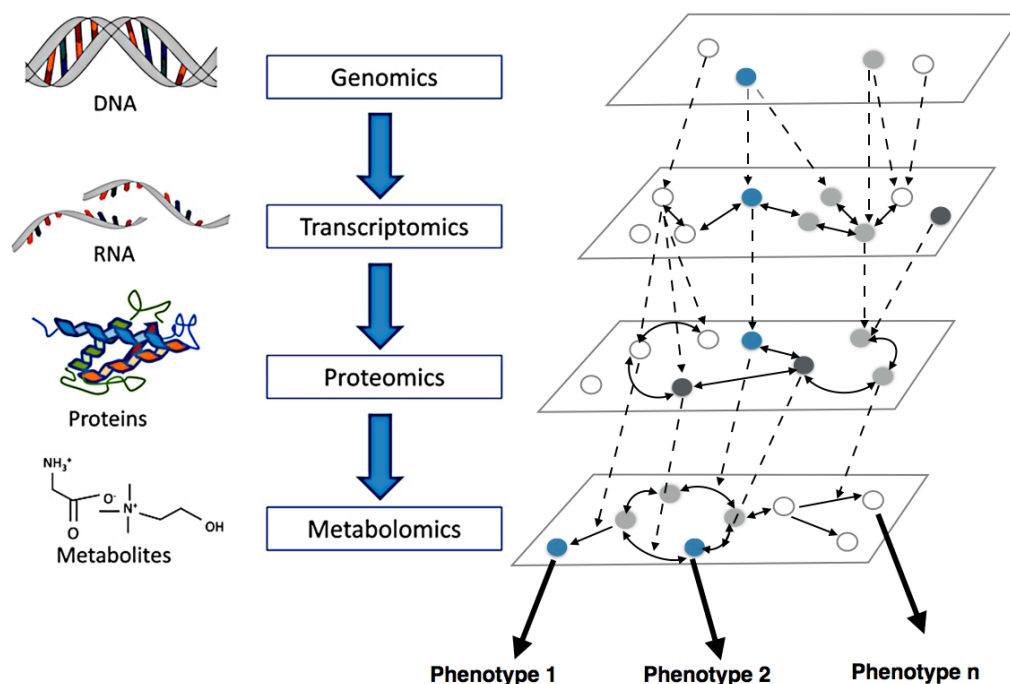


FIGURE 2.2: The omics cascade starting with the level of the genes to the last level of the metabolites triggering a particular phenotype (adapted from Haukaas et al. (2017)). For example, type 2 diabetes mellitus is a multi-factorial disease resulting from the breakdown of the insulin secretion pathway and the insulin sensitivity pathway along with some genetic (and environmental) factors.

are some of the mostly used microarray platforms. Recently, microarrays have been competing with next-generation sequencing techniques in gene expression analysis. DNA sequencing is the process of characterizing a sequence of bases in a given DNA strand. Steps of DNA sequencing framework involve preparing templates of the sample DNA fragments which are then joined with known adapter sequences that can be bound by universal primers. Template sequence are generally amplified by producing millions of copies. An additional type of sequencing is RNA sequencing. It can be realized by reverse transcribing RNA molecules which are profiled as in DNA sequencing or by directly profiling the RNA. Depending on the experiment aims, mRNA or rRNA can be selected.

Next-generation sequencing techniques are advantageous thanks to their larger range, lower background noise, higher sensitivity and increased throughput. Genomic data are extremely informative however they are not usually fully characterized and need complex pre-processing steps prior to statistical analysis.

For microarray experiments, as for all other omics platforms, normalization is crucial for background and instrument noise removal. Numerous normalization methods for microarrays are used depending on the microarray platform, for example, normalization methods provided by Affymetrix in Microarray Suite 5.0 (Roberts, 2008) and robust multi-array average (RMA) (Irizarry et al., 2003) for Affymetrix microarrays.

Gene annotation and ontology is a critical step towards results interpretation. Gene ontology ensures that gene information derived from genomic analyses is related to its biological role (Ashburner et al., 2000; Wadi et al., 2016). Genomic databases benefit from a comprehensiveness that is missing in metabolomics (see next section) and that facilitates the functional analysis and annotations of genes in a transcriptomic approach (Garber et al., 2011; Consortium and Team, 2002).

Gene annotation consists in three steps: Gene finding and gene function (functional analysis). The principal activity in gene finding is identifying known landmarks into the genome i.e. the precise start and stop position of a gene. Different softwares have been developed to this end including GENSCAN (Burge and Karlin, 1997), GeneScan (Tiwari et al., 1997), AUGUSTUS (Stanke and Morgenstern, 2005) and SNAP (Korf, 2004). A nice review on gene annotation computational tools can be found in Stein (2001). The most intricate part of genome annotation is linking genes to biological processes (gene function or gene ontology). The Gene Ontology project (Ashburner et al., 2000) aims at unifying the representation of genes via identifying its molecular function, biological process and cellular component. It allowed the implementation of various computational ontology tools such as DAVID (Dennis et al., 2003), GOstat (Beißbarth and Speed, 2004), amiGO (Carbon et al., 2008), FatiGO (Al-Shahrour, Díaz-Uriarte, and Dopazo, 2004) and PANTHER (Mi and Thomas, 2009).

2.3 Metabolomics

Metabolomics aims to obtain a complete molecular profile of the measured biological sample (Lindon, Nicholson, and Holmes, 2011). Metabolites usually represent “the chemical fingerprints that specific cellular processes leave behind” (Daviss, 2005). They are monitored through examination of e.g. tissues,

cells or biofluids such as urine, serum, breast milk and are most commonly interrogated in terms of concentration, structure and transformation pathways.

A common pipeline in metabolomic studies consists in i) Sample collection and preparation ¹ ii) Data acquisition ¹, iii) Data preprocessing ², iv) Statistical analysis (chapter 3), v) Metabolite annotation (Section 2.3.4) vi) Enrichment analysis or Pathway analysis (Section 2.3.5) vii) Biological interpretation or rationale. “*Data integration*”, as the focus of this thesis, is an additional level that can be introduced in steps (iv), (v) or (vi). Figure 2.3 depicts the different steps of metabolomic studies in the targeted as well as in the untargeted modes. Steps (iv) and (v) are often carried out simultaneously or in a loop where statistical analysis is performed firstly, significant metabolites are annotated secondly, then further statistical analysis is carried out using the significant metabolites. The shaded steps (steps (iv), (v) and (vi)) indicate where data integration can be possibly introduced during analysis. These steps will be the focus of the thesis. Specifically, step iv is at the core of chapter 5. Step (v) is the main concern of chapter 6. In chapter 4, I try to address steps (iv) and (vi) jointly.

In metabolomics, data acquisition techniques are defined according to the study objectives and priorities. Typical experimental platforms for data acquisition purposes use analytical techniques such as ¹H nuclear magnetic resonance spectroscopy (NMR) (Nicholson, Lindon, and Holmes, 1999; Reo, 2002) and mass spectrometry (MS) (Dettmer, Aronov, and Hammock, 2007) to generate appropriate spectral metabolomic profiles of the studied biological system. Metabolic phenotyping has generally been restricted to experts in analytical chemistry because of its analytical challenges. However, recently, the development of analytical chemistry techniques has become more and more reproducible and standardized with an increasing dependence on mathematical and computational models. In the consecutive sections, I will briefly describe some basic aspects of NMR and MS, followed by a short introduction of “metabolite annotation” and “enrichment analysis”. The “Statistical analysis” step will be discussed in details in Chapter 3.

¹Experimental step

²Note that “Data preprocessing” involves a huge range of computational methods but is not in the scope of this thesis and will only be very briefly introduced.

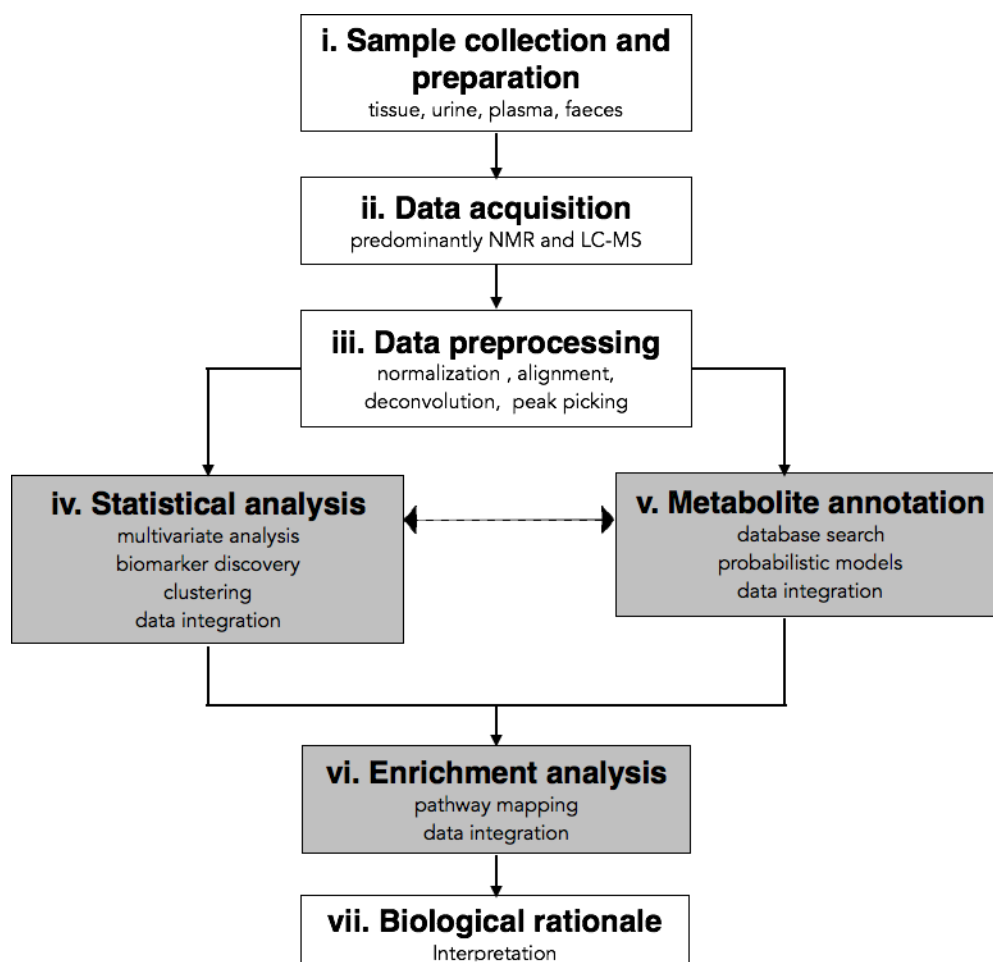


FIGURE 2.3: Representative analysis pipeline of metabolomics data

2.3.1 ^1H Nuclear Magnetic Resonance Spectroscopy

NMR is a highly reproducible and reliable technique for metabolomics applications. It offers the ability to detect a wide range of metabolites simultaneously (^1H containing metabolites in this case) providing a snapshot of the metabolic profile of a biological system. NMR spectroscopy is applied on solution-state samples to generate spectral profiles based on local magnetic fields around atomic nuclei. In fact, when placed in a magnetic field, the atomic nuclei (in this case ^1H i.e proton) is excited at a resonant frequency proportional to the magnetic field. These resonance frequencies change according to the intramolecular (local) magnetic field which makes NMR signals very specific to characterize molecular structures and functional groups. These changes translate into the NMR spectrum as chemical shifts. The chemical shift is defined as the effect of the chemical structure of the molecule on the resonance frequency of the nuclei.

The x-axis of a NMR spectrum represents the chemical shift denoted in “parts per million” (ppm). It does not only provide information about the position of peaks but also the structure of the molecules. A typical NMR spectrum will only contain signals from a few hundred of the most abundant metabolites. Low weight molecules are usually well resolved and distinguishable (sharp peaks) whereas large molecules have broader overlapping peaks (See figure 2.4).

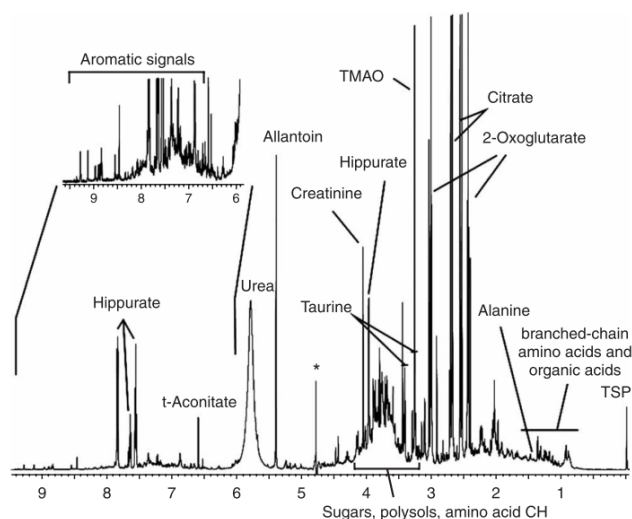


FIGURE 2.4: Example of a typical NMR spectrum of a rat urine sample with some annotated metabolites (Beckonert et al., 2007). All peaks are referenced to the resonance of TSP (sodium salt of 3-trimethylsilylpropionic acid) at 0 ppm. Sharp peaks indicate presence of low weight molecular compounds whilst broad ones indicate presence of larger molecules. Some aromatic signals in the 6-9 ppm region are low but they can be resolved. The asterisk refers to a suppressed water signal.

NMR spectra can be used as a set of ppm values and corresponding signal intensities (approx 30.000 variables depending on the resolution) or a set of ppm values and corresponding integrated peak area values (roughly 6000 variables if binning is used and 100 variables if peak picking³ is used). Integrated peak area values allow quantification of the metabolite relative concentrations and will be the only form of NMR data used in this thesis.

Appraised for its reproducibility, NMR is advantageous as it needs minimal sample preparation, is non-destructive and offers high throughput. In addition, NMR is a convenient method to investigate most abundant metabolites in metabolomic studies since it offers an overall snapshot of the organism. It is especially advantageous if no prior information about metabolites that are linked

³explained in next section

to a factor such as disease, smoking status or a specific diet is known. However, NMR procures a poor sensitivity compared to that of MS which allows signal detection of a larger number of metabolites.

2.3.2 Liquid Chromatography Mass Spectrometry

Mass spectrometry is an analytical technique that separates ionized molecules according to their mass-to-charge ratio (m/z values). It is widely used for metabolite profiling (metabolomics) or protein measurements (proteomics). Mass Spectrometry is usually coupled with a chromatography technique to help separate isomers (i.e molecules with the same mass but different structures) in an additional dimension. This results in a three dimensional spectrum (See figure 2.5) where the x-axis represents the mass-to-charge ratio, the y-axis represents retention time and the z-axis represents signal intensities.

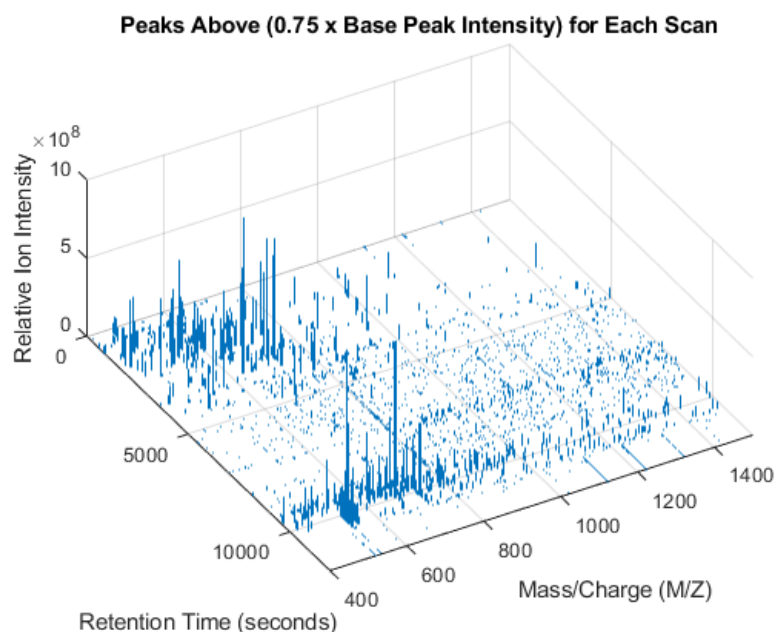


FIGURE 2.5: Example of a LC-MS spectrum of a mix of 7 proteins obtained from Sashimi data repository (file 7MIX_STD_110802_1 from <http://sashimi.sourceforge.net/repository.html>). The x-axis of a mass spectrum represents the m/z values denoting the mass-to-charge ratio formed by dividing the mass of an ion by the unified atomic mass unit and by its charge number (positive absolute value). The y-axis represents retention times of the ions. The z-axis represents signal intensities of the ions.

When MS is coupled with liquid chromatography (LC-MS technique), the sample is first pumped into mobile phase which is then delivered under high pressure to the column or stationary phase. Differentiation occurs when sample mixture components interact with either or both of the mobile and stationary phases and so elute from the column at different times. The mobile phase is a solution containing sample analytes water and organic solvents such as methanol and isopropanol. The stationary or solid phase consists in a chemical that can selectively retain sample components. The time from when the injection is made to when components exit the column is referred to as “retention time”.

Upon elution from the chromatographic column, compounds enter the ion source where they are fragmented and transformed into charged ions. Ionization can be performed by electrospray ionization (ESI) or bombarding the compounds by e.g. electron beams (EI), photon beams or laser beams. The ions are then sent to the mass analyzer. The latter applies both electric and magnetic fields to sort the ions according to their mass. Meanwhile, a detector amplifies the mass-resolved ions current and creates an electronic signal proportional to the number of ions. This signal is then processed by a computer.

In metabolomics research, both NMR and MS can form a multi-platform approach by combining their respective merits in order to boost metabolite identification and propose meaningful data interpretation (Pan and Raftery, 2007; Marshall et al., 2015). In studies where no prior information about a phenotype effect on the metabolic profile is available NMR is more favorable whereas MS is preferred for targeted analysis. Recent developments in LC-MS have broadened its applicability with a far superior sensitivity to that of NMR (more than 1000 detectable metabolites compared to 30-100 detectable metabolites for NMR), especially for complex mixtures. NMR, however, benefits from high reproducibility and robustness against different conditions. Despite its high variability due to a greater number of experimental variables including sample preparation procedures, chromatographic separation and ion suppression effects, MS methods are more popular as they offer higher sensitivity with relatively cheaper instruments.

2.3.3 Data preprocessing

Diverse preprocessing pipelines and softwares are used for raw metabolomics data depending on the adopted technique and the experiment purposes (Smith et al., 2006; Pluskal et al., 2010; Lommen, 2009; Wei et al., 2011; Dietrich,

Rüdel, and Neumann, 1991; Xi and Rocke, 2008). Common preprocessing steps consist in noise filtering and baseline correction, peak detection, alignment and normalization (Castillo et al., 2011). The aim of baseline correction is adjustment of signal intensities and background noise removal (instrumental noise, chemical noise derived from components other than the target analytes). For this purpose various thresholding methods (Castillo et al., 2011; Gan, Ruan, and Mo, 2006), filtering methods such as Gaussian filtering and Savitzky-Golay filtering (Savitzky and Golay, 1964) and wavelet transformations (Tautenhahn, Boettcher, and Neumann, 2008; Walczak and Massart, 1997; Cappadona et al., 2008; Günther, Ludwig, and Rüterjans, 2000) can be used to improve the signal-to-noise ratio in raw NMR and MS spectra.

In MS data, the peak detection or peak picking can be performed simultaneously with the noise filtering step. Peak picking refers to the procedure of detecting peaks that correspond to actual molecules in the sample. It is usually accompanied by deconvolution in order to separate poorly-resolved peaks. Two of the most popular algorithms for peak detection is the centWave algorithm (Tautenhahn, Boettcher, and Neumann, 2008) used in the XCMS package (Smith et al., 2006; Gowda et al., 2014; Tautenhahn, Boettcher, and Neumann, 2008) and MZmine 2 (Pluskal et al., 2010). Peak picking methods often need user defined input parameters which may produce different results if parameters are changed.

In common MS pipelines, depending on the softwares, peak alignment can be carried out following peak picking. An arsenal of computational methods are implemented and publicly available for MS data such as metAlign (Lommen, 2009), OpenMS (Sturm et al., 2008), msInspect (Fitzgibbon et al., 2008) and MET-XAlign (Zhang et al., 2015). In NMR, alignment is usually carried out after baseline correction. The objective is to establish correspondence between peaks that are not perfectly aligned across different samples. An extensive list of alignment methods in NMR can be found in (Bloemberg et al., 2013).

In addition to alignment, normalization methods are necessary to correct for differences in overall metabolite concentrations, for instance, dilution factors for solution-based samples but also global changes in instrument response, global sample preparation issues sample weight or detector response. Various normalization approaches include area normalization, range normalization, probabilistic quotient normalization (PQN) and normalization to a reference metabolite (e.g creatinine in urine spectra) (Dieterle et al., 2006). Normalization in LC-MS

can be realized by using an internal standard or by scaling the spectrum by a normalization factor (Smith et al., 2006; Pluskal et al., 2010).

Scaling is commonly performed as the last step in the NMR preprocessing pipeline (Craig et al., 2006). As metabolites exhibit variance in abundance, scaling is used to adjust signal intensities between metabolites of low abundance and metabolites of high abundance. Hence, changes in low abundant metabolites are not masked by changes in high abundant ones. Typical scaling methods include autoscaling (unit variance scaling) and Pareto scaling which uses the square root of the standard deviation instead of the standard deviation by contrast to autoscaling.

An additional steps in the LC-MS pipeline is deisotoping (Snider, 2007). Molecules are naturally present as isotopes i.e nuclei sharing the same number of protons which is reflected by different peaks for one particular molecule in a mass spectrum. In the deisotoping step, peaks are grouped into isotopic groups to remove redundant information and thus facilitate peak identification. The metabolite identification or annotation step, can be carried out before or after alignment in MS (See Section 2.3.4). Metabolite identification is still a challenging task for both NMR and MS but is the key to suggesting a biological rationale from the data.

Metabolomics is widely seen as complementary to other omics offering some special advantages. Namely, the theory of Metabolic Control Analysis (MCA) (Kell, 1999) and experimental efforts have both shown that, as downstream products of genes, changes in enzyme concentrations might not be significant but might have significant effects on metabolite concentrations during a biochemical reaction. Furthermore, it is estimated that metabolomics experiments cost two to three times less than proteomic and transcriptomic experiments (Harrigan and Goodacre, 2003). However, two main bottlenecks limit metabolomics experiments: the limited coverage and comprehensiveness of the metabolome as a single metabolomics technique is not able to span the whole metabolome of the organism of interest and the lack of reproducibility as slight changes in experimental conditions might have significant impact on the metabolome (Keun et al., 2002) (See chapter 1). These limitations have non negligible effects on metabolite identification and pathway analysis discussed in the following sections.

2.3.4 Metabolite annotation

Metabolites are in dynamic equilibrium in biological organisms. As a consequence, state changes due to a disease, toxin, environmental effects or life style choices are rapidly reflected into the metabolic profile. Metabolites that are found to be strongly associated with changes of the phenotypic state of a system are called “biomarkers”. Biomarkers are of major interest in metabolomics as they can be used as indicators of that specific phenotype. However, due to the vast structural heterogeneity of the metabolome and limited sensitivity of the analytical platforms, metabolites are not easily annotated or identified which represents a considerable challenge to understanding the biological rationale behind changes and comparing results between different research groups. The Chemical Analysis Working Group of the Metabolomics Standards Initiative has proposed 4 confidence levels for metabolite identification (Sumner et al., 2007). The most rigorous level of confidence is where metabolites are confidently identified by applying two analytical techniques under identical conditions within the same laboratory such as accurate mass and fragmentation data from tandem MS. The second and third levels of confidence do not require extensive experimental validation and are based on similarity with open or commercial libraries. These levels are termed as “putative annotations”. The fourth level of confidence designs the unknowns. In the metabolomics literature unknowns are categorized into two categories: the “known unknowns” which are already mapped to reference libraries and metabolic networks and the “unknown unknowns” i.e novel compounds that have not been previously quantified and integrated into a database. Various strategies to achieve unknown unknowns identification involve structure elucidation. A related survey can be found in Wishart (2009). Currently, no complete characterization of any metabolome-specific organism has been reported, not even as putative annotations. This will potentially cause downstream issues as all metabolites collaborating for a specific cellular function may have not been quantified. It is strongly recommended that researchers communicate the level of identification and standards used in their publications for data to be systematically shared and used.

Even though the annotation task has a lower confidence than the identification task, it is still an important step towards data interpretation and can be realized without further experimental setup. Due to chemical heterogeneity, annotation is not straightforward and require in depth chemical knowledge and tedious manual search prompting for the development of computer-aided

methods. The most successful and popular annotation methods use authentic material such as standard libraries and repositories to match quantified peak information to known features. For NMR data, metabolite annotation can involve the application of a range of techniques with peak curve fitting being one of the most used techniques. However typically ^1H NMR spectra of some biofluids allow many resonances to be assigned directly based on their chemical shifts, signal multiplicities and by adding known standards and authentic material from databases (Cui et al., 2008).

In mass spectrometry, a variety of repositories comprise KEGG (Kanehisa and Goto, 2000), BRENDA (Schomburg et al., 2002), MetaCyc (Caspi et al., 2013), HMDB 3.0 (Wishart et al., 2012), METLIN (Smith et al., 2005) and LIPID MAPS (Fahy et al., 2007) to facilitate both manual and computational annotation of metabolomics data. For LC-MS data, comprehensive softwares such as XCMS, Mzmine2, mzMatch (Scheltema et al., 2011) and MetSign (Wei et al., 2011) use the above databases, but also additional spectral databases, to perform metabolite annotations as a preliminary step to peak alignment. Kuhl et al. (2011) have introduced CAMERA as an integrated tool for metabolite annotation in LC-MS. The tool uses retention time to group peaks originating from the same metabolite and peak shape to identify metabolites along with isotopic patterns and adducts given a rule table. In a similar fashion, AStream creates peak clusters based on retention time and correlation across samples. Isotopic peaks are then identified based on m/z pair differences and concentration ratios. Lastly, adducts and metabolites are annotated using a provided list of expected adducts. In a different approach, MetaboAnalyst (Chong et al., 2018) a popular comprehensive software for metabolomics data analysis, supports mummichog-based MS peak analysis. Mummichog (Li et al., 2013) forms “activity networks” to predict functional activity from spectral features. The software estimates the null distribution of module activities in order to compute the statistical significance of each module. The statistical significance of functional modules is then used to predict metabolite identities. Beside annotation tasks, MetaboAnalyst offers eight components: Statistical Analysis and Time-Series Analysis modules, Enrichment Analysis, Pathway Analysis and Integrated Pathway Analysis modules for both genes and metabolites, Biomarker Analysis, Sample Size Estimation and Power Analysis modules.

Metabolome Searcher (Dhanasekaran et al., 2015) is an online tool for genome-scale metabolic models that is compound search is limited to genome plausible

metabolites. Based on its genome, only the compounds that the organism of interest is able to produce are queried. The tool is very flexible and allows inputs and outputs of multiple formats. The MassTRIX web tool (Suhre and Schmitt-Kopplin, 2008) performs putative annotations of metabolites by comparing the experimental masses against all KEGG compounds and linking the chemical compounds to the annotated genome. Polyomics integrated Metabolomics Pipeline (PiMP) (Gloaguen et al., 2017) annotates both MS metabolites and identifies corresponding pathways. In a similar vein, MetaMapp (Barupal et al., 2012) integrates biochemical pathways and chemical relationships to map metabolites. Attributing a biological interest to chemical entities is known as chemical ontology, such information is likely to boost performance of metabolite annotation tools. Chemical ontology terms have been organized by several teams comprising LipidMAPS (Fahy et al., 2007) and ChEBI (Chemical Entities of Biological Interest) (Degtyarenko et al., 2007) which offer both ontology of biological activity classes and chemical structural classes based on chemical compound databases such as ChemSpider (Pence and Williams, 2010) and PubChem (Kim et al., 2015).

Since the majority of tools relies on existing databases, it is important for the users to have some appreciation of the reliability of these databases. Incompleteness is one of the critical caveats of public databases as it is common to find missing, ambiguous or redundant entries which strongly affects the results of enrichment analyses (Marco-Ramell et al., 2018). Besides completeness, reliability also depends highly on regular updates to improve the quality and consistency of the databases. Wadi et al. (2016) produced an interesting review suggesting that newer database versions improve annotations compared to outdated ones. An additional aspect concerns incoherence of chemical identifiers involving various referencing schemes which can be addressed by means of identifiers standardization tools such as MetMask (Redestig et al., 2010).

Despite the characteristic ambiguity of the annotation task, very few probabilistic models have been proposed to deal with uncertainty in assignments. To the best of our knowledge, the unique available probabilistic models for LC-MS data are Rogers et al. (2008), Daly et al. (2014) and Silva et al. (2014). These models discussed in further details in Chapter 6. In NMR, the unique works of Hao et al. (2012) and Zheng et al. (2011) propose a relatively novel Bayesian framework for curve fitting to provide probabilistic metabolite quantification.

2.3.5 Metabolic pathways

Metabolites interact within networks of complex chemical reactions or pathways. Chemical reactions can be reversible (bidirectional) or irreversible (one way). Almost all reactions in living organisms require enzymes for catalysis at a rate compatible with life (See figure 2.6). Faust, Croes, and Helden (2011) define a metabolic pathway as “a set of interconnected reactions that can be activated coordinately to ensure a particular cellular function”. Most metabolic pathways are rather arbitrarily defined – due to historical reasons which results in multiple different definitions for the “same” pathway in different databases. However, these definitions remain biologically consistent with a specific cellular function.

When a biological system is perturbed due to an external factor i.e. treatment, disease or environmental exposure, metabolic reactions might be altered. This is of major interest to practitioners and researchers in metabolomics as it allows to trace back the mechanisms induced as a result of exposure. Much of this information can be extracted by inference of significantly altered metabolic pathways. This is termed as “pathway analysis” or “metabolite set analysis”. It has been first applied in genomics under a specific approach known as Gene Set Enrichment Analysis (GSEA) where differentially expressed genes are matched with genes that are associated with a specific biological function. Similarly, MSEA is used in metabolomics to designate Metabolite Set Enrichment Analysis (Xia and Wishart, 2010).

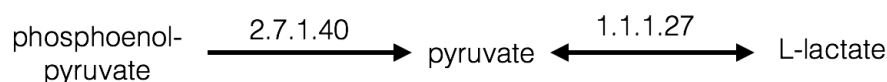


FIGURE 2.6: Transformation of phosphoenol-pyruvate into pyruvate (irreversible reaction) and pyruvate into L-lactate (reversible reaction). A substrate is the chemical substance that generates a product. For example, in the irreversible reaction, phosphoenol-pyruvate is the substrate whilst pyruvate is the product. Each reaction is catalyzed by an enzyme encoded here by its KEGG sequence 2.7.1.40 and 1.1.1.27 respectively. If the enzyme is inhibited, the reaction rate is low and the substrate tends to accumulate.

Networks are a natural integrative way to link the layers of the omics cascade namely genomics, transcriptomics and proteomics. Inherently, learning about metabolic pathway has attracted many researchers to develop various databases and computational tools in order to facilitate deriving biological meaning from the data. The most popular functional databases include MetaCyc with 14,971

enzymatic reactions, KEGG with about 11.004 biochemical reactions as of August 2018, BRENDA which comprises non-redundant reactions from both databases and Reactome.

Functional databases can be parsed by different tools such as MetaboSignal (Rodriguez-Martinez et al., 2017), MetaboNetworks (Posma et al., 2013) and Metscape (Karnovsky et al., 2011) to map KEGG pathways to networks for ease of visualization and manipulation. For more comprehensive pathway analysis, numerous computational tools are also available. These tools require metabolites to be annotated prior to pathway analysis. MetPA (Xia and Wishart, 2010) performs enrichment analysis by applying network centrality measures to determine metabolite importance and hence impact of a specific pathway by summing metabolites importance within a pathway. Significance is computed based on univariate statistical measures such as Fisher’s exact test and hypergeometric test. MetaboAnalyst (Chong et al., 2018) considers that mean level changes of one or more metabolite concentrations involved in a pathway indicate perturbation of the latter pathway. MetaboAnalyst also supports other pathway analysis methods such as MSEA and MetPA. Another web tool, IMPaLA (Kamburov et al., 2011) proposes an integrated framework for enrichment analysis. The tool is able to achieve over-representation analysis for metabolomic, proteomic and transcriptomic data simultaneously. In IMPaLA, a pathway is considered dysregulated if one of its compounds is differentially expressed or all of its components are jointly differentially expressed. In fact, pathway alterations might result from very small metabolite concentration changes that jointly affect the cellular corresponding function. The lack of reproducibility in metabolomics might affect these types of enrichment methods based on metabolite level changes as slight differences in the experimental conditions can cause important changes in the observed metabolome. However, methods that are based on network measures and do not involve a quantitative assessment of the change in metabolite concentrations are more robust to the lack of reproducibility. The majority of enrichment methods strongly depend on the list of detected metabolites given by the user. This suggests that the choice of the experimental platform is crucial to the observed metabolite enrichment in view of the different detection powers of analytical platforms (Rosato et al., 2018; Riekeberg and Powers, 2017).

Pathway analysis has also greatly benefited from integrating different types of datasets (Kim et al., 2018). Other better characterized omics data can be

used to infer perturbed metabolic pathways. For instance ConsensusPathDB (Herwig et al., 2016) combines molecular interaction information with gene levels to identify significant pathways. Metabotype patterns can be closely associated with the enzymatic functions of the perturbed genes in the metabolic pathways, hence knowledge of better established functional properties of given a gene can give critical insights on metabolic pathway perturbation (Cruickshank-Quinn et al., 2018; Vaske et al., 2010; Koc, Yuksel, and Caetano-Anollés, 2018). A review of more resources for pathway analysis is available in Chagoyen and Pazos (2012).

Although pathway knowledge is far from comprehensive, recent developments in computational methods have attempted to achieve metabolite annotations using prior knowledge from metabolic pathways and observed improvements over existent methods. For example, Rogers et al. (2008) report better performance in annotation of LC-MS data when using a connectivity matrix derived from pathways compared to using mass information alone. Gipson et al. (2008) propose instrumental clustering based on high correlation patterns of metabolite peak shapes in the temporal domain (retention times) and mapping to reactant pairs. They show that it increases the accuracy of metabolite assignment compared to a search restricted on mass inputs.

2.4 Summary and discussion

This chapter has provided the necessary background knowledge to understand the basic principles of the central dogma in molecular biology and metabolomics based analysis but it is far from comprehensive. For further readings on the basics of molecular biology the reader is directed to the textbook of biochemistry (Devlin, 2011). Both NMR and MS techniques have been successfully applied to investigate a wide range of effects that can have an influence on the metabolic composition of biological organisms. These effects may include diet (Winning, 2009), adiposity (Elliott, 2015), pregnancy (Maitre, 2014) and a variety of diseases including diabetes, coronary heart disease (Elias-Smale, 2010), Alzheimer’s disease and cancer tumors (Rob, 2011). For a complete introduction NMR and MS as analytical techniques, textbooks of Friebolin and Becconsall (1993) and Gross (2006) are recommended.

Metabolomics deals with multiple and complex chemical reactions within

living organisms and how these are influenced by external or internal perturbations. The resulting metabolotypes (i.e. metabolic profiles) are detected through examination of tissues and biofluids as urine, serum, breast milk. Thanks to the recent efforts towards standardization and reproducibility, a vast array of computational methods and chemometric techniques have been developed to analyze metabolomics data and uncover latent information about the mechanisms leading to phenotypic changes. Due to the inherent heterogeneous nature of these mechanisms, integrative analysis methods for multi-omic data are emerging and gaining popularity among researchers. A detailed survey of data fusion methods for metabolomics and other omics data is covered in Chapter 3.

Metabolite annotation is a necessary step to understanding the underlying biological processes of a studied case. Most of the proposed models need several input parameters such as expected transformations (condensation, hydroxylation, *etc*) or potential adduct forms. This makes the algorithms highly customizable and hinders standardization of the results. Moreover, methods that make use of spectral databases are generally criticized for their slow performance (Gipson et al., 2008). On another hand, despite the uncertainty of the task, metabolomics literature suffers from a lack of probabilistic models for metabolite annotation. In addition to the above limitations, the available methods are generally not applicable for high-throughput metabolomics datasets that have been produced in a fee-for-service manner or where NMR or LC-MS raw data is not available. Following their work on metabolite annotation, Gipson et al. (2008) propose further perspectives to address these shortcomings, mainly the use of context-based annotation methods i.e. integration of metabolite concentrations to the model. Upon the authors recommendation, we introduce a context-based Bayesian model for metabolite annotation in Chapter 6. The model will additionally benefit from the availability of extra omics data to use prior pathway knowledge to enhance the assignments accuracy. A more detailed background on context-based methods can also be found in Chapter 6.

Another additional way to improve biochemical understanding resides in pathway analysis and uncovering associations between variables. Unfortunately, probabilistic pathway analysis in metabolomics is subject to scarce literature. One option consists in introducing biologically consistent knowledge inspired by frameworks described in Section 2.3.5. In Chapter 4, we propose an integrative Bayesian model for longitudinal analysis of time course data which is able to perform pathway analysis and simultaneously uncover associations

between heterogeneous variables in a multivariate framework. The model allows for enrichment analysis in a straightforward manner by construction. As a matter of fact, the chapter is an attempt to enrich the scarce literature not only on probabilistic enrichment analysis, but also on probabilistic inference of linked variables from short time series data. The second option consists in inferring these associations in a purely statistical based approach which is what we try to accomplish in Chapter 5. In order to provide the reader with essential background to Chapter 5, state-of-the-art on related statistical analysis for data integration methods will follow in the next chapter. The next chapter also covers a range of challenges in data integration and gives grounds to the rest of our statistical integrative models so that the reader can appreciate our own contributions to the metabolomics field.

Chapter 3

Statistical multi-omics data integration

The aim of this chapter is to provide the reader with the relevant statistical material to shed light on our research focus and appreciate our own contributions in the appropriate context. The chapter opens up with an overview of types of integrative analysis and strategies for omics data. In Section 3.2.1 of this chapter we briefly cover the concept of Bayesian inference as the principal framework used in this thesis. It follows on with an introduction on popular integrative statistical methods used in metabolomic and omics data analysis with a specific emphasis on integrative regression models (Section 3.2.2) and integrative latent variable models (Section 3.2.3) such as partial least squares (PLS) or canonical correlation analysis (CCA). Integrative clustering is another class of integrative models that will be used in this thesis and will be introduced in Section 3.2.4.

3.1 Multi-omics data integration and challenges

Phenotypic traits of a living organism resulting from internal or external perturbations are of particular interest in omics studies. Along with the appropriate study design, the combination of metabolomics data with the information available from genomics, transcriptomics, and proteomics provides an enhanced readout of the cellular processes related to the phenotype of interest. The need for data fusion ¹ and integration is naturally explained by the complex biological interrelationship between the different layers of the omics cascade where genetic information is transferred from the DNA to the RNA which is then translated

¹data fusion is the process of combining multiple datasets from different sources to reach one dataset that has the properties of the input datasets. It can be seen as a specific case of data integration.

into proteins (Chapter 2). The proteins and the metabolites represent the end products that are closer to the phenotype than genomics and transcriptomics and mainly characterize the cellular biochemistry (Fiehn, 2002). Integrative analysis of these biological entities offers unprecedented possibilities to explore the cooperation between multiple levels to maintain the function of the cell.

Ritchie et al. (2015) define multi-omic data integration as “the incorporation of multi-omic information in a *meaningful* way to provide a more comprehensive analysis of a biological point of interest”. Hence, data integration in omics does not only concern data concatenation, linking, coupling or correlation but most importantly the biological consistency of the combined information. For instance, if variation is assumed to be hierarchical that is unidirectional from the genome to the metabolome, a multi-staged integrative analysis should be privileged (Ritchie et al., 2015). Multi-staged analysis stands for the process of combining data in consecutive steps where, for example, genomic variables are first associated with transcriptomic variables. Significant transcriptomic variables are then associated with metabolotypes. An additional example is where genomic and metabolomic data are separately filtered and associated with a specific phenotype e.g. via GWAS (Genome-Wide Association Studies) and MWAS (Metabolome-Wide Association Studies). The resulting datasets are then tested for mutual association e.g. via metabolome GWAS (Valcárcel et al., 2014; Krumsiek et al., 2011). Multi-staged analysis is often seen as restrictive as omics regulation is not fully hierarchical and involves complex regulatory cascades within and across omics layers, which means that the outcome-predictor structure while combining omics is not established and is variable. However, this approach has successfully been carried out to identify changes in phenotypic traits that are induced by changes in the metabolome which in turn are caused by variation in the genome (Gieger et al., 2008; Nicholson et al., 2011) (Figure 3.1). On the contrary, meta-dimensional analysis supports the hypothesis of simultaneous variation in the genome, transcriptome, proteome and metabolome leading to the phenotype. In other words, the meta-dimensional approach assumes that it is the combination of multiple variables from various data types that results in the phenotype (Smolinska et al., 2012; Witten and Tibshirani, 2009; Cavill et al., 2011).

In addition to the underlying hypothesis, data integration approaches can be defined according to multiple parameters including data types, strategy, study

design, and study focus.²

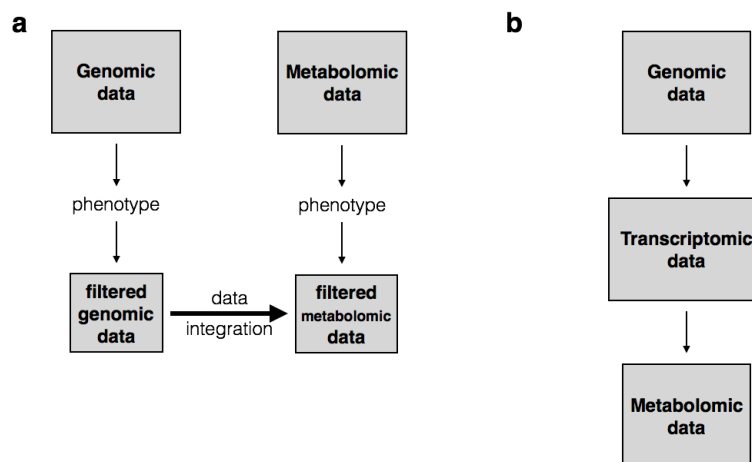


FIGURE 3.1: Examples of a multi-stage integrative analysis approach. Example (a) illustrates a three-step framework where genomic and metabolomic datasets are concurrently tested for association with the phenotype resulting in smaller datasets. These datasets are then investigated to infer linked variables. Example (b) illustrates a typical scenario where genomic variables are tested for association with transcripts which are in turn associated to metabolotypes. These metabolotypes might for instance explain the expression of phenotype.

Data types

Generally speaking, the research community is following two kinds of integrative analysis : *horizontal* or *homogeneous meta-analysis* versus *vertical* or *heterogeneous multi-omics analysis* (Richardson, Tseng, and Sun, 2016; Hamid et al., 2009; Evangelou and Ioannidis, 2013). Horizontal analysis concerns integration of data where similar entities are measured across different studies, cohorts or labs. On the contrary, vertical analysis deals specifically with different entities such as metabolites and proteins measured on the same set of samples. In this thesis, we will mainly focus on vertical multi-omics analysis (as per data availability).

Data integration strategies

Integrative analysis can also be broadly categorized with respect to different strategies : *low level* or *early integration*, *intermediate integration* and *high level*

²Integrative models based on metabolic fluxes and kinetic models are not in the scope of this thesis.

or *late integration* (Figure 3.2). In early integration, all inputs are concatenated at raw or processed level to form a single dataset, gathering all the provided information with minimal loss. Hence, one major challenge in early integration is to use an appropriate common representation for datasets from different scales (Fridley et al., 2012; Lanckriet et al., 2004a). Nonetheless, early integration benefits from two main advantages. First, statistical methods as used for a single data can be applied with slight modifications to the obtained data matrix. Second, it usually preserves information of interaction between omics variables of the input datasets. However, this strategy is subject to increasing the high dimensionality of the data by concatenating the total number of variables from all input data into a single one. Hence, dimensionality reduction techniques or pre-filtering signals that are known to be related to the outcome of interest might be required before performing early integration.

Whereas in early integration transformation shouldn't change the nature of the data, intermediate integration deals with finding a suitable mapping into another representation prior to data combination. This approach covers models that make use of kernel functions or network models of the data. Kernels have been widely used to capture and transform implicit patterns into explicit schemes by representing (projecting) the data into a relevant space (Lanckriet et al., 2004a; Lanckriet et al., 2004b; Smolinska et al., 2012). By contrast to their superior predictive accuracy, a major disadvantage of kernel-based methods is that they are often difficult to interpret. On another hand, network based methods are popular in omics data integration as they offer easy integration (by merging edges for example) and enhanced interpretability (Davis and Chawla, 2011; Guo et al., 2011). In metabolomics, ease of interpretability is a major concern and ensures high functionality of the model. Ease of interpretability is, thus, one of the key aspects we will consider when developing our integrative models.

In late integration, each data type is modeled independently then, the resulting models are used to build an integrative or high level model. In the metabolomics literature, latent variable models, namely PCA and PLS variants, are very popular. They can be used for integrative analysis according to the following procedure (Forshed, Idborg, and Jacobsson, 2007) : Separate models are fitted for each dataset and score matrices are extracted. These score matrices are then concatenated and used as input for an additional model. The latter is termed as high level model. The main limitation of late integration lies

in the fact that correlation structures between different datasets might not be reflected in data-specific models as they are first fitted separately (Žitnik and Zupan, 2015; Acharjee et al., 2016).

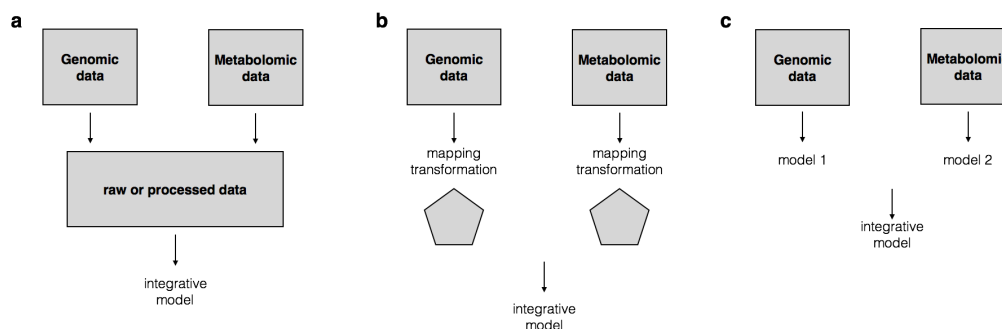


FIGURE 3.2: Different data integration strategies. Figure (a) illustrates early integration where data is combined into a single data matrix before modeling. Figure (b) depicts the intermediate data integration level where data matrices are transformed or mapped into a common meaningful representation before modeling. In figure (c), each data model is generated separately and is then combined with models based on other data types to generate the integrative or high-level model.

Study design

Cavill et al. (2015) have identified four types of study design that should be considered when performing data integration: *repeated study*, *replicate matched study*, *split sample study* and *source matched study*. Briefly speaking, in a repeated study, one omics dataset is first generated following a specific experimental protocol. A second dataset is then obtained by repeating the same experimental protocol usually on a different time or lab. This study design is likely to introduce batch effects into the integrative analysis and is usually complicated to correct for. In certain studies, separate samples from different biological replicates (in the same experiment) are needed to generate metabolomics and transcriptomics data depending on the experimental extraction protocol. This is referred to as replicate matched study. The split sample study is subject to less variation than the repeated or replicate matched study. As its name suggests, it consists in splitting the same biological sample e.g. tissue or biofluid into two samples profiled with different omics technologies. The last case is the source matched study where different fractions of the biological system are used for different analyses for example urine, plasma or faeces. The datasets used in

this thesis advantageously benefit from either the source matched study design (All chapters) or the split sample study design (Chapter 6).

Study focus

It goes without saying that the integrative analysis process is greatly influenced by the primary statistical or biological focus of the study. Three non exhaustive categories of integrative analysis types according to study objectives have been identified by Wang et al. (2012) and Daemen et al. (2009). The first category, called *sequential integration*, attempts to uncover the development of a phenotype e.g disease as opposed to its prediction. This category mainly answers questions on how does data fusion deepen our understanding of the disease? Does the additional data type confirm the findings of the first data type? Does the additional data type enhance our understanding of the first data type? Similarly to multi-staged analysis, sequential analysis comprises integrative approaches that assume a systems hierarchy which is not true in real life due to e.g. regulatory loops. Such analyses were conducted by e.g. Kleemann et al. (2007), Santos et al. (2009), and Verhoeckx et al. (2004). In general, the authors firstly identify genes that are associated to external perturbations or disease. Secondly, genes are linked to metabolites and corresponding enriched pathways. Sequential analysis also comprises disease subtype discovery, commonly achieved via clustering approaches (Sun et al., 2018; Tang et al., 2014; Shen et al., 2012; Le Van et al., 2016; Yuan, Savage, and Markowetz, 2011). For instance, Le Van et al. (2016) propose a clustering model that simultaneously identifies features related to each subtype. In this approach data is integrated via ranked transformation. In Yuan, Savage, and Markowetz (2011), the designed model, termed as PSDF, allows clustering of different types of discrete genomic data to identify cancer subtypes, feature selection and infer whether patients exhibit similar profiles across data types. The PSDF model makes use of the Dirichlet process to infer probabilistic cluster assignments and Bayesian hierarchical modeling to integrate genomic data (See Section 3.2.4). As it uses discrete data with similar scales, initial data transformation is not required for the PSDF model, however, typically data transformation needs to be realized before applying integrative clustering models.

An important literature body in multi-omics analysis involves two additional types of integrative analysis: *biological integration* and *model-based integration*. The aim of biological integration is to uncover the biological mechanisms of

interaction between heterogeneous variables including metabolic pathways, regulatory mechanisms and signaling mechanisms (Hirai et al., 2005; Cavill et al., 2011; Safo, Li, and Long, 2018; Hong et al., 2013)³. In particular, Safo, Li, and Long (2018) developed a sparse canonical correlation analysis (CCA) model to uncover hidden association patterns between heterogeneous data where sparsity is adjusted based on structural information of biological networks, particularly edge information among variables within pathways. In simple words, prior network structural information (extracted from KEGG for metabolites and mum-michog for genes) is used to guide selection of relevant genes and metabolites in sparse CCA. CCA is a popular integrative method and will be discussed in more details in Section 3.2.2 and chapter 5.

In model-based integration, researchers are faced with a range of statistical questions such as which omics variables are associated with the phenotypic changes? Which groups of variables from the different datasets are interacting? Does data fusion improve predictive accuracy of phenotype, disease, temporal behavior? Is information expressed by the different data types redundant? In this context, an important range of statistical and machine learning methods have been developed in the literature. By way of illustration, kernel-based approaches were proposed to integrate multi-platform metabolomics data such as NMR and GC-MS (Smolinska et al., 2012) and multi-platform genomics data (Daemen et al., 2009). Both authors show that predictive ability of the integrative model significantly outperforms predictive ability of models based on a single type of data. Žitnik and Zupan (2015) used matrix factorization to integrate 11 data types to predict gene function in *D.discoideum* and similarly shows that the integrative model significantly improves prediction compared to single models and is more robust to technical and methodological biases. In order to find significant linear combinations between copy number and gene expression data, Witten and Tibshirani (2009) developed a supervised sparse canonical correlation model where datasets were linked to a clinical outcome.

To conclude, different multi-omics integration approaches can be further classified according to multiple dimensions. Broadly speaking, *data types* and *study design* are parts of the experimental dimension whereas the *strategy* types are parts of the methodological dimension. Finally, *study focus* and underlying

³By contrast to sequential analysis where one data is analyzed then a second one is used to confirm or deepen results from the first analysis (the emphasis is not specifically about cellular mechanisms), the focus in biological integration is directly related to underlying cellular mechanisms.

Integrative analysis	Description	Examples
<i>hypothesis</i>		
Multi-staged	Inter-omic variation is hierarchical	Gieger et al. (2008), Nicholson et al. (2011), and Krumstiek et al. (2011)
Meta-dimensional	Inter-omic variation is simultaneous	Smolinska et al. (2012), Witten and Tibshirani (2009), and Daemen et al. (2009)
<i>data types</i>		
Horizontal or homogeneous	Same omics entities are measured across various cohorts, labs or studies	Yuan, Savage, and Markowitz (2011) and Richardson, Tseng, and Sun (2016)
Vertical or heterogeneous	Entities from different omics levels are measured using different platforms	Richardson, Tseng, and Sun (2016) and Evangelou and Ioannidis (2013)
<i>strategy</i>		
Early integration	Concatenation-based integration	Fridley et al. (2012) and Le Van et al. (2016)
Intermediate integration	Data transformation needs to be applied prior to modeling	Lancriet et al. (2004a), Smolinska et al. (2012), and Guo et al. (2011)
Late integration	Based on combining single data models into a high level model	Acharjee et al. (2016)
<i>study design</i>		
Repeated study	Study is repeated in another time or place to generate a second type of data	Cavill et al. (2015)
Replicate matched study	Biological replicates are used to generate additional data	Cavill et al. (2015)
Source matched study	Samples from the same organism e.g. animal are extracted	Cavill et al. (2015)
Split sample study	Samples e.g. tissues are split for profiling with different omics technologies	Cavill et al. (2015)
<i>study focus</i>		
Sequential analysis	Does the additional data type enhance understanding of the first data type?	Shen et al. (2012), Le Van et al. (2016), and Yuan, Savage, and Markowitz (2011)
Biological analysis	What are the underlying processes leading to phenotypical changes?	Hirai et al. (2005), Cavill et al. (2011), Safo, Li, and Long (2018), and Hong et al. (2013)
Model-based analysis	Which variables are phenotypically relevant? significantly associated? Can predictive ability be improved?	Smolinska et al. (2012), Witten and Tibshirani (2009), and Daemen et al. (2009)

TABLE 3.1: Classification of different data integration approaches. The examples list is by no means exhaustive.

variation *hypothesis* reflect the biological dimension regardless of the adopted statistical method. Table 3.1 summarizes different data integration classes depending on *hypothesis*, *data types*, *strategy*, *study design* and most importantly *study focus*.

Study focus is of crucial importance to performing *meaningful* and efficient integrative analysis. In this thesis we mainly focus on *biological analysis* (Chapter 4), *model-based analysis* (Chapter 5) and *sequential analysis* (Chapter 6) frameworks. Metabolomics is a highly suitable source for deriving biomarkers under these frameworks as it is the closest layer of the omics cascade that is related to the phenotype. In fact, metabolic profiling is widely used to study genotype-metabotype interactions or metabotype-phenotype interactions such as disease-relevant phenotypes or external stimuli. Identifying interactions between omics variables either in terms of significant statistical associations, biomarker discovery or biological networks enhances data interpretability and represents the end goal of many studies. For the sake of interpretability, an arsenal of mathematical and computational techniques has been developed to achieve such analyses. These techniques include correlation analysis (Safo, Li, and Long, 2018; Witten and Tibshirani, 2009), integrative regression models (Gieger et al., 2008; Nicholson et al., 2011; Krumsiek et al., 2011) and Bayesian integrative clustering of gene profiles (Yuan, Savage, and Markowitz, 2011; Kirk et al., 2012; Savage et al., 2010). A particular emphasis is given to each one of these techniques in the next section.

3.2 Statistical background and gaps

This section comprises a review of the relevant statistical state-of-the-art techniques for the audience of this thesis to place our own contributions in context of others. The section also highlights the areas where potential shortcomings present research opportunities. Part 3.2.1 serves as a tutorial on Bayesian inference as the main inferential framework used in this thesis. Part 3.2.2 covers different integrative regression techniques in the context of biomarker discovery across multiple omics datasets (i.e. cross-omics biomarker discovery). Integrative regression is used as a component of the model developed in Chapter 4 to perform statistical analysis and enrichment analysis (Figure 2.3 in Chapter 2). Part 3.2.3 introduces latent variable chemometrics techniques used to uncover

linear cross-omics dependencies (Chapter 5). Finally, part 3.2.4 introduces integrative clustering models applied to functional analysis of multi-omics data (Chapter 6).

3.2.1 Bayesian inference

There are at least two different main approaches in statistical inference: the frequentist approach and the Bayesian approach. The frequentist approach is the most favored approach in the metabolomics field. However, the Bayesian approach is recently gaining popularity amongst metabolomics scientists thanks to its ability to quantify uncertainty. In fact, in the frequentist approach, generally the model likelihood is considered. In other words, if θ represent the model parameters and $\mathbf{x}$ represent the observation values, then commonly frequentist inference reduces to finding θ^* which maximizes the likelihood $p(\mathbf{x}|\theta)$.

