Integration of heterogeneous data: a multi-omics application

Ana Gabriela Pereira de Vasconcelos

MASTER THESIS PRESENTED TO THE
INSTITUTE OF MATHEMATICS AND STATISTICS
OF THE UNIVERSITY OF SÃO PAULO
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCES

Program: Statistics

Advisor: Prof. Dr. Júlia Maria Pavan Soler

The author acknowledges financial support provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

São Paulo, August 2020

Integration of heterogeneous data: a multi-omics application

This is the corrected version of the master thesis
by Ana Gabriela Pereira de Vasconcelos,
as suggested by the committee members
during the defense of the original document
on August 25, 2020.

Committee members:

- Profa. Dra. Júlia Maria Pavan Soler - IME-USP
- Profa. Dra. Joanlise Marco de Leon Andrade - EST-UNB
- Profa. Dra. Mariza de Andrade - Mayo Clinic

Acknowledgments

I would like to, not only acknowledge, but also dedicate this project to my parents. To my dad, who is the reason why I am so interested in biostatistics, and is my daily motivation to fight each battle and follow my dreams. And to my mom, my main inspiration, the person responsible for making me the woman I am. Also, I would like to thank my friends, specially my family from São Paulo, for always being there for me and being part of the best moment of my life so far. Thank you also to my partner Paulo, for being always so patient and caring with me through this whole process.

Next I would like to specially acknowledge Prof. Júlia Maria Pavan Soler, a woman with the greatest heart I have met, always full of joy and concerned with everyone around. Thank you for believing in me since the first moment and instigating me to try for the master's program at USP. Thank you for all the guidance and help throughout the past two years. Also, a special thanks for Dra. Silvia Titan for providing the data and being always open and willing to help and teach me more about CKD and the data. In addition, thank you for the judging committee for taking a time to read and assist on this projects.

Lastly but not least, I would like to thank my professor from University of Brasília and University of São Paulo for providing me the best of knowledge to allow me to finish this project. Also I would like to acknowledge CAPES for the financial support.

Resumo

Vasconcelos, Ana G. P. **Integração de dados heterogêneos: uma aplicação em dados multi-ômicos**. 2020. Dissertação de Mestrado - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Atualmente, uma enorme quantidade de dados tem sido coletada em diversas áreas do conhecimento, como saúde, agropecuária, marketing, fazendo com que dados de alta dimensão se tornem cada vez mais comuns. Mais especificamente, com os avanços da tecnologia muitas informações biológicas estão disponíveis por preços acessíveis – como dados do genoma, miRNA (micro RNA), mRNA (RNA mensageiro), expressão gênica e proteica, metilação, lipídeos, metabólicos e de fenótipos, por exemplo. Diversos estudos têm sido feitos para análise de cada tipo de dados individualmente, entretanto, recentemente vem se tornando interessante integrar diferentes tipos de dados para obter mais informação. Porém, muitas das metodologias clássicas utilizadas com esse objetivo assumem que a matriz de dados é completa e numérica. Portanto, a heterogeneidade de dados com variáveis de diversos tipos não está sendo considerada. Alternativamente, os *Generalized Low Rank Models (GLRM)* são modelos capazes de lidar com grandes bancos de dados com variáveis heterogêneas. Apesar desse método ser destinado para um único banco de dados, mostramos neste trabalho que ele é flexível o bastante para lidar com dados abstratos, de diferentes fontes, ao atribuir funções perdas diferentes, adequadas para cada tipo de variável. Com isso, o GLRM é uma ferramenta para trabalhar com problemas de diversas naturezas, mas, por ser muito recente, seu potencial para trabalhar com dados multi-ômicos ainda está sendo descoberto. Neste contexto, no presente trabalho O GRLM é introduzido e são exploradas diferentes possibilidades de usar o GLRM para redução de dimensionalidade e integração de bancos de dados em análises supervisionadas e não supervisionadas utilizando dados multi-ômicos simulados e reais.

Palavras-chave: dados multi-ômicos, generalized low rank models, fatorização de matrizes, análise multivariada.

Abstract

Vasconcelos, Ana G. P. **Integration of heterogeneous data: a multi-omics application**. 2020. Masters thesis - Institute of Mathematics and Statistics, University of São Paulo, São Paulo.

Nowadays, a huge amount of data has being collected in different research areas, such as public health, agriculture, marketing, so high-dimension databases are becoming very common to encounter. More specifically, with the advance of technology many biological information are now available at low costs – data from genome, miRNA (MicroRNA), mRNA (messenger RNA), gene expression, protein, methylation, lipids, metabolism, phenotypes and so on. Several different studies have been done individually with each type of data, but more recently there is an increasingly interest in integrating different data to gather more information. However, many classical methodologies used to this end assume the data matrix to be completed and numerical. Therefore, the heterogeneity of dataset with different variable types is not considered. Alternatively, the Generalized Low Rank Models (GLRM) is a tool capable of dealing with large datasets of heterogeneous data. Although its use is destined for a single database, this projects shows that it is flexible enough to handle abstract data, from different sources, by using different loss functions, adequate to each variable type. GLRM is a very powerful method that can deal with problems from different natures, but it is very recent, so its potential to work with multi-omics is still being discovered. In this context, the present work introduces GLRM and explores its possibilities for dimensionality reduction on supervised and unsupervised analysis using simulated and real multi-omics datasets.

Keywords: multi-omics, generalized low rank models, matrix factorization, multivariate analysis.

Contents

Chapter 1

Introduction

Nowadays, a huge amount of data is being collected in several areas, such as health, agriculture and marketing, so high-dimension databases are becoming very common. More specifically, with the advance of technology many biological information are now available at low costs, for example, the omics data. Omics studies focus on a global assessment of a set of molecules(see Hasin, Seldin, and Lusis, 2017, Box 1 for more references). The first and most common omics field are the *Genomics*. This field studies the genome - complete set of DNA - and focuses on identifying genetic variants associated with diseases, responses to treatment or future patience diagnosis, among others. To do so it deals with data of Single Nucleotide Polymorphisms (SNPs) [1] or Copy Number Variate (CNVs) [2]. *Epigenomic* uses different strategies for genome-wide characterization of reversible modifications of DNA or DNA-associated proteins. These modifications are regulators of gene transcription (information going from DNA to RNA) and could indicate a disease status. One example of epigenomic data are DNA methylation, in which a methyl group is added to the DNA modifying the gene expression, but without changing the sequence. A possible variable quantifying the methylation event is the percentage of methylation at each genomic site. Next, another omics field is the *Transcriptomics*. It exams RNA levels in a genome-wide dimension, either verifying which transcripts are present, by the number of RNA sequence present (RNAseq), or how much of each transcript is being expressed (gene expression), for example. Another omics class of interest is the *Proteomics*, that studies the set of proteins in an organism. It quantifies peptide abundance, modification and interaction, which may reflect in some biological process related to some disease. Finally, *Metabolomics* is another omics platform drived to quantify molecules – carbohydrates or amino acids, for example – that reflect metabolic functions, and could indicate disease status if out of range. There are also other types of omics studies such as *Lipidomics* (quantify lipids), *Microbiomics* (quantify microorganism), *Phenomics* (quantify phenotypes) and *Foodomics* (quantify food and nutrition).

On the last decades, many different studies have been conducted individually, with each omics

[1] One SNP is a nucleotide mutation that occurs on at least 1% of the population. The variable codifying the SNP genotyping information has values 0, 1 and 2 to indicate the number of the allele of minor frequency in a genomic locus, called as homozygous for the major allele frequency, heterozygous and homozygous for the minor allele frequency, respectively.

[2] CNV indicates the number of copies of a gene that an individual has in a genomic region.

data separately, and showed potential to understand and explain certain diseases. However, common and complex diseases does not result from changes on only one omics layer; they usually result from an intricate system, involving different types of association, interaction effects, not only between different omics, but also with the environment. Therefore, recently there is an increasing interest in combining different data to gather more information to better explain diseases. The idea of data integration is natural due to the known relations between omics layers. One example of this relationship is explained by the central dogma of molecular biology (Crick, 1970). This dogma basically describes the flow of information from DNA to RNA, to protein and to phenotypes. So there's a factual relation among genomics, transcriptomics and proteomics. This is a well known and accepted idea, however in practice the relation among omics is not necessarily linear and can be a lot more complex.

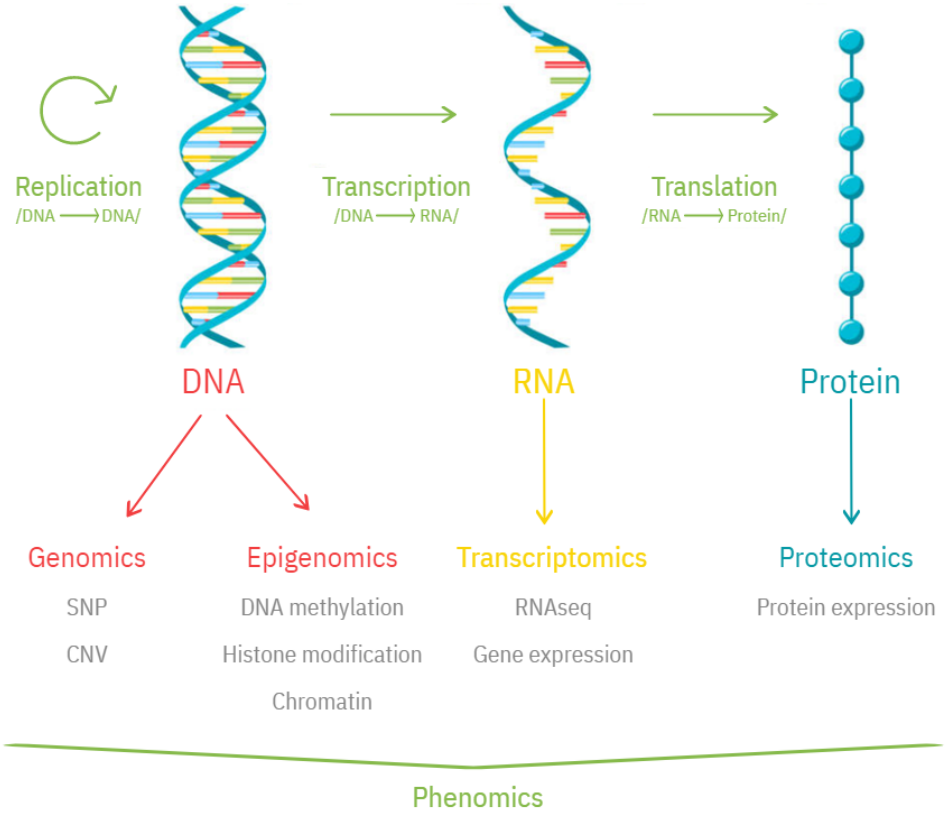


Figure 1.1: *The central dogma of biology and multi-omics data.*
Source: elaborated by the author.

1.1 Data integration

When dealing with data integration there are two main ways to do it. The first is the p-integration, where independent datasets or studies are available for different individuals, but with the same p variables, which is the case of multi-center studies, for example. The second is the n-integration, where each dataset provides a different information, but all for the same individuals. An example is the integration of multi-omics data, in which there are different datasets, one for

each omics platform, all observed for the same patients. Throughout this project the term *integration* will be used to represent an n-integration situation, although some methods could be applied on p-integration cases.

Different techniques can be used to integrate multiple dataset depending whether the analysis is supervised or unsupervised. Unsupervised analysis focuses on understanding the relationship among datasets and grouping individuals, while also performing dimension reduction and variable selection in high-dimensional data. When dealing with multi-omics, unsupervised integration techniques could be used to identify novel phenotypic clusters (Wang et al., 2014; Kirk et al., 2012; S. Zhang, C.-C. Liu, et al., 2012; Lock et al., 2013; W. Li et al., 2012; A. Tenenhaus and M. Tenenhaus, 2011; A. Tenenhaus, Philippe, et al., 2014,Argelaguet et al., 2018), or to reconstruct and understand biological networks (Glass et al., 2013; S. Zhang, Q. Li, et al., 2011, Zhu et al., 2012), for example. In contrast, supervised analysis approaches have the interest of classifying sample groups and predicting a class of new samples, based on a response variable or outcome. Much has been done in a multi-omics context, specially regarding disease studies. To summarize, there are two main approaches: multi-staged and meta-dimensional analysis. Each analysis reflects an assumption made on the disease etiology (Marylyn D. Ritchie et al., 2015).

The **Multi-staged analysis** tries to identify the relationship in multiple stages: firstly between the data, then between the data and the phenotype. It has a linear assumption of disease etiology, where a variation in DNA leads to a variation in RNA, that leads to variation in protein, that leads to variation on the phenotype (Figure 1.2), for example.

Figure 1.2: *Representation of linear assumption of disease etiology.*
Source: reproduction of Figure 2 by Marylyn D. Ritchie et al., 2015.

In general, there are three main methods to perform multi-stage analysis. The *Genomic Variation analysis* approach uses the triangle method that consists on: first identifying SNPs associated with the phenotype and filtering them based on genome-wide significant threshold; then the significant SNPs are tested for association with gene expression; finally, these gene expression levels are tested for correlation with the phenotype of interest (Holzinger, Grady, et al., 2012). The second method is the *Domain Knowledge-guided approach*, which is similar to the triangle approach. The difference is that, on the first step, the SNPs are annotated with knowledge from external database resources, and only those overlapping with functional units are taken to the next steps. And lastly, is the *Allele-specific Expression approach*. It consists on analyzing the products of each parental allele and compare their association with gene expression or methylation. The resulting allele is then tested for correlation with the phenotype of interest. These approaches have the advantage of modelling causal relationships between multi-omics data, as it has been done by Schadt et al., 2005. They are fairly powerful, as long as the linear hypothesis holds. However, this hypothesis fails to effectively model complex phenotypes.

Alternatively, the **Meta-dimensional analysis** combines different data types in a simultaneous analysis associated with an phenotype. The approach reflects the hypothesis that different level of omics interact in a nonlinear and complex way, and the combination of variation across all layers contribute to the phenotype variation (Figure 1.3).

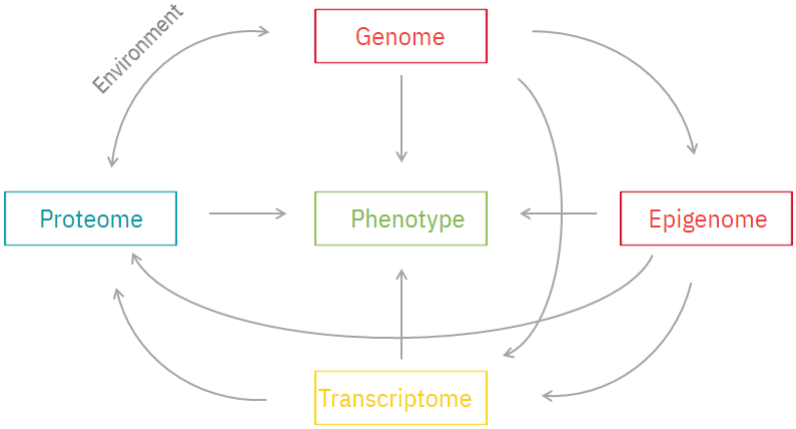


Figure 1.3: *Representation of complex and non linear assumption of disease etiology.*
Source: reproduction of Figure 2 by Marylyn D. Ritchie et al., 2015.

There are several different approaches to perform meta-dimensional analysis. *Concatenation-based integration* combines all datasets into one matrix prior to applying a classification model. It has the advantage of being useful to consider interaction among different type of data, and once the single matrix is assembled it is easy to apply any statistical method for analysis. Some possible statistical techniques to apply are: Bayesian integrative models to predict phenotype (Fridley et al., 2012), Cox regression to predict time recurrence and survival (Mankoo et al., 2011) or factor and discriminant analysis (Y. Liu et al., 2013). There is also a multi functional software called The Analysis Tool for Heritable and Environmental Network Associations (ATHENA), that carries meta-dimensional analysis with statistical techniques such as symbolic regression, Support Vector Machine (SVM), Bayesian Networks and Neural Networks (Holzinger, Dudek, Frase, Pendergrass, et al., 2014; Holzinger, Dudek, Frase, Krauss, et al., 2013; Kim et al., 2013). However, a big challenge of this approach is to find the best way to combine multiple datasets of different data types, without driving bias due to it.

In contrast, the *Transformation-based integration* first transforms each data type into a graph or kernel matrix, then combines the multiple data. It has the advantage of preserving data-type specific properties, so it is robust to different scales. This technique has already been used to predict protein function in different contexts: via kernel-based integration (Lanckriet et al., 2004), graph kernels (Borgwardt et al., 2005) and graph-based semi-supervised learning (Tsuda, Shin, and Schölkopf, 2005). ATHENA also performs graph-based integration to predict cancer clinical outcomes (Kim et al., 2013), for example. The disadvantage of transformation-based integration is that it can be difficult to identify interaction between different datasets, since each is transformed independently. To guarantee that these interactions are not missed, it is necessary to find transformations that maintain most of data type specific properties.

Finally, there is also *model-based integration* where each dataset is used as training sets for a model independently, and a final integrative model is generated from the multiple models. This method has a huge advantage of being able to integrate dataset collected from different sets of individuals (all with the same outcome of interest). It is also useful when data are so heterogeneous that they cannot be transformed or concatenated. Some examples of approaches are: the *majority voting approach* (Drăghici and Potter, 2003), where the final model is determined by the one that preforms better among multiples models constructed; *ensemble classifiers*, constructed by using multiple learning methods (Mankoo et al., 2011; Günther et al., 2012); *network-based approach*, where multiple networks can be combined to construct probabilistic causal networks, for example (Akavia et al., 2010). A more recent approach is the DIABLO framework (Rohart et al., 2017; Singh et al., 2018), which transforms each dataset into latent components and maximizes the correlation between them. However, when applying models to each dataset, all the specific hypothesis of each analysis must be considered, and it is necessary to find a meaningful way to combine all the models, and, in general, it only allows for analysis of quantitative datasets.

1.2 Heterogeneous data

All the approaches described previously have the same common challenge: dealing with datasets with different data types. Some may deal with this problem better than others. However most of them can only work with continuous and discrete data. In practice, each omics dataset comes from different platforms and represents different biological processes. For example, genomics data could be represented as a counting variable (Copy Number Variations) or discrete (Single Nucleotide Polymorphism), as well as, trinomial. Gene expression values generally follow asymmetric distributions. Proteomics data can describe the gene expression, which is an intensity (real values), or a RNAseq, that gives counts of different RNA sequences, and usually has a Poisson or Binomial Negative Distribution. The solution usually involves representing the data in a quantitative way, log-transformed, when possible, and performing normalization to guarantee that all the data is at the same scale. Although this solution is an alternative to allow the analysis with classical methodologies, in many cases this transformation is not possible nor correct.

Thus, this issue has gained interest lately, and several studies propose methodologies to maintain the original scale of variables. Chavent et al., 2014 extends Principal Component Analysis and Factor Analysis to deal with mixture of numerical and categorical variables in a single dataset. Bonat, 2017 deals with multivariate non-Gaussian outcomes in a statistical genetics context. He deals with different data types by using covariance link functions to specify the marginal covariance structure of multivariate covariance generalized linear models (McGLM). Song, 2019 proposes a generalization of simultaneous component analysis (SCA) to an exponential family simultaneous component analysis (ESCA) for unsupervised data integration.

Another novel methodology that focuses on dealing with different data types, outside the context of omics, is the Generalized Low Rank Models (GLRM), developed by Udell et al., 2016. This

model is capable of dealing with large datasets, with different types of variables and also missing data. It basically decomposes the dataset into numerical matrices and can also handle abstract data by using different loss functions, adequate to each variable type. However, even though GLRM is a very powerful tool that can deal with many different problems, it is very recent; so only a few people have used it to try to solve old problems, including dealing with multi-omics data. Hu et al., 2020 used GLRM to segment gene expression data in pathways and clusters for further inferring omics causal networks to find common paths from genetic variants to Alzheimer's dementia and Type 2 diabetes. GLRM was also used for feature selection, missing and imputing missing data to identify subtypes of disease for phenotype characterization of traumatic brain injury (Masino and Folweiler, 2018) and autism spectrum disorders (Paskov and Wall, 2018). Also, Xiong, 2018a and Xiong, 2018b have introduced and applied GLRM in the context of eQTL and eQTL epistasis analysis to understand gene regulation, representation of groups of gene expression, and for gene expression deconvolution to identify constituent cell types in a tissue. This shows that GLRM can be applied to studies that used only one type of omics, however, these references have not discussed the potential use of GLRM for data integration.

Therefore, this project focuses on exploring the GLRM in a multi-omics integration context. As it is a methodology used only for one dataset it is a meta-dimensional, concatenation and component-based integration that can be used in a supervised and unsupervised analysis. First, on Chapter 2 we will go through Principal Component Analysis, which is the basis for many multivariate techniques, followed by methods for supervised analysis (Discriminant Analysis) and unsupervised data integration (Canonical Correlation Analysis). Then, generalizations for dealing with high-dimensional databases will be presented to introduce the idea of regularizers. Finally the methodology of Generalized Low Rank Models will be briefly explained, from the optimization problem to how to choose parameters. On chapter 3 the GLRM will be applied on two datasets. The first is a simulated dataset of a hypothetical breast cancer pathway (Chung and Kang, 2019), and the second is a real Chronic Kidney Disease (CKD) by the PROGREDIR cohort study (Domingos et al., 2017). The simulated data will be used to evaluate how well GLRM can be used to integrate multiple omics datasets to predict a disease outcome. The GLRM will be compared with other integration methods based on their predictive abilities. Alternatively, the CKD dataset will be used to explore different applications of GLRM, including understand the impact of missing data on metabolomics, integrate variables to gain information for data imputation and obtain latent variables for associating it with mortality risk. Finally, Chapter 4 presents a discussion about the results presented on Chapter 3 and considerations for future works.

References

Akavia, Uri David et al. (2010). "An integrated approach to uncover drivers of cancer". In: *Cell* 143.6, pp. 1005–1017.

Argelaguet, Ricard et al. (2018). "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets". In: *Molecular systems biology* 14.6.

Bonat, Wagner Hugo (2017). "Modelling Mixed Types of Outcomes in Additive Genetic Models". In: *The international journal of biostatistics* 13.2, pp. 1–16. ISSN: 15574679. DOI: 10.1515/ijb-2017-0001.

Borgwardt, Karsten M et al. (2005). "Protein function prediction via graph kernels". In: *Bioinformatics* 21.suppl_1, pp. i47–i56.

Chavent, Marie et al. (2014). "Multivariate Analysis of Mixed Data: The R Package PCAmixdata". In: 4. arXiv: 1411.4911. URL: http://arxiv.org/abs/1411.4911.

Chung, Ren Hua and Chen Yu Kang (2019). "A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification". In: *GigaScience* 8.5, pp. 1–12. ISSN: 2047217X. DOI: 10.1093/gigascience/giz045.

Crick, Francis (1970). "Central dogma of molecular biology". In: *Nature* 227.5258, p. 561.

Domingos, Maria Alice Muniz et al. (2017). "Doença renal crônica-determinantes de progressão e risco cardiovascular. Coorte PROGREDIR: desenho de estudo e métodos". In: *Sao Paulo Medical Journal* 135.2, pp. 133–139. ISSN: 15163180. DOI: 10.1590/1516-3180.2016.0272261116.

Drăghici, Sorin and R Brian Potter (2003). "Predicting HIV drug resistance with neural networks". In: *Bioinformatics* 19.1, pp. 98–107.

Fridley, Brooke L et al. (2012). "AB ayesian Integrative Genomic Model for Pathway Analysis of Complex Traits". In: *Genetic epidemiology* 36.4, pp. 352–359.

Glass, Kimberly et al. (2013). "Passing messages between biological networks to refine predicted interactions". In: *PloS one* 8.5.

Günther, Oliver P. et al. (2012). "A computational pipeline for the development of multi-marker bio-signature panels and ensemble classifiers". In: *BMC Bioinformatics* 13.1. ISSN: 14712105. DOI: 10.1186/1471-2105-13-326.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis (2017). "Multi-omics approaches to disease". In: *Genome Biology* 18.1, pp. 1–15. ISSN: 1474760X. DOI: 10.1186/s13059-017-1215-1.

Holzinger, Emily R, Scott M Dudek, Alex T Frase, Ronald M Krauss, et al. (2013). "ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels". In: *Biocomputing 2013*. World Scientific, pp. 385–396.

Holzinger, Emily R, Scott M Dudek, Alex T Frase, Sarah A Pendergrass, et al. (2014). "ATHENA: the analysis tool for heritable and environmental network associations". In: *Bioinformatics* 30.5, pp. 698–705.

Holzinger, Emily R, Benjamin Grady, et al. (2012). "Genome-wide association study of plasma efavirenz pharmacokinetics in AIDS Clinical Trials Group protocols implicates several CYP2B6 variants". In: *Pharmacogenetics and genomics* 22.12, p. 858.

Hu, Zixin et al. (2020). “Shared Causal Paths underlying Alzheimer’s dementia and Type 2 Diabetes”. In: *Scientific Reports* 10.1, pp. 1–15. ISSN: 20452322. DOI: 10.1038/s41598-020-60682-3.

Kim, Dokyoon et al. (2013). “ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network”. In: *BioData mining* 6.1, p. 23.

Kirk, Paul et al. (2012). “Bayesian correlated clustering to integrate multiple datasets”. In: *Bioinformatics* 28.24, pp. 3290–3297.

Lanckriet, Gert RG et al. (2004). “A statistical framework for genomic data fusion”. In: *Bioinformatics* 20.16, pp. 2626–2635.

Li, Wenyuan et al. (2012). “Identifying multi-layer gene regulatory modules from multi-dimensional genomic data”. In: *Bioinformatics* 28.19, pp. 2458–2466.

Liu, Yuanhua et al. (2013). “Multilevel omic data integration in cancer cell lines: Advanced annotation and emergent properties”. In: *BMC Systems Biology* 7, pp. 1–13. ISSN: 17520509. DOI: 10.1186/1752-0509-7-14.

Lock, Eric F et al. (2013). “Joint and individual variation explained (JIVE) for integrated analysis of multiple data types”. In: *The annals of applied statistics* 7.1, p. 523.

Mankoo, Parminder K et al. (2011). “Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles”. In: *PloS one* 6.11.

Masino, Aaron J. and Kaitlin A. Folweiler (2018). “Unsupervised learning with GLRM feature selection reveals novel traumatic brain injury phenotypes”. In: arXiv: 1812.00030. URL: http://arxiv.org/abs/1812.00030.

Paskov, Kelley M and Dennis P Wall (2018). “A Low Rank Model for Phenotype Imputation in Autism Spectrum Disorder.” In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* 2017, pp. 178–187. ISSN: 2153-4063.

Ritchie, Marylyn D. et al. (2015). “Methods of integrating data to uncover genotype-phenotype interactions”. In: *Nature Reviews Genetics* 16.2, pp. 85–97. ISSN: 14710064. DOI: 10.1038/nrg3868.

