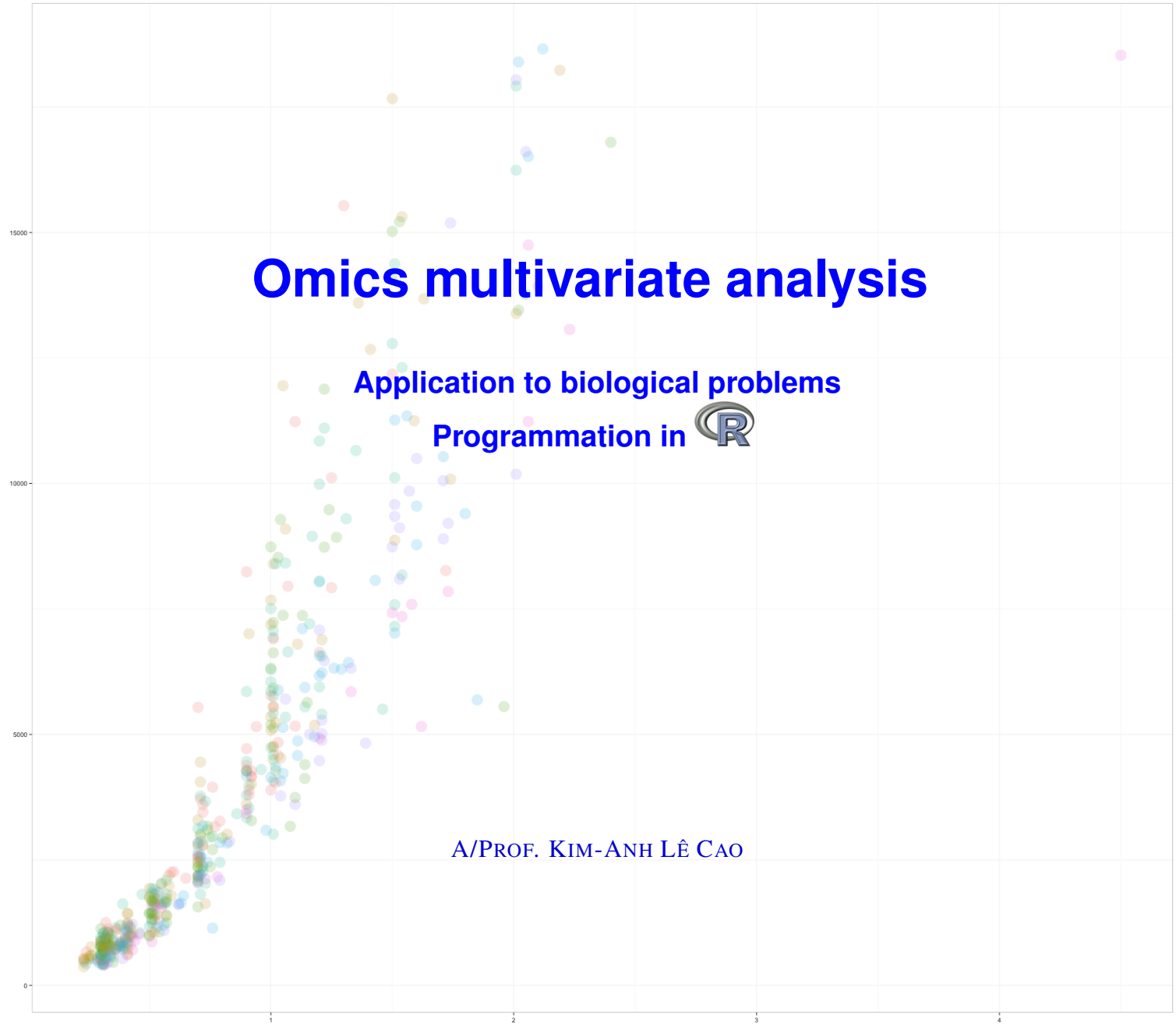
ELIXIR OMICS INTEGRATION AND SYSTEMS BIOLOGY

YEAR 2021



Omics multivariate analysis

Application to biological problems

Programmation in R

A/PROF. KIM-ANH LÊ CAO

Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne. Email: kimanh.lecao@unimelb.edu.au, Website: http://lecao-lab.science.unimelb.edu.au

Resources

Foreword

The course is an excerpt from the material developed by A/prof Kim-Anh Lê Cao as part of her 3-day mixOmics workshop. Formal permission must be sought to redistribute, or reuse part of any material provided during this workshop. Acknowledgements will also be required if any of this work is quoted or cited.

Learning objectives

Theory

- Understand the main concepts of multivariate dimension reduction methods
- Choose the 'right' method for the 'right' biological question
- Be aware of the benefits and limitations of all methods presented
- Interpretation of the visualisation outputs

Practice

- Ability to use provided `R` code on own data, or data for assignment
- Perform several types of multivariate analyses ranging from data exploration to biomarker selection using `mixOmics`
- Be critical of the results obtained

Your intructor

A/Prof Kim-Anh Lê Cao

A/Prof Kim-Anh Lê Cao (University of Melbourne, Australia) was awarded her PhD in 2008 at Université de Toulouse, France. She then moved to Australia as a post-doctoral fellow at the University of Queensland, Brisbane. Since the beginning of her PhD Kim-Anh has initiated a wide range of valuable collaborative and research opportunities in both statistics and molecular biology. Her main research focus is on variable selection for biological data ('omics' data) coming from different functional levels by the means of multivariate dimension reduction approaches. Since 2009, her team has been working on developing a statistical software dedicated to the integrative

analysis of 'omics' data, to help researchers make sense of biological big data. Kim-Anh is a senior lecturer at the University of Melbourne (Melbourne Integrative Genomics, School of Mathematics and Statistics), and regularly runs statistical training workshops and short series seminars as well as mixOmics multi-day workshops.
Kim-Anh is specifically interested in developing novel computational and statistical approaches for the analysis of microbiome and single cell data. More details on Kim-Anh current research themes and collaborative projects: `http://lecao-lab.science.unimelb.edu.au/`.
During her spare time, Kim-Anh is a keen rock climber and bushwalker. Her favourite or common travel destinations are France and South Africa.