On the contrary, Bayesian models represent an elegant way to consider uncertainty in the model by introducing prior distributions. Henceforth, instead of finding a point estimate θ^* , we now want to sample from a posterior distribution defined by the Bayes theorem as follows:

$$\underbrace{p(\theta|\mathbf{x})}_{\text{posterior}} = \frac{\underbrace{p(\mathbf{x}|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}}{\underbrace{p(\mathbf{x})}_{\text{marginal likelihood}}}$$

The Bayes theorem shows how from a generative model $p(\mathbf{x}|\theta)$ of the dataset combined with a prior knowledge or belief $p(\theta)$, one can make inferences on the variable of interest in the light of the observed data (Barber, 2012). The use of generative models enhances interpretability and fits well with models in metabolomics especially where prior information such as pathways, biochemical reactions, chemical properties are available. Bayesian nonparametric models represent an extension of Bayesian models. In fact, in the Bayesian framework, prior distributions represent uncertainty over model parameters, whereas in the Bayesian nonparametric framework uncertainty is modeled over probability density functions. We remind the reader that, in addition to previously exposed aims, this thesis also aspires to encourage use and integration of Bayesian and probabilistic models by metabolomics scientists as it has been limited so far.

3.2.2 Integrative regression

One of the major questions that a metabolomics scientist is faced with in cross-omics studies is : How much do we know about $\mathbf{Y}$ given a value of $\mathbf{X}$? Recall that $\mathbf{X}$ and $\mathbf{Y}$ are two centered heterogeneous datasets with dimensions $p \times n$ and $q \times n$ respectively as per the used notation in this thesis. Linear models have proven efficiency into identifying relations between homogeneous⁴ and heterogeneous⁵ omics variables (Schmidt et al., 2007; Steuer, 2006; Camacho, De La Fuente, and Mendes, 2005). In the case where $\mathbf{Y}$ is a metabolomics dataset and $\mathbf{X}$ is a genomic dataset, one way to answer the question above is to simply use metabolome GWAS models (Gieger et al., 2008; Nicholson et al., 2011; Krumsiek et al., 2011). Metabolome GWAS is an integrative regression technique that is used in the context of biomarker discovery where the outcome variables are metabolic features. This is referred to as cross-omics biomarker discovery. The main limitation of these models is that estimations are achieved in a univariate way without taking into account the existing interaction between covariates in the $\mathbf{X}$ or $\mathbf{Y}$ matrix.

Due to the “large p , small n ” problem characteristic of omics data, multi-variate regression by introducing the whole $\mathbf{X}$ matrix in the regression model can not be performed due the high correlation structure inherent to omics data. Hence several penalization terms were introduced to form shrinkage models. In simple regression one aims to minimize the mean squared prediction error:

$$\text{MSE} = \text{Tr}(\text{E}(\mathbf{Y} - \mathbf{b}^T \mathbf{X})(\mathbf{Y} - \mathbf{b}^T \mathbf{X})^T) = \|\mathbf{Y} - \mathbf{b}^T \mathbf{X}\|_2^2$$

where $\|\cdot\|_2$ is the l_2 norm, E is the expectation and $\mathbf{b}$ is a $p \times q$ matrix of regression coefficients. The least squares estimator of $\mathbf{b}$ can be written as $\mathbf{b}^* = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}\mathbf{Y}}$ but is not applicable in high dimensional setting due to the singularity of the empirical estimation of $\boldsymbol{\Sigma}_{\mathbf{X}}$. Adding regularization terms, the objective function in penalized regression is such as :

$$F(\mathbf{b}) = \text{Tr}(\text{E}(\mathbf{Y} - \mathbf{b}^T \mathbf{X})(\mathbf{Y} - \mathbf{b}^T \mathbf{X})^T) + R(\boldsymbol{\lambda}, \mathbf{b})$$

where $R(\boldsymbol{\lambda}, \mathbf{b})$ is a positive regularization term depending on $\mathbf{b}$ and the shrinkage vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_q)$. The shrinkage parameter $\boldsymbol{\lambda}$ controls the amount of shrinkage i.e. the larger is λ_i the larger the penalty is on the regression

⁴variables representing the same omics entities.

⁵variables representing entities from different omics layers.

coefficients $\mathbf{b}_i$ and the more they are shrunk towards zero. Some of the popular penalty terms include the ridge penalty (Hoerl and Kennard, 1970) with $R(\lambda, \mathbf{b}) = \sum_{i=1}^q \lambda_i \|\mathbf{b}_i\|_2^2$ and the lasso penalty (Tibshirani, 1996) with $R(\lambda, \mathbf{b}) = \sum_{i=1}^q \lambda_i \|\mathbf{b}_i\|_1$ where $\|\cdot\|_1$ is the l_1 norm. The lasso regression is commonly used as a penalized variable selection method since it introduces sparsity for high values of λ_i . The term lasso stands for “least absolute shrinkage and selection operator” but is also motivated by the picture of catching variables with a lasso. Ridge regression, on the other hand, offers a smooth estimate of $\Sigma_{\mathbf{X}}$ to ensure non-singularity. It also induces shrinkage on the regression coefficients providing the following solution $\mathbf{b}_i^{ridge} = (\Sigma_{\mathbf{X}} + \lambda_i \mathbf{I}_p)^{-1} \Sigma_{\mathbf{X}} \mathbf{y}_i$. Such enclosed form can not be expressed in case of lasso estimator. It is possible, however, to show that depending on the choice of λ_i , the l_1 penalty sets some regression coefficients to zero. Compared to the lasso, the ridge regression only shrinks coefficients towards zero but they are generally not exactly zero regardless of the value of λ_i . Unlike ridge regression, lasso will select at most n variables for high dimensional data $n \ll p$. In addition, variable selection via lasso performs poorly in the case where variables are collinear or highly correlated (Fu, 1998). In fact, lasso will only select one variable from a group of correlated variables whereas it is preferable to include the whole group in the model (Meier, Van De Geer, and Bühlmann, 2008). To address these limitations, elastic net conciliates both penalties $R(\lambda, \mathbf{b}) = \sum_{i=1}^q \lambda_i^1 \|\mathbf{b}_i\|_1 + \sum_{i=1}^q \lambda_i^2 \|\mathbf{b}_i\|_2^2$. By combining both the lasso and the ridge penalties, elastic net stabilizes the l_1 regularization path, removes the limitation on the number of selected variables and encourages grouping effect (Zou and Hastie, 2005).

The Bayesian framework benefits from the flexibility to define priors on model parameters and thus is not subject to heavy tuning procedures. A natural Bayesian approach is to impose sparsity via a “spike-and-slab” prior (Mitchell and Beauchamp, 1988) where the spike is modeled by a delta function at zero and the slab by a heavy tailed continuous distribution. This model was heavily criticized because of its poor scalability to high-dimensional data (Van Der Pas, Kleijn, and Van Der Vaart, 2014). For applications on large datasets, several other priors have been investigated. For instance, the Laplacian or double exponential prior on the regression coefficients represents the underlying Bayesian model of the lasso. Another shrinkage model is the relevance vector machine of Tipping (2001) in kernel regression. The model is mainly composed of Student-t priors with mean 0 on the regression coefficients, and a prior on

the scales of the Student-t distributions leading to a scale mixture model. The shrinkage amount in this model is controlled by the scale parameters of the Student-t distribution. Namely, the smaller are the scales, the more important is the shrinkage and the closer the coefficients are to 0.

In a similar fashion, (Carvalho, Polson, and Scott, 2009) proposed the horseshoe prior based on a scale mixture of normals where scale parameters are modeled as the product of a global shrinkage (scale) parameter and a local shrinkage (scale) parameter. This definition allows for an additional flexibility where sparsity can be controlled at a global level (i.e. how many non-zero coefficients?) and a local level (i.e. which coefficients are non-zero?). The horseshoe prior has been widely recognized and extended by the statistical community since its introduction by Carvalho, Polson, and Scott (2009) as it benefits from various desirable properties such as simple analytic form, easy computation and preservation of significant coefficients (no over-shrinkage) (Van Der Pas, Kleijn, and Van Der Vaart, 2014; Bhadra et al., 2017). For a more data specific review on variable selection in high dimensional data in the context of genomics, we refer the reader to Richardson, Bottolo, and Rosenthal (2010). In Chapter 4, we make use of the horseshoe prior as part of a more comprehensive model for integrative analysis, especially cross-omics biomarker discovery, in time-course metabolomics data.

3.2.3 Latent variable dimension reduction methods for data integration

Integrative regression introduced in the previous section focuses on inference of inner linear relationships mainly between heterogeneous variables. We now turn our attention to a more general setting of linear relationships in chemometric methods, called latent variable models where both homogeneous and heterogeneous relationships are taken into account. The goal of these methods is to find linear relationships between data matrices $\mathbf{X}$ and $\mathbf{Y}$ based on projections. In order to reduce the dimensionality of the data, the general scenario involves projecting the omics variables into a lower dimensional space or latent space. The $\mathbf{X}$ and $\mathbf{Y}$ variables are then transformed into a set of few latent variables in new coordinates usually termed as *components*. In mathematical terms this consists in finding matrices $\mathbf{W}_\mathbf{X}$ and $\mathbf{W}_\mathbf{Y}$ of dimension $m \times p$ and $m \times q$, where m is the number of components and $m \leq \min(p, q)$, such that the p -dimensional (and q -dimensional) data variables are mapped into m -dimensional variables i.e.

$\mathbf{Z}^{\mathbf{X}} = \mathbf{W}_{\mathbf{X}}\mathbf{X}$ and $\mathbf{Z}^{\mathbf{Y}} = \mathbf{W}_{\mathbf{Y}}\mathbf{Y}$. Some of the most well recognized latent variable method in metabolomics are the PCA (*principal component analysis*) and PLS (*partial least squares*) derived methods. In PCA (Pearson, 1901; Hotelling, 1933), the $\mathbf{Z}^{\mathbf{X}}$ and $\mathbf{Z}^{\mathbf{Y}}$ are *score* matrices that account for much of the variance in $\mathbf{X}$ and $\mathbf{Y}$ respectively whilst $\mathbf{W}_{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{Y}}$ are the matrices containing the eigenvectors of $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{Y}}$. They are called PCA *whitening* or *sphering* matrices (Chapter 5). Scaled eigenvectors are called *loadings*.

PLS (Wold et al., 1984; Wold, 1975) is a latent variable regression model that aims to maximize the covariance between the x -scores and y -scores under the constraint that either score components are orthogonal or loadings are orthogonal. PLS remains one of the most popular methods used in chemometrics mainly thanks to its ability to distinguish variability due to the y -variables which significantly enhances interpretability. It encompasses many variants including PLS1, PLS2 and O2-PLS (Trygg, 2002; Trygg and Wold, 2003). PLS1 is used for regression with a single y -variable whereas PLS2 assumes multiple y -variables i.e. a data matrix $\mathbf{Y}$. However, rather than estimating an inner relationship between $\mathbf{X}$ and $\mathbf{Y}$, both data matrices are modeled by the following linear latent variable models:

$$\mathbf{X} = \underbrace{\mathbf{Z}^{\mathbf{X}^T}}_{x\text{-scores}} \underbrace{\mathbf{A}}_{x\text{-loadings}} + \underbrace{\mathbf{E}_{\mathbf{X}}}_{\text{error matrix}} \quad (3.1)$$

$$\mathbf{Y} = \underbrace{\mathbf{Z}^{\mathbf{Y}^T}}_{y\text{-scores}} \underbrace{\mathbf{B}}_{y\text{-loadings}} + \underbrace{\mathbf{E}_{\mathbf{Y}}}_{\text{error matrix}} \quad (3.2)$$

with $\mathbf{A} = (\alpha_1, \dots, \alpha_p)^T$, $\mathbf{B} = (\beta_1, \dots, \beta_q)^T$. The scores $\mathbf{Z}^{\mathbf{X}}$ and $\mathbf{Z}^{\mathbf{Y}}$ are then connected through the following linear regression model:

$$\mathbf{Z}^{\mathbf{Y}} = \mathbf{Z}^{\mathbf{X}}\mathbf{D} + \mathbf{E} \quad (3.3)$$

where $\mathbf{D}$ is a diagonal matrix and $\mathbf{E}$ is the error matrix. Considering that $\mathbf{Z}^{\mathbf{X}}$ is constructed, $\mathbf{B}$ is obtained by plugging equation 3.3 into equation 3.2. The result is the least squares solution of the regression between $\mathbf{Z}^{\mathbf{X}}$ and $\mathbf{Y}$, namely $\mathbf{B} = (\mathbf{Z}^{\mathbf{X}}\mathbf{Z}^{\mathbf{X}^T})^{-1} \mathbf{Z}^{\mathbf{X}}\mathbf{Y}$. The most complicated step in PLS consists in finding $\mathbf{Z}^{\mathbf{X}}$. First, one needs to notice that finding the first PLS components coincides with finding α_1 and β_1 that maximize the objective function:

$$\text{cov}(\alpha_1^T \mathbf{X}, \beta_1^T \mathbf{Y}) \quad \text{s.t.} \quad \|\alpha_1^T \mathbf{X}\|_2 = 1 \quad \text{and} \quad \|\beta_1^T \mathbf{Y}\|_2 = 1 \quad (3.4)$$

In addition to the length constraint, usually additional constraints on orthogonality of either loading vectors or score vectors is required. The solution of the maximization problem corresponds to the largest singular value of $\Sigma_{\mathbf{X}\mathbf{Y}}$.

Different algorithms have been developed to infer the remaining components such as NIPALS (Wold, Sjöström, and Eriksson, 2001) and SIMPLS (De Jong, 1993). However, these algorithms are based on iterative procedures where the general scenario is to remove the variability of the component found at the previous iteration from either the datasets (NIPALS) or the covariance matrix (SIMPLS) by projecting into a plane that is orthogonal to the found component.

PLS remains one of the most popular methods used in chemometrics mainly thanks to i) its efficiency for small sample data with highly correlated variables, ii) ability to distinguish variability due to the y-variables which significantly enhances visualization and interpretability. In addition, O2-PLS fulfills another objective for integrative analysis compared to PLS2. By explicitly modeling orthogonal variation to $\mathbf{Y}$ through extra components $\mathbf{Z}_o^{\mathbf{X}^T} \mathbf{A}_o$ such that $\mathbf{X} = \mathbf{Z}^{\mathbf{X}^T} \mathbf{A} + \mathbf{Z}_o^{\mathbf{X}^T} \mathbf{A}_o + \mathbf{E}_\mathbf{X}$ (and similarly for $\mathbf{Y}$ using $\mathbf{Z}_o^{\mathbf{Y}^T} \mathbf{B}_o$), O2PLS guarantees straightforward interpretation on 3 levels : i) common variation to $\mathbf{X}$ and $\mathbf{Y}$ via $\mathbf{Z}^{\mathbf{X}}$, ii) specific variation to $\mathbf{X}$ via $\mathbf{Z}_o^{\mathbf{X}}$ and iii) specific variation to $\mathbf{Y}$ via $\mathbf{Z}_o^{\mathbf{Y}}$. Other PLS derived methods have been successfully used for data integration. OnPLS (Löfstedt and Trygg, 2011) extends O2-PLS to more than 2 datasets. Sparse PLS (Lê Cao et al., 2008; Fonville et al., 2010) extends PLS for variable selection using lasso penalization for integrated omics. Sparse PLS optimizes the square error terms with a penalty term of loading vectors of response matrix $\mathbf{Y}$ and predictor matrix $\mathbf{X}$. Group PLS (gPLS) and sparse group PLS (sgPLS) (Liquet et al., 2015; Lê Cao et al., 2009) are PLS variants to take into account the natural grouping structure of omics variables such as pathways or time points. The sparse group PLS approach allows selection at both groups of omics variables and single variable levels.

Canonical correlation analysis (CCA) (Hotelling, 1935) is another essential tool employed to discover shared trends between two data matrices. In a similar fashion to PLS, CCA is concerned with maximizing the correlation between $\mathbf{X}$ and $\mathbf{Y}$ rather than the covariance. Particularly, in canonical correlation analysis the aim is to find mutually orthogonal pairs of maximally correlated linear combinations of the components of $\mathbf{X}$ and of $\mathbf{Y}$. Specifically, we seek canonical directions α_i and β_j (i.e. vectors of dimension p and q , respectively) for which $\text{cor}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \Lambda$ is maximal such that $\text{cor}(\mathbf{A}\mathbf{X}) = \mathbf{I}$

and $\text{cor}(\mathbf{BY}) = \mathbf{I}$. The projected vectors $\mathbf{AX}$ and $\mathbf{BY}$ are the CCA *scores* or the *canonical variables*, $\mathbf{A}$ contains the *canonical correlations*.

Classical CCA uses the empirical covariance matrix to obtain canonical correlations and directions. Since the latter can only be safely employed if the number of observations is much larger than the dimensions of either of the two random vectors $\mathbf{X}$ and $\mathbf{Y}$, different variants of CCA that are applicable to high dimensional problems have been developed. For example, Parkhomenko, Tritchler, and Beyene (2009) and Waaijenborg, Witt Hamer, and Zwinderman (2008) show that penalized CCA is able to identify associations between heterogeneous omics variables. Other variants of CCA make use of probabilistic models to facilitate application of Bayesian inference and other regularized statistical procedures (Bach and Jordan, 2005).

Benefiting from enhanced interpretability, probabilistic models motivate our approach to CCA in Chapter 5. In our work, we also link CCA to PLS-derived methods and latent multivariate regression and provide an optimized implementation for high dimensional settings. Furthermore, we illustrate our approach on integrative statistical analysis of metabolomics datasets by extending it to network modeling ⁶.

3.2.4 Integrative clustering

Cluster analysis involves identifying groups in a data set such that data points within a group are similar according to a chosen criteria and groups are dissimilar according to the same criteria. Clustering has become increasingly important in a variety of applications in exploratory analysis—to summarize the data—as well as advanced analysis—to uncover hidden structure in the data—. For multi-omics data analysis, integrative clustering has been successfully used for disease subtype discovery (Shen, Olshen, and Ladanyi, 2009; Shen et al., 2012; Yuan, Savage, and Markowetz, 2011) and functional analysis (Kirk et al., 2012; Savage et al., 2010). For instance, Shen, Olshen, and Ladanyi (2009) have used a latent factor representation of the K-means clustering algorithm to build an integrative clustering tool, termed as *iCluster*. The potential of applying *iCluster* for multi-omics data is shown in e.g. Shen et al. (2012) where the authors have identified three distinct molecular cancer subtypes that showed different survival patterns when further investigated. However, one of the limitations of

⁶Model-based integrative analysis (table 3.1) that can be performed as part of step iv in figure 2.3 in Chapter 2.

iCluster lies in the fact that a single common clustering structure is imposed on all datasets, regardless of whether or not they exhibit similar signals. In addition, iCluster is subject to the major classic issue in cluster analysis: The choice of the number of clusters. Generally, one needs to assume that there are exactly K clusters representing the data or resort to heuristic procedures to select the number of clusters.

In the Bayesian nonparametric (BNP) paradigm, a clustering problem consists of inferring both the number of clusters and the data assignments. The Dirichlet process (DP) has been widely applied for that purpose. Generally speaking, the idea is to assume that the number of potential clusters is infinite⁷ but the data only exhibits some of the clusters (a finite number). A range of applications of the DP in biostatistics were explored in Dunson, 2010. These approaches seek to benefit from the power of the Bayesian and Bayesian nonparametric settings to increase the flexibility of clustering models. Mixture models play an important role as the generative models underlying the clusters. The traditional density function of a mixture model is given by:

$$p(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{k=1}^K \pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k) \quad (3.5)$$

where π_k are non negative mixing weights of each component (cluster) k and $p(\mathbf{x}_i|\boldsymbol{\theta}_k)$ are the mixture densities with parameters $\boldsymbol{\theta}_k$. It is often useful to introduce a latent indicator variable z_i that conditionally assigns $\mathbf{x}_i$ to one and only component such that $p(\mathbf{x}_i|z_i = k, \boldsymbol{\Theta}) = p(\mathbf{x}_i|\boldsymbol{\theta}_k)$. The model below and (3.5) are equivalent :

$$p(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_k p(z_i = k) p(\mathbf{x}_i|\boldsymbol{\theta}_k) \quad (3.6)$$

For integrative analysis, one of the popular applications under the BNP framework is to combine mixture models with the hierarchical Dirichlet process (HDP) (Teh et al., 2006). The latter was used by Yuan, Savage, and Markowetz (2011) and Savage et al. (2010) for data integration of gene expression with copy number variation and gene expression with chIP-chip to build the PSDF model. This model uses a two-level hierarchy of Dirichlet processes : First each dataset is modeled by a separate Dirichlet process then these are integrated by means of a third Dirichlet process. The principal innovation of the model is to include an indicator variable for each gene that determines whether the

⁷This means that the number of clusters can potentially grow with new observations

gene should join a cluster based on both data sources combined (via a product of likelihoods) or whether it should be clustered separately for each dataset; which allows to identify concordant and discordant signals between the datasets (Savage et al., 2010). In a similar fashion, Kirk et al. (2012) introduced MDI for “Multi Dataset Integration” where integrative clustering is not performed via hierarchical Dirichlet process but directly through “fusion” parameters ϕ_{dm} of data d and data m such that, for a total of D datasets, $p(z_{d1}, \dots, z_{Di}) = \prod_{d=1}^D \pi_{d, z_{di}} \prod_{d=1}^{D-1} \prod_{m=d+1}^D (1 + \phi_{dm} \mathbf{I}_{[z_{di}=z_{mi}]})$. The larger is ϕ_{dm} the more likely groups of co-clustering genes in dataset d will have the same label in dataset m . Although, fusion parameters allow a rich interpretation of integrative clustering models, they are only eligible for horizontal data integration where the same entities i.e. genes are profiled by means of different technologies. Yet, to the best of our knowledge, these approaches have not been applied to metabolomics data.

Our contribution to integrative clustering takes advantage of both the flexibility of the Bayesian nonparametric paradigm to infer clusters and the hierarchical models to integrate metabolomics data with genomics data. In the first place, we use functional clustering to group similar omics profiles into modules. Next, we explore the HDP model to integrate data-specific clusters. On top of that, we investigate combining the HDP with biological information to help metabolite identification⁸.

3.3 Summary and discussion

The parallel acquisition of high-throughput datasets from the genome, metabolome, proteome, and transcriptome of the same biological system has rapidly emerged in molecular biology. The integrated analysis multi-omics data is motivated by finding synergies in different data types. It principally deals with the association, correlation, combination, coupling and linking of data from multiple sources to improve cross-omics biomarker discovery and deepen our understanding of living organisms and underlying biological mechanisms.

In this chapter we have provided an overview on different data integration types which we categorized according to different parameters involving underlying hypothesis, data types, integration strategy, study design and study focus. The chapter also covered three state-of-the-art statistical frameworks used for

⁸Table 3.1. Sequential analysis : Does one data type improve our understanding of the second one?

integrative analysis and which will be adopted in this thesis. These frameworks encompass integrative regression, integrative latent variable models and integrative clustering. They have mainly been applied for integration of different types of genomic, transcriptomic and proteomic data, but rarely with metabolomic data. Further surveys on other multi-omics integrative methods are available in Richardson, Tseng, and Sun (2016), Hamid et al. (2009), and Tseng, Ghosh, and Zhou (2015). In this thesis, we will develop three different statistical models to integrate metabolomics data under each of these frameworks to investigate different aspects, namely uncovering hidden associations, biomarker discovery and metabolite identification. Besides the ability to deal with metabolomics data, additional challenges arise depending on the focus of the integrative analysis:

- *Integrative regression for cross-omics biomarker discovery* : Metabolomics is a highly suitable source for deriving phenotype biomarkers as well as cross-omics biomarkers since it integrates genetic as well as non-genetic factors. Regression is one of the elementary methods used for biomarker discovery. In a similar fashion to GWAS and MWAS, metabolome GWAS was widely applied to integrate of metabolomic and genomic data but is criticized as analysis is performed in a univariate way. Due the correlation structures inherent to omics data, multivariate regression can be achieved by introducing penalty terms in the frequentist setting or shrinkage priors in the Bayesian setting. Yet, these methods ignore dependence between metabolites in favor of genotype-metabotype dependence. Hence, one of the fundamental challenges that arise in this class of models it to simultaneously model metabolite-metabolite associations and metabolite associations with other omics entities. Biological entities are complex by nature and are arguably regulated by sequences of actions and complex interactions. In this sense, modeling a sequence of observations naturally regulated by chemical processes has proven successful in computational biology. Accordingly, our iCARH model developed in Chapter 4 seeks to take into account heterogeneous and homogeneous associations by combining the horseshoe prior with current biochemical pathway knowledge. The latter allows to model metabolite-metabolite associations in a biologically consistent way.
- *Integrative latent variable models to uncover multi-omics associations*: Metabolomics data are characterized by a high number of metabolic profiles compared to the number of biological samples. Moreover, metabolomic

variables are also regulated by complex and strong correlation patterns as explained in Section 2.3. Henceforth, dimensionality reduction techniques are of fundamental importance in chemometrics for ease of visualization and interpretation. Dimensionality reduction techniques such as PCA and PLS derived techniques usually involve maximizing a covariance function under orthogonality constraints. Although these methods offer rich interpretation in terms of shared and orthogonal components, it is not straightforward to quantify associations between the different variables and thus limits interpretation with regards to variables' mutual interrelationships. CCA is principally distinguished from PLS by two aspects : First, the objective function: CCA seeks to maximize a correlation function instead of a covariance function. Second, CCA involves the maximization of a single objective function whilst PLS is based on iterative heuristics. Hence, CCA is more liable to improve with regards to its interpretability. We address this issue by constructing a probabilistic model for CCA in Chapter 5. Our main contribution consists in providing a latent variable model for CCA based on whitening transformations⁹. We show that this model enhances interpretability as it offers easy derivation of association measures between heterogeneous variables.

- *Integrative clustering for metabolite identification* : Clustering for functional analysis was explored by (Manikandan, Ramyachitra, and Banupriya, 2016; Becker et al., 2012; Yi, Sze, and Thon, 2007) in the Bayesian parametric setting and proven to provide more understanding of the forces underlying cellular processes and an unbiased method for researchers to identify related functional clusters. In the BNP setting more flexible models were implemented by (Savage et al., 2010; Yuan, Savage, and Markowitz, 2011; Kirk et al., 2012) via hierarchical models where the notion of “fusion” state was introduced. Hierarchical models offer more degrees of freedom than one-level models and thus allows to define for each data its own parameters that might (or not) be shared. Nonetheless, these models are only applicable for homogeneous integrative analysis. One of the fundamental challenges in this context, hence, lies in the heterogeneity of multiple data types. In addition, to the best of our knowledge, infinite functional clustering has not been investigated for metabolomics data. In Chapter 6 we combine network information from pathway databases with

⁹whitening will be covered in Chapter 5

infinite clustering via hierarchical Dirichlet processes. We will investigate the potential of this model to improve metabolite identification.

Furthermore, one of our major contributions in this thesis consists in increasing the offer of probabilistic models tailored for metabolomics data as the corresponding literature is scarce. We focus on modeling uncertainty by exploiting probabilistic, Bayesian and Bayesian nonparametric models. Bayesian models extend probabilistic models by specifying prior probability distributions on model parameters based on prior information available from data or expert knowledge. Whereas Bayesian nonparametrics directly extend Bayesian models by defining priors on probability functions. The power of Bayesian models lies mainly in their flexibility to contain different levels depending on the complexity of the data. In the next chapter, we illustrate the flexibility inherent to the Bayesian framework by developing a single model, in the context of integrative regression, where different sources of variation are taken into account. For the sake of interpretability, the model also allows for probabilistic enrichment analysis.

Chapter 4

Integrative Longitudinal Analysis and Metabolic Biomarker Discovery ¹

4.1 Background

There is a growing interest in longitudinal experiments for heterogeneous omics data and statistical models to infer biomarkers of a particular treatment or disease over time. Longitudinal data or time-course data involves repeated measurements of the same variables at different time points. Longitudinal metabolic phenotyping experiments carried along with additional longitudinal omics experiments are key experiments to characterize a biological system's response to perturbations. Observing the temporal evolution of metabolic profiles, in addition to other omics profiles, helps understand their roles in the underlying biological processes and elucidate how they interact with other biological entities. Identifying metabolites with significant association patterns with a particular phenotype (disease, clinical variables, physical traits, etc) is of major interest in metabolic phenotyping experiments and is referred to as *biomarker discovery* stage. Identified biomarkers can be used as an indicator of that specific phenotype.

As mentioned in chapter 2, metabolomic data sets are characterized by high correlation structures in that many spectral peaks can arise from the same metabolite and metabolites operate within networks of chemical reactions (pathways). In addition, further correlation structure is present in longitudinal metabolomic studies due to repeated measurements of observations over time. Additional challenges include not only the low number of time points

¹Material in this chapter is based on Jendoubi and Ebbels, 2018

and samples compared to the number of profiled metabolic variables, but also integration of a different omics data to the metabolomic data.

First, metabolomic time series are often short due to experimental costs or ethical considerations. Typically, less than 10 time points are available compared to a large number of metabolic variables profiled at each time point e.g. hundreds of metabolic variables for targeted experiments and thousands of metabolic variables for untargeted experiments. Taking into account the small number of time points compared to the large number of metabolic variables profiled, the number of temporal patterns that will arise is limited (due to the limited number of degrees of freedom). Some temporal patterns will be repeated and thus these patterns can be induced by randomness. Second, models fitted to a small number of data points are often prone to overfitting i.e. the model is very sensitive to small fluctuations. This can lead to a poor fitting to unseen data and a high generalization error. Third, it is also important to consider the number of parameters of the statistical model and make use of the simplest models in order to avoid overparametrization. Finally, monitoring heterogeneous omics variables can substantially enhance the understanding of the underlying biological mechanism and provide a systems biology approach as these omics variables represent entities that are often involved in related cellular processes (Joyce and Palsson, 2006; Ebrahim et al., 2016).

Presently, available tools for dynamic metabolomic data analysis can be divided to two main categories : univariate methods and multivariate methods. In fact, different approaches attempt to infer influential or significant metabolites using dynamic metabolomic data under the assumption that metabolites are independent. These models include fitting smooth splines mixed effects models (SME) to time curves (Berk, Ebbels, and Montana, 2011) and linear mixed effects models augmented with a variable selection approach (Mei, Kim, and Tsui, 2009). However between-metabolite correlation is richly structured and biologically relevant and should be modeled.

Seemingly unrelated regression accounts for metabolite correlation by using correlated regression errors and can be used to identify biologically significant metabolites (Chen et al., 2015; Chen et al., 2017). In gene expression data analysis, Pham et al., 2015 recently proposed to use confirmatory factor analysis to capture gene-pathway relationship and a conditional autoregressive model to capture relationships between a set of pathways where pathway network has

been constructed based on KEGG (Kanehisa and Goto, 2000) pathways. The latter accounts for biological variation in the data and allows ease of interpretation.

In the metabolomics literature, traditional frameworks for metabolomic data analysis use dimensionality reduction techniques, namely principal component analysis (PCA), partial least squares (PLS) (Wold, Martens, and Wold, 1983) and PLS derived models (O-PLS (Trygg and Wold, 2002), O2-PLS (Trygg and Wold, 2003), On-PLS (Löfstedt and Trygg, 2011)) to take into account high correlation between metabolites. Extension of PLS to O2-PLS and On-PLS allows for integrative analysis of heterogeneous omics data. PCA and PLS models are very popular among metabolomicists as tools for exploratory data analysis, visualization purposes and ease of interpretation. One of the interests of PCA (PLS) derived models is to be able to visually assess whether or not there is a time effect in the data and identify metabolites that change over time by looking into time trajectories of each metabolite (Antti et al., 2002).

Extensions to PCA (PLS) for longitudinal analysis include lagged PCA (PLS) and dynamic PCA (PLS) where a backshift matrix is introduced to take into account time dependency. Another extension to PLS is termed as multilevel PLS (Westerhuis et al., 2010; Lique et al., 2012; Jain et al., 2018) where variance decomposition is first performed prior to applying PLS on the temporal variation matrix. In the more general context of repeated measurements, temporal variation is also called the within subject variation. Group (or sparse group) PLS is an additional variant of PLS that was developed by (Lique et al., 2015; Lê Cao et al., 2009) to take into account groups of omics variables in the context of biomarker discovery. The group PLS approach enables linear modeling of interrelationships between omics variables of the same group e.g. biological pathways, functional modules or of the same time point. The sparse group PLS approach allows selection at both groups of omics variables and single variable levels. Group (sparse group) PLS might be combined with multilevel PLS in order to fully model metabolite dependencies and temporal correlations. In a different fashion, Rantalainen et al., 2008 used a set piecewise orthogonal projections latent structures to describe changes between neighboring time points. PARAFAC (Bro, 1997) is a multi-linear unsupervised decomposition method that can account for the multi-way variation seen in dynamic metabolomics data. Similar models such as ASCA (Smilde et al., 2005)

and APCA (Harrington et al., 2005) also seek to account for temporal variation by combining analysis of variance (ANOVA) with PCA. Recently, dynamic probabilistic PCA (DPPCA) was proposed in Nyamundanda, Gormley, and Brennan, 2014 as a probabilistic generative model of longitudinal metabolomic data where a stochastic volatility model is used for the latent variables. Probabilistic generative models are used to describe how the data is generated in terms of probability distributions. Generative models do not attempt to explain the generative process, they are rather an attempt to mimic as closely as possible the unknown underlying process that generated the data for e.g. prediction purposes or to generate new data points. The main inconvenience of these approaches is that further techniques such as variable importance scoring methods have to be separately applied to the data in order to identify biomarkers and they do not take heterogeneous data (i.e. data from different omics techniques) into account.

Our model is based on a hierarchical shrinkage prior to provide a natural framework for sparse problems inherent to omics data and allow for integrative analysis. Under the Bayesian framework different priors can be used for variable selection. For instance, the spike-and-slab (Mitchell and Beauchamp, 1988; Efron, 2008) prior is weighted mixture of a point mass at zero (spike) and a zero-centered Gaussian with a high variance (slab). The zero-centered Gaussian allows for non-zero coefficients whilst the point mass at zero puts coefficients at exactly zero. This prior, however, has poor performance for high dimensional data. Further development on the spike-and-slab prior that can scale reasonably well to high dimensional data was proposed by Scheipl, Fahrmeir, and Kneib, 2012 where a bimodal distribution is defined on the variance of the coefficients rather than the coefficients themselves. The authors also developed the `SpikeSlabGAM` package (Scheipl, 2011) combining the proposed prior and generalized additive mixed models. Bhattacharya et al., 2015 introduced the Dirichlet-Laplace prior allowing for straightforward posterior computations for high dimensional data. In the same context, Carvalho, Polson, and Scott, 2009 introduced the horseshoe estimator for sparsity handling. The horseshoe prior is based on a scale mixture of normals where the mixing parameters are half-Cauchy distributed (Polson and Scott, 2010). Carvalho, Polson, and Scott, 2009 showed that the prior performed well in shrinking the low coefficients and preserving the high coefficients. The horseshoe prior was extended by Bhadra et al., 2017 to the “horseshoe+” prior where an additional mixing parameter

with a standard half-Cauchy distribution is introduced. This additional mixing parameter allows for ultra sparsity i.e. better shrinkage of noise terms than the standard horseshoe prior. In this work, we will use a different natural extension of the horseshoe prior by using a half t-distributed rather than a half-Cauchy distributed mixing parameter.

To summarize, many statistical methods currently used to analyze short time-series omics data are

- prone to overfitting, or
- do not take into account the experimental design, or
- do not make full use of the multivariate information intrinsic to the data, or
- do not provide statistical integrative analysis

Hence, appropriate models are needed to integrate heterogeneous omics data and take into account both experimental conditions and biological variations in order to extract important information from the data. In addition, cautious estimates of model parameters are needed in order to ensure robust interpretation of the results.

This chapter is concerned with dose-response time course experiments where additional omics variables (bacteria, genes, transcripts, etc) are monitored along with metabolites in the context of biomarker discovery. Our aim is to build a *single* model for integrative analysis of time course metabolic data in an attempt to:

- overcome overfitting by using a weakly informative Bayesian model
- capture experimental design conditions through a mixed-effects model
- models metabolite interactions by using pathway information through a conditional auto-regressive (CAR) component
- uncovers multiple associations between metabolites and other omics variables by using a horseshoe prior

An additional benefit of our approach is that it naturally yields a list of perturbed metabolic pathways since it is based both on a mixed effects component and a CAR component. Since it is based on both a CAR model and

a horseshoe prior, we call our model the integrative CAR Horseshoe (iCARH) model. Model implementation is available in the iCARH R package on CRAN.

This chapter is based on the integrative CAR Horseshoe (iCARH) model published in Jendoubi and Ebbels, 2018.

4.2 Model Framework

Given a metabolomics data $\mathbf{X} \in \mathbb{R}^{n \times h \times p}$ where n is the number of observations, h the number of time points and p the number of metabolites. $\mathbf{Y} \in \mathbb{R}^{n \times h \times q}$ is an additional continuous omics data measured along with $\mathbf{X}$ where q is the number of associated omics variables. The set of n observations consists of a set of cases and controls. Throughout the paper, index i always runs through observations, index t runs through time points, index m runs through metabolites and index k runs through $\mathbf{Y}$ variables. Vector quantities are written in bold. Matrices are written in bold capitals. The important feature of the study designs we are interested in is the existence of repeated measures of the same observations through time. Our goal is to build a simple model that can identify biomarkers relative to a specific treatment in time taking into account the multiple sources of variations in the data.

The model is built on three levels: First, a CAR component to capture dependencies between metabolites. Second, a variable selection model to uncover associations between metabolites and $\mathbf{Y}$ data. Third, a mixed effects component to model experimental design. We give more details about each level of our model respectively in each of the following sections.

4.2.1 Metabolite dependencies

In any integrative biological model, it is useful to be able to interpret the model at a systems level, e.g. according to functional groups of biological molecules, rather than attempting to interpret results for individual molecules. Metabolic pathways are the most widely used groupings for this type of analysis in metabolomics, and have been widely used to interpret experimental data, usually by performing over-representation or enrichment analyses (Xia and Wishart, 2010; Kamburov et al., 2011; Chagoyen and Pazos, 2012; Kankainen et al., 2011; Gao et al., 2010). Since pathways are regulated in a coordinated fashion, it is natural to assume that the levels of metabolites which are members of the same pathway may be correlated. This dependence, though weak, is

observable in associations between correlations and network distance (Walther et al., 2010; Gipson et al., 2008; Krumsiek et al., 2011). For example, figure 4.1 from Walther et al. (2010) show that pairwise correlation levels in yeast decrease when the shortest pairwise pathway distances increase whether yeast is exposed to high or low temperature. A similar pattern is also observable in the metformin data we used as a case study in section 4.3.2. We therefore incorporate a pathway-based correlation component into our model via a CAR approach, in a similar fashion to Pham et al., 2015. The extent of pathway specific correlations will vary according to the experimental system and assay, and may in some cases provide little extra information. Nonetheless, including such a pathway based component can greatly increase the interpretability of the resulting model beyond one which does not include such grouping information.

2

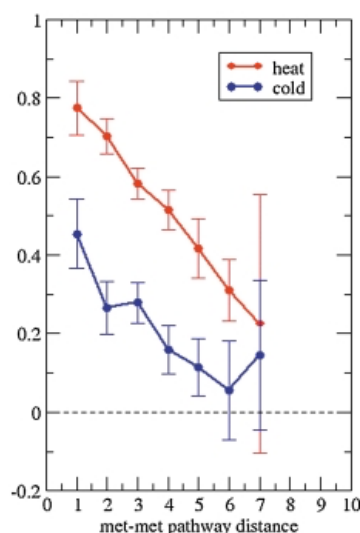


FIGURE 4.1: Figure from Walther et al. (2010). The figure shows that pairwise correlation levels (y axis) decrease when the shortest pathway distance between metabolites increase for different temperature settings.

The Conditional Autoregressive (CAR) model

Conditional autoregressive (CAR) models (Besag, 1974) are commonly used to describe the spatial variation of quantities of interest. These models have been successfully used in multiple areas and are very popular in spatial statistics

²For non annotated metabolites, pathway information for these metabolites is ignored but inference of treatment effect and association with other omics variables can be seamlessly carried out (See implementation details in section 4.4).

including demography, economy and epidemiology as a tool to uncover and quantify spatial relations in the data. The term “conditional” in the CAR model refers to the conditional neighboring structure where each element is specified conditionally on its neighbors. Defining a neighboring structure is in fact key to determining the dependence structure of the CAR model (Cressie and Wikle, 2015; Cressie, 1993; Banerjee, Carlin, and Gelfand, 2014). In spatial statistics, a site of interest is partitioned into different sub-sites indexed by $m = 1, \dots, p$ where variables of interest are monitored and measurements s_m are recorded³. The collection of sub-sites is assumed to follow a well defined neighborhood system where, in a well known emblematic example, neighboring regions to sub-site i are those sharing a boundary with i . Multiple other neighborhood systems can be defined based on distance or shared features (Cressie, 1993; Case, Rosen, and Hines Jr, 1993). Assuming that observations s_m have zero mean, the following formulation holds under the CAR model:

$$s_m | \mathbf{s}_{-m} \sim N \left(\sum_{\substack{m'=1 \\ m' \neq m}}^M c_{mm'} s_{m'}, \sigma^2 \right) \quad (4.1)$$

where $\mathbf{s}_{-m}$ represents measurements in all sub-sites excluding sub-site i , $c_{mm'} \geq 0$ are covariance parameters with $c_{mm} = 0$ for all m and $\sigma^2 > 0$. Besag (1974) demonstrated that equation 4.1 can lead to a full joint distribution for all random variables:

$$s_m | \mathbf{s}_{-m} \sim N \left(0, (\mathbf{I}_p - \mathbf{C})^{-1} \sigma^2 \right) \quad (4.2)$$

where $\mathbf{I}_p$ is the p th order identity matrix. This full joint distribution is only well defined if $\mathbf{C}$ is symmetric and $\mathbf{I}_p - \mathbf{C}$ is positive definite. Specifying the neighborhood structure gives rise to a “neighborhood” or “adjacency” matrix $\mathbf{A}$ closely related to the matrix $\mathbf{C} = (c_{mm'})$ ⁴. Different specifications of $\mathbf{C}$ as a function of $\mathbf{A}$ have been considered in the literature (Cressie, Perrin, and Thomas-Agnan, 2005). A common example determines $\mathbf{C}$ to satisfy $\mathbf{C} = \phi \mathbf{G} \mathbf{A}$ with $a_{mm'} > 0$ if m and m' are neighbors and 0 otherwise. $\mathbf{G}$ is a scaling diagonal matrix containing the reciprocal of the number of neighbors for each sub-site.

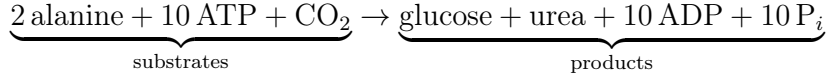
³We will use the notation s_m for measurements in this paragraph to avoid confusion with the notation adopted for our model framework.

⁴Based on these neighborhood matrices, CAR models can also be seen as network-based models or graphical models.

The model is valid only if $\mathbf{I}_p - \mathbf{C} = \mathbf{I}_p - \phi \mathbf{G} \mathbf{A}$ is positive definite i.e. ϕ is required to belong to $(1/xi^1, 1/xi^2)$ the maximum and minimum eigenvalues of $\mathbf{G} \mathbf{A}$. More general frameworks of the CAR model comprises Markov random fields which also appear in Cressie and Wikle (2015), Cressie, Perrin, and Thomas-Agnan (2005), Cressie (1993), and Banerjee, Carlin, and Gelfand (2014). In the next section we will describe how the CAR model can be used to model metabolite dependencies.

Using CAR to model metabolite dependencies

First, we assume that the concentration of each metabolite is linearly influenced by concentration levels of metabolites in the same pathway. Such an assumption is legitimate as, if unperturbed, chemical reactions impose linear relation between substrates (right hand side) and products (left hand side). For example, the chemical reaction below represents glucose synthesis from alanine in the liver. Two alanine molecules are metabolized to produce one glucose molecule i.e. considering this reaction alone, a decrease by two units in alanine leads to an increase by one unit in glucose.



Although this assumption can not be verified for all chemical reactions e.g. redox reactions, linear models in general have proven to be efficient in capturing metabolite-metabolite associations by means of e.g. linear mixed effects models or PLS methods (Siddiqui et al., 2018; Wold et al., 1984; Mei, Kim, and Tsui, 2009)⁵.

In the following, let $\mathbf{C} \in \mathbb{R}^{p \times p}$ be the design matrix quantifying metabolite interactions such that matrix elements $c_{mm} = 0$, $c_{mm'} \neq 0$ if metabolites m and m' are in the same pathway and 0 otherwise. Thus, metabolite intensity levels can be expressed as:

$$x_{itm} | \mathbf{x}_{it,-m}, \boldsymbol{\mu}_{it}, \mathbf{C}, \sigma \sim N(\mu_{itm} + \sum_{\substack{m'=1 \\ m' \neq m}}^M c_{mm'}(x_{itm'} - \mu_{itm'}), \sigma^2) \quad (4.3)$$

⁵In order to capture non linear effects, non linear conditional autoregressive models can be applied (Lee and Mitchell, 2013)

where $\mathbf{x}_{it,-m}$ represents measurements of metabolites of sample i at time point t excluding metabolite m , μ_{itm} is a function of covariates of sample i for metabolite m at time point t . With $\mathbf{I}_p$ the p th order identity matrix, the joint distribution of $\mathbf{x}_{it}$ can be explicitly written as (Cressie and Wikle, 2015):

$$\mathbf{x}_{it}|\mu_{it}, \mathbf{C}, \sigma \sim N(\mu_{it}, (\mathbf{I}_p - \mathbf{C})^{-1} \sigma^2) \quad (4.4)$$

Chemists are most interested by identifying which pathways are “on” or “off” as an effect of treatment. In the CAR literature, the design matrix $\mathbf{C}$ can be modeled as a scaled product of a diagonal weight matrix and an adjacency matrix. In order to allow for pathway perturbation inference we construct the distance matrix based on the individual contribution of each pathway. To be precise, we define $\mathbf{C}(\phi) = \sum_{j=1}^l \phi_j \mathbf{G}_j \mathbf{A}_j$ where l is the number of pathways. The distance matrices $\mathbf{A}_j$ are a zero-diagonal symmetric matrices with elements $a_{mm'}^p$ equal to the inverse of the length of the shortest path between metabolites m and m' if they are in pathway j and 0 otherwise. A path between two metabolites consists in the number of reactions that lead from one metabolite to the other. The shortest path is the path that contains the smallest number of reactions⁶. This prior information on the distances matrices can be extracted from pathway databases. A variety of pathway databases comprise KEGG (Kanehisa and Goto, 2000), BRENDA (Schomburg et al., 2002) and MetaCyc (Caspi et al., 2013). Pathway information can vary from database to database depending on its definition in a given database. Although a pathway remains biologically consistent with a specific cellular function across databases, pathway length and distances between a pair of metabolites can be different across databases. In this work we solely used KEGG pathways, however, this could be extended to other databases⁷.

In the diagonal matrices $\mathbf{G}_j$ we use the reciprocal of the number of neighbors of each metabolite in pathway j i.e. $(g_{mm'}^j)^{-1} = \sum_{m'=1}^M (a_{mm'} > 0)$ so that the squared partial correlation $\text{cor}(x_{itm}, x_{itm'} | \mathbf{x}_{it,-(m,m')})^2 \propto \phi_j^2 g_{mm}^j g_{mm'}^j$ is reduced when more metabolites from the same pathway are profiled (Cressie and Wikle,

⁶The distance matrices $\mathbf{A}_j$ can also be modeled in a binary way where $a_{mm'}^p = 1$ if metabolites m and m' are in the same pathway and 0 otherwise. Both models gave similar results for the metformin data as a study case (See section 4.3.2).

⁷The number of non-annotated metabolites can constitute an important proportion of the metabolomics data. In this case, pathway information will not be considered for non-annotated metabolites which is equivalent to modeling these metabolites as independent from the rest of the annotated metabolites.

2015). The model parameter $\phi = \{\phi_j\}_{j=1}^l$ is estimated. It quantifies pathway contribution and is referred to as *spatial*-dependence parameter in the CAR literature.

The model needs to comply with the condition that $\mathbf{I}_p - \mathbf{C}(\phi)$ is positive definite. If we assume that pathways are *a priori* equally perturbed, ϕ_j must fall in the interval $\left(\frac{1}{l\xi_j^1}, \frac{1}{l\xi_j^2}\right)$ where ξ_j^1 and ξ_j^2 are the minimum and maximum eigenvalues of $\mathbf{G}_j\mathbf{A}_j$, respectively. In practice, strong interaction between observed metabolites of pathway j is reproduced in CAR models only when the scaling parameter ϕ_j is quite close to one of the boundaries $\frac{1}{l\xi_j^1}, \frac{1}{l\xi_j^2}$. Hence, we use a beta-type prior for ϕ_j that places substantial mass on large values of $|\phi_j|$ (Banerjee, Carlin, and Gelfand, 2014) :

$$p(\phi_j) = \frac{1}{\mathbf{B}\left(\frac{1}{2}, \frac{1}{2}\right)} \left(\phi_j - \frac{1}{l\xi_j^1}\right)^{-\frac{1}{2}} \left(\frac{1}{l\xi_j^2} - \phi_j\right)^{-\frac{1}{2}} \quad (4.5)$$

where $\mathbf{B}$ is the beta function. The parameter σ^2 captures variance heterogeneity in metabolite intensities and is given an inverse gamma prior $\mathbf{G}(\psi, \psi - 1)$. This prior provides 2ψ pseudo-observations in addition to $N \times T$ available observations. In order to build a reasonably informative prior we set $\psi = N \times T/4$.

4.2.2 Integrative analysis

Interactions between heterogeneous omics variables such as transcripts and metabolites or gene expression and metabolites is modeled via the following hierarchical shrinkage model:

$$\mu_{itm} = \alpha_m + \gamma_{im} + \beta_m \mathbf{y}_{it} + \nu_{itm} \quad (4.6)$$

$$\beta_{mk} | \lambda_{mk}, \sigma_{\beta_m} \sim N(0, \lambda_{mk}^2 \sigma_{\beta_m}^2) \quad (4.7)$$

$$\lambda_{mk} | \tau \sim \text{St}^+(\tau, 0, 1) \quad (4.8)$$

A summary of these model parameters is depicted in table 4.1. St^+ denotes the half Student-t distribution with τ degrees of freedom, α_m represents treatment effect for metabolite m , $\gamma_{im} \sim N(0, \sigma_{\gamma_m}^2)$ represents individual perturbations for metabolite m , $\nu_{itm} | \nu_{i,t-1,m} \sim N(\theta_m \nu_{i,t-1,m}, \sigma_{\nu_m}^2)$ follows an auto-regressive process and represents temporal effects for metabolite m of individual i at time point t (See Section 4.2.3). Finally, β_m is a vector of size q that quantifies associations between metabolite m and the other q omics variables. λ_{mk} is called local shrinkage parameter whilst $\sigma_{\beta_m}^2$ is the global shrinkage parameter.

For $\tau = 1$, this prior reduces to the horseshoe prior (Carvalho, Polson, and Scott, 2009). Intuitively, for small values of λ_{mk} the coefficient β_{mk} is very close to 0 while for relevant variables λ_{mk} will be large. In addition, σ_{β_m} controls the overall shrinkage level i.e sparsity of the vector β_m is more important for small values of σ_{β_m} . We will simply refer to this prior as the “horseshoe like” prior.

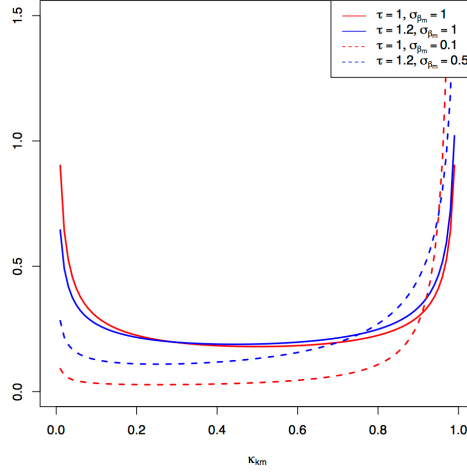


FIGURE 4.2: Prior $p(\kappa_{mk}|\tau, \sigma_{\beta_m})$ on κ_{mk} for different values of σ_{β_m} and τ . The prior distribution skews towards 1 if τ increases or σ_{β_m} decreases.

Define $\kappa_{mk} = \frac{1}{1 + \lambda_{mk}^2 \sigma_{\beta_m}^2 / \tau}$ a random shrinkage coefficient such that $\kappa_{km} \approx 0$ when λ_{mk} is large and $\kappa_{km} \approx 1$ when λ_{mk} is small. This transformation implies the following prior distribution on κ_{mk} :

$$p(\kappa_{mk}|\tau, \sigma_{\beta_m}) = \frac{1}{2\sqrt{\pi}\mathbf{B}(\frac{\tau}{2}, \frac{1}{2})} \frac{\sigma_{\beta_m}^\tau \kappa_{mk}^{\tau/2-1} (1 - \kappa_{mk})^{-1/2}}{(1 - \kappa_{mk} + \kappa_{mk} \sigma_{\beta_m}^2)} \quad (4.9)$$

This prior density is shown in figure 4.2 for different values of σ_{β_m} and τ . It reduces to a $\text{Beta}(\tau/2, 1/2)$ distribution if $\sigma_{\beta_m} = 1$ and to a $\text{Beta}(1/2, 1/2)$ which looks like a horseshoe, if in addition $\tau = 1$. When τ increases, $\text{Beta}(\tau/2, 1/2)$ skews towards 1 which increases the global shrinkage power. The expectation of β_m given $\mathbf{Y}, \kappa_m, \tau, \mu_{tm}$ can be expressed as:

$$\mathbb{E}(\beta_m|\mathbf{Y}, \kappa_m, \tau, \mu_{tm}) = \left(\sum_{t=1}^T \mathbf{Y}_t^T \Sigma_m^{-1} \mathbf{Y}_t + \tau \Upsilon_m \right)^{-1} \sum_{t=1}^T \mathbf{Y}_t^T \Sigma_m^{-1} \mu_{tm} \quad (4.10)$$

where $\Sigma_m = \left(\frac{\sigma_{\nu_m}^2}{1-\theta_m^2} + \sigma_{\gamma_m}^2 \right) \mathbf{I}_p$ and Υ_m is a diagonal matrix of order K with elements $1/\kappa_{mk} - 1$. Equation (4.10) introduces a penalty term $\tau \Upsilon_m$ where Υ_m is a metabolite specific penalty term introduced by the horseshoe prior and τ is a global penalty term. Precisely, τ captures the overall sparsity level amongst all metabolites. The expectation of β_m given $\mathbf{Y}, \kappa_m, \tau, \mu_{tm}$ is very similar to the estimate of β_m under ridge regression where $\tau \Upsilon_m$ simply reduces to $\tau \mathbf{I}_p$. The global sparsity level can be controlled using τ . Increasing the global sparsity level is a desired property in omics studies, as usually we deal with a large number of omics variables where only few are important. Moreover, when there is prior knowledge available, specifying τ *a priori* can optimize the inference and additionally, provide a more informative prior on λ_{mk} . If we fix $p(\sigma_{\beta_m}^2) \propto 1/\sigma_{\beta_m}^2$, integrating over σ_{β_m} gives the expected value of κ_{mk} as :

$$\mathbb{E}(\kappa_{km}|\tau) = \frac{\Gamma(1/2)^{-1}}{2\sqrt{\pi}\Gamma(\tau/2)} \mathbf{G}_{3,3}^{2,3} \left(\begin{matrix} 1, \tau/2, 0 \\ \tau/2, \tau/2 - 1/2, 0 \end{matrix} \middle| 1 - \sigma_{\beta_m}^2 \right)$$

where $\mathbf{G}_{3,3}^{2,3}$ is Meijer's G-function (Meijer, 1936). The equation above can be used to fix τ *a priori* by defining the expected proportion of shrunk coefficients. In practice, different values of τ are plugged into the equation above to get the desired proportion of shrunk coefficients. However, many definite integrals can be obtained using the tables of Meijer functions in Brychkov (2008) for special values of parameters.

4.2.3 Experimental design

Biological organisms are dynamical by nature hence the necessity to perform longitudinal experiments in omics. In this sense, modeling a sequence of consecutive observations naturally ordered through time is fundamental in computational biology. Here, we use "Markov models" to model temporal variation ν_{itm} in equation (4.6). "Markov models" is a very popular class of statistical models that represent a generative model where a variable is influenced by its past. Due to the limited number of longitudinal measurements, we use the first order Markov model for the temporal effects ν_{itm} for metabolite m of individual i at time point t i.e. ν_{itm} is influenced by its immediate previous state such that $p(\nu_{itm}|\nu_{i,t-1,m}, \dots, \nu_{i,1,m}) = p(\nu_{itm}|\nu_{i,t-1,m})$. Under a widely used Markov model, known as the first order auto-regressive model, we further assume that

$\nu_{itm} | \nu_{i,t-1,m} \sim N(\theta_m \nu_{i,t-1,m}, \sigma_{\nu_m}^2)$. In other words, ν_{itm} is a noisy linear function of its previous observation where the noise term is assumed to be zero-mean Gaussian with variance $\sigma_{\nu_m}^2$.

The model can be further extended to take into account other aspects of the experimental design. As specified in the previous section, α_m captures the treatment effect for metabolite m and γ_{im} represents individual perturbations for metabolite m in equation (4.6).