Rohart, Florian et al. (2017). “mixOmics: An R package for ‘omics feature selection and multiple data integration”. In: *PLoS Computational Biology* 13.11, pp. 1–14. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005752LK. URL: http://www.embase.com/search/results?subaction=viewrecord%7B%5C&%7Dfrom=export%7B%5C&%7Did=L619520782%7B%5C%%7D0Ahttp://dx.doi.org/10.1371/journal.pcbi.1005752.

Schadt, Eric E et al. (2005). “An integrative genomics approach to infer causal associations between gene expression and disease”. In: *Nature genetics* 37.7, pp. 710–717.

Singh, Amrit et al. (2018). “DIABLO: from multi-omics assays to biomarker discovery, an integrative approach”. In: *bioRxiv* 61, p. 067611.

Song, Yipeng (2019). “Fusing heterogeneous data sets”. PhD thesis. ISBN: 9789463755313. arXiv: 1908.09653. URL: http://arxiv.org/abs/1908.09653.

Tenenhaus, Arthur, Cathy Philippe, et al. (2014). "Variable selection for generalized canonical correlation analysis". In: *Biostatistics* 15.3, pp. 569–583. ISSN: 14684357. DOI: 10.1093/biostatistics/kxu001.

Tenenhaus, Arthur and Michel Tenenhaus (2011). "Regularized Generalized Canonical Correlation Analysis". In: *Psychometrika* 76.2, pp. 257–284. ISSN: 00333123. DOI: 10.1007/s11336-011-9206-8.

Tsuda, Koji, Hyunjung Shin, and Bernhard Schölkopf (2005). "Fast protein classification with multiple networks". In: *Bioinformatics* 21.suppl_2, pp. ii59–ii65.

Udell, Madeleine et al. (2016). "Generalized low rank models". In: *Foundations and Trends in Machine Learning* 9.1, pp. 1–118. ISSN: 19358245. DOI: 10.1561/2200000055. arXiv: 1410.0342.

Wang, Bo et al. (2014). "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature methods* 11.3, p. 333.

Xiong, Momiao (2018a). *Big Data in Omics and Imaging: Association Analysis*. ISBN: 9781498725781.

— (2018b). *Big Data in Omics and Imaging: Integrated Analysis and Causal Inference*. Vol. 1. Chapman and Hall/CRC. ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Zhang, Shihua, Qingjiao Li, et al. (2011). "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules". In: *Bioinformatics* 27.13, pp. i401–i409.

Zhang, Shihua, Chun-Chi Liu, et al. (2012). "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data". In: *Nucleic acids research* 40.19, pp. 9379–9391.

Zhu, Jun et al. (2012). "Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation". In: *PLoS biology* 10.4.

Chapter 2

Materials and Methods

Before explaining the GLRM methodology we will discuss some multivariate techniques. Principal Components are the basis of many multivariate techniques. Its analytical solution of dimensionality reduction has made it a useful and popular method. However, it is used for a single dataset as an unsupervised approach. Thus, techniques such as Discriminant Analysis and Canonical Correlation Analysis can solve these issues. All three methodologies have analytical solutions, but are not directly feasible on high dimension datasets. Thus, it will be explained how to generalize this methods for dealing with data with more variables than observations, called n«p problem. Based on this explanation, we follow with the development of GLRM as proposed by Udell et al., 2016 and, at the end of this chapter, the problem of how to choose the number of principal components it is considered presenting some references, which can be useful for determining the rank of the model.

2.1 Principal Component Analysis

Principal Components Analysis (PCA), first introduced by Pearson K., 1901, and then developed by H. Hotelling, 1933, is a multivariate technique useful when working with one single dataset with a large number of interrelated quantitative variables. PCA focuses on reducing dimensionality by finding linear combination of variables of $\boldsymbol{Y}$. The idea is to find a new set of variables – the Principal Components (PCs) – that are uncorrelated and the first few retain most of the variation present in the original ones.

Thus, considering a numerical data matrix $\boldsymbol{Y}_{n\times p}$ derived from a single population, with no groups separation, where $\boldsymbol{Y}_{\boldsymbol{i}p\times 1} \in \mathcal{R}^p$ and $\boldsymbol{Y_i} \overset{\text{iid}}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we want to find independent linear functions, $\boldsymbol{a}^\top \boldsymbol{Y}$, that has maximum variance and are uncorrelated with others. Then, the j-th sample principal component can be found by

$$\max_{\boldsymbol{a_j}} \frac{\boldsymbol{a_j^\top S a_j}}{\boldsymbol{a_j^\top a_j}}, \quad s.t. \quad \begin{aligned} &\boldsymbol{a_j^\top a_j} = 1, j = 1, \ldots, K \\ &\boldsymbol{a_j^\top S a_l} = 0, \ \forall l < j, \end{aligned} \tag{2.1}$$

where $\boldsymbol{S} = \frac{1}{n}\tilde{\boldsymbol{Y}}^\top\tilde{\boldsymbol{Y}}$, with $\tilde{\boldsymbol{Y}} = (Y_{ij} - \frac{1}{n}\sum_{i=1}^{n} Y_{ij})_{ij}$ and the denominator $(n-1)$ can also be used. The dimension K is given by the $\min(n, p)$. From Lagrange multipliers it follows that $\boldsymbol{a_1}, \ldots, \boldsymbol{a_K}$ that maximizes Equation (2.1) correspond to $\boldsymbol{a_1}, \ldots, \boldsymbol{a_K}$ given by the eigenvectors associated with the biggest to smallest eigenvalues of $\boldsymbol{S}$, respectively. So, we have

- $\boldsymbol{Z_k} = \boldsymbol{a_k}^\top \boldsymbol{Y_i}$: k-th Principal Component;
- z_{ik}: score for the i-th observation on the k-th Pincipal Component;
- $\boldsymbol{a_k}$: eigenvector corresponding to the k-th largest eigenvalue of the covariance matrix.

It follows from the Singular Value Decomposition of $\boldsymbol{Y}$ that

$$\boldsymbol{Y}_{n\times p} = \boldsymbol{L}_{n\times n}\boldsymbol{\Lambda}^{1/2}_{n\times p}\boldsymbol{R}^\top_{p\times p},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the singular values in decreasing order, and $\boldsymbol{L}$ and $\boldsymbol{R}$ are orthonormal matrices with columns corresponding to the singular vectors associated with the observations and variables, respectively (see Ian T. Jolliffe, 2002, pg 44). Or even, $\boldsymbol{R}$ is the matrix of eigenvectors of the covariance matrix and coefficients of the PC, $\boldsymbol{\Lambda}^{1/2}$ is a diagonal matrix with the square roots of the eigenvalues and standard deviation of the PCs, and $\boldsymbol{L}$ is the matrix with the scaled version of the PC scores. Thus, $\boldsymbol{Z} = \boldsymbol{L}\boldsymbol{\Lambda}^{1/2}$ is the matrix of scores for the PCs.

Therefore, by selecting only the first m singular values, $m <= K$, the data Y can be represented in a lower dimension as

$$\boldsymbol{Y}_{n\times p} \approx \boldsymbol{L}_{\boldsymbol{m}_{n\times m}}\boldsymbol{\Lambda}^{1/2}_{\boldsymbol{m}_{m\times m}}\boldsymbol{R}^\top_{\boldsymbol{m}_{m\times p}}.$$

This is the best possible rank m approximation to $\boldsymbol{Y}$ (Gabriel, 2008; Young and Householder, 1938) in the sense of minimizing

$$\min_{\boldsymbol{A},\boldsymbol{B}} ||\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{B}||^2_F,$$

where $||\boldsymbol{w}||^2_2 = \sum_{j=1}^{p} w_j^2$, therefore

$$\min_{\boldsymbol{a},\boldsymbol{b}} \sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \boldsymbol{a_i}\boldsymbol{b_j})^2. \tag{2.2}$$

2.2 Discriminant Analysis

The term "discrimination", first introduced by Fisher, 1938, relates to a multivariate technique used to separate objects or observations. Suppose there are G populations or groups, $\tau_1, \ldots, \tau_G$, $G \geq 2$, and the observations of the data matrix $\boldsymbol{Y}_{n\times p}$ are partitioned into these groups, such that $n = \sum_{g=1}^{G} n_g$. Then, there is an interest in allocate an individual to one of these G groups based on the p measures, while making as few mistakes as possible. One approach to do so is

the Fisher's Linear Discriminant Function, whose goal is to find a low dimensional representation while separating the groups as much as possible.

Then, the rows of the data matrix $\boldsymbol{Y}_{n\times p}$ are partitioned into groups, in which $\boldsymbol{Y_{ig}}|\tau_g \overset{\text{iid}}{\sim} (\boldsymbol{\mu_g}, \boldsymbol{\Sigma_W})$. Note that, although it does not assume normality of the data, it does assume that all the groups have the same covariance matrix $\boldsymbol{\Sigma_W}$, called the within-group covariance matrix. If we consider the total covariance matrix $\boldsymbol{\Sigma}$ of data $\boldsymbol{Y}$, it can be decomposed into the within-group covariance matrix $\boldsymbol{\Sigma_W}$ and the between-group covariance matrix $\boldsymbol{\Sigma_B}$. So, now we want to find independent linear functions $\boldsymbol{a}^\top \boldsymbol{Y}$ that maximizes the between-groups covariance relative do their within-group covariance. Considering $\boldsymbol{B} = \sum_{i=1}^{g} n_g(\bar{\boldsymbol{Y}}_{\boldsymbol{i}} - \bar{\boldsymbol{Y}})((\bar{\boldsymbol{Y}}_{\boldsymbol{i}} - \bar{\boldsymbol{Y}}))^\top$ as the sample between-group matrix and $\boldsymbol{W} = (n-G)^{-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\boldsymbol{Y}}_{\boldsymbol{i}})(Y_{ij} - \bar{\boldsymbol{Y}}_{\boldsymbol{i}})^\top$ the sample within-group matrix, where, the j-th sample discriminant vector can be found by

$$\begin{aligned} \max_{\boldsymbol{a_j}} \frac{\boldsymbol{a_j^\top B a_j}}{\boldsymbol{a_j^\top W a_j}}, \quad s.t. \quad & \boldsymbol{a_j^\top W a_j} = 1, j = 1, \ldots, k \\ & \boldsymbol{a_j^\top W a_l} = 0, j = 1, \ldots, k, \forall l < j, \end{aligned} \tag{2.3}$$

where $\boldsymbol{a_1}, \ldots, \boldsymbol{a_k}$ are the eigenvectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$.

It is easier to separate groups when the between-group sum of squares $(\boldsymbol{B})$ is bigger than the within-group $\boldsymbol{W}$. That is, when groups are very similar within and different from each other.

One advantage of Fisher's Linear Discriminant Function is that it can perform reduced-rank classification by recasting the classification problem as a regression problem (Hastie, Buja, and Tibshirani, 1995; Clemmensen et al., 2011; Han and Clemmensen, 2016). Let $\boldsymbol{D}_{n\times k}$ be a matrix of dummy variables for the G groups, thus

$$D_{ig} = \begin{cases} 1, \text{if the } i\text{-th observation belongs to the } g\text{-th group;} \\ 0, \text{otherwise} \end{cases}.$$

So it follows

$$\begin{aligned} \min_{\boldsymbol{a_j}, \boldsymbol{\theta_j}} ||\boldsymbol{D\theta_j} - \boldsymbol{Y a_j}||^2, \quad s.t. \quad & \frac{1}{n}\boldsymbol{\theta_j^\top D^\top D \theta_j} = 1, \\ & \boldsymbol{\theta}_j^\top \boldsymbol{D}^\top \boldsymbol{D\theta_l} = 1, \forall l < j, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\theta_j}$ is a G-vector of scores and $\boldsymbol{a_j}$ is a p-vector of variable coefficients. The vector $\boldsymbol{a_j}$ that minimizes (2.4) is a discriminant vector equivalent to the one that maximizes (2.3).

2.3 Canonical Correlation Analysis

Initially developed by Harold Hotelling, 1935; Harold Hotelling, 1936, the Canonical Correlation Analysis focuses on summarizing the association between two datasets, by maximizing the

correlation between linear combinations of the variables of each dataset. For two datasets $\boldsymbol{Y}^{(1)}_{n\times p}$ and $\boldsymbol{Y}^{(2)}_{n\times q}$, so that

$$\boldsymbol{Y}_{i(p+q)\times 1} = \begin{bmatrix} \boldsymbol{Y_i^{(1)}} \\ \hdashline \boldsymbol{Y_i^{(2)}} \end{bmatrix} \overset{\text{iid}}{\sim} \left(\boldsymbol{\mu} = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix} \right).$$

we want to find independent linear combinations for each dataset, say $\boldsymbol{U} = \boldsymbol{a}^\top \boldsymbol{Y}^{(1)}$ and $\boldsymbol{V} = \boldsymbol{b}^\top \boldsymbol{Y}^{(2)}$. Then, considering the sample covariance matrix

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S_{11}} & \boldsymbol{S_{12}} \\ \boldsymbol{S_{21}} & \boldsymbol{S_{22}} \end{pmatrix}, \quad \text{with} \quad \boldsymbol{S_{kl}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{Y_i^{(k)}} - \boldsymbol{\bar{Y}^{(k)}})(\boldsymbol{Y_i^{(l)}} - \boldsymbol{\bar{Y}^{(l)}})^\top,$$

the j-th sample canonical variables can be found by the follow bilinear optimization problem

$$\max_{\boldsymbol{a_j}, \boldsymbol{b_j}} Corr(\boldsymbol{U}, \boldsymbol{V}) = \max_{\boldsymbol{a_j}, \boldsymbol{b_j}} \frac{\boldsymbol{a_j^\top S_{12} b_j}}{\sqrt{\boldsymbol{a_j^\top S_{11} a_j}} \sqrt{\boldsymbol{b_j^\top S_{22} b_j}}}, j = 1, \ldots, k, \tag{2.5}$$

which is equivalent to solve two separate optimization problems(Mardia, Kent, and Bibby J M, 1979)

$$\max_{\boldsymbol{a_j}} \frac{\boldsymbol{a_j^\top S_{12} S_{22}^{-1} S_{21} a_j}}{\boldsymbol{a_j^\top S_{11} a_j}},$$

and

$$\max_{\boldsymbol{b_j}} \frac{\boldsymbol{b_j^\top S_{12} S_{11}^{-1} S_{21} b_j}}{\boldsymbol{b_j^\top S_{22} b_j}}.$$

Thus, by solving each separately we obtain that $\boldsymbol{a_j} = \boldsymbol{e_j^\top S_{11}^{-1/2}}$ and $\boldsymbol{b_j} = \boldsymbol{f_j^\top S_{22}^{-1/2}}$, where $\boldsymbol{e_1}, \ldots, \boldsymbol{e_k}$ are the eigenvectors associated to $\boldsymbol{S_{11}^{-1/2} S_{12} S_{22}^{-1} S_{21} S_{11}^{-1/2}}$ and $\boldsymbol{f_1}, \ldots, \boldsymbol{f_k}$ the eigenvectors associated to $\boldsymbol{S_{22}^{-1/2} S_{21} S_{11}^{-1} S_{12} S_{22}^{-1/2}}$.

If we consider the spectral decomposition of the covariance matrix S_{11}, we have that $S_{11}^{-1/2} = \boldsymbol{R_{(1)} \Lambda_{(1)}^{-1/2} R_{(1)}^\top}$, where $\boldsymbol{R_{(1)}}$ and $\Lambda_{(1)}$ are the matrices of eigenvectors and eigenvalues of the covariance matrix of dataset $\boldsymbol{Y_{(1)}}$. So, it follows that the canonical variable of dataset $\boldsymbol{Y^{(1)}}$ can be interpreted as rotations of the standardized principal component. A similar interpretation can be given for $\boldsymbol{V}$

$$\boldsymbol{U_1 = a_1^\top Y^{(1)}} = \boxed{\boldsymbol{e_1^\top L_{(1)}} \boxed{\boldsymbol{\Lambda_{(1)}^{-1/2}} \boxed{\boldsymbol{L_{(1)}^\top Y^{(1)}}}_{\text{Principal Component}}}_{\text{Standardized Principal Component}}}_{\text{Rotation}}$$

Also, Canonical Correlation Analysis is closely related to Discriminant Analysis (Bartlett, 1938) when one of the datasets consist of a dummy variable.

2.4 High-dimensional Datasets

The previous solutions work well for numerical and homogeneous datasets on cases in which the number of observations is greater than the number of variables and the matrix is complete. However, since they consist on finding linear combinations of all variables, as this number starts to increase, it becomes harder to interpret the vectors, once most loadings are non-zero. Also, when the number of variables increases, multicollinearity is more likely to be present; thus the covariance matrices won't have full rank and it would not be possible to find the solutions.

2.4.1 Principal Component Analysis

Different solutions have been proposed to overcome this problem for Principal Components, such as rotation techniques (Ian T. Jolliffe, 1995), linear combinations using some simplicity measure (Ian T. Jolliffe and Uddin, 2000), restrict the loading to -1, 0 or 1 (Vines, 2000), artificially set loadings to zero if their absolute value are below a certain threshold (Cadima and Ian T. Jolliffe, 1995), and many others. Another possible solution is to include regularization functions to impose a certain structure to the matrices and also improve the algorithm performance. Thus, we can modify the optimization model in (2.2) by adding regularizers to obtain a more general form such as

$$\min \sum_{i=1}^{n}\sum_{j=1}^{p}(Y_{ij} - \boldsymbol{a_i}\boldsymbol{b_j})^2 + \gamma \sum_{i=1}^{n} r_i(\boldsymbol{a_i}) + \gamma \sum_{j=1}^{p} \tilde{r}_j(\boldsymbol{b_j}),$$

with $\gamma \in (0, 1)$ a regularization parameter and the regularizer $r_i : \mathcal{R}^{1\times m} \to \mathcal{R}$, for $i = 1, \ldots, n$, of the rows of A, and of the columns of B $\tilde{r}_j : \mathcal{R}^{1m} \to \mathcal{R}$, for $j = 1, \ldots, p$.

The regularizers imposes a specific structure to the data accordingly to the function used, as shown at Table 2.1.

Table 2.1: *Regularizers and data structure*

Imposed Structure	$r(\boldsymbol{a})$	$\tilde{r}(\boldsymbol{b})$
Small	$\|\|\boldsymbol{a}\|\|_2^2$	$\|\|\boldsymbol{b}\|\|_2^2$
Sparse	$\|\|\boldsymbol{a}\|\|_1$	$\|\|\boldsymbol{b}\|\|_1$
Non-negative	$\mathbb{I}(\boldsymbol{a} \geq 0)$	$\mathbb{I}(\boldsymbol{b} \geq 0)$
Cluster	$\mathbb{I}(\boldsymbol{card}(\boldsymbol{a}) = 1)$	0

It is defined that $\mathbb{I}(\boldsymbol{a} \geq 0) = 0$ if $\boldsymbol{a} \geq 0$, and $\mathbb{I}(\boldsymbol{a} \geq 0) = \infty$ otherwise. Also, the function $\boldsymbol{card}(\boldsymbol{a})$ denotes the number of nonzero entries of a vector. So, for example, if the interest is the selection of variables, a sparse structure could be imposed at the columns of $\boldsymbol{B}$. Hence, by using $\tilde{r}(\boldsymbol{b}) = ||\boldsymbol{b}||_1$ many loadings of the variables in the PC would be equal to zero, so only a smaller number of variables would contribute to the problem.

2.4.2 Discriminant Analysis

As for Discriminant Analysis, when the number of variables is much greater then the observations, the within-class covariance matrix, $\boldsymbol{W}$, is singular, so Fisher's Discriminant Analysis can not be directly applied. Also, when we are in a high-dimensional setting we may find a classifier that performs feature selection as well. Some solutions have been proposed to deal with this situation. Some proposed Multivariate Gaussian for Linear Discriminant Analysis, such as Dudoit, Fridlyand, and Speed, 2002 and Bickel and Levina, 2004, who assumed independence on features, and Friedman, 1989, who applied a ridge penalty to the within-class covariance matrix. Others focused on sparse classifiers such as Tibshirani et al., 2002; Guo, Hastie, and Tibshirani, 2007 and Witten and Tibshirani, 2011.

One possible solution proposed by Clemmensen et al., 2011 generalizes the Equation (2.4) to allow the calculation of discriminant vectors, while making it sparse. Thus, the j-th sparse discriminant analysis solution pair (θ_j, a_j) solves

$$\begin{aligned} \min_{\boldsymbol{a_j},\boldsymbol{\theta_j}} ||\boldsymbol{D\theta_j} - \boldsymbol{Ya_j}||^2 + \gamma \boldsymbol{a_j^\top \Omega j} + \lambda ||\boldsymbol{a_j}||_1, \quad s.t. \quad & \frac{1}{n}\boldsymbol{\theta_j^\top D^\top D\theta_j} = 1, \\ & \boldsymbol{\theta}_j^\top \boldsymbol{D}^\top \boldsymbol{D\theta_l} = 1, \forall l < j, \end{aligned} \tag{2.6}$$

with $\boldsymbol{\theta_j}, \boldsymbol{a_j}$ and $\boldsymbol{D}$ as described in Section 2.2, $\boldsymbol{\Omega}$ is a positive definite matrix and γ and λ are non-negative tuning parameters.

2.4.3 Canonical Correlation Analysis

In the case of Canonical Correlation Analysis one possible solution to deal with high dimensional data is to use a penalized matrix decomposition (Witten, Tibshirani, and Hastie, 2009). When $\boldsymbol{Y^{(1)}}$ and $\boldsymbol{Y^{(2)}}$ have standardized columns, the Equation (2.5) can be rewritten as (Harold Hotelling, 1936)

$$\begin{aligned} \max_{\boldsymbol{a_j},\boldsymbol{b_j}} \boldsymbol{a_j^\top Y^{(1)\top} Y^{(2)} b_j}, \quad s.t. \quad & \boldsymbol{a_j^\top Y^{(1)\top} Y^{(1)} a_j} \leq 1, j = 1, \ldots, k \\ & \boldsymbol{b_j^\top Y^{(1)\top} Y^{(1)} b_j} \leq 1, j = 1, \ldots, k. \end{aligned} \tag{2.7}$$

So, to obtain sparse canonical variables we can include some penalty functions r and $\tilde{r}$ in Equation (2.8), such that

$$\begin{aligned} \max_{\boldsymbol{a_j},\boldsymbol{b_j}} \boldsymbol{a_j^\top Y^{(1)\top} Y^{(2)} b_j}, \quad s.t. \quad & \boldsymbol{a_j^\top Y^{(1)\top} Y^{(1)} a_j} \leq 1, j = 1, \ldots, k, \\ & \boldsymbol{b_j^\top Y^{(1)\top} Y^{(1)} b_j} \leq 1, j = 1, \ldots, k, \\ & r(\boldsymbol{a_j}) \leq c_1, \tilde{r}(\boldsymbol{b_j}) \leq c_2, \end{aligned} \tag{2.8}$$

where c_1 and c_2 are chosen by cross-validation to maximize the correlation between $\boldsymbol{a_j Y^{(1)}}$ and $\boldsymbol{b_j Y^{(2)}}$.

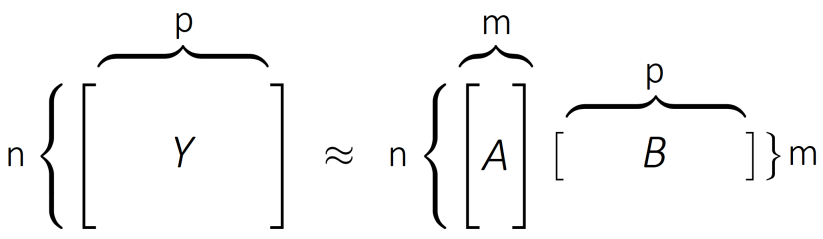
Another possible solution that deals with different high dimensional datasets is the Sparse Generalized Canonical Correlation Analysis (sGCCA) (Tenenhaus et al., 2014). Consider that now we have Q normalized, centered and scaled dataset $\boldsymbol{Y^{(1)}_{n\times p_1}}, \boldsymbol{Y^{(2)}_{n\times p_2}}, \ldots, \boldsymbol{Y^{(Q)}_{n\times p_Q}}$. Then, for each dimension $h = 1, \ldots, H$, sGCCA solves

$$\max_{\boldsymbol{a^{(1)}_h},\ldots,\boldsymbol{a^{(Q)}_h}} \sum_{i,j=1,i\neq j}^{Q} c_{i,j}\, cov(\boldsymbol{Y^{(i)}_h a^{(i)_h}}, \boldsymbol{Y^{(j)}_h a^{(j)_h}}), \quad s.t. \quad ||\boldsymbol{a^{(q)_h}}||_2 = 1, \\ ||\boldsymbol{a^{(q)_h}}||_1 \leq \lambda^{(q)}, \forall 1 \leq q \leq Q, \tag{2.9}$$

where $C_{Q\times Q} = c_{i,j_i}, j$ is a design matrix indicating the relation between datasets, λ is a nonnegative parameter of shrikage and $\boldsymbol{a^{(q)_h}}$ is the loading vector on dimension h associated to the residual matrix $X^{(1)}_h$ of the dataset $X^{(q)}$. It is an iterative problem: first solve (2.9) for $h = 1$ to obtain the first set of coefficients $(\boldsymbol{a^{(1)_1}}), \ldots, \boldsymbol{a^{(Q)_1}}$; then maximize (2.9) for $h = 2$ using the residual matrices $\boldsymbol{Y^{(q)}_2} = \boldsymbol{Y^{(q)}_1} - \boldsymbol{t^{(q)}_1 a^{(q)}_1}$, $1 \leq q \leq Q$, where $\boldsymbol{t^{(q)}_1} = \boldsymbol{Y^{(q)}_1 a^{(q)}_1}$is the component score; and so on, until a sufficient number of components is obtained. By including the shrinking parameter it allows to perform variable selection, thus identifying those which are highly correlated between and within datasets.

2.5 Generalized Low Rank Model (GRLM)

The previous techniques can deal well with data, as long as they are quantitative. However, in many real problems we need to work with data of different abstract types. Thus, now consider $\boldsymbol{Y}_{n\times p}$ to be a data matrix with n individuals and p variables where each entry Y_{ij} was obtained from a space $\mathcal{F}_j$. Therefore, $\mathcal{F}_j$ can represent data of many different types such as boolean ($\mathcal{F}_j = V, F$), integers ($\mathcal{F}_j = 1, 2, 3, \ldots$), ordinal ($\mathcal{F}_j = low, medium, high$) or intervals ($\mathcal{F}_j = (z, w) : z \in \mathcal{R}, w \in \mathcal{R}$). The data can be decomposed in two numeric matrices $\boldsymbol{A}_{n\times m}$ and $\boldsymbol{B}_{m\times p}$. Note that $\boldsymbol{A} \in \mathcal{R}^{n\times m}$ and $\boldsymbol{B} \in \mathcal{R}^{m\times p}$, were $m \leq min\{n, p\}$. Therefore, it is possible to approximate the values of Y_{ij} by the inner product of the ith line of matrix $\boldsymbol{A}$ and the j column of matrix $\boldsymbol{B}$, i.e., $Y_{ij} \approx a_i b_j$. A representation of such dimensionality reduction is



.

Matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be seen as compact representations of the sample units and the variables, respectively. Once there is no restrictionon data type for $\boldsymbol{Y}$ – they could be either numerical, categorical, boolean, ordinal or counting, for example – it is interesting to represent them this way, so that abstract variables can be represented by numerical vectors. Hence, if the interest is prediction or clustering, for instance, it would be possible to do it from the numerical vector with the already known techniques.

To obtain the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ for a dataset $\boldsymbol{Y}$ with different data types, Udell et al., 2016 developed the Generalized Low Rank Model (GLRM), given by the optimization problem

$$\min \sum_{i=1}^{n} \sum_{j=1}^{p} L_{ij}(Y_{ij}, \boldsymbol{a_i}\boldsymbol{b_j}) + \gamma_a \sum_{i=1}^{n} r_i(\boldsymbol{a_i}) + \gamma_b \sum_{j=1}^{p} \tilde{r}_j(\boldsymbol{b_j}),$$

with the variables $\boldsymbol{A} = (\boldsymbol{a_i}) \in \mathcal{R}^{n\times m}$, $\boldsymbol{B} = (\boldsymbol{b_j}) \in \mathcal{R}^{m\times p}$, loss function $L_{ij} : \mathcal{R} \times \mathcal{F}_j \to \mathcal{R}$ and regularizers $r_i : \mathcal{R}^{1\times m} \to \mathcal{R}$ and $\tilde{r}_j : \mathcal{R}^{1m} \to \mathcal{R}$. In this case the loss function $L_{ij} : \mathcal{R} \times \mathcal{F}_j \to \mathcal{R}$ describes the error committed by approximating a variable in $\mathcal{F}_j$ by a real number.

Therefore, abstract variables can be introduced to the problem by choosing a corresponding Loss Function for each type. Even for a specific data type there are different functions to also account for different problems. Some loss functions are described at Table 2.2.

Table 2.2: *Loss function and data type*

Data type	Loss	$L(u, v)$
Real	Quadratic	$(u - v)^2$
	Absolute	$\lvert u - v\rvert$
	Huber	$huber(u - v)$
	Fractional	$max\left(\frac{v-u}{u}, \frac{u-v}{v}\right)$
	Logarithmic	$log^2\left(\frac{u}{v}\right)$
Boolean	Hinge	$(1 - uv)_+$
	Logistic	$log(1 + exp(-uv))$
Integer	Poisson	$exp(u) - uv + vlog(v) - v$
Ordinal	Hinge Ordinal	$\sum_{v'=1}^{v-1}(1 - u + v')_+ + \sum_{v'=v+1}^{d}(1 + u - v')_+$
Categorical	One-vs-all	$(1 - u_v)_+ + \sum_{v'\neq v}(1 + u_{v'})_+$

Let's consider the case where the data is numeric. The Quadratic Loss reduces the problem to PCA with regularization. But if there are many outliers to the data, a more robust function could be used, such as the Absolute or Huber losses. When the interest is to find approximations of Y, whose entries are close to the original matrix on a relative scale, rather then an absolute scale, the Fractional or Logarithmic losses should be used.

Also, by combining different loss function and regularizers it is possible to reproduce already known models, such as shown in the following table.

Table 2.3: *Models obtained by combinations of different loss functions and regularizers*

Model	$L_j(u, v)$	$r(\boldsymbol{a})$	$\tilde{r}(\boldsymbol{b})$
PC	$(u-v)^2$	0	0
Regularized PC	$(u-v)^2$	$\|\|\boldsymbol{a}\|\|_2^2$	$\|\|\boldsymbol{b}\|\|_2^2$
NNMF	$(u-v)^2$	$\mathbb{I}(\boldsymbol{a} \geq 0)$	$\mathbb{I}(\boldsymbol{b} \geq 0)$
Sparse PC	$(u-v)^2$	$\|\|\boldsymbol{a}\|\|_1$	$\|\|\boldsymbol{b}\|\|_1$
Robust PC	$\|u-v\|$	$\|\|\boldsymbol{a}\|\|_2^2$	$\|\|\boldsymbol{b}\|\|_2^2$
Logistic PC	$log(1+esp(-vu))$	$\|\|\boldsymbol{a}\|\|_2^2$	$\|\|\boldsymbol{b}\|\|_2^2$
Boolean PC	$(1-vu)_+$	$\|\|\boldsymbol{a}\|\|_2^2$	$\|\|\boldsymbol{b}\|\|_2^2$
K-means	$(u-v)^2$	$\mathbb{I}(\boldsymbol{card}(\boldsymbol{a}) = 1)$	0

2.5.1 Dealing with Missing data

Now consider there are some missing data, with only the observed entries were Y_{ij} for $(i, j) \in \Omega \subset 1, \ldots, n \times 1, \ldots p$ of matrix $\boldsymbol{Y}$. Then, the optimization problem is

$$\min \sum_{(i,j)\in\Omega} L_{ij}(Y_{ij}, \boldsymbol{a_i b_j}) + \gamma_a \sum_{i=1}^{n} r_i(\boldsymbol{a_i}) + \gamma_b \sum_{j=1}^{p} \tilde{r}_j(\boldsymbol{b_j}).$$

The solution of this problem gives the approximation $\hat{Y}_{ij} = \boldsymbol{a_i b_j}$ for the missing entries $(i, j) \notin \Omega$. So, if the matrix Y contains only numerical values, the missing data can be imputed simply by using the estimative $\hat{Y_{ij}} = \boldsymbol{a_i b_j}$. However, if the data is of any other type, this solution is not of the same type, since it is a representation of the data in the subspace of real matrices. Thus, the missing data can be obtained by

$$\hat{Y}_{ij} = \arg\min_{u} L_{ij}(\boldsymbol{a_i b_j}, u),$$

therefore, $\hat{Y}_{ij}$ belongs to the domain $\mathcal{F}_j$ of L_{ij}. In some cases this solution is straightforward, for example, when using the Hinge loss the imputation can be done with $\hat{Y_{ij}} = sign\{\boldsymbol{a_i b_j}\}$. This approximation $\hat{Y}_{ij}$ is useful for imputation or prediction, and can even be used when the entry was observed, which can be seen as a denoised value.

2.5.2 Scaling and Offset

When dealing with different data types their different scales may interfere in the analysis. An approach to compensate is to standardize the data. However, this solution can be inappropriate when dealing with abstract data, since it requires the calculation of means and variances. Thus, to avoid changing the original dataset the loss functions are rescaled in a way that generalizes standardization. Consider the generalization of mean and variance of each column by (Udell et al.,

2016) (see section 4.3)

$$\mu_j = \arg\min_{\mu} \sum_{i:(i,j)\in\Omega} L_{ij}(\mu, Y_{ij}), \quad \sigma_j^2 = \frac{1}{n_j - 1} \sum_{i:(i,j)\in\Omega} L_{ij}(\mu_j, Y_{ij}). \tag{2.10}$$

Then the loss function can be rescaled by σ_j^2 to fit a scaled GLRM by solving

$$\min \sum_{(i,j)\in\Omega} L_{ij}(Y_{ij}, \boldsymbol{a_i}\boldsymbol{b_j})/\sigma_j^2 + \gamma_a \sum_{i=1}^{n} r_i(\boldsymbol{a_i}) + \gamma_b \sum_{j=1}^{p} \tilde{r}_j(\boldsymbol{b_j}).$$

Another possibility to deal with different scales of data is to include and offset. This can be done modifying the regularizers and increasing the size of the rank $m' = m + 1$. Then, let

$$r'(\boldsymbol{a}) = \begin{cases} r(a_2, \ldots, a_{m+1}, a_1 = 1; \\ \infty, \text{otherwise} \end{cases}, \quad \text{and} \quad \tilde{r}'(\boldsymbol{b}) = \tilde{r}(b_2, \ldots, b_{m+1}),$$

where $r' : \mathcal{R}^{m+1} \to \mathcal{R}$ and $\tilde{r}' : \mathcal{R}^{m+1} \to \mathcal{R}$. This guarantees that the first columns of $\boldsymbol{A}$ is constant and the first row of $\boldsymbol{B}$ will be the vector of columns mean $\boldsymbol{\mu}$ from (2.10).

2.5.3 Interpretation

The interpretation of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ for GLRM is very similar to that of the Principal Components. Here each row of $\boldsymbol{B}$ gives an archetype, and each column is a numerical representation of the variable in a m-dimensional space. The lines $\boldsymbol{a_i}$ of matrix $\boldsymbol{A}$ gives a representation of the individual i in terms of all archetypes. Also, all the entries of $\boldsymbol{A}$ and $\boldsymbol{B}$ are real numbers, differently from the original data. So there are numerical representations of the individuals and the variables. Now they are easily plotted or clustered. For example, if there is the interest of clustering the individuals based on the variables, the lines of $\boldsymbol{A}$ could be used in any regular clustering algorithms. Or even if the interest is to find similar variables, the columns of $\boldsymbol{B}$ could be clustered.

2.5.4 Implementation

In this project, GLRM was implemented using *Julia*. Alternating proximal gradient method was used to solve the optimization problem.

When adjusting the model, it is necessary to specify the rank m, the regularization parameter γ and the loss function L_{ij}. The loss function is chosen based on prior knowledge of the data type, to best reflect its nature. The parameter can be chosen based on cross validation to avoid over or underfitting the data, as explained next.

2.6 Choosing parameters

2.6.1 Choosing the number of Principal Components

A common discussion when performing Principal Components analysis is choosing the optimal number of principal components that reduces dimension loosing as few information as possible. Ian T. Jolliffe, 2002 describes on chapter 6 a few possibilities for choosing the number of PCs.

First are the **Ad-hoc measures**, which are intuitively plausible and work in practice, but do not have a formal basis.

1. **Cumulative Percentage of Total Variance:** it consists on first calculating the percentage of variation accounted for by the first m PCs

$$t_m = 100 \cdot \frac{\sum_{k=1}^{m} l_k}{\sum_{k=1}^{p} l_k},$$

   where l_k is the variance of the kth PC (k-th eigenvalue). Then a cut off t^* is chosen. The value of m for which this percentage is exceeded, is the number of PCs to be chosen. The choice of this cut-off is not straightforward, so many autors have studied the distribution and values of t_m to facilitate on the choice of t^* (see chap 6 Ian T. Jolliffe, 2002, for references)

2. **Size of Variances of PCs:** when using standardized variables, the Kaiser's rule (Kaiser, 1960) can be applied. It retain only those PCs with l_k greater than 1 - Ian T Jolliffe, 1972 discusses that it retains too few variables and suggests using the 0.7 threshold instead. This follows that only 1 PC associated with each group of variable is retained. Adapting for when variables are not standardized, the PCs contained are those with l_k greater then 0.7 times the average value of the eigenvalues.

3. **Scree Graph and Log-Eigenvalue Diagram:** consists on looking at a plot of l_k (or $log(l_k)$ for the log-eigenvalue diagram) against k and choosing the "elbow" (Figure 2.1) on the graph as the number of components to be retained. That is, choosing the smalest value of k for which the difference between variances $l_{k-1} - l_k$ becomes smaller. It is a very subjective criterion and not always clear. An alternative is to compare it with the plot of 95th percentile of the distribution of each eigenvalue, when PCA is done in a "random" matrix.

Another possibility is by **Hypothesis Testing**. This approach sequentially tests hypothesis of equality of last $p-m$ eigenvalues against alternative hypothesis of, at least, two unequal to identify how many are simply noise.

$$H_0 : \lambda_{m+1} = \cdots = \lambda_p.$$

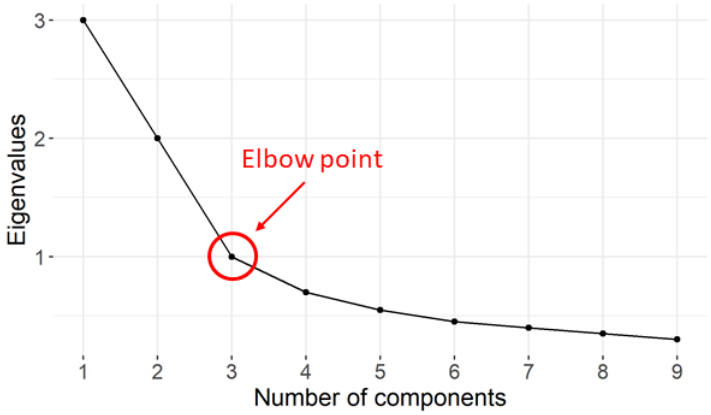


Figure 2.1: *Scree plot showing elbow point and slow decrease of variance difference.*

The number m of components to be retained is that in which the hypothesis of equality of last $p - m$ are rejected, after all others – $H_0 : \lambda_{p-1} = \lambda_p, \ldots, H_0 :_{m+1}= \cdots = \lambda_p$– were not rejected. However, this approach assumes multivariate normality of data and seems to retain more variables then necessary.

Finally, there are also **Cross-validatory of computationally intense methods**. It consists on predicting each element y_{ij} of the data Y based on a submatrix of Y that does not include y_{ij}. There are two methods to determine a criterion for choosing the number of components to be retained: one by Eastment and Krzanowski, 1982 and one by Wold, 1978. The idea of both is to successively take the number of PCs starting from 1 until the prediction of y_{ij} is no longer improved by the addition of new terms. Both calculate the Prediction Sum of Squares

$$PRESS(m) = \sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{y}_{ij} - y_{ij})^2$$

as a metric to decide the number of components to retain. The methods differ on how the subset to predict y_{ij} is chosen, and how the Prediction Sum of Squares is used to find the number of components. However, there is not a formal significance test to define this number and this methods become computationally expensive for large datasets.

2.6.2 Choosing parameters of the GLRM

According to Udell et al., 2016, the choice of the parameters depends on the goal of the analysis. When using GLRM to compress data, two metrics can be used to define the parameters

$$\text{Compression ratio} = \frac{\text{number of nonzero numbers to represent data}}{\text{data size}}$$

and

$$\text{error} = \sum_{(i,j)\in\Omega} L_j(Y_{ij} - a_i b_j).$$

Based on them, there are two options: define a required error and choose the model parameters (rank m and regularization γ) with highest compression rate to achieve it; or choose a required compression rate and choose parameters with lowest error rate. Also, as information criterion, analogous to AIC or BIC can be constructed, with degree of freedom being the difference between the number of nonzeros in the model and the dimensionality of the symmetry group of the problem.

Alternatively, if we suppose that the data is contaminated by random noise, $Y_{ij} = Y_{ij}^{true} + \varepsilon_{ij}$, then the focus is to choose parameters that minimizes

$$\sum_{(i,j)\in\Omega} L_j(Y_{ij}^{true} - a_i b_j)$$

to remove noise. To do so, the scree graph and log-eigenvalue diagram from PCA can be adapted to GLRM. However, using cross validation to this denoising problem can be tricky, because leaving out too few entries can overfit the noise, whereas leaving out too many can underfit the signal.

Finally, when wanting to predict new entries, the parameters cannot be chosen considering only the objective function; it needs to be chosen based on a cross validation to obtain the performance of the model with non observed entries. The resampling for the cross validation can be done in three different ways, based on the source of the variability. If the rows are drawn independently and identically distributed from a population, then resample the rows – the same applies for columns. If the rows or columns are fixed, but the indices of entries are drawn randomly, then resample the observed entries of the matrix. Finally, if the indices are fixed, but the observed values contain errors of measurement, then the errors of the model must be resampled using a bootstrap or jackniffe. Thus, with the cross validation, an overall performance metric can be used to chose parameters. A common approach, as suggested by Boehmke and Greenwell, 2019, is to perform a grid search with different values for each parameter and choosing the one that minimizes a specific metric of interest.

In general, GLRMs with higher rank is able to fit the noisy data better; however a GLRM with many parameter may overfit the noise. But higher number of observations can make it more difficult to overfit data, allowing higher ranks. If the GLRM is being used to describe the data to gain more understanding of observations and variables relation, then there is no need for regularization. Alternatively, if the model will be used to predict new observations, then regularization must be used to generalize the model for unseen data Boehmke and Greenwell, 2019. Also, regularization is more important when fewer entries have been observed.

References

Bartlett, M. S. (1938). "Further aspects of the theory of multiple regression". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 34.1, pp. 33–40. ISSN: 14698064. DOI: 10.1017/S0305004100019897.

Bickel, Peter J. and Elizaveta Levina (2004). "Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations". In: *Bernoulli* 10.6, pp. 989–1010. ISSN: 13507265. DOI: 10.3150/bj/1106314847.

Boehmke, Brad and Brandon M Greenwell (2019). *Hands-On Machine Learning with R*. CRC Press.

Cadima, Jorge and Ian T. Jolliffe (1995). "Loadings and correlations in the interpretation of principal components". In: *Journal of Applied Statistics* 22.2, pp. 203–214. ISSN: 13600532. DOI: 10.1080/757584614.

Clemmensen, Line et al. (2011). "Sparse discriminant analysis". In: *Technometrics* 53.4, pp. 406–413. ISSN: 00401706. DOI: 10.1198/TECH.2011.08118.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data". In: *Journal of the American Statistical Association* 97.457, pp. 77–86. ISSN: 01621459. DOI: 10.1198/016214502753479248.

Eastment, HT and WJ Krzanowski (1982). "Cross-validatory choice of the number of components from a principal component analysis". In: *Technometrics* 24.1, pp. 73–77.

Fisher, Ronald Aylmer (1938). "The Statistical Utilization of Multiple Measurements". In: *Annals of Eugenics*, pp. 1–9. DOI: https://doi.org/10.1111/j.1469-1809.1938.tb02189.x.

Friedman, Jerome H. (1989). "Regularized discriminant analysis". In: *Journal of the American Statistical Association* 84.405, pp. 165–175. ISSN: 1537274X. DOI: 10.1080/01621459.1989.10478752.

Gabriel, K R (2008). "Least Squares Approximation of Matrices by Additive and Multiplicative Models". In: *Journal of the Royal Statistical Society . Series B (Methodological)* 40.2, pp. 186–196.

Guo, Yaqian, Trevor Hastie, and Robert Tibshirani (2007). "Regularized linear discriminant analysis and its application in microarrays". In: *Biostatistics* 8.1, pp. 86–100. ISSN: 14654644. DOI: 10.1093/biostatistics/kxj035.

Han, Xixuan and Line Clemmensen (2016). "Regularized generalized eigen-decomposition with applications to sparse supervised feature extraction and sparse discriminant analysis". In: *Pattern Recognition* 49, pp. 43–54.

Hastie, Trevor, Andreas Buja, and Robert Tibshirani (1995). "Penalized Discriminant Analysis". In: *The Annals of Statistics* 23.1, pp. 73–102.

Hotelling, H. (1933). "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6, pp. 417–441. ISSN: 00220663. DOI: 10.1037/h0071325.

Hotelling, Harold (1935). “The most predictable criterion”. In: *Journal of Educational Psychology* 26.2, pp. 139–142. ISSN: 00220663. DOI: 10.1037/h0058165.

— (1936). “Relations Between Two Sets of Variates”. In: *Biometrika* 28.3, pp. 321–377. URL: https://www.jstor.org/stable/2333955.

Jolliffe, Ian T (1972). “Discarding variables in a principal component analysis. I: Artificial data”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 21.2, pp. 160–173.

— (1995). “Rotation of principal components: Choice of normalization constraints”. In: *Journal of Applied Statistics* 22.1, pp. 29–35. ISSN: 13600532. DOI: 10.1080/757584395.

— (2002). *Principal Component Analysis*. Vol. 2.

Jolliffe, Ian T. and Mudassir Uddin (2000). “The simplified component technique: An alternative to rotated principal components”. In: *Journal of Computational and Graphical Statistics* 9.4, pp. 689–710. ISSN: 15372715. DOI: 10.1080/10618600.2000.10474908.

Kaiser, Henry F (1960). “The application of electronic computers to factor analysis”. In: *Educational and psychological measurement* 20.1, pp. 141–151.

Mardia, K V, T J Kent, and Bibby J M (1979). *Multivariate Analysis*, p. 519. ISBN: 0124712525. URL: http://dlx.b-ok.org/genesis/756000/87bd84eac4c0b06e1403c996e0531e2d/%7B%5C_%7Das/[K.%7B%5C_%7DV.%7B%5C_%7DMardia,%7B%5C_%7DJ.%7B%5C_%7DT.%7B%5C_%7DKent,%7B%5C_%7DJ.%7B%5C_%7DM.%7B%5C_%7DBibby]%7B%5C_%7DMultivaria(b-ok.org).pdf.

Pearson K. (1901). “Pearson, K. 1901. On lines and planes of closest fit to systems of points in space.” In: *Philosophical Magazine* 2, pp. 559–572.

Tenenhaus, Arthur et al. (2014). “Variable selection for generalized canonical correlation analysis”. In: *Biostatistics* 15.3, pp. 569–583. ISSN: 14684357. DOI: 10.1093/biostatistics/kxu001.

Tibshirani, Robert et al. (2002). “Diagnosis of multiple cancer types by shrunken centroids of gene expression”. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.10, pp. 6567–6572. ISSN: 00278424. DOI: 10.1073/pnas.082099299.

Udell, Madeleine et al. (2016). “Generalized low rank models”. In: *Foundations and Trends in Machine Learning* 9.1, pp. 1–118. ISSN: 19358245. DOI: 10.1561/2200000055. arXiv: 1410.0342.

Vines, S. K. (2000). “Simple principal components”. In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 49.4, pp. 441–451. ISSN: 00359254. DOI: 10.1111/1467-9876.00204.

Witten, Daniela M. and Robert Tibshirani (2011). “Penalized classification using Fisher’s linear discriminant”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 73.5, pp. 753–772. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2011.00783.x.

Witten, Daniela M., Robert Tibshirani, and Trevor Hastie (2009). “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis”. In: *Biostatistics* 10.3, pp. 515–534. ISSN: 14654644. DOI: 10.1093/biostatistics/kxp008.

Wold, Svante (1978). “Cross-validatory estimation of the number of components in factor and principal components models”. In: *Technometrics* 20.4, pp. 397–405.

Young, Gale and A. S. Householder (1938). “Discussion of a set of points in terms of their mutual distances”. In: *Psychometrika* 3.1, pp. 19–22. ISSN: 00333123. DOI: 10.1007/BF02287916.

Chapter 3

Application

Now, considering that the methodology of GLRM was introduced, we will apply it on biological data. First, the predictive ability of GLRM will be evaluated on a supervised data integration analysis through simulated dataset. It will be compared with multiple other data integration techniques of different types, on a multi-omics context. Then, the GLRM will be used to solve problems of a real biological dataset. In this context GLRM and data integration will be used for imputing missing data and compared with other imputation methods. The impact of data imputation in variable selection and explained variance will be studied. Finally, data of different types will be combined into latent variables and its association with mortality risk will be assessed.

3.1 Simulated Data

As stated before, the GLRM is a very powerful and flexible model and we are interested in exploring its potential to solve data integration problems. Thus, to verify its ability of integrating high dimensional data in a supervised approach, we fit it on simulated data. However, many datasets have complex interaction among data, specially in the context of disease, as represented by Figure 1.3.

3.1.1 Dataset

An example given by Ritchie et al., 2015 is the hypothetical pathway for Breast Cancer, represented in Figure 3.1. It is an example of how different omics levels interact to characterize a complex trait. This pathway was one of the examples generated by OmicsSIMLA (Chung and Kang, 2019), a tool that simulates multiple omics data types while modeling the interactions among them and the outcome of interest. In the hypothetical pathway, the outcome of Breast Cancer (Case and Control) is modelled. It generates 2022 SNPs of genes CYP1B1, CYP1A1, COMT, GSTM1, GSTT1 and from a regulatory region – where is located the meQTL (SNP that influences on methylation), eQTL (SNP that influences on gene expression) and the trans-eQTL or SNPr (SNP that is associated with a cluster of other SNPs). Among these there is a probability of deletion of a genotype (resulting on copy number variations (CNVs)) on gene CYP1A1. Both the CNV and 3

common variants on gene CYP1B1 are directly related to the outcome. Also, there are 5 rare SNPS on gene COMT related to the outcome and the meQTL, which influences the methylation rate of 688 CpGs sites of gene XRCC1. There are also 4 rare SNPs of genes GSTM1 and GSTT1 relating both to the outcome, and to the eQTL and SNPr, respectively. The eQTL and SNPr influence the gene expression, which influences the protein expression consequently, of gene XRCC3 and 99 others. The methylation, gene expression and protein expression were simulated retrospectively conditional on the disease status.

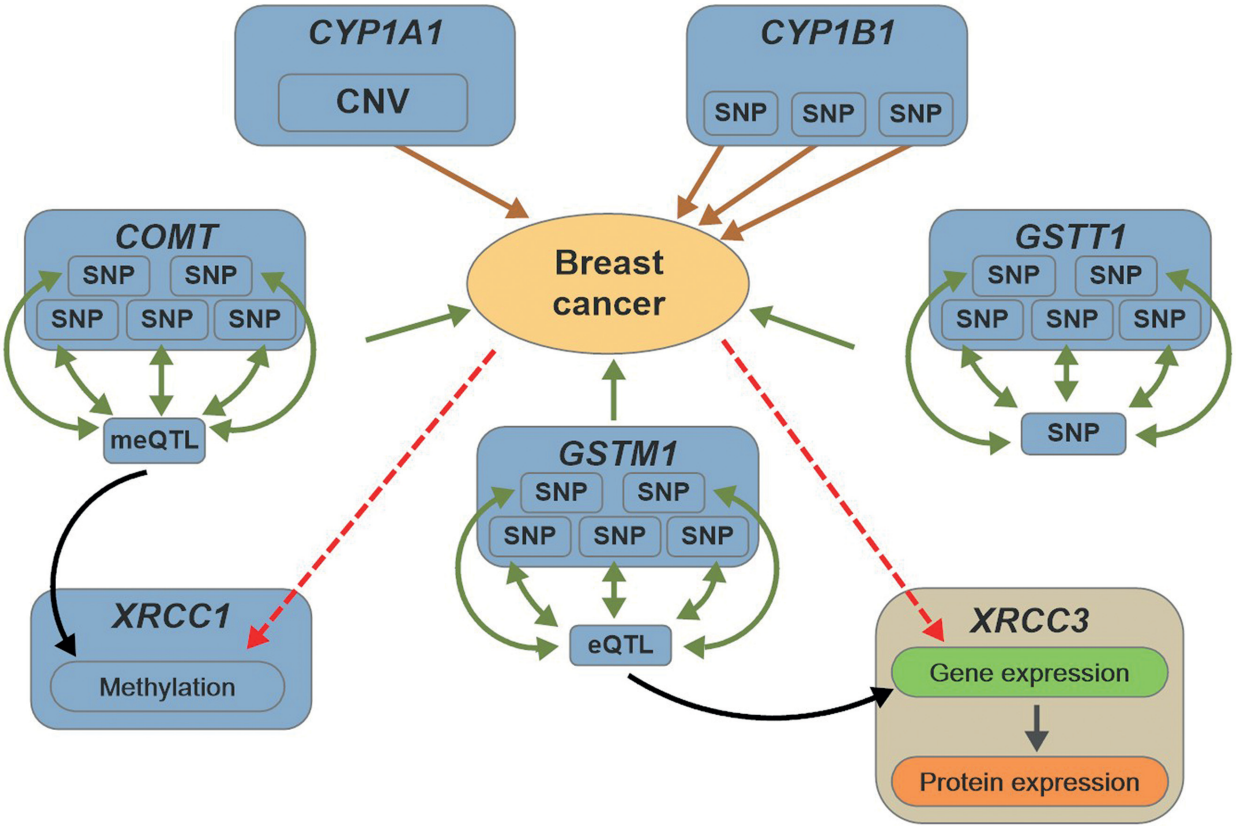


Figure 3.1: *Hypothetical pathways involved in breast cancer. The brown solid arrows represent the main effects of SNPs and CNVs on the disease, while the green solid arrows represent the interaction effects of SNPs on the disease. The black solid arrows represent the regulatory effects of the meQTL and eQTL on methylation and gene expression, respectively. The red dotted arrows represent the retrospective simulations of the methylation, gene expression, and protein expression levels conditional on the disease status.*

Source: Chung and Kang, 2019

Chung and Kang, 2019 simulated 1000 batches on three scenarios:

1. **Scenario 1:** stronger genetic effect on disease status. Data with 500 cases and 500 controls on training dataset and 100 cases and 100 controls on validation dataset.