Moreover, the covariance structure between metabolites might change drastically as a result of treatment if the latter affects metabolic pathways. In this sense, we allow covariance structures $\mathbf{C}(\phi^e)$ to be different for the control samples and the cases where $e \in \{\text{cases}, \text{controls}\}$ designates experimental groups. This yields the overall hierarchical model:

$$\mathbf{x}_{it}^e | \boldsymbol{\mu}_{it}, \mathbf{C}, \sigma \sim N(\boldsymbol{\mu}_{it}, (\mathbf{I}_p - \mathbf{C}(\phi^e))^{-1} \sigma^2) \quad (4.11)$$

$$\mu_{itm} = \alpha_m + \gamma_{im} + \boldsymbol{\beta}_m \mathbf{y}_{it} + \nu_{itm} \quad (4.12)$$

$$\beta_{mk} | \lambda_{mk}, \sigma_{\beta_m} \sim N(0, \lambda_{mk}^2 \sigma_{\beta_m}^2) \quad (4.13)$$

$$\lambda_{mk} | \tau \sim \text{St}^+(\tau, 0, 1) \quad (4.14)$$

$$\gamma_{im} | \sigma_{\gamma_m} \sim N(0, \sigma_{\gamma_m}^2) \quad (4.15)$$

$$\nu_{itm} | \theta_m, \sigma_{\nu_m} \sim N(\theta_m \nu_{i,t-1,m}, \sigma_{\nu_m}^2) \quad (4.16)$$

Note that by specifying different dependence parameters for metabolite interactions in cases and controls metabolism, the model is able to identify perturbed pathways by comparing ϕ^{cases} and ϕ^{controls} .

As discussed in chapter 2, most of the current pathway analysis approaches are based on evaluating differential expression of a group of functionally related metabolites. These approaches, however, examine each functional module individually and often produce a list of modules with high similarity and redundancy. In addition, usually identification of one differentially expressed metabolite is sufficient to estimate all the corresponding pathways as perturbed. As a result, it is very difficult to derive biological meaning from the obtained list, especially for modules containing very few metabolites. Compared to these methods, our approach using the CAR model is more advantageous in the sense that it evaluates perturbation as a change in the correlation structure ϕ^{cases} and ϕ^{controls} between metabolites. Most importantly, it preserves the aspect of

collaboration between metabolites and is consistent with the functional definition of a metabolic pathway (Faust, Croes, and Helden, 2011) resulting in a more intuitive and straightforward interpretation. All models parameters are depicted in the table below.

Parameters of interest	
$\phi^e, e \in \{\text{cases, controls}\}$	quantifies metabolite interactions, Vector of size p $(\phi^{\text{cases}} - \phi^{\text{controls}})$ quantifies pathway perturbation
β_m	Vector of size q quantifies association between metabolite m and other omics variables from different omics techniques
α_m	scalar, quantifies treatment effect for metabolite m
Other inferred parameters	
σ^2	metabolite variance
γ_{im}	represents individual perturbations for metabolite m
ν_{itm}	represents temporal effects for metabolite m of individual i at time point t
θ_m	temporal dependence for metabolite m
λ_{mk}	local shrinkage parameter
$\sigma_{\beta_m}^2$	global shrinkage parameter
$\sigma_{\gamma_m}^2$	variance of individual perturbations
$\sigma_{\nu_m}^2$	temporal variance
Fixed parameters	
C	between metabolite adjacency matrix
τ	overall sparsity level amongst all metabolites

TABLE 4.1: iCARH model summary

4.3 Results

In this section we perform experiments on both synthetic and real data to investigate whether our algorithm gives reasonable solutions. We first try our method on a simulated dataset in Section 4.3.1 to get an understanding of the performance of our method. In Section 4.3.2, we test our method on a data set

using metabolomic and bacterial composition in a drug treatment experiment. In the following, we refer to our model as “ iCARH ” model for “ integrative CAR Horseshoe ” model.

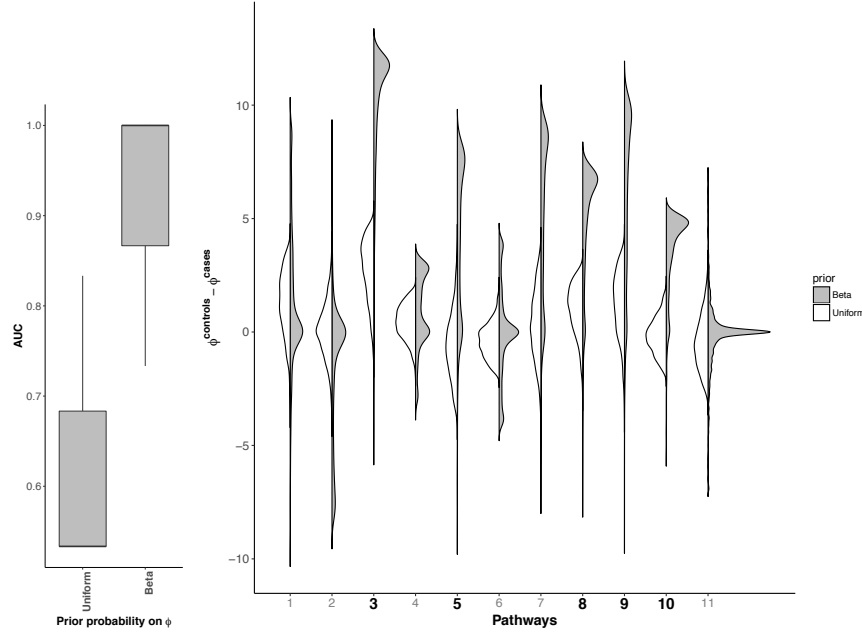


FIGURE 4.3: Left : Boxplots of Area Under the Curve (AUC) for pathway perturbation inference with uniform prior on ϕ compared to a beta like prior on ϕ for 10 simulated datasets. We infer perturbations based on the posterior distribution of $\phi^{\text{controls}} - \phi^{\text{cases}}$. Right : Distribution of $\phi^{\text{controls}} - \phi^{\text{cases}}$ under the uniform prior (white) and the beta like prior (grey). True perturbed pathways are printed in bold on the x axis.

4.3.1 Simulation study

To get better understanding of our method and test its applicability, we first perform our approach on synthetic datasets. We will mainly focus on the ability of our model to infer pathway perturbation and association between heterogeneous variables.

Assessing pathway inference with beta like prior

In this simulation our objective is to assess how the beta like prior in equation (4.5) improves the iCARH model. We first fixed the number of pathways P to 11, then simulate the design matrices $\mathbf{A}_j$. Specifically, a membership matrix $\mathbf{Z}$ with dimensions $p \times l$ is randomly generated based on the density of the number of KEGG pathways in which a single metabolite is involved. Each

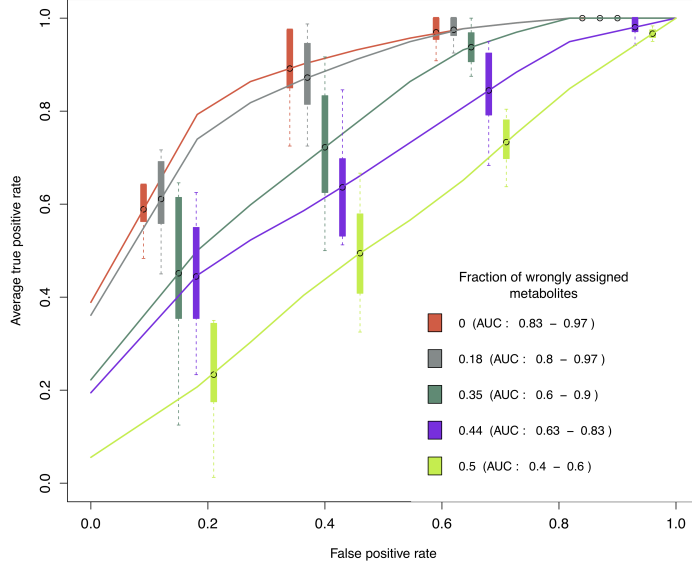


FIGURE 4.4: Average Receiver Operating Characteristic (ROC) curves for pathway perturbation inference across 10 datasets for different fractions of “wrongly” assigned metabolites.

design matrix $\mathbf{A}_j$ is then equal to $\mathbf{z}_j \mathbf{z}_j^T$ where $\mathbf{z}_j$ is the j th column of $\mathbf{Z}$. We simulate pathway perturbation according to an indicator variable ω as follows: If $\omega = 1$ then pathways are perturbed hence ϕ_p^{controls} and ϕ_p^{cases} are simulated from normal distributions with different means. If $\omega = 0$ then $\phi_p^{\text{controls}} = \phi_p^{\text{cases}}$ and pathways are not perturbed. The rest of the parameters is set as follows : number of bacterial variables $q = 1$, number of metabolites $p = 40$, number of time points $h = 7$, number of samples $n = 22$, parameters ν_{itm} , γ_{im} , μ_{itm} , $\mathbf{x}_{it}^e$ simulated according to equations (4.16), (4.15), (4.12), (4.11) respectively. The global sparsity level can be controlled using τ . τ is fixed to 1.2 in order to approach the Horseshoe prior behavior.

We set non-informative uniform priors on α_m , $\sigma_{\gamma_{im}}$, θ_m , $\sigma_{\mu_m}^2$. We set an informative prior on $\sigma_{\gamma_m}^2 \sim \text{inverse-gamma}(1, 0.1)$ as we expect low variability amongst biological samples of the same group. This assumption is often suitable for experiments where conditions are well controlled (e.g. cell lines, rat, mice, yeast) but might not be suitable for human studies due to multiple sources of heterogeneity (Gelman, 2006). In the latter case non-informative priors on $\sigma_{\gamma_m}^2$ should be preferred. We fix π_ω to 0.7 the proportion of expected perturbed pathways. ρ is fixed to a value of 0.05, σ_ϕ^2 to 0.2 and ψ to a large value. We compare inference of the model under a uniform prior for ϕ_j^e and the prior in

equation (4.5). Inference is done using 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations.

The left plot in figure A.1 shows the boxplots of the Area Under the Curve (AUC) for pathway perturbation inference for 10 simulated datasets with uniform prior on ϕ and a beta like prior on ϕ . We infer perturbations based on the posterior probability that ϕ^{controls} and ϕ^{cases} are different i.e. the 95% credible interval of $\phi^{\text{controls}} - \phi^{\text{cases}}$ does not contain zero. The AUC values for the beta like distribution, reaching 1, are significantly higher than the AUC values for the uniform distribution. On average pathway perturbation inference under the uniform distribution reduces to a random guess with an average AUC of 0.53. This is likely due to the lack of variance of the uniform distribution. The right plot in figure A.1 shows the posterior distributions of $\phi^{\text{controls}} - \phi^{\text{cases}}$ under the uniform prior (white) and the beta like prior (grey) for each pathway. True perturbed pathways are printed in bold on the x axis. Using the beta like distribution as a prior clearly outperforms the uniform distribution as all of the perturbed pathways have been recovered using the beta distribution whilst the uniform distribution fails to recover these perturbations.

Assessing pathway inference against design inaccuracies

It is very common in metabolomics data to find metabolites that are correlated but not in the same KEGG pathway. In the following simulation we assess how inaccuracies in the covariance structure between metabolites and the design matrices $\mathbf{A}_j$ affect the iCARH model. We used the 10 datasets from the previous simulation and perturbed the design matrices by selecting a random fraction of metabolites in each pathway. We then randomly (wrongly) assign these metabolites to no pathway, or to different pathways⁸. We similarly run the model for 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations for each of the fractions 0, 0.18, 0.35, 0.44, 0.5 of wrongly assigned metabolites. Finally, in the same fashion as previously, we assess perturbations based on the 95% credible interval of $\phi^{\text{controls}} - \phi^{\text{cases}}$. Figure 4.4 is a series of average Receiver Operating Characteristic (ROC) curves across 10 datasets for each of the fractions y . On average, the performance of our model reduces to a random guess (AUC of 0.5) if 50% of the metabolites in each pathway is

⁸In other words, this is equivalent to simulating metabolites correlated with each other but are not in the same pathway.

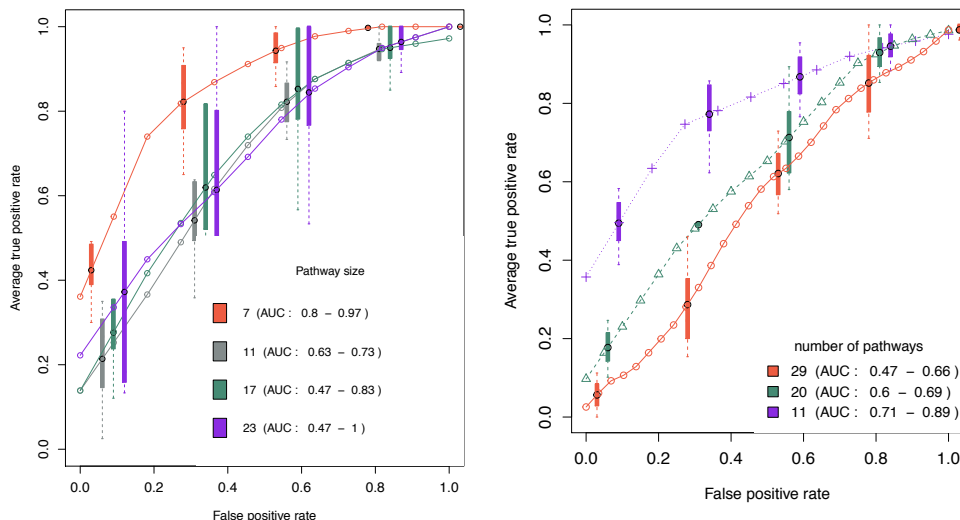


FIGURE 4.5: Left: Average Receiver Operating Characteristic (ROC) curves reflecting the effect of pathway size across 10 datasets for 18 % of “wrongly” assigned metabolites. Pathway size is the average of the number of metabolites per pathway for each simulation for a total of 11 pathways. Right: Average Receiver Operating Characteristic (ROC) curves reflecting the effect of the number of pathways across 10 datasets for 18 % of “wrongly” assigned metabolites.

perturbed. The AUC of our model reaches 0.97 if no metabolites are perturbed and is about 0.88 if 18% of the metabolites in each pathway are perturbed.

In order to assess the effect of the number of pathways and pathway size (number of metabolites in the pathway) on the above results, we perform additional simulations on different 10 datasets each time while perturbing a fixed value of 18% in each pathway (wrongly assigned metabolites). In the first simulation, the number of pathways was fixed to 11 and the total number of metabolites to 40⁹. The average number of metabolites in each simulation i.e. pathway size varies from 7 to 23 metabolites. Figure 4.5 (right) shows that AUC decreases from 7 to 11 with a similar performance for pathway sizes of 11, 17 and 23. This might be due the existent relationship between pathway size, the number of pathways and total number of metabolites which induces an important overlap between different pathways in this simulation (pathways have more than 50% overlap in metabolites starting from pathway size of 11).

⁹All the parameters are similar to previous simulations except pathway size.

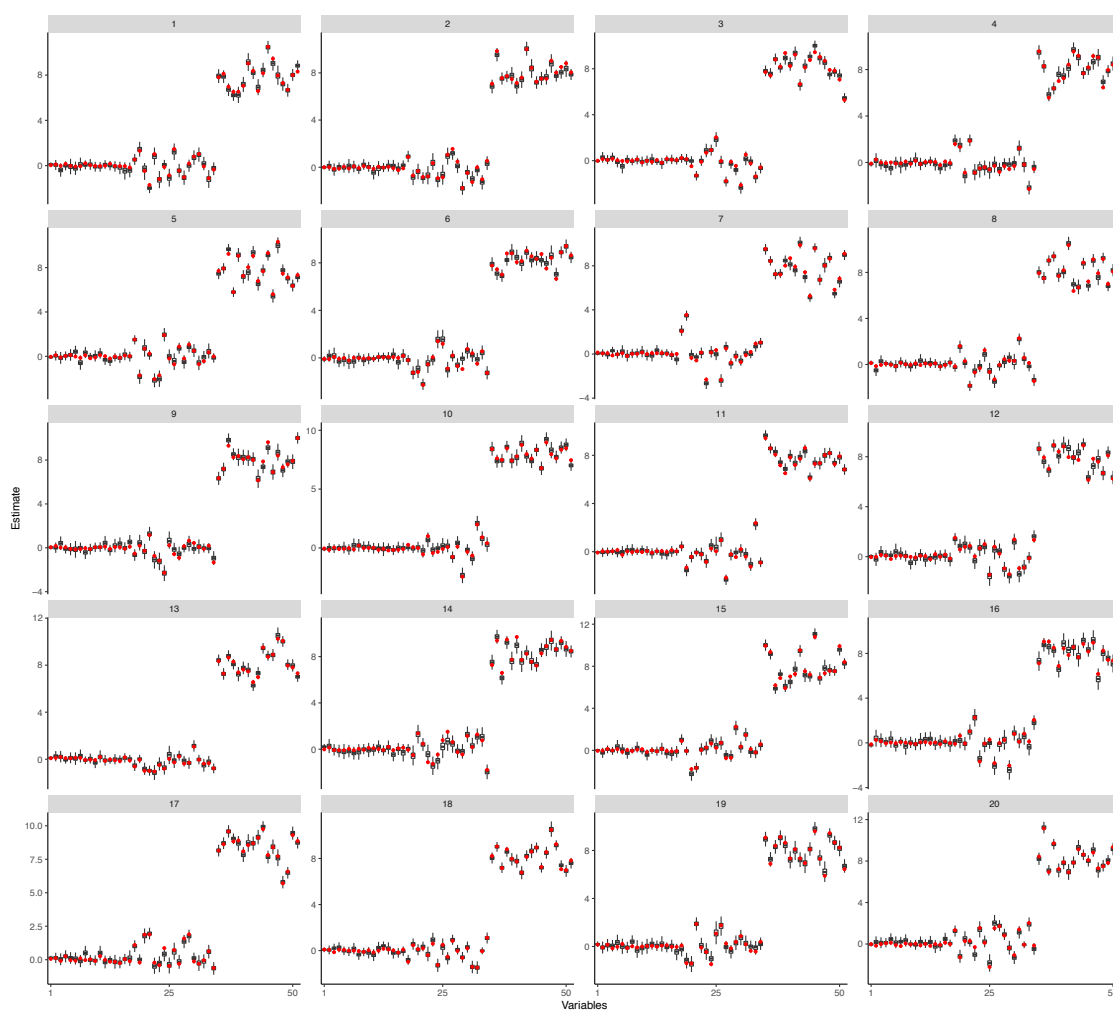


FIGURE 4.6: Assessing the horseshoe like prior behavior : box-plots for estimated β coefficients (black) compared to true simulated coefficients (red). The horseshoe like prior performs well in terms of shrinking the noise terms to zero ($\kappa_{km} \approx 1$) and preserving the high terms ($\kappa_{km} \approx 0$).

In the following simulation we investigate the effect of the total number of pathways. We keep all other parameters as set previously and generate 10 different datasets for each given number of pathways. The total number of pathways spans three values 11, 20 and 29. The pathway size in this simulation was not fixed to a specific values and ranges from 7 to 23 for the total of simulated pathways. For each of these values 18% of the metabolites in each pathway were perturbed. As expected, the performance decreases as the number of pathways increases. This is also explained by the increasing overlap between pathways as the number of total metabolites was also fixed to 40 in this case as well.

Assessing integrative ability

Modeling intra-metabolite interactions and interactions between metabolites and bacteria differently is the key to allow for integrative analysis under the iCARH model. Our choice to use the horseshoe like prior is to allow shrinkage only on small coefficients. In fact, unlike Laplacian or spherical Gaussian priors (MacKay, 1992) which are commonly used for LASSO or ridge regression, the horseshoe like prior does not impose shrinkage on all coefficients thanks to its horseshoe shaped density. Hence, if coefficients are high no shrinkage is applied. In order to assess the horseshoe like prior behavior, we first generate 51 β coefficients for each metabolite randomly from the normal distributions $N(0, 0.1)$ (for the first 18 coefficients), $N(0, 1)$ (for the middle 17 coefficients) and $N(8, 1)$ (for the last 17 coefficients). We then simulate 10 synthetic dataset based on equations (4.16), (4.15), (4.12) and (4.11). The rest of the parameters were fixed as follows : number of bacterial variables $q = 51$, number of metabolites $p = 20$, number of time points $h = 7$, number of samples $n = 18$, global parameter τ fixed to 1.2. Similarly to previous simulations, inference is done using 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations. Figure 4.6 shows boxplots for estimated β coefficients. Each subgraph corresponds to one metabolite where boxplots of estimated β coefficients for each of the 51 bacterial profiles are depicted. Red dots correspond to the true simulated coefficients. By looking at the estimates, it is clear that the horseshoe like prior does a good job in terms of shrinking the noise terms to zero ($\kappa_{km} \approx 1$) and preserving the high terms ($\kappa_{km} \approx 0$). In fact, the horseshoe prior has been widely praised, in the Bayesian variable selection literature in terms of predictive ability, for recovering true zero coefficients without actually shrinking the larger coefficients (the last 34 coefficients here) by contrast to e.g. the Laplace prior, lasso and ridge penalties (Van Der Pas, Kleijn, and Van Der Vaart, 2014; Park and Casella, 2008; Carvalho, Polson, and Scott, 2009). These results are well verified in this experiment.

4.3.2 Case study

In this section, we test our model on an actual metabolomic data and 16S data for bacterial profiles. In this study we are interested in the influence of metformin on a non-diabetic model. Metformin is the first-line medicine to treat type 2 diabetes. It has also been suggested that metformin has anti-cancer, cardiovascular and anti-aging effects. Because of their very large metabolic

capacity, the gut bacteria can influence toxicity and metabolism of drugs. Here, we are particularly looking for metabolic biomarkers indicative of microbiota changes as result of treatment.

The study design is as follows: metabolic profiles of 24 rats are acquired once a week using different mass spectrometry techniques from plasma samples over a period of 9 weeks. Bacterial profiles are acquired using miSeq. The study has allowed for two groups of 12 rats where metformin has been administered to the second group (weeks 3 to 7) allowing for 2 weeks of acclimation (weeks 1 and 2) and 2 weeks of recovery (weeks 8 and 9). After data processing and metabolite identification, a total of 56 metabolites are further examined by looking into the empirical intra-pathway and inter-pathway correlations. Figure 4.7 shows boxplots of intra- and inter-pathway correlations. A higher empirical pairwise correlation signature is observed if metabolites are in the same pathway compared to pairwise correlation between metabolites assigned to different pathways which is consistent with the CAR approach adopted in our model. The data containing a total of 56 metabolites and 6 bacteria species are then analyzed using our model. Some metabolites in the data could not be mapped to KEGG identifiers. Pathway information for these metabolites was ignored but inference of treatment effect and association with other omics variables was seamlessly carried out. Inference is done using 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations. It took about 62 minutes.

We assess performance of our model for different values of τ using the Watanabe-Akaike information criterion (WAIC) which is asymptotically equivalent to Leave-One-Out cross validation (LOO) (Watanabe, 2010). In particular, this criterion evaluates a measure of predictive accuracy a function of the posterior predictive distribution. Tested values of τ comprise 1, 1.2, 5, 10 with corresponding WAIC values of 7317.296 , 7322.798 , 7317.457 , 7316.476 respectively. As for BIC (Bayesian information criterion) and AIC (Akaike information criterion), the lower the WAIC value the best the model fit. The model is not very sensitive to different input values for τ in this case. The change in WAIC for these values of τ is indeed not informative which suggests to rather use the most selective model with $\tau = 10$ as it is the simplest i.e with the smallest number of selected variables.

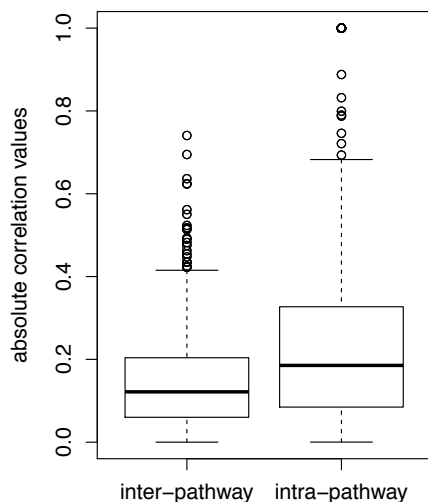


FIGURE 4.7: Metformin data: Figure shows boxplots of correlations between metabolites that are part of the same pathway (intra-pathway correlations) compared to correlations between metabolites that belong to different pathways (inter-pathway correlations). Although correlations are not very close to 1, there is an empirically observed correlation signature related to whether metabolites are in the same pathway or not. In fact average correlations of intra-pathway metabolites are higher than average correlations of inter-pathway metabolites.

Assessing model fit

In order to assess our model fit, we compare of our model to two alternative approaches: DPPCA (Nyamundanda, Gormley, and Brennan, 2014) and SME (Berk, Ebbels, and Montana, 2011). Firstly, we chose DPPCA as one of the few *single* Bayesian models available for longitudinal metabolomics data analysis accounting for time correlations, multivariate aspect and experimental design and hence can be fairly compared to our model. Secondly, the SME approach is chosen as one of the most popular univariate methods used within the metabolomics community. By using spline mixed effects, SME takes into account both temporal variation and experimental design. Other alternative approaches such as PLS derived methods were investigated but they poorly performed for this dataset with a R squared value around 0.3.

We first perform posterior predictive checks of our model compared to DPPCA (Nyamundanda, Gormley, and Brennan, 2014). The DPPCA model is a

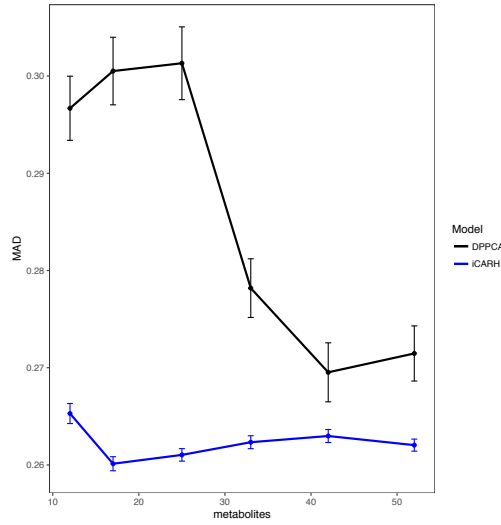


FIGURE 4.8: Metformin data: Posterior predictive checks for mean absolute deviation (MAD) compared to DPPCA for different numbers of metabolites used for inference. The MAD decreases as the number of metabolites increases. Our model performs clearly better than the DPPCA model.

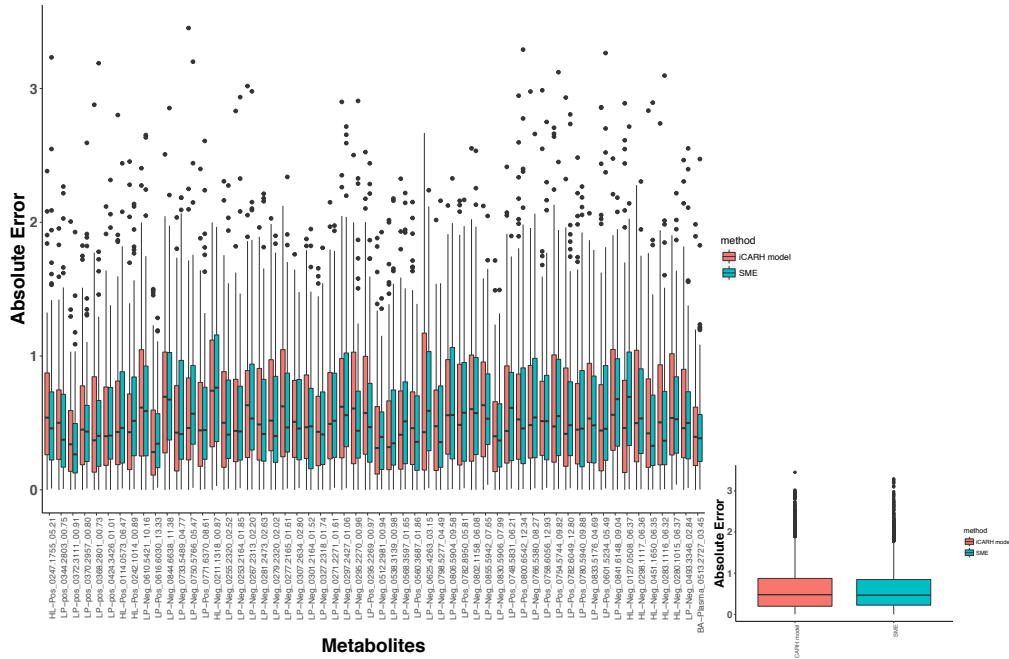


FIGURE 4.9: Metformin data: Absolute mean error of estimated metabolic profiles using our model (red) and the SME model (green) for each metabolite (right) and on average (left). Our model performs in a similar way to SME which is indicative of a good fit.

multivariate model using PCA, where PCA scores are modeled via a stochastic volatility model (Jacquier, Polson, and Rossi, 2002; Kim, Shephard, and Chib,

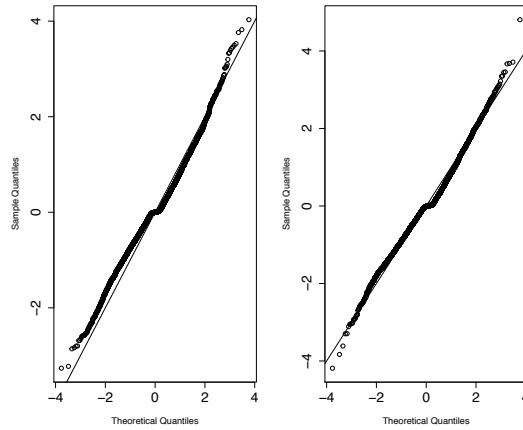


FIGURE 4.10: Metformin data: Right and left panels show model fit assessment for controls and cases data. Left : quantile-quantile normal plot of $\Psi_{\text{cases}}^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$. Right : quantile-quantile normal plot of $\Psi_{\text{controls}}^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$.

1998). Stochastic volatility models are models where the variance (volatility) follows a latent stochastic process. In particular, in the DPPCA model, the log variances of the latent variables follow a stationary first order vector autoregressive process ¹⁰ to account for potential time dependence in the data. In the Bayesian framework, posterior predictive checks consist in comparing data simulated from the posterior predictive distribution with the observed data. We compared the simulated data and the observed data by means of mean absolute deviations (MADs) between the observed and the simulated covariance matrices repeatedly for different numbers of metabolites included i.e only part of the data corresponding to these metabolites is used for inference. The same process was repeated for inference using the DPPCA model (Nyamundanda, Gormley, and Brennan, 2014). Time performance of the iCARH and the DPPCA model was very similar e.g. about 1 hour if the whole dataset is used. Figure A.2 shows MADs of our model and the DPPCA model. As expected, MADs for both models decrease when the number of metabolites increases as predictive ability of both models is strengthened. Additionally, for all experiments, our model clearly outperforms the DPPCA model.

Further to posterior predictive checks, we compare predictive ability of our model with the spline mixed effects (SME) model (Berk, Ebbels, and Montana, 2011). The SME model is one of the popular algorithms in the metabolomics

¹⁰Vector autoregressive models generalize the autoregressive (AR) model by allowing for multiple times series.

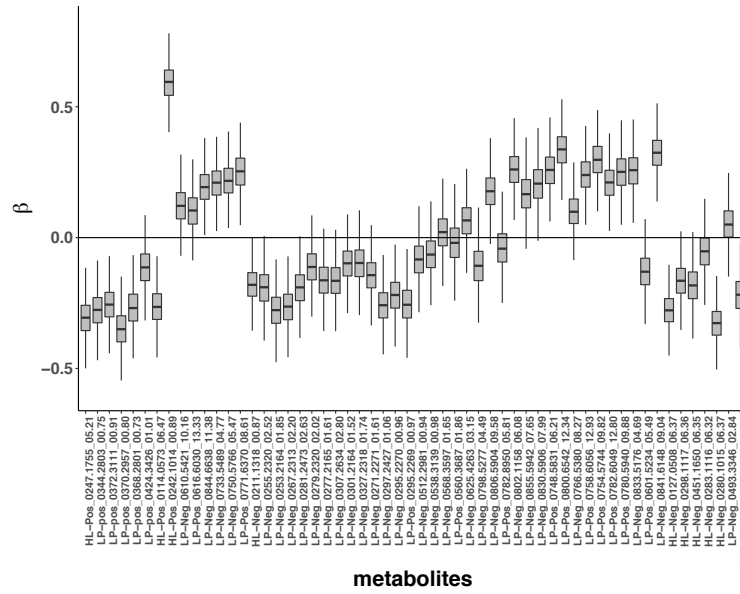


FIGURE 4.11: Estimates of effects of treatment on metabolite profiles are captured by β_m^α . Only part of the data is plotted as we are mainly interested in the feature “LP-Neg_0295.2270_00.96” which is a hydroxy octadecadienoic acid.

community for analyzing time course data. It was designed in order to prevent under- and overfitting by setting the right smoothing penalty. However, fitting is done in a univariate way. We perform cross-validation for both models by predicting metabolites each time using the previous time points. We compare performance of each model using the absolute mean error over predicted time points. Figure 4.9 shows the absolute mean error of our model and the SME model for each metabolite (right) and on average (left). Both models perform in a very similar way. This is indicative of a good fit of our model.

Finally, goodness of fit was also checked by using $\Psi_e^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it}) \sim N(0, \mathbf{I}_p)$ where Ψ_e denotes the Cholesky factor of $(\mathbf{I}_p - \mathbf{C}(\boldsymbol{\phi}^e))^{-1}\sigma^2$. Zero-mean and normality were thus checked for $\Psi_e^{-1}(\mathbf{x}_{it} - \boldsymbol{\mu}_{it})$ (See figure 4.10).

Data results

Since the administrated drug was also profiled using mass spectrometry, we fit the iCARH model with $\boldsymbol{\alpha}_m = \beta_m^\alpha \mathbf{y}_{\text{drug}}$ where β_m^α is a scalar. Figure 4.11 is a series of boxplots for β_m^α for all metabolites. We are mainly interested in the feature “LP-Neg_0295.2270_00.96” which is a hydroxy octadecadienoic acid as it is associated with the parasuterella bacteria species.

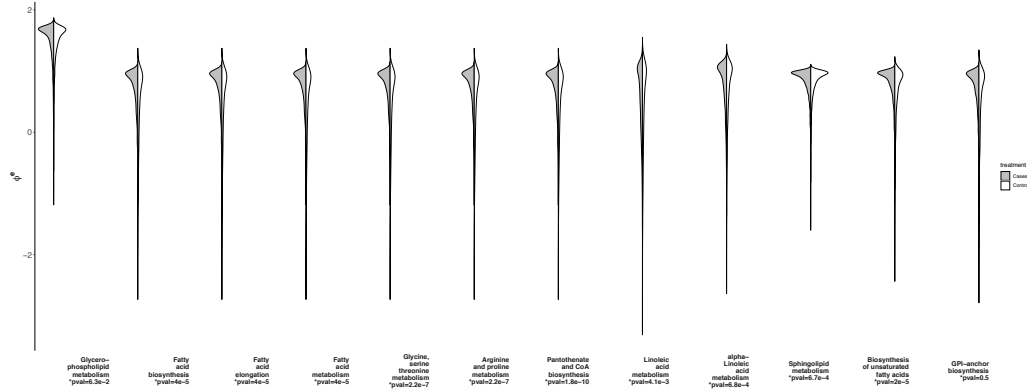


FIGURE 4.12: Metformin data: Posterior distributions of ϕ^e for each pathway for each treatment group. Posterior distributions of ϕ^e are very similar for both controls and cases which is indicative of no -or mild- pathway alterations. These pathways correspond as well to the top 11 pathways inferred by MetaboAnalyst (Chong et al., 2018). The p-values are the p-values returned by MetaboAnalyst for each pathway.

Figures 4.12 and 4.14 show posterior distributions of ϕ^e for each pathway and estimates of effects of bacteria on metabolites. In “assessing pathway inference against design inaccuracies” in Section 4.3.1, we have investigated for similar data parameters the performance of iCARH in pathway inference where metabolites have been wrongly assigned to different or no pathways. This comes down to simulating metabolites that are correlated with each other but are not in the same pathway. Based on this, we can refer to figure 4.4 to give a flavor of the performance of the iCARH model *a priori* depending on the behavior of correlation structure of metabolites compared to the design matrices $\mathbf{A}_j$. Hence, in order to have an *a priori* idea on the robustness of pathway inference, results suggest to compare the correlation structure of metabolites in the observed data with the design matrices.

For a correlation threshold of 0.35, about 25% of the metabolites that are correlated in the metformin data are not assigned to the same pathway in the design matrices. According to figure 4.4, 25% of wrongly assigned metabolites corresponds to an AUC around 0.8. This supports the use of the iCARH model for pathway perturbation inference for this data. If we set a higher correlation threshold, a lower number of metabolites will be misspecified and hence a higher AUC can be obtained.

Estimates of effects of bacteria on metabolite profiles are captured by β_m . Some metabolites present significant changes along with the bacterial profiles.

For example, the hydroxy octadecadienoic acid, previously identified, is associated with alterations in abundance of 3 bacteria species: parasuterella, bacteroidales and prevotellaceae. Figure 4.12 shows that, as a result of treatment, KEGG pathways are not significantly altered. However, distributions of ϕ^{controls} for “fatty acids biosynthesis” and “biosynthesis of unsaturated fatty acids” KEGG pathways are remarkably flatter than the distributions of ϕ^{cases} . These pathways involve the previously identified hydroxy fatty acid. Our analysis confirms previously reported studies that hydroxy fatty acids might be produced by the gut microbiome (Kishino et al., 2013; Kimura et al., 2013).

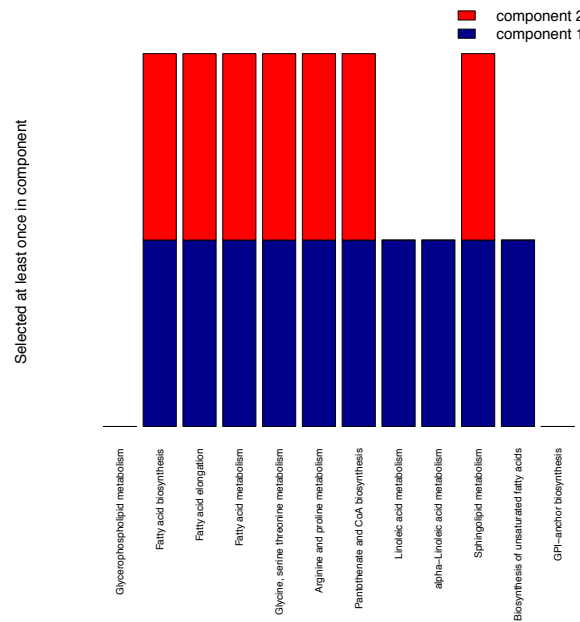
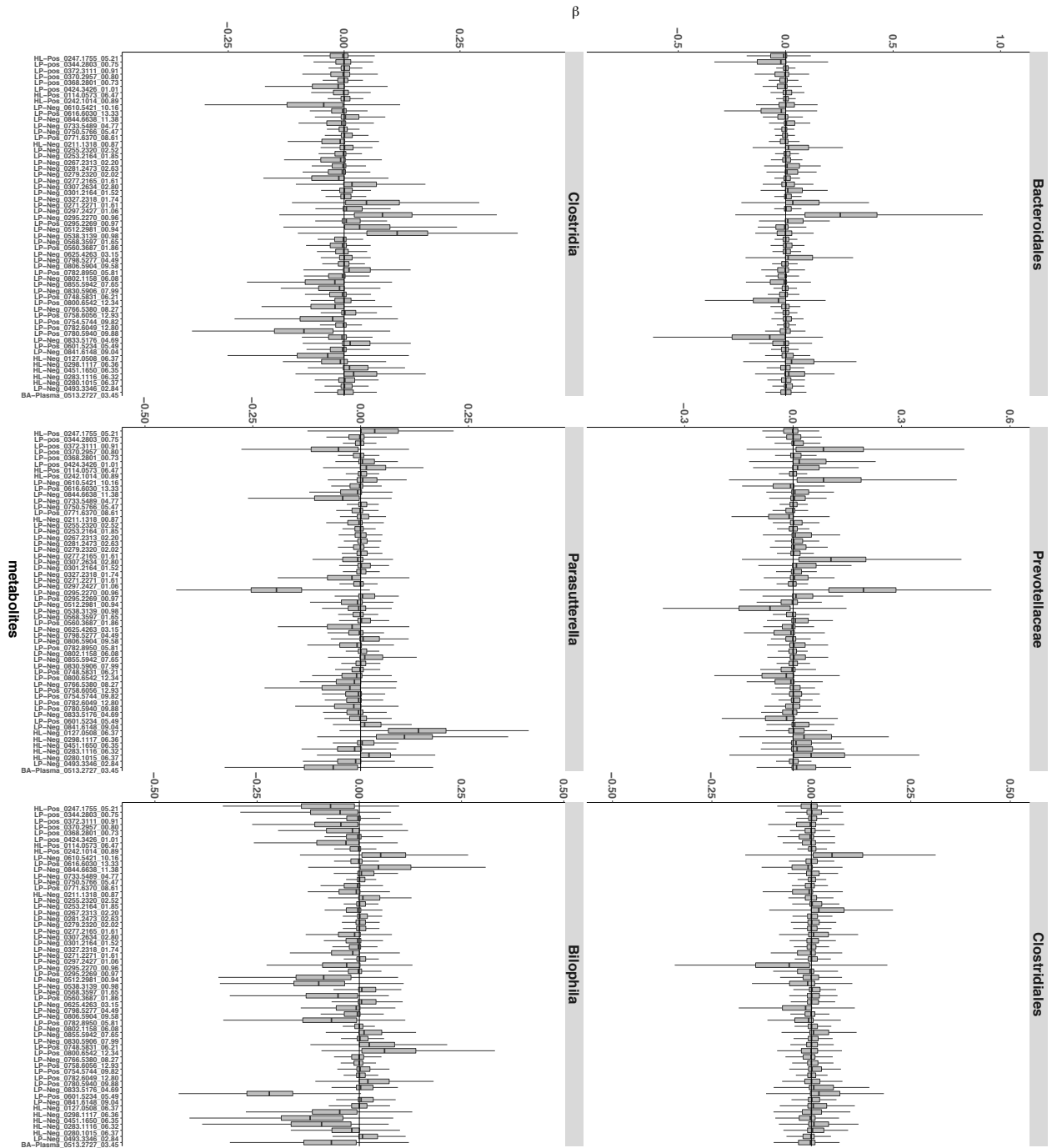


FIGURE 4.13: Metformin data: Pathway inference using sparse group PLS. For each pathway, bars indicate whether or not the pathway was selected in the first (blue) or second (red) component. Non selected pathways indicate that all related genes correspond to zero weights in a given component. “Glycerophospholipid metabolism” and “GPI-anchor biosynthesis” were not selected in both components whilst other pathways were selected at least once.

On the other hand, results from MetaboAnalyst (Chong et al., 2018) give p-values between 6.3×10^{-2} and 1.8×10^{-10} indicating that all pathways are significant to changes of the treatment except one; glycerophospholipid metabolism. We think that this discrepancy in the results is due to the way iCARH and MetaboAnalyst model pathway perturbation i.e. whilst MetaboAnalyst considers that mean level changes of one or more metabolite concentrations involved in a pathway indicate perturbation of the latter pathway, iCARH considers that

changes in the covariance structure of metabolites in the same pathway are indicative of pathway perturbation. In addition to comparing with MetaboAnalyst, integrative analysis of both the metabolomic and bacterial profiles was performed using sparse group PLS (Liquet et al., 2015) in order to simultaneously select the contributing group of metabolites or pathways¹¹. A sparse group PLS model with 2 components was fitted to the data. The number of components was selected based on a scree plot whereas optimal sparsity parameters and number of groups were selected to minimize the cross validation error. Figure 4.13 shows whether or not a pathway has been selected for a given component. Blue indicates that a given pathway has been selected in the first component whilst red indicates that a given pathway has been selected in the second component. Although posterior distributions of ϕ^e for “GPI-anchor biosynthesis” indicate potential pathway perturbation, “GPI-anchor biosynthesis” was not selected in either PLS components. By contrast, “Sphingolipid metabolism” was selected in both PLS component whereas posterior distributions of ϕ^e for “Sphingolipid metabolism” are not indicative of pathway perturbation. Results of both methods are concordant to a great extent for the rest of the pathways. For instance, iCARH does not indicate perturbation of “Glycerophospholipid metabolism” as a result of treatment as posterior distributions of ϕ^e for “Glycerophospholipid metabolism” are similar for both cases and controls. In accordance with this result, “Glycerophospholipid metabolism” was not selected in any of the PLS components. Other pathways were selected at least once in the PLS model and exhibit different posterior distributions of ϕ^e for cases and controls in the iCARH model.

¹¹This approach does not account for temporal correlations.

FIGURE 4.14: Metformin data: Estimates of effects of bacteria on metabolite profiles in iCARH are captured by β_m .

4.4 Implementation

The core of my work on this chapter focuses not only on the iCARH model but also on ease of use and free availability of software. Implementation of the model is available in an open source R (R Core Team, 2017) package iCARH on CRAN. The package comprises various useful functions ranging from constructing pathways adjacency matrices to plotting inferred parameters and checking of assumptions.

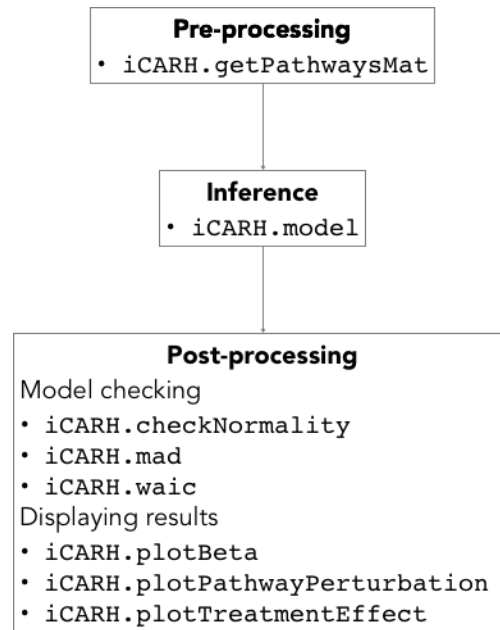


FIGURE 4.15: iCARH's simple pipeline and functions. The package also provides the `iCARH.simulate` function used to simulate data presented in Section 4.3. This function allows for different setting parameters.

iCARH.simulate

This function provides the user mainly with two synthetic longitudinal datasets X and Y where X is a three $T \times N \times M$ dimensional array and Y a three $T \times N \times K$ dimensional array. Amongst various parameters, the function takes as input `num.corr.y` parameter which determines the number of correlated omics profiles in the Y dataset. The last `num.corr.y` variables will be highly correlated to the first `num.corr.y` variables. It is important to note that the β coefficients are simulated based on normal distributions with varying means and variances rather than being simulated via the horseshoe like prior. The β

coefficients can also be specified by the user via the `beta.val` input parameter.

In addition, the function also simulates pathway distance matrices in 3 ways depending on the value of the input parameters `path.names`, `path.probs` and `pathway.perturb.ratio`. Particularly `path.names` represents the pathways to sample from as specified in KEGG. If `path.names` is not specified, `path.probs` will be used. If `path.probs` is `TRUE`, KEGG like density of pathways per metabolite is used for sampling. If `path.probs` is a scalar, then `path.probs` is the expected ratio of metabolites belonging to each pathway. `pathway.perturb.ratio` is the ratio of perturbed pathways. Finally, the function returns a set of variables of interest most importantly including $\mathbf{X}, \mathbf{Y}, \mathbf{y}_{drug}, \boldsymbol{\beta}, \mathbf{A}_p$ for $p = 1 \cdots P, \boldsymbol{\phi}$ and θ_m .

iCARH.getPathwaysMat

The function `iCARH.getPathwaysMat` builds pathways distance matrices from specified KEGG compounds (metabolites) in data $\mathbf{X}$ and organism. It returns a list of pathway distance matrices based on shortest paths between every pair of metabolites. This function mainly requires the `KEGGgraph` R package.

iCARH.model

This is the main function of the package where model parameters are inferred (See table 4.1). The model is specified using the `rstan` package (Stan Development Team, 2017). The `rstan` package is an R interface to the Stan probabilistic programming language. Stan has been mainly selected thanks to its ability to perform Hamiltonian Monte Carlo sampling by contrast with BUGS (Lunn et al., 2012) or jags (Hornik, Leisch, and Zeileis, 2003) that mainly use Gibbs sampling. In fact, Hamiltonian Monte Carlo sampling allows for faster convergence and better stability as it uses the gradient of the log probability function whereas Gibbs sampling is only based on random exploration of the parameter space given the remaining parameters.

Expected inputs of the function are composed of the observed metabolomics time-course data $\mathbf{X}$, the additional omics time-course data $\mathbf{Y}$, class information or treatment variable `drug` and pathway adjacency matrices `pathways`. The global sparsity parameter `tau` (i.e. τ in Section 4.2) is set to 1.2 by default but

can be changed by the user. In our experience with different simulated data, the value of 1.2 for τ provides the highest WAIC value.

For vectorization and indexing efficiency in `Stan`, each observed data X and Y is a three $T \times N \times M$ dimensional array and $T \times N \times K$ dimensional array respectively. As it is very common to have missing values in omics data, missing values are only supported for X and Y and are specified by the input parameter `NA_value`. Missing data in X are inferred based on equation (4.11) whereas missing values in Y are inferred using a Markov process i.e $\mathbf{y}_{it} \sim N(\mathbf{y}_{i,t-1}, \sigma_y)$ where σ_y has a weakly informative Gamma prior. `drug` is a $T \times N$ matrix containing values of the administrated drug or treatment class for each sampling time point. `pathways` is a list of length P of $M \times M$ pathway distance matrices. Pathway distance matrices can be specified in many ways. The function `iCARH.getPathwaysMat` supports building pathway matrices based on the shortest KEGG path between two metabolites.

It is important to note that in this function, β coefficients are initialized using a univariate regression model ignoring time and metabolite dependence. Another noteworthy feature of `iCARH.model` is that it is compatible with unknown KEGG identifiers or non-annotated metabolites. In either cases, pathway information for these metabolites is ignored but inference of treatment effect and association with other omics variables will still be performed. Average time performance of the function for each of the simulated datasets presented in Section 4.3 with 2000 iterations of Hamiltonian Monte Carlo sampling and 1000 warm-up iterations was about 2 hours.

`iCARH.checkNormality`

As its name intuitively suggests, this function checks normality assumptions. It returns the normalized data and performs quantile-quantile normal plot for each class. See figure 4.10.

`iCARH.mad`

This function performs posterior predictive checks. It computes and returns MADs (Mean Absolute Deviation) between the true covariance matrix and the inferred covariance matrix.

iCARH.waic

This function computes the Watanabe-Akaike Information Criterion (WAIC) which is asymptotically equivalent to Leave-One-Out cross validation (Watanabe, 2010).

iCARH.plotBeta,**iCARH.plotTreatmentEffects,****iCARH.plotPathwayPerturbation**

These functions respectively display boxplots of the posterior distributions of the β coefficients (figure 4.14), boxplots of the posterior distributions of β_m^α (figure 4.11) and the posterior distributions of ϕ (figure 4.12). These plots can be altered (e.g. adding a title, changing the axes labels, changing the background) as they are generated using ggplot2 commands. Additional parameters from the Stan object returned by the `iCARH.model` can also be extracted and plotted.

4.5 Discussion

Identifying biomarkers in time course metabolic data and inferring significant associations with heterogeneous omics variables is extremely challenging due to the several sources of variations of the data. In addition, existing methods developed to analyze such data are very scarce and have the limitations of i) overfitting to the few available data points or ii) confounding the experimental and longitudinal variation or iii) ignoring the metabolite interactions or iv) ignoring effects of other omics variables. In this paper, the model we have developed combines several approaches to take into account the different aspects of the data namely the number of time points, the experimental variation captured by μ_{it} , interactions between metabolites captured by ϕ and interactions with additional omics variables captured by β_m .

Our results demonstrate that our model successfully addresses the main questions of a metabolomic study. Most importantly, our model is able to identify metabolic biomarkers related to treatment, infer perturbed pathways as a result of treatment and find significant associations with additional omics variables. We have shown that providing an informative prior on metabolic pathways and an informative prior over the parameter ϕ is a significant improvement

over the DPPCA model. Particularly, our model is more robust to slight variations usually observed in short time series data thanks to the small number of covariance parameters needed to estimate compared to DPPCA. We have also shown through simulation that an informative beta like prior compares better than a non-informative uniform prior in inferring significant pathways. On real data, we have investigated how the number of profiled metabolites can affect the predictive ability of the model. Using all available measurements, we have also shown that our model yields a remarkably similar predictive performance to SME.

In addition, iCARH provides a natural way to handling missing values in omics data. Missing values are in fact considered as random variables which are directly inferred from the model itself (See Appendix A). Nevertheless, the iCARH model is not scalable to more than few hundreds of omics variables since it has been designed for annotated or partially annotated metabolomics data. For completely non annotated data other models such as multilevel PLS, dynamic PLS or PARAFAC can be used. Moreover, iCARH is in practice limited to experimental designs with 2 groups i.e. cases and controls and will not scale well to more than a dozen of time points. Additional sensitivity analyses with regards to the size of pathways, number of pathways and number of missing values could be further performed to assess the performance of the model. A sensitivity analysis relating to the number of time points has been performed on the metformin dataset and can be found in Appendix A.

From a practical point of view, the model has been fitted using HMC sampling and has an average time performance (about 1 hour) mostly because of the variable selection procedure and metabolites interdependence. This could be addressed by using variational Bayes. In fact, variational Bayes inference procedures offer cost-effective inference by means of principled approximations and appealing computational time for high dimensional data. A variational Bayes inference of CAR models was proposed by Harrison and Green (2010) for high dimensional data, and a variational Bayes approach for variable selection was recently proposed by Ormerod, You, and Müller (2017).

Several potential extensions arise naturally from our model. In terms of the metabolite interactions component, many research questions can arise. Alternative strategies to modeling metabolite interactions can be examined such as modeling the non-zero elements of the adjacency matrix $\mathbf{C}$ of each pathway as random variables. This strategy was adopted in the CAR literature by

Lee (2013) and Rushworth, Lee, and Sarrao (2017) to take into account step changes in spatial variation. Step changes can potentially be useful to model changes in metabolites correlations as a result of treatment. Lee (2011) provide an overview of different CAR models used in spatial modeling. The proposed models can be adapted to fit into the metabolomics literature.

Another potential extension is linked to pathway annotation and modeling. In this research paper, the KEGG pathways are used but this poses certain limitations regarding the performance of the model, due to some shortcomings in the database (missing compounds, inaccuracies in the database etc). In this sense extending the tool to support software formats and applications that enable assembly of superpathways (SBML (Hucka et al., 2003), OWL (Bleasby, Akrigg, and Attwood, 1994), KEGGConverter (Moutselos et al., 2009) or KENEV (Pilalis et al., 2015)) would probably increase its performance in terms of accuracy and computational efficiency, as the adjacency matrix would be known a priori and its sparsity would be stronger.

Metabolomics longitudinal profiling techniques are imperative to understand the effect of a drug or a disease across time and can provide enhanced understanding of the underlying biology of the system. In a data integration framework, we have illustrated the use of the CAR model to incorporate metabolites associations in the model and the horseshoe prior to identify association with heterogeneous omics variables obtained by other omics techniques. The combination of the CAR and horseshoe levels yields the “integrative CAR Horseshoe” (iCARH) model which we presented in this chapter.

Although, it is computationally expensive, the iCARH model has various appealing features such that it is able to identify metabolic biomarkers related to treatment, infer perturbed pathways as a result of treatment and identify potential associations between heterogeneous omics variables. Clearly, these appealing features open up further research topics.

Chapter 5

A new approach for canonical correlation analysis and correlation-based network model¹

5.1 Background

In the previous chapter we have investigated linear relationship between heterogeneous variables via a Bayesian regression model. In this chapter, we turn our attention to exploring linear relationships between two sets of variables via Canonical correlation analysis (CCA) (Hotelling, 1936; Härdle and Simar, 2015). CCA is a classic and highly versatile statistical approach that helps to decode complex dependency structures in multivariate data and to identify groups of interacting variables. Consequently, it has numerous practical applications in molecular biology, for example omics data integration (Lê Cao et al., 2009) and network analysis (Hong et al., 2013), but also in many other areas such as econometrics or social science.

In its original formulation CCA is viewed as an algorithmic procedure optimizing a set of objective functions, rather than as a probabilistic model for the data. Only relatively recently this perspective has changed. Bach and Jordan (2005) proposed a latent variable model for CCA building on earlier work on probabilistic principal component analysis (PCA) by Tipping and Bishop (1999). The probabilistic approach to CCA not only allows to derive the classic CCA algorithm but also provide an avenue for Bayesian variants (Wang, 2007; Klami and Kaski, 2007).

In parallel to establishing probabilistic CCA, the classic CCA approach has also been further developed in the last decade by introducing variants of the

¹Most material in this chapter is based on Jendoubi and Strimmer, 2018

CCA algorithm that are more pertinent for high-dimensional data sets now routinely collected in the life and physical sciences. In particular, the problem of singularity in the original CCA algorithm is resolved by introducing sparsity and regularization (Haroon and Shawe-Taylor, 2011; Wilms and Croux, 2015) and, similarly, large-scale computation is addressed by new algorithms (Cruz-Cano and Lee, 2014; Ma, Lu, and Foster, 2015).

In this note, we revisit both classic and probabilistic CCA from the perspective of whitening of random variables (Kessy, Lewin, and Strimmer, 2018). As a result, we propose a simple yet flexible probabilistic model for CCA linking together multivariate regression, latent variable models, and high-dimensional estimation. Crucially, this model for CCA not only facilitates comprehensive understanding of both classic and probabilistic CCA via the process of whitening but also extends CCA by allowing for negative canonical correlations and providing the flexibility to include non-normal latent variables.

The remainder of this chapter is as follows. First, after reviewing classical CCA in Section 5.2.1 we demonstrate that the classic CCA algorithm is special form of whitening (Section 5.2.2). Subsequently, we show that the link of CCA with multivariate regression leads to a probabilistic two-level latent variable model for CCA that directly reproduces classic CCA without any rotational or other ambiguity (Section 5.2.3). Since the model is fully probabilistic any preferred method for inference or regularization can be employed. We propose an entropy-based measure linked to CCA to quantify multivariate associations and identify underlying network structure in Section 5.3. We discuss simple computationally and statistically effective procedures for high-dimensional application of CCA in Section 5.5.1. Finally, in Section 5.4 we demonstrate and investigate our proposed approach, which is implemented and freely available as R package, both in simulations as well as by analyzing multiple integrated omics data.

5.2 A whitening approach for probabilistic canonical correlation analysis

In this section, we first introduce CCA from a classical perspective, then we demonstrate that CCA is best understood as a special and uniquely defined type of whitening transformation. Next, we investigate the close link of CCA

with multivariate regression. This not only allows to interpret CCA as regression model and to better understand canonical correlations, but also provides the basis for a probabilistic generative latent variable model of CCA based on whitening. This model is introduced in the last subsection.

5.2.1 Classical CCA

We now briefly introduce the objective of classical CCA as well as specific notation. Most of the discussion in this chapter is entirely framed in terms of random vectors and their properties, rather than data matrices. This allows to cleanly separate the study of probabilistic model for CCA from statistical procedures for estimating the corresponding parameters.

Notation

As per the notation used in this thesis, the starting point are two random vectors $\mathbf{X} = (X_1, \dots, X_p)^T$ and $\mathbf{Y} = (Y_1, \dots, Y_q)^T$ of dimension p and q . Their respective multivariate distributions $F_{\mathbf{X}}$ and $F_{\mathbf{Y}}$ have expectation $E(\mathbf{X}) = \boldsymbol{\mu}_{\mathbf{X}}$, $E(\mathbf{Y}) = \boldsymbol{\mu}_{\mathbf{Y}}$, and covariance $\text{var}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{X}}$ and $\text{var}(\mathbf{Y}) = \boldsymbol{\Sigma}_{\mathbf{Y}}$. The cross-covariance between $\mathbf{X}$ and $\mathbf{Y}$ is given by $\text{cov}(\mathbf{X}, \mathbf{Y}) = \boldsymbol{\Sigma}_{\mathbf{XY}}$. The corresponding correlation matrices are denoted by $\mathbf{P}_{\mathbf{X}}$, $\mathbf{P}_{\mathbf{Y}}$, and $\mathbf{P}_{\mathbf{XY}}$. By $\mathbf{V}_{\mathbf{X}} = \text{diag}(\boldsymbol{\Sigma}_{\mathbf{X}})$ and $\mathbf{V}_{\mathbf{Y}} = \text{diag}(\boldsymbol{\Sigma}_{\mathbf{Y}})$ we refer to the diagonal matrices containing the variances only, allowing to decompose covariances as $\boldsymbol{\Sigma} = \mathbf{V}^{1/2} \mathbf{P} \mathbf{V}^{1/2}$. The composite vector $(\mathbf{X}^T, \mathbf{Y}^T)^T$ has therefore mean $(\boldsymbol{\mu}_{\mathbf{X}}^T, \boldsymbol{\mu}_{\mathbf{Y}}^T)^T$ and covariance

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{XY}}^T & \boldsymbol{\Sigma}_{\mathbf{Y}} \end{pmatrix}.$$

The joint random vector $(X_1, \dots, X_p, Y_1, \dots, Y_q)^T$ has distribution $F_{\mathbf{X}, \mathbf{Y}}$. Under independence of $\mathbf{X}$ and $\mathbf{Y}$ the joint distribution becomes $F_{\mathbf{X} \perp \mathbf{Y}}$.

For a symmetric positive definite matrix $\mathbf{M}$ with positive eigenvalues κ_i and eigenvectors $\mathbf{U}$ the matrix power is given by $\mathbf{M}^\alpha = \mathbf{U} \text{diag}(\kappa_i^\alpha) \mathbf{U}^T$ and the matrix logarithm by $\log(\mathbf{M}) = \mathbf{U} \text{diag}(\log \kappa_i) \mathbf{U}^T$. The log-determinant of $\mathbf{M}$ is $\log(\det(\mathbf{M})) = \text{Tr}(\log(\mathbf{M})) = \sum_i \log \kappa_i$. Series expansion of the log function yields $-\log(\mathbf{I} - \mathbf{M}) = \sum_{k=1}^{\infty} \frac{\mathbf{M}^k}{k}$ where for convergence a spectral radius $\max(\kappa_i) < 1$ is needed. Thus in first order approximation $-\log(\mathbf{I} - \mathbf{M}) \approx \mathbf{M}$ and $-\log(\det(\mathbf{I} - \mathbf{M})) = \text{Tr}(-\log(\mathbf{I} - \mathbf{M})) \approx \text{Tr}(\mathbf{M}) = \sum_i \kappa_i$.

Vector-valued samples of the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are denoted by $\mathbf{x}_i$ and $\mathbf{y}_i$ so that $(\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n)^T$ is the $n \times p$ data matrix for $\mathbf{X}$ containing n observed samples (one in each row). Correspondingly, the empirical mean

for $\mathbf{X}$ is given by $\hat{\boldsymbol{\mu}}_{\mathbf{X}} = \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, the unbiased covariance estimate is $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{\mathbf{X}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$, and the corresponding correlation estimate is denoted by $\hat{\mathbf{P}}_{\mathbf{X}} = \mathbf{R}_{\mathbf{X}}$.

CCA objective

In canonical correlation analysis the aim is to find mutually orthogonal pairs of maximally correlated linear combinations of the components of $\mathbf{X}$ and of $\mathbf{Y}$. Specifically, we seek canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ (i.e. vectors of dimension p and q , respectively) for which

$$\text{cor}(\boldsymbol{\alpha}_i^T \mathbf{X}, \boldsymbol{\beta}_j^T \mathbf{Y}) = \begin{cases} \lambda_i & \text{maximal for } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (5.1)$$

where λ_i are the canonical correlations, and simultaneously

$$\text{cor}(\boldsymbol{\alpha}_i^T \mathbf{X}, \boldsymbol{\alpha}_j^T \mathbf{X}) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (5.2)$$

and

$$\text{cor}(\boldsymbol{\beta}_i^T \mathbf{Y}, \boldsymbol{\beta}_j^T \mathbf{Y}) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise.} \end{cases} \quad (5.3)$$

In matrix notation, with $\mathbf{A} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_p)^T$, $\mathbf{B} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_q)^T$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$, the above can be written as $\text{cor}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \boldsymbol{\Lambda}$ as well as $\text{cor}(\mathbf{A}\mathbf{X}) = \mathbf{I}$ and $\text{cor}(\mathbf{B}\mathbf{Y}) = \mathbf{I}$. The projected vectors $\mathbf{A}\mathbf{X}$ and $\mathbf{B}\mathbf{Y}$ are also called the CCA scores or the *canonical variables*.

Hotelling (1936) showed that there are, assuming full rank covariance matrices $\boldsymbol{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}}$, exactly $m = \min(p, q)$ canonical correlations and pairs of canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$, and that these can be computed analytically from a generalized eigenvalue problem (e.g. Härdle and Simar, 2015). In Section 5.2.2 we will see how canonical directions and correlations follow almost effortlessly from a whitening perspective of CCA.

Since correlations are invariant against rescaling, optimizing Eq. 5.1 determines the canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ only up to their respective lengths, and we can thus arbitrarily fix the magnitude of the vectors $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$. A common choice is to simply normalize them to unit length so that $\boldsymbol{\alpha}_i^T \boldsymbol{\alpha}_i = 1$ and $\boldsymbol{\beta}_i^T \boldsymbol{\beta}_i = 1$.

Similarly, the overall sign of the canonical directions α_i and β_j is also undetermined. As a result, different implementations of CCA may yield canonical directions with different signs, and depending on the adopted convention this can be used either to enforce positive or to allow negative canonical correlations, see Section 5.2.3 for further discussion in the light of CCA as a regression model.

Because it optimizes correlation, CCA is invariant against location translation of the original vectors $\mathbf{X}$ and $\mathbf{Y}$, yielding identical canonical directions and correlations in this case. However, under scale transformation of $\mathbf{X}$ and $\mathbf{Y}$ only the canonical correlations λ_i remain invariant whereas the directions will differ as they depend on the variances $\mathbf{V}_\mathbf{X}$ and $\mathbf{V}_\mathbf{Y}$. Therefore, to facilitate comparative analysis and interpretation the canonical directions the random vectors $\mathbf{X}$ and $\mathbf{Y}$ (and associated data) are often standardized.

Classical CCA uses the empirical covariance matrix $\mathbf{S}$ to obtain canonical correlations and directions. However, $\mathbf{S}$ can only be safely employed if the number of observations is much larger than the dimensions of either of the two random vectors $\mathbf{X}$ and $\mathbf{Y}$, since otherwise $\mathbf{S}$ constitutes only a poor estimate of the underlying covariance structure and in addition may also become singular. Therefore, to render CCA applicable to small sample high-dimensional data two main strategies are common: one is to directly employ regularization on the level of the covariance and correlation matrices to stabilize and improve their estimation; the other to devise probabilistic models for CCA to facilitate application of Bayesian inference and other regularized statistical procedures.

5.2.2 Whitening transformations and CCA

In this section we now demonstrate that CCA is best understood as a special and uniquely defined type of whitening transformation.

Background on whitening

Whitening, or sphering, is a linear statistical transformation that converts a random vector $\mathbf{X}$ with covariance matrix $\Sigma_\mathbf{X}$ into a random vector

$$\widetilde{\mathbf{X}} = \mathbf{W}_\mathbf{X} \mathbf{X} \tag{5.4}$$

with unit diagonal covariance $\text{var}(\widetilde{\mathbf{X}}) = \boldsymbol{\Sigma}_{\widetilde{\mathbf{X}}} = \mathbf{I}_p$. The matrix $\mathbf{W}_{\mathbf{X}}$ is called the *whitening matrix* or *sphering matrix* for $\mathbf{X}$, also known as the *unmixing matrix*.

Whitening is commonly used as a preprocessing step in image processing and neuroscience mainly in order to remove spatial correlation effects. In general, whitening transforms a set of correlated random variables into a set of uncorrelated random variables giving rise to a simpler dependence structure between the whitened random variables. Whitening also scales the variables to have unit variance which is useful to treat variables with equal importance. Using these properties, we will use whitening to answer to a CCA objective with particular scaling constraints. Most importantly we will show that the canonical directions, solution of the CCA objective function, are the basis of the whitening matrix. This whitening approach to CCA allows to build a probabilistic model of CCA identifying shared and non shared latent variables between $\mathbf{X}$ and $\mathbf{Y}$ and hence achieving enhanced interpretability.

In order to achieve whitening the matrix $\mathbf{W}_{\mathbf{X}}$ has to satisfy the condition $\mathbf{W}_{\mathbf{X}} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{W}_{\mathbf{X}}^T = \mathbf{I}_p$, but this by itself is not sufficient to completely identify $\mathbf{W}_{\mathbf{X}}$. There are still infinitely many possible whitening transformations, and the family of whitening matrices for $\mathbf{X}$ can be written as

$$\mathbf{W}_{\mathbf{X}} = \mathbf{Q}_{\mathbf{X}} \mathbf{P}_{\mathbf{X}}^{-1/2} \mathbf{V}_{\mathbf{X}}^{-1/2}. \quad (5.5)$$

Here, $\mathbf{Q}_{\mathbf{X}}$ is an orthogonal matrix; therefore the whitening matrix $\mathbf{W}_{\mathbf{X}}$ itself is not orthogonal unless $\mathbf{P}_{\mathbf{X}} = \mathbf{V}_{\mathbf{X}} = \mathbf{I}_p$. The choice of $\mathbf{Q}_{\mathbf{X}}$ determines the type of whitening (Kessy, Lewin, and Strimmer, 2018). For example, using $\mathbf{Q}_{\mathbf{X}} = \mathbf{I}_p$ leads to ZCA-cor whitening, also known as Mahalanobis whitening based on the correlation matrix. PCA-cor whitening, another widely used sphering technique, is obtained by setting $\mathbf{Q}_{\mathbf{X}} = \mathbf{G}^T$, where $\mathbf{G}$ is the eigensystem resulting from the spectral decomposition of the correlation matrix $\mathbf{P}_{\mathbf{X}} = \mathbf{G} \boldsymbol{\Theta} \mathbf{G}^T$. Since there is a sign ambiguity in the eigenvectors $\mathbf{G}$ we adopt the convention of Kessy, Lewin, and Strimmer (2018) to adjust columns signs of $\mathbf{G}$, or equivalently row signs of $\mathbf{Q}_{\mathbf{X}}$, so that the rotation matrix $\mathbf{Q}_{\mathbf{X}}$ has a positive diagonal.

The corresponding inverse relation $\mathbf{X} = \mathbf{W}_{\mathbf{X}}^{-1} \widetilde{\mathbf{X}} = \boldsymbol{\Phi}_{\mathbf{X}}^T \widetilde{\mathbf{X}}$ is called a *coloring transformation*, where the matrix $\mathbf{W}_{\mathbf{X}}^{-1} = \boldsymbol{\Phi}_{\mathbf{X}}^T$ is the *mixing matrix*, or *coloring matrix* that we can write in terms of rotation matrix $\mathbf{Q}_{\mathbf{X}}$ as

$$\boldsymbol{\Phi}_{\mathbf{X}} = \mathbf{Q}_{\mathbf{X}} \mathbf{P}_{\mathbf{X}}^{1/2} \mathbf{V}_{\mathbf{X}}^{1/2} \quad (5.6)$$

Like $\mathbf{W}_X$ the mixing matrix Φ_X is not orthogonal. The entries of the matrix Φ_X are called the *loadings*, i.e. the coefficients linking the whitened variable $\widetilde{\mathbf{X}}$ with the original $\mathbf{X}$. Since $\widetilde{\mathbf{X}}$ is a white random vector with $\text{cov}(\widetilde{\mathbf{X}}) = \mathbf{I}_p$ the loadings are equivalent to the covariance $\text{cov}(\widetilde{\mathbf{X}}, \mathbf{X}) = \Phi_X$. The corresponding correlations, also known as *correlation-loadings*, are

$$\text{cor}(\widetilde{\mathbf{X}}, \mathbf{X}) = \Psi_X = \Phi_X \mathbf{V}_X^{-1/2} = \mathbf{Q}_X \mathbf{P}_X^{1/2}. \quad (5.7)$$

Note that the sum of squared correlations in each column of Ψ_X sum up to 1, as $\text{diag}(\Psi_X^T \Psi_X) = \text{diag}(\mathbf{P}_X) = \mathbf{I}_p$.

Our approach to CCA whitening

We will show now that CCA has a very close relationship to whitening. In particular, the objective of CCA can be seen to be equivalent to simultaneous whitening of both $\mathbf{X}$ and $\mathbf{Y}$, with a diagonality constraint on the cross-correlation matrix between the whitened $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$.

First, we make the choice to standardize the canonical directions α_i and β_i according to $\text{var}(\alpha_i^T \mathbf{X}) = \alpha_i^T \Sigma_X \alpha_i = 1$ and $\text{var}(\beta_i^T \mathbf{Y}) = \beta_i^T \Sigma_Y \beta_i = 1$. As a result α_i and β_i form the basis of two whitening matrices, $\mathbf{W}_X = (\alpha_1, \dots, \alpha_p)^T = \mathbf{A}$ and $\mathbf{W}_Y = (\beta_1, \dots, \beta_q)^T = \mathbf{B}$, with *rows* containing the canonical directions. The length constraint $\alpha_i^T \Sigma_X \alpha_i = 1$ thus becomes $\mathbf{W}_X \Sigma_X \mathbf{W}_X^T = \mathbf{I}_p$ meaning that $\mathbf{W}_X$ (and $\mathbf{W}_Y$) is indeed a valid whitening matrix.

Second, after whitening $\mathbf{X}$ and $\mathbf{Y}$ individually to $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ using $\mathbf{W}_X$ and $\mathbf{W}_Y$, respectively, the joint covariance of $(\widetilde{\mathbf{X}}^T, \widetilde{\mathbf{Y}}^T)^T$ is $\begin{pmatrix} \mathbf{I}_p & \mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}} \\ \mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}}^T & \mathbf{I}_q \end{pmatrix}$. Note that whitening of $(\mathbf{X}^T, \mathbf{Y}^T)^T$ simultaneously would in contrast lead to a fully diagonal covariance matrix. In the above $\mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}} = \text{cor}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) = \text{cov}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ is the cross-correlation matrix between the two whitened vectors and can be expressed as

$$\mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}} = \mathbf{W}_X \Sigma_{XY} \mathbf{W}_Y^T = \mathbf{Q}_X \mathbf{K} \mathbf{Q}_Y^T = (\tilde{\rho}_{ij}) \quad (5.8)$$

and

$$\mathbf{K} = \mathbf{P}_X^{-1/2} \mathbf{P}_{XY} \mathbf{P}_Y^{-1/2} = (k_{ij}). \quad (5.9)$$

Following the terminology in Zuber and Strimmer (2011) we may call $\mathbf{K}$ the correlation-adjusted cross-correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$.

With this setup the CCA objective can be framed simply as the demand that $\text{cor}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \mathbf{P}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$ must be diagonal. Since in whitening the orthogonal matrices $\mathbf{Q}_{\mathbf{X}}$ and $\mathbf{Q}_{\mathbf{Y}}$ can be freely selected we can achieve diagonality of $\mathbf{P}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$ and hence pinpoint the CCA whitening matrices by applying singular value decomposition to

$$\mathbf{K} = (\mathbf{Q}_{\mathbf{X}}^{\text{CCA}})^T \mathbf{A} \mathbf{Q}_{\mathbf{Y}}^{\text{CCA}}. \quad (5.10)$$

This provides the rotation matrices $\mathbf{Q}_{\mathbf{X}}^{\text{CCA}}$ and the $\mathbf{Q}_{\mathbf{Y}}^{\text{CCA}}$ of dimensions $m \times p$ and $m \times q$, respectively, and the $m \times m$ matrix $\mathbf{A} = \text{diag}(\lambda_i)$ containing the singular values of $\mathbf{K}$, which are also the singular values of $\mathbf{P}_{\tilde{\mathbf{X}}\tilde{\mathbf{Y}}}$. Since $m = \min(p, q)$ the larger of the two rotation matrices will not be a square matrix but it can nonetheless be used for whitening via Eq. 5.4 and Eq. 5.5 since it still is semi-orthogonal with $\mathbf{Q}_{\mathbf{X}}^{\text{CCA}}(\mathbf{Q}_{\mathbf{X}}^{\text{CCA}})^T = \mathbf{Q}_{\mathbf{Y}}^{\text{CCA}}(\mathbf{Q}_{\mathbf{Y}}^{\text{CCA}})^T = \mathbf{I}_m$. As a result, we obtain $\text{cor}(\tilde{X}_i^{\text{CCA}}, \tilde{Y}_i^{\text{CCA}}) = \lambda_i$ for $i = 1 \dots m$, i.e. the canonical correlations are identical to the singular values of $\mathbf{K}$.

Hence, CCA may be viewed as the outcome of a uniquely determined whitening transformation with underlying sphering matrices $\mathbf{W}_{\mathbf{X}}^{\text{CCA}}$ and $\mathbf{W}_{\mathbf{Y}}^{\text{CCA}}$ induced by the rotation matrices $\mathbf{Q}_{\mathbf{X}}^{\text{CCA}}$ and $\mathbf{Q}_{\mathbf{Y}}^{\text{CCA}}$. Thus, the distinctive feature of CCA whitening, in contrast to other common forms of whitening described in Kessy, Lewin, and Strimmer (2018), is that by construction it is not only informed by $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{Y}}$ but also by $\mathbf{P}_{\mathbf{XY}}$, which fixes all remaining rotational freedom.

Relationship to O2-PLS and Orthogonal CCA

O2-PLS (Trygg and Wold, 2003) and, more recently, Orthogonal CCA (Cunningham and Ghahramani, 2015) are two further proposals for modeling multiple multivariate data measured on the same individual samples. Both are very closely related to CCA and use similar linear projection models. O2-PLS is widely used for data integration in the field of chemometrics.

The crucial difference of these two approaches in comparison with CCA is that both O2-PLS and Orthogonal CCA enforce the orthogonality of the projection matrices ($\mathbf{A}$ and $\mathbf{B}$) and the corresponding loading matrices. In contrast, in CCA the canonical directions α_i and β_j are *not* orthogonal, since the whitening matrices $\mathbf{W}_{\mathbf{X}}^{\text{CCA}}$ and $\mathbf{W}_{\mathbf{Y}}^{\text{CCA}}$ and the loadings $\Phi_{\mathbf{X}}^{\text{CCA}}$ and $\Phi_{\mathbf{Y}}^{\text{CCA}}$ are not orthogonal either.

The orthogonality constraints of O2-PLS and Orthogonal PLS lead to corresponding projected score vectors that have correlated components. Contrarily,

CCA scores $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ by construction have identity covariance matrix and thus uncorrelated components.

The benefits of whitened canonical variables will become evident when constructing a probabilistic model for CCA since independence of the underlying latent variables is a very desirable property as it simultaneously aids model building, computation, and interpretation.

5.2.3 CCA and latent variable multivariate regression

Next, we investigate the close link of CCA with multivariate regression. This not only allows to interpret CCA as special case of regression and to better understand canonical correlations, and but also offers the basis for a probabilistic latent variable model of CCA based on whitening.

Optimal linear multivariate predictor

In multivariate regression the aim is to build a model that, given an input vector $\mathbf{X}$, predicts a vector $\mathbf{Y}$ as well as possible according to a specific measure such as squared error. Assuming a linear relationship $\mathbf{Y}^* = \mathbf{a} + \mathbf{b}^T \mathbf{X}$ is the predictor random variable, with mean $E(\mathbf{Y}^*) = \boldsymbol{\mu}_{\mathbf{Y}^*} = \mathbf{a} + \mathbf{b}^T \boldsymbol{\mu}_{\mathbf{X}}$. The expected squared difference between $\mathbf{Y}$ and $\mathbf{Y}^*$, i.e. the mean squared prediction error

$$\begin{aligned} \text{MSE} &= \text{Tr}(E((\mathbf{Y} - \mathbf{Y}^*)(\mathbf{Y} - \mathbf{Y}^*)^T)) \\ &= \sum_{i=1}^q E((Y_i - Y_i^*)^2), \end{aligned} \tag{5.11}$$

is a natural measure of how well $\mathbf{Y}^*$ predicts $\mathbf{Y}$. As a function of the model parameters $\mathbf{a}$ and $\mathbf{b}$ the predictive MSE becomes

$$\begin{aligned} \text{MSE}(\mathbf{a}, \mathbf{b}) &= \text{Tr}((\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}^*})(\boldsymbol{\mu}_{\mathbf{Y}} - \boldsymbol{\mu}_{\mathbf{Y}^*})^T + \\ &\quad \boldsymbol{\Sigma}_{\mathbf{Y}} + \mathbf{b}^T \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{b} - 2\mathbf{b}^T \boldsymbol{\Sigma}_{\mathbf{XY}}). \end{aligned} \tag{5.12}$$

Optimal parameters for best linear predictor are found by minimizing this MSE function. For the offset $\mathbf{a}$ this yields

$$\mathbf{a} = \boldsymbol{\mu}_{\mathbf{Y}} - (\mathbf{b})^T \boldsymbol{\mu}_{\mathbf{X}} \tag{5.13}$$

which regardless of the value of $\mathbf{b}$ ensures $\boldsymbol{\mu}_{Y^*} - \boldsymbol{\mu}_Y = 0$. Likewise, for the matrix of regression coefficients minimization results in

$$\mathbf{b}^{\text{all}} = \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{XY} \quad (5.14)$$

with minimum achieved $\text{MSE}(\mathbf{a}^{\text{all}}, \mathbf{b}^{\text{all}}) = \text{Tr}(\boldsymbol{\Sigma}_Y) - \text{Tr}(\boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{XY})$.

If we exclude predictors from the model by setting regression coefficients $\mathbf{b}^{\text{zero}} = 0$ then the corresponding optimal intercept is $\mathbf{a}^{\text{zero}} = \boldsymbol{\mu}_Y$ and the minimum achieved $\text{MSE}(\mathbf{a}^{\text{zero}}, \mathbf{b}^{\text{zero}}) = \text{Tr}(\boldsymbol{\Sigma}_Y)$. Thus, by adding predictors $\mathbf{X}$ to the model the predictive MSE is reduced, and hence the fit of the model correspondingly improved, by the amount

$$\begin{aligned} \Delta &= \text{MSE}(\mathbf{a}^{\text{zero}}, \mathbf{b}^{\text{zero}}) - \text{MSE}(\mathbf{a}^{\text{all}}, \mathbf{b}^{\text{all}}) \\ &= \text{Tr}(\boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_{XY}) \\ &= \text{Tr}(\text{cov}(\mathbf{Y}, \mathbf{Y}^{\text{all}*})) . \end{aligned} \quad (5.15)$$

If the response Y is univariate ($q = 1$) then Δ reduces to the variance-scaled coefficient of determination $\sigma_Y^2 \mathbf{P}_{YX} \mathbf{P}_X^{-1} \mathbf{P}_{XY}$. Note that in the above no distributional assumptions are made other than specification of means and covariances.

Regression view of CCA

The first step to understand CCA as a regression model is to consider multivariate regression between two whitened vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ (considering whitening of any type, including but not limited to CCA-whitening). Since $\boldsymbol{\Sigma}_{\widetilde{\mathbf{X}}} = \mathbf{I}_p$ and $\boldsymbol{\Sigma}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}} = \mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}}$ the optimal regression coefficients to predict $\widetilde{\mathbf{Y}}$ from $\widetilde{\mathbf{X}}$ are given by

$$\mathbf{b}^{\text{all}} = \mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}} , \quad (5.16)$$

i.e. the pairwise correlations between the elements of the two vectors $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. Correspondingly, the decrease in predictive MSE due to including the predictors $\widetilde{\mathbf{X}}$ is

$$\begin{aligned} \Delta &= \text{Tr}(\mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}}^T \mathbf{P}_{\widetilde{\mathbf{X}}\widetilde{\mathbf{Y}}}) = \sum_{i,j} \widetilde{\rho}_{ij}^2 \\ &= \text{Tr}(\mathbf{K}^T \mathbf{K}) = \sum_{i,j} k_{ij}^2 \\ &= \text{Tr}(\boldsymbol{\Lambda}^2) = \sum_i \lambda_i^2 . \end{aligned} \quad (5.17)$$

In the special case of CCA-whitening the regression coefficients further simplify to $b_{ii}^{\text{all}} = \lambda_i$, i.e. the canonical correlations λ_i act as the regression coefficients linking CCA-whitened $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$. Furthermore, as the decrease in predictive MSE Δ is the sum of the squared canonical correlations (cf. Eq. 5.17), each λ_i^2 can be interpreted as the variable importance of the corresponding variable in $\tilde{\mathbf{X}}^{\text{CCA}}$ to predict the outcome $\tilde{\mathbf{Y}}^{\text{CCA}}$. Thus, CCA directly results from multivariate regression between CCA-whitened random vectors, where the canonical correlations λ_i assume the role of regression coefficients and λ_i^2 provides a natural measure to rank the canonical components in order of their respective predictive capability.

A key difference between classical CCA and regression is that in the latter both positive and negative coefficients are allowed to account for the directionality of the influence of the predictors. In contrast, in classical CCA only positive canonical correlations are permitted by convention. To reflect that CCA analysis is inherently a regression model we advocate here that canonical correlations should indeed be allowed to assume both positive and negative values, as fundamentally they are regression coefficients. This can be implemented by exploiting the sign ambiguity in the singular value decomposition of $\mathbf{K}$ (Eq. 5.10). In particular, the rows signs of $\mathbf{Q}_X^{\text{CCA}}$ and $\mathbf{Q}_Y^{\text{CCA}}$ and the signs of λ_i can be revised simultaneously without affecting $\mathbf{K}$. We propose to choose $\mathbf{Q}_X^{\text{CCA}}$ and $\mathbf{Q}_Y^{\text{CCA}}$ such that both rotation matrices have a positive diagonal, and then to adjust the signs of the λ_i accordingly. Note that orthogonal matrices with positive diagonals are closest to the identity matrix (e.g. in terms of the Frobenius norm) and thus constitute minimal rotations.

Generative latent variable model for CCA

With the link of CCA to whitening and multivariate regression established it is straightforward to arrive at simple and easily interpretable generative probabilistic latent variable model for CCA. This model has two levels of hidden variables: it uses uncorrelated latent variables $\mathbf{Z}^{\mathbf{X}}, \mathbf{Z}^{\mathbf{Y}}, \mathbf{Z}^{\text{shared}}$ (level 1) with zero mean and unit variance to generate the CCA-whitened variables $\tilde{\mathbf{X}}^{\text{CCA}}$ and $\tilde{\mathbf{Y}}^{\text{CCA}}$ (level 2) which in turn produce the observed vectors $\mathbf{X}$ and $\mathbf{Y}$ – see Fig. 5.1

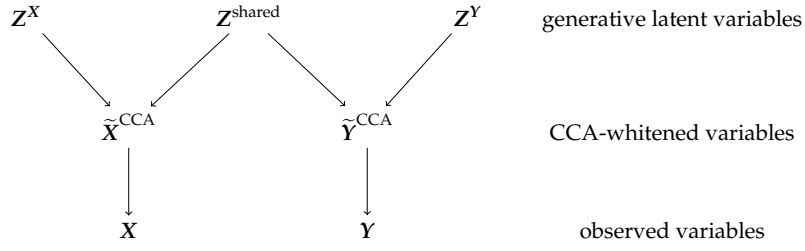


FIGURE 5.1: Probabilistic CCA as a two layer latent variable generative model. The middle layer contains the CCA-whitened variables $\tilde{\mathbf{X}}^{\text{CCA}}$ and $\tilde{\mathbf{Y}}^{\text{CCA}}$, and the top layer the uncorrelated generative latent variables $\mathbf{Z}^{\mathbf{X}}$, $\mathbf{Z}^{\mathbf{Y}}$, and $\mathbf{Z}^{\text{shared}}$.

Specifically, on the first level we have latent variables

$$\begin{aligned}\mathbf{Z}^{\mathbf{X}} &\sim F_{\mathbf{Z}^{\mathbf{X}}}, \\ \mathbf{Z}^{\mathbf{Y}} &\sim F_{\mathbf{Z}^{\mathbf{Y}}}, \text{ and} \\ \mathbf{Z}^{\text{shared}} &\sim F_{\mathbf{Z}^{\text{shared}}},\end{aligned}\tag{5.18}$$

with $\mathbb{E}(\mathbf{Z}^{\mathbf{X}}) = \mathbb{E}(\mathbf{Z}^{\mathbf{Y}}) = \mathbb{E}(\mathbf{Z}^{\text{shared}}) = 0$ and $\text{var}(\mathbf{Z}^{\mathbf{X}}) = \mathbf{I}_p$, $\text{var}(\mathbf{Z}^{\mathbf{Y}}) = \mathbf{I}_q$, and $\text{var}(\mathbf{Z}^{\text{shared}}) = \mathbf{I}_m$ and no mutual correlation among the components of $\mathbf{Z}^{\mathbf{X}}$, $\mathbf{Z}^{\mathbf{Y}}$, and $\mathbf{Z}^{\text{shared}}$. The second level latent variables are then generated by mixing shared and non-shared variables according to

$$\begin{aligned}\tilde{X}_i^{\text{CCA}} &= (1 - |\lambda_i|)^{1/2} Z_i^{\mathbf{X}} + |\lambda_i|^{1/2} Z_i^{\text{shared}} \\ \tilde{Y}_i^{\text{CCA}} &= (1 - |\lambda_i|)^{1/2} Z_i^{\mathbf{Y}} + |\lambda_i|^{1/2} Z_i^{\text{shared}} \text{sgn}(\lambda_i)\end{aligned}\tag{5.19}$$

where the parameters $\lambda_1, \dots, \lambda_m$ can be positive as well as negative and range from -1 to 1 (See Appendix B for further details.). The components $i > m$ are always non-shared and taken from $\mathbf{Z}^{\mathbf{X}}$ or $\mathbf{Z}^{\mathbf{Y}}$ as appropriate, i.e. as above but with $\lambda_{i>m} = 0$. By construction, this results in $\text{var}(\tilde{\mathbf{X}}^{\text{CCA}}) = \mathbf{I}_p$, $\text{var}(\tilde{\mathbf{Y}}^{\text{CCA}}) = \mathbf{I}_q$ and $\text{cov}(\tilde{X}_i^{\text{CCA}}, \tilde{Y}_i^{\text{CCA}}) = \lambda_i$ (see Appendix A for further details). Finally, the observed variables are produced by a coloring transformation (cf. Section 5.2.2) and subsequent translation

$$\begin{aligned}\mathbf{X} &= \Phi_{\mathbf{X}}^T \tilde{\mathbf{X}}^{\text{CCA}} + \mu_{\mathbf{X}} \\ \mathbf{Y} &= \Phi_{\mathbf{Y}}^T \tilde{\mathbf{Y}}^{\text{CCA}} + \mu_{\mathbf{Y}}\end{aligned}\tag{5.20}$$

Note that the above probabilistic model for CCA is in fact not a single model

but a family of models, since we do not completely specify the underlying distributions, only their means and (co)variances. While in practice we will typically assume normally distributed generative latent variables, and hence normally distributed observations, it is equally possible to employ other distributions for the first level latent variables. For example, a rescaled t -distribution with a wider tail than the normal distribution may be employed to obtain a robustified version of CCA (Adrover and Donato, 2015).

5.3 Entropy-based criteria for multivariate association

In this section we show that the matrix $\mathbf{K}$ can be linked to a measure of multivariate association between the two groups of random variables in $\mathbf{Y}$ and $\mathbf{X}$ and can thus be used to uncover structural relations of integrated variables. More specifically, i) we demonstrate the $\mathbf{K}$ is related to *the vector correlation coefficient* (Rozeboom, 1965) by entropy derivation and ii) we propose to use vector correlation coefficient as alternative to the RV coefficient (Escoufier, 1973) to measure association between two random vectors.

5.3.1 The RV coefficient

Multivariate association measures are also known as matrix correlation measures. Many measures of multivariate association have been suggested in the literature e.g. (Cramer and Nicewander, 1979) with recent review conducted by Josse and Holmes (2016). Using entropy we have recovered a number of these multivariate measures. In addition, some further have been proposed, in particular, the RV coefficient (Escoufier, 1973) defined by:

$$\text{RV}(\mathbf{X}, \mathbf{Y}) = \frac{\text{Tr}(\mathbf{\Sigma}_{\mathbf{XY}}\mathbf{\Sigma}_{\mathbf{YX}})}{\text{Tr}(\mathbf{\Sigma}_{\mathbf{X}}^2)\text{Tr}(\mathbf{\Sigma}_{\mathbf{Y}}^2)} \quad (5.21)$$

appears to be one of the most widely used criteria in practice. Its popularity is due to the fact that it is considered as a natural multivariate generalization of the correlation coefficient to describe associations between two groups of variables instead of just two variables. As is desirable, for the bivariate case with $q = p = 1$ the RV coefficient reduces to the simple pairwise squared correlation. However, the RV coefficient is subject to the following limitations:

- Although it reduces to Pearson correlation for $p = q = 1$, the RV coefficient does not reduce to the squared multiple correlation coefficient $\text{cor}(Y, \mathbf{X})^2$ for $q = 1$ and p arbitrary.
- The RV coefficient can exhibit large values even if the two random vectors are independent. To address this limitation two variants were proposed: the modified RV coefficient by Smilde et al. (2009) and adjusted RV coefficient by Mayer, Lorent, and Horgan (2011).
- No decomposition available to the best of our knowledge.

5.3.2 Approach

Expected relative entropy to measure predictive importance

The relative entropy, or equivalently the Boltzmann-Kullback-Leibler entropy, which is defined by $I^{\text{BKL}}(F; G)$ such that:

$$I^{\text{BKL}}(F; G) = \mathbb{E}_F \left(\log \frac{f(X)}{g(X)} \right) \quad (5.22)$$

is an information-theoretic measure of the divergence between two distributions F and G over the same sample space. It was introduced in Kullback and Leibler (1951) but discovered already in the preceding century in 1878 by Boltzmann (Akaike, 1985). Essentially, $-I^{\text{BKL}}(F; G)$ quantifies the loss that occurs when the model F is approximated by G . In most applications² in statistics and machine learning F represents the true underlying model — or at least the more complex and realistic model — over which is the expectation in Eq. 5.22 is taken, whereas G is a more simplistic approximate distribution.

By construction, relative entropy is non-symmetric with regard to F and G and it is invariant against reparameterization since the volume element cancels out in the ratio of the densities. $I^{\text{BKL}}(F; G)$ is always non-negative and equals zero only for $F = G$. Furthermore, it is also the only divergence for probability measures that is simultaneously both a Bregman and an f -divergence and thus inherits many favorable mathematical properties of both classes (Amari, 2009; Csiszar, 1991).

We follow Lindley (1956) and Bernardo (1979) to measure the importance of $\mathbf{X}$ to predict $\mathbf{Y}$ by computing the expected relative entropy from the conditional distribution $F_{\mathbf{Y}|\mathbf{X}}$ (i.e. the model for $\mathbf{Y}$ using $\mathbf{X}$ as predictors) to the

²Variational Bayes methods being a notable exception.

unconditioned distribution $F_{\mathbf{Y}}$ (i.e. the marginal distribution of $\mathbf{Y}$ without any predictors). The corresponding BKL-entropy is given by

$$I^{\text{BKL}}(F_{\mathbf{Y}|\mathbf{X}}; F_{\mathbf{Y}}) = E_{F_{\mathbf{Y}|\mathbf{X}}} \log \left(\frac{f(\mathbf{Y}|\mathbf{X})}{f(\mathbf{Y})} \right), \quad (5.23)$$

which is itself a random variable since it depends on $\mathbf{X}$. Thus, taking the expectation with regard to $F_{\mathbf{X}}$ results in

$$\begin{aligned} E_{F_{\mathbf{X}}} I^{\text{BKL}}(F_{\mathbf{Y}|\mathbf{X}}; F_{\mathbf{Y}}) &= E_{F_{\mathbf{X}}} E_{F_{\mathbf{Y}|\mathbf{X}}} \log \left(\frac{f(\mathbf{Y}|\mathbf{X})f(\mathbf{X})}{f(\mathbf{Y})f(\mathbf{X})} \right) \\ &= E_{F_{\mathbf{X},\mathbf{Y}}} \log \left(\frac{f(\mathbf{X},\mathbf{Y})}{f(\mathbf{Y})f(\mathbf{X})} \right) \\ &= I^{\text{BKL}}(F_{\mathbf{X},\mathbf{Y}}; F_{\mathbf{X} \perp \mathbf{Y}}). \end{aligned} \quad (5.24)$$

This quantity is the mutual information $\text{MI}(\mathbf{X}, \mathbf{Y})$ between the two multivariate random vectors $\mathbf{X}$ and $\mathbf{Y}$ — note that in most references only $\text{MI}(X, Y)$ is considered assuming univariate X and Y .

$\text{MI}(\mathbf{X}, \mathbf{Y})$ is symmetric with regard to the roles of $\mathbf{X}$ and $\mathbf{Y}$. Since I^{BKL} is invariant against coordinate transformations $\text{MI}(\mathbf{X}, \mathbf{Y})$ summarizes both linear and nonlinear dependencies between the two groups of variables in $\mathbf{X}$ and $\mathbf{Y}$. If $\text{MI}(\mathbf{X}, \mathbf{Y}) = 0$ then $F_{\mathbf{X},\mathbf{Y}} = F_{\mathbf{X} \perp \mathbf{Y}}$ and thus $\mathbf{X}$ and $\mathbf{Y}$ are independent.

Mutual information and vector correlation coefficient

For univariate X and Y and both having a joint bivariate normal distribution it is well-known (e.g. Cover and Thomas, 2006) that

$$\text{MI}(X, Y) = -\frac{1}{2} \log(1 - \rho^2) \approx \frac{\rho^2}{2} \quad (5.25)$$

so that mutual information is a monotonic function of the squared correlation $\rho = \text{cor}(X, Y)$ between X and Y .

With $\mathbf{X}$ a vector and Y univariate, and both being joint multivariate normal random variables, the corresponding mutual information is

$$\text{MI}(\mathbf{X}, Y) = -\frac{1}{2} \log(1 - \rho_{\mathbf{X}Y}^2) \approx \frac{\rho_{\mathbf{X}Y}^2}{2} \quad (5.26)$$

which is identical to the above except that now the function depends on the squared multiple correlation coefficient $\rho_{\mathbf{X}Y}^2 = \mathbf{P}_{Y\mathbf{X}} \mathbf{P}_{\mathbf{X}}^{-1} \mathbf{P}_{\mathbf{X}Y}$ between $\mathbf{X}$ and Y .

Most generally, with both $\mathbf{X}$ and $\mathbf{Y}$ multivariate, the mutual information becomes

$$\text{MI}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{2} \log(1 - \rho_{\mathbf{XY}}^2) \quad (5.27)$$

$$\approx \rho_{\mathbf{XY}}^2/2 \quad (5.28)$$

which shows again the same functional relationship but this time with

$$\begin{aligned} \rho_{\mathbf{XY}}^2 &= 1 - \det \begin{pmatrix} \mathbf{P}_{\mathbf{X}} & \mathbf{P}_{\mathbf{XY}} \\ \mathbf{P}_{\mathbf{XY}}^T & \mathbf{P}_{\mathbf{Y}} \end{pmatrix} / (\det \mathbf{P}_{\mathbf{X}} \det \mathbf{P}_{\mathbf{Y}}) \\ &= 1 - \det (\mathbf{I}_q - \mathbf{P}_{\mathbf{Y}}^{-1} \mathbf{P}_{\mathbf{YX}} \mathbf{P}_{\mathbf{X}}^{-1} \mathbf{P}_{\mathbf{XY}}) \\ &= 1 - \det (\mathbf{I}_p - \mathbf{P}_{\mathbf{X}}^{-1} \mathbf{P}_{\mathbf{XY}} \mathbf{P}_{\mathbf{Y}}^{-1} \mathbf{P}_{\mathbf{YX}}) . \end{aligned} \quad (5.29)$$

This formulation makes it easy to see that $\rho_{\mathbf{XY}}^2$ reduces to $\rho_{\mathbf{XY}}^2 = \mathbf{P}_{\mathbf{YX}} \mathbf{P}_{\mathbf{X}}^{-1} \mathbf{P}_{\mathbf{XY}}$ the squared multiple correlation coefficient if $q = 1$. On the other hand, we can also write it equivalently as

$$\begin{aligned} \rho_{\mathbf{XY}}^2 &= 1 - \det (\mathbf{I}_q - \mathbf{K}^T \mathbf{K}) \\ &= 1 - \det (\mathbf{I}_p - \mathbf{K} \mathbf{K}^T) \\ &= 1 - \prod_{i=1}^m (1 - \lambda_i^2) . \end{aligned} \quad (5.30)$$

with $\mathbf{K} = \mathbf{P}_{\mathbf{X}}^{-1/2} \mathbf{P}_{\mathbf{XY}} \mathbf{P}_{\mathbf{Y}}^{-1/2} = (k_{ij})$ and $\lambda_1, \dots, \lambda_m$ are non-vanishing singular values of $\mathbf{K}$.

Using $\rho_{\mathbf{XY}}^2$ as measure of association between two sets of variables was first proposed in Rozeboom (1965) as vector correlation coefficient and corresponds to measure γ_2 in Cramer and Nicewander (1979). The complementary quantity $1 - \rho_{\mathbf{XY}}^2$ is the vector alienation coefficient of Hotelling (1936). A more recent application of $\rho_{\mathbf{XY}}^2$ is found in (e.g. Roverato and Castelo, 2017), but in general it seems to be largely ignored.

In terms of the above $\rho_{\mathbf{XY}}^2$, we replace the expression of $\rho_{\mathbf{XY}}^2$ in equation 5.30 into the expression of $\text{MI}(\mathbf{X}, \mathbf{Y})$ in equation 5.27 to get the mutual information

between $\mathbf{X}$ and $\mathbf{Y}$ as (see e.g. Theil and Chung (1988)):

$$\begin{aligned} \text{MI}(\mathbf{X}, \mathbf{Y}) &= -\frac{1}{2} \log \det (\mathbf{I}_q - \mathbf{K}^T \mathbf{K}) \\ &= -\frac{1}{2} \log \det (\mathbf{I}_p - \mathbf{K} \mathbf{K}^T) \\ &= -\frac{1}{2} \sum_{i=1}^m \log(1 - \lambda_i^2) . \end{aligned} \quad (5.31)$$

Equations 5.31 and 5.30 show that the matrix $\mathbf{K}$ being directly related to both the vector correlation coefficient $\rho_{\mathbf{X}\mathbf{Y}}^2$ and the mutual information $\text{MI}(\mathbf{X}, \mathbf{Y})$. In the following paragraph, we will discuss how $\mathbf{K}$ can be used as a network model and its link with the whitening approach to CCA.

Network decomposition model and link to CCA

The approximations in equations (5.25), (5.26) and (5.28) hold under independence assumptions $\mathbf{X} \perp \mathbf{Y}$ where $\rho_{\mathbf{X}\mathbf{Y}}^2$ is small. Hence by means of equations 5.28 and 5.31 we have:

$$\rho_{\mathbf{X}\mathbf{Y}}^2 \approx -\sum_{i=1}^m \log(1 - \lambda_i^2) \quad (5.32)$$

In the case where $\mathbf{X} \perp \mathbf{Y}$, $\text{MI}(\mathbf{X}, \mathbf{Y})$ is also small and hence λ_i are also small as per equation 5.31. We can then write $-\sum_{i=1}^m \log(1 - \lambda_i^2) = \sum_{i=1}^m \lambda_i^2$ which gives the following approximation of the vector correlation coefficient :

$$\begin{aligned} \rho_{\mathbf{X}\mathbf{Y}}^2 &\approx \sum_{i=1}^m \lambda_i^2 \\ &= \text{Tr}(\mathbf{K}^T \mathbf{K}) \\ &= \text{Tr}(\mathbf{K} \mathbf{K}^T) \\ &= \sum_{i=1}^p \sum_{j=1}^q k_{ij}^2 \end{aligned} \quad (5.33)$$

In this context, note that $\sum_{i=1}^m \lambda_i^2/m$ is the “generalized coefficient of determination” (GCD) in Yanai (1974) and Cramer and Nicewander (1979).

Similarly, for $\text{MI}(\mathbf{X}, \mathbf{Y})$ the approximation is $\text{MI}(\mathbf{X}, \mathbf{Y}) \approx \sum_{i=1}^p \sum_{j=1}^q k_{ij}^2/2$. Consequently, $\rho_{\mathbf{X}\mathbf{Y}}^2$ and $2\text{MI}(\mathbf{X}, \mathbf{Y})$ can be decomposed, at least in first order approximation, as the sum of the squared individual pairwise cross-correlations

between the components of two independent variables. The independence assumption always holds for whitened variables $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$, hence, $\mathbf{K}$ can be safely used to compute $\rho_{\mathbf{XY}}^2$ and $2\text{MI}(\mathbf{X}, \mathbf{Y})$ offering a decomposition property that can be exploited to construct a network with pairwise contributions. We should note here that for $q = 1$ or $p = 1$ the decomposition property is not an approximation but exact.

The use of whitened variables $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ yields a direct link to CCA. In fact, the matrix $\mathbf{K}$ is simply the cross-correlation matrix between the CCA whitened versions $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ and λ_i are the $m = \min(p, q)$ canonical correlations.

Note that Theil and Chung (1988) used the fact that mutual information can be expressed as $\text{MI}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^m \text{MI}(\tilde{X}_i^{\text{CCA}}, \tilde{Y}_i^{\text{CCA}})$ to derive decompositions which are difficult to interpret. In contrast, we decompose $\rho_{\mathbf{XY}}^2$ over the squared pairwise contributions k_{ij}^2 easily interpreted as the cross-correlations between the CCA whitened versions of $\mathbf{X}$ and $\mathbf{Y}$.

5.4 Results

We now analyze simulated and three experimental omics data sets to illustrate our approach. Specifically, on synthetic data we demonstrate the capability of our variant of CCA to identify negative canonical correlations among canonical variates. On real data we investigate the application of CCA datasets where the number of samples n is smaller than the number of variables p and q and explore the network structure resulting from the decomposition of the vector correlation coefficient.

5.4.1 Synthetic data

In order to test whether our algorithm allows to correctly identify negative canonical correlations we conducted simulations using synthetic data. Specifically, we generated data $\mathbf{x}_i$ and $\mathbf{y}_i$ from a $p + q$ dimensional multivariate normal distribution with zero mean and covariance matrix $\begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{XY}}^T & \boldsymbol{\Sigma}_{\mathbf{Y}} \end{pmatrix}$ where $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathbf{I}_p$, $\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{I}_q$ and $\boldsymbol{\Sigma}_{\mathbf{XY}} = \text{diag}(\lambda_i)$. The canonical correlations were set to have alternating positive and negative signs $\lambda_1 = \lambda_3 = \lambda_5 = \lambda_7 = \lambda_9 = \lambda$ and $\lambda_2 = \lambda_4 = \lambda_6 = \lambda_8 = \lambda_{10} = -\lambda$ with varying strength $\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. A similar setup was used in Cruz-Cano and Lee (2014). The dimensions were fixed at $p = 60$ and $q = 10$ and the sample

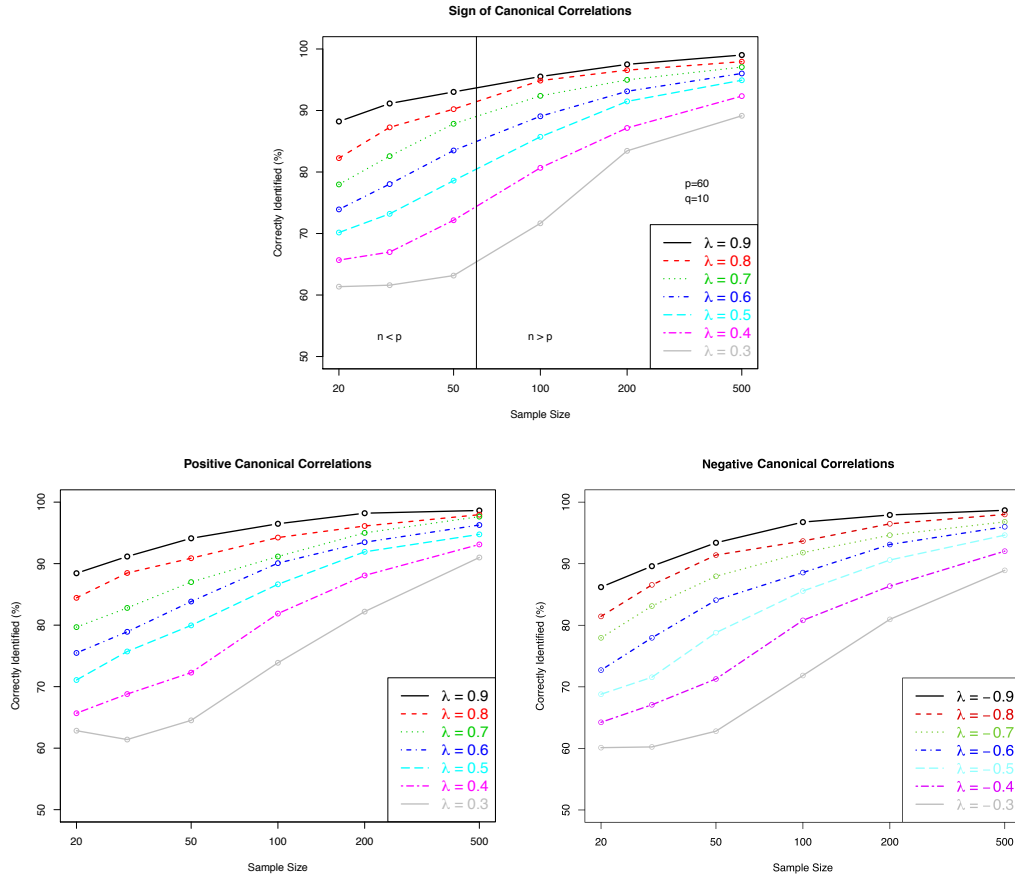


FIGURE 5.2: Top: Percentage of estimated canonical correlations with correctly identified signs in dependence of the sample size and the strength of the true canonical correlation. Bottom: Percentage of estimated positive and negative canonical correlations respectively (from left to right) in dependence of the sample size and the strength of the true canonical correlation. Performance in recovering the signs separately for positive and negative canonical correlations is very similar for all sample sizes.

size was $n \in \{20, 30, 50, 100, 200, 500\}$ so that both the small and large sample regime was covered. For each combination of n and λ the simulations were repeated 500 times, and our algorithm using shrinkage estimation of the underlying covariance matrices was employed to each of the 500 data sets to fit our CCA model. The resulting estimated canonical correlations were then compared with the corresponding true canonical correlations, and the proportion of correctly estimated signs was recorded.

The outcome from this simulation study is summarized graphically in Fig. 5.2. The key finding is that, depending on the strength of correlation λ and sample

size n , our algorithm correctly determines the sign of both negative and positive canonical correlations³. As expected, the proportion of correctly classified canonical correlations increases with sample size and with the strength of correlation. Remarkably, even for comparatively weak correlation such as $\lambda = 0.5$ and low sample size still the majority of canonical correction were estimated with the true sign. In short, this simulation demonstrates that negative canonical correlations present among pairs of canonical variables these will be detected by our approach. For more quantitative assessment of the estimated canonical correlations see Appendix B. Results in Appendix B also show how estimated canonical correlations compare to sparse CCA.

5.4.2 Nutrimouse data

We now analyze two experimental omics data sets to illustrate our approach. Specifically, we demonstrate the capability of our variant of CCA to identify negative canonical correlations among canonical variates as well its application to high-dimensional data where the number of samples n is smaller than the number of variables p and q .

This data set is due to Martin et al. (2007) and results from a nutrigenomic study in the mouse studying $n = 50$ animals. The $\mathbf{X}$ variable collects the measurements of the gene expression of $p = 120$ genes in liver cells. These were selected a priori as potentially involved in nutritional problems. The $\mathbf{Y}$ variable contains lipid concentrations of $q = 21$ hepatic fatty acids, measured on the same animals. Before further analysis we standardized both $\mathbf{X}$ and $\mathbf{Y}$.

Since the number of available samples n is smaller than the number of genes p we used shrinkage estimation to obtain the joint correlation matrix which resulted in a shrinkage intensity of $\lambda_{\text{cor}} = 0.16$ ⁴. The resulting correlations are visualized in Fig. 5.3. Subsequently, we computed canonical directions and associated canonical correlations $\lambda_1, \dots, \lambda_{21}$. The canonical correlations are shown in Fig. 5.4, and range in value between -0.96 and 0.87. As can be seen, 16 of the 21 canonical correlations are negative, including the first three top ranking correlations. In Fig. 5.5 we depict the squared correlation loadings between the first 5 components of the canonical covariates $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ and the corresponding observed variables $\mathbf{X}$ and $\mathbf{Y}$. This visualization shows that

³Note that other CCA models are not able to identify both negative and positive canonical correlations.

⁴ $\lambda_{\text{cor}} = 0.16$ is estimated using an analytic formula from Schäfer and Strimmer (2005) via the `cor.shrink` function in `corpcor` R package.