2. **Scenario 2:** weaker genetic effect on disease status. Data with 500 cases and 500 controls on training dataset and 100 cases and 100 controls on validation dataset.

3. **Scenario 3:** stronger genetic effect on disease status. Data with 1500 cases and 1500 controls on training dataset and 500 cases and 500 controls on validation dataset.

The genetic effect of Scenarios 1 and 2 reflect on the difference of gene and protein expression between cases and controls. Figure 3.2, by using a heatmap, presents the gene and protein expression for gene XRCC3 (which is influenced by the SNPs) and 3 random ones. On Scenario 1 it is clear that individuals with breast cancer have higher gene and protein expression then those without. However, for scenario 2, since the genetic effect on gene expression is smaller, there is no clear separation between individuals with different outcomes.

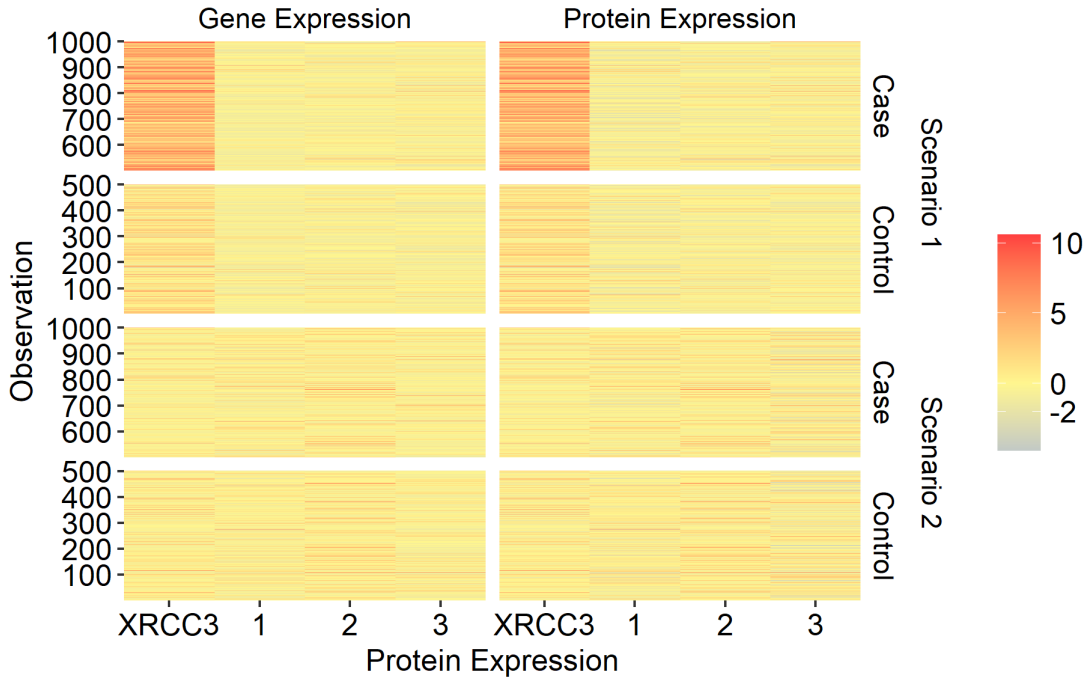


Figure 3.2: *Gene and Protein expression for Scenarios 1 and 2. The x axis represents a sample of genes and the y axis represents the individuals from each Scenario, separated by their outcomes. The color then represents the value of the variables for each individual.*

For each scenario Chung and Kang, 2019 compare different multi-omics analysis methods. First, a multi-staged method (3-stage (Holzinger, 2012)), where SNPs and CNV are associated with the disease, methylation and gene and protein expression were selected and used on a logistic model to predict disease status. Then, a meta-dimensional random forest-base method (RFomics) was used, which consist on calculating a risk score based on SNPs for each gene and combine them with the other multi-omics normalized data to be evaluated by the Random Forest algorithm. Further, the strategy was followed by a meta-dimensional graph-based method (CANetwork (Yan, Zhao, and Pang, 2017)) that uses a graph-based semi-supervised learning algorithm to predict disease outcome based on a graph matrix, which represents the distance between samples. Finally, it was considered a meta-dimensional model-based method (ATHENA (Holzinger et al., 2014)) that uses grammatical evolution neural networks (GENNs) to construct a prediction model from multi-omics data. More information about model implementation and parameters used are shown in Chung and Kang, 2019. The Area Under the Curve (AUC) was calculated for 100 batches and compared between each method.

3.1.2 Results

Now we wish to compare the predictive ability of GLRM, here considered as a meta-dimensional component-based integration method, with the same simulated data used in Chung and Kang, 2019. To do so, the data for CNV, Gene Expression, Methylation, SNP and Protein Expression was concatenated. Variables with constant values for both training and validation dataset were excluded. Gene and Protein expression data were already normalized. To better visualize each data type, Figure 3.3 presents a sample of variables from each dataset of the first batch of Scenario 1. The phenotype data (disease status) is a boolean type variable, with case being 1 and control being 0. CNV data indicates the number of copies of a specific gene, where -1 indicates a deletion in one of the 2 chromosomes, -2 indicates a deletion on both, and 0 indicates a normal number of copies. Gene and protein expression represent an intensity, thus they are real values with negative indicating a gene that is underexpressed, and positive values indicating an over expressed gene, or protein, compared to default values. Gene XRCC3 was simulated conditionally on the disease status; therefore, there is a clear separation between the expression for cases and control (in this scenario), with them being more expressed for cases. The methylation dataset represents the percentage of methylated runs for each site, thus being a variable between 0 and 1. Finally, the SNPs are coded as -1 and 1 for minor and major homozygous genotypes, respectively, and 0 for heterozygous.

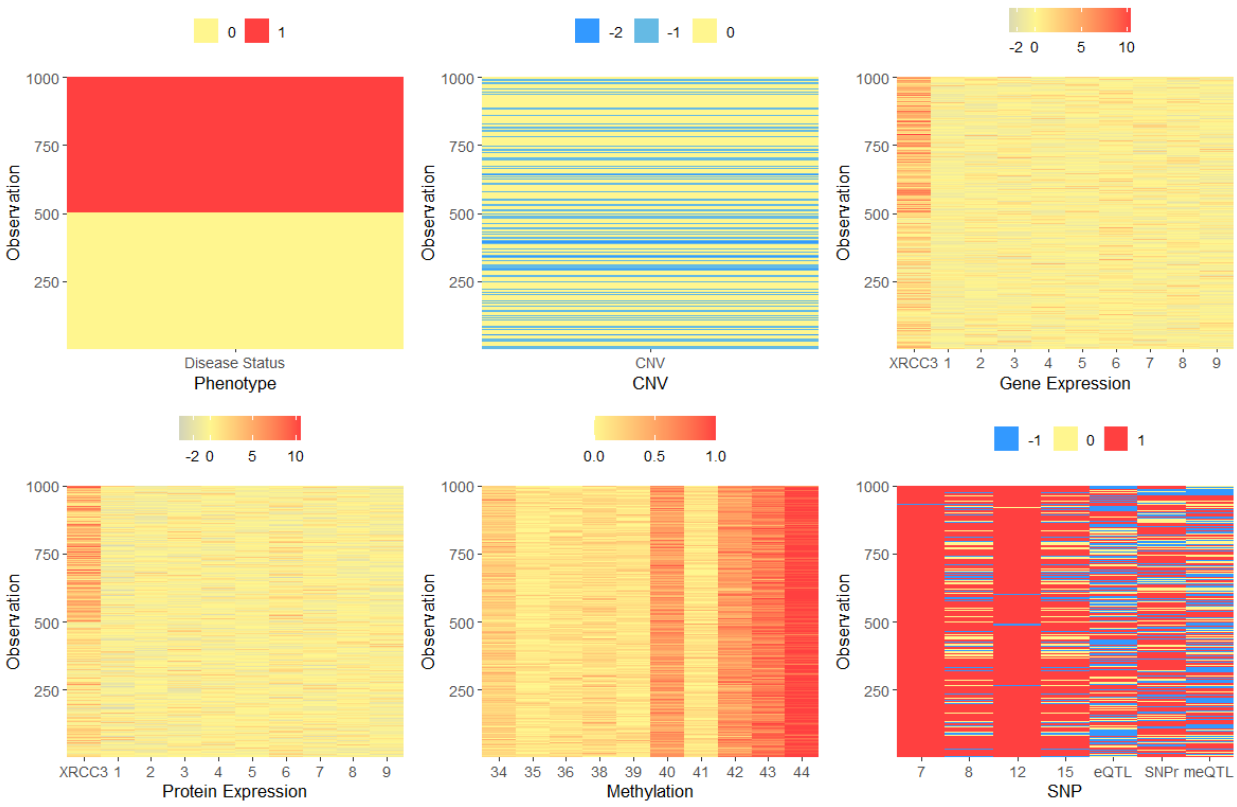
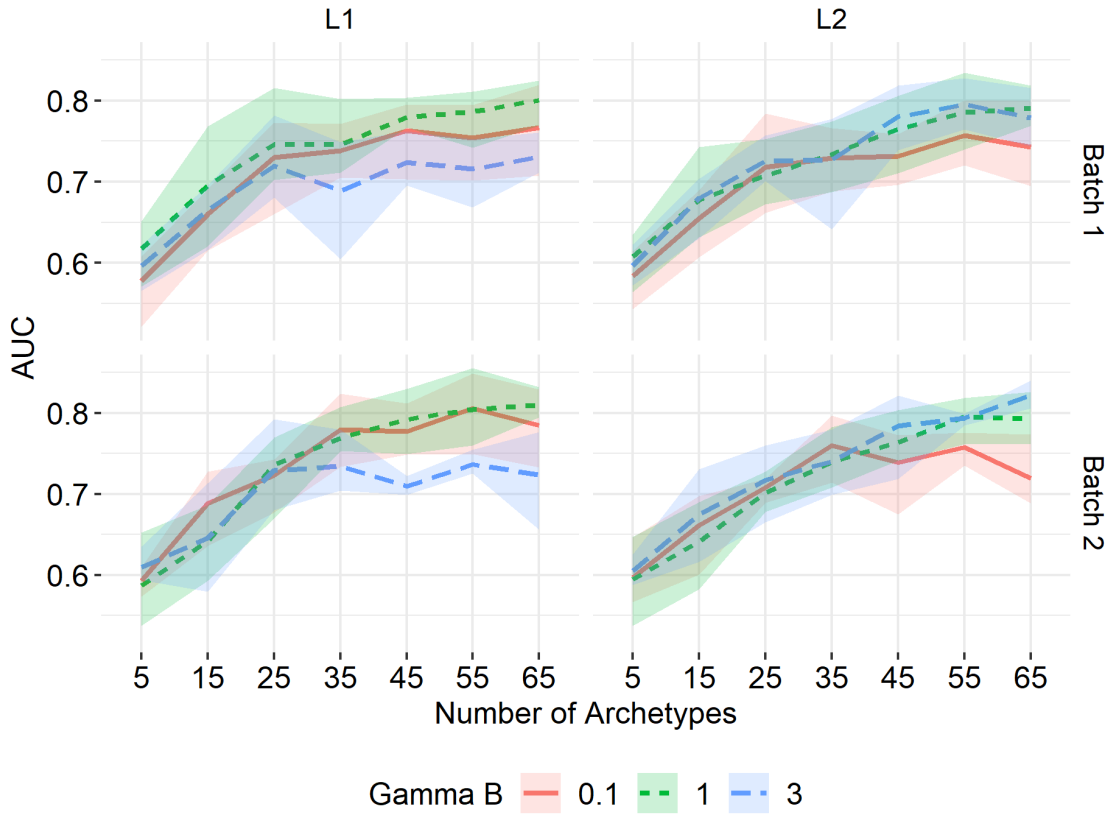
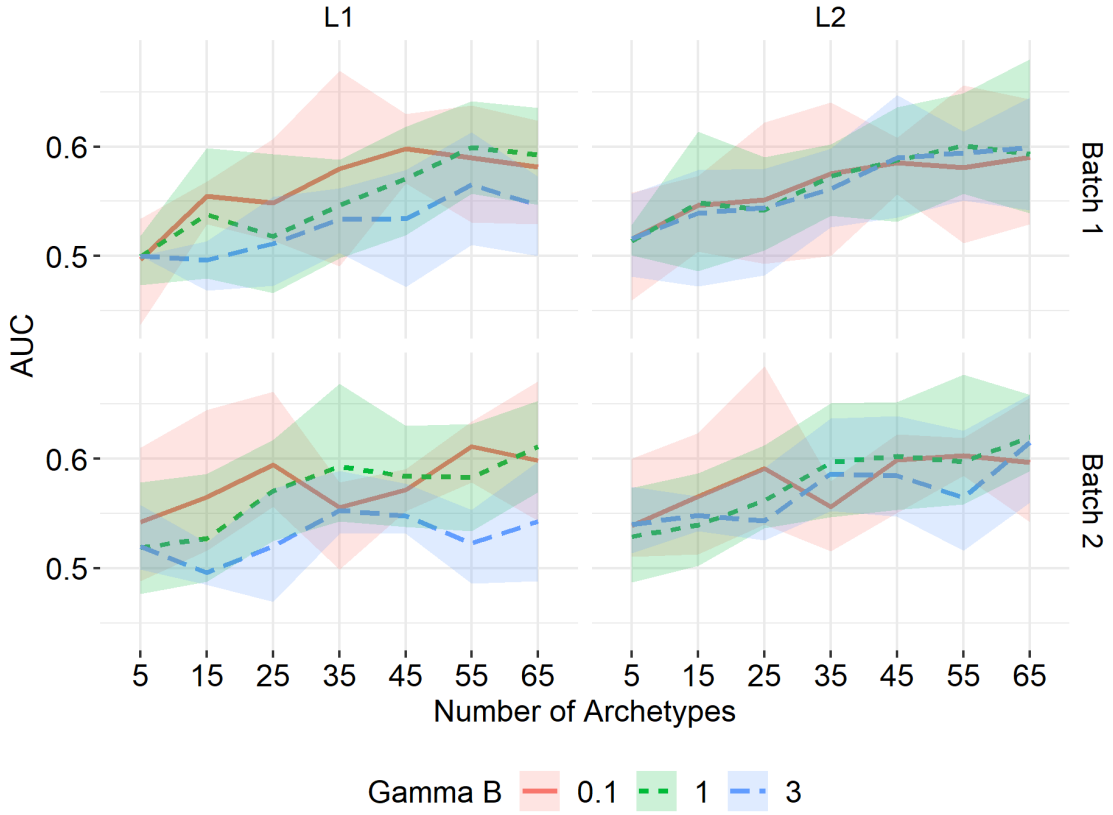


Figure 3.3: *Representation of Simulated Data. The x axis represents a sample of variables for each dataset and the y axis represents all 1000 individuals from training data of Scenario 1. The color then represents the value of the variables for each individual.*

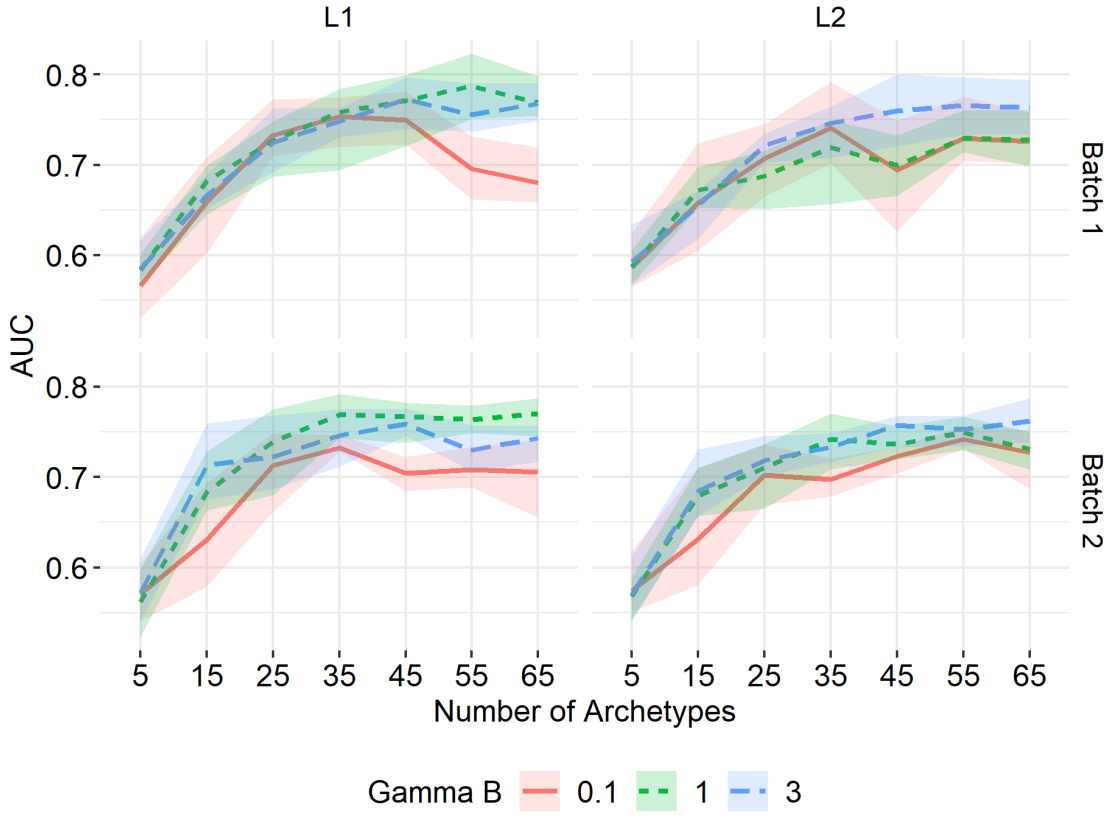
Therefore, considering each data type, for GLRM adjustment, adequate loss functions must be selected for each variable. A logistic loss was used for Disease status (boolean variable), and a Huber loss for the methylation. As for CNV, SNP, gene and protein expression the Quadratic loss was used. Since our goal is to predict data not observed, we tried the L1 and L2 regularization functions on the matrix of variables (matrix $\boldsymbol{B}$) to verify which one results on a better prediction. For the matrix of the observations (matrix $\boldsymbol{A}$), a L2 regularizer was used. The choice of parameters was made based on two randomly chosen validation batches. All variables were scaled and an offset was added. A grid search was performed with rank values of 5, 15, 25, 35, 45, 55 and 65 (where 1 of each corresponds to the offset) , γ_a values of 0.1 and 1 and γ_b of 0.1,1 and 3. Other parameter in between and higher than these were also adjusted (results not presented), they were not as informative but time consuming. Data was divided on 5 fold, and, for each fold, training and validation data were concatenated with disease Status missing for validation data. Then, imputed values for Disease Status were compared with real observed values for the validation dataset and the Area Under the Curve (AUC) metric was computed. The chosen parameters are then those that maximizes the AUC.



(a) *Scenario 1.*



(a) *Scenario 2.*



(b) *Scenario 3.*

Figure 3.5: *Values of AUC of different parameters (γ_a equals to 1). Different colors represent different values of γ_b. Lines represent the mean value for the 5 folds, and the bands represent the minimum and maximum value obtained on the 5 folds.*

In general, different values of γ_a resulted on very similar metrics, thus only those for γ_a equals to 1 (default value) were represented on Figure 3.5. As expected, both batches showed similar behavior. In all cases, the γ_b parameter did not impact the AUC; that is, metrics were similar across all ranks for different values of the regularization parameter, specially lower ones. In contrast, for Scenarios 1 and 2 (n = 1200), when using the L1 function, higher regularization did not increase the AUC, whereas γ_b of 0.1 and 1 showed similar results, for some ranks more than others. As for Scenario 3, the one with more observations (n = 4000), the GLRM was able to predict disease outcome better for higher regularization parameter (γ_b of 1 and 3). For all scenarios, GLRM with higher ranks tended to predict the disease outcome better. For Scenarios 1 and 3, the AUC had almost a linear increase from rank 5 to 25, then it stabilized and the increase rate has become lower – or declined as observed for γ_b on scenario 3 with L1 loss. As for Scenario 2 there is a gradual increase, however the mean AUC varied from 0.5 to 0.6 only. The same parameter were chosen for all three scenarios: the default regularization parameter, γ_a and γ_b equal to 1; and rank of 55, since for all schemes it appeared as one of the higher AUC.

To compare with results from Chung and Kang, 2019, 100 out of the 1000 batches were randomly selected for each scenario, with disease status not observed for validation data. Imputed values were compared with observed ones to evaluate the predictive ability of the model once again. The mean AUC and its Standard Deviation is presented on Table 3.1 alongside the same metric obtained for the 4 other integration techniques by Chung and Kang, 2019. For the scenarios with stronger genetic effect, the GLRM was better than the CANetwork approach only, but with mean AUC closer to 0.8, which is close to the 3-stage and the RFomics consequently. ATHENA was by far the best integration method to predict the disease outcome. However, it has a high computational cost, and when the genetic effect is weaker (scenario 2), the GLRM outperformed all the other methods, with a mean AUC of 0.58. An interesting result is that, for all the other 4 methods, the prediction of disease outcome was better with more observations (scenario 3 vs scenario 1), as expected. However, the GLRM prediction was better for the scenario with fewer observations (scenario 1), than for scenario 3, which did not happen o the other methods. That could be explained by the fact that the parameter choice for scenario 3 was made to be the same as the other two, but probably other parameter would yield on better results, such as a higher regularization parameter, for example.

Table 3.1: *Mean and standard deviation of the Area Under the Curve (AUC) for different methods.*

Scenario	**3-stage**	**RFomics**	**CANetwork**	**ATHENA**	**GLRM**
1	0.821 (0.028)	0.825 (0.028)	0.626 (0.037)	0.964 (0.017)	0.7803 (0.036)
2	0.501 (0.042)	0.511 (0.038)	0.529 (0.027)	0.509 (0.041)	0.5834 (0.035)
3	0.825 (0.013)	0.845 (0.013)	0.679 (0.019)	0.969 (0.005)	0.7433 (0.024)

Source: Data from Chung and Kang, 2019.

In practice, it is important to have a good prediction ability, but it must also be computationally feasible to do so. That means that the runtime is an important aspect and needs to be considered.

Table 3.2 represents the mean runtime in seconds of all five methods. It is important to highlight that the information for the 3-stage, RFomics, CANetwork and ATHENA were obtained from Chung and Kang, 2019, and not reproduced by us. So, due to differences on computer speed, memory, processing and many other factors, the numbers cannot be directly compared. However, it is possible to note that ATHENA takes more than an hour to run on the smallest database (Scenario 1) and more than a day for the larger one (Scenario 3). In contrast, GLRM takes only 5 to 10 minutes to run, which is still longer than the other methods, but a lot faster than ATHENA, which performs better than any other.

Table 3.2: *Processing runtime (in seconds) for different methods based on 100 batches of training and validation datasets.*

Scenario	3-stage	RFomics	CANetwork	ATHENA	GLRM
1	72.9	37.78	40.54	39,533.91	350.47
3	230.32	143.91	94.16	113,872.50	627.97

Source: Data from Chung and Kang, 2019.

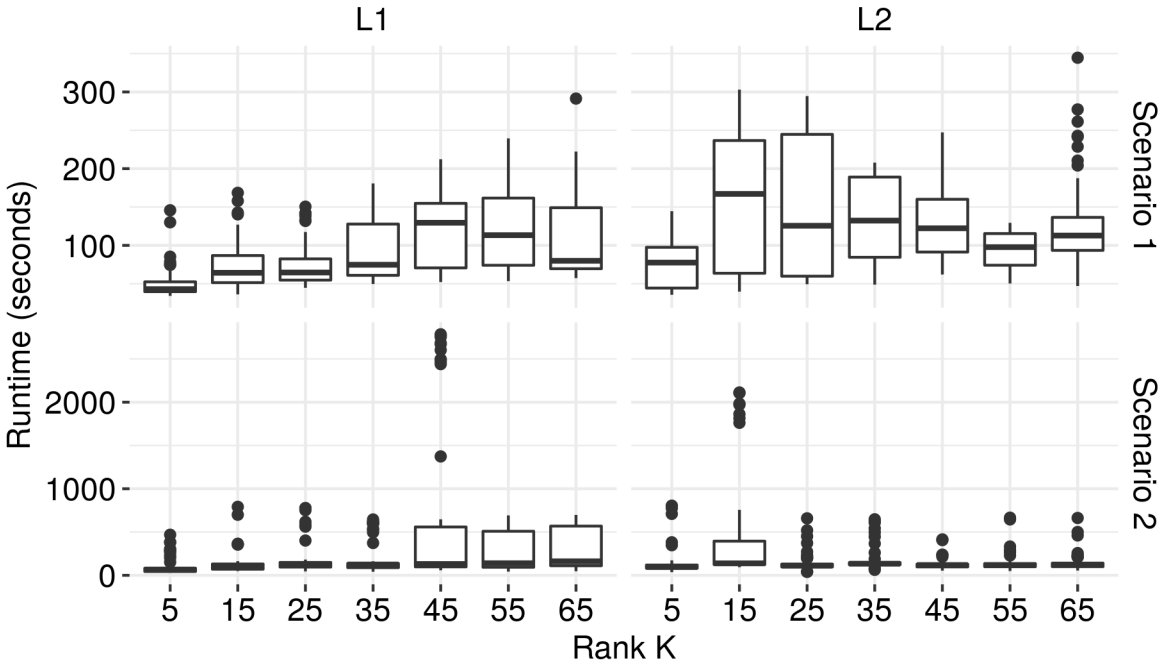


Figure 3.6: *Processing runtime (in seconds) for each rank based on two random batches for training data.*

In some cases, the GLRM can even be faster, since it is directly related to the rank. To illustrate this, Figure 3.6 shows the mean runtime by rank and regularizer function, calculated based on 30 values (combination of 2 batches, 2 values of γ_a and 3 values of γ_b). When using a L1 loss, as the rank increases the run time also increases. Alternatively, when using the L2 loss, ranks from 15 to 35 took longer to run. Also, note that the scales are different, so even though both scenarios have the same number of observations, the second one took longer to process. Therefore, the runtime should be an aspect taken into consideration when choosing the parameters. In this small example,

it was feasible to use higher ranks; but when working with high-dimensional datasets choosing a lower rank could allow for GLRMs to run at a feasible time.