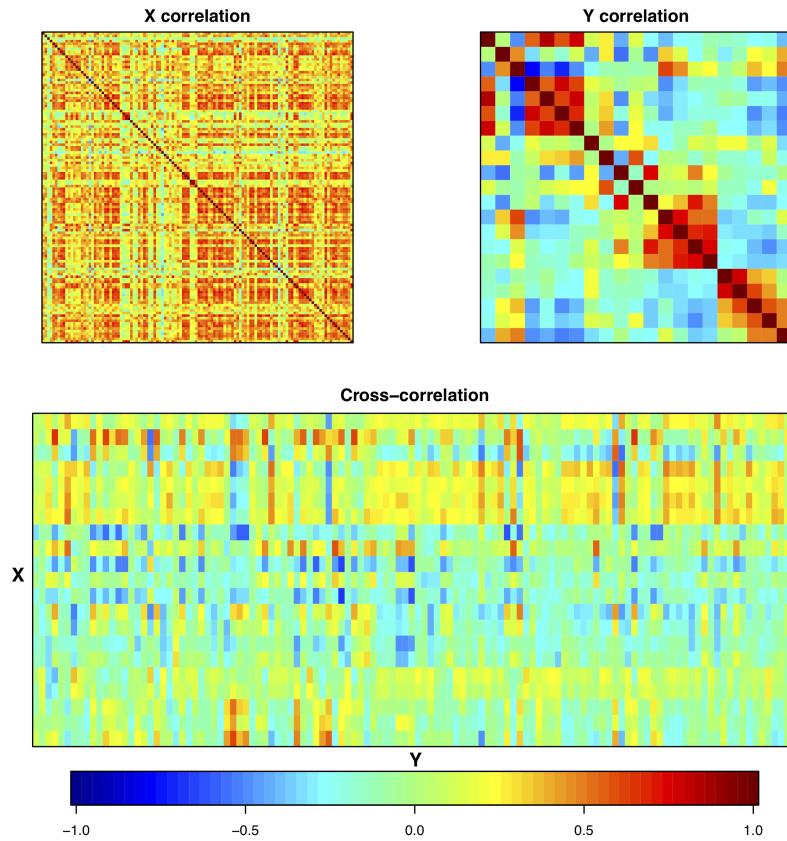


FIGURE 5.3: Visualization of the correlations within and between gene expression levels ($\mathbf{X}$) and lipid concentrations ($\mathbf{Y}$) from the Nutrino mouse study. The graphs show strong correlation patterns (≥ 0.5) within the gene expression data compared to the lipids data. The cross-correlation matrix indicates mild correlation patterns between the genes and the lipids which will be further exploited by CCA.

most information about the correlation structure within and between the two data sets (gene expression and lipid concentrations) is concentrated in the first few latent components. Particularly, most information seems to be concentrated in the first component for gene expression and in the first 4 to 5 components for the lipids data. This imbalance between both datasets is expected as fatty acids profiled in the lipids data are directly related to the oils used for the preparation of the different diets i.e. corn and colza oils for a reference diet (REF), hydrogenated coconut oil (COC), sunflower oil (SUN), linseed oil LIN) and fish oils for the FISH diet. Each of the $\mathbf{Y}$ components actually explains specific diets as confirmed by further investigation of the scatter plots both between corresponding pairs of $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ canonical variates (Fig. 5.6)

as well as within each variate (Fig. 5.7).

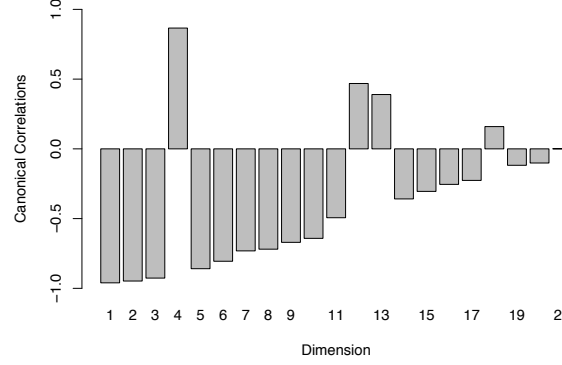


FIGURE 5.4: Plot of the estimated canonical correlations for the Nutrimouse data. The majority of the correlations indicate a negative association between the corresponding canonical variables.

Specifically, the first CCA component allow to identify the genotype of the mice (wt: wild type; ppar: PPAR- α deficient) whereas the subsequent few components reveal the imprint of the effect of the various diets (COC: coconut oil; FISH: fish oils; LIN: linseed oils; REF: reference diet; SUN: sunflower oil) on gene expression and lipid concentrations. Correspondingly, in figure B.3 in Appendix B we observe that the whitening approach yields better correlation between pairs of canonical covariates than sparse or classical CCA as implemented in the CCA R package. For a more detailed comparison of these results with PCA, classical and sparse CCA see Appendix B.

In the next step we compare the network structure induced by the K matrix in the CCA whitening approach to that of sparse CCA. From figure 5.8, we observe that whitening CCA yields better sparsity which facilitates interpretation.

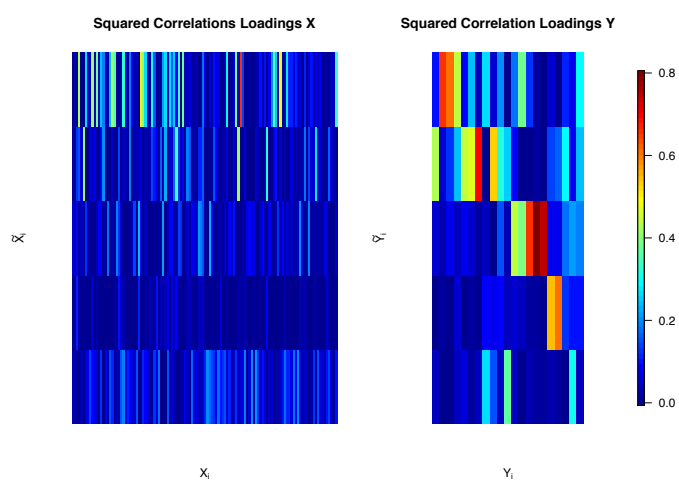


FIGURE 5.5: Squared correlations loadings between the first 5 components of the canonical covariates $\tilde{\mathbf{X}}^{\text{CCA}}$ and $\tilde{\mathbf{Y}}^{\text{CCA}}$ and the corresponding observed variables $\mathbf{X}$ and $\mathbf{Y}$ for the Nutrimouse data.

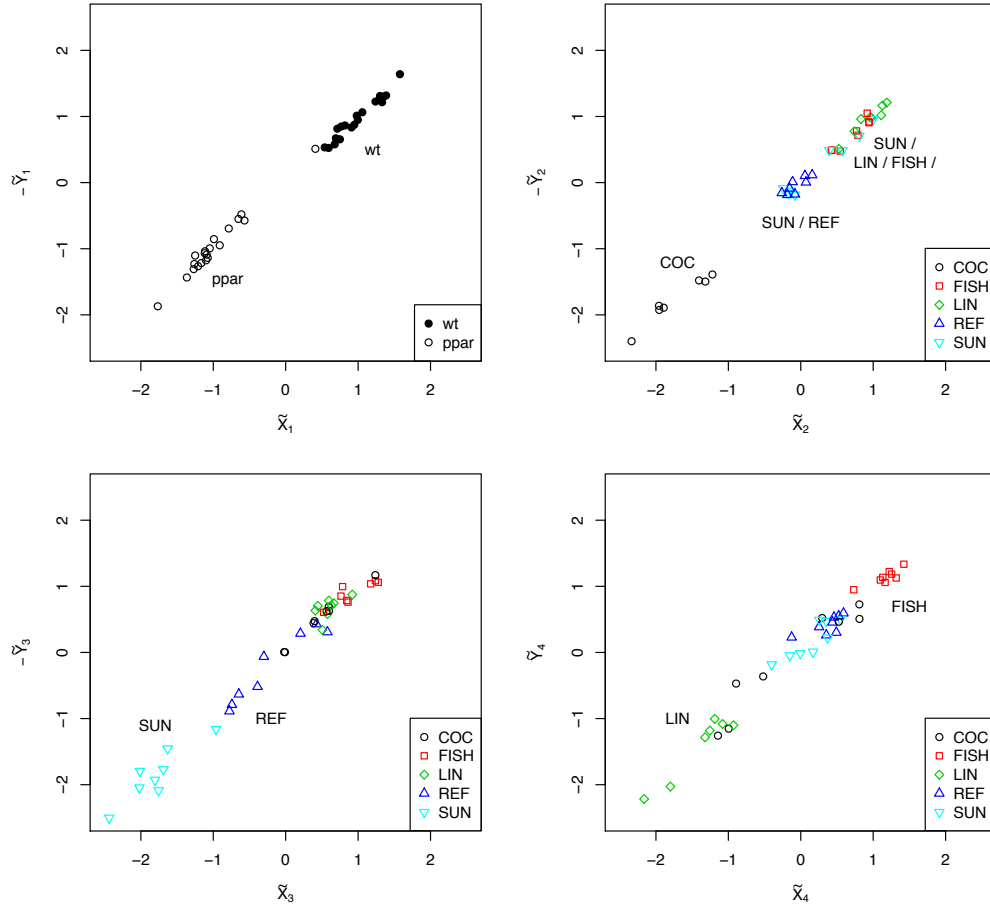


FIGURE 5.6: Scatter plots between corresponding pairs of canonical covariates for the Nutrimouse data.

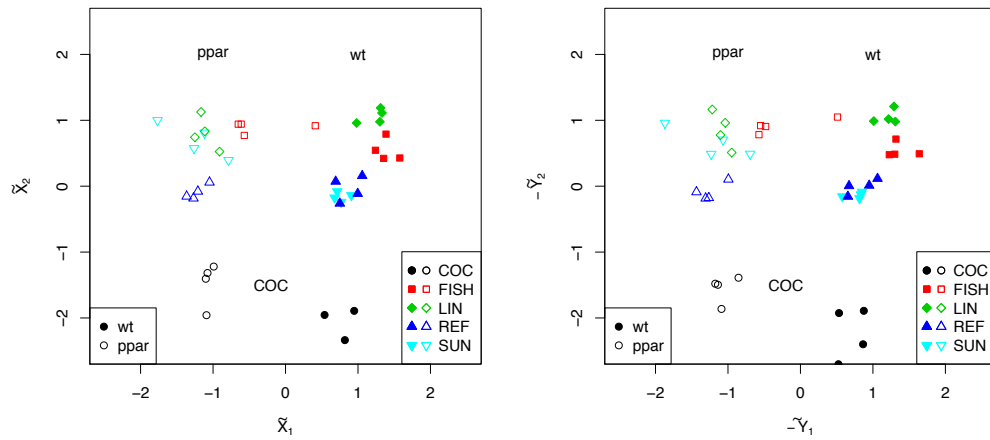
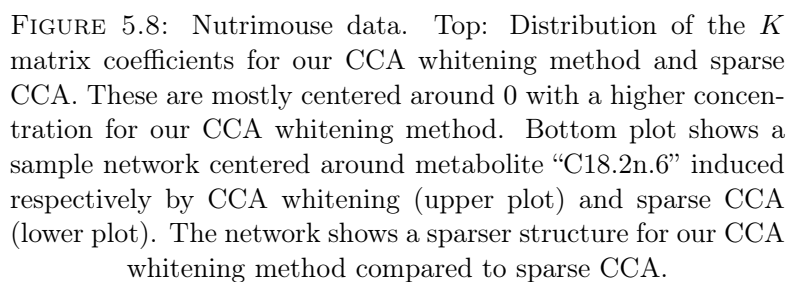


FIGURE 5.7: Scatter plots between first and second components within each canonical covariate for the Nutrimouse data.



5.4.3 NCI-60 microRNA data

In order to elucidate whether this might be a common pattern we also reanalyzed a high dimensional data from Liu et al. (2010), see also Cruz-Cano and Lee (2014). In this data set the $\mathbf{X}$ variable describes the expression level of $p = 365$ microRNAs for the $n = 60$ cell lines in the NCI-60 study, which are grown under controlled condition by the U.S. National Cancer Institute. The NCI-60 cell lines include experimental units from the breast, central nervous system, colorectal, lung, prostate, ovarian and renal cancers. It also includes leukemia and melanoma cell lines. $\mathbf{X}$ includes microRNA measurements from these 9 different types of cancer tissues, see the above original references for more details. The $\mathbf{Y}$ variables collects the growth inhibitory responses of the $q = 15$ Topoisomerase II inhibitor drugs for the investigated cell lines.

As above, we standardized the data set and obtained a joint correlation matrix by using shrinkage estimation, with an optimal shrinkage intensity of $\lambda_{\text{cor}} = 0.25$. The resulting correlations are depicted in Fig. 5.9. For these data the estimated canonical correlations $\lambda_1, \dots, \lambda_{15}$ range from -0.49 to 0.94, with only 3 of 15 being negative (see Fig. 5.10).

Thus, we observe that in experimental data both positive and negative canonical correlations can be prevalent, and presumably knowledge of the directionality is likely to be a helpful guide in further interpretation. Correspondingly, in figure 5.11 we observe that the whitening approach yields better correlation between pairs of canonical covariates than sparse or classical CCA as implemented in the CCA R package. In the next step we compare the network structure induced by the K matrix in the CCA whitening approach to that of sparse CCA. From figure 5.12, we observe that whitening CCA yields better sparsity which facilitates interpretation.

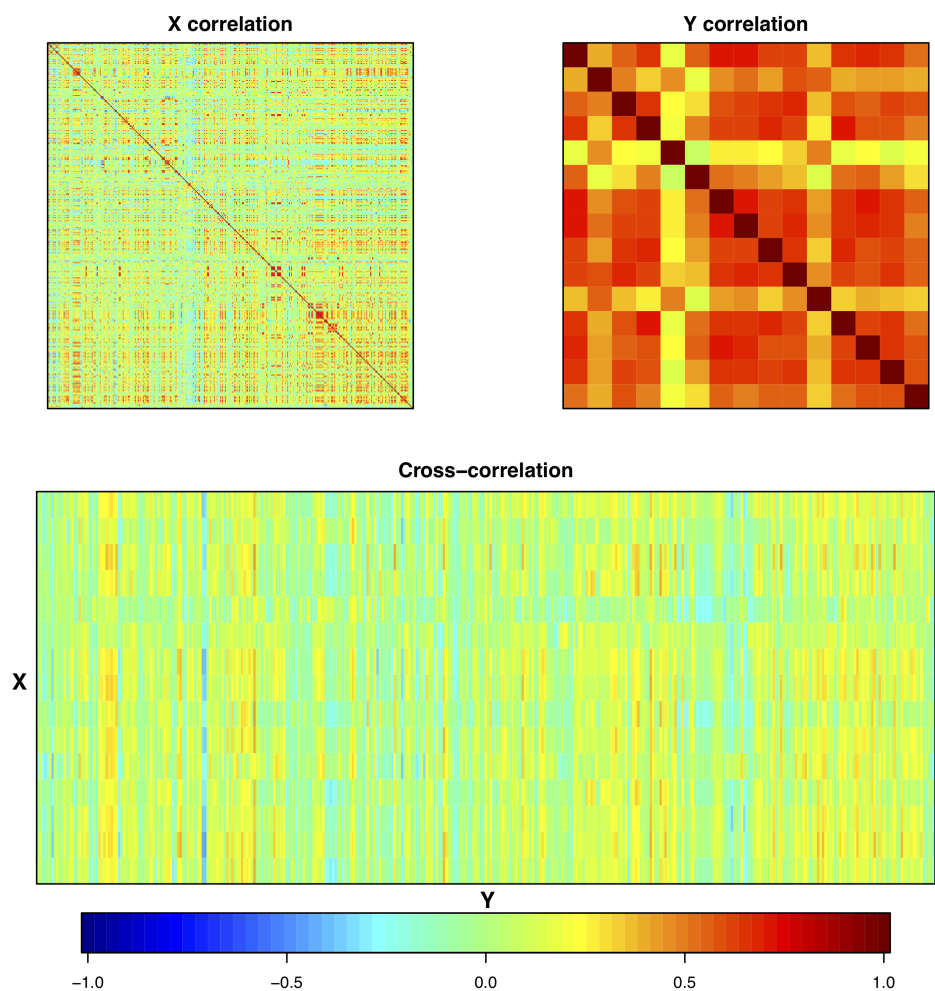


FIGURE 5.9: Visualization of the correlations within and between the two investigated variables (microRNA expression and growth inhibitory response) from the NCI-60 data (Liu et al., 2010). The graphs show strong correlation patterns (≥ 0.5) within $\mathbf{Y}$. The cross-correlation matrix indicates mild correlation patterns (≤ 0.5) between $\mathbf{Y}$ and $\mathbf{X}$ which will be further exploited by CCA.

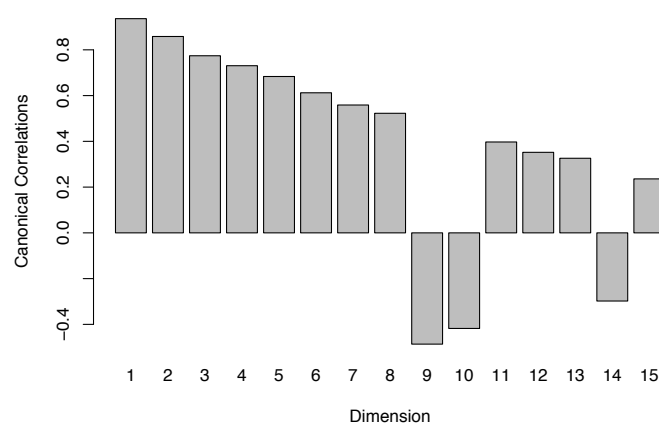


FIGURE 5.10: Plot of the estimated canonical correlations for the NCI-60 data.

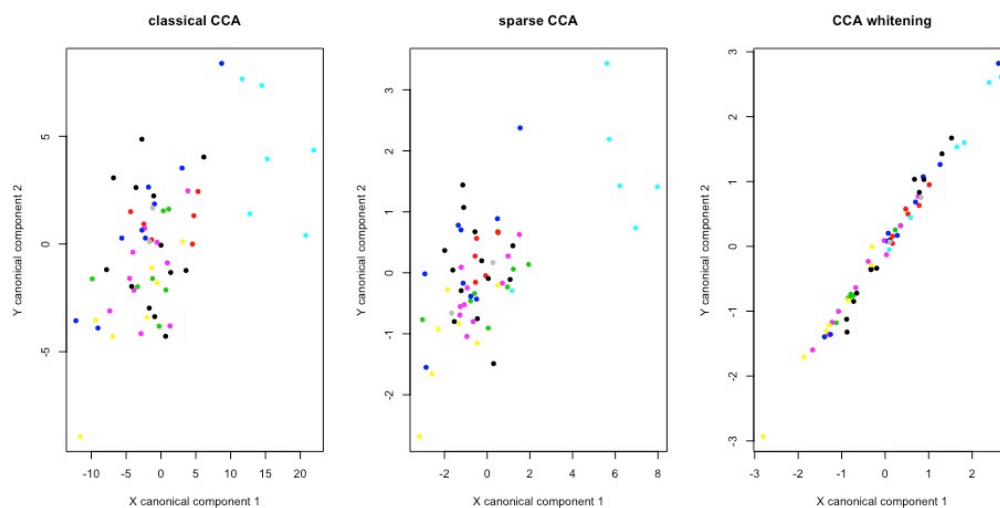


FIGURE 5.11: Scatter plots between corresponding pairs of canonical covariates for the NCI-60 data with different CCA models applied. Samples are colored according to their cancer type.

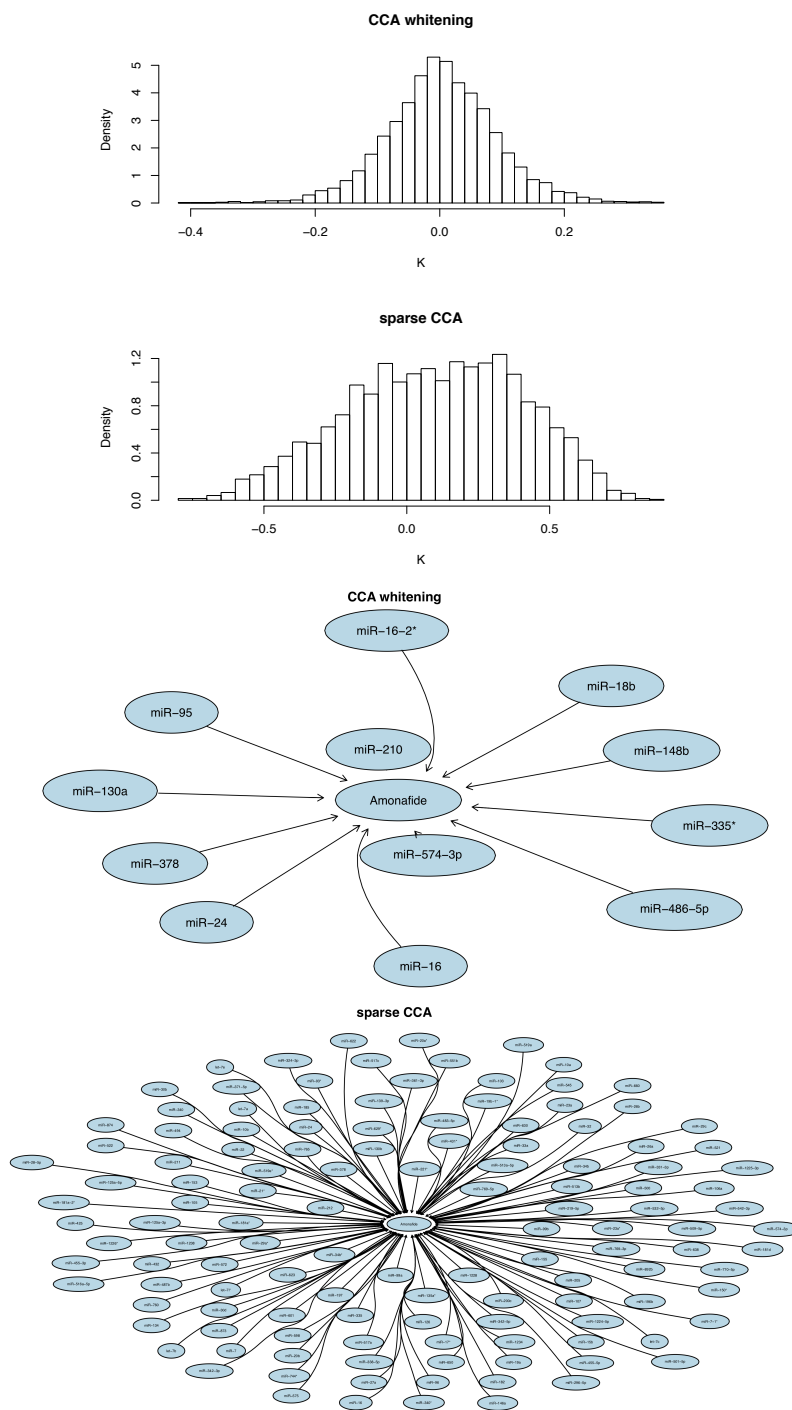


FIGURE 5.12: NCI-60 data. Top: Distribution of the K matrix coefficients for our CCA whitening method and sparse CCA. These are mostly centered around 0 with a higher concentration for our CCA whitening method. Bottom plot shows a sample network centered around drug “Amonafide” induced respectively by CCA whitening (upper plot) and sparse CCA (lower plot). The network shows a sparser structure for our CCA whitening method compared to sparse CCA.

5.4.4 The Cancer Genome Atlas LUSC data

As a further illustrative example we studied genomic data from The Cancer Genome Atlas (TCGA), a public resource that catalogs clinical data and molecular characterizations of many cancer types (Kandoth et al., 2013). We used the TCGA2STAT tool to access the TCGA database from within R (Wan, Allen, and Liu, 2016)

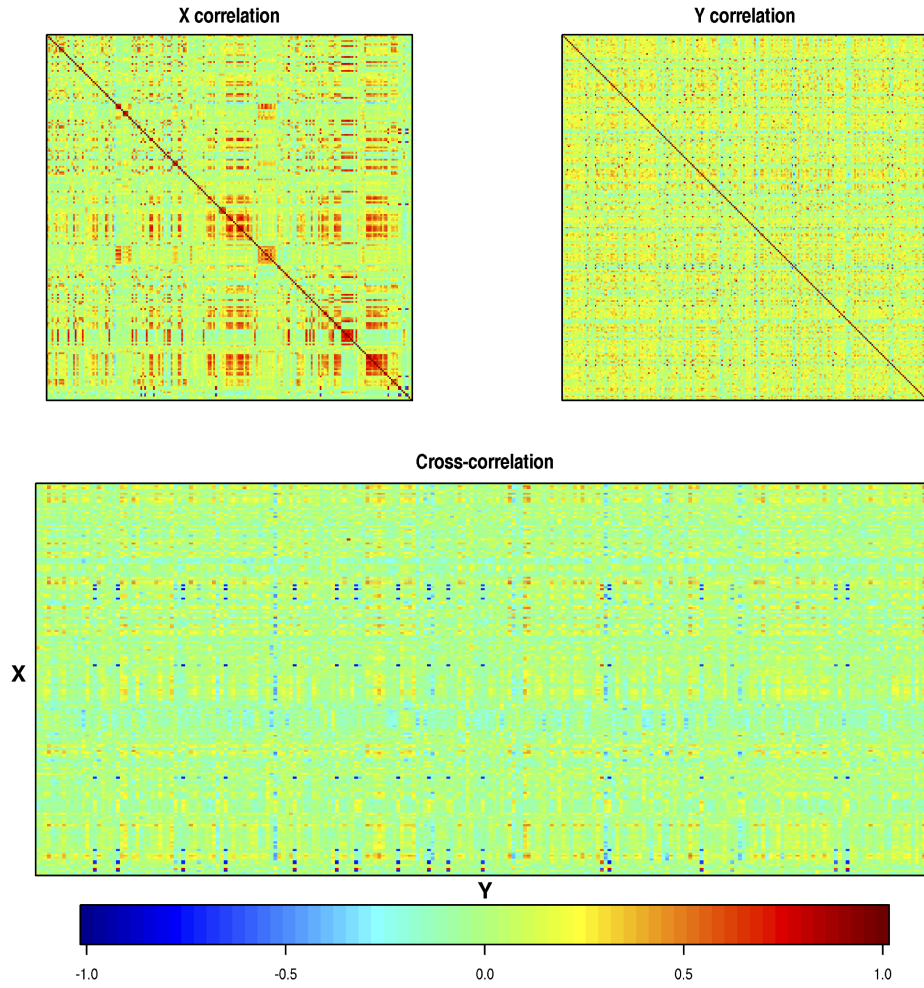


FIGURE 5.13: Visualization of the correlations of the gene expression ($\mathbf{X}$) and methylation levels ($\mathbf{Y}$) for the TCGA LUSC data. The graphs show strong correlation patterns (≥ 0.5) within $\mathbf{Y}$. The cross-correlation matrix indicates mild correlation patterns (≤ 0.5) between $\mathbf{Y}$ and $\mathbf{X}$ which will be further exploited by CCA.

Specifically, we retrieved gene expression (RNASeq2) and methylation data

for lung squamous cell carcinoma (LUSC) which is one of the most common types of lung cancer. After download, calibration and filtering as well as matching the two data types to 130 common patients following the guidelines in Wan, Allen, and Liu (2016) we obtained two data matrices, one ($\mathbf{X}$) measuring gene expression of $p = 206$ genes and one ($\mathbf{Y}$) containing methylation levels corresponding to $q = 234$ probes. Following Wan, Allen, and Liu (2016), the $p = 206$ genes were obtained by selecting the ones that are at the top 1% of variance across patients and the $q = 234$ probes by only including CpG sites. As clinical covariates the sex of each of the 130 patients (97 males, 33 females) was downloaded as well as the vital status (46 events in males, and 11 in females) and cancer end points, i.e. the number of days to last follow-up or the days to death. In addition, since smoking cigarettes is a key risk factor for lung cancer, the number of packs per year, pack-years, smoked was also recorded. Pack-years is defined as the packs of cigarette smoked per day multiplied by the duration of smoking. The number of packs ranged from 7 to 240, so all of the patients for which this information was available were smokers.

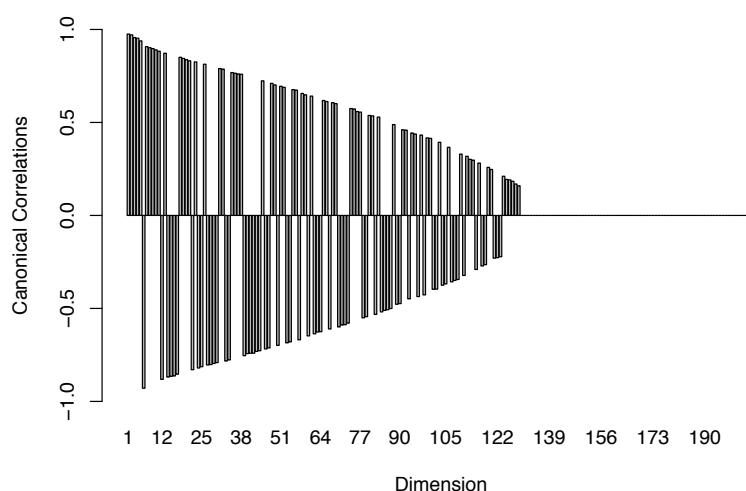


FIGURE 5.14: Plot of the estimated canonical correlations for the TCGA LUSC data. The graphs show some strong correlation patterns (≥ 0.5) within $\mathbf{X}$. The cross-correlation matrix indicates few strong correlations between $\mathbf{Y}$ and $\mathbf{X}$. As most of the correlations are very mild in this example (green), CCA will be mainly used for exploratory analysis.

As above we applied the shrinkage CCA approach to the LUSC data which

resulted in a correlation shrinkage intensity of $\lambda_{\text{cor}} = 0.19$. The resulting estimated correlations are shown in Fig. 5.13. The cross-correlation matrix indicates few strong correlations between $\mathbf{Y}$ and $\mathbf{X}$. As most of the cross-correlations are very mild, CCA will be mainly used for exploratory analysis in this example. Subsequently, we computed canonical directions and associated canonical correlations $\lambda_1, \dots, \lambda_{21}$. The canonical correlations are shown in Fig. 5.14, and range in value between -0.92 and 0.98. Among the top 10 strongest correlated pairs of canonical covariates only one has a negative coefficient. The plot of the squared correlation loadings (Fig. 5.15) for these 10 components already indicates that the data can be sufficiently summarized by a few canonical covariates e.g. 4 to 5 components for $\mathbf{X}$ and 1 to 2 components for $\mathbf{Y}$.

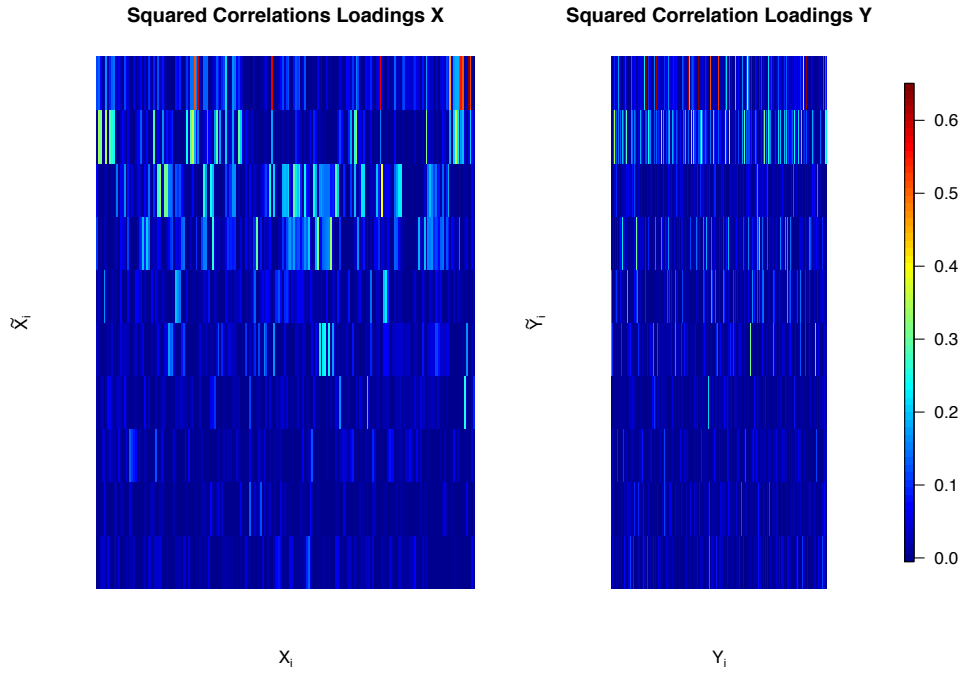


FIGURE 5.15: Squared correlations loadings between the first 10 components of the canonical covariates $\widetilde{\mathbf{X}}^{\text{CCA}}$ and $\widetilde{\mathbf{Y}}^{\text{CCA}}$ and the corresponding observed variables $\mathbf{X}$ and $\mathbf{Y}$ for the TCGA LUSC data. These indicate that around 4 components are needed for $\mathbf{X}$ and around 2 components are needed for $\mathbf{Y}$.

Scatter plots between the first pair of canonical components and between the first two components of $\widetilde{\mathbf{X}}^{\text{CCA}}$ are presented in Fig. 5.16. These plots show that the first canonical component corresponds to the sex of the patients,

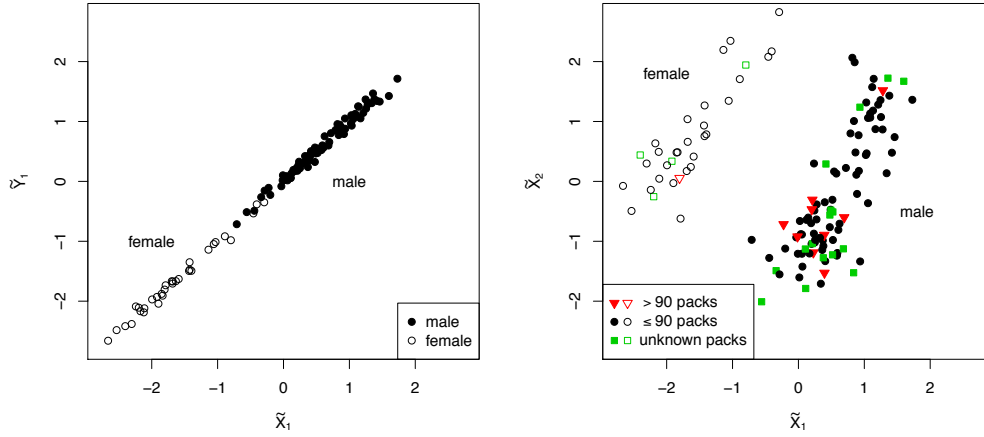


FIGURE 5.16: Scatter plots between first component of $\tilde{\mathbf{X}}^{\text{CCA}}$ and $\tilde{\mathbf{Y}}^{\text{CCA}}$ (left) and within the first two components of $\tilde{\mathbf{X}}^{\text{CCA}}$ (right) for the TCGA LUSC data.

with males and females being clearly separated by underlying patterns in gene expression and methylation. Strong smokers as shown in Fig. 5.16 (right) are over-represented in males compared to females. The survival probabilities computed for both groups show that there is a statistically significant different risk pattern between males and females (Fig. 5.17) which explains the clustering pattern in Fig. 5.16.

5.5 Software

5.5.1 Computational aspects

High-dimensional estimation

Practical application of CCA, in both the classical and probabilistic variants, requires estimation of the joint covariance of $\mathbf{X}$ and $\mathbf{Y}$ from data, as well as the computation of the corresponding underlying whitening matrices $\mathbf{W}_{\mathbf{X}}^{\text{CCA}}$ and $\mathbf{W}_{\mathbf{Y}}^{\text{CCA}}$ (i.e. canonical directions) and canonical correlations λ_i .

In moderate dimensions and large sample size n , i.e. when both p and q are not excessively big and n is larger than both p and q the classic CCA algorithm is applicable and empirical or maximum likelihood estimates may be used. Conversely, if the sample size n is small compared to p and q then there exist numerous effective Bayesian, penalized likelihood and other related regularized

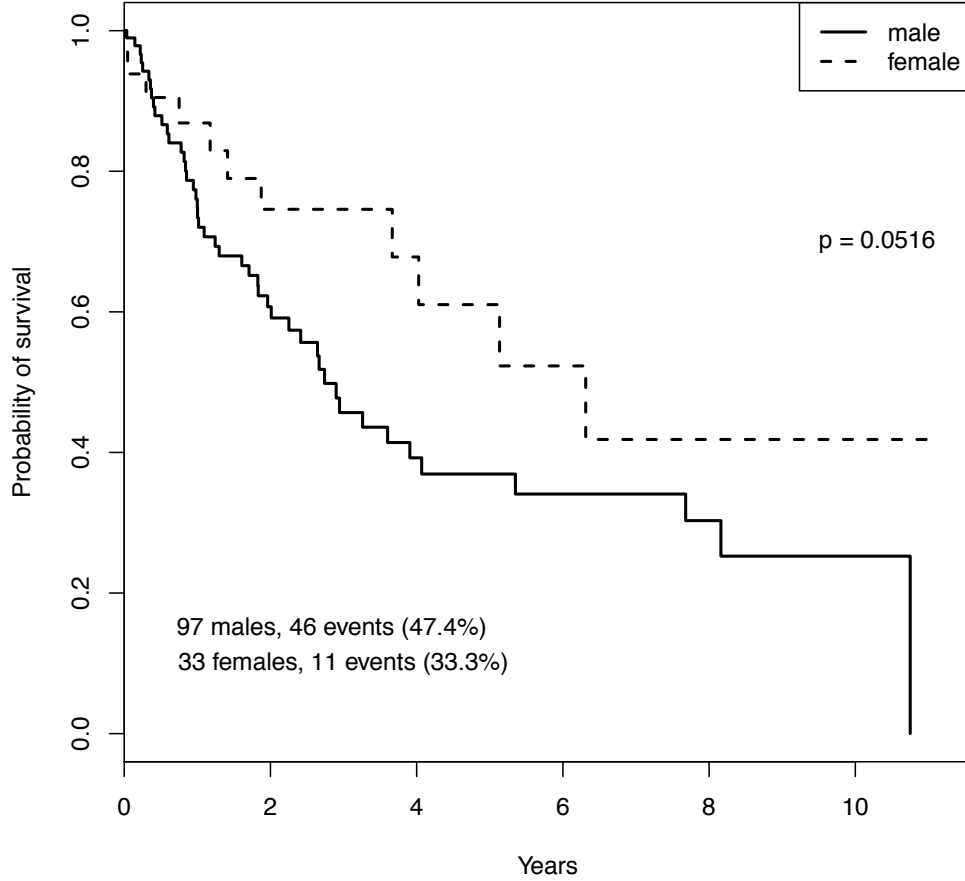


FIGURE 5.17: Plot of the survival probabilities for male and female patients for the TCGA LUSC data.

estimators to obtain *statistically efficient estimates* of the required covariance matrices (e.g. Touloumis, 2015; Hannart and Naveau, 2014; Bickel and Levina, 2008). In our implementation and in the analysis below we use the shrinkage covariance estimation approach developed in Schäfer and Strimmer (2005) and also employed for CCA analysis in Cruz-Cano and Lee (2014). However, in principle any other preferred covariance estimator may be applied.

Algorithmic efficiency

In addition to statistical issues concerning accurate estimation, high dimensionality also poses substantial challenges in *algorithmic* terms, with regard both

to memory requirements as well as to computing time. Specifically, for large values of p and q directly performing the matrix operations necessary for CCA, such as computing the matrix square root or even simple matrix multiplication, will be prohibitive since these procedures typically scale in cubic order of p and q .

In particular, in a CCA analysis this affects i) the computation and estimation of the matrix $\mathbf{K}$ (Eq. 5.9) containing the adjusted cross-correlations, and ii) the calculation of the whitening matrices $\mathbf{W}_X^{\text{CCA}}$ and $\mathbf{W}_Y^{\text{CCA}}$ with the canonical directions $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ from the rotation matrices $\mathbf{Q}_X^{\text{CCA}}$ and $\mathbf{Q}_Y^{\text{CCA}}$ (Eq. 5.5). These computational steps involve multiplication and square-root calculations involving possibly very large matrices of dimension $p \times p$ and $q \times q$.

Fortunately, in the small sample domain with $n \leq p, q$ there exist computational tricks to perform these matrix operations in a very effective and both time- and memory-saving manner that avoids to directly compute and handle the large-scale covariance matrices and their derived quantities (e.g. Hastie and Tibshirani, 2004). Note this requires the use of regularized estimators, e.g. shrinkage or ridge-type estimation.

Here, we capitalize on an algorithm described in more detail in Zuber, Silva, and Strimmer (2012) that allows to compute the matrix product of the inverse matrix square root of the shrinkage estimate of the correlation matrix $\mathbf{R}$ with a matrix $\mathbf{M}$ without the need to store or compute the full estimated correlation matrices. Specifically, writing the correlation estimator in the form

$$\underbrace{\mathbf{R}}_{p \times p} = \lambda(\mathbf{I}_p + \underbrace{\mathbf{U}}_{p \times n} \underbrace{\mathbf{N}}_{n \times n} \mathbf{U}^T) \quad (5.34)$$

allows for algorithmically effective matrix multiplication of

$$\underbrace{\mathbf{R}^{-1/2}}_{p \times p} \underbrace{\mathbf{M}}_{p \times d} = \lambda^{-1/2}(\mathbf{M} - \mathbf{U} \underbrace{(\mathbf{I}_n - (\mathbf{I}_n + \mathbf{N})^{-1/2})}_{n \times n} (\underbrace{\mathbf{U}^T \mathbf{M}}_{n \times d})). \quad (5.35)$$

Note that on the right-hand side of these two equations no matrix of dimension $p \times p$ appears; instead all matrices are of much smaller size. We apply this procedure to Eq. 5.5 to obtain the whitening matrix and the canonical directions and also to Eq. 5.9 to efficiently compute $\mathbf{K}$. The computational savings due to effective matrix operations for $n < p$ and $n < q$ can be substantial, going from $O(p^3)$ and $O(q^3)$ down to $O(n^3)$ in terms of algorithmic complexity. Correspondingly, for example for $p/n = 3$ this implies time savings of factor 27

compared to “naive” direct computation.

5.5.2 Implementation in R

We have implemented our method for high-dimensional CCA allowing for potential negative canonical correlations in the R package `whitening` that is freely available from <https://CRAN.R-project.org/package=whitening>. The functions provided in this package incorporate the computational savings described in Section 5.5.1. The whitening approach to canonical correlation analysis is implemented in the functions `cca` (empirical estimator) and `scca` (shrinkage estimator). The R package also includes example scripts. The whitening package has been used to conduct the simulations and data analysis in this chapter. Further information and R code to reproduce the analysis is available at <http://strimmerlab.org/software/whitening/>.

5.6 Conclusions

CCA is crucially important procedure for integration of multivariate data. Here, we have revisited CCA from the perspective of whitening that allows a better understanding of both classical CCA and its probabilistic variant. In particular, our main contributions in this chapter are:

- first, we show that CCA is procedurally equivalent to a special whitening transformation, that unlike other general whitening procedures, is uniquely defined and without any rotational ambiguity;
- second, we demonstrate the direct connection of CCA with multivariate regression and demonstrate that CCA is effectively a linear model between whitened variables, and that correspondingly canonical correlations are best understood as regression coefficients;
- third, the regression perspective advocates for permitting both positive and negative canonical correlations and we show that this also allows to resolve the sign ambiguity present in the canonical directions;
- forth, we propose an easily interpretable probabilistic generative model for CCA as a two-layer latent variable framework that not only admits canonical correlations of both signs but also allows non-normal latent variables;

- fifth, we provide a computationally effective computer implementation in the “whitening” R package based on high-dimensional shrinkage estimation of the underlying covariance and correlation matrices and show that this approach performs well both for simulated data as well as in application to the analysis of various types of omics data; and
- sixth, we also briefly discuss the relationship of CCA with related latent variable projection methods such as O2-PLS or Orthogonal CCA.
- seventh, we provide a decomposition of an entropy-based measure and link it to CCA to evaluate the relative contributions and to uncover the network structure of two groups of variables

Our algorithm was compared to PCA, classical and sparse CCA and showed better performance (see Appendix B). From a practical point of view, we have investigated the behavior of our CCA method on high dimensional datasets i.e $p \geq n$ where p reached a few hundreds of variables in the NCI-60 data and the LUSC data. However, the algorithm is scalable to higher dimensions and further investigation on higher values of p and q are needed to cover higher dimensional omics datasets. One of the main limitations of our model is that it only explores correlation between omics variables and does not take into account the correlation within omics. Other interesting aspects that can improve CCA also include group effects and multi-level analysis. These models are known as group (sparse) CCA where group effects usually correspond to structural effects such as functional modules or SNPs in the same gene (Virtanen, Klami, and Kaski, 2011; Lin et al., 2013). An additional limitation of CCA whitening is the lack of robust estimators. However, further developments can benefit from the flexibility of this approach to extend it to support robust CCA estimates by introducing robust estimates for the covariance matrix $\mathbf{S}$ (Adrover and Donato, 2015; Wilms and Croux, 2016). For instance, Wilms and Croux (2016) found that robust sparse CCA outperforms sparse CCA on the Nutrimouse data. The paper gives a simple, initial robustification procedure –which can be adopted in our approach– since the normal distribution can be easily replaced by a rescaled t-distribution with a wider tail. Another appealing feature consists in extending the model to analysis of more than two datasets. In fact, interest in integrating multiple data is increasing with the availability of cheaper experimental tools stimulating, for example, the extension of the popular O2-PLS to On-PLS which is applicable for simultaneous analysis of several datasets (Löfstedt and

Trygg, 2011). Under the framework we propose here, the method can be e.g. almost directly upgraded to support more than two datasets by simultaneously whitening the data.

In short, this work provides a unifying perspective on CCA, linking together sphering procedures, multivariate regression and corresponding probabilistic generative models, and also offers a practical tools for practitioners in applied data analysis for high-dimensional CCA which is easily interpretable on the level of the latent variables.

Chapter 6

Investigating infinite functional clustering for metabolite annotation

6.1 Introduction

In chapter 5, we have investigated a probabilistic model for CCA and explored underlying multivariate measures for association between heterogeneous variables. These measures might uncover potential biologically consistent relationships between variables when further investigated. Regardless of whether or not metabolites are annotated, this type of integrative statistical analysis can be performed on the set of heterogeneous omics variables as part of step iv of the analysis pipeline (Figure 2.3). Uncovering key metabolites by means of statistical analysis minimizes experimental costs as it helps targeting metabolites for further analysis. In practice, as long as the metabolites remain unknown their use as a biomarker for example in clinical applications is rather restrained. Metabolite annotation has seen a tremendous boost lately thanks to increased development of automated methods, especially in LC-MS based metabolomics data. Nevertheless, it is still an intriguing task that mainly depends on costly matching of spectral patterns e.g. fragmentation or isotopic patterns in MS and patterns in the chemical shift in NMR. Furthermore, in several cases where datasets are produced in a fee-for-service manner, the raw spectra might not be available and thus characteristic features of MS or NMR peaks cannot be used to identify metabolites. Often in these cases, the available data is a two dimensional matrix where rows and columns correspond respectively to ppm values and integrated peak areas in NMR or single pair (m/z , retention time) values and metabolite signals after peak picking in LC-MS as obtained in the

preprocessing phase explained previously in chapter 2. Note that the only parameter that can be mapped to a database here is the m/z value for LC-MS data.

Some recent attempts have aimed to approach the problem from different perspectives to improve computational metabolite annotations, for instance pathway-based approaches and context-based approaches. In this chapter, we cover metabolite annotation as step v of the analysis pipeline of metabolomics data in the case where scarce information about the raw data is at hand. Specifically, we propose and investigate a Bayesian model for context-based metabolite annotation that allows researchers to derive sensible hypotheses on metabolite identities. The model benefits from three types of accessible data: i) metabolite signals, ii) additional measurements from genomic, transcriptomic or proteomic data, and iii) metabolic pathways.

Context-based methods are methods that use variable similarities to explore related functional modules. In our approach, we first employ infinite clustering models to group variables with similar profiles into the same cluster based on the Dirichlet process. Given the intrinsic complexity of cellular processes, it is more realistic to assume that the number of functional modules is unknown. As a consequence, we will investigate membership clustering models based on the Dirichlet process.

We then extend our model to functional clustering of metabolites and other omics variables at hand. In fact, signals from other omics technologies are easily identifiable since they only have to span four bases in the genomics case and 20 amino acids in the proteomics case, compared to the broad chemical diversity in metabolomics. Hence, by linking metabolotypes to genes, transcripts, proteins and metabolic pathway information, we are able to generate sensible candidates for metabolite identities. Integrating metabolic and other omics variables is done by means of a hierarchical Dirichlet process (HDP) where a high level Dirichlet process (DP) links the different data-specific clustering configurations induced by each data-specific Dirichlet process. Finally, probabilistic assignment of metabolites to components will be performed by combining clustering and pathway information in a biologically consistent way. We show that our model can offer potential improvements on available annotations of a fee-for-service metabolomics data.

To sum up, our main aim in this chapter is to investigate an integrative method for metabolite annotation when raw metabolomics data is not available.

More specifically, our contributions in this chapter consist in

- Investigating the hierarchical Dirichlet process as an integrative clustering method to help metabolite annotation,
- Combining pathway information with integrative clustering by means of 3 scoring functions: homogeneity score, heterogeneity score, log-likelihood based homogeneity score.

6.2 Background

Metabolite annotation for metabolomics data is a demanding and inherently complex task. As discussed in chapter 2, three aspects (amongst others) should be considered when performing metabolite annotation. Firstly, detection bias and sensitivity of the assay. For instance, NMR is characterized by a poor sensitivity (30-100 detectable metabolites) compared to that of MS which allows signal detection of a larger number of metabolites (more than 1000 metabolites). Secondly, many metabolites are still unidentified or ill-identified (unknown unknowns) and so incompleteness of search databases needs to be taken into account by, for instance, introducing uncertainty in assignments. Thirdly, metabolite annotation is prone to another level of ambiguity depending on the standards used and the available data parameters as defined by The Chemical Analysis Working Group of the Metabolomics Standards. In consideration of the intrinsic complexity of the task, it is an achievement for scientists in metabolomics to reach of a coverage of 100-150 identified metabolites.

Bioinformatics tools designed to perform metabolite annotation have been under rapid development recently. However, only few propose probabilistic frameworks. Probabilistic frameworks are actually more advantageous and well suited for the problem as annotation is subject to a certain level of uncertainty. Recently, Daly et al. (2014) and Silva et al. (2014) developed MetAssign and ProbMetab as probabilistic frameworks improving over previous works. MetAssign uses infinite Gaussian mixtures for clustering peaks resulting from the same metabolite. A cluster is defined by a specific retention time and a group of masses resulting from specific adducts and isotopes. Each cluster represents a metabolite and might contain several peaks. A particular adduct-isotope peak is assigned to a cluster if the characteristic mass, retention time and intensity is similar to known peak features in the library provided by the user. This information is simply inferred by assuming normal distribution on each of the

parameters to the one given in the database, allowing for a predefined standard deviation. For example, a Gaussian prior is given to the log of the mass. Another Gaussian prior is given to peak intensities assuming specific adduct and isotopic patterns and a last Gaussian prior is given to the retention time. The authors demonstrate that their software outperforms *mzMatch* and *CAMERA*, yet it needs a library to be defined by the user. The format of these libraries is XML and should include mass and retention time which can be extracted from major databases publicly available (KEGG, METLIN, etc). The user is also required to provide a list of adduct types that will be used in the generation of theoretical peaks. Whilst an exhaustive list can be provided, it is better to stick with few adducts, as spurious adducts can generate more false positives.

In another elegant probabilistic approach, Rogers et al. (2008) propose a probabilistic model based uniquely on mass-to-ratio values and prior pathway information to perform putative metabolite annotations. In their work, metabolic pathways have been introduced by means of a binary membership matrix of metabolites to pathways. This matrix is then used to define the prior probability of a peak to be assigned to a specific chemical formula. The authors show that introducing pathway information reduces the number of mismatches compared to using m/z values alone. As an improvement on Rogers et al. (2008), *ProbMetab* proposes integration of noise error, isotope profile error and retention time error into the likelihood function under the assumption that the error terms are independent. Furthermore, the software offers network based visualization of the annotations based on known reaction networks. However, both of these approaches rely on a library of active transformations and metabolic pathways which is mainly extracted from KEGG via a function provided by the software and requires every observed peak to be matched to a metabolite in the library.

Although pathway databases are far from comprehensive, several automated computational analyses following a pathway-based approach similar to Rogers et al. (2008) have been suggested. Weber and Viant (2010) use transformation mapping based on KEGG reactions to match experimental formula difference of a reaction peak pair to a known empirical formula difference between reactant pairs in a chemical reaction derived from KEGG. For several mass ranges, errors associated with peak differences are estimated to construct a mass error surface. Evaluations of the algorithm shows that the false positive rate is reduced

compared to using a single peak search instead of peak pair differences. Although this approach demonstrate how prior knowledge of metabolic pathways can increase the accuracy of metabolite identification, it is rather limited to two types of reactant pairs : direct reactant pairs that are directly linked by an enzyme or indirect reactant pairs where the substrate needs two enzymatically catalyzed reactions to form the product. This is a central bottleneck as it is not often possible to detect all directly and indirectly related metabolites with an experiment.

In a similar fashion, to Weber and Viant (2010), Gipson et al. (2008) combine instrumental clustering based on metabolite correlations and mapping to KEGG reactant pairs. Peaks exhibiting high correlations are termed interaction pairs. First, these peaks are assigned to the same instrumental cluster if they exhibit the same temporal patterns. Then, putative assignments are obtained by intersecting peak interaction pair clusters with biochemical reactions.

Using correlations as the underlying measure to construct functional metabolic networks is a controversial subject. Although neighboring entities in metabolic pathways have been observed to be significantly correlated, investigations of both synthetic and experimental data suggests that observed correlations might be the result of yet unknown processes given the limited current biochemical knowledge and so do not necessarily explain the underlying pathway structure (Camacho, De La Fuente, and Mendes, 2005; Steuer et al., 2003; Steuer, 2006). We acknowledge that metabolic pathways have only been partially covered and that not all metabolite correlations fully reflect described pathway relationships, however, we assume that correlations which intersect with known biochemical interactions are likely to explain these reactions (Gipson et al., 2008; Krumsiek et al., 2011). As a matter of fact, we believe that actual coverage of biochemical networks intersected with expression of metabolite concentrations is a route to providing useful insights to metabolite identities. Under these assumptions, Krumsiek et al. (2012) combined Gaussian graphical models based on partial correlations (Krumsiek et al., 2011), metabolomics GWAS (Suhre et al., 2011b) and metabolic pathways to generate hypotheses on peak assignments to chemical entities. This integrative context-based model remarkably allowed to generate concrete hypotheses for 106 unknown metabolites out of a total of 225 unknowns. The main feature that distinguishes this model from others is that it can be applied for either NMR or LC-MS data as it does not require the raw data.

In genomics, context-based models augmented with mapping to metabolic pathways, are well established and widely used for functional analysis. For instance, Kuehne et al. (2017) provide a probabilistic model for network segmentation to identify sites of metabolic regulation. The idea is based on the assumption that metabolic reactions that connect the modules of segmented network are the most likely sites of regulation. The authors extracted all KEGG reactant pairs in order to gather all neighborhood information. They used Markov random fields (MRF) to infer network modules based on two objective functions : The observation factor potential which assumes that genes expressing similar patterns are generated from the same normal distribution and the neighborhood factor potential which is based on connectivity statistics from pathway information. GLOBUS is another software developed by Plata et al. (2012) for gene function prediction. The software aims to assign genes to their correct network location by using an objective function based simultaneously on sequence identity to assigned location and context correlation to neighboring genes e.g. via Z-score. In a similar vein, Wu et al. (2002) proposed the parallel application of a variety of functional clustering methods to obtain a database of clusters that are then annotated using standard libraries to perform prediction. Clustering is in fact an additional well established method to group co-expressed genes, infer functional modules and understand gene functions and cellular processes (Eisen et al., 1998).

Our new approach has the advantage that it does not require the raw metabolomics data but relies on valuable information at hand namely metabolite intensities, which also exploits the joint information between samples. In this fashion, Bayesian models offer a natural way to integrate heterogeneous data through hierarchical models. The first level of our model attributes a Dirichlet process to each of the datasets to perform functional clustering on metabolites and genes separately. Clusters with similar profiles are then pooled together via a third Dirichlet process. By integrating metabolomic and genomic clusters in this way, we obtain heterogeneous clusters containing both metabolites and genes exhibiting similar profiles. Genes and metabolites with high mutual context correlation tend to be located closer in metabolic network (Chen and Vitkup, 2006). These entities are then likely to intervene in the same biological function (context) and thus belong to the same pathway. Using this information (e.g. knowing the enzyme, looking for possible pathways) our model can help annotate the non-annotated metabolites of interest. As the

number of clusters (functional modules) is unknown a priori, we use Bayesian non parametric methods. This framework, not only allows for integrative analysis via Bayesian hierarchical models, but also allows functional modules to be inferred adaptively from the data.

We develop our probabilistic metabolite annotation model under the same assumptions as Krumsiek et al. (2012), Gipson et al. (2008), and Kuehne et al. (2017). Namely, we assume that clusters of co-expressed genes intersect with known biochemical interactions e.g metabolic pathways (Chen and Vitkup, 2006). We acknowledge that this assumption does not entirely reflect the true complexity of cellular processes and we discuss its limitations later (see Section 6.6). In addition we do not make assumptions on spatial structure of genes and hence we do not distinguish between *cis*- and *trans*-effects¹. Our approach also relies on the assumption that metabolites and genes are co-regulated so that their expression provide information about their activity level. This means that we also assume that inter-omic variation is simultaneous (See Chapter 2). This is reflected in our model as each of the genomic and metabolomic data is represented by a different Dirichlet process to infer data-specific clusters. A third Dirichlet process simultaneously integrates the data-specific clusters into heterogeneous clusters (see figure 6.1).

In this chapter, we investigate use of infinite integrative clustering models by means of the hierarchical Dirichlet process. The hierarchical Dirichlet process is introduced in Section 6.3.1. Section 6.3.2 discusses different functions to incorporate pathway information. It is followed by Section 6.4 which describes inference of our model where the hierarchical Dirichlet process is combined with pathway information. In Section 6.5 we use both synthetic and real data to investigate whether our model can potentially increase accuracy of metabolite annotation. Finally, in Section 6.6 we discuss model limitations and propose further improvements.

¹cis-effects refers to intramolecular effects i.e elements that act from the same molecule by contrast to trans-effects which refers to intermolecular effects i.e elements that act from different molecules.

6.3 Infinite functional clustering model for metabolite annotation

In this section we introduce our context-based model for metabolite identification. The proposed model consists in two levels depicted respectively in each of the following sections: First a HDP level to integrate different omics data into functional clusters is described in section 6.3.1, then a second level to incorporate pathway information via different scoring functions is described in section 6.3.2. A summary of model parameters is shown in table 6.1 and graphical representation of the model is shown in figure 6.1.

6.3.1 The HDP mixture model

The hierarchical Dirichlet process is defined by Teh et al. (2006) as follows:

$$\begin{aligned}
 \mathbf{w} | \gamma &\sim \text{stick}(\gamma) \\
 \boldsymbol{\pi}_d | \alpha_d, \mathbf{w} &\sim \text{DP}(\alpha_d, \mathbf{w}) \\
 z_{dj} | \boldsymbol{\pi}_d &\sim \boldsymbol{\pi}_d \\
 \boldsymbol{\theta}_k | H_0 &\sim H_0 \\
 \mathbf{x}_{dj} | \boldsymbol{\theta}, z_{dj} &\sim p(\mathbf{x}_{dj} | \boldsymbol{\theta}_{z_{dj}})
 \end{aligned} \tag{6.1}$$

The above equations describe a generative process where mixing weights $\boldsymbol{\pi}_d$ for each dataset are drawn from a data-specific Dirichlet process. These weights are pooled into high level mixing weights $\mathbf{w}$ drawn from a stick breaking distribution. In other words, due to the discreteness of the DP, datasets will share the atom locations (cluster parameters) $\boldsymbol{\theta}_k$ with probability π_{dk} for each data d . The concentration parameters α_d control data-specific cluster weights $\boldsymbol{\pi}_d$ and how these realizations are shrunk towards the base measure $\mathbf{w}$. The high level concentration parameter γ controls the number of atoms (clusters) that are shared amongst datasets (heterogeneous clusters) as well as the non-shared ones (homogeneous clusters). For further details on the above equations and their link to the DP and finite Gaussian mixtures see Appendix C.

There are different scenarios for the concentration coefficients γ and α_d . We here focus on two relevant scenarios to our application:

- γ is small and α_d 's are small: Since the α_d 's are small, we obtain data-specific clusters containing respectively a high proportion of metabolites or genes. These clusters will likely be merged since γ is small. However, large clusters might affect the metabolite identification process since metabolites or genes from different metabolic pathways are likely to be in the same component.
- γ is small and α_d 's are large : Since the α_d 's are large, the data-specific DP will favor a higher number of small clusters than the previous scenario. With γ being small, different clusters will likely be merged resulting in heterogeneous and homogeneous clusters containing a smaller number of variables than the previous scenario. This is expected to provide useful insights to metabolite identities since neighboring entities in metabolic pathways are expected to be clustered together.

We should note here that we are not interested in the scenario where γ is large since it favors a large number of clusters and does not encourage data-specific clusters to be merged.

6.3.2 Modeling pathway information

In this section, we present different types of functions to incorporate pathway information in our model. It is important to note here that in the case we consider where data has been already preprocessed and raw information is not available, some metabolites are identified while others are still unknown. First, we assume that clusters of metabolic profiles that intersect with biochemical knowledge will likely provide sensible candidates for unknown metabolites in these clusters. Hence, we need to define a prior distribution for metabolite annotation conditioned on all other metabolite annotations and cluster assignments that accounts for proximity in metabolic pathways. We denote by h the number of potential candidates for the p metabolites where $h \geq p$. In addition, let $\boldsymbol{\psi} = (\psi_1, \dots, \psi_p)$ where $\psi_j = j'$, with $j' = 1 \dots h, j = 1 \dots p$ is the indicator variable that encodes the assignment of metabolite j to candidate j' . The prior distribution for metabolite annotation conditioned on all other metabolite annotations and cluster assignments is then defined by $p(\psi_j | \boldsymbol{\psi}_{-j}, \mathbf{z})$. We will use $d = 1$ for metabolomics data and $d = 2$ for the additional omics data.

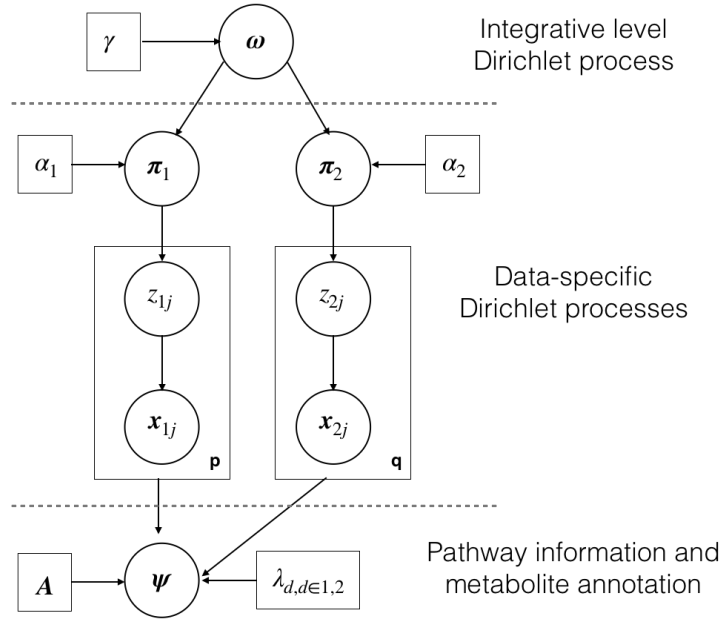


FIGURE 6.1: Graphical representation of infinite functional clustering for metabolite annotation model. The cluster parameters θ_k are deliberately removed from the figure for clarity. The model consists in three components: the integrative level Dirichlet process, the data-specific Dirichlet processes and the metabolite annotation level. Fixed parameters are represented by squares and random variables are represented by circles. Depending on the score function used, metabolite annotations are conditioned on cluster assignments z_{dj} or cluster assignments z_{dj} and data points x_{dj} .

Based on the assumption that neighboring variables in a metabolic pathway should be in the same functional module, we define the conditional distribution of metabolite annotations by the following exponential decay function:

$$p(\psi_j = j' | \psi_{-j}, \mathbf{z}) \propto \exp \left(- \sum_{d=1}^D \lambda_d \frac{\sum_{l=1}^{l=p+q} a_{\psi_{dl}j'} + 1}{\sum_{d=1}^D \sum_{l=1}^{p+q} a_{\psi_{dl}j'} \mathbf{1}_{z_{dl} \neq z_{1j}} + 1} \right) \quad (6.2)$$

Pathway information is introduced via a $(h + q) \times h$ zero-diagonal connectivity matrix $\mathbf{A}$ where $a_{ll} = 0$, $a_{ll'} = 1, l \neq l'$ if entities l and l' are in the same pathway and 0 otherwise. The rows of $\mathbf{A}$ represent metabolomic and genomic variables concatenated whilst the columns of $\mathbf{A}$ represent metabolite

candidates. For notation homogeneity, ψ_{dj} represent metabolomic variables annotations for $d = 1$ and the additional omics variables identities for $d > 1$. λ_d is a factor that controls the strength of metabolite-metabolite neighborhood influence ($d = 1$) and neighborhood influence between metabolites and other omics variables ($d = 2$). The factor $\frac{\sum_{l=1}^{p+q} a_{\psi_{dl}j'} + 1}{\sum_{d=1}^D \sum_{l=1}^{p+q} a_{\psi_{dl}j'} \mathbf{1}_{z_{dl} \neq z_{1j}} + 1}$ reflects the proportion of neighboring omics variables which belong to a different functional module. Hence, candidate j' is likely to be assigned to metabolite j if most of its neighboring omics variables belong to the same functional module as metabolite j . A similar function was adopted in Rogers et al. (2008) on probabilistic assignment of mass peaks to formulas where a connectivity matrix encoding specific chemical transformations is used. We will refer to equation 6.2 as the *homogeneity score*.

The second conditional prior we will investigate is termed as the *heterogeneity score*. It is inspired from the work of Kuehne et al. (2017) to identify sites of regulation in metabolic networks based on correlation structure and neighborhood influence. The conditional prior is as follows:

$$p(\psi_j = j' | \psi_{-j}, \mathbf{z}) \propto \exp \left(- \sum_{d=1}^D \lambda_d \frac{\sum_{l=1}^{p+q} (1 - a_{\psi_{dl}j'}) \mathbf{1}_{z_{dl}=z_{1j}} + 1}{\sum_{d=1}^D \sum_{l=1}^{p+q} \mathbf{1}_{z_{dl}=z_{1j}} + 1} \right) \quad (6.3)$$

where λ_d again controls the strength of neighborhood influence. This function describes the proportion of heterogeneity in the cluster with respect to metabolite j' by considering the proportion of variables that are not connected to metabolite j' but have been assigned to the same cluster as j i.e. their indicator variable is equal to z_{1j} .

The last conditional prior we will investigate is inspired from the work of Argiento et al. (2015) to associate genomic clusters with disease phenotypic traits. The score we use now depends on the data $\mathbf{X}$; we thus refer to equation

6.4 as *log likelihood-based homogeneity score* defined by:

$$\log p(\psi_j | \mathbf{X}, \psi_{-j}, \mathbf{z}) \propto \frac{\sum_{d=1, l=1}^{d=D, l=p+q} a_{\psi_{dl}j'} \log p(\mathbf{x}_{dl} | \boldsymbol{\theta}_{z_{1j}}) \mathbf{1}_{z_{dl} \neq z_{1j}}}{\sum_{d=1, l=1}^{d=D, l=p+q} a_{\psi_{dl}j'} \mathbf{1}_{z_{dl} \neq z_{1j}}} \quad (6.4)$$

where $\boldsymbol{\theta}_{z_{1j}}$ denote the parameters of component z_{1j} . The above equation can be simply interpreted as the standardized log probability of observing the neighboring entities of ψ_j in the same functional module as metabolite j . A summary of all model parameters and a graphical representation of the model are depicted in table 6.1 and figure 6.1 respectively.

Parameters of interest	
z_{dj}	Cluster indicator of observation j in data d
ψ_j	Potential candidate for metabolite j ² .
Other inferred parameters	
$\boldsymbol{\pi}_d$	Vector of data-specific cluster weights
$\boldsymbol{\theta}_k$	Cluster parameters i.e. mean and variance.
Fixed parameters	
γ	Concentration parameter. This parameter controls the number of heterogeneous and homogeneous clusters.
α_d	Data-specific concentration parameters. These parameters control data-specific cluster weights.
λ_d	Neighborhood influence for data d . $\lambda_d = 1$ is for strong neighborhood influence.
$\mathbf{A}$	Binary connectivity matrix encoding pathway information.

TABLE 6.1: Model summary for infinite functional clustering for metabolite annotation.

6.4 Inference

Before deriving posterior inference formulas of cluster assignments, we derive the data log likelihood term $\log p(\mathbf{x}_{dj} | \mathbf{X}_{1:2,-j}, z_{dj} = k, \mathbf{z}_{-dj})$ assuming Gaussian clusters and a Gaussian Inverse Wishart prior on cluster parameters (see Appendix C for details). Cluster assignments can then be inferred using a given sampling scheme for HDP.

Different sampling schemes for HDP are discussed in Teh et al. (2006). We will particularly focus on Markov chain Monte Carlo (MCMC) inference based on the Chinese Restaurant Franchise metaphor as it allows for inference of both data-specific assignments (homogeneous clusters i.e. tables) and assignments to merged clusters (heterogeneous clusters i.e. dishes). In addition, it also allows for pathway information to be easily introduced.

The Chinese Restaurant Franchise (CRF) process is closely related to the HDP. The CRF process was made popular by Teh et al. (2006) as an analog of the Chinese restaurant process. The underlying metaphor now involves a shared menu across restaurants. Restaurants represent different datasets, customers represent observations, tables represent data-specific clusters and dishes of the menu represent merged clusters. Under the CRF metaphor, customers sitting at the same table must have the same dish and different tables can order the same dish. Similarly the Chinese restaurant process, a customer i enters a restaurant d and chooses a table t with probability $\frac{p_{dt}}{\alpha_d + p_d - 1}$ where p_{dt} is the number of customers already sitting at that table and p_d is the total number of customers in restaurant d ³. The customer will then have the dish served at table t . The customer can also choose a new table with probability $\frac{\alpha_d}{\alpha_d + p_d - 1}$. In this case, the customer can either choose i) a dish k that has been already ordered for other tables with probability $\frac{m_k}{\gamma + m}$ where m is the total number of tables and m_k is the number of tables serving dish k , or ii) a new dish with probability $\frac{\gamma}{\gamma + m}$. For full details on the sampling probabilities see Appendix C.

Sampling algorithm

The full inference algorithm is implemented in R software and run on a standard Mac laptop. A typical run time of 3 MCMC chains with 2 datasets of dimensions 24×140 and 24×41 respectively is less than 10 mins. The algorithm for each MCMC iteration is depicted in Algorithm 1.

6.5 Results

In this section we attempt to answer the question: Does genomic enhance our understanding of metabolomics in terms of metabolite annotation? by presenting multiple results relative to varying model parameters, clustering validation and inference of metabolite identities. We assess our model performance on

³ $p_1 = p$ for metabolomics data ($d=1$) and $p_2 = q$ for the additional omics data ($d=2$).

```

for each data  $d$  do
  for each variable  $j$  do
    for each table  $t$  do
      Remove  $j$  and update component mean and variance
      (Appendix C equation C.10);
      if table  $t$  is empty then
        | remove table (and remove dish);
      end
      Sample  $t_{dj}$  (Appendix C equations C.11 and C.12) ;
      Update component mean and variance (Appendix C equation
      C.10) ;
    end
  end
end
for each data  $d$  do
  for each table  $t$  do
    for each dish  $k$  do
      Remove  $t$  and update component mean and variance
      (Appendix C equation C.10) ;
      if dish  $k$  is empty then
        | remove dish;
      end
      Sample  $k_{jt}$  (Appendix C equation C.13) ;
      Update component mean and variance (Appendix C equation
      C.10) ;
    end
  end
end
Sample metabolite annotations from 6.2 or 6.4 or 6.3;

```

Algorithm 1: Collapsed Gibbs sampler for each MCMC iteration

the NCI-60 dataset and the nutrimouse dataset. In practice, all metabolite names were manually compared with KEGG and assigned a KEGG compound identifier whenever possible. Gene names are mapped to their respective gene enzymes, also according to the KEGG database and associated with EC (Enzyme Commission) numbers whenever possible.

Prior to model fitting, we rank transform⁴ and normalize the heterogeneous datasets to obtain comparable data values in all experiments. The MCMC output of infinite clustering algorithms is very rich and contains additional useful information compared to standard clustering methods. As the main output

⁴rank transformation simply consists in mapping the ranked data values to normal quantiles.

would be cluster assignments, it is important to highlight use of other information such as the distribution of the number of clusters and uncertainty around the proposed clustering. A range of methods has been developed to infer a representative clustering from such an output. In our experiments we use the `minbinder` from the `mcclust` (Fritsch, 2012) R package which implements an optimization algorithm that minimizes the posterior expectation of Binders loss function to find a single best clustering of the data based on the posterior similarity matrix (see Fritsch and Ickstadt (2009) for further details). In Bayesian cluster analysis, the posterior similarity matrix is a matrix whose entries contains the posterior probability that corresponding row and column observations are assigned to the same cluster.

6.5.1 NCI-60 metabolomics data

The NCI-60 data was made available by the collaborative NCI60 project from the Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI) for cancer drug screening and have both microarray and metabolomics data downloadable from the DTP website at <https://wiki.nci.nih.gov/display/ncidtpdata/molecular+target+data>⁵. The data comprises 57 cell lines in 9 cancer classes measuring 353 metabolites with 177 technical replicates and 17,987 genes. We followed the same preprocessing steps as in Siddiqui et al. (2018) by filtering the genes with mean below the 10th percentile or with less than 2-fold variation and removing metabolites with more than 80% missing values. Genes and metabolites that were not associated with KEGG identifiers were also removed as they can't be used for pathway mapping. This results in 119 final metabolites and 644 genes.

We run our model with fixed values of $\gamma = 5$, $(\alpha_1, \alpha_2) = (12, 14)$ for 40000 MCMC iterations allowing for 10000 burn-in and compared it with the `mclust` model using the same validity measures as in the previous section (Table 6.2). In figures 6.3 and 6.2 we depict the distribution of the number of clusters following MCMC inference and the posterior similarity matrix showing small cluster sizes. The distribution of the number of clusters is centered around 200 clusters with a cluster mean size of 3-4 omics variables. As a result, only one heterogeneous cluster is obtained compared to 5 heterogeneous clusters for `mclust`. However `mclust` clusters have a mean size of 152.2 making them difficult to investigate.

⁵We here use different data types from the NCI-60 study as the ones used in the previous chapter

model	HDP	mclust
Best number of clusters	200	6
Heterogeneous clusters	1	5
Mean size of heterogeneous clusters	2	152.2
FSI	38917.46	38206.59
XBI	0.0204	0.0684
PC	0.3	0.92

TABLE 6.2: NCI-60 dataset: This table shows different measures of interest comparing our model with `mclust`. Validity measures show that they perform similarly.

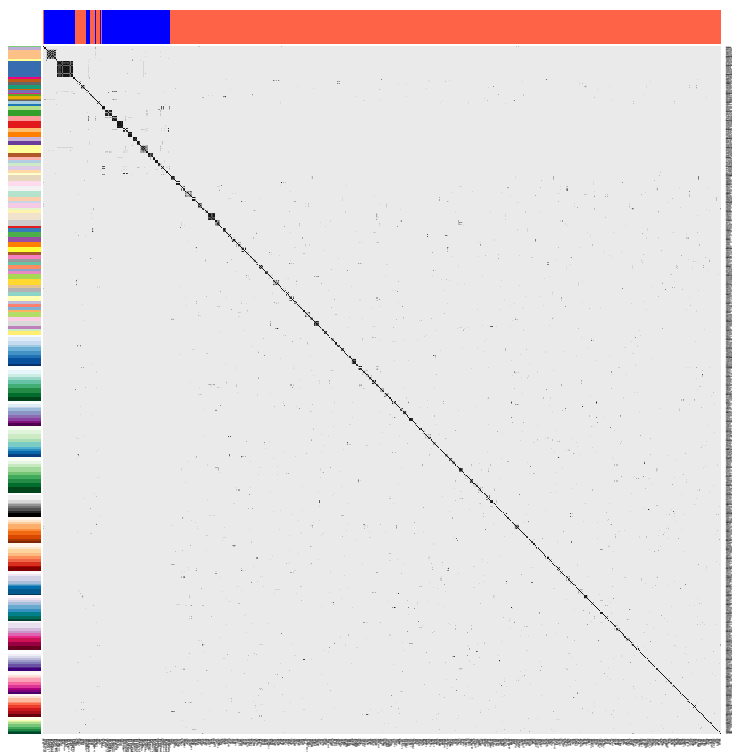


FIGURE 6.2: Similarity matrix for HDP clustering of the NCI-60 data. The matrix shows a diagonal pattern suggesting small size clusters.

In order to have an idea about the ability of our model to propose sensible candidates, we investigate the correlation structure of the data (figure 6.4 left) and the proportion of metabolite-metabolite pathway connections recovered in each cluster (figure 6.4 right). As all metabolite identities in the NCI-60 data are known, we can check how well our assumption is supported by the data i.e. is

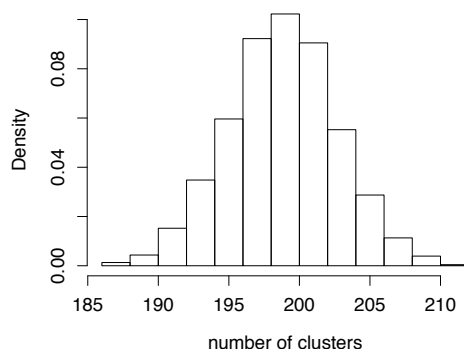


FIGURE 6.3: Histogram of the number of clusters for NCI-60 data following MCMC inference.

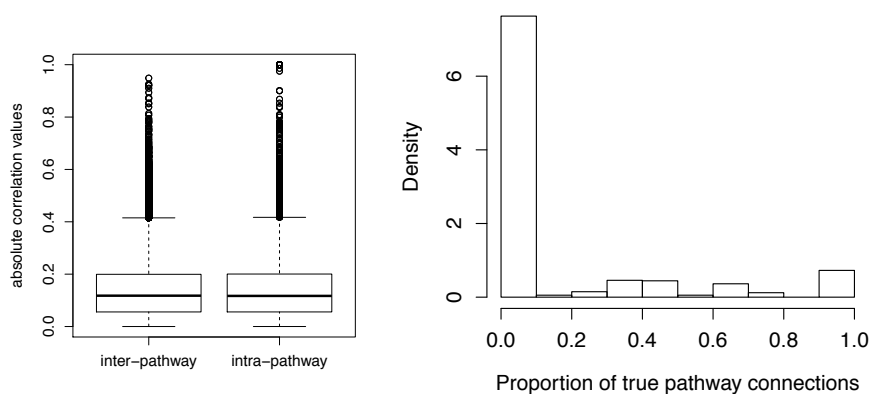


FIGURE 6.4: Left: Boxplot of intra-pathway and inter-pathway correlations for the NCI-60 data. Intra-pathway correlation are correlation between metabolites in the same pathway by contrast to inter-pathway correlations designating metabolites in different pathways. Right: The histogram of the proportion of "true" pathway connections in each cluster shows that 0 is the dominant value suggesting that the clustering is not in accordance with metabolic pathways.

the obtained clustering in accordance with pathway structure? In simple words this comes down to verifying if a pair of metabolites assigned to the same cluster share a pathway connection. Boxplots in figure 6.4 (left) show no difference between inter- and intra-pathway correlations and hence we do not expect our model to perform well on this dataset. Figure 6.4 (right) depicts a histogram of

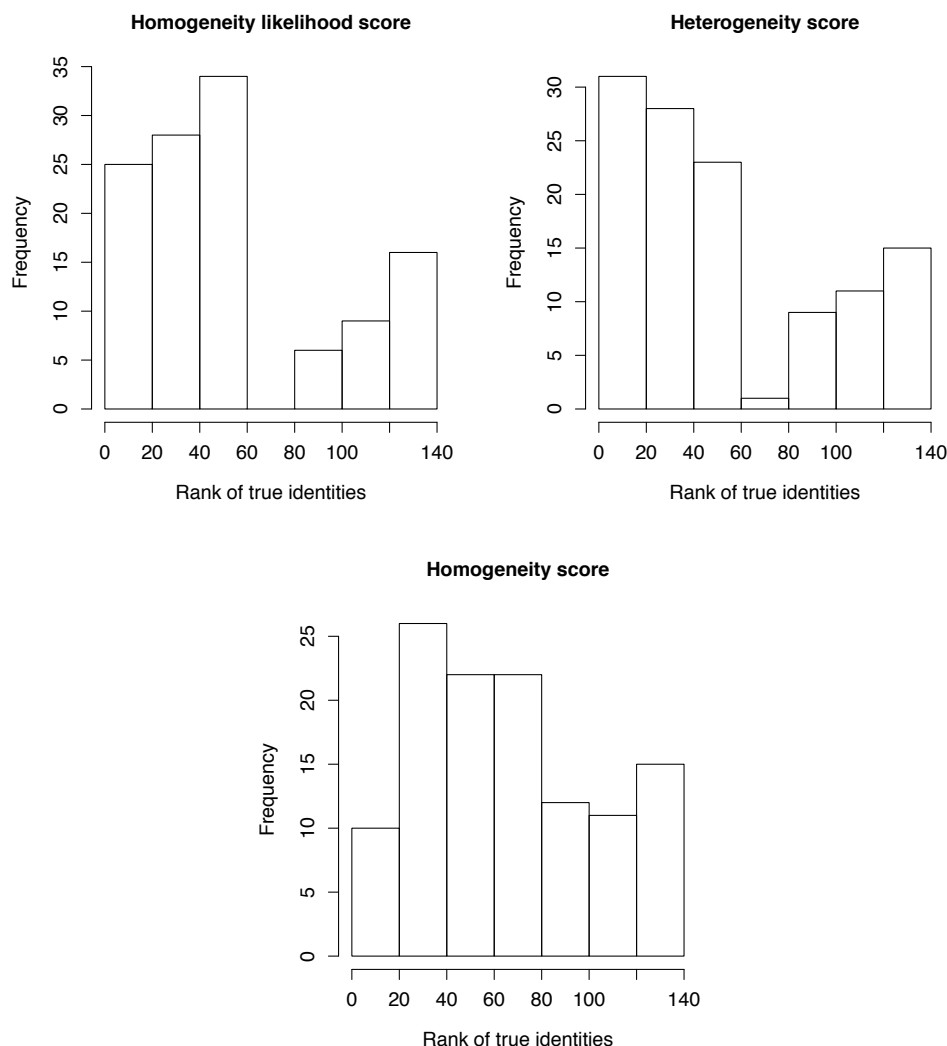


FIGURE 6.5: Mean rank (out of 128) of true recovered metabolite identities for NCI-60 data based on decreasing score functions. The plot shows a poor performance of the model for all score functions.

the proportion of true pathway connections recovered per cluster. In this example, 0 means that no metabolites assigned to the same cluster share a pathway connection. With 0 being the dominant value, we conclude that the NCI-60 data might not be well fitted for context-based models for metabolite identification as clustering is not in accordance with metabolic pathways. In fact, this dataset was designed for drug screening involving many complex biochemical processes that are mostly unknown or that require knowledge of cancer mechanisms from available literature and hence, will poorly perform when mapped to standard databases that partially cover these processes. This is confirmed by examining

figure 6.5 which was generated based on a provided database of 128 potential metabolites. In figure 6.5, we report the position of the “true” candidate (true identity) where all the potential candidates are ranked in decreasing order of the posterior probabilities (or pathway score functions) given in equations 6.2 or 6.4 or 6.3. Figure 6.5 depicts histograms of the ranks of true metabolite identities as given by each of the pathway score functions. The figure shows that true metabolite identities are ranked 10th-60th based on decreasing posterior probability of each candidate metabolite according to the different score functions. This indicates a rather poor performance as true metabolite identities are not amongst the top 10 potential candidates.

6.5.2 Nutrimouse data

In this section we perform different experiments on the nutrimouse data to investigate whether our algorithm gives reasonable clustering solutions. We expect our model to perform better on this dataset as the organism being studied and experiment benefit from a more controlled setting compared to NCI-60.

We run our model for different values of the low level DP parameters (α_1, α_2) and the high level DP parameter γ with 4000 MCMC iterations and 1000 for burn-in then we checked convergence using the `coda` R package. Considering that $\log(p+q) \approx 5$, values of γ were fixed to $\{1, 3, 5, 7, 14\}$ so that both settings of high and low number of expected clusters are covered (see Section 6.3.1). On the other hand we fix $(\alpha_1, \alpha_2) \in \{(3, 3), (6, 9), (7, 10), (8, 12), (12, 14)\}$ so that they are at least equal to $\log(p) \approx \log(q) \approx 3$.

Figure 6.6 shows boxplots of the total number of clusters across different configurations of (α_1, α_2) . First we note that the number of inferred clusters is significantly higher than the number of expected clusters ranging from 30 to 43. Overall, the number for clusters increases when γ increases and the size of each cluster decreases accordingly as depicted in table 6.3. This is explained by the fact that the concentration parameter γ defines the mixing weights of the HDP. Table 6.3 shows the mean size of all clusters i.e. the average number of metabolites per cluster. The higher is γ the more clusters are created and hence the less items are assigned to existing clusters which explains why the mean size of total clusters decreases from 4.1 to 3.5. Table 6.3 also shows the average number of heterogeneous clusters i.e. comprising both genes and metabolites. For $\gamma = 7$ or $\gamma = 14$ data-specific clusters are less favorable to integration than when $\gamma = 3$ or $\gamma = 5$ as a low γ encourages merging

γ	1	3	5	7	14
Average number of heterogeneous clusters	0.4	1.4	1.4	0.8	1
Mean size of total clusters	4.1	3.8	3.8	3.6	3.5

TABLE 6.3: Total clusters refer to the total number of clusters. Average number of heterogeneous clusters refer to the number of clusters containing variables from both data averaged over the different values of α_d . As expected, the clusters sizes are highest when γ is low, however, the average number of heterogeneous clusters is highest when $\gamma \in \{3, 5\}$. The exception here being $\gamma = 1$ and $\gamma \in \{7, 14\}$ since these values do not favor integrative analysis as explained in Section 6.3.1.

data-specific (homogeneous) clusters into heterogeneous clusters. A special case arises when $\gamma = 1$, the average number of heterogeneous clusters is 0.4. In this case, the total number of inferred clusters is small which in turn affects the number of integrated heterogeneous clusters.

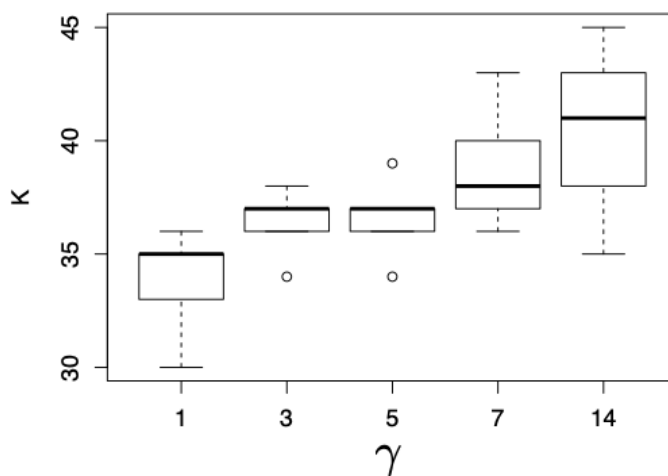


FIGURE 6.6: Boxplot of the number of clusters for different values of γ across different configurations of (α_1, α_2) for the nutrimeuse data.

Clustering validity

To show how well our method compares against competing methods in discovering a *meaningful* data structure, we compare our implementation to `mclust`

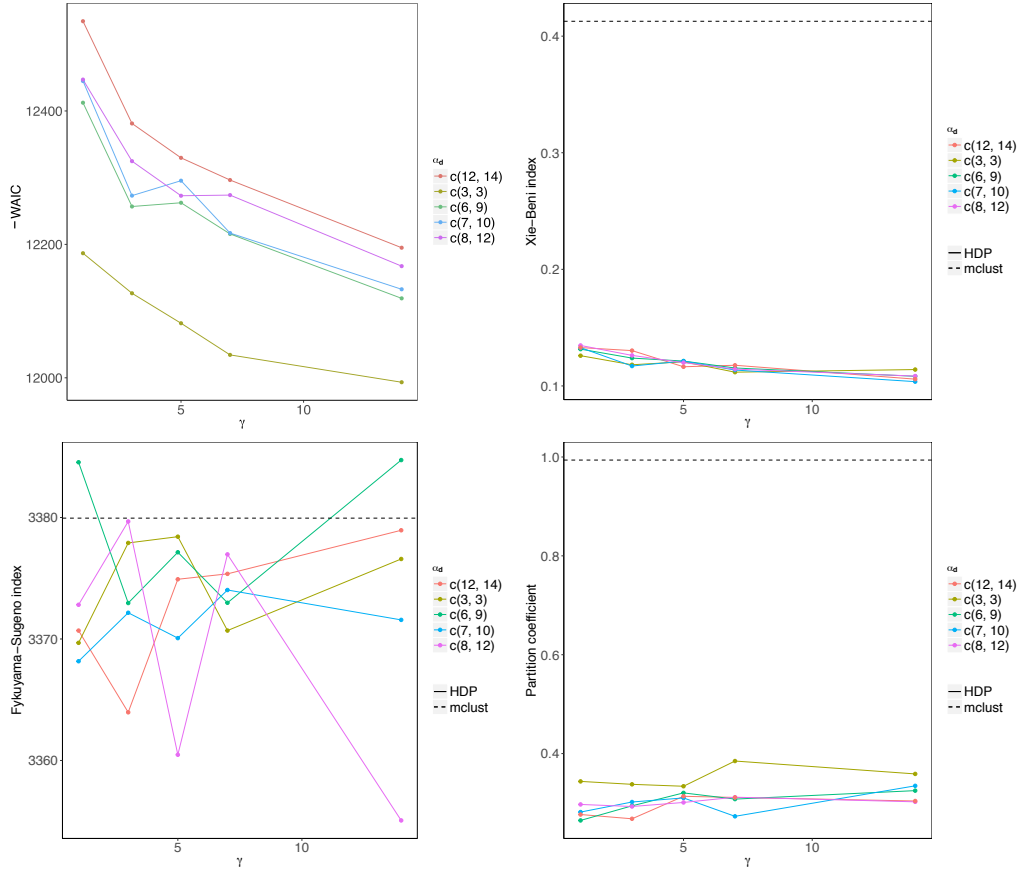


FIGURE 6.7: Figure shows different validity measures used to assess clustering of nutrimouse data and compare it with mclust model. Values obtained for mclust are represented by a dashed line. Top left panel shows -WAIC values on a positive scale. Best performance is obtained for highest values. Top left: -WAIC values are at maximum when γ is at lowest values. Top right and bottom left : Fukuyama-Sugeno index and Xie-Beni index plots show that all HDP models perform similarly but remarkably better than mclust. Bottom right : Partition coefficient values show less uncertainty in the clustering induced by mclust (partition coefficient equal to 1) compared to the HDP model (partition coefficient less than 0.4) . Best model is given by $\gamma = 3$ and $(\alpha_1, \alpha_2) = (12, 14)$.

(Fraley and Raftery, 2006) R package and report three cluster validity measures. The partition coefficient $PC = \sum_{ik} u_{ik}^2$ measures uncertainty of the clustering membership matrix where u_{ik} is the probability of element i to be clustered in k . If PC is close to 1, the clustering is close to hard membership clustering. Values close to 0 are indicative of uncertainty in the clustering. Fukuyama-Sugeno index $FSI = v_c - v_s$ measures the trade-off between cluster compactness $v_c = u_{ik} \sum_k \sum_i \| \mathbf{x}_i - s_k \|^2$ and separability $v_s = \sum_k u_{ik} \| s_k - \bar{s} \|^2$ where

s_k are cluster centroids and $\bar{s}$ is the mean of the cluster centroids (Fukuyama and Sugeno, 1989). Lower values of the Fukuyama-Sugeno index indicate high compactness and high separability thereof, better clustering. The Xie-Beni index (XBI) (Xie and Beni, 1991) considers the ratio between compactness and separation with a different definition of separation to that of FSI i.e. $v_s = \min_{k,k'} \|s_k - s_{k'}'\|$.

Figure 6.7 shows assessment of the obtained clustering according to the benchmarked values of γ and (α_1, α_2) with a comparison to `mclust` where fitting is performed via finite Gaussian mixture modeling fitted using the Expectation-Maximization algorithm. In finite modeling the Bayesian Information Criterion (BIC) was used to select the number of clusters from a range spanning the values 1 to 43. The best model according to BIC contained only 3 clusters with no heterogeneous clusters. In a similar fashion, we evaluate the performance of the HDP model via WAIC estimation (Watanabe, 2010) which is well suited for MCMC sampling. WAIC values are negative with the best model given by the lowest WAIC. Figure 6.7 top left shows -WAIC values on a positive scale and, hence, the best model is given by the highest -WAIC values. -WAIC values show a decreasing trend as a function of γ and, similarly, a decreasing overall trend as a function of (α_1, α_2) with a clear advantage to models where $\gamma = 3$. Plots of the XBI, FSI and PC measures show performance of the best finite `mclust` model (dashed black lines) compared to the different benchmarked HDP models. The FSI and XBI plots (top right and bottom left) display empirical evidence that the HDP model outperforms the `mclust` model in terms of relative compactness and separation of the clusters. On another hand the HDP model is characterized by higher uncertainty than the `mclust` model as depicted by the PC measure i.e assignment probabilities are not close to 1 in the HDP model (bottom right). However, this should be interpreted carefully as uncertainty can lead to poorer as well as richer conclusions.

Assessing pathway incorporation

HDP models with $\gamma = 3$ are fairly similar, subsequently, for the rest of the analysis we choose the model with the relatively lowest WAIC value i.e with $(\alpha_1, \alpha_2) = (12, 14)$. The similarity matrix of the corresponding clustering is depicted in figure 6.8. Firstly, we compare correlation levels of the nutrilmouse data to pathway structure. Figure 6.9 (left) displays slightly higher correlations

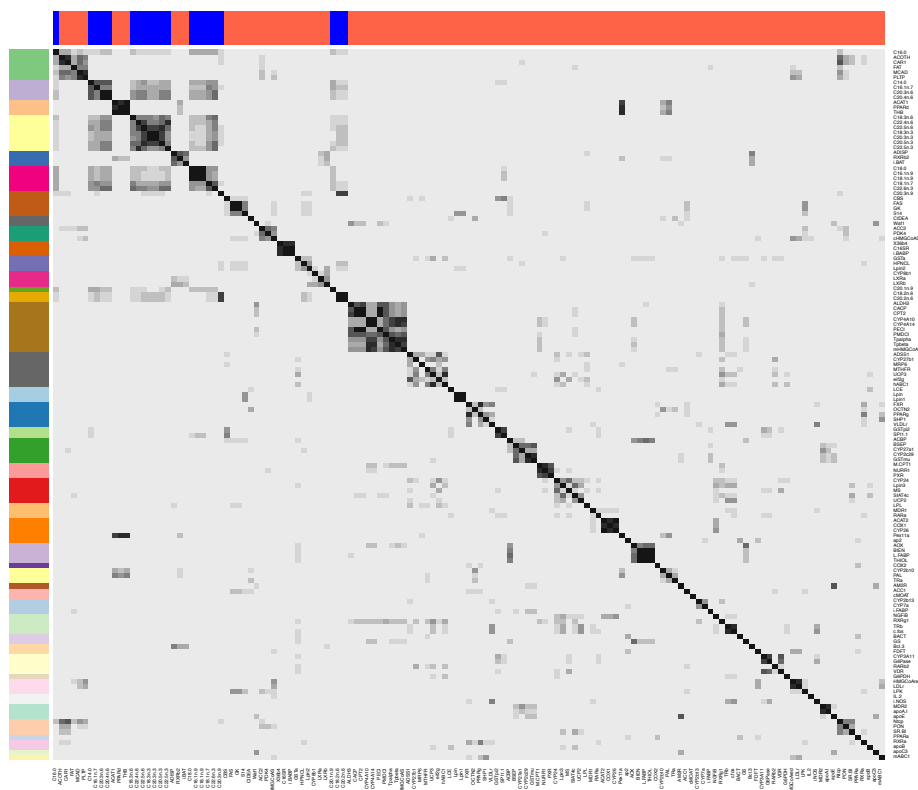


FIGURE 6.8: Similarity matrix for HDP clustering of the nutraceutical data for $\gamma = 3$ and $(\alpha_1, \alpha_2) = (12, 14)$. Darker colors indicate higher posterior probabilities that two omics variables are in the same cluster. Matrix rows and columns are arranged by cluster labels. Horizontal side bar encodes data types while the vertical side bar encodes the different cluster labels.

between metabolites in the same pathway (intra-pathway correlations) compared to correlations between metabolites on different pathways (inter-pathway correlations). In figure 6.9 (right) we plot the histogram of the proportion of metabolites within the same cluster that are also in the same pathway, similarly to what have been carried out for NCI-60 data. For instance, 1 means that all metabolites in a given cluster are in the same pathway. With 1 being the dominant value, we conclude that the clustering is in general agreement with pathway structure.

When fitting was performed, we provided the model with a database of 30 potential candidates for each of 20 metabolites we wish to annotate. In figure 6.10, we report the position of the “true” candidate (true identity) where all the

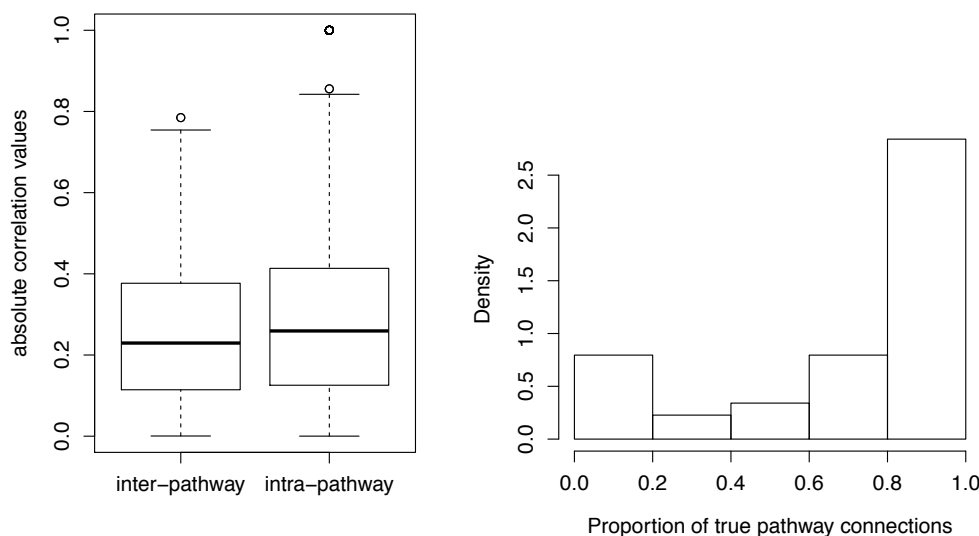


FIGURE 6.9: Left: Boxplot of intra-pathway and inter-pathway correlations for the nutrino dataset. Intra-pathway correlations are correlation between metabolites in the same pathway by contrast to inter-pathway correlations designating metabolites in different pathways. Right: The histogram of the proportion of “true” pathway connections in each cluster shows that 1 is the dominant value suggesting that the clustering is to some extent consistent with metabolic pathways.

potential candidates are ranked in decreasing order of the posterior probabilities (or pathway score functions) given in equations 6.2 or 6.4 or 6.3. Figure 6.10 depicts histograms of the ranks of true metabolite identities as given by each of the pathway score functions. For the homogeneity likelihood score and the heterogeneity score, the true metabolite identities were mostly recovered in the first 5 candidates amongst 30 proposed candidates. For the homogeneity score however, positions of the right identities ranked mostly between 1-10. These results show promising potential when combined with other models to perform metabolite annotation. As an example, metabolite “C22.6n.3” was assigned to “C20.1n.9”⁶ with probability 0.052 then to “C22.6n.3” with probability 0.051. If at most mass information is also available, one can reject “C20.1n.9” as a potential candidate as its mass is around 310.28 Da whilst the mass of “C22.6n.3”

⁶both metabolites are in the same pathway (KEGG identifier: map01040) and same cluster.

is about 328.24 Da. However, it is worthy to note that using only mass information can be restrictive as it is very common to have two different metabolites sharing the same chemical formula and hence the same mass i.e $C_{16}H_{30}O_2$. Consequently, incorporating mass with pathway information will most likely improve results.

As a matter of illustration, the graph in figure 6.11 shows the network structure in “Biosynthesis of unsaturated fatty acids” with metabolites and genes annotated by their KEGG identifiers. This pathway was represented by two clusters in our model: a heterogeneous cluster (left) and a homogeneous cluster (right). In fact, palmitic acid, previously called “C22.6n.3”, was not assigned to the same cluster as other metabolites belonging to the pathway but was merged with a cluster containing the ACOT2 gene (referred by its EC identifier 3.1.2.2), which is also part of the same pathway. We think that this might be due to specific regulation processes induced by the different diets of the mice.

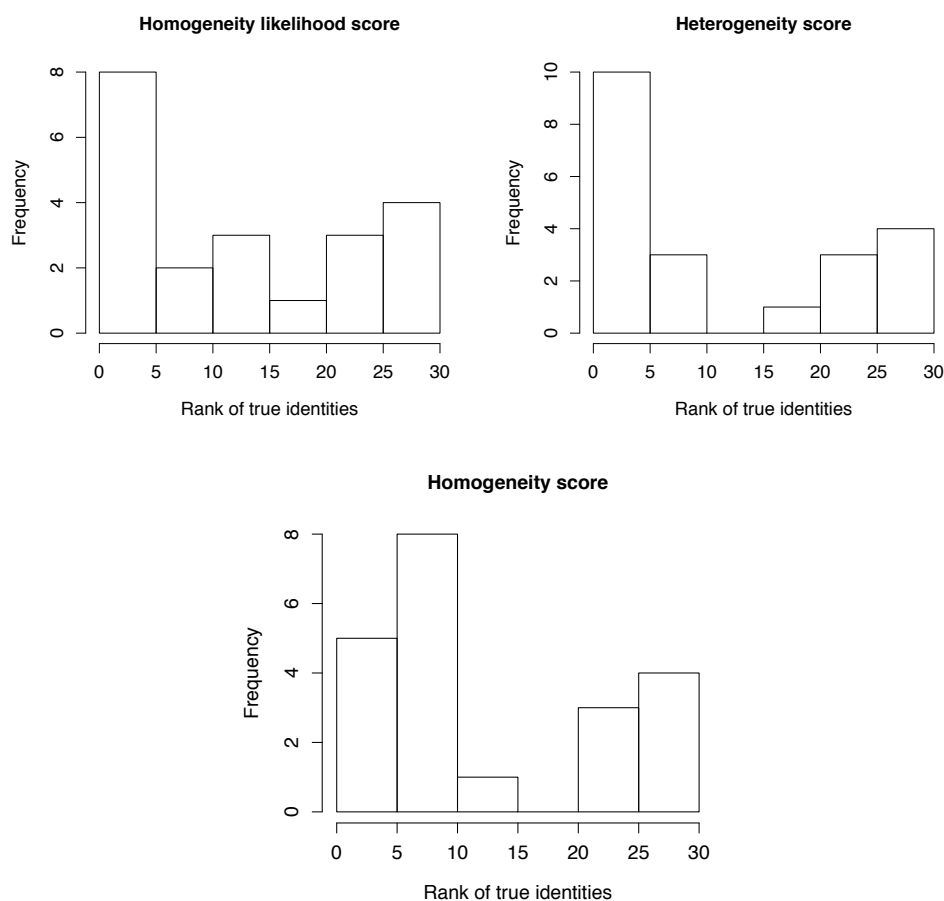


FIGURE 6.10: Mean rank (out of 30) of true recovered metabolite identities for nutrimouse data based on decreasing score functions. All score functions seem to give similar results with the homogeneity score slightly underperforming compared to the two other score functions.

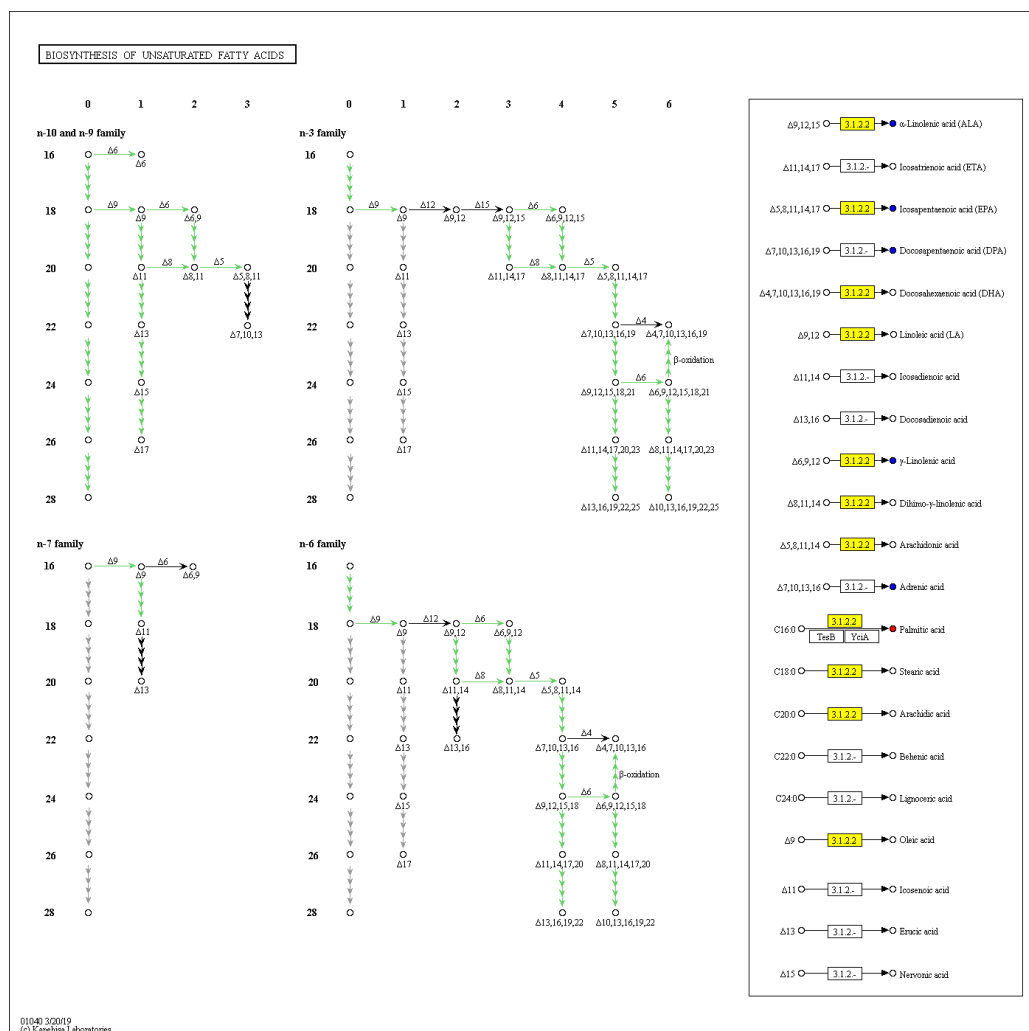


FIGURE 6.11: KEGG pathway: Biosynthesis of unsaturated fatty acids. Highlighted metabolites (red and blue) and genes (yellow) are those who correspond to detected variables in the nutr mouse dataset. The pathway comprises two clusters: A homogeneous cluster represented by metabolites in blue and a heterogeneous cluster represented by the ACOT2 gene in yellow (EC 3.1.2.2) and palmitic acid in red.

6.6 Discussion

The possibilities that metabolomics offers to evaluate the response of a biological system to an internal or external perturbation is extremely interesting. However, in order to derive biological meaning from the data, metabolite identities need to be known to practitioners. In this chapter we investigated a Bayesian model for context-based metabolite annotation that makes use of available biological information to propose sensible candidates for unknown metabolite identities, with an emphasis on clustering models under the Bayesian non parametric (BNP) paradigm. One of the fundamental challenges we consider is the limited access to raw data restricting the use of common computational annotation methods.

The model we have developed is based on two levels: The functional clustering level which consists in integrating metabolomics and genomics variables via the HDP and the pathway modeling level based on metabolite connectivity information. We have tested the potential of the model to generate reasonable hypotheses on metabolite identities using three types of score functions that incorporate both pathway information and cluster assignments. We also investigated the effect of the concentration parameter of the Dirichlet distribution on the number of heterogeneous clusters and metabolite assignments. We finally showed that our method is able to propose sensible candidates for unknown metabolites based on simulations from real data and consequently that genomic data contributes to furthering our knowledge of metabolomics data.

Hierarchical models are relevant models for integrative analysis thanks to their superior flexibility compared to one-level models. One of the main advantages lies in the fact that these models provide more degrees of freedom than one-level models and thus permits, for each data a definition of its own cluster components that might or not be shared with the rest of the datasets. In our approach, we combined hierarchical models with infinite clustering models, resulting in the HDP model in order to offer an unbiased framework for researchers to identify functional modules where both metabolic and genomic variables are important. The set of functional modules obtained is then compared to network information derived from pathway databases via heterogeneity, likelihood-based heterogeneity and homogeneity function scores to generate hypotheses on unknown metabolite identities. To the best of our knowledge, infinite integrative functional clustering has not been investigated for metabolomics data.

Nevertheless, our results need to be carefully appraised as the model depends

on various assumptions. A substantial hypothesis in our research consists in assuming that clusters of genes and metabolites coincide, partially at least, with known metabolic reactions. In our model, cluster assignments are not an output but only a target for further exploration. Our score functions for pathway modeling use the degree of overlap between pathways and clusters to propose candidates for unknown compounds. Conceivably, this assumption is rather biased. Although, the mapping of pathways to clusters dramatically reduces the search space for the unknown compound however it also increases false positives as candidate metabolites in a smaller search space have higher probability to be selected. Unfortunately, overlap between pathways and clusters is only partial and depends on both experimental and statistical factors. Experimental factors are various and include detection bias, partial coverage of the metabolome of the organism of interest, complexity of the organism of interest, type of experiment (targeted or untargeted) and the biological sample. For instance, for cell lines which are simplified models of more complex organisms, pathway mapping can be more efficient than for urine or plasma samples. From a statistical point of view, other clustering models and measures ⁷ can be utilized to improve the degree of overlap (Su et al., 2011; Krumsiek et al., 2012). Integrative analysis of omics layers often raise the question of effect direction and causality. Our model simultaneously integrates genomics and metabolomics via a hierarchical Dirichlet Process in a sequential fashion which inherently assumes a systems hierarchy. However, in reality, dependencies between genes and metabolites are inherently complex involving multiple bi-directional reactions and feedback loops which can not be incorporated into Bayesian models and require advanced graph-based models to improve pathway mapping (Shen-Orr et al., 2002; Xia, Gill, and Hancock, 2015; Abou-Jaoudé, Chaves, and Gouzé, 2014).

The investigation we have conducted in this chapter does not only allow to shed light on different shortcomings of the proposed model but also encourages further directions of research. First, in our model a variable is assigned to only one functional module, however, metabolites, enzymes, or proteins can participate in multiple biological processes and are not restricted to one functional module. In this context, a richer model addressing multiple membership of investigated variables in functional modules or pathways is a natural extension of our model. This can be done, for instance, by replacing the HDP with the hierarchical beta process. The beta process has been widely applied as

⁷The clustering approach we have used does not take into account the spatial structure of genes.

an infinite latent feature model (Paisley and Carin, 2009; Thibaux and Jordan, 2007; Teh and Jordan, 2009; Griffiths and Ghahramani, 2005) and only recently for clustering purposes with a multiplicative likelihood following the paper of Heller and Ghahramani (2007), as in Ross et al. (2014) for example, where authors showed that the beta process is able to uncover latent clusters in chronic obstructive pulmonary disease. Nevertheless, it is particularly challenging to address multiple membership mixture models as they give rise to 2^K possible assignments instead of K possible assignments. We are currently exploring this research route.

Another challenge resides in biochemical reaction modeling. In this model, we investigated linear relationships as an overall trend between metabolic changes and proximal enzymes. Nevertheless, enzymatic reactions are also characterized by specific non linear relationships (Steuer et al., 2003; Camacho, De La Fuente, and Mendes, 2005). These non-linear relationships are commonly modeled via partial differential equations when metabolic fluxes are measured using an experimental setup allowing to do so. Unfortunately, such experiments can not be conducted in cohort studies involving highly complex living organisms e.g. animals, humans or plants. Therefore, ideally one seeks to develop different models that take into account different reaction types. For instance, Tamaddoni-Nezhad, Muggleton, and Bang (2003) have examined different Bayesian models for metabolic reactions where they introduce random variables to indicate whether an enzyme inhibitor is present or absent as the inhibitor that deactivates a specific enzyme, acts also on the production rate of the reaction products. In practice, the influence of enzymes on products and different product-substrate and inhibitor-enzyme relationships are quantified by conditional probability tables. The authors show that their probabilistic model outperforms non probabilistic models. In addition, non linear models for metabolic reaction modeling have been investigated by various authors (see Karlebach and Shamir (2008) for an overview). For example Sriyudthsak, Shiraiishi, and Hirai (2013) use non linear least squares method and show that it efficiently predicts a probable structure of a metabolic reaction network. In accordance with the metabolomics literature, non linear methods are promising to uncover patterns related to pathway reactions, however, determining the types of non-linear patterns that result from sharing the same chemical reaction remains a challenging task.