3.1.3 Flexibility

The analyzes presented in Section 3.1.2 considered simple loss functions and regularizers. However, GLRM is very flexible by allowing different loss and regularizer functions, as weights for each variable. In this example, there are data from SNPs, that could be interpreted either as a discrete quantitative variable or as categorical. In the results presented before, a quadratic loss function was chosen for SNPs data, but another possibility would be to use a categorical loss. We then fitted a GLRM considering a multinomial loss function for SNPs. In this case, there was not a significant gain on predictive ability, and the processing time increased substantially (data not shown).

Then, still trying to find a better model, different weights were chosen. In some studies it may be known that some variables are more important to understand a phenotype than others. For example, let's assume the pathway for breast cancer is unknown; and we note that gene XRCC3 is differently expressed for cases and controls (Figure 3.2). Thus, a possibility is to give larger weights for the variables of gene and protein expression for this gene. The processing time in this case did not change, however it did not lead to an increase in predictive ability; in fact, in some cases there was a decrease. Another possibility would be to assign different weights to the outcome, since, in this case, the goal is to predict it. Such strategy would indicate the importance of this variable. By doing so there is an increase on the mean AUC, compared to the default, where all variables have weights equal to 1 (Figure 3.7).

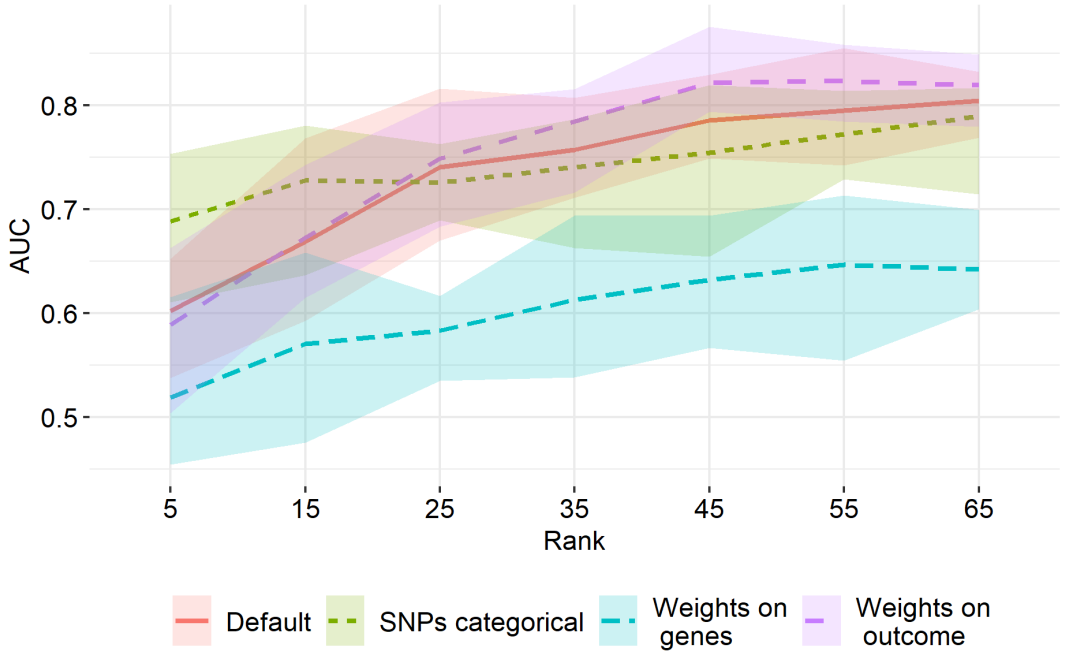


Figure 3.7: *Values of AUC for different possibilities of flexibilization. For scenario 1 with L1 loss, γ_a and γ_b equals to one.*

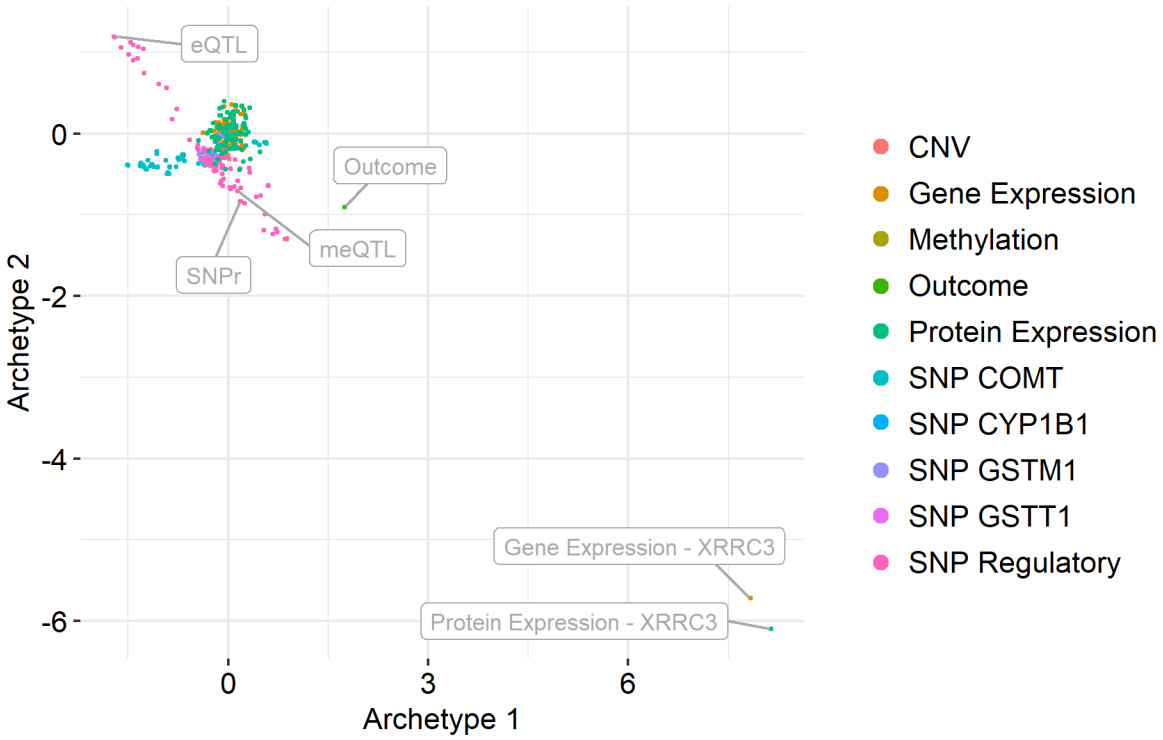
Therefore, in this specific example the possibilities of modifications did not lead to significant increases on the predictive ability. However, it served as an illustration of the GLRM flexibility, allowing for several possibilities of adjustments. It is worth noting that, when dealing with real datasets, choosing different loss function, or considering weights for variables of interest might be necessary to guarantee a good fit of the data analyzed. That is, this modifications might be necessary in order for the GLRM to achieve its best performance.

3.1.4 Data visualization

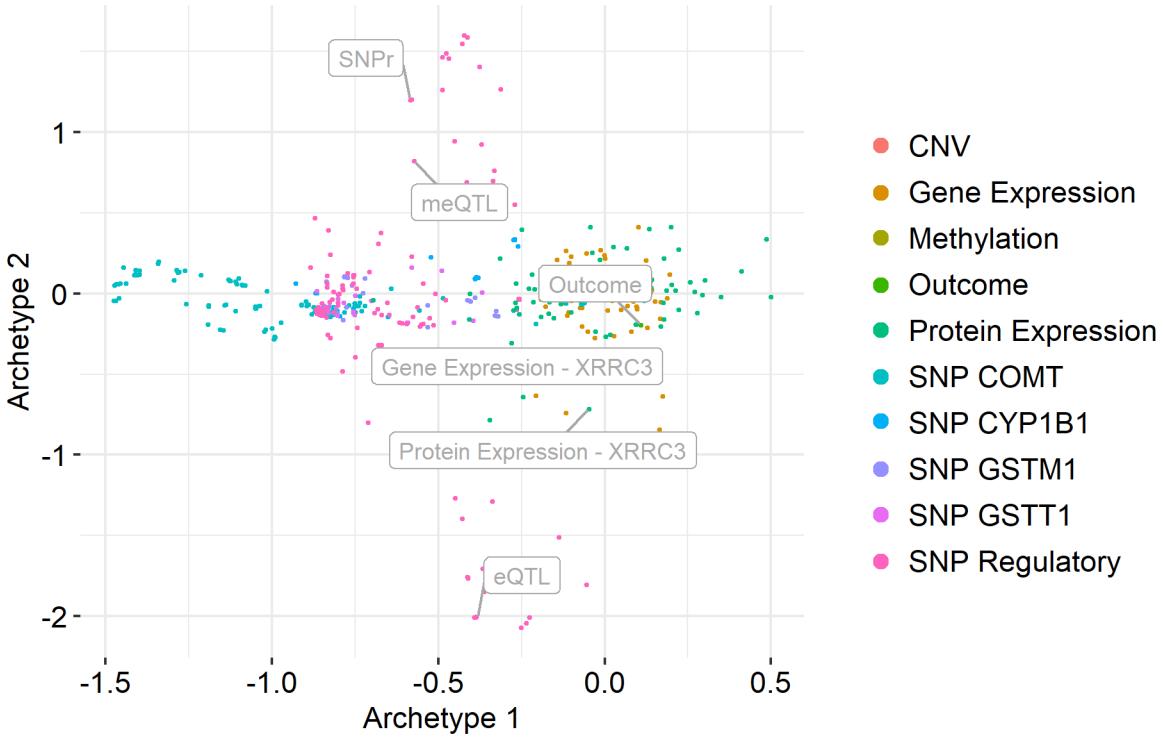
Besides being able to predict disease status, the GLRM has the advantage of allowing data visualization. From the matrix factorization we have the numerical representation of observations (columns of matrix $\boldsymbol{A}$) or variables (lines of matrix $\boldsymbol{B}$) for each archetype. Thus, Figure 3.9 has the plot for the variable representations on the first two archetypes for each scenario. Dots represent the variables of interest, with colors varying by its data type (either CNV, SNP, gene expression, methylation, etc.). All three figures display many values close or equal to zero, as a result of the L1 regularization.

Figures 3.8a and 3.9a show the representation of scenarios 1 and 3, both with a stronger genetic effect, but with different number of observations. In both cases the variable representations are similar, indicating that the number of observations does not influence on how variables are represented. Also, there is clustering of variables based on their types and regions. That is, SNPs from the regulatory region (in pink) were grouped differently than SNPs from the COMT gene (in blue), for example. Still, on Figures 3.8a and 3.9a, gene and protein expression of gene XRRC3 have the highest absolute value on both archetypes, which makes sense since cases and controls have different expression levels (see Figure 3.2), when genetic effect is high. That being said, the outcome also has values greater than zero, on the same direction of expressions of gene XRRC3. All SNPs from the regulatory region, including eQTL, meQTL and SNPr are the line $y = -x$, again on the same direction of gene XRRC3 expression. SNPs from gene COMT (in which some are related to the meQTL, see Figure 3.1) all have negative values on the second archetypes, as does the meQTL. SNPs from other genes, methylation, and gene and protein expression (for all the 99 genes other than XRRC3) have values floating around zero for both archetypes.

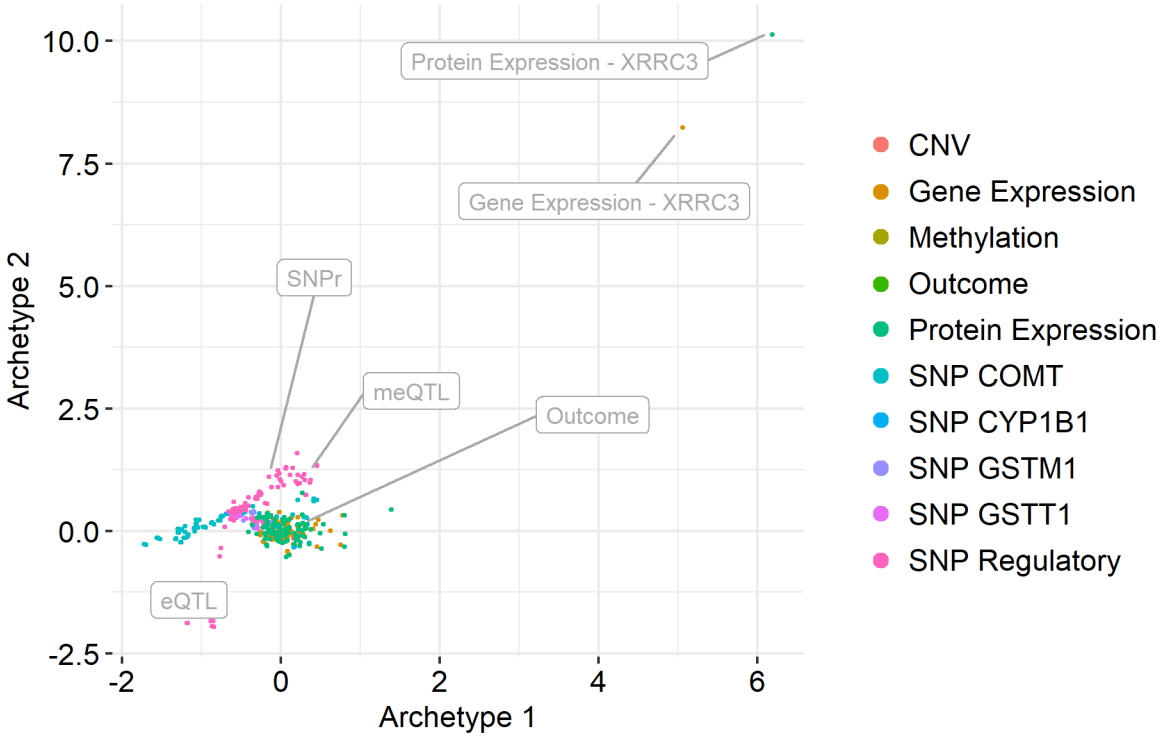
Now, for Scenario 2 (Figure 3.8b), the genetic effect is weaker; thus, there is not a clear difference between gene and protein expression of gene XRRC3 for case and control. Consequently, gene and protein expression of gene XRRC3 received values closer to the expression of other genes, which are closer to zero. In this case, SNPs from the regulatory region have values close to -0.5 on archetype 1 and from -2 to 2 on archetype 2. SNPs from the COMT are all close to zero on the second archetype, but all negatives for the first one, as does the meQTL. Methylation again has values around zero for both archetypes. In all three cases, we note that the eQTL has values in the opposite direction of meQTL and SNPr, which is curious, since all three of them have the same fold change between cases and controls.



(a) *Scenario 1.*



(b) *Scenario 2.*



(a) *Scenario 3.*

Figure 3.9: *Numerical representation of variables (lines of matrix B) for different Scenarios.*

3.2 Real Data

The second application is on a real dataset, focused on understanding and analyzing Chronic Kidney Disease (CKD). CKD consists on a reduced or damaged renal function, leading to abnormal pressure and filtering, reabsorption and degradation of metabolics, for example. Nowadays, CKD is a major public health concern, which makes preventive measures highly necessary. This disease is defined by biomarkers of kidney damage or renal function, such as urinary protein levels, albumin excretion or estimated glomerular filtration rate (eGFR) (Hill et al., 2016). However, these measures are still limited and dependent of other variables. Therefore, there is an interest on finding new markers of kidney functions to allow for a better understanding of CKD progression. Since kidneys function impact directly on circulating metabolites, changes on metabolic levels may biomarkers for kidney disease. Thus, the study of the human metabolome can be promising in this context.

Considering the goal of finding new markers useful for assessing kidney function, untargeted metabolites are an interesting possibility. They give the opportunity of finding new molecules associated with the outcome, by not targeting a specific metabolite. Analysis of metabolomics data in CKD studies have been performed in order to find associated metabolites, either with other biomarkers via regression models (Sekula et al., 2016; Yu et al., 2014; Luo et al., 2019; Rhee et al., 2019; S. M. Titan et al., 2019) or with risk of death or replacement therapy via Cox models (Silvia M. Titan et al., 2019; Hu et al., 2020). In both cases, finding metabolites associated with

the decrease in eGFR, or increase in mortality risk, may allow for the identification of metabolic pathways related to CKD.

Even though working with untargeted metabolites has the advantages of allowing the discovery of different associated metabolites, considering the CKD study it has the disadvantage of having many missing data. According to Armitage et al., 2015, missing data in metabolomics are receiving a lot of attention, since they can comprise more than 70% of a single metabolite, which can lead to inaccurate statistical analysis and data interpretation. A flexible way to deal with missing data is by imputation; however, different imputation methods can affect data interpretation (Gromski et al., 2014). Some possibilities are to impute a constant low value, such as zero or the lowest value on the dataset (Xia et al., 2009); impute values drawn from an estimated distribution (Chen et al., 2011; Richardson and Ciampi, 2003); impute by the median or mean of each metabolite (Gromski et al., 2014); or impute by using multivariate methods, such as MICE K-nearest neighbors (KNN) (Troyanskaya et al., 2001; Tutz and Ramzan, 2015) or multiple imputation by chained equations (MICE) (Van Buuren, 2007).

Therefore, in this application different imputation methods will be compared with the GLRM in two contexts. First, the impact of imputation on the selection of metabolites associated with the eGFR will be evaluated. Secondly, the explained variance of a Cox proportional hazard model on the mortality risk will be compared under different scenarios, with distinct imputation methods.

3.2.1 Dataset

This project analyzes data from the PROGREDIR cohort (Domingos et al., 2017). This cohort comprises 454 patients with moderate to advance CKD from São Paulo, Brazil. The data was collected to allow researches to test the performance of new biomarkers and high throughput technologies for CKD progression. Therefore, clinical variables were collected, focused on cardiovascular and renal parameters. Besides data on time to event (death and renal replacement therapy), there are other 357 variables, including

- 9 anthropometrics (circunferences, weights, ratios);
- 41 laboratory (serum fasting samples, two-hour glucose-tolerance test, spot urine, 24-hour urine, gasometry, total and partial proteins);
- 14 cardiac and vascular evaluation (transthoracic echocardiography, heart rate variability, retinography);
- 14 clinical (age, sex, race, smoking, drinking, presence of diabetes, AVC, heart attack);
- APOL gene (Genovese, Friedman, and Pollak, 2013) and ancestrality scores;
- 10 medications (number of medications and 9 different medications);
- Batch and 264 (after preprocessing) untargeted metabolites (see Table A.1).

Data will be used in different contexts. Metabolites and some clinical and laboratory variables will be considered important predictors. Time to event and eGFR will each be considered response variables on the next two applications. Also, the idea is to integrate all the variables to gather information for data imputation. However, different data types and scales are included and must be considered for our analysis.

3.2.2 Impact of imputation on selected metabolites

One of the laboratory variables is the estimated glomerular filtration rate (eGFR) which tells how well the kidneys are working; for this reason it is one of the main biomarkers for CKD. In fact, different stages of CKD are defined based on eGFR levels:

- **Stage 1:** eGFR greater than 90, with other signs of kidney damage;
- **Stage 2:** eGFR between 89 and 60, with other signs of kidney damage;
- **Stage 3:** eGFR between 59 and 30, moderate kidney damage;
- **Stage 4:** eGFR between 29 and 15, advanced kidney damage;
- **Stage 5:** eGFR less than 15, kidneys close to failure or already failured.

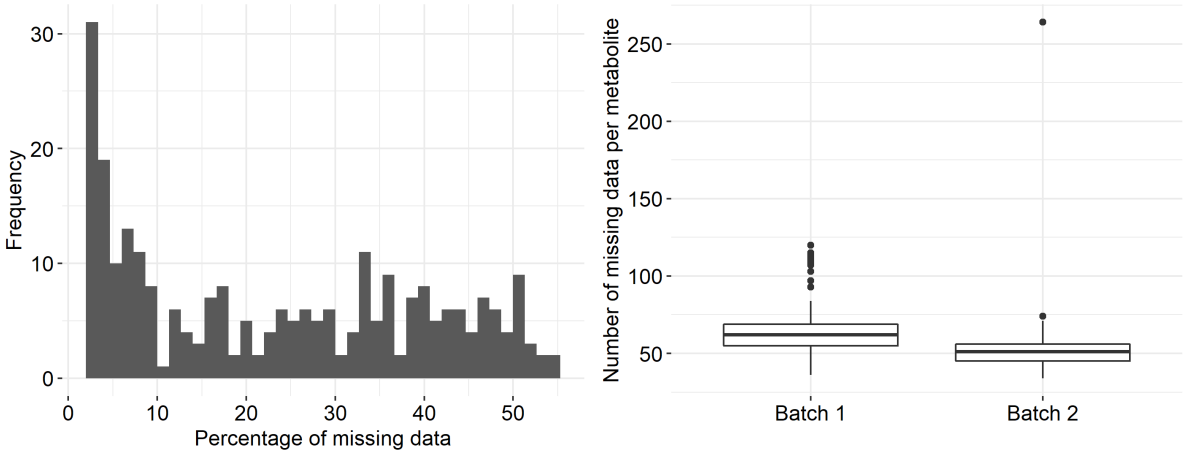
Thus, finding new markers related to eGFR may allow for the identification of biomarkers related to CKD progression. As stated before, metabolomics are a possibility of markers, since kidney functions are directly related to it. Silvia M. Titan et al., 2019 showed potential metabolites related to CKD using the PROGREDIR cohort data. In the article, missing data were considered unavailable, but it is suspected that they can affect data interpretation (Armitage et al., 2015). So, motivated by this, there is an interest in studying the impact of missing data in the selected metabolites. Since there are many variables of different types available, we can use the ability of GLRM to deal with missing data to impute values based on the information of all 357 variables.

Therefore, to achieve our goal, first missing metabolites will be imputed either by GLRM, or mean, or median or KNN (recommended by Armitage et al., 2015; Di Guida et al., 2016 and Do et al., 2018). Then, linear regression models will be fitted for eGFR for each metabolite (log2), adjusting by batch, sex, age, diabetes status, systolic blood pressure (SBP) and smoking status. In addition, after a Bonferroni correction, metabolites with p-values smaller than $\frac{0.05}{264}$ were considered associated. Finally, selected metabolites will be compared for different imputation methods, and contrasted with those selected without imputation.

To verify whether the correct metabolites are being selected, a simulation framework was proposed. A complete dataset (without missing data) was obtained by excluding all metabolites with more than 3.5% missings, resultinging a dataset with 432 individuals and 31 metabolites. Then, missing data were incorporated to the dataset to be further imputed and results were compared for each method. By doing so, we assume that the metabolites selected with the complete data are the

ones that are truly associated. This is a strong assumption that does not necessarily hold, but it is a common approach (Armitage et al., 2015) to allow for better interpretation.

When incorporating missing data it is important that the data reflects the real data to avoid biased evaluations (Do et al., 2018). Figure 3.10 shows a characterization of missing metabolites on the dataset. It is possible to verify that metabolites have up to 55% of missing data (Figure 3.10a), and around 30 have less than 3.5%. Also, the metabolomics data were collected in two different batches. Do et al., 2018 presents the importance of analyzing the batch effect on missing data, and it seems to be the case here. Figure 3.10b presents the number of missing data per metabolite for each batch, and it shows that there are more missing for batch 1 than 2.



(a) *Histogram of percentage of missing data per metabolite.*

(b) *Boxplot of number of missing data per metabolite for each batch.*

Figure 3.10: *Characterization of missing metabolites on dataset.*

Therefore, missing data were included in the complete dataset in a simple scheme. First, the proportion of missing metabolites was calculated for each individual from each batch. Then, for each metabolite, one proportion was randomly selected from the individuals from batch 1, and the same procedure was done for batch 2. Finally, the corresponding number of individuals for each proportion was randomly selected from each batch and considered missing. This scheme was repeat 100 times to obtain different configuration of missing data.

For the imputation methods, parameters must be chosen for KNN and GLRM. The KNN depends on the number of nearest neighbors, so we evaluated numbers of 2,5,10,15 and 20. Results were equivalent in all scenarios for 10 or more nearest neighbors, which is consistent with results from Do et al., 2018. For the GLRM, non-negative regularization was used on both the scores of the individuals (matrix $\boldsymbol{A}$) and loadings of the variables (matrix $\boldsymbol{B}$), to guarantee that only positive values would be imputed for metabolites. Loss functions used for each type of variables are described in Table 3.3. Data were scaled and an offset was added. As for the rank, values of 2, 4, 10, 15, 20 and 25 were considered and results for each of them will be presented next. Again, since an offset was added, the true rank is one less than the ones considered.

Table 3.3: *Loss functions used for each variable.*

Variables		**Loss Function**
eGFR		Quadratic
Anthropometry		Huber
Laboratory		Huber
Heart		Huber
Clinical	Quantitative	Huber
	Categoricals	Multinomial
	Boolean	Logistic
APOL Gene		Quadratic
Ancestrality		Huber
Medications	Number of oral medications	Poisson
	Indicative of use	Logistic
Batch	Batch	Logistic
	Metabolites	Huber

When there where no missing data, 11 out of the 31 metabolites were associated with the response – which is represented by the horizontal line in Figure 3.11. When missing data is incorporated into the dataset, the ability to detect associated metabolites decreases, since only between 7 and 10 are selected. When imputing data with mean or median of each phenotype, there is an increase of variability of the number of selected metabolites, getting up to 11. However, the median number of metabolites associated decreases when compared to the scenario without imputation. When data is imputed with the KNN 50 out of the 100 repetitions resulted in 10 or 11 metabolites selected. Thus, these results show that missing data impacts the ability of finding associations. Depending on the missing pattern, imputing metabolites with its mean or median can improve or not the results. KNN seems to impute data closer to the original, based on the number of selected metabolites.