In practice, we limited the use of pathway databases to the KEGG database

in the results of this chapter. However the algorithm can be easily extendable, as it principally takes in input a connectivity matrix, to more queryable pathway databases to alleviate the shortcomings of the KEGG database. This will enable construction of superpathways that gives further information into the metabolite and gene connectivity in order to enhance the accuracy of the model.

Finally, we should emphasize here that this method should not be considered as a standalone method but should be used to improve metabolite annotation in conjunction with other statistical and experimental identification methods in accordance with the discipline requirements.

Chapter 7

Conclusion

The methods we have developed in this thesis contribute another stone to the construction and diversification of models that can be used to address two major bottlenecks faced by practitioners within metabolomics: The first is the identification of unknown metabolites and the second is the attribution of biological significance to results. The challenges in metabolite identification arise mainly from the large chemical heterogeneity in the metabolome, the inability of current experimental techniques to fully characterize the whole metabolome of a biosystem and the incompleteness of reference databases.

Following the identification step, the end goal of most studies is to extract biological meaning from the data by unraveling the role of these metabolites in the biological system. This is where pathway analysis tools can prove to be very useful. Most of the current pathway analysis approaches are based on evaluating differential expression of a group of functionally related metabolites. These approaches, however, examine each functional module individually and often produce a list of modules with high similarity and redundancy. As a result, it is difficult to derive biological meaning from the resulting hierarchical list of metabolites, especially for modules containing only a few metabolites.

In chapter 2, we reviewed some of these biological and experimental limitations motivating data integration for metabolite identification and pathway analysis in the field. We also addressed features, limitations of metabolomics and also values of other omics than metabolomics. In chapter 3, we covered relevant methods, assumptions and related challenges used for data integration from a statistical perspective. The main statistical challenge that was identified was to propose flexible probabilistic and Bayesian models tailored to the analysis of metabolomics data, in addition to enhancing the interpretability of results. Based on these reviews, this thesis has primarily undertaken research in

three different directions in an attempt to widen the metabolomics bottleneck and promote the incorporation of integrative analysis into the data analysis pipelines. In fact, each of the methods we have developed can be easily used in specific steps towards enriching the analysis compared to using metabolomics data alone. Table 7.1 summarizes features of our methods in each chapter to better appreciate the diversity of the contributions of this thesis and their relationship.

Chapters	Chapter 4	Chapter 5	Chapter 6
Analysis steps (Figure 2.3)	Statistical analysis (step iv) & Enrichment analysis (step vi)	Statistical analysis (step iv)	Metabolite annotation (step v)
Research questions	Associated variables and perturbed pathways	Associated variables or group of variables	Potential of genomic data to help annotate metabolites
Type of analysis (Table 3.1)	Biological analysis	Model-based analysis	Sequential analysis
Framework	Bayesian	Probabilistic	Bayesian nonparametric
Methods	iCARH model	CCA	HDP

TABLE 7.1: Summary of different features of the result chapters in this thesis

Chapter 4 was not only limited to integrative analysis of two different datasets but also covered analysis of short time course data and inference of perturbed pathways by explicitly modeling different sources of variation in the data. Metabolomics time-course experiments provide the opportunity to understand the changes to an organism by observing the evolution of metabolic profiles in response to internal or external stimuli. Along with other omics longitudinal profiling technologies, these techniques have great potential to complement the analysis of complex relations between variations across diverse omics variables and provide unique insights into the underlying biology of the system. However, many statistical methods currently used to analyze short time-series omics data are either prone to overfitting or only model part of the variation in the data. The model we proposed seeks to alleviate these shortcomings by using

weakly informative Bayesian priors to overcome overfitting and explicitly model known sources of variation while allowing for integrative analysis. Additionally, the model we propose is also able to infer metabolic biomarkers (significant metabolites) related to treatment and infer perturbed pathways as a result of treatment.

Since variations in the data are modeled explicitly, our model is able to provide enhanced interpretation of the results. We assessed the performance of our model on synthetic and real datasets and showed that it outperforms existing models for longitudinal metabolomics data analysis. We also showed through simulation that our model is fairly robust against inaccuracies of metabolite assignments as pathway perturbation is jointly inferred based on the expression of functionally related metabolites rather than individual expression. Moreover, our model is advantageously robust against slight variations usually observed in short time series data thanks to the small number of covariance parameters we need to estimate. On real data, we investigated how the number of profiled metabolites can affect the predictive ability of the model. On the same dataset, We have also performed additional sensitivity analyses mainly relating to the number of time points and how missing values are handled. Keeping in mind the need for probabilistic software, the iCARH model is augmented by a user-friendly R package called `iCARH`. The current implementation of iCARH model is limited to shortest pathway based distance matrix which is subject to further investigation. Although this distance measure have proven to be robust, predictive ability could be improved by using more accurate representations of the pathway matrices. However, one should keep in mind that pathways are a subjective segmentation of the true underlying network implying that predictive ability under this framework will be always limited to available knowledge to the advantage of providing biologically consistent meaning to the data. Different caveats related to database based tools should also be taken into account; namely the reliability, incompleteness and ambiguity of these databases. It is also important to critically appraise the assumption of linearity as some chemical reactions do not obey a linear relationship.

In chapter 5 we turned our attention to specifically to modeling dependence relationships between heterogeneous variables. This issue is in fact of critical importance in integrative analysis and at the core of multiple multi-omics

studies. Our work in chapter 5 contributes to the CCA literature by revisiting it from a whitening perspective and proposing a two-layer latent variable probabilistic generative model. In this fashion, our CCA variant can be easily linked to PLS-derived methods and latent variable multivariate regression which allows for easy derivation of negative canonical correlations. In addition, the latent variables underlying the CCA model construction benefit remarkably from straightforward interpretation. In fact, we demonstrated that CCA can be obtained by mixing non correlated data-specific latent variables and common latent variables shared across heterogeneous data in a fairly similar manner to O2-PLS, a well established integrative method in the metabolomics literature. We show that our approach also allows for straightforward derivation of dependency measures between two variables based on an entropy criterion offering the advantage that we can use network representation strategies to model the data.

By means of some simplifying derivations and regularization procedures, we implemented our CCA variant in the R package `whitening` using a highly scalable algorithm well suitable for high-dimensional data as part of the statistical analysis step that can be performed before or after the metabolite annotation step. Lastly, we tested our approach by means of simulated data to judge its ability to recover negative and positive canonical correlations. We also applied it to two omics data sets illustrating the integration of gene expression data, metabolomics data and methylation levels. We compared our model to the widely used sparse CCA models and performed quantitative assessment of canonical correlations to investigate the ability of our model to recover these correlations. We proved that our model outperforms sparse CCA in recovering negative correlations. Overall, we proved that our whitening approach to CCA provides a unifying perspective on CCA, linking together multivariate regression and corresponding probabilistic generative models. We also discussed limitations of our method with respect to robust estimators since we assumed light-tailed generative distributions i.e. Gaussian distribution, and extension to integration of more than 2 datasets. An additional limitations of CCA models is that it only explores correlation between omics variables and does not take into account the correlation within omics such as in group CCA models (Virtanen, Klami, and Kaski, 2011).

Finally, in Chapter 6 we addressed step (v) i.e. the metabolite annotation step by exploiting pathway information. A variety of repositories are available

and are regularly updated to meet the needs of metabolomics scientists. The continuous evolution of both spectral and pathway databases is certainly a valuable step forward to facilitating metabolite identification. Specifically, pathway databases provide a natural and efficient way to integrate metabolomics and other types of genomics data to help in metabolite identification as they incorporate mapping jointly to metabolites, enzymes and genes involved in metabolic pathways. In a similar fashion to Krumsiek et al. (2011), we addressed the problem where the raw metabolomics data is not provided. The method we proposed makes use of the information at hand in the context of data integration, namely, other types of measured data and prior biological knowledge. Specifically, we investigated the incorporation of additional omics data and pathway information into a hierarchical Bayesian nonparametric clustering model and showed that our method can be applied as an auxiliary tool to established methods to provide reasonable hypotheses on metabolite identities. We should pinpoint again that this method should not be used in isolation but only as an auxiliary method to conventional annotation tools. Our model simultaneously integrates genomics and metabolomics via a hierarchical Dirichlet Process in a sequential fashion which inherently assumes a systems hierarchy that is not true in real life due to feedback loops. Our approach is not specifically designed for *cis*-effects but might be more suitable to detect these effects as they affect the same DNA strand as the genes they control, and hence might respond well to clustering methods (Frith, Li, and Weng, 2003; Frith, Hansen, and Weng, 2001). We also discussed that experimental factors should be critically taken into account for this method especially detection bias, coverage of the metabolome of the organism of interest and type of experimental platform.

This approach obviously promotes several avenues of further research involving scalability, computational cost and incorporation of other types of information about raw data whenever available. A potential research direction arises from the fact that variables are limited to one functional module whereas omics entities are regulated within multiple biological processes and hence should not be assigned to only one component. Such methods were recently explored in the Bayesian nonparametric literature by means of the Beta process and proven to provide consistent functional modules and to concurrently unravel latent disease subtypes. Henceforth, we are interested in exploiting the potential of the beta process in this context.

An additional limitation is common to pathway modeling procedures in Chapters 4 and 6. As a matter of fact, both models could benefit from the integration of different pathway structures as defined by various pathway databases compared to the use of one database which is far from comprehensive. This allows assembly of superpathways covering a wider range of biological knowledge. Furthermore, in all three chapters we have developed models that particularly capture linear relationships between heterogeneous omics variables. Nevertheless, non linear relationships are inevitably part of the complex interrelationships between omics variables (Camacho, De La Fuente, and Mendes, 2005; Daemen et al., 2009) and should be considered in integrative models. As an example, CCA can be generalized to kernel CCA (Lai and Fyfe, 2000), the integrative regression model in Chapter 4 can be upgraded to support generalized regression models as in Sriyudthsak, Shiraishi, and Hirai (2013) and the clustering model in Chapter 6 can be improved by using non linear similarity measures (Daub et al., 2004, e.g.).

To sum up, metabolomics is gaining importance in the field of computational medicine with continuing progress in both experimental technologies and computational methods. Growing biochemical knowledge is challenging established data modeling techniques and leading to the exploration of novel approaches that are revolutionizing our understanding of the metabolome and its interconnection with other omics. The methods we have developed in this thesis add to the construction of such knowledge by enhancing interpretability and guiding researchers to pursue promising directions of research. Research on integrating multi-omics data across systems biology is still evolving and is undeniably crucial to enabling discoveries in metabolomics within a biologically meaningful context.

Appendix A

Appendix: iCARH model

A.1 Inferring missing values in iCARH

Since iCARH is a Bayesian model, it handles missing values by directly inferring them from the model. This means that missing values in the metabolomics data are inferred from the following statement:

$$\mathbf{x}_{it}^{missing} | \boldsymbol{\mu}_{it}, \mathbf{C}, \sigma \sim N(\boldsymbol{\mu}_{it}, (\mathbf{I}_p - \mathbf{C}(\phi))^{-1} \sigma^2) \quad (\text{A.1})$$

This takes into account metabolites dependencies via $\mathbf{C}$ and all other covariates via $\boldsymbol{\mu}$. Missing values in the metformin data mainly result from sacrificing half of the animals in the controls and cases groups after the 7th measurement (i.e. 7th timepoint).

Figure [A.1](#) provides an example on missing values inference in iCARH. The figure mainly shows that iCARH provides sensible estimates of posterior means of inferred missing values for acylcarnitine. Each subgraph represents time series of a biological sample. Observed values are encoded in blue dots and posterior means in green. Posterior distributions for missing values are encoded in orange violin plots and span a wide range.

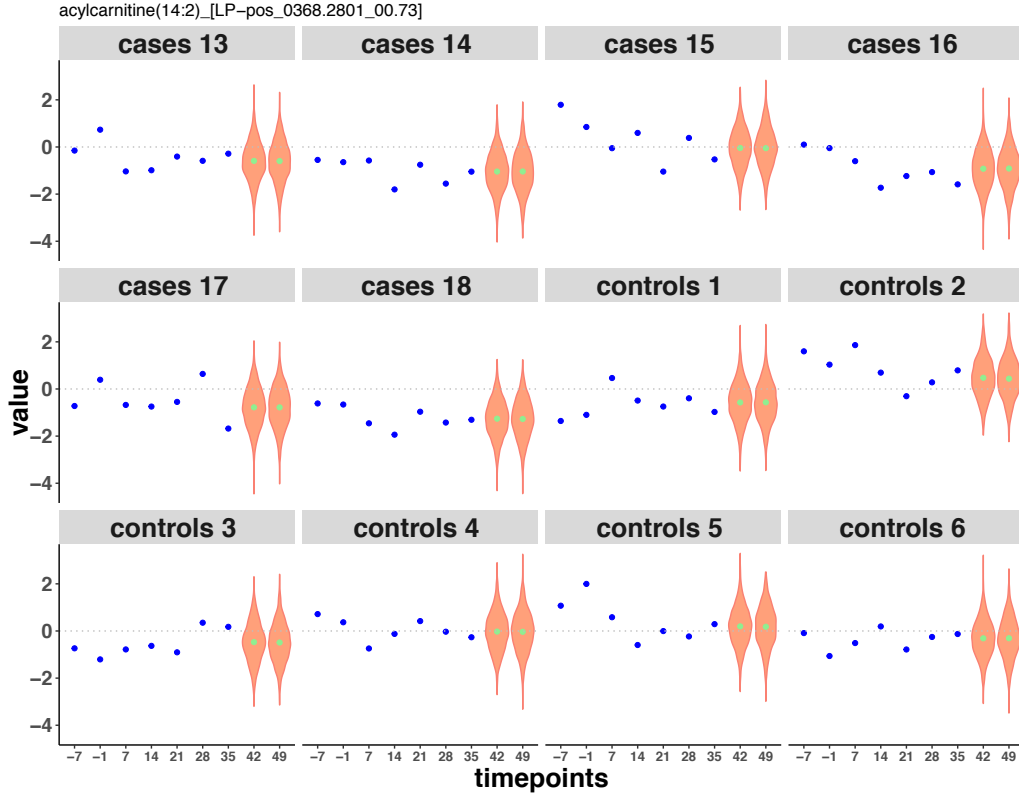


FIGURE A.1: Metformin data: Posterior distributions and means of missing values in acylcarnitine for the last two time-points. Posterior means in green seem to provide reasonable estimates of missing values. Orange violin plots show the posterior distributions and span a wide range.

A.2 Sensitivity analysis

Sensitivity analysis on effect of the number of observed time points has been performed for iCARH via time series cross validation of the Metformin data. Each time a number of consecutive time points is used to predict the following one and the prediction error is computed. Figure A.2 shows boxplots of the prediction error starting from 4 observed time points progressively increasing the number of observed time points to 8. The prediction behavior of iCARH is similar for all time points and shows an error centered around 0. It is important to note that the prediction error is slightly reduced from 4 to 6 observed time points, predicting time points 5 to 7 respectively, as the latter are in the treatment period. The prediction error gets slightly higher for 7 and 8 observed time points, predicting time points 8 and 9 respectively, as the latter are in the recovery period.

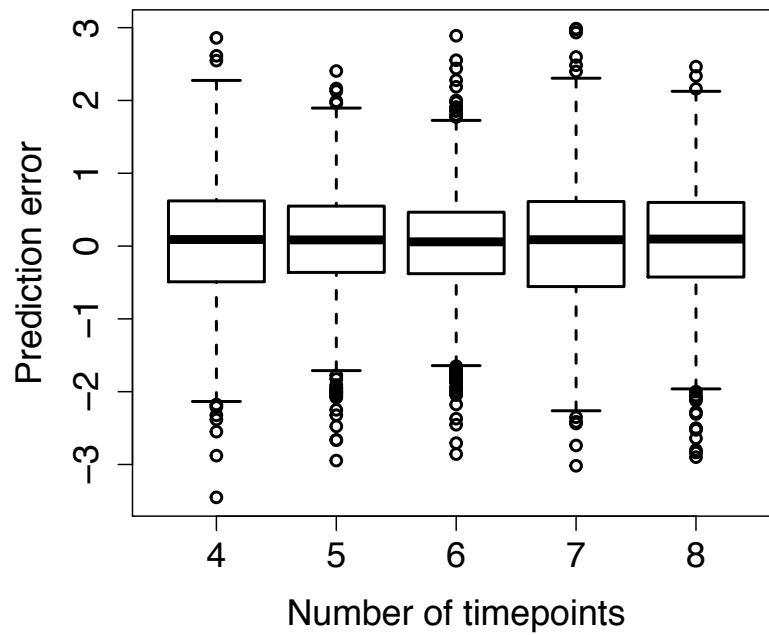


FIGURE A.2: Metformin data: Boxplots of prediction error for 4 observed time points to 8 observed time points. iCARH performs similarly in predicting future timepoints with a prediction error always centered around 0.

Appendix B

Appendix: CCA whitening approach

B.1 Induction of correlation by a shared latent variable

Here we show the workings behind Eq. 5.19. Assume there are three uncorrelated random variables Z_1 , Z_2 , and Z_3 with mean 0 and variance 1. We construct X_1 as a mixture of Z_1 and Z_3 according to $X_1 = \sqrt{1-\alpha}Z_1 + \sqrt{\alpha}Z_3$ where $\alpha \in [0, 1]$, and, correspondingly, X_2 as a mixture of Z_2 and Z_3 via $X_2 = \sqrt{1-\alpha}Z_2 + \sqrt{\alpha}Z_3$. If $\alpha = 0$ then $X_1 = Z_1$ and $X_2 = Z_2$, and if $\alpha = 1$ then $X_1 = X_2 = Z_3$.

By design, the new variables have mean zero ($E(X_1) = E(X_2) = 0$) and unit variance ($\text{var}(X_1) = \text{var}(X_2) = 1$). Crucially, the weight α of the latent variable Z_3 common to both mixtures induces a correlation between X_1 and X_2 . The covariance between X_1 and X_2 is $\text{cov}(X_1, X_2) = \text{cov}(\sqrt{\alpha}Z_3, \sqrt{\alpha}Z_3) = \alpha$, and since X_1 and X_2 have variance 1 we have $\text{cor}(X_1, X_2) = \alpha$.

In Eq. 5.19 this is further extended to allow a signed α and hence negative correlations.

B.2 Synthetic data : Assessment of estimated canonical correlations

In this section we show the results from simulations performed in Section 5.4.1 for the synthetic data. We remind the reader that the canonical correlations were set to have alternating positive and negative signs $\lambda_1 = \lambda_3 = \lambda_5 = \lambda_7 = \lambda_9 = \lambda$ and $\lambda_2 = \lambda_4 = \lambda_6 = \lambda_8 = \lambda_{10} = -\lambda$ with varying strength $\lambda \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. A similar setup was used in Cruz-Cano and Lee (2014). The dimensions were fixed at $p = 60$ and $q = 10$ and the sample size was $n \in \{20, 30, 50, 100, 200, 500\}$ so that both the small and large sample regime was covered. For each combination of n and λ the simulations were repeated 500 times, and our algorithm using shrinkage estimation of the underlying covariance matrices was employed to each of the 500 data sets to fit the CCA model. The resulting estimated canonical correlations were then compared with the corresponding true canonical correlations.

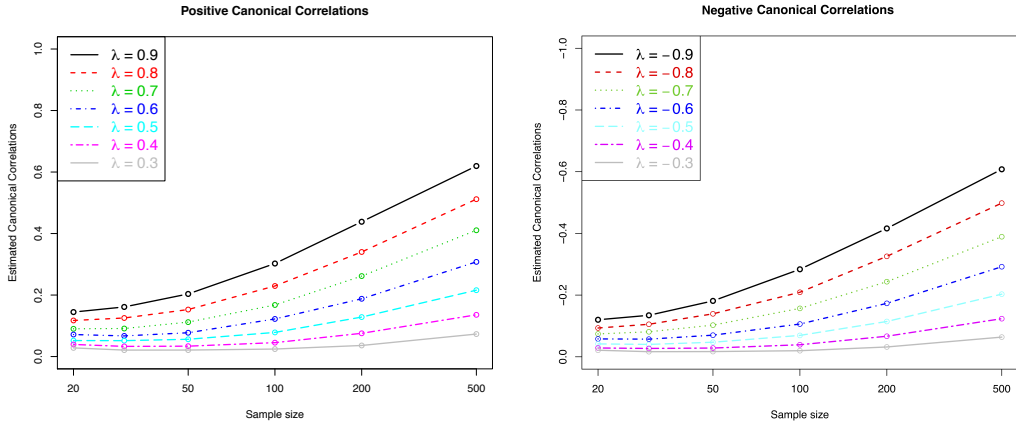


FIGURE B.1: Estimated canonical correlations with correctly identified signs in dependence of the sample size and the strength of the true canonical correlation for both positive (left) and negative (right) signs.

The outcome from this simulation study is summarized graphically in Fig. B.1. The figure mainly shows that for higher sample sizes the canonical correlations are recovered better than for small sample sizes. We compare these results to the results of fast regularized CCA using the `frcc` function in the `FRCC` R package. Figure B.2 depicts the estimated canonical correlation in FRCC for the same values of sample size and strength of canonical correlations. FRCC

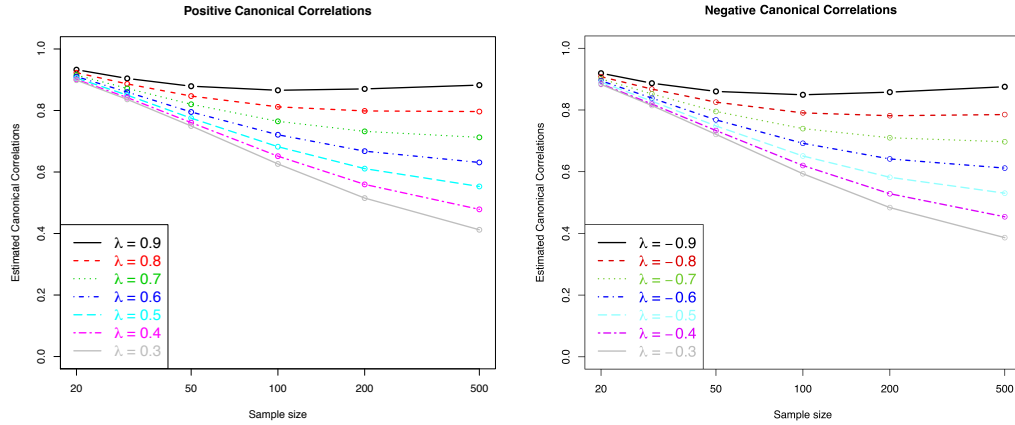


FIGURE B.2: Estimated canonical correlations from fast regularized CCA in dependence of the sample size and the strength of the true canonical correlation for both positive (left) and negative (right) signs.

performs poorly as well in recovering the true canonical correlations strengths except for a positive strong canonical correlation of 0.9. Similarly to our CCA model, FRCC also shows better recovery of positive canonical correlations for higher sample sizes. However, as FRCC has positivity constraints on canonical correlations it is not able to recover negative correlations and hence performs poorly than our CCA algorithm for negative correlations.

B.3 Analysis of nutrimouse data

In this section we perform a brief comparative analysis of our CCA whitening method to classical and sparse CCA depicted in figure B.3 and PCA in figures B.5 and B.6. Results in B.3 confirm the findings of Figure 5.7 in Chapter 5 that our CCA whitening method discriminates well between the wild type (WT) and PPAR- α deficient type. In addition, it shows higher correlation between the first and second canonical components than for classical and sparse CCA.

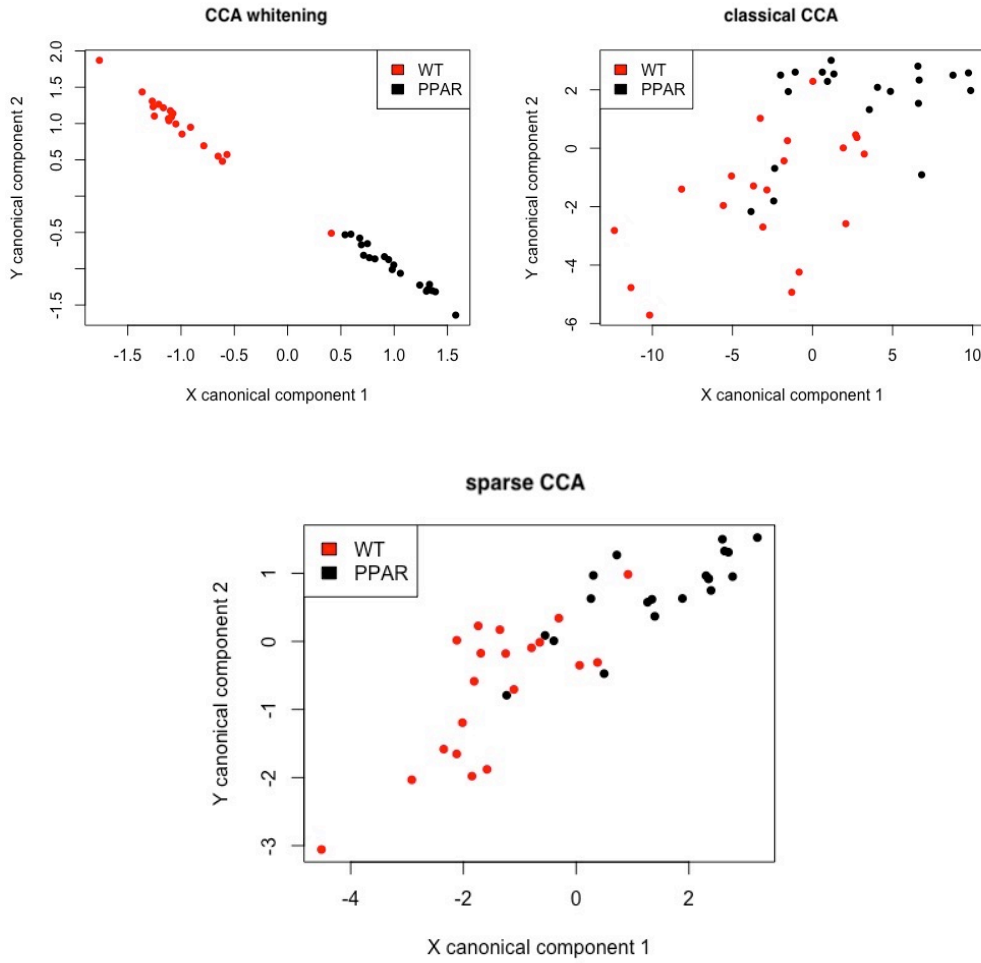


FIGURE B.3: Scatter plots between corresponding pairs of canonical covariates for the nutrimouse data. Samples are colored according to wild type (WT) and PPAR labels. Plots show that our CCA model outperforms classical and sparse CCA for nutrimouse data for exploratory analysis purposes by exhibiting better correlation and better discrimination across labels.

Figures B.5, B.6 and B.4 show results from PCA analysis of each of the gene expression data ($\mathbf{X}$) and the lipids data ($\mathbf{Y}$) separately. The scree plot in figure

B.4 indicates that 3 to 4 components are needed for the gene expression data compared to 1 canonical component as indicated by the correlation loadings plots (Figure 5.5) in Chapter 5. For the lipids data both methods indicate 4 to 5 components (See Figure 5.5 in Chapter 5). Overall, CCA indicates that less components are needed compared to PCA. This explained by the complementary information between the two datasets that gets captured in CCA (Figure 5.7 in Chapter 5). Plots in figures B.5 and B.6 give a flavor of which aspects are captured by each of the two PCA models. The first pair of PCA components of the gene data encode the genotype of the mice whilst component (3,4) and (4,5) are able to discriminate mainly two of the five diets imposed i.e COC and SUN. For the lipids data, the first pair of PCA components encodes the diet information. This is explained by the fact that lipids are directly related to oils used in the imposed diets. Components 2 and 3 explain to some extent the genotype of the mice and further components explain some diet types. The canonical components obtained from CCA benefit of a better visualization as shown in Figures 5.6 and 5.7 in Chapter 5 with the advantage that : i) CCA is an integrative approach performed once on both datasets, ii) overall, CCA results in less components needed and iii) CCA takes into account linear interaction between the omics variables.

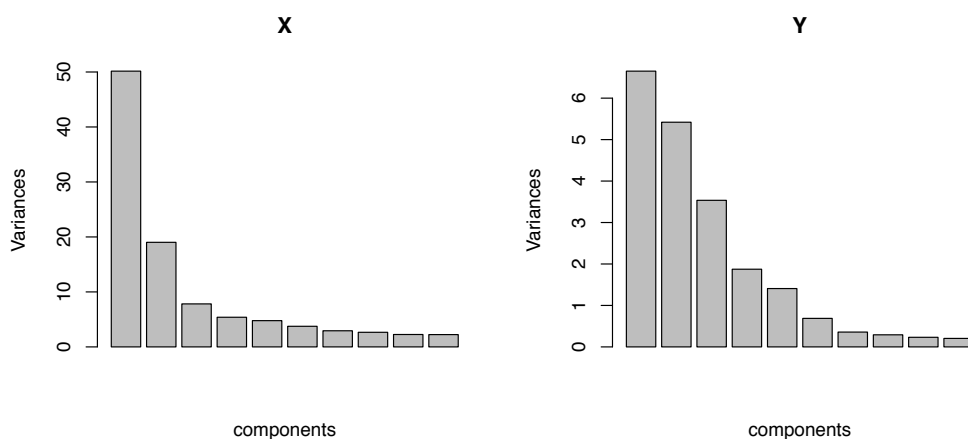


FIGURE B.4: PCA explained variance for gene expression data (left) and lipids data (right).

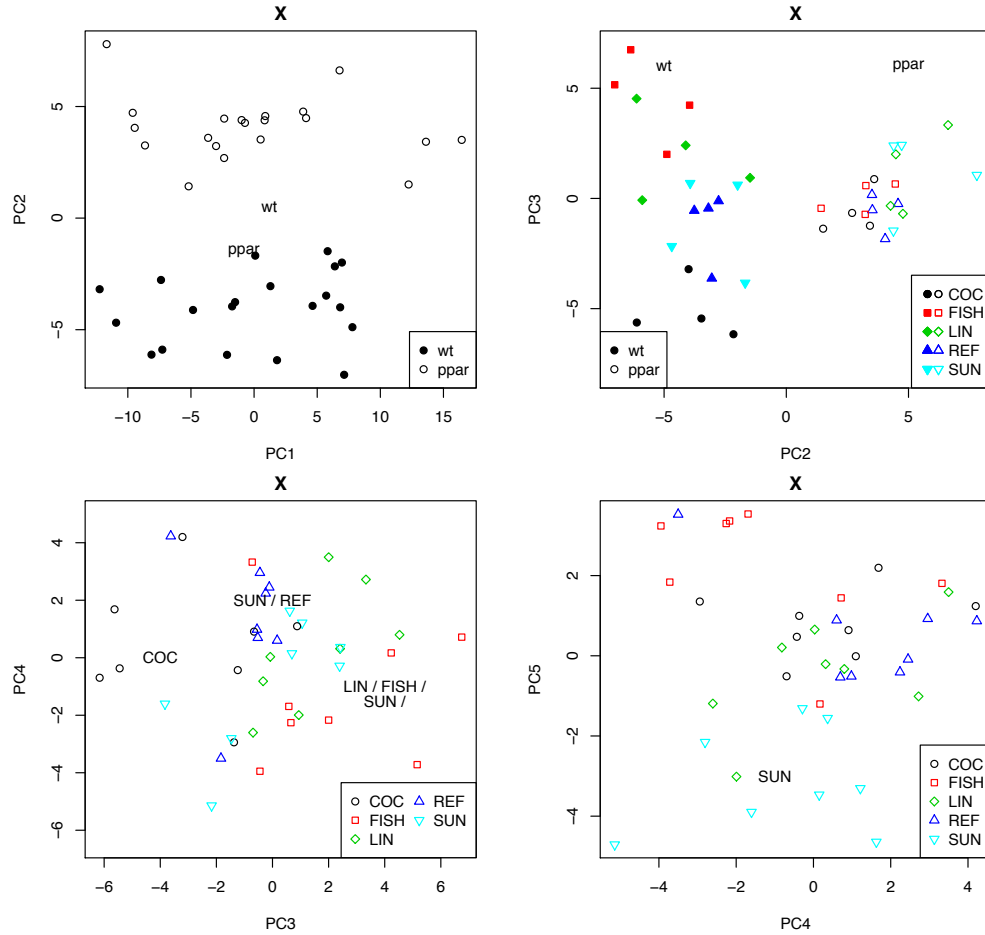


FIGURE B.5: Scatter plots between different pairs of components for the nutrinoise data. Samples are colored according to mice genotypes and diet labels. Plots show that the first two PCA components are able to discriminate between the genotype of the mice whilst components (3,4) and (4,5) are able to identify COC and SUN diets respectively.

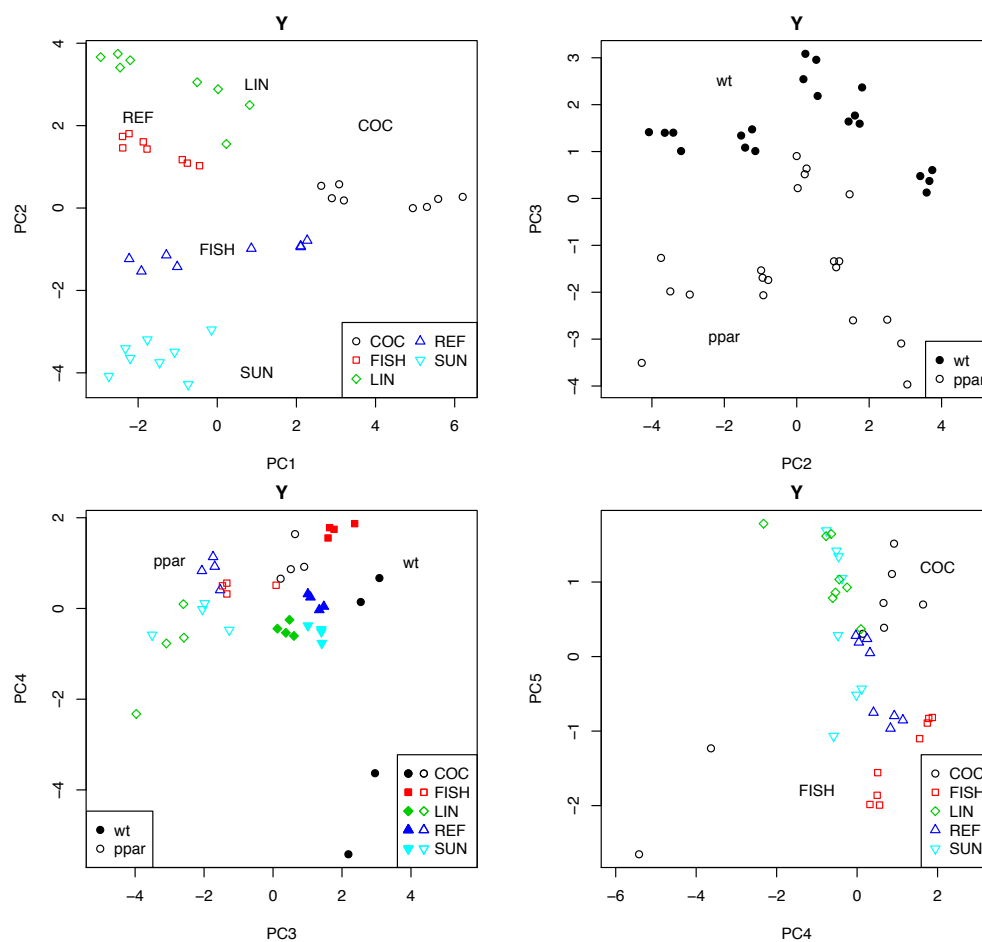


FIGURE B.6: Scatter plots between different pairs of components for the nutrimouse data. Samples are colored according to mice genotypes and diet labels. Plots show that the first two PCA components are able to discriminate between the different diets whilst components 2 and 3 show some information relative to the genotype of the mice. Further components potentially identify some diet types.

Appendix C

Appendix: Investigating infinite functional clustering for metabolite annotation

In this appendix we introduce Gaussian mixture models both in the finite (Section C.1) and infinite (Section C.2) settings. Technical details explained in this appendix are important to understand different properties of the DP applied to clustering, especially how Gaussian DP mixtures can be derived from finite Gaussian mixtures .

C.1 Finite Gaussian mixture models

Mixture models are one of the most popular generative models for clustering applications. Sampling from a mixture model concerns choosing a mixture component k with parameters $\boldsymbol{\theta}_k$ based on probabilities defined by the mixing proportions π_k (equation 3.5, chapter 2). If we place a symmetric Dirichlet prior on the mixing proportions and a discrete prior on the indicator variables z_i , we obtain the following mixture model :

$$\begin{aligned}
 \boldsymbol{\pi} | \gamma &\sim \text{Dirichlet} \left(\frac{\gamma}{K} \dots \frac{\gamma}{K} \right) \\
 z_j | \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\
 \boldsymbol{\theta}_k | H_0 &\sim H_0 \\
 \boldsymbol{x}_j | \boldsymbol{\Theta}, z_j &\sim \text{p}(\boldsymbol{x}_j | \boldsymbol{\theta}_{z_j})
 \end{aligned} \tag{C.1}$$

where K is the total number of clusters ¹. We should note here that θ_k are a set of parameters that define a distribution, specifically the mean and standard deviation $(\boldsymbol{\xi}, \boldsymbol{\Sigma})$ of a Gaussian. As such, the above generative model associates each object i with a set of latent parameters $\boldsymbol{\theta}_k = (\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k)$ that fully define the density $p(\mathbf{x}_i|\boldsymbol{\theta}_k)$. To achieve conjugacy, we define H_0 as the Gaussian Inverse Wishart (GIW) distribution which factorizes into a Gaussian distribution on $\boldsymbol{\xi}_k$ s.t $\boldsymbol{\xi}_k \sim G\left(\boldsymbol{\xi}_0, \frac{1}{\kappa_0} \boldsymbol{\Sigma}_k\right)$ and an Inverse Wishart distribution on $\boldsymbol{\Sigma}_k$ s.t $\boldsymbol{\Sigma}_k \sim IW(\nu_0, \boldsymbol{\Sigma}_0)$.

The Dirichlet distribution (Antoniak, 1974) is specified by its concentration parameters which are equal to $\frac{\gamma}{K}$ in this case. This leads to a symmetric Dirichlet distribution where, when γ is small, the density tends to concentrate on few components i.e. a large proportions of the observations is assigned to few components. On the other hand, if γ is large mixture components will contain similar proportions of the observations. This is best understood by looking into the conditional distribution $p(z_j|\mathbf{z}_{-j}, \gamma)$. In fact, integrating out the weights π_k , we get

$$p(z_j|\mathbf{z}_{-j}, \gamma) = \frac{n_k + \gamma/K}{n + \gamma} \quad (\text{C.2})$$

where n_k is the number of observations in component k . If γ is large, n_k becomes negligible and thus all components will have the same probability. Otherwise, if γ is small, the component containing a large proportion of samples will become larger.

We wish to cluster variables from metabolomic and genomic datasets mainly to help metabolite identification. As genomic variables are usually known, we wish to mainly obtain clusters that contain both heterogeneous variables. We will term these clusters as *heterogeneous clusters*. By investigating these clusters in accordance with pathway information, one might be able to derive potential candidates for unknown metabolites. Hence, we are interested in a clustering model able to merge clusters from different datasets into heterogeneous clusters. One popular approach consists in defining different Dirichlet priors for mixing weights in each data. Clusters induced by these priors are merged via another Dirichlet hyper-prior (Model C.3).

¹We use the terms “clusters”, “components” and “functional modules” interchangeably although they are slightly different as discussed in Yau and Holmes (2011)

$$\begin{aligned}
\mathbf{w}|\gamma &\sim \text{Dirichlet}\left(\frac{\gamma}{K}, \dots, \frac{\gamma}{K}\right) \\
\boldsymbol{\pi}_d|\mathbf{w} &\sim \text{Dirichlet}(\alpha_d \mathbf{w}) \\
z_{dj}|\boldsymbol{\pi}_d &\sim \boldsymbol{\pi}_d \\
\boldsymbol{\theta}_k|H_0 &\sim H_0 \\
\mathbf{x}_{dj}|\boldsymbol{\Theta}, z_{dj} &\sim p(\mathbf{x}_i|\boldsymbol{\theta}_{z_{dj}})
\end{aligned} \tag{C.3}$$

We introduce index d to denote data-specific variables with D the total number of data types. For the sake of simplicity, the j th variable from the d th dataset is denoted by x_{dj} . The index d is also introduced for the mixing proportions π_{dk} and indicator variables z_{dj} . The data-specific weights $\boldsymbol{\pi}_d$ control the size of data-specific clusters whilst $\mathbf{w}$ control the size of merged clusters and, thus, heterogeneous clusters. Since for our application the number of functional clusters is unknown, we will focus on an “advanced” version of the model that is able to infer the number of clusters. In fact, models C.1 and C.3 are representative finite models of the Dirichlet process (DP) and hierarchical Dirichlet processe (HDP) respectively (Teh et al., 2005).

C.2 Latent class model and the Dirichlet process

It is very important in the context of our application to understand different properties of the DP. The DP (Ferguson, 1973) is a normalised random measure and represents a cornerstone of Bayesian nonparametrics (Rasmussen, 2000; Escobar and Mike, 1995). The clustering property, related to the Dirichlet process, presents a draw from it as a perfect choice and the resulting generative model is as follows:

$$\begin{aligned}
D = \sum_{k \geq 1} \pi_k \delta_{\boldsymbol{\theta}_k} &\sim \text{DP}(\gamma, H_0) \\
\phi_j|D &\sim D \\
\mathbf{x}_j|\phi_j &\sim p(\mathbf{x}_j|\phi_j) \quad \text{for } j = 1 \dots p
\end{aligned} \tag{C.4}$$

where $\pi_k \in [0, 1]$ and $\sum_{k \geq 1} \pi_k = 1$. We should note here that use of π_k as in model C.1 is an abuse of notation as finite and infinite weights are equivalent up to a biased random permutation (Teh et al., 2005). The class assignments are formed through the clustering induced by ties in ϕ_j ’s. The choice of the

Dirichlet process as the prior over the latent measure D , allows for a possible infinite number of classes. Given the discreteness of D , i.e. $D = \sum_{k \geq 1} \pi_k \delta_{\theta_k}$, each sample ϕ_j , generated by the model in equation (C.5), coincides with an atom θ_k in D and there can be ties among the samples in the sense that $p(\phi_{j'} = \phi_j) > 0$ for any $j' \neq j$. These possible ties induce a *partition structure* Ψ_N among the samples ϕ_j 's. Suppose, for instance, that there are $K \leq p$ distinct values $\phi_1^*, \dots, \phi_K^*$ among the samples $\phi_1, \dots, \phi_p$ and let $C_k := \{j : \phi_j = \phi_k^*\}$ for $k = 1, \dots, K$. Any two different indices j and j' belong to the same group C_k if and only if $\phi_j = \phi_{j'} = \phi_k^*$. Consequently, the C_k 's describe a clustering scheme for the ϕ_j 's. If we consider $[p] := \{1, \dots, p\}$ as representing the indices of the p datapoints in the population, then a *partition* Ψ_p of $[p]$ is defined to be a collection of mutually exclusive, exhaustive, and non-empty subsets of $[p]$, also called clusters, i.e. $\Psi_p = \{C_1, \dots, C_K\}$ with K being the number of cluster present in the population. In the case of $[p] = \mathbb{N} = \{1, 2, 3, \dots\}$, the number of possible clusters might be infinite and we write $\Psi_{\mathbb{N}} = \{C_1, C_2, \dots\}$. However, in a finite population $\{1, \dots, p\}$, only a finite number of these are present. We should note here that $\theta_k \sim H_0$ where H_0 is now called *base measure*. The model defined in equation (C.5) is a hierarchical model and it is known as Dirichlet Process Mixture Model (DPMM) with numerous applications in biostatistics (Dunson, 2010; Mitra and Müller, 2015). To sum up, the DPMM has three levels; the random measure level where the directing measure is sampled, the combinatorial level where the cluster assignments are drawn, and the data generation level where the observations are drawn.

The equivalence between the DPMM and model C.1 is best understood in terms of the latent indicator variable z_j and weights π . In the infinite setting, the weights π result from a *stick breaking construction* (Sethuraman, 1994) where :

$$\pi'_k \sim \text{Beta}(1, \gamma) \tag{C.5}$$

$$\pi_k = \pi'_k \prod_{l=1}^{k-1} (1 - \pi'_l) \tag{C.6}$$

The process defined by steps C.5 and C.6 generates an infinite collection of weights for the Dirichlet process. It will be denoted in the following by stick(γ). Thus, an equivalent representation of DPMM can be simply specified

by the model below:

$$\begin{aligned}
 \boldsymbol{\pi} | \gamma &\sim \text{stick}(\gamma) \\
 z_j | \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\
 \boldsymbol{\theta}_k | H_0 &\sim H_0 \\
 \mathbf{x}_j | \boldsymbol{\theta}, z_j &\sim p(\mathbf{x}_j | \theta_{z_j})
 \end{aligned} \tag{C.7}$$

Sampling from the DPMM is closely related to the *chinese restaurant process* (Aldous, 1985). The process relates to a chinese restaurant with an infinite number of tables i.e $K \rightarrow \infty$. The j th customer $\mathbf{x}_j$ enters the restaurant and sits at a table $\boldsymbol{\theta}_k$ with probability $\frac{n_k}{\gamma + n - 1}$ where n_k is the number of customers already sitting at table $\boldsymbol{\theta}_k$. This is the limit of equation C.2 when $K \rightarrow \infty$. The customer might also choose a new table with probability $\frac{\gamma}{\gamma + n - 1}$. Through this metaphor, it becomes clear that the Dirichlet process inherits the properties of the Dirichlet distribution. Particularly, if γ is large, new customers tends to choose new tables which results in many clusters with small size. If γ is small, new customers tend to choose a table where a large number of customers is sitting already. This results in few clusters with a large size. This is referred to as “the rich gets richer” property of the Dirichlet process (See figure C.1). In fact, the number of components inferred by the Dirichlet process is approximately $O(\gamma \log(n))$ (Teh et al., 2005).

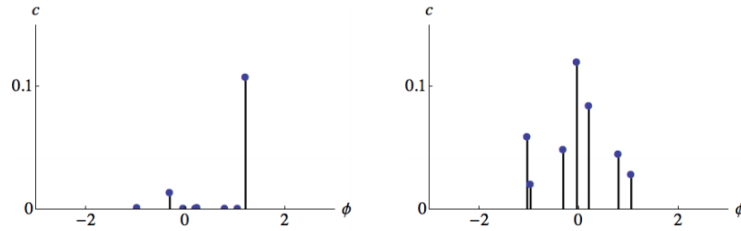


FIGURE C.1: Figure from Orbanz (2012). The figure shows realizations of Dirichlet processes with different concentration parameters γ . The x-axis shows atoms locations and y-axis shows atom weights. Left: For $\gamma = 1$ the density is concentrated on few atoms who have large weights compared to others. Thus most of the observations will likely belong to small number of clusters. Right : For $\gamma = 10$ the weights are more even and a higher number of clusters is expected.

C.3 Hierarchical Dirichlet process and collapsed Gibbs sampling

The hierarchical Dirichlet process Teh et al. (2006) is derived by adding a DP level to model C.7 as follows:

$$\begin{aligned}
\mathbf{w}|\gamma &\sim \text{stick}(\gamma) \\
\boldsymbol{\pi}_d|\alpha_d, \mathbf{w} &\sim \text{DP}(\alpha_d, \mathbf{w}) \\
z_{dj}|\boldsymbol{\pi}_d &\sim \boldsymbol{\pi}_d \\
\boldsymbol{\theta}_k|H_0 &\sim H_0 \\
\mathbf{x}_{dj}|\boldsymbol{\theta}, z_{dj} &\sim \text{p}(\mathbf{x}_{dj}|\boldsymbol{\theta}_{z_{dj}})
\end{aligned} \tag{C.8}$$

Collapsed Gibbs sampling

We put a Gaussian Inverse Wishart (GIW) distribution on $\boldsymbol{\theta}_k$ s.t $\boldsymbol{\xi}_k \sim \text{G}\left(\boldsymbol{\xi}_0, \frac{1}{\kappa_0} \boldsymbol{\Sigma}_k\right)$ and $\boldsymbol{\Sigma}_k \sim \text{IW}(\nu_0, \boldsymbol{\Sigma}_0)$. Since the GIW is a conjugate prior to the normal distribution $\text{p}(\mathbf{x}_{dj}|\boldsymbol{\theta}_k) = \text{G}(\mathbf{x}_{dj}|\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k)$, we can marginalize out the component parameters using an integral over their posterior. Hence, the data log posterior predictive can be written as below:

$$\begin{aligned}
\log \text{p}(\mathbf{x}_{dj}|\mathbf{X}_{1:2,-j}, z_{dj} = k, \mathbf{z}_{-dj}) &= -\frac{n}{2}\pi + \log \Gamma\left(\frac{\nu_k + 1}{2}\right) \\
&\quad - \log \Gamma\left(\frac{\nu_k + 1 - n}{2}\right) + \frac{\nu_k}{2}|\mathbf{S}_k| \\
&\quad + \frac{\nu_k + 1}{2}|\mathbf{S}_k + \frac{\kappa_k}{\kappa_k + 1}(\mathbf{x}_j - \boldsymbol{\xi}_k)(\mathbf{x}_j - \boldsymbol{\xi}_k)^T|
\end{aligned} \tag{C.9}$$

where the following equations hold for each component:

$$\begin{aligned}
\boldsymbol{\xi}_k &= \frac{\kappa_0 \boldsymbol{\xi}_0 + \sum_{j': \mathbf{z}_{dj'}=k} \mathbf{x}_{j'}}{\kappa_k} \\
\kappa_k &= \kappa_0 + p_k \\
\nu_k &= \nu_0 + p_k \\
\mathbf{S}_k &= \mathbf{S}_0 + \sum_{j': \mathbf{z}_{dj'}=k} \mathbf{x}_{j'} \mathbf{x}_{j'}^T + \kappa_0 \boldsymbol{\xi}_0 \boldsymbol{\xi}_0^T - \kappa_k \boldsymbol{\xi}_k \boldsymbol{\xi}_k^T
\end{aligned} \tag{C.10}$$

To sample from the CRF (Teh et al., 2006), we first need to sample table assignments (data-specific assignments) for each observation j . Assignment of

an observation j to a table t in restaurant d is denoted by t_{dj} . If the table $t_{dj} = t$ has already been used before, the probability of choosing that table is proportional to p_{dt} i.e. the number of observations already assigned to that table². Observation j is then assigned the dish already served in table t_{dj} denoted by k_{jt} . A new table $t_{dj} = t_{new}$ can also be chosen with a probability proportional to α_d . In this case, all possible values for the dish choice should be integrated out. Combining assignment prior probabilities with the data likelihood in equation yields the following conditional posterior for table assignments:

$$p(t_{dj} | \mathbf{X}_{1:2,-j}, \mathbf{z}, \mathbf{t}_{-dj}) \propto \begin{cases} p_{dt} p(\mathbf{x}_j | \mathbf{X}_{1:2,-j}, z_{dj} = k_{jt}, \mathbf{z}_{-dj}) & \text{if } t_{dj} \text{ used} \\ \sum_k \frac{\alpha_d m_k}{\gamma + m} p(\mathbf{x}_j | \mathbf{X}_{1:2,-j}, z_{dj} = k, \mathbf{z}_{-dj}) \\ + \frac{\alpha_d \gamma}{\gamma + m} p(\mathbf{x}_j | \mathbf{X}_{1:2,-j}, z_{dj} = k_{new}, \mathbf{z}_{-dj}) & \text{if } t_{dj} \text{ new} \end{cases} \quad (\text{C.11})$$

If a new table is sampled, a dish needs to be assigned by sampling from

$$p(z_{dj} = k_{jt_{new}} | \mathbf{X}_{1:2,-j}, \mathbf{z}_{-dj}, \mathbf{t}) \propto \begin{cases} m_{k_{jt_{new}}} p(\mathbf{x}_j | \mathbf{X}_{1:2,-j}, z_{dj} = k_{jt_{new}}, \mathbf{z}_{-dj}) & \text{if } k_{jt_{new}} \text{ used} \\ \gamma p(\mathbf{x}_j | \mathbf{X}_{1:2,-j}, z_{dj} = k_{new}, \mathbf{z}_{-dj}) & \text{if new dish} \end{cases} \quad (\text{C.12})$$

Once table assignments are sampled, dish assignments need to be updated. Under the CRF, a new dish is chosen with probability proportional to m_k if already used or γ if the dish is new. Hence, sampling dish assignments follows similar equations as C.12 by replacing $\mathbf{x}_j$ by $\mathbf{X}^t$ where $\mathbf{X}^t$ is the subset of observations assigned to table t . Updates are specified by equation C.13

$$p(z_{dj} = k_{jt} | \mathbf{X}_{1:2,-t}, \mathbf{z}_{-dj}, \mathbf{t}) \propto \begin{cases} m_{k_{jt}} p(\mathbf{X}^t | \mathbf{X}_{1:2,-t}, z_{dj} = k_{jt}, \mathbf{z}_{-dj}) & \text{if } k_{jt} \text{ used} \\ \gamma p(\mathbf{X}^t | \mathbf{X}_{1:2,-t}, z_{dj} = k_{new}, \mathbf{z}_{-dj}) & \text{if new dish} \end{cases} \quad (\text{C.13})$$

In each MCMC iteration, once all cluster assignments are sampled, probabilistic metabolite annotation is performed by computing $p(\psi_j | \mathbf{X}, \psi_{-j}, \mathbf{z})$ via equation 6.3 or $p(\psi_j | \psi_{-j}, \mathbf{z})$ via equations 6.2 or 6.4.

²exchangeability property allows to consider variable j as the last variable being sampled.

Bibliography

- Abou-Jaoudé, Wassim, Madalena Chaves, and Jean-Luc Gouzé (2014). “Links between topology of the transition graph and limit cycles in a two-dimensional piecewise affine biological model”. In: *Journal of mathematical biology* 69.6–7, pp. 1461–1495.
- Acharjee, Animesh et al. (2016). “Integration of metabolomics, lipidomics and clinical data using a machine learning method”. In: *BMC bioinformatics* 17.15, p. 440.
- Adrover, Jorge G and Stella M Donato (2015). “A robust predictive approach for canonical correlation analysis”. In: *Journal of Multivariate Analysis* 133, pp. 356–376.
- Akaike, Hirotugu (1985). “Prediction and Entropy”. In: *A Celebration of Statistics*. Ed. by Anthony C. Atkinson and Stephen E. Fienberg. New York, NY: Springer New York, pp. 1–24. ISBN: 978-1-4613-8560-8.
- Al-Shahrour, Fátima, Ramón Díaz-Uriarte, and Joaquín Dopazo (2004). “FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes”. In: *Bioinformatics* 20.4, pp. 578–580.
- Aldous, David J (1985). “Exchangeability and related topics”. In: *École d’Été de Probabilités de Saint-Flour XIII—1983*. Springer, pp. 1–198.
- Amari, S. (2009). “ α -Divergence Is Unique, Belonging to Both f -Divergence and Bregman Divergence Classes”. In: *IEEE Transactions on Information Theory* 55.11, pp. 4925–4931. ISSN: 0018-9448. DOI: [10.1109/TIT.2009.2030485](https://doi.org/10.1109/TIT.2009.2030485).
- Antoniak, Charles E (1974). “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems”. In: *The annals of statistics*, pp. 1152–1174.
- Antti, Henrik et al. (2002). “Batch statistical processing of ¹H NMR-derived urinary spectral data”. In: *Journal of Chemometrics* 16.8–10, pp. 461–468.
- Argiento, Raffaele et al. (2015). “Modeling the association between clusters of SNPs and disease responses”. In: *Nonparametric Bayesian Inference in Biostatistics*. Springer, pp. 115–134.

- Ashburner, Michael et al. (2000). “Gene ontology: tool for the unification of biology”. In: *Nature genetics* 25.1, p. 25.
- Bach, Francis R and Michael I Jordan (2005). “A probabilistic interpretation of canonical correlation analysis”. In:
- Banerjee, Sudipto, Bradley P Carlin, and Alan E Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.
- Barber, David (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
- Barupal, Dinesh K et al. (2012). “MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity”. In: *BMC bioinformatics* 13.1, p. 99.
- Baumeier, Christian et al. (2017). “Elevated hepatic DPP4 activity promotes insulin resistance and non-alcoholic fatty liver disease”. In: *Molecular metabolism* 6.10, pp. 1254–1263.
- Becker, Emmanuelle et al. (2012). “Multifunctional proteins revealed by overlapping clustering in protein interaction network”. In: *Bioinformatics* 28.1, pp. 84–90. ISSN: 13674803.
- Beckonert, Olaf et al. (2007). “Metabolic profiling, metabolomic and metabolomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts.” In: *Nature protocols* 2.11, pp. 2692–2703. ISSN: 17502799. DOI: [10.1038/nprot.2007.376](https://doi.org/10.1038/nprot.2007.376).
- Beißbarth, Tim and Terence P Speed (2004). “GOstat: find statistically over-represented Gene Ontologies within a group of genes”. In: *Bioinformatics* 20.9, pp. 1464–1465.
- Berk, Maurice, Timothy Ebbels, and Giovanni Montana (2011). “A statistical framework for biomarker discovery in metabolomic time course data”. In: *Bioinformatics* 27.14, pp. 1979–1985.
- Bernardo, Jose M. (1979). “Expected Information as Expected Utility”. In: *Ann. Statist.* 7.3, pp. 686–690. DOI: [10.1214/aos/1176344689](https://doi.org/10.1214/aos/1176344689). URL: <https://doi.org/10.1214/aos/1176344689>.
- Besag, Julian (1974). “Spatial interaction and the statistical analysis of lattice systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225.
- Bhadra, Anindya et al. (2017). “The horseshoe+ estimator of ultra-sparse signals”. In: *Bayesian Analysis* 12.4, pp. 1105–1131.