Moving on to results for the GLRM imputation, Figure 3.11 shows that, once again, the choice of the rank interferes on results. Lower ranks (2 and 4) do not seem to impute data informative enough to improve results from the case without imputation. As the rank increases so does the number of metabolites associated found. Compared with other imputation methods, GLRM apparently adds more information to missing data, which leads to more associated metabolites than with the complete dataset. If we assume that the 11 metabolites selected by the complete model are truly the ones associated with eGFR, this would mean that imputing data with the GLRM results in more false positives than by using other methods. However, since the real biological process of metabolites and CKD is still unknown, we cannot really access if they are really false positives, or if the imputation adds information that allows for the identification of other metabolites.

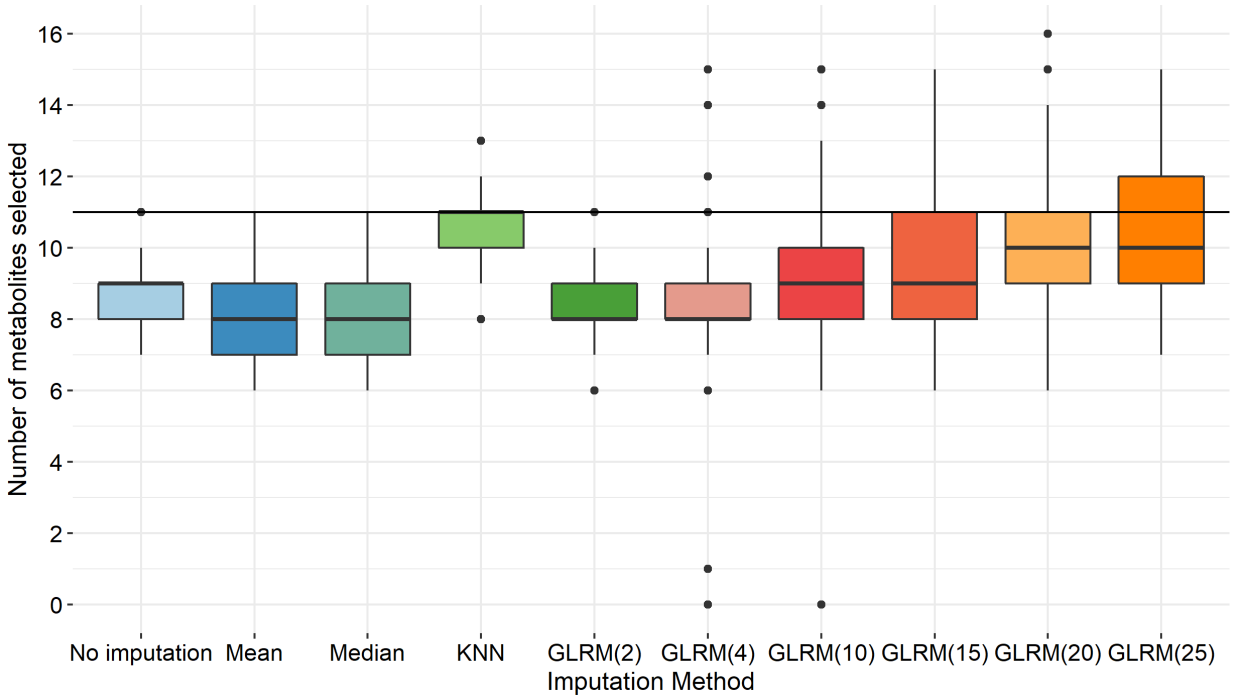


Figure 3.11: *Boxplot of the number of metabolites selected in association with eGFR for each type of imputation. Number of selected metabolites (11) on complete dataset is represented by the horizontal line.*

Now that the number of selected metabolites were compared, we will verify if the correct metabolites were being selected. Figure 3.11 shows how many variables were associated with the response. However, it is important to assess if the correct metabolites are being selected, assuming results from the complete dataset as the gold standard. To do so, the selected metabolites with each imputation were compared with the correct ones (from the complete dataset) – for each of the 100 replicates. Figure 3.12 relates this number with the total number of selected metabolites. The ideal scenario would be that all 11 metabolites were selected. It is possible to note that the results without imputation or imputed with the mean, median or GLRM with rank 2 are similar: less than 11 metabolites selected in most cases, and from those selected not all are the correct ones. When imputing with the KNN algorithm, the correct variables seem to be correctly selected, in the sense that, 8 out of 8 associated metabolites were the correct ones, for example. As for the GLRM, on Figure 3.11 the one with rank 25 resulted on more metabolites selected. Considering this, Figure 3.12 shows that for 11 or more metabolites selected, at least 9 of them are correct. In contrast, by imputing data with GLRM with rank 20 all correct metabolites are selected when the total number is greater than 12.

In conclusion, the presence of missing data reduced the power of identification of associated predictors for the eGFR. Imputing these values by the mean or median does not seem to bring improvement, and in some cases can mask the effect of metabolites, depending on the missing pattern. KNN was the most consistent imputation method and was able to correctly select all the metabolites in most of the cases. For the GLRM, the rank once again has an impact on the results and must be carefully chosen. Higher ranks seem to incorporate more information into the data, allowing for the correct metabolites to be all selected in more cases. GLRM results showed a

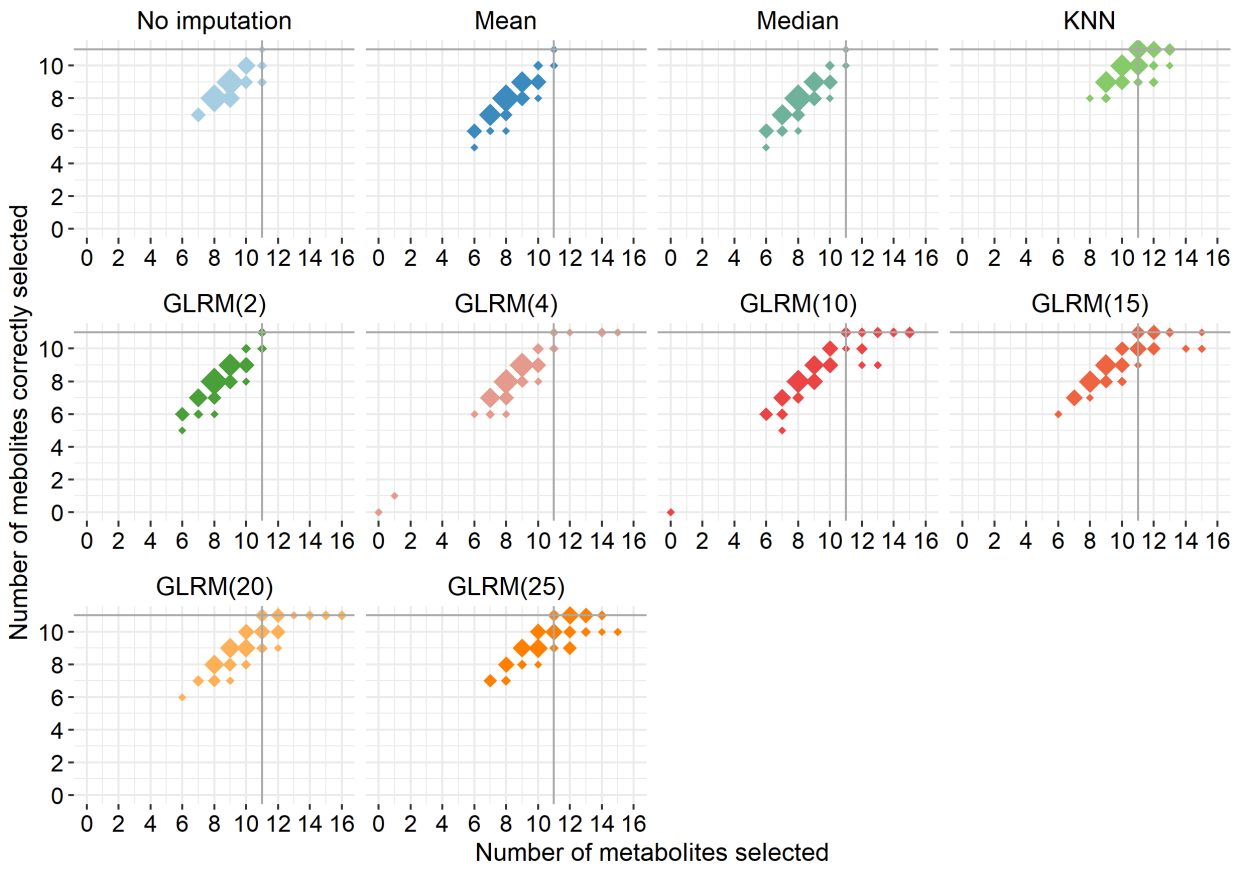


Figure 3.12: *Comparison between the number of correct metabolites selected and the total number for each imputation method. Results are for 100 repetitions, and point sizes are proportional to the number of occurrences.*

higher variability, indicating that it is dependent on the missing pattern, in the sense that it can perform better or worse under different scenarios.

After understanding how each imputation method performs in a simulated context we can move on to the complete dataset analysis. Once again, our goal is to find metabolites associated with the eGFR that could lead to a better understanding of CKD. Metabolites found in this analysis would need to be better explored on future studies, for example, by using targeted metabolites. Thus, different imputation methods were used and compared: median, KNN with 10 nearest neighbours and GLRM with rank of 25 and same loss functions as before. Without imputation 73 metabolites were associated with the response, as in Silvia M. Titan et al., 2019. From the simulation done before we expect that most of these, if not all, would also be associated if all the information were present. In fact, as it was the case before, probably some information was lost with the missing data that we aimes to gain with imputation. When data was imputed with the median, KNN and GLRM, 67, 71 and 80 metabolites were selected, respectively. Figure 3.13 shows a Venn diagram for the selected metabolites for each imputation method. It shows that 63 metabolites were commonly selected, independently of he imputation method. Both median and KNN imputation resulted in 65 metabolites in common with those selected without imputation. KNN selected 6 more than without imputation, 2 of which are common with the median and 4 in common with the GLRM.

When imputing data with the GLRM, all 73 metabolites (from without imputation) were also selected, with an addition of 7 others.

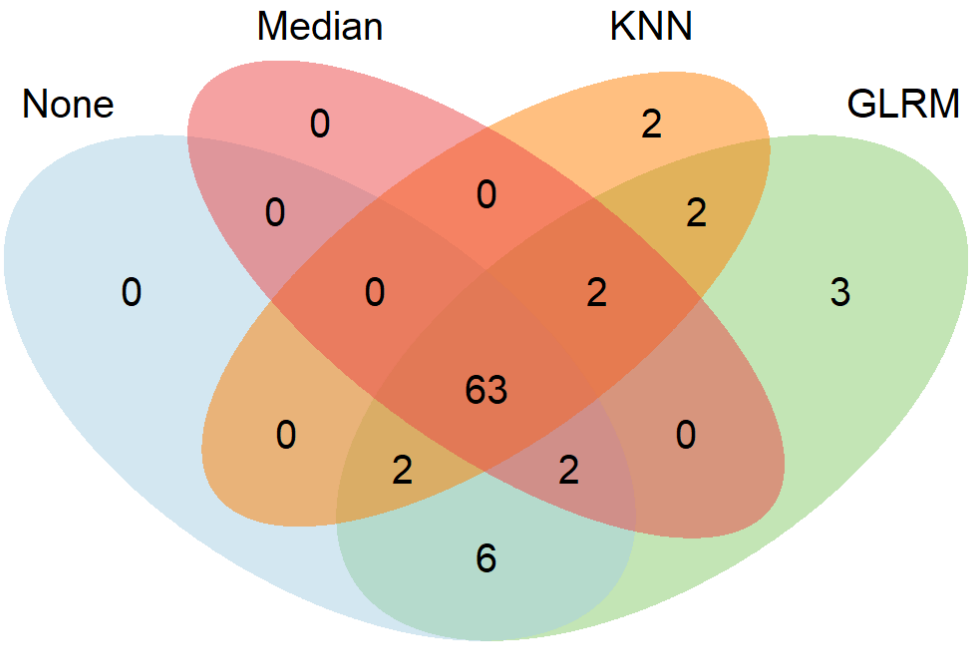


Figure 3.13: *Venn diagram comparing selected metabolites with each imputation type.*

To better compare the selected metabolites, Table 3.4 shows metabolites (see Table A.1 for correspondent names) that were selected by some models, but not by others; that is, metabolites selected, or not, by all methods are not shown. From all the metabolites, 19 had different behaviors depending on the imputation method. Out of the 19, 6 were only selected without imputation or imputing with GLRM (X4, X22, X72, X73, X162 and X178). Additionally, 4 more metabolites were also selected when imputing with GLRM and without imputation, however were not selected by one of the other two imputation methods. More specifically, X89 (5α-Cholestanol) and X232 (Glycine) were considered associated in one of the two Brazilian replicate studies presented by Silvia M. Titan et al., 2019, but were not selected when imputing with the median and KNN, respectively. Another interesting result is regarding metabolite X109 (Ribitol). Ribitol was selected in both replicate studies, but for this dataset its association can only be detected when imputing data using KNN or GLRM.

Table 3.4: *Selected metabolites for each imputation type.*

	None	Median	KNN	GLRM
Number of selected metabolites	**73**	**67**	**71**	**80**
X4	✓			✓
X22	✓			✓
X28				✓
X34				✓
X56				✓
X71	✓	✓		✓
X72	✓			✓
X73	✓			✓
X89*	✓		✓	✓
X109**			✓	✓
X138			✓	✓
X142	✓		✓	✓
X162	✓			✓
X178	✓			✓
X190			✓	
X224		✓	✓	✓
X232*	✓	✓		✓
X281		✓	✓	✓
X289			✓	

* Present in at least one of the Brazilian replicate studies from Silvia M. Titan et al., 2019.

** Present in both Brazilian replicate studies from Silvia M. Titan et al., 2019.

Therefore, when studying the entire dataset, imputing data with median or KNN lead to the discovery of less metabolites associated with eGFR than without imputation, some of which are not the same. There is not an exact way to affirm which metabolites are the correct ones to be selected. In contrast, when imputing data with the GLRM, all metabolites selected by the other imputation methods were also selected, except for X190 and X289 selected with KNN. In this specific example, by imputing with GLRM more information is added to the dataset, which allows for the identification of more metabolites. Since the goal is to find possible biomarkers related to CKD, now there are 6 (or 8, if X190 and X289 are considered) more metabolites to be studied,

besides the already 73 found in the literature (Silvia M. Titan et al., 2019).

An important observation about the analysis done before is that different imputation methods were compared in a dataset with up to 50% missing data. However, articles, such as Armitage et al., 2015, show that an important aspect for the performance of imputation methods is the amount of missing data present. In the article, all imputation methods performed better when less data was missing. Also, missing percentages from 40 to 70% were considered in a "grey area", in the sense that some of the missing data might actually be truly missing and should not be imputed (or imputed with zero). Since KNN and GLRM can allow for imputation of zeros, this might explain why their performance was better than imputing by the mean or median. For comparison, the same analysis was performed for the dataset with only metabolites with up to 20% missing data (see Appendix Figures A.1,A.2 and A.3). By doing so, 136 metabolites were excluded. In this case, similar results and comparison among imputation methods were obtained, but with more consistent results. In the entire dataset analysis, data imputation did not lead to the identification of more metabolites than the ones already selected without imputation (Figure A.3). Thus, when data is available for more than 80% of the individuals, different imputation methods have similar performances, including the GLRM. However, when less data is available (i.e., the proportion of missing is high), some imputation methods may provide better results than other depending on the goal of analysis.

3.2.3 Impact of imputation and latent variables on explained variance

In addition to finding metabolites associated with eGFR, another possibility would be finding biomarkers associated with an outcome event in a CKD population. Both (eGFR or event status as the phenotype of interest) analysis can allow the identification of new biomarkers for better understanding CKD progression. For example, once metabolomic markers are consistently discovered, then checking individual levels of specific metabolites could provide a more accurate medical diagnostic. Thus, to achieve this goal, we will be analyzing the follow up data collected on death, on the PROGREDIR cohort dataset. Considering the same data, S. M. Titan et al., 2019 identified metabolomics biomarkers related to overall mortality and incident end-stage renal disease. Now, there is an interest in verifying if these metabolomic biomarkers have an effect on the explained variance of risk prediction models. However, the amount of missing data can harm the analysis, since all individuals have at least one missing metabolite, which requires imputation. Therefore, in the following analysis, the impact of different imputation methods on the predictive ability of models will be compared. Also, an archetype approach will be tested; that is, data integration will be done for finding latent variables, which will be predictors for the risk of overall mortality, in a CKD population.

As in S. M. Titan et al., 2019, first a Cox proportional hazards regression was performed on the death outcome (from 2019), for each metabolite, adjusted only by batch. Then, metabolites with false discovery rate q-values less than 0.05 were selected. Since the impact of imputation was already analyzed in the section before, for this example the metabolites were selected with

missing values considered unavailable. From the 264 metabolites 9 were selected: X50, X78, X83, X104, X111, X114, X233, X269 and X274 (see Table A.1 for the correspondent names). Then, Cox regression models were adjusted for the overall mortality with different covariates:

1. **Basic clinical model:** adjusted for sex, age, eGFR and diabetes status;
2. **Extended clinical model:** adjusted for sex, age, eGFR, diabetes status, race, hypertension status, previous myocardial infarction, previous stroke, smoking status, systolic blood pressure, diastolic blood pressure, body-mass index, waist to hip ratio, and levels of potassium, urea, creatinine, microalbuminuria, phosphorus, total calcium, parathormone, glycemia, Ggycated hemoglobin, total cholesterol, LDL-cholesterol, HDL-cholesterol, triglycerides, bicarbonade, hemoglobin and albumin;
3. **Metabolites model:** batch and previous selected metabolites only;
4. **Metabolites and basic clinical model:** batch, previous selected metabolites and the same covariates as the basic clinical model;
5. **Metabolites and extended clinical model:** batch, previous selected metabolites and the same covariates as the extended clinical model;
6. **Archetypes:** archetypes (scores) from the GLRM.

In the models considering metabolites, different imputation methods were considered: mean, median, KNN and GLRM. Once again parameters must be chosen. For the KNN, results for 2,5,10,15 and 20 nearest neighbors were evaluated and once again results were consistent for 10 neighbors. For the GLRM, the same variables and loss functions as in Table 3.3 were used with the addition of 2 variables: incidence of mortality (logistic loss) and time to event (absolute loss). The choice of the loss function for the time of event was done based on Korn and Simon, 1990; Henderson, 1995, which gives less weight to discrepant times of survival. Again, a non negative regularization function was used for both matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Data was standardized (scaled and with an offset), and ranks of 2,4, 10, 15, 20 and 25 were evaluated. In this example, different ranks provided similar metrics (see Appendix Figure A.4), therefore a rank of 15 was considered.

Additionally, instead of selecting specific covariates and metabolites to predict risk of mortality, we could integrate all possible variables to obtain archetypes, latent variables for each individual, and use them as predictors. To do so, GLRM with the same variables and loss functions from Table 3.3 were considered. In this case, a L2 regularization (with $\gamma_a = 1$) was used on the lines of the matrix related to the individuals (matrix $\boldsymbol{A}$) for better generalization, and a L1 regularizer (with default regularizer (γ_b) of one) was used on the loadings of the variables to impose a sparse structure – force some loadings to zero. Data was standardized and after a grid search of values of ranks (number of archetypes consequently) of 5,10,15,20,25,30,35,40 and 45 (not including the offset), were evaluated.

Two metrics were used to verify models predictive ability. The first is the R^2, which is based on the global likelihood test (Walker, 2003):

$$R^2 = \frac{1 - exp(-(L_0 - L)/n)}{1 - exp(-L_0)/n},$$

where n is the number of observations and L and L_0 are -2 times the log likelihood for the fitted model and a model with no predictive information, respectively. The R^2 varies from 0 to 1, where 1 indicates a predictive model that discriminates survival times perfectly. This metric is helpful to understand the association between the response variable and predictors, however it does not quantifies model's predictive power. Thus, another metric that will be used as a discrimination index is derived from Sommer's D_{xy} rank correlation (Korn and Simon, 1990). Given a random pair of observations from the population, their observed and predicted survival times are compared. Probability of concordance (individual who survived longer has higher predicted survival time), discordance and ties are calculated. Then, based on the probabilities calculated, the Sommer's D_xy is given by:

$$D_{xy} = \frac{\mathbb{P}(\text{concordance}) - \mathbb{P}(\text{discordance})}{1 - \mathbb{P}(\text{ties on observed values})}.$$

Then, Korn and Simon, 1990 chooses as a metric of explained variance the square of D_xy, to be equivalent to the squared of Pearson correlation on linear regression analysis. To asses the explained variance with D^2_{xy} a 30-fold cross validation was performed – and rank choice was made based on a 10-fold cross validation. Different metrics were then calculated: first, predicted values for the entire dataset were obtained from a model with all data (original); then, based on a model fitted on the validation dataset values predicted for the training and test data for each fold and the median was calculated; and finally the original metric was corrected by the difference between the median value for the training and test. During the cross validation, imputation on the training data was made without data from the test, and for the test the imputation was done with information from all observations – except for GLRM, in which survival was considered missing for test data. As for the archetypes, GLRM was modelled only for validation data and the loading of archetypes for each variable were obtained (matrix $\boldsymbol{B}$). Then, test data was transformed into a numerical matrix (boolean variables coded as 0 and 1 and categorical transformed to dummies), missing data was imputed with GLRM based on entire data, and scores were obtained.

So, before comparing different models, the rank of GLRM, i.e. the number of archetypes must be chosen. Figure 3.14 shows boxplots of values of D^2_{xy} for each fold, calculated on training and test data for different number of archetypes. For the training data, the metric does not vary much between folds, and as the number of archetypes increases, so does the explained variance, tending to 0.3. However, when calculating on the test data there is an increase on metrics variation and most of the explained information is lost. The rank that seems to better adjust the data is 15, since it has the highest median on the test data, and was chosen for the following analysis.

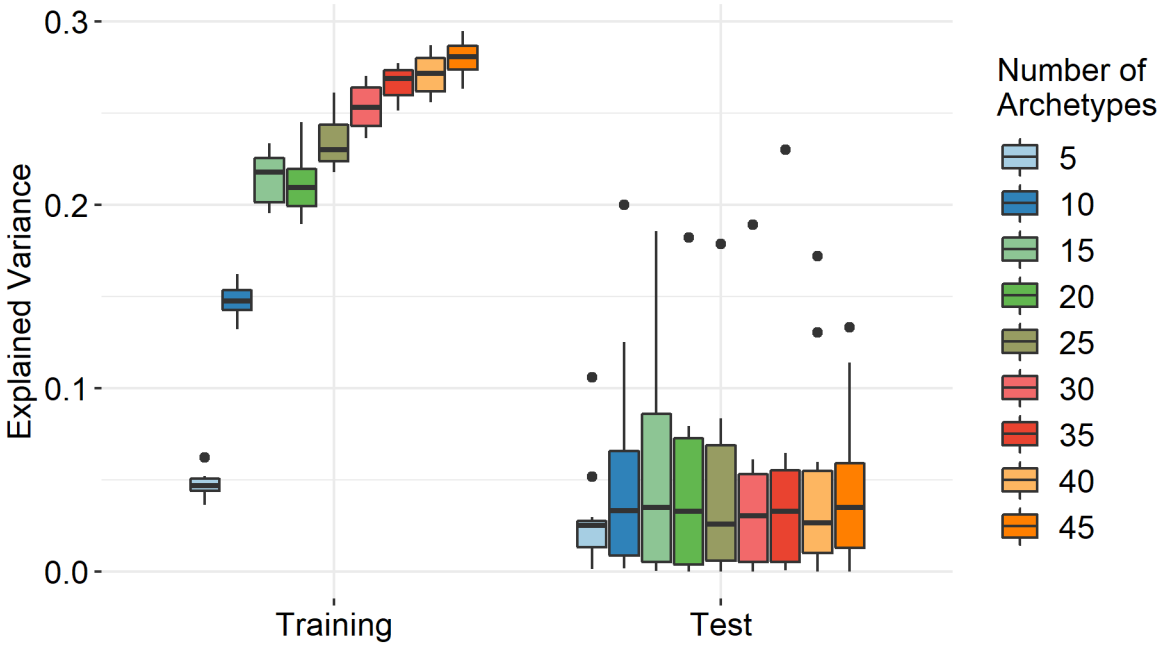


Figure 3.14: *Boxplot of Explained variance (D^2_{xy}) on training and test dataset, for 10 folds, for different number of archetypes (ranks of GLRM).*

With parameters already chosen, results for each model with different imputation methods can be compared. Considering R^2 first, Table 3.5 shows that for all three metabolites models imputation with GLRM and KNN provide higher metrics. However, overall, there is no significant difference among them. Also, all metabolomics models have higher R^2 than the basic clinical model. When contrasted with the extended clinical model, adding metabolomics leads to a slightly increase on the metric. Moreover, the archetypes model has a R^2 closer to the extended clinical model, which indicates that both models present similar discrimination of survival times. Thus, judging only by this metric, we would believe that the combination of many clinical variables and metabolomics is the one that explains most of the mortality risk. But, when considering the metric based on Sommer's D, that might not be the case.

In this case, GLRM presented the highest metrics in most cases, specially when metrics are corrected. Variation among imputation methods becomes more evident when median D^2_{xy} is obtained from test data, in which the GLRM is consistently among the highest ones. Comparing the original and corrected explained variance of the extended clinical model with metabolites and the archetypes model, we see that the high values of explained variance does not hold. That is, the models apparently fit well the training data, but are overfitting it and loosing part of explained variance when analyzing new data. The archetypes model, in particular, basically loses all explained variance. The metabolomics model, extended clinical and the basic clinical (with and without metabolites), present an increase of explained variance on test data, compared to the training. This happens due to variability between test data. Each fold has around 15 observations, and, in some cases, models predict very well and in others not so much, as can be seen in Figures 3.14 and A.4. This is the reason why the median of folds was chosen over the mean to summarize metrics.

Finally, by looking at the corrected D^2_{xy}, we have that the extended clinical model has the ability to better explain the data, specially when dealing with new data.

Table 3.5: *Comparison of R^2 and D^2_{xy} for each model.*

Model		R^2	D^2_{xy}			
			Complete	Training	Test	Corrected
Basic clinical model		0.094	0.068	0.067	0.072	0.073
Extended clinical model		0.253	0.206	0.208	0.214	0.212
Metabolomics only	**Mean**	0.137	0.096	0.096	0.096	0.095
	Median	0.138	0.101	0.100	0.107	0.108
	KNN	0.142	0.104	0.104	0.094	0.094
	GLRM	0.147	0.115	0.113	0.126	0.127
Metabolomics and basic clinical model	**Mean**	0.190	0.141	0.141	0.137	0.137
	Median	0.191	0.144	0.144	0.133	0.133
	KNN	0.194	0.146	0.146	0.136	0.136
	GLRM	0.198	0.150	0.152	0.140	0.138
Metabolomics and extended clinical model	**Mean**	0.259	0.195	0.196	0.138	0.138
	Median	0.262	0.195	0.197	0.160	0.158
	KNN	0.268	0.202	0.202	0.130	0.129
	GLRM	0.266	0.201	0.202	0.152	0.152
Archetypes		0.251	0.218	0.216	0.053	0.055

In summary, the nine metabolites previously selected do not improve much the predictive ability of overall mortality risk. The difference between metrics of basic clinical model, with and without metabolomics, can indicate an association between metabolomics and survival times; however, clinical variables can still explain more variation. As for imputation methods, metrics were similar between each of them, but when predicting new observations, imputing with GLRM avoids some overfitting of data. It is important to state that, just because the nine metabolomics did not increase the explained variance, does not mean that there is no association between metabolomics and overall mortality in a CKD population. More studies must be done in this context, such as analyzing other possibilities of combinations of metabolites, for example. Also, we can note that, even the highest metric still indicate a low percentage of variation explained, Henderson, Jones, and Stare, 2001; Henderson, 1995; Korn and Simon, 1990 and others discuss this difficulty of predicting data in survival models and assessing this predictive power.

In this specific example, the archetypes model did not show a good predictive ability when extrapolating to new data. However, its R^2 indicates an association between archetypes and survival rate. Thus, we will explore it more to better understand this relation. So, Table 3.6 shows the coefficients for the archetypes model on the entire dataset. Model adequacy and basic assumptions were tested and are considered valid (data not shown). Based on their p-values, archetypes 2, 7, 10, 11, 12, 13 and 15 were significantly associated, in which, archetype 2 slows the hazard ratio function, while the others (7, 10, 11, 12, 13, and 15) present positive coefficients. That is, individuals with higher values of scores from archetype 2, with other archetypes constant, have lower risk of mortality, whereas higher values on archetype 7 indicate lower survival, or higher risks.

Table 3.6: *Coefficient and significance of archetypes.*

	Coefficient	Standard error	p-value
Archetype 1	0.149	0.094	0.113
Archetype 2	−0.541	0.110	**<0.0001**
Archetype 3	−0.143	0.120	0.2346
Archetype 4	0.111	0.124	0.3691
Archetype 5	−0.017	0.136	0.8979
Archetype 6	−0.057	0.124	0.6454
Archetype 7	0.927	0.143	**<0.0001**
Archetype 8	−0.103	0.198	0.6025
Archetype 9	0.021	0.123	0.8678
Archetype 10	0.476	0.138	**0.0006**
Archetype 11	0.453	0.135	**0.0008**
Archetype 12	0.408	0.146	**0.0053**
Archetype 13	0.302	0.134	**0.0248**
Archetype 14	−0.051	0.131	0.6976
Archetype 15	0.481	0.135	**0.0004**

To understand and interpret each latent variable, Figure 3.15 presents the loadings of dataset variables for each one. For a better visualization, each variable is coded based on the type of data – anthropometric, laboratory, heart related (cardiac and vascular evaluation), clinical, gene related (with APOL gene or ancestrality scores calculated based on genetic information), medications or metabolites. The x and y axis represent the variables and their loadings, respectively. For each archetype, the 8 variables with higher positive and negative loadings were identified.

So, archetype 2 does not have many discrepant loadings, metabolites and variables, all contribute a little for the score. Still, it is the archetype in which metabolites have the highest loading in both directions, that is, some metabolites with positive and other with negative loadings. The metabolite X95 (2-ketoisocaproic acid 2) is among the variables with higher positive loadings for this archetype, alongside with sex (coded as 0 for female and 1 for males), eGFR and hemoglobin levels. In the Cox model this archetype presented a negative coefficient, which could be an indication that this variables contribute for the increase of survival function. Variables with the higher negative coefficients are two medications (calcitriol and CaCO3) and Urea and Creatinine levels. Archetype 10 also have metabolites with higher loadings than other archetypes, but this one has more discrepant clinical variables, such as, the indicator of previous myocardial infarction (MI), for example. Note that, on Table 3.6, archetypes 2 and 10 have coefficients with different signs, and the variables they have in common (sex and calcitriol) follow this behavior, with loadings with different signs as well.

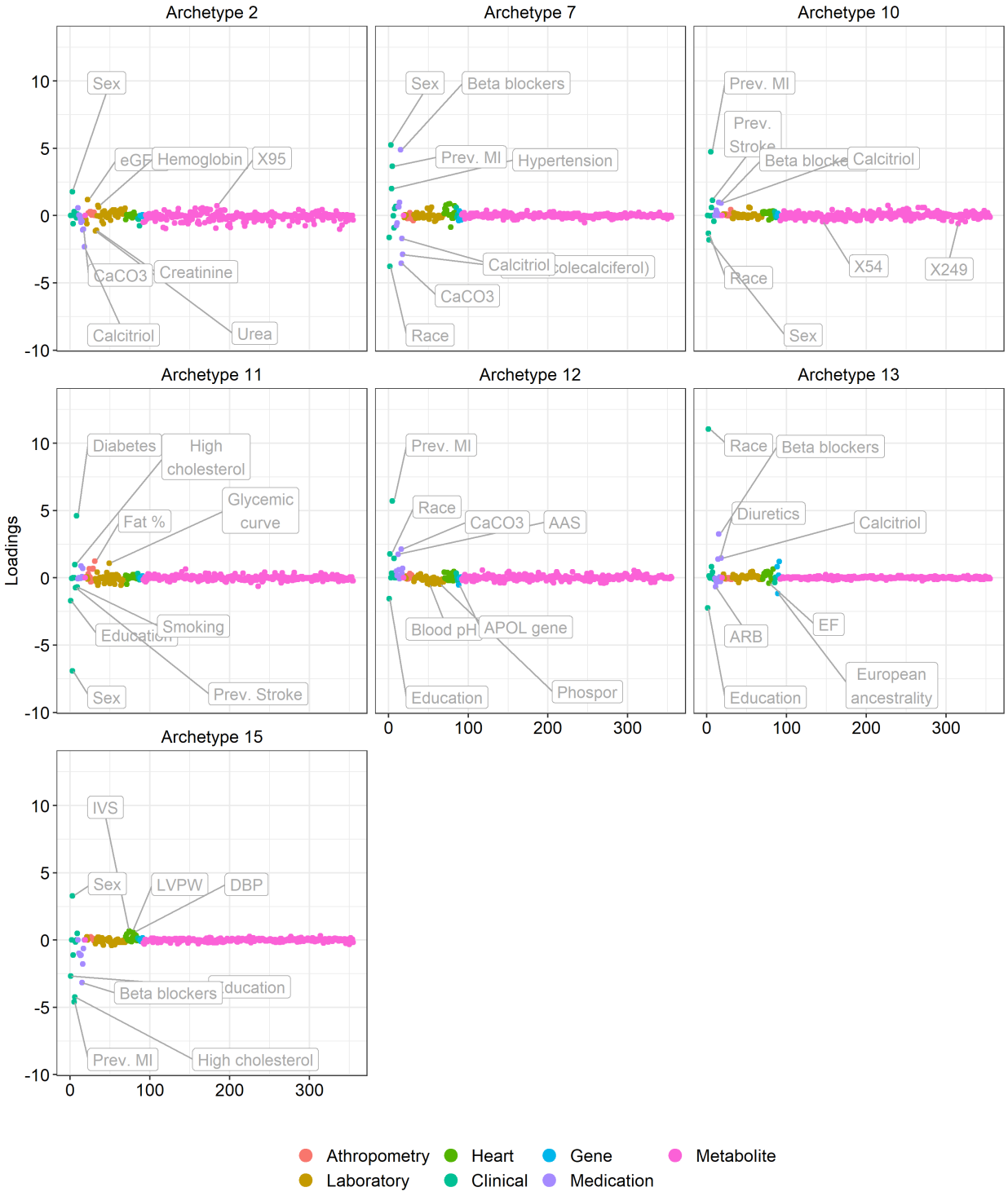


Figure 3.15: *Loadings for variables with significant archetypes. X-axis represent the variables, and y-axis, their loadings.*

In archetype 7, clinical and medication variables have the highest absolute loading values. Considering the variables identified, those who are males (sex = 1), have had previous myocardial infarction (MI), have hypertension's and are on beta blockers present higher scores for archetype 7. Moreover, those who are caucasians (race = 0) and do not take vitamin D, CaCO3 and calcitriol, have even higher scores, since the opposite of this categories would lead to a decrease on the

scores – due to negative loadings. So, an example of interpretation would be that an individual with these characteristics exemplified has higher risk of mortality, since archetype 7 has a hazard ratio of 2.53 (exponential of 0.9272). In addition, archetype 11 is the one with higher loadings for anthropometric and laboratorial variables, such as fat percentage and glycemic curve. Whereas archetype 13 has the highest loadings for gene and ancestrality scores, the 15th archetype presents heart related variables with only positive loadings, such as left ventricular posterior wall (LVPW), interventricular septum (IVS) and mean diastolic blood pressure (DBP), contrasting with clinical and medication variables, like previous MI, high cholesterol and beta blockers.

An interesting evaluation is the presence of educational level as one of the highest coefficient in several archetypes. Since it is a socioeconomic variable, it does not necessarily relate to overall mortality with CKD. It could be an indication of environmental effect, due to different living conditions or just a spurious association. Also, it was commented about the coefficients of archetypes 2 and 10 with opposite signs, and loadings for sex and calcitriol also with loadings with different signs. However, that is not the case for every archetype. For example, archetypes 10 and 15 have same signs of coefficients, but the variable sex has a positive loading on archetype 15. This also happens with other variables, such as CaCO3, beta blockers and race, for example. This means that the interpretation of each archetype is not straightforward and must be done carefully, considering the total combination of variables.

References

Armitage, Emily Grace et al. (2015). "Missing value imputation strategies for metabolomics data". In: *Electrophoresis* 36.24, pp. 3050–3060. ISSN: 15222683. DOI: 10.1002/elps.201500352.

Chen, Haiying et al. (2011). "A distribution-based multiple imputation method for handling bivariate pesticide data with values below the limit of detection". In: *Environmental Health Perspectives* 119.3, pp. 351–356.

Chung, Ren Hua and Chen Yu Kang (2019). "A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification". In: *GigaScience* 8.5, pp. 1–12. ISSN: 2047217X. DOI: 10.1093/gigascience/giz045.

Di Guida, Riccardo et al. (2016). "Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling". In: *Metabolomics* 12.5, pp. 1–14. ISSN: 15733890. DOI: 10.1007/s11306-016-1030-9.

Do, Kieu Trinh et al. (2018). "Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies". In: *Metabolomics* 14.10, pp. 1–18. ISSN: 15733890. DOI: 10.1007/s11306-018-1420-2. URL: http://dx.doi.org/10.1007/s11306-018-1420-2.

Domingos, Maria Alice Muniz et al. (2017). "Doença renal crônica-determinantes de progressão e risco cardiovascular. Coorte PROGREDIR: desenho de estudo e métodos". In: *Sao Paulo Medical Journal* 135.2, pp. 133–139. ISSN: 15163180. DOI: 10.1590/1516-3180.2016.0272261116.

Genovese, Giulio, David J Friedman, and Martin R Pollak (2013). "APOL1 variants and kidney disease in people of recent African ancestry". In: *Nature reviews Nephrology* 9.4, p. 240.

Gromski, Piotr et al. (2014). "Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data". In: *Metabolites* 4.2, pp. 433–452. ISSN: 2218-1989. DOI: 10.3390/metabo4020433.

Henderson, Robin (1995). "Problems and prediction in survival-data analysis". In: *Statistics in Medicine* 14.2, pp. 161–184. ISSN: 10970258. DOI: 10.1002/sim.4780140208.

Henderson, Robin, Margaret Jones, and Janez Stare (2001). "Accuracy of point predictions in survival analysis". In: *Statistics in Medicine* 20.20, pp. 3083–3096. ISSN: 02776715. DOI: 10.1002/sim.913.

Hill, Nathan R et al. (2016). "Global prevalence of chronic kidney disease–a systematic review and meta-analysis". In: *PloS one* 11.7, e0158765.

Holzinger, Emily R (2012). "Integrating heterogeneous high-throughput data for meta- Pharmacogenomics". In: 13.2, pp. 213–222. DOI: 10.2217/pgs.11.145.Integrating.

Holzinger, Emily R et al. (2014). "ATHENA: the analysis tool for heritable and environmental network associations". In: *Bioinformatics* 30.5, pp. 698–705.

Hu, Zixin et al. (2020). "Shared Causal Paths underlying Alzheimer's dementia and Type 2 Diabetes". In: *Scientific Reports* 10.1, pp. 1–15. ISSN: 20452322. DOI: 10.1038/s41598-020-60682-3.

Korn, Edward L. and Richard Simon (1990). "Measures of explained variation for survival data". In: *Statistics in Medicine* 9.5, pp. 487–503. ISSN: 10970258. DOI: 10.1002/sim.4780090503.

Luo, Shengyuan et al. (2019). "Serum metabolomic alterations associated with proteinuria in CKD". In: *Clinical Journal of the American Society of Nephrology* 14.3, pp. 342–353. ISSN: 1555905X. DOI: 10.2215/CJN.10010818.

Rhee, Eugene P. et al. (2019). "Variability of two metabolomic platforms in CKD". In: *Clinical Journal of the American Society of Nephrology* 14.1, pp. 40–48. ISSN: 1555905X. DOI: 10.2215/CJN.07070618.

Richardson, David B and Antonio Ciampi (2003). "Effects of exposure measurement error when an exposure variable is constrained by a lower limit". In: *American journal of epidemiology* 157.4, pp. 355–363.

Ritchie, Marylyn D. et al. (2015). "Methods of integrating data to uncover genotype-phenotype interactions". In: *Nature Reviews Genetics* 16.2, pp. 85–97. ISSN: 14710064. DOI: 10.1038/nrg3868.

Sekula, Peggy et al. (2016). "A metabolome-wide association study of kidney function and disease in the general population". In: *Journal of the American Society of Nephrology* 27.4, pp. 1175–1188. ISSN: 15333450. DOI: 10.1681/ASN.2014111099.

Titan, S. M. et al. (2019). "Metabolites related to eGFR: Evaluation of candidate molecules for GFR estimation using untargeted metabolomics". In: *Clinica Chimica Acta* 489, pp. 242–248. ISSN: 18733492. DOI: 10.1016/j.cca.2018.08.037.

Titan, Silvia M. et al. (2019). "Metabolomics biomarkers and the risk of overall mortality and ESRD in CKD: Results from the PRoGREDIR cohort". In: *PLoS ONE* 14.3, pp. 1–14. ISSN: 19326203. DOI: 10.1371/journal.pone.0213764. URL: http://dx.doi.org/10.1371/journal.pone.0213764.

Troyanskaya, Olga et al. (2001). "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6, pp. 520–525.

Tutz, Gerhard and Shahla Ramzan (2015). "Improved methods for the imputation of missing data by nearest neighbor methods". In: *Computational Statistics & Data Analysis* 90, pp. 84–99.

Van Buuren, Stef (2007). "Multiple imputation of discrete and continuous data by fully conditional specification". In: *Statistical methods in medical research* 16.3, pp. 219–242.

Walker, Esteban (2003). *Regression Modeling Strategies*. Vol. 45. 2, pp. 170–170. ISBN: 9783319194240. DOI: 10.1198/tech.2003.s158.

Xia, Jianguo et al. (2009). "MetaboAnalyst: a web server for metabolomic data analysis and interpretation". In: *Nucleic acids research* 37.suppl_2, W652–W660.

Yan, Kang K, Hongyu Zhao, and Herbert Pang (2017). "A comparison of graph-and kernel-based–omics data integration algorithms for classifying complex traits". In: *BMC bioinformatics* 18.1, p. 539.

Yu, Bing et al. (2014). "Serum metabolomic profiling and incident ckd among african americans". In: *Clinical Journal of the American Society of Nephrology* 9.8, pp. 1410–1417. ISSN: 1555905X. DOI: 10.2215/CJN.11971113.

Chapter 4

Discussion and conclusion

Dealing with several different datasets is a difficult task that has gained a lot of interest lately. Many different techniques of multi-staged and meta-dimensional integration analysis have been proposed, however they all have a common problem: dealing with heterogeneous data. Our proposition was to use Generalized Low Rank Models (GLRM) for dealing with multiple dataset in a multi-omics context. The use of different loss functions and regularization functions makes GLRM a very flexible model, capable of dealing with data of different kinds and missing data. After introducing the methodology of some multivariate analysis and GLRM on Chapter 2, the model was applied on a simulation dataset, as well as on a real dataset on Chapter 3.

The first application consists on verifying how well the GLRM performs in a supervised analysis context. As seen in Section 2.5, the GLRM finds the numerical matrices A and B that better approximate the abstract data matrix Y. In this case, the optimization is not made with the goal of better predicting a variable, as it is done by Discriminant Analysis. Also, an essential step of fitting a GLRM is the parameter choice, in which there is not a specific way of doing so. In fact, Udell et al., 2016 and Boehmke and Greenwell, 2019 state that the choice should be done differently, depending on the analysis objective. Therefore, even though it is not a supervised method, we can find the best model that predicts an outcome by choosing adequate parameters. On Section 3.1 we showed how this can be used. Our goal was to predict disease outcome, which is a boolean variable. Thus, the predictive metric Area Under the Curve (AUC) is appropriate to measure how well a classification is done. Then, a grid search was made and models with different parameters were selected based on this metric, with cross validation. As a result, we got a model that is competitive with other integration methods of different kind. So, in different contexts other metrics could be used, either to predict a quantitative variable, more than one variable, or to impute or denoise data. Two examples are in Section 3.2, where in each applications a grid search was performed and the rank was chosen based on metrics of interest: number of variables selected or explained variance.

The analysis of simulated data on Section 3.1 compared a meta-dimensional concatenation component-based method (GLRM) with other integration techniques, and it had some advantages upon them. The 3-stage approach, a multi-stage one, assumes that the disease etiology is linear (Figure 1.2). However, it is known that in this case the disease is complex, and a result of interac-

tion between several omics layers (Figure 3.1). So it may not be the best integration method. Next, CANetwork is a meta-dimensional graph-based method. Transformation-based methods, such as CANetwork, normally can preserve data-type-specific properties if an adequate transformation is done (Holzinger, 2012). However, since transformation is done independently on each dataset, it can be difficult to identify interaction between them. In the example, CANetwork obtained the lowest metrics when genetic effect was strong (Scenarios 1 and 3), but it was able to predict the outcome better than other methods when the genetic effect was weaker. Either way, lower metrics can be due to this difficulty in identifying relations among data. Then, two meta-dimensional concatenating analysis were adopted: random forest-based (RFomics) and model-based (ATHENA). As many other concatenating integration methods, they had the challenge of combining multiple datasets of different data types without driving bias. One solution used by RFomics was to calculate risk scores based on SNPs and normalizing the other omics data. However, it still considers data to be quantitative, which, in this example, is valid, but does not always happens. As for ATHENA, it was by far the best method to predict disease outcome with strong effect, but when the effect was weaker, it did not succeed as much. Finally, we have GLRM, which can solve the problem of combining data of different types while also being able to predict disease outcome well, even better than the others (in Scenario 2).

In the simulated data from Chung and Kang, 2019 all data (besides disease status outcome) can be seen as a quantitative value, therefore allowing many methods to be applied. However, when dealing with other variable types, such as clinical and demographic, for example, it becomes difficult to work with classical methods. In addition, even on this example, variables such as SNPs and CNV could be considered as categorical or ordinal, for example. In this behalf, GLRM is flexible enough to deal with variables of different forms. To exemplify this, a multinomial loss was used for SNP data, assuming they are categories of genotypes instead of number of allele copies. Also, other loss could have been selected for the variables. For example, methylation data is a proportion, between 0 and 1, so an adequate loss function could be implemented, such as a Beta loss, for example. Moreover, GLRM still offers other flexibility, which is to give different weights for each variable. By doing so, the model becomes more informative by indicating variables of interest. This can be useful when external information is available, for example in the hypothetical breast cancer analysis (Ritchie et al., 2015; Chung and Kang, 2019) where the relationship between gene XRCC3 and the outcome could have been known. Another utility for using different weights is shown in Figure 3.7. In a supervised analysis context, giving higher weight to the response variable (breast cancer outcome in this example) lead to slightly higher predictive metrics for most of rank options. In summary, GLRM flexibility allows for analysis of different data and offers different adjustments to better explain the data.

Results from Section 3.1 also showed a trade off between predictive ability and processing time. When analyzing 100 batches, GLRM took a mean time of 5 to 10 minutes, depending on the scenario. This is longer than 3-stage, RFomics and CANetwork, and a lot less than ATHENA, even though this comparison should not be directly made. In this small example, processing time was not a problem, but as the data gets bigger, the longer it takes for models to run. So, it is important

to save time as much as possible. In addition, when dealing with big data, it can become unfeasible to run an extend grid search, which we verified is important to find the best model. For that, it is important to understand the behavior of GLRM. In the simulated data example, we learned that the choice of the regularization parameter γ_b did not influence the AUC as much as when using the L1 regularization. Also, for the L1 the processing mean time increased as the rank got higher, but with L2 losses ranks from 15 to 45 showed the higher run times. Therefore, in future works all this information could be used to facilitate parameter choice, for example.

Another interesting aspect of GLRM is the data visualization. As shown in Figure 3.9, GLRM was able to identify groups of variables and variables of interest. In the simulated example, we know exactly the relation between omics data and the outcome of interest, such as how eQTL and SNPr influence gene and protein expression of a specific gene and the QTL influence methylation rates. Then, by looking at the numerical representation of variables, we observed that these variables had values greater than zero. That is, they have an influence on the model, while those with values of zero do not. Now, consider that the relation between omics data and the outcome is unknown. First, we found a model capable of predicting disease outcome, but there is also an interest on which variables are related to the outcome. Say we observe Figure 3.8a of Scenario 1. There are two points separated from the rest, which are the gene and protein expression of the same gene. This could be an indication of a gene of interest, and could lead to more studies relating this gene with the disease status. Also, several SNPs from a regulatory region were identified, among them the eQTL, meQTL and SNPr. Then, with more studies these SNPs could be also related with the outcome. These are just some examples on how data visualization could help generate hypothesis about a certain disease.

On the other hand, the GLRM still has some disadvantages as an integration technique. First, it is a model that requires parameter, loss and regularization functions to be specified prior to the optimization. The loss functions should be chosen based on data type, and parameter choice must be chosen with a grid search. However, in some cases, it is unfeasible to fit many different combinations of parameters and functions, which could reflect on the performance of the model. Second, as some other concatenation analysis, it can be biased toward certain omics. In the data visualization, for example, we see that SNPs from regulatory region and the COMT received the highest weights, whereas other important variables such as some SNPs from CYP1B1 or methylation did not showed much influence when they should. Also, the interaction among datasets is considered by concatenating them, but there is not an objective of maximizing the correlation between them, as it is done in Canonical Correlation Analysis or in the DIABLO framework (Rohart et al., 2017; Singh et al., 2018) (Figure 4.1. When relationship among omics layers is already known, it can be incorporated into the model to better explain the data. Even though GLRM is able to specify different weights for different variables, it is still unable to incorporate the relationship or correlation between variables. A possibility to generalize GLRM in this sense would be to fit GLRM individually into each dataset, and choose the parameters that maximizes a metric of correlation between dataset. An adequate metric should be defined, and the optimization problem should be modified to guarantee the achievement of the goal and a good performance of the model. This may

be the subject of future researches.

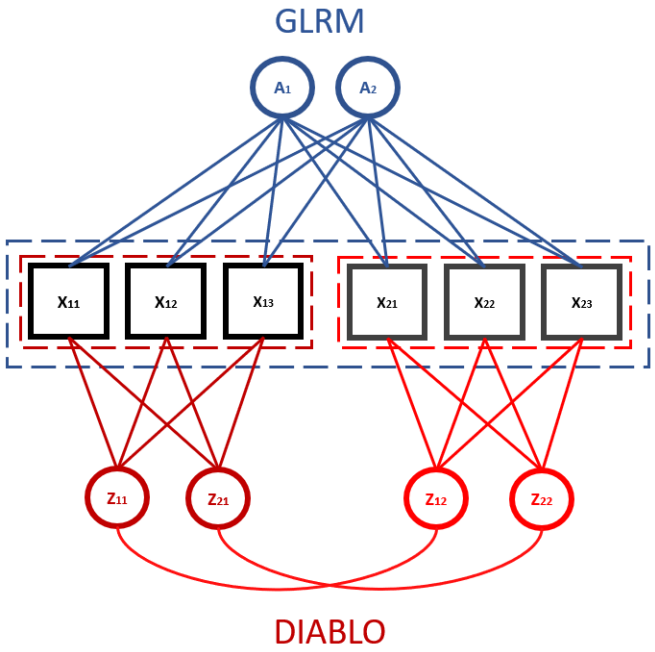


Figure 4.1: *Structural difference between GLRM and DIABLO.*

In addition, besides predicting an outcome, data integration can also be useful for imputing data or finding latent variables, which was the focus of the second application. It is motivated by a dataset from the PROGREDIR cohort (Domingos et al., 2017), which gathers data from two omics fields: metabolomics and phenomics. The phenomics data consists on phenotypes from different origins: measurements of different body parts, length, weights or their rations; quantification of molecules of blood or urine; characterization of heart sound waves; calculation of ancestrality scores based on genetic information; information on medications use or previous health conditions; socio-demographic characteristics and so on. Thus, there are many different data types available, each one with a different measurement scale. The goal was to find new metabolomic biomarkers to allow for a better understanding of Chronic Kidney Disease (CKD). However, as in many real data analysis there were many missing data.

Results from Section 3.2.2 shows that imputing data with GLRM, based on different phenotype variables, results on identifying more metabolites associated with the eGFR than with any other imputation methods compared. From the simulation study (Figures 3.11 and 3.12), the presence of missing data implicated on less metabolites being identified than with the complete dataset. When imputing with the GLRM, more metabolites than the correct ones were selected; however those which were supposed to be selected in most cases were. Thus, it might be the case that the 73 (out of 264) metabolites selected considered missing data unavailable are not the only ones truly associated with the eGFR. So, there is a possibility that the 7 new metabolites identified by the GLRM (Figure 3.13 and Table 3.4) are related to eGFR, and to CKD, consequently. In fact, one of them, metabolite X109, was also identified on both Brazilian replicate studies compared in Titan et al., 2019 (see Table 3.4). Another possibility is that the imputation by GLRM leads

to a spurious association between metabolites and eGFR. Nonetheless, this is an initial study to identify potential metabolites, further researches will be focused on verifying the real association between them. So, if there is a possibility that these could be biomarkers, then they should be considered.

In the context of survival analysis, imputation based on data integration stood out in the sense of avoiding overfitting of the model, which leads to better metrics when extrapolating to new data. A possible explanation is due to the fact that imputing data by the mean, median or KNN takes into consideration the information of only metabolomics, whereas GLRM considers the relationship among all variables. Another possibility is that the model is less overfitted due to the addition of regularization function. However, more studies should be conducted for a better understanding of this behavior. As stated before, assessing predictive ability of survival models is still a hard task, so in future work more sophisticated metrics could be used, such as the ones stated by Korn and Simon, 1990 and Henderson, 1995. There is still much to be explored of GLRM in the survival analysis context. In this study, in order to simplify, Logistic and Absolute loss functions were considered for the outcome and survival time, respectively, but there are other loss functions that might model survival time better. Furthermore, censoring was only indirectly considered by including the indicative variable of event, but a loss function for survival time that incorporates censoring information could be implemented.