- Bhattacharya, Anirban et al. (2015). “Dirichlet–Laplace priors for optimal shrinkage”. In: *Journal of the American Statistical Association* 110.512, pp. 1479–1490.
- Bickel, Peter J, Elizaveta Levina, et al. (2008). “Regularized estimation of large covariance matrices”. In: *The Annals of Statistics* 36.1, pp. 199–227.
- Bird, Adrian (2007). “Perceptions of epigenetics”. In: *Nature* 447.7143, p. 396.
- Bleasby, AJ, D Akrigg, and TK Attwood (1994). “OWL—a non-redundant composite protein sequence database.” In: *Nucleic acids research* 22.17, p. 3574.
- Bloemberg, Tom G et al. (2013). “Warping methods for spectroscopic and chromatographic signal alignment: a tutorial”. In: *Analytica chimica acta* 781, pp. 14–32.
- Bottolo, Leonardo et al. (2013). “GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm”. In: *PLoS genetics* 9.8, e1003657.
- Bro, Rasmus (1997). “PARAFAC. Tutorial and applications”. In: *Chemometrics and intelligent laboratory systems* 38.2, pp. 149–171.
- Brychkov, Yury A (2008). *Handbook of special functions: derivatives, integrals, series and other formulas*. CRC Press.
- Burge, Chris and Samuel Karlin (1997). “Prediction of complete gene structures in human genomic DNA”. In: *Journal of molecular biology* 268.1, pp. 78–94.
- Camacho, Diogo, Alberto De La Fuente, and Pedro Mendes (2005). “The origin of correlations in metabolomics data”. In: *Metabolomics* 1.1, pp. 53–63.
- Cappadona, Salvatore et al. (2008). “Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry”. In: *Analytical Chemistry* 80.13, pp. 4960–4968.
- Carbon, Seth et al. (2008). “AmiGO: online access to ontology and annotation data”. In: *Bioinformatics* 25.2, pp. 288–289.
- Carvalho, Carlos M, Nicholas G Polson, and James G Scott (2009). “Handling sparsity via the horseshoe”. In: *Artificial Intelligence and Statistics*, pp. 73–80.
- Case, Anne C, Harvey S Rosen, and James R Hines Jr (1993). “Budget spillovers and fiscal policy interdependence: Evidence from the states”. In: *Journal of public economics* 52.3, pp. 285–307.
- Caspi, Ron et al. (2013). “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic acids research* 42.D1, pp. D459–D471.

- Castillo, Sandra et al. (2011). “Algorithms and tools for the preprocessing of LC–MS metabolomics data”. In: *Chemometrics and Intelligent Laboratory Systems* 108.1, pp. 23–32.
- Cavill, Rachel et al. (2011). “Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells”. In: *PLoS computational biology* 7.3, e1001113.
- Cavill, Rachel et al. (2015). “Transcriptomic and metabolomic data integration”. In: *Briefings in bioinformatics* 17.5, pp. 891–901.
- Chadeau-Hyam, Marc et al. (2010). “Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification”. In: *Journal of proteome research* 9.9, pp. 4620–4627.
- Chadeau-Hyam, Marc et al. (2013). “Deciphering the complex: Methodological overview of statistical models to derive OMICS-based biomarkers”. In: *Environmental and molecular mutagenesis* 54.7, pp. 542–557.
- Chagoyen, Monica and Florencio Pazos (2012). “Tools for the functional interpretation of metabolomic experiments”. In: *Briefings in bioinformatics* 14.6, pp. 737–744.
- Chatziioannou, Aristotelis et al. (2017). “Blood-based omic profiling supports female susceptibility to tobacco smoke-induced cardiovascular diseases”. In: *Scientific reports* 7, p. 42870.
- Chen, Chen et al. (2015). “Exploring metabolic profile differences between colorectal polyp patients and controls using seemingly unrelated regression”. In: *Journal of Proteome Research* 14.6, pp. 2492–2499. ISSN: 15353907. DOI: [10.1021/acs.jproteome.5b00059](https://doi.org/10.1021/acs.jproteome.5b00059).
- Chen, Chen et al. (2017). “Altered metabolite levels and correlations in patients with colorectal cancer and polyps detected using seemingly unrelated regression analysis”. In: *Metabolomics* 13.11, p. 125. ISSN: 1573-3882. DOI: [10.1007/s11306-017-1265-0](https://doi.org/10.1007/s11306-017-1265-0). URL: <http://link.springer.com/10.1007/s11306-017-1265-0>.
- Chen, Lifeng and Dennis Vitkup (2006). “Predicting genes for orphan metabolic activities using phylogenetic profiles”. In: *Genome biology* 7.2, R17.
- Chong, Jasmine et al. (2018). “MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis”. In: *Nucleic acids research* 46.W1, W486–W494.
- Consortium, Fantom, RIKEN Genome Exploration Research Group Phase I & II Team, et al. (2002). “Analysis of the mouse transcriptome based on

- functional annotation of 60,770 full-length cDNAs". In: *Nature* 420.6915, p. 563.
- Coughlin, Steven S (2014). "Toward a road map for global-omics: a primer on-omic technologies". In: *American journal of epidemiology* 180.12, pp. 1188–1195.
- Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience. ISBN: 0471241954.
- Craig, Andrew et al. (2006). "Scaling and normalization effects in NMR spectroscopic metabonomic data sets". In: *Analytical chemistry* 78.7, pp. 2262–2267.
- Cramer, Elliot M and W Alan Nicewander (1979). "Some symmetric, invariant measures of multivariate association". In: *Psychometrika* 44.1, pp. 43–54.
- Cressie, Noel (1993). "Regional mapping of incidence rates using spatial Bayesian models". In: *Medical care*, YS60–YS65.
- Cressie, Noel, Olivier Perrin, and Christine Thomas-Agnan (2005). "Likelihood-based estimation for Gaussian MRFs". In: *Statistical Methodology* 2.1, pp. 1–16.
- Cressie, Noel and Christopher K Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Cruickshank-Quinn, Charmion I et al. (2018). "Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD". In: *Scientific reports* 8.1, p. 17132.
- Cruz-Cano, Raul and Mei-Ling Ting Lee (2014). "Fast regularized canonical correlation analysis". In: *Computational Statistics & Data Analysis* 70, pp. 88–100.
- Csiszar, Imre (1991). "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems". In: *Ann. Statist.* 19.4, pp. 2032–2066. DOI: [10.1214/aos/1176348385](https://doi.org/10.1214/aos/1176348385). URL: <https://doi.org/10.1214/aos/1176348385>.
- Cui, Qiu et al. (2008). "Metabolite identification via the madison metabolomics consortium database". In: *Nature biotechnology* 26.2, p. 162.
- Cunningham, John P and Zoubin Ghahramani (2015). "Linear dimensionality reduction: Survey, insights, and generalizations". In: *The Journal of Machine Learning Research* 16.1, pp. 2859–2900.

- Cushman, John C and Hans J Bohnert (2000). “Genomic approaches to plant stress tolerance”. In: *Current opinion in plant biology* 3.2, pp. 117–124.
- Daemen, Anneleen et al. (2009). “A kernel-based integration of genome-wide data for clinical decision support”. In: *Genome medicine* 1.4, p. 39.
- Daly, Rónán et al. (2014). “MetAssign: probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach”. In: *Bioinformatics* 30.19, pp. 2764–2771.
- Daub, Carsten O et al. (2004). “Estimating mutual information using B-spline functions—an improved similarity measure for analysing gene expression data”. In: *BMC bioinformatics* 5.1, p. 118.
- Davis, Darcy A and Nitesh V Chawla (2011). “Exploring and exploiting disease interactions from multi-relational gene and phenotype networks”. In: *PloS one* 6.7, e22670.
- Daviss, Bennett (2005). “Growing pains for metabolomics: the newest’omic science is producing results—and more data than researchers know what to do with”. In: *The Scientist* 19.8, pp. 25–29.
- De Jong, Sijmen (1993). “SIMPLS: an alternative approach to partial least squares regression”. In: *Chemometrics and intelligent laboratory systems* 18.3, pp. 251–263.
- Degtyarenko, Kirill et al. (2007). “ChEBI: a database and ontology for chemical entities of biological interest”. In: *Nucleic acids research* 36.suppl_1, pp. D344–D350.
- Dennis, Glynn et al. (2003). “DAVID: database for annotation, visualization, and integrated discovery”. In: *Genome biology* 4.9, R60.
- Dettmer, Katja, Pavel A Aronov, and Bruce D Hammock (2007). “Mass spectrometry-based metabolomics”. In: *Mass spectrometry reviews* 26.1, pp. 51–78.
- Devlin, Thomas M (2011). *Textbook of biochemistry*. John Wiley & Sons.
- Dhanasekaran, A Ranjitha et al. (2015). “Metabolome searcher: a high throughput tool for metabolite identification and metabolic pathway mapping directly from mass spectrometry and using genome restriction”. In: *BMC bioinformatics* 16.1, p. 62.
- Dieterle, Frank et al. (2006). “Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ¹H NMR metabonomics”. In: *Analytical chemistry* 78.13, pp. 4281–4290.

- Dietrich, Wolfgang, Christian H Rüdél, and Markus Neumann (1991). “Fast and precise automatic baseline correction of one-and two-dimensional NMR spectra”. In: *Journal of Magnetic Resonance (1969)* 91.1, pp. 1–11.
- Duensing, Stefan and Karl Münger (2004). “Mechanisms of genomic instability in human cancer: insights from studies with human papillomavirus oncoproteins”. In: *International Journal of Cancer* 109.2, pp. 157–162.
- Dunn, Warwick B and David I Ellis (2005). “Metabolomics: current analytical platforms and methodologies”. In: *TrAC Trends in Analytical Chemistry* 24.4, pp. 285–294.
- Dunson, David B (2010). “Nonparametric Bayes applications to biostatistics”. In: *Bayesian nonparametrics* 28, p. 223.
- Ebrahim, Ali et al. (2016). “Multi-omic data integration enables discovery of hidden biological regularities”. In: *Nature communications* 7.
- Efron, Bradley (2008). “Microarrays, empirical Bayes and the two-groups model”. In: *Statistical science*, pp. 1–22.
- Eisen, Michael B et al. (1998). “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14863–14868.
- Elias-Smale, S. E. et al. (2010). “Coronary Calcium Score improves classification of coronary heart disease risk in the elderly: The Rotterdam Study”. In: *Journal fo the American College of Cardiology* 56.17, pp. 1407–1414.
- Elliott, P. et al. (2015). “Urinary metabolic signatures of human adiposity”. In: *Science Translational Medicine* 7.285, 285ra62.
- Escobar, Michael D and West Mike (1995). “Bayesian Density Estimation and Inference Using Mixtures”. In: pp. 0–22.
- Escoufier, Yves (1973). “Le Traitement des Variables Vectorielles”. In: *Biometrics* 29.4, pp. 751–760. ISSN: 0006341X, 15410420. URL: <http://www.jstor.org/stable/2529140>.
- Evangelou, Evangelos and John PA Ioannidis (2013). “Meta-analysis methods for genome-wide association studies and beyond”. In: *Nature Reviews Genetics* 14.6, p. 379.
- Fahy, Eoin et al. (2007). “LIPID MAPS online tools for lipid research”. In: *Nucleic acids research* 35.suppl_2, W606–W612.
- Fasanelli, Francesca et al. (2015). “Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts”. In: *Nature communications* 6, p. 10192.

- Faust, Karoline, Didier Croes, and Jacques van Helden (2011). "Prediction of metabolic pathways from genome-scale metabolic networks". In: *Biosystems* 105.2, pp. 109–121.
- Ferguson, Thomas S (1973). "A Bayesian analysis of some nonparametric problems". In: *The annals of statistics*, pp. 209–230.
- Fiehn, Oliver (2002). "Metabolomics—the link between genotypes and phenotypes". In: *Functional genomics*. Springer, pp. 155–171.
- Fitzgibbon, Matthew et al. (2008). "Open-source platform for the analysis of liquid chromatography-mass spectrometry (LC-MS) data". In: *Clinical Proteomics*. Springer, pp. 369–381.
- Fonville, Judith M et al. (2010). "The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping". In: *Journal of Chemometrics* 24.11-12, pp. 636–649.
- Forshed, Jenny, Helena Idborg, and Sven P Jacobsson (2007). "Evaluation of different techniques for data fusion of LC/MS and 1H-NMR". In: *Chemometrics and intelligent laboratory systems* 85.1, pp. 102–109.
- Fraley, Chris and Adrian E Raftery (2006). *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. Tech. rep. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS.
- Fridley, Brooke L et al. (2012). "A Bayesian Integrative Genomic Model for Pathway Analysis of Complex Traits". In: *Genetic epidemiology* 36.4, pp. 352–359.
- Friebolin, Horst and Jack K Beconsall (1993). *Basic one-and two-dimensional NMR spectroscopy*. VCH Weinheim.
- Frith, Martin C, Ulla Hansen, and Zhiping Weng (2001). "Detection of cis-element clusters in higher eukaryotic DNA". In: *Bioinformatics* 17.10, pp. 878–889.
- Frith, Martin C, Michael C Li, and Zhiping Weng (2003). "Cluster-Buster: Finding dense clusters of motifs in DNA sequences". In: *Nucleic acids research* 31.13, pp. 3666–3668.
- Fritsch, Arno (2012). "mcclust: Process an MCMC Sample of Clusterings". In: *R package version 1*.
- Fritsch, Arno, Katja Ickstadt, et al. (2009). "Improved criteria for clustering based on the posterior similarity matrix". In: *Bayesian analysis* 4.2, pp. 367–391.

- Fu, Wenjiang J (1998). “Penalized regressions: the bridge versus the lasso”. In: *Journal of computational and graphical statistics* 7.3, pp. 397–416.
- Fukuyama, Y. and M. Sugeno (1989). “A new method of choosing the number of clusters for fuzzy c-means method”. In: *Proceedings of the 5th Fuzzy Systems Symposium*.
- Gan, Feng, Guihua Ruan, and Jinyuan Mo (2006). “Baseline correction by improved iterative polynomial fitting with automatic threshold”. In: *Chemo-metrics and Intelligent Laboratory Systems* 82.1-2, pp. 59–65.
- Gao, Jing et al. (2010). “Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks”. In: *Bioinformatics* 26.7, pp. 971–973.
- Garber, Manuel et al. (2011). “Computational methods for transcriptome annotation and quantification using RNA-seq”. In: *Nature methods* 8.6, p. 469.
- Gelman, Andrew et al. (2006). “Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)”. In: *Bayesian analysis* 1.3, pp. 515–534.
- Gevaert, Kris and Joël Vandekerckhove (2000). “Protein identification methods in proteomics”. In: *ELECTROPHORESIS: An International Journal* 21.6, pp. 1145–1154.
- Gieger, Christian et al. (2008). “Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum”. In: *PLoS genetics* 4.11, e1000282.
- Gipson, Geoffrey T et al. (2008). “Assignment of MS-based metabolomic datasets via compound interaction pair mapping”. In: *Metabolomics* 4.1, pp. 94–103.
- Gloaguen, Yoann et al. (2017). “PiMP my metabolome: an integrated, web-based tool for LC-MS metabolomics data”. In: *Bioinformatics* 33.24, pp. 4007–4009.
- Gowda, Harsha et al. (2014). “Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses”. In: *Analytical chemistry* 86.14, pp. 6931–6939.
- Griffiths, Thomas L. and Zoubin Ghahramani (2005). “Infinite Latent Feature Models and the Indian Buffet Process”. In: *Adv. Neural Inf. Process. Syst.* 18, pp. 475–482. ISSN: 1049-5258.
- Gross, Jürgen H (2006). *Mass spectrometry: a textbook*. Springer Science & Business Media.

- Guida, Florence et al. (2015). “Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation”. In: *Human molecular genetics* 24.8, pp. 2349–2359.
- Günther, Ulrich L, Christian Ludwig, and H Rüterjans (2000). “NMRLAB—advanced NMR data processing in MATLAB”. In: *Journal of Magnetic Resonance* 145.2, pp. 201–208.
- Guo, Xingli et al. (2011). “A computational method based on the integration of heterogeneous networks for predicting disease-gene associations”. In: *PloS one* 6.9, e24171.
- Hamid, Jemila S et al. (2009). “Data integration in genetics and genomics: methods and challenges”. In: *Human genomics and proteomics: HGP 2009*.
- Hannart, Alexis and Philippe Naveau (2014). “Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework”. In: *Journal of Multivariate Analysis* 131, pp. 149–162.
- Hao, Jie et al. (2012). “BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model”. In: *Bioinformatics* 28.15, pp. 2088–2090.
- Härdle, Wolfgang Karl and Léopold Simar (2015). “Canonical correlation analysis”. In: *Applied multivariate statistical analysis*. Springer, pp. 443–454.
- Hardoon, David R and John Shawe-Taylor (2011). “Sparse canonical correlation analysis”. In: *Machine Learning* 83.3, pp. 331–353.
- Harrigan, George G and Royston Goodacre (2003). *Metabolic profiling: its role in biomarker discovery and gene function analysis: its role in biomarker discovery and gene function analysis*. Springer Science & Business Media.
- Harrington, Peter de B et al. (2005). “Analysis of variance–principal component analysis: A soft tool for proteomic discovery”. In: *Analytica chimica acta* 544.1-2, pp. 118–127.
- Harrison, Lee M and Gary GR Green (2010). “A Bayesian spatiotemporal model for very large data sets”. In: *NeuroImage* 50.3, pp. 1126–1141.
- Hastie, Trevor and Robert Tibshirani (2004). “Efficient quadratic regularization for expression arrays”. In: *Biostatistics* 5.3, pp. 329–340.
- Haukaas, Tonje H et al. (2017). “Metabolic portraits of breast cancer by HR MAS MR spectroscopy of intact tissue samples”. In: *Metabolites* 7.2, p. 18.
- Heard, Edith and Robert A Martienssen (2014). “Transgenerational epigenetic inheritance: myths and mechanisms”. In: *Cell* 157.1, pp. 95–109.

- Heller, Katherine and Zoubin Ghahramani (2007). “A Nonparametric Bayesian Approach to Modeling Overlapping Clusters”. In: *Proc. 11th Int. Conf. Artif. Intell. Stat.* 2, pp. 187–194.
- Herwig, Ralf et al. (2016). “Analyzing and interpreting genome data at the network level with ConsensusPathDB”. In: *Nature protocols* 11.10, p. 1889.
- Hirai, Masami Yokota et al. (2005). “Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics”. In: *Journal of Biological Chemistry*.
- Hoerl, Arthur E and Robert W Kennard (1970). “Ridge regression: Biased estimation for nonorthogonal problems”. In: *Technometrics* 12.1, pp. 55–67.
- Holliday, Robin (1990). “DNA methylation and epigenetic inheritance”. In: *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 326.1235, pp. 329–338.
- Holmes, Elaine et al. (2008). “Human metabolic phenotype diversity and its association with diet and blood pressure”. In: *Nature* 453.7193, p. 396.
- Hong, Shengjun et al. (2013). “Canonical correlation analysis for RNA-seq co-expression networks”. In: *Nucleic acids research* 41.8, e95–e95.
- Hornik, Kurt, Friedrich Leisch, and Achim Zeileis (2003). “JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling”. In: *Proceedings of DSC*. Vol. 2, pp. 1–1.
- Hotelling, H (1935). “Canonical correlation analysis (cca)”. In: *Journal of Educational Psychology*.
- (1936). “CCA: an r package to extend canonical correlation analysis”. In: *Biometrika* 28, pp. 321–377.
- Hotelling, Harold (1933). “Analysis of a complex of statistical variables into principal components.” In: *Journal of educational psychology* 24.6, p. 417.
- Hucka, Michael et al. (2003). “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models”. In: *Bioinformatics* 19.4, pp. 524–531.
- Irizarry, Rafael A et al. (2003). “Exploration, normalization, and summaries of high density oligonucleotide array probe level data”. In: *Biostatistics* 4.2, pp. 249–264.
- Jacquier, Eric, Nicholas G Polson, and Peter E Rossi (2002). “Bayesian analysis of stochastic volatility models”. In: *Journal of Business & Economic Statistics* 20.1, pp. 69–87.

- Jain, Pooja et al. (2018). “A multivariate approach to investigate the combined biological effects of multiple exposures”. In: *J Epidemiol Community Health* 72.7, pp. 564–571.
- Jendoubi, Takoua and Timothy Ebbels (2018). “Integrative analysis of time course metabolic data and biomarker discovery”. In: *arXiv preprint arXiv:1801.07767*.
- Jendoubi, Takoua and Korbinian Strimmer (2018). “A whitening approach to probabilistic canonical correlation analysis for omics data integration”. In: *arXiv preprint arXiv:1802.03490v2*.
- Ji, Yuan et al. (2011). “Glycine and a glycine dehydrogenase (GLDC) SNP as citalopram/escitalopram response biomarkers in depression: pharmacometabolomics-informed pharmacogenomics”. In: *Clinical Pharmacology & Therapeutics* 89.1, pp. 97–104.
- Josse, Julie and Susan Holmes (2016). “Measuring multivariate association and beyond”. In: *Statistics surveys* 10, p. 132.
- Joyce, Andrew R and Bernhard Ø Palsson (2006). “The model organism as a system: integrating ‘omics’ data sets”. In: *Nature reviews Molecular cell biology* 7.3, pp. 198–210.
- Kamburov, Atanas et al. (2011). “Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA”. In: *Bioinformatics* 27.20, pp. 2917–2918.
- Kandoth, Cyriac et al. (2013). “Mutational landscape and significance across 12 major cancer types”. In: *Nature*.
- Kanehisa, Minoru and Susumu Goto (2000). “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1, pp. 27–30.
- Kankainen, Matti et al. (2011). “MPEA—metabolite pathway enrichment analysis”. In: *Bioinformatics* 27.13, pp. 1878–1879.
- Karlebach, Guy and Ron Shamir (2008). “Modelling and analysis of gene regulatory networks”. In: *Nature Reviews Molecular Cell Biology* 9.10, p. 770.
- Karnovsky, Alla et al. (2011). “Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data”. In: *Bioinformatics* 28.3, pp. 373–380.
- Kell, Douglas (1999). “Screensavers: trends in high-throughput analysis”. In: *Trends in biotechnology* 17.3, pp. 89–91.
- Kessy, Agnan, Alex Lewin, and Korbinian Strimmer (2018). “Optimal whitening and decorrelation”. In: *The American Statistician*, pp. 1–6.

- Keun, Hector C et al. (2002). “Analytical reproducibility in ^1H NMR-based metabonomic urinalysis”. In: *Chemical research in toxicology* 15.11, pp. 1380–1386.
- Kim, Sangjoon, Neil Shephard, and Siddhartha Chib (1998). “Stochastic volatility: likelihood inference and comparison with ARCH models”. In: *The review of economic studies* 65.3, pp. 361–393.
- Kim, So Yeon et al. (2018). “Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer”. In: *BMC medical genomics* 11.3, p. 68.
- Kim, Sunghwan et al. (2015). “PubChem substance and compound databases”. In: *Nucleic acids research* 44.D1, pp. D1202–D1213.
- Kimura, Ikuo et al. (2013). “The gut microbiota suppresses insulin-mediated fat accumulation via the short-chain fatty acid receptor GPR43”. In: *Nature communications* 4, p. 1829.
- Kirk, Paul et al. (2012). “Bayesian correlated clustering to integrate multiple datasets”. In: *Bioinformatics* 28.24, pp. 3290–3297.
- Kishino, Shigenobu et al. (2013). “Polyunsaturated fatty acid saturation by gut lactic acid bacteria affecting host lipid composition”. In: *Proceedings of the National Academy of Sciences* 110.44, pp. 17808–17813.
- Klami, Arto and Samuel Kaski (2007). “Local dependent components”. In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 425–432.
- Kleemann, Robert et al. (2007). “Atherosclerosis and liver inflammation induced by increased dietary cholesterol intake: a combined transcriptomics and metabolomics analysis”. In: *Genome biology* 8.9, R200.
- Knaap, Jan A van der and C Peter Verrijzer (2016). “Undercover: gene control by metabolites and metabolic enzymes”. In: *Genes & development* 30.21, pp. 2345–2369.
- Koc, Ibrahim, Isa Yuksel, and Gustavo Caetano-Anollés (2018). “Metabolite-Centric Reporter Pathway and Tripartite Network Analysis of Arabidopsis Under Cold Stress”. In: *Frontiers in bioengineering and biotechnology* 6, p. 121.
- Kopczynski, Dominik, Albert Sickmann, and Robert Ahrends (2017). “Computational proteomics tools for identification and quality control”. In: *Journal of biotechnology* 261, pp. 126–130.

- Korf, Ian (2004). “Gene finding in novel genomes”. In: *BMC bioinformatics* 5.1, p. 59.
- Krumsiek, Jan et al. (2011). “Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data”. In: *BMC systems biology* 5.1, p. 21.
- Krumsiek, Jan et al. (2012). “Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information”. In: *PLoS genetics* 8.10, e1003005.
- Kuehne, Andreas et al. (2017). “Metabolic network segmentation: A probabilistic graphical modeling approach to identify the sites and sequential order of metabolic regulation from non-targeted metabolomics data”. In: *PLoS computational biology* 13.6, e1005577.
- Kuhl, Carsten et al. (2011). “CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets”. In: *Analytical chemistry* 84.1, pp. 283–289.
- Kullback, Solomon and Richard A Leibler (1951). “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1, pp. 79–86.
- Lai, Pei Ling and Colin Fyfe (2000). “Kernel and nonlinear canonical correlation analysis”. In: *International Journal of Neural Systems* 10.05, pp. 365–377.
- Lanckriet, Gert RG et al. (2004a). “A statistical framework for genomic data fusion”. In: *Bioinformatics* 20.16, pp. 2626–2635.
- Lanckriet, GRG et al. (2004b). “Kernel-based data fusion and its application to protein function prediction in yeast”. In: *Biocomputing 2004, Proceedings of the Pacific Symposium, Hawaii, USA*, pp. 300–311.
- Lê Cao, Kim-Anh et al. (2008). “A sparse PLS for variable selection when integrating omics data”. In: *Statistical applications in genetics and molecular biology* 7.1.
- Lê Cao, Kim-Anh et al. (2009). “Sparse canonical methods for biological data integration: application to a cross-platform study”. In: *BMC bioinformatics* 10.1, p. 34.
- Le Van, Thanh et al. (2016). “Simultaneous discovery of cancer subtypes and subtype features by molecular data integration”. In: *Bioinformatics* 32.17, pp. i445–i454.
- Lee, Duncan (2011). “A comparison of conditional autoregressive models used in Bayesian disease mapping”. In: *Spatial and Spatio-temporal Epidemiology* 2.2, pp. 79–89.

- (2013). “CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors”. In: *Journal of Statistical Software* 55.13, pp. 1–24.
- Lee, Duncan and Richard Mitchell (2013). “Locally adaptive spatial smoothing using conditional auto-regressive models”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 62.4, pp. 593–608.
- Li, Shuzhao et al. (2013). “Predicting network activity from high throughput metabolomics”. In: *PLoS computational biology* 9.7, e1003123.
- Lin, Dongdong et al. (2013). “Group sparse canonical correlation analysis for genomic data integration”. In: *BMC bioinformatics* 14.1, p. 245.
- Linda, PC et al. (2009). “A symmetrical tetramer for *S. aureus* pyruvate carboxylase in complex with coenzyme A”. In: *Structure* 17.6, pp. 823–832.
- Lindley, D. V. (1956). “On a Measure of the Information Provided by an Experiment”. In: *Ann. Math. Statist.* 27.4, pp. 986–1005. DOI: [10.1214/aoms/1177728069](https://doi.org/10.1214/aoms/1177728069). URL: <https://doi.org/10.1214/aoms/1177728069>.
- Lindon, John C, Jeremy K Nicholson, and Elaine Holmes (2011). *The handbook of metabonomics and metabolomics*. Elsevier.
- Liquet, Benoit et al. (2012). “A novel approach for biomarker selection and the integration of repeated measures experiments from two assays”. In: *BMC bioinformatics* 13.1, p. 325.
- Liquet, Benoit et al. (2015). “Group and sparse group partial least square approaches applied in genomics context”. In: *Bioinformatics* 32.1, pp. 35–42.
- Liu, Hongfang et al. (2010). “mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities”. In: *Molecular cancer therapeutics*, pp. 1535–7163.
- Löfstedt, Tommy and Johan Trygg (2011). “OnPLS—a novel multiblock method for the modelling of predictive and orthogonal variation”. In: *Journal of Chemometrics* 25.8, pp. 441–455.
- Lommen, Arjen (2009). “MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing”. In: *Analytical chemistry* 81.8, pp. 3079–3086.
- Lunn, David et al. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC press.

- Ma, Zhuang, Yichao Lu, and Dean Foster (2015). “Finding linear structure in large datasets with scalable canonical correlation analysis”. In: *International Conference on Machine Learning*, pp. 169–178.
- MacKay, David JC (1992). “Bayesian interpolation”. In: *Neural computation* 4.3, pp. 415–447.
- Maitre, L. et al. (2014). “Urinary metabolic profiles in early pregnancy are associated with preterm birth and fetal growth restriction in the Rhea mother–child cohort study”. In: *BMC Medicine* 12.
- Manikandan, P., D. Ramyachitra, and D. Banupriya (2016). “Detection of overlapping protein complexes in gene expression, phenotype and pathways of *Saccharomyces cerevisiae* using Prorank based Fuzzy algorithm”. In: *Gene* 580.2, pp. 144–158. ISSN: 03781119.
- Marco-Ramell, Anna et al. (2018). “Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data”. In: *BMC bioinformatics* 19.1, p. 1.
- Marshall, Darrell D et al. (2015). “Combining DI-ESI-MS and NMR datasets for metabolic profiling”. In: *Metabolomics* 11.2, pp. 391–402.
- Martin, Pascal G. P. et al. (2007). “Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study.” In: *Hepatology* 45 3, pp. 767–77.
- Mayer, Bernd (2011). *Bioinformatics for omics data: methods and protocols*. 57: 004 BIO. Springer.
- Mayer, Claus-Dieter, Julie Lorent, and Graham W. Horgan (2011). “Exploratory analysis of multiple omics datasets using the adjusted RV coefficient”. In: *Statistical Applications in Genetics and Molecular Biology* 10.1. DOI: [10.2202/1544-6115.1540](https://doi.org/10.2202/1544-6115.1540).
- Mei, Yajun, Seoung Bum Kim, and Kwok Leung Tsui (2009). “Linear-mixed effects models for feature selection in high-dimensional NMR spectra”. In: *Expert Systems with Applications* 36.3 PART 1, pp. 4703–4708. ISSN: 09574174. DOI: [10.1016/j.eswa.2008.06.032](https://doi.org/10.1016/j.eswa.2008.06.032). URL: <http://dx.doi.org/10.1016/j.eswa.2008.06.032>.
- Meier, Lukas, Sara Van De Geer, and Peter Bühlmann (2008). “The group lasso for logistic regression”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 53–71.

- Meijer, C. S. (1936). “Über Whittakersche bzw. Besselsche Funktionen und deren Produkte.” German. In: *Nieuw Arch. Wiskd., II. Ser.* 18.4, pp. 10–39. ISSN: 0028-9825.
- Mi, Huaiyu and Paul Thomas (2009). “PANTHER pathway: an ontology-based pathway database coupled with data analysis tools”. In: *Protein Networks and Pathway Analysis*. Springer, pp. 123–140.
- Mitchell, Toby J and John J Beauchamp (1988). “Bayesian variable selection in linear regression”. In: *Journal of the American Statistical Association* 83.404, pp. 1023–1032.
- Mitra, Riten and Peter Müller (2015). *Nonparametric Bayesian inference in biostatistics*. Springer.
- Morozova, Olena and Marco A Marra (2008). “Applications of next-generation sequencing technologies in functional genomics”. In: *Genomics* 92.5, pp. 255–264.
- Mostafavi, Nahid et al. (2017). “Associations between genome-wide gene expression and ambient nitrogen oxides”. In: *Epidemiology* 28.3, pp. 320–328.
- Moutselos, Konstantinos et al. (2009). “KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database”. In: *BMC bioinformatics* 10.1, p. 324.
- Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Nicholson, George et al. (2011). “A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection”. In: *PLoS genetics* 7.9, e1002270.
- Nicholson, Jeremy K, John C Lindon, and Elaine Holmes (1999). “‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data”. In: *Xenobiotica* 29.11, pp. 1181–1189.
- Nicholson, Jeremy K and Ian D Wilson (2003). “Understanding ‘global’ systems biology: metabonomics and the continuum of metabolism”. In: *Nature Reviews Drug Discovery* 2.8, p. 668.
- Niedzwiecki, Megan M et al. (2019). “The exposome: molecules to populations”. In: *Annual review of pharmacology and toxicology* 59, pp. 107–127.

- Nyamundanda, Gift, Isobel Claire Gormley, and Lorraine Brennan (2014). “A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63.5, pp. 763–782.
- Orbanz, Peter (2012). “Lecture Notes on Bayesian Nonparametrics”. In: *Journal of Mathematical Psychology* 56, pp. 1–12.
- Ormerod, John T, Chong You, Samuel Müller, et al. (2017). “A variational Bayes approach to variable selection”. In: *Electronic Journal of Statistics* 11.2, pp. 3549–3594.
- Paisley, John and Lawrence Carin (2009). “Nonparametric Factor Analysis with Beta Process Priors”. In: *Proc. 26th Int. Conf. Mach. Learn.* Pp. 1–8.
- Pan, Zhengzheng and Daniel Raftery (2007). “Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics”. In: *Analytical and bioanalytical chemistry* 387.2, pp. 525–527.
- Park, Trevor and George Casella (2008). “The bayesian lasso”. In: *Journal of the American Statistical Association* 103.482, pp. 681–686.
- Parkhomenko, Elena, David Tritchler, and Joseph Beyene (2009). “Sparse canonical correlation analysis with application to genomic data integration”. In: *Statistical applications in genetics and molecular biology* 8.1, pp. 1–34.
- Pearson, Karl (1901). “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.
- Pence, Harry E and Antony Williams (2010). *ChemSpider: an online chemical information resource*.
- Pham, Lisa M. et al. (2015). “Perturbation Detection Through Modeling of Gene Expression on a Latent Biological Pathway Network: A Bayesian hierarchical approach”. In: *Journal of the American Statistical Association* 1459.July 2016, pp. 1–61. ISSN: 0162-1459. DOI: [10.1080/01621459.2015.1110523](https://doi.org/10.1080/01621459.2015.1110523). arXiv: [arXiv:1409.0503v1](https://arxiv.org/abs/1409.0503v1). URL: <http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2015.1110523>{\#.VyfdxT99haU}.
- Pilalis, Eleftherios et al. (2015). “KENeV: A web-application for the automated reconstruction and visualization of the enriched metabolic and signaling super-pathways deriving from genomic experiments”. In: *Computational and structural biotechnology journal* 13, pp. 248–255.

- Plata, Germán et al. (2012). “Global probabilistic annotation of metabolic networks enables enzyme discovery”. In: *Nature chemical biology* 8.10, p. 848.
- Pluskal, Tomáš et al. (2010). “MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data”. In: *BMC bioinformatics* 11.1, p. 395.
- Polson, Nicholas G and James G Scott (2010). “Shrink globally, act locally: Sparse Bayesian regularization and prediction”. In: *Bayesian statistics* 9, pp. 501–538.
- Posma, Joram M et al. (2013). “MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG”. In: *Bioinformatics* 30.6, pp. 893–895.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Rantalainen, Mattias et al. (2006). “Statistically integrated metabonomic- proteomic studies on a human prostate cancer xenograft model in mice”. In: *Journal of proteome research* 5.10, pp. 2642–2655.
- Rantalainen, Mattias et al. (2008). “Piecewise multivariate modelling of sequential metabolic profiling data”. In: *BMC bioinformatics* 9.1, p. 105.
- Rasmussen, Carl E. (2000). “The Infinite Gaussian Mixture Model”. In: *Adv. Neural Inf. Process. Syst.* 12, pp. 554–560. ISSN: 10495258.
- Ratray, Nicholas JW et al. (2018). “Beyond genomics: understanding exposotypes through metabolomics”. In: *Human genomics* 12.1, p. 4.
- Redestig, Henning et al. (2010). “Consolidating metabolite identifiers to enable contextual and multi-platform metabolomics data analysis”. In: *BMC bioinformatics* 11.1, p. 214.
- Reo, Nicholas V (2002). “NMR-based metabolomics”. In: *Drug and chemical toxicology* 25.4, pp. 375–382.
- Richardson, Sylvia, Leonardo Bottolo, and Jeffrey S Rosenthal (2010). “Bayesian models for sparse regression analysis of high dimensional data”. In: *Bayesian Statistics* 9, pp. 539–569.
- Richardson, Sylvia, George C Tseng, and Wei Sun (2016). “Statistical methods in integrative genomics”. In:
- Richelle, Anne, Chintan Joshi, and Nathan E Lewis (2019). “Assessing key decisions for transcriptomic data integration in biochemical networks”. In: *PLoS computational biology* 15.7, e1007185.

- Riekeberg, Eli and Robert Powers (2017). “New frontiers in metabolomics: from measurement to insight”. In: *F1000Research* 6.
- Ritchie, Marylyn D et al. (2015). “Methods of integrating data to uncover genotype–phenotype interactions”. In: *Nature Reviews Genetics* 16.2, p. 85.
- Rob, A. C. et al. (2011). “Regulation of cancer cell metabolism”. In: *Nature Reviews Cancer* 11, pp. 85–95.
- Roberts, Peter C (2008). “Gene expression microarray data analysis demystified”. In: *Biotechnology annual review* 14, pp. 29–61.
- Robinette, Steven L, John C Lindon, and Jeremy K Nicholson (2013). “Statistical spectroscopic tools for biomarker discovery and systems medicine”. In: *Analytical chemistry* 85.11, pp. 5297–5303.
- Rodriguez-Martinez, Andrea et al. (2017). “MetaboSignal: a network-based approach for topological analysis of metabotype regulation via metabolic and signaling pathways”. In: *Bioinformatics* 33.5, pp. 773–775.
- Rogers, Simon et al. (2008). “Probabilistic assignment of formulas to mass peaks in metabolomics experiments”. In: *Bioinformatics* 25.4, pp. 512–518.
- Rosato, Antonio et al. (2018). “From correlation to causation: analysis of metabolomics data using systems biology approaches”. In: *Metabolomics* 14.4, p. 37.
- Ross, James C et al. (2014). “Dual beta process priors for latent cluster discovery in chronic obstructive pulmonary disease”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 155–162.
- Roverato, Alberto and Robert Castelo (2017). “The networked partial correlation and its application to the analysis of genetic interactions”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66.3, pp. 647–665.
- Rozeboom, William W. (1965). “Linear correlations between sets of variables”. In: *Psychometrika* 30.1, pp. 57–71. ISSN: 1860-0980.
- Rushworth, Alastair, Duncan Lee, and Christophe Sarran (2017). “An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk”. In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 66.1, pp. 141–157. ISSN: 14679876. DOI: [10.1111/rssc.12155](https://doi.org/10.1111/rssc.12155). arXiv: [1411.0924](https://arxiv.org/abs/1411.0924).
- Safo, Sandra E, Shuzhao Li, and Qi Long (2018). “Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis

- with incorporation of biological information”. In: *Biometrics* 74.1, pp. 300–312.
- Santos, Eduarda M et al. (2009). “Identifying health impacts of exposure to copper using transcriptomics and metabolomics in a fish model”. In: *Environmental science & technology* 44.2, pp. 820–826.
- Savage, Richard S et al. (2010). “Discovering transcriptional modules by Bayesian data integration”. In: *Bioinformatics* 26.12, pp. i158–i167.
- Savitzky, Abraham and Marcel JE Golay (1964). “Smoothing and differentiation of data by simplified least squares procedures.” In: *Analytical chemistry* 36.8, pp. 1627–1639.
- Schäfer, Juliane and Korbinian Strimmer (2005). “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics”. In: *Statistical applications in genetics and molecular biology* 4.1.
- Scheipl, Fabian (2011). “spikeSlabGAM: Bayesian Variable Selection, Model Choice and Regularization for Generalized Additive Mixed Models in R”. In: *Journal of Statistical Software* 43.14, pp. 1–24. URL: <http://www.jstatsoft.org/v43/i14/>.
- Scheipl, Fabian, Ludwig Fahrmeir, and Thomas Kneib (2012). “Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models”. In: *Journal of the American Statistical Association* 107.500, pp. 1518–1532. URL: <http://dx.doi.org/10.1080/01621459.2012.737742>.
- Scheltema, Richard A et al. (2011). “PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis”. In: *Analytical chemistry* 83.7, pp. 2786–2793.
- Schmidt, Michael W et al. (2007). “Comparative proteomic and transcriptomic profiling of the fission yeast *Schizosaccharomyces pombe*”. In: *Molecular systems biology* 3.1, p. 79.
- Schomburg, Ida et al. (2002). *BRENDA: a resource for enzyme data and metabolic information*.
- Sethuraman, Jayaram (1994). “A constructive definition of Dirichlet priors”. In: *Statistica sinica*, pp. 639–650.
- Shen, Ronglai, Adam B Olshen, and Marc Ladanyi (2009). “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis”. In: *Bioinformatics* 25.22, pp. 2906–2912.

- Shen, Ronglai et al. (2012). “Integrative subtype discovery in glioblastoma using iCluster”. In: *PloS one* 7.4, e35236.
- Shen-Orr, Shai S et al. (2002). “Network motifs in the transcriptional regulation network of *Escherichia coli*”. In: *Nature genetics* 31.1, p. 64.
- Shin, So-Youn et al. (2014). “An atlas of genetic influences on human blood metabolites”. In: *Nature genetics* 46.6, p. 543.
- Siddiqui, Jalal K et al. (2018). “IntLIM: integration using linear models of metabolomics and gene expression data”. In: *BMC bioinformatics* 19.1, p. 81.
- Silva, Ricardo R et al. (2014). “ProbMetab: an R package for Bayesian probabilistic annotation of LC–MS-based metabolomics”. In: *Bioinformatics* 30.9, pp. 1336–1337.
- Smilde, A. K. et al. (2009). “Matrix correlations for high-dimensional data: the modified RV-coefficient”. In: *Bioinformatics* 25.3, pp. 401–405. DOI: [10.1093/bioinformatics/btn634](https://doi.org/10.1093/bioinformatics/btn634). eprint: [/oup/backfile/content_public/journal/bioinformatics/25/3/10.1093/bioinformatics/btn634/2/btn634.pdf](http://oup/backfile/content_public/journal/bioinformatics/25/3/10.1093/bioinformatics/btn634/2/btn634.pdf). URL: <http://dx.doi.org/10.1093/bioinformatics/btn634>.
- Smilde, Age K et al. (2005). “ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data”. In: *Bioinformatics* 21.13, pp. 3043–3048.
- Smith, Colin A et al. (2005). “METLIN: a metabolite mass spectral database”. In: *Therapeutic drug monitoring* 27.6, pp. 747–751.
- Smith, Colin A et al. (2006). “XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification”. In: *Analytical chemistry* 78.3, pp. 779–787.
- Smolinska, Agnieszka et al. (2012). “Interpretation and visualization of nonlinear data fusion in kernel space: Study on metabolomic characterization of progression of multiple sclerosis”. In: *PLOS One* 7 (6).
- Snider, Ross K (2007). “Efficient calculation of exact mass isotopic distributions”. In: *Journal of the American Society for Mass Spectrometry* 18.8, pp. 1511–1515.
- Sriyudthsak, Kansuporn, Fumihide Shiraishi, and Masami Yokota Hirai (2013). “Identification of a metabolic reaction network from time-series data of metabolite concentrations”. In: *PLoS One* 8.1, e51212.
- Stan Development Team (2017). *RStan: the R interface to Stan*. R package version 2.17.2. URL: <http://mc-stan.org/>.

- Stanke, Mario and Burkhard Morgenstern (2005). “AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints”. In: *Nucleic acids research* 33.suppl_2, W465–W467.
- Stein, Lincoln (2001). “Genome annotation: from sequence to biology”. In: *Nature reviews genetics* 2.7, p. 493.
- Steuer, Ralf (2006). “On the analysis and interpretation of correlations in metabolomic data”. In: *Briefings in bioinformatics* 7.2, pp. 151–158.
- Steuer, Ralf et al. (2003). “Observing and interpreting correlations in metabolomic networks”. In: *Bioinformatics* 19.8, pp. 1019–1026.
- Stringhini, Silvia et al. (2015). “Life-course socioeconomic status and DNA methylation of genes regulating inflammation”. In: *International journal of epidemiology* 44.4, pp. 1320–1330.
- Sturm, Marc et al. (2008). “OpenMS—an open-source software framework for mass spectrometry”. In: *BMC bioinformatics* 9.1, p. 163.
- Su, Gang et al. (2011). “Integrated metabolome and transcriptome analysis of the NCI60 dataset”. In: *BMC bioinformatics* 12.1, S36.
- Suhre, Karsten and Philippe Schmitt-Kopplin (2008). “MassTRIX: mass translator into pathways”. In: *Nucleic acids research* 36.suppl_2, W481–W484.
- Suhre, Karsten et al. (2011a). “A genome-wide association study of metabolic traits in human urine”. In: *Nature genetics* 43.6, p. 565.
- Suhre, Karsten et al. (2011b). “Human metabolic individuality in biomedical and pharmaceutical research”. In: *Nature* 477.7362, p. 54.
- Sumner, Lloyd W et al. (2007). “Proposed minimum reporting standards for chemical analysis”. In: *Metabolomics* 3.3, pp. 211–221.
- Sun, Xuezheng et al. (2018). “Correlated metabolomic, genomic, and histologic phenotypes in histologically normal breast tissue”. In: *PloS one* 13.4, e0193792.
- Tamaddoni-Nezhad, A, S Muggleton, and J Bang (2003). “A bayesian model for metabolic pathways”. In: *International Joint Conference on Artificial Intelligence (IJCAI03) Workshop on Learning Statistical Models from Relational Data*, pp. 50–57.
- Tang, Xiaohu et al. (2014). “A joint analysis of metabolomics and genetics of breast cancer”. In: *Breast cancer research* 16.4, p. 415.
- Tautenhahn, Ralf, Christoph Boettcher, and Steffen Neumann (2008). “Highly sensitive feature detection for high resolution LC/MS”. In: *BMC bioinformatics* 9.1, p. 504.

- Teh, Y and Michael I Jordan (2009). “Hierarchical Bayesian nonparametric models with applications”. In: *Bayesian Nonparametrics*, pp. 1–48. ISSN: 0162-1459.
- Teh, Yee W et al. (2005). “Sharing clusters among related groups: Hierarchical Dirichlet processes”. In: *Advances in neural information processing systems*, pp. 1385–1392.
- Teh, Yee Whye et al. (2006). “Hierarchical Dirichlet Processes”. In: *Journal of the American Statistical Association* 101.476, pp. 1566–1581. ISSN: 0162-1459. DOI: [10.1198/016214506000000302](https://doi.org/10.1198/016214506000000302). URL: <http://www.tandfonline.com/doi/abs/10.1198/016214506000000302>.
- Theil, Henri and Ching-Fan Chung (1988). “Information-Theoretic Measures of Fit for Univariate and Multivariate Linear Regressions”. In: *The American Statistician* 42.4, pp. 249–252. ISSN: 00031305. URL: <http://www.jstor.org/stable/2685131>.
- Thibaux, R and M I Jordan (2007). “Hierarchical Beta Processes and the Indian Buffet Process”. In: *Proc. Elev. Int. Conf. Artif. Intell. Stat.* ISSN: 15324435.
- Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.
- Tipping, Michael E (2001). “Sparse Bayesian learning and the relevance vector machine”. In: *Journal of machine learning research* 1.Jun, pp. 211–244.
- Tipping, Michael E and Christopher M Bishop (1999). “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.
- Tiwari, Shrish et al. (1997). “Prediction of probable genes by Fourier analysis of genomic sequences”. In: *Bioinformatics* 13.3, pp. 263–270.
- Touloumis, Anestis (2015). “Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings”. In: *Computational Statistics & Data Analysis* 83, pp. 251–261.
- Trygg, Johan (2002). “O2-PLS for qualitative and quantitative analysis in multivariate calibration”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 16.6, pp. 283–293.
- Trygg, Johan and Svante Wold (2002). “Orthogonal projections to latent structures (O-PLS)”. In: *Journal of chemometrics* 16.3, pp. 119–128.
- (2003). “O2-PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter”. In: *Journal of Chemometrics* 17.1, pp. 53–64.

- Tseng, George, Debashis Ghosh, and Xianghong Jasmine Zhou (2015). *Integrating Omics Data*. Cambridge University Press.
- Valcárcel, Beatriz et al. (2014). “Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity”. In: *Journal of The Royal Society Interface* 11.94, p. 20130908.
- Van Der Pas, SL, BJK Kleijn, AW Van Der Vaart, et al. (2014). “The horseshoe estimator: Posterior concentration around nearly black vectors”. In: *Electronic Journal of Statistics* 8.2, pp. 2585–2618.
- Vaske, Charles J et al. (2010). “Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM”. In: *Bioinformatics* 26.12, pp. i237–i245.
- Verhoeckx, Kitty CM et al. (2004). “Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis”. In: *International Immunopharmacology* 4.12, pp. 1499–1514.
- Vineis, Paolo et al. (2013). “Advancing the application of omics-based biomarkers in environmental epidemiology”. In: *Environmental and molecular mutagenesis* 54.7, pp. 461–467.
- Virtanen, Seppo, Arto Klami, and Samuel Kaski (2011). “Bayesian CCA via group sparsity”. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, pp. 457–464.
- Vogel, Christine and Edward M Marcotte (2012). “Insights into the regulation of protein abundance from proteomic and transcriptomic analyses”. In: *Nature reviews genetics* 13.4, p. 227.
- Waaijenborg, Sandra, Philip C Verselewe de Witt Hamer, and Aeilko H Zwinderman (2008). “Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis”. In: *Statistical applications in genetics and molecular biology* 7.1.
- Wadi, Lina et al. (2016). “Impact of outdated gene annotations on pathway enrichment analysis”. In: *Nature methods* 13.9, p. 705.
- Walczak, B and DL Massart (1997). “Noise suppression and signal compression using the wavelet packet transform”. In: *Chemometrics and Intelligent Laboratory Systems* 36.2, pp. 81–94.

- Walther, Dirk et al. (2010). “Metabolic pathway relationships revealed by an integrative analysis of the transcriptional and metabolic temperature stress-response dynamics in yeast”. In: *Omics: a journal of integrative biology* 14.3, pp. 261–274.
- Wan, Ying-Wooi, Genevera I. Allen, and Zhandong Liu (2016). “TCGA2STAT: simple TCGA data access for integrated statistical analysis in R”. In: *Bioinformatics* 32.6, pp. 952–954. DOI: [10.1093/bioinformatics/btv677](https://doi.org/10.1093/bioinformatics/btv677). eprint: [/oup/backfile/content_public/journal/bioinformatics/32/6/10.1093_bioinformatics_btv677/4/btv677.pdf](https://oup/backfile/content_public/journal/bioinformatics/32/6/10.1093_bioinformatics_btv677/4/btv677.pdf). URL: <http://dx.doi.org/10.1093/bioinformatics/btv677>.
- Wang, Chong (2007). “Variational Bayesian approach to canonical correlation analysis”. In: *IEEE Transactions on Neural Networks* 18.3, pp. 905–910.
- Wang, Wenting et al. (2012). “iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data”. In: *Bioinformatics* 29.2, pp. 149–159.
- Watanabe, Sumio (2010). “Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory”. In: *Journal of Machine Learning Research* 11.Dec, pp. 3571–3594.
- Weber, Ralf JM and Mark R Viant (2010). “MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways”. In: *Chemometrics and Intelligent Laboratory Systems* 104.1, pp. 75–82.
- Wei, Xiaoli et al. (2011). “MetSign: a computational platform for high-resolution mass spectrometry-based metabolomics”. In: *Analytical chemistry* 83.20, pp. 7668–7675.
- Westerhuis, Johan A et al. (2010). “Multivariate paired data analysis: multilevel PLSDA versus OPLSDA”. In: *Metabolomics* 6.1, pp. 119–128.
- Wilms, Ines and Christophe Croux (2015). “Sparse canonical correlation analysis from a predictive point of view”. In: *Biometrical Journal* 57.5, pp. 834–851.
- (2016). “Robust sparse canonical correlation analysis”. In: *BMC systems biology* 10.1, p. 72.
- Winning, H. et al. (2009). “An exploratory NMR nutri-metabonomic investigation reveals dimethyl sulfone as a dietary biomarker for onion intake”. In: *Analyst* 134, pp. 2344–2351.

- Wishart, David S (2009). “Computational strategies for metabolite identification in metabolomics”. In: *Bioanalysis* 1.9, pp. 1579–1596.
- Wishart, David S et al. (2012). “HMDB 3.0—the human metabolome database in 2013”. In: *Nucleic acids research* 41.D1, pp. D801–D807.
- Witten, Daniela M and Robert J Tibshirani (2009). “Extensions of sparse canonical correlation analysis with applications to genomic data”. In: *Statistical applications in genetics and molecular biology* 8.1, pp. 1–27.
- Wold, Herman (1975). “Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach”. In: *Journal of Applied Probability* 12.S1, pp. 117–142.
- Wold, Svante, Harold Martens, and H Wold (1983). “The multivariate calibration problem in chemistry solved by the PLS method”. In: *Matrix pencils*, pp. 286–293.
- Wold, Svante, Michael Sjöström, and Lennart Eriksson (2001). “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and intelligent laboratory systems* 58.2, pp. 109–130.
- Wold, Svante et al. (1984). “The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses”. In: *SIAM Journal on Scientific and Statistical Computing* 5.3, pp. 735–743.
- Wu, Lani F et al. (2002). “Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters”. In: *Nature genetics* 31.3, p. 255.
- Xi, Yuanxin and David M Rocke (2008). “Baseline correction for NMR spectroscopic metabolomics data analysis”. In: *BMC bioinformatics* 9.1, p. 324.
- Xia, Jianguo, Erin E Gill, and Robert EW Hancock (2015). “NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data”. In: *Nature protocols* 10.6, p. 823.
- Xia, Jianguo and David S Wishart (2010). “MetPA: a web-based metabolomics tool for pathway analysis and visualization”. In: *Bioinformatics* 26.18, pp. 2342–2344.
- Xie, Xuanli Lisa and Gerardo Beni (1991). “A validity measure for fuzzy clustering”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 8, pp. 841–847.
- Yanai, H (1974). “Unification of various techniques of multivariate analysis by means of generalized coefficient of determination (GCD)”. In: *J. Behavior-metrics* 1.1, pp. 45–54.

- Yau, Christopher and Chris Holmes (2011). “Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination”. In: *Bayesian Analysis* 6.2, pp. 329–352. ISSN: 19360975. DOI: [10.1214/11-BA612](https://doi.org/10.1214/11-BA612).
- Yi, G, S H Sze, and M R Thon (2007). “Identifying clusters of functionally related genes in genomes”. In: *Bioinformatics* 23.9, pp. 1053–1060. ISSN: 1367-4811.
- Yuan, Yinyin, Richard S Savage, and Florian Markowetz (2011). “Patient-specific data fusion defines prognostic cancer subtypes”. In: *PLoS computational biology* 7.10, e1002227.
- Yugi, Katsuyuki et al. (2016). “Trans-omics: how to reconstruct biochemical networks across multiple ‘omic’ layers”. In: *Trends in biotechnology* 34.4, pp. 276–290.
- Zhang, Wenchao et al. (2015). “MET-XAlign: a metabolite cross-alignment tool for LC/MS-based comparative metabolomics”. In: *Analytical chemistry* 87.18, pp. 9114–9119.
- Zheng, Cheng et al. (2011). “Identification and quantification of metabolites in ¹H NMR spectra by Bayesian model selection”. In: *Bioinformatics* 27.12, pp. 1637–1644.
- Žitnik, Marinka and Blaž Zupan (2015). “Data fusion by matrix factorization”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.1, pp. 41–53.
- Zou, Hui and Trevor Hastie (2005). “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.
- Zuber, Verena, A Pedro Duarte Silva, and Korbinian Strimmer (2012). “A novel algorithm for simultaneous SNP selection in high-dimensional genome-wide association studies”. In: *BMC bioinformatics* 13.1, p. 284.
- Zuber, Verena and Korbinian Strimmer (2011). “High-dimensional regression and variable selection using CAR scores”. In: *Statistical Applications in Genetics and Molecular Biology* 10.1.