Hence, imputing data with the GLRM showed potential improvements in both scenarios. Still it is not yet clear if the enhancement on metrics is due to the information gathered by data integration or due to another aspect. Forthcoming works can begin to investigate this by comparing results with imputation from other data integration methods, such as the ensemble based one from Subramanian et al., 2020. However, since it handles only numerical data, different transformation and normalization must be performed, specially on categorical data, which are not required with GLRM.

Still in the survival analysis context, another application of data integration was studied. Latent variables from variable combinations were obtained via GLRM. Results from Table 3.5 showed that the model with archetypes as predictors loses almost all explained variance when extrapolating to new data, but they can be useful to find association between scores and the overall risk of mortality. Visualization of variables weights on loadings of each archetype lead to some interesting interpretation. For example, the identification of cardiac and vascular evaluation variables on archetype 15, in which, from Table 3.6, higher values lead to increase risk of death. Even though interpretation of latent variable may be ambiguous and not straightforward, it can be useful for formulating new hypothesis of association between variables. Future research with collaboration of physicians experts in CKD would probably be beneficial to identify connections between each archetypes and the survival time, by looking into the literature.

In conclusion, this project discusses integration techniques and the problem of combining data of different types. A solution for this problem was proposed using resources from the GLRM, as well as a new application for it. The research started out focusing on studying the potential of GLRM in a supervised analysis, but, motivated by real data analysis problem, other purposes were

found for heterogeneous data integration with GLRM. Not only GLRM can be used in supervised integration analysis – shown by results from simulated data and compared with other known techniques – but can also integrate data to solve missing data problems and find latent variable to account for phenotype explained variation. Thus, being fast, flexible and easy to implement, GLRM has a lot of potential to solve biological problems, including multi-omics applications. Although it is an easy model to implement, there is still few or no information available about minor details useful for modelling data, such as specifying data domains, adding different weights to variable, implementing new loss functions and so on. Thus, we plan to prepare a GitHub documentation with step by step to fit a GLRM in Julia.

References

Boehmke, Brad and Brandon M Greenwell (2019). *Hands-On Machine Learning with R*. CRC Press.

Chung, Ren Hua and Chen Yu Kang (2019). "A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification". In: *GigaScience* 8.5, pp. 1–12. ISSN: 2047217X. DOI: 10.1093/gigascience/giz045.

Domingos, Maria Alice Muniz et al. (2017). "Doença renal crônica-determinantes de progressão e risco cardiovascular. Coorte PROGREDIR: desenho de estudo e métodos". In: *Sao Paulo Medical Journal* 135.2, pp. 133–139. ISSN: 15163180. DOI: 10.1590/1516-3180.2016.0272261116.

Henderson, Robin (1995). "Problems and prediction in survival-data analysis". In: *Statistics in Medicine* 14.2, pp. 161–184. ISSN: 10970258. DOI: 10.1002/sim.4780140208.

Holzinger, Emily R (2012). "Integrating heterogeneous high-throughput data for meta- Pharmacogenomics". In: 13.2, pp. 213–222. DOI: 10.2217/pgs.11.145.Integrating.

Korn, Edward L. and Richard Simon (1990). "Measures of explained variation for survival data". In: *Statistics in Medicine* 9.5, pp. 487–503. ISSN: 10970258. DOI: 10.1002/sim.4780090503.

Ritchie, Marylyn D. et al. (2015). "Methods of integrating data to uncover genotype-phenotype interactions". In: *Nature Reviews Genetics* 16.2, pp. 85–97. ISSN: 14710064. DOI: 10.1038/nrg3868.

Rohart, Florian et al. (2017). "mixOmics: An R package for 'omics feature selection and multiple data integration". In: *PLoS Computational Biology* 13.11, pp. 1–14. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005752LK. URL: http://www.embase.com/search/results?subaction=viewrecord%7B%5C&%7Dfrom=export%7B%5C&%7Did=L619520782%7B%5C%%7D0Ahttp://dx.doi.org/10.1371/journal.pcbi.1005752.

Singh, Amrit et al. (2018). "DIABLO: from multi-omics assays to biomarker discovery, an integrative approach". In: *bioRxiv* 61, p. 067611.

Subramanian, Indhupriya et al. (2020). "Multi-omics Data Integration, Interpretation, and Its Application". In: *Bioinformatics and Biology Insights* 14, pp. 7–9. ISSN: 11779322. DOI: 10.1177/1177932219899051.

Titan, Silvia M. et al. (2019). "Metabolomics biomarkers and the risk of overall mortality and ESRD in CKD: Results from the PRoGREDIR cohort". In: *PLoS ONE* 14.3, pp. 1–14. ISSN: 19326203. DOI: 10.1371/journal.pone.0213764. URL: http://dx.doi.org/10.1371/journal.pone.0213764.

Udell, Madeleine et al. (2016). "Generalized low rank models". In: *Foundations and Trends in Machine Learning* 9.1, pp. 1–118. ISSN: 19358245. DOI: 10.1561/2200000055. arXiv: 1410.0342.

Appendix A

Real data application

Table A.1: *ID and name of 264 metabolites*

ID	Metabolome
X1	(R*,R*)-2,3-Dihydroxybutanoic acid
X2	(R*,S*)-2,3-Dihydroxybutanoic acid
X3	(R*,S*)-3,4-Dihydroxybutanoic acid tritms
X4	(t-Butyldimethylsilyl)[3-methyl-3-(4-methyl-pent-3-enyl)-oxiran-2-yl]-methanone
X5	α-D-Glucopyranose
X6	α-Tocopherol
X7	β-D-Galactofuranose
X8	β-D-Galactofuranoside
X9	β-Gentiobiose
X10	β-Tocopherol
X11	γ-Tocopherol
X12	Phosphoric acid
X13	O-phosphocolamine
X14	Pyrophosphate
X15	Heptadecanoic acid
X16	Arachidic acid
X17	L-(+) lactic acid
X18	Succinic acid
X19	2-hydroxybutyric acid
X20	Uric acid 1
X21	Uric acid 2
X22	Galactitol
X23	Xanthine
X24	Tetracosane
X25	4-hydroxyphenylacetic acid

Table A.1 continued from previous page

ID	Metabolome
X26	Galactonic acid 2
X27	L-proline 2
X28	N-acetyl-D-glucosamine 2
X29	D-threitol
X30	Acetohydroxamic acid
X31	L-norleucine 1
X32	1,5-anhydro-D-sorbitol
X33	L-asparagine 2
X34	Beta- alanine 1
X35	1-stearoyl-rac-glycerol
X36	D-glucose 1
X37	D-glucose 2
X38	1-hexadecanol
X39	Tagatose 1
X40	p-Cresol
X41	Capric acid
X42	Cholesterol
X43	Citric acid
X44	L-glutamic acid 1
X45	L-glutamic acid 2
X46	L-glutamic acid 3
X47	Lauric acid
X48	Trans-aconitic acid
X49	Palmitoleic acid
X50	Galacturonic acid 1
X51	Eicosapentaenoic acid
X52	D-allose 1
X53	Serotonin 1
X54	Linoleic acid
X55	Stearic acid
X56	1 monoolein
X58	Threonic acid
X59	Trans-4-hydroxy-L-proline 1
X60	Creatinine
X61	L-alanine 1
X62	L-alanine 2
X63	L-serine 1
X64	L-serine 2

Table A.1 continued from previous page

ID	Metabolome
X65	Aspartic acid 1
X66	Aspartic acid 2
X67	L-lysine 1
X68	L-lysine 2
X69	Inosine
X70	Xylose 1
X71	Uridine
X72	Gluconic acid 2
X73	L-tyrosine 2
X74	Tyrosine 1
X75	Tyrosine 2
X76	L-leucine 2
X77	L-methionine 1
X78	L-methionine 2
X79	D-mannitol
X80	L-valine 2
X81	L-threonine 1
X82	L-threonine 2
X83	L-tryptophan 2
X84	Elaidic acid
X86	Quinic acid
X87	Norvaline 2
X88	Arabinose
X89	5alpha-Cholestanol
X90	L-cystine 1 [20.526]
X91	L-cystine 2
X92	L-cystine 3
X93	Guanosine 2
X94	Xylitol
X95	2-ketoisocaproic acid 2
X96	Ethanolamine
X97	L-glutamine 1
X98	L-glutamine 2
X99	Glycine
X100	Glycerol
X101	Glycerol 1-phosphate
X102	Glycolic acid
X103	Hypoxanthine

Table A.1 continued from previous page

ID	Metabolome
X104	DL-isoleucine 1
X105	DL-isoleucine 2
X106	3-indoleacetic acid
X107	1-octadecene
X108	Eicosane
X109	Ribitol
X110	Dioctyl phthalate
X111	Lactose 1
X112	Myo-inositol
X113	Psicose 1
X114	D-malic acid
X115	3-indolelactic acid 2
X117	Oxalic acid
X118	Citrulline 2
X119	Palmitic acid
X120	Ribose
X121	Phenylalanine 1
X131	Tetramethylbenzene
X132	1-(Trifluoroacetyl)-L-proline
X133	1,1'-Biphenyl, 4,4'-dinitro
X134	1,2,5-Thiadiazolo[3,4-c]coumarine, 8-nitro
X135	1,2-Benzenedicarboxylic acid
X136	1,2-Dipalmitin
X137	1,3-Dipalmitin
X138	10-Heptadecenoic acid
X139	11-Methyldodecanol
X140	13β-Ethyl-3-oximinogon-4-en-17-one 3-TMS
X141	13-Docosenoic acid
X142	17-Methoxy-d-homo-18-norandrosta-4,8,13,15,17-pentaen-3-one
X143	18-Methyl-nonadecane-1,2-diol
X144	1-Chloromethyl-1-hexadecyloxy-1-silacyclohexane
X145	1-Decanol, 2-hexyl
X146	1-Dodecanol, 2-hexyl
X147	1-Dodecanol
X148	1-Ethyl-1-tridecyloxy-1-silacyclohexane
X149	1-Heptanol
X150	1-Hexadecanol
X151	1H-Pyrazole, 4-nitro

Table A.1 continued from previous page

ID	Metabolome
X152	1-Hydroxy-3-methoxy-6-methylanthraquinone
X153	1-Monomyristin
X154	1-Monopalmitin
X155	1-Nonene, 4,6,8-trimethyl
X156	1-Octadecanol
X157	1'-Oxocannabinol
X158	2(1H)-Pyridinone, 3-acetyl-4-hydroxy-6-methyl
X159	2-(Methylamino)ethanol
X160	2,3-Dihydroxypropyl icosanoate
X161	2,4-Thiazolidinedione
X162	2,5-Di-tert-butyl-4-((trimethylsilyl)oxy)phenol
X163	2,6-Bis(tert-butyl)phenol
X164	2,6-Diisopropylnaphthalene
X166	2-α-Mannobiose
X167	2-Ethyl-3-hydroxypropionic acid
X168	2-Hydrazino-4,6-dimethylpyrimidine
X169	2-Hydroxybutyric acid
X171	2-Methoxy-6-methyl-4-phenyl-quinazoline
X172	2-Methoxyestradiol
X173	2-Monomyristin
X174	2-Monostearin
X175	2-O-Glycerol-α-d-galactopyranoside
X176	2-Palmitoylglycerol
X177	2-Phenyl-1,3-oxazol-2-ine
X178	2-Propenoic acid
X179	2-Pyrrolidinone
X180	3-α-Mannobiose, octakis(trimethylsilyl) ether (isomer 1)
X181	3-α-Mannobiose, octakis(trimethylsilyl) ether (isomer 2)
X183	3-Bromo-1-propanol
X184	3-Pyridinol
X185	4-Bromo-1-butanol
X186	4-Pyridinol
X187	5,5'-Biphthalide
X188	9-Decenoic acid
X189	9-Hexadecenoic acid
X190	9-Tetradecenoic acid
X191	Acetamide
X192	Acetyl chloride

Table A.1 continued from previous page

ID	Metabolome
X193	Aminomalonic aci
X194	Anthraquinone
X195	Arachidonic acid
X196	Benzoic acid, ethyl ester
X197	Benzoic acid, trimethylsilyl ester
X198	Benzoic Acid
X199	Butanoic acid
X200	Butylphosphonic acid
X201	Campesterol
X202	Carbonic acid, eicosyl vinyl ester
X203	Carbonic acid, octadecyl vinyl ester
X204	Cholest-7-en-3-ol
X212	D-(+)-Turanose
X213	D-Arabino-Hexonic acid
X214	Decane
X215	Dibenzo-1,4,8,11-tetraazacyclotetradecine, 5,6,7,8,9,14,15,16,17,18-decahydro
X216	Diethanolamine
X217	Dihydrouracil
X220	DL-Ornithine
X221	Doconexent
X222	Dodecane
X223	Dodecane
X224	Ethanol
X225	Ethanolamine
X226	Ethylα-D-glucopyranoside
X227	Galactose oxime
X228	Gluconic acid
X229	Glucose
X230	Glyceric acid
X231	Glycine, 2TMS derivative
X232	Glycine, TMS derivative
X233	Glycolic acid
X234	Heneicosane
X235	Heptacosane
X236	Heptadecanoic acid
X238	Hexadecane
X239	Hexadecane, 2-methyl
X240	Indol-5-ol

Table A.1 continued from previous page

ID	Metabolome
X241	Lactic Acid
X242	L-Alanine
X243	L-Alanine
X244	L-Asparagine
X245	Leu-Trp, N-trimethylsilyl
X246	L-Phenylalanine
X247	L-Proline, 2TMS derivative
X248	L-Proline, TMS derivative
X249	L-Serine, 2TMS derivative
X250	L-Valine
X251	m-Cresol
X252	Methoxyamine
X253	Methyl 2-[2-(4-chlorophenyl)-5-methyl-1H-imidazol-1-yl]dithiobenzoate
X254	Methyl galactoside
X255	Monolaurin
X256	N-(6-Quinolinyl)phthalimide
X257	N-(Hydroxymethyl)trifluoroacetamide
X258	N-Acetyl glucosamine methoxime
X259	Nonanoic acid
X260	Octadecane
X261	Oxalic acid
X262	Pentadecanoic acid
X263	Pentadecanoic acid
X265	Phenol, 2,4-bis(1,1-dimethylethyl)
X266	Phenol
X267	Phosphoric acid, 2-(trimethylsiloxy)-1-[(trimethylsiloxy)methyl]ethyl bis(trimethylsilyl) ester
X268	Phosphoric acid, 2-[(trimethylsilyl)oxy]-1,3-propanediyl tetrakis(trimethylsilyl) ester
X269	Phosphoric acid, bis(trimethylsilyl)monomethyl ester
X271	Propanoic acid
X272	Propylene glycol
X273	Propylparaben
X274	Pseudo uridine penta-tms
X275	p-Tolyl-β-D-glucuronide
X276	Pyroglutamic acid
X277	Ribonic acid
X278	Sedoheptulose
X279	Silane, diethyldecyloxyhexadecyloxy
X280	Silane, dimethyl(octadecyloxy)propyl

Table A.1 continued from previous page

ID	Metabolome
X281	Silane, dimethyl-2-propenyl(tetradecyloxy)
X282	tert-Butyl(2-(tert-butyl)-4-methoxyphenoxy)dimethylsilane
X283	Tetradecane
X284	Tetradecanoic acid
X286	Thiazolidine-2,5-dione
X287	Tricosanoic acid
X289	Undecane, 2,10-dimethyl
X290	Undecane, 3,8-dimethyl
X291	Undecane, 4,7-dimethyl
X292	Urea
X293	Uridine

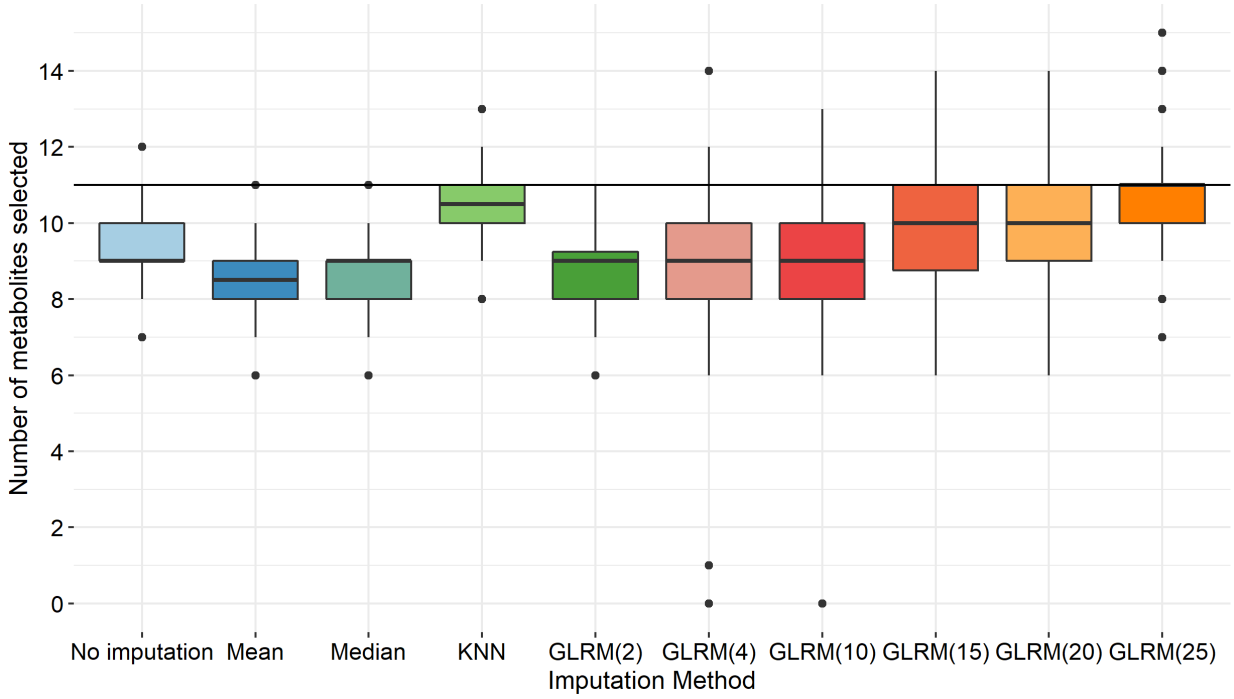


Figure A.1: *Boxplot of the number of metabolites selected in association with eGFR for each type of imputation with less than 20% missing. Number of selected metabolites (11) on complete dataset is represented by the horizontal line.*

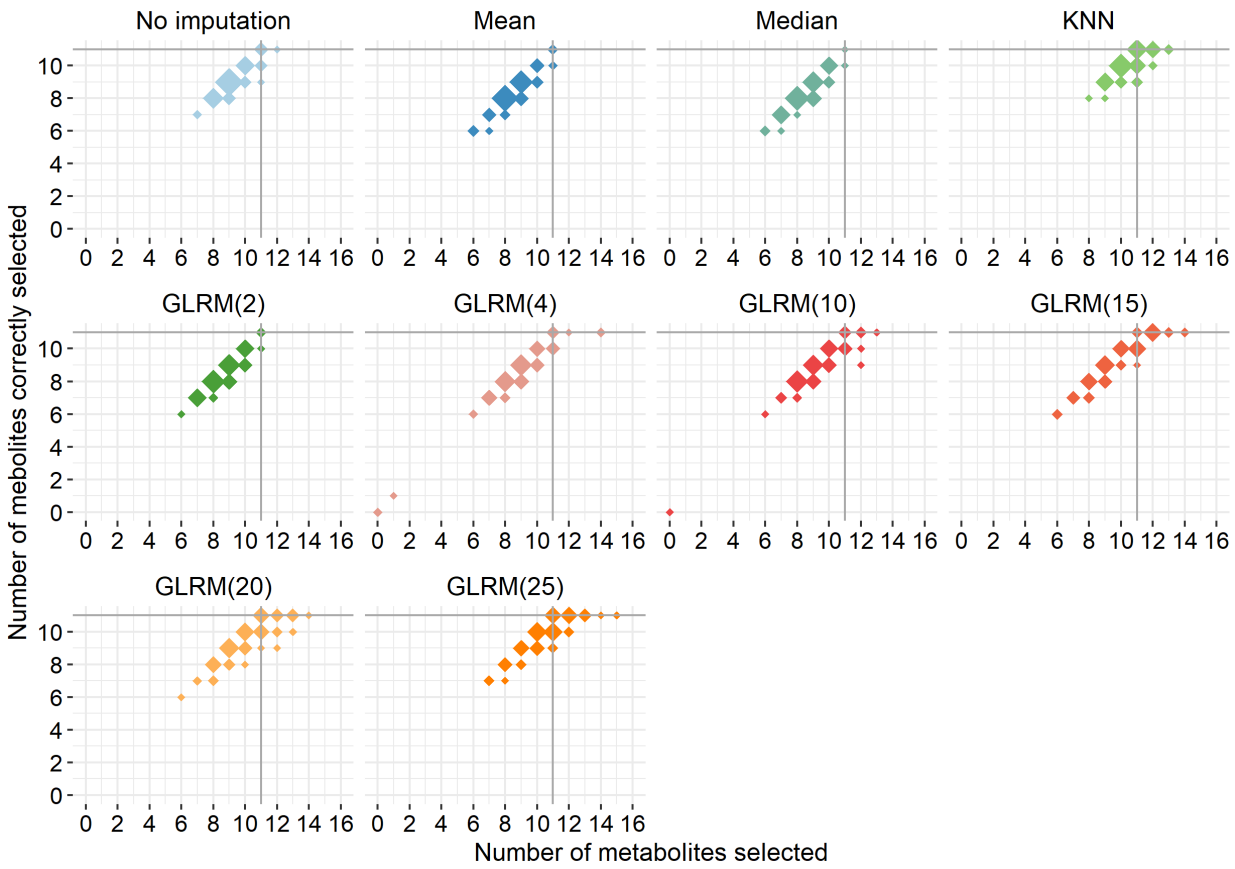


Figure A.2: *Comparison between the number of correct metabolites selected and the total number for each imputation method with less than 20% missing. Results are for 100 repetitions, and point sizes are proportional to the number of occurrences.*

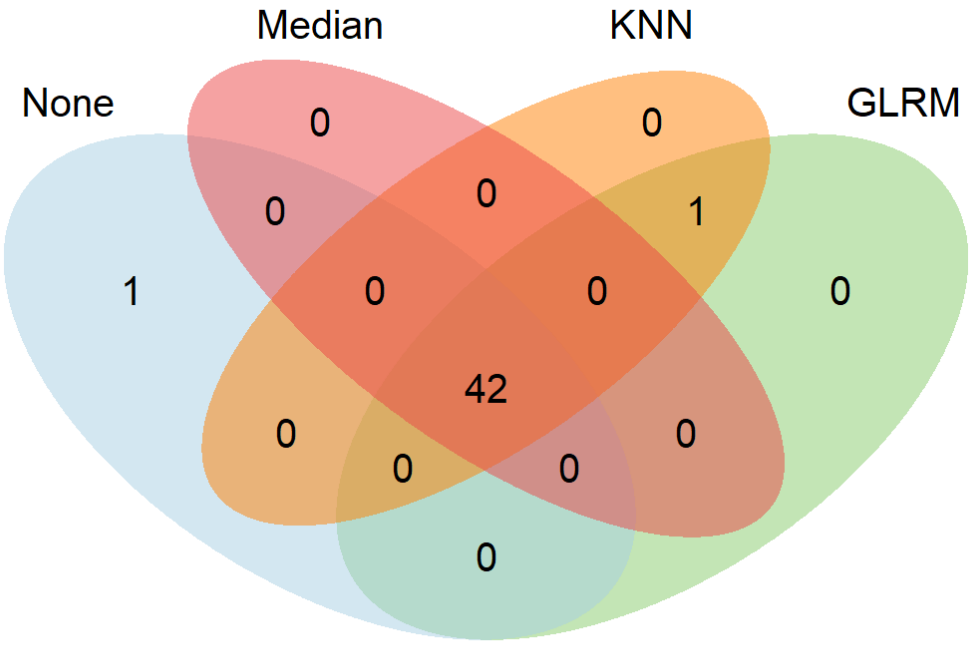


Figure A.3: *Venn diagram comparing selected metabolites with each imputation type (less than 20% missing).*

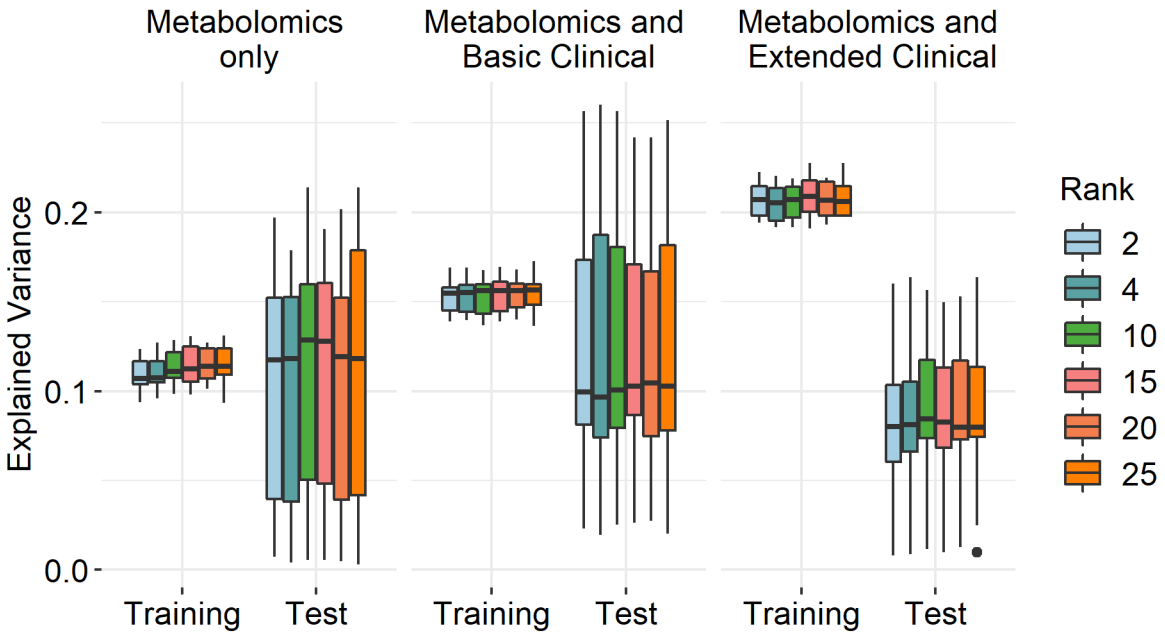


Figure A.4: *Boxplot of Explained variance (D^2_{xy}) of GLRM on training and test dataset, for 10 folds, for different ranks (including offset).*