Some background reading

- Hervé, M.R., Nicolè, F. and Lê Cao, KA. Multivariate Analysis of Multiple Datasets: a Practical Guide for Chemical Ecology J Chem Ecol (2018) 44: 215.
  A gentle and applied introduction to multivariate methods.

- Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics. 2017;8:84.
  An overview of how the 'omics research field is moving towards data integration.

- Rohart F, Gautier B, Singh A, Lê Cao K-A (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13(11): e1005752.
  Presentation of the `mixOmics` R package we will be using during the prac and some different integrative frameworks (not covered in this course).

Contents

Chapter 1

Exploring and classifying large scale data using multivariate methods

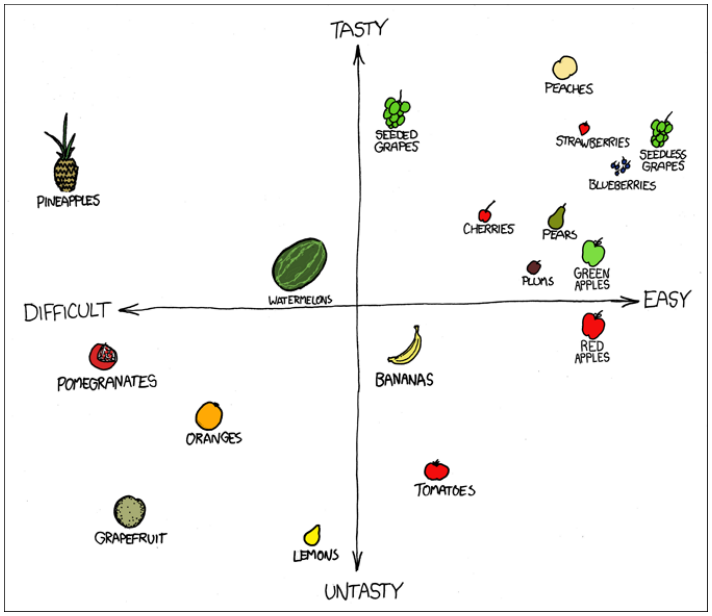


Figure 1.1: Principal Component Analysis is a multivariate method to summarise information from large data sets. This Chapter will introduce some underlying concepts in Principal Component Analysis and other types of multivariate methods to get more insight from large (or larg-ish) omics biological data.

1.1 What are 'omics data?

'Omics refer to any type of biological study that ends in -omics. Those include transcriptomics – the study of transcripts, proteomics – the study of proteins, metabolomics – the study of metabolites. Sometimes we refer to the objects of study, the 'omes (transcriptome, proteome, metabolome, *microbiome*). The aim of omics studies is to characterise and quantify biological molecules at each 'omics level.

Studying the 'omes require a *holistic approach*, as opposed to a reductionist approach, in order to better understand a biological system. With the advent of high-throughput sequencing technologies, these large data set need to be bioinformatically processed, then statistically analysed using novel techniques that often shift the classical analysis paradigm.

1.2 Illustrative data set: metabolome analysis of yeast

The illustrative example that will be used in the remaining of this chapter is a data set from a yeast study (Villas-Boâs *et al.*, 2005). In this data set, two Saccharomyces cerevisiae strains were used: a reference strain (wild-type: WT) and a mutant (MT) were carried out in batch cultures under two different environmental conditions in standard mineral media with glucose as the sole carbon source. The authors assayed metabolite levels in the two yeast strains (WT and MT) and two different environmental conditions, aerobic and anaerobic perturbations (AER and ANA). After normalisation and pre processing, the metabolome data results in 37 metabolites and 55 samples which include 13 MT-AER, 14 MT-ANA, 15 WT-AER and 13 WT-ANA samples.

Biological question. One of the main questions when analyzing high throughput data is whether the information provided by the metabolites spectra relate to the experimental conditions, or rather, to some interfering signals. In this chapter, we are focusing on different techniques to vizualize datasets in a 'blind' (unsupervised) way, i.e. when the biological background information, such as group affiliation or class label is not used in the statistical approaches. The aim is to represent the major or global information from the data sets without experimental knowledge.

1.3 Principal Component Analysis

A well-established technique for *visualisation* and *extraction of relevant information* is the popular Principal Component Analysis (PCA).

1.3.1 Principle

The aim of PCA (Jolliffe, 2005) is to reduce the dimensionality of the data while retaining as much information as possible. 'Information' is referred here as *variance*. The idea is to create uncorrelated artificial variables called principal components (PCs) that combine in a linear manner the existing and possibly correlated variables (here the genes, or the metabolites). The dimension is reduced by projecting the data into the smaller subspace spanned by the PCs, while capturing the largest sources of variation between the samples.
The principal components are obtained by maximising the variance-covariance matrix of the data, finding eigenvalue of the variance-covariance matrix or using singular value decomposition when the number of variables is very large. The data are usually centered, and sometimes scaled. Missing

values are not allowed, unless using the NIPALS (nonlinear iterative partial least squares Wold *et al.* (1987) algorithm which also enables an estimation of the missing values.
The first PC is defined as the linear combination of the original variables that explains the greatest amount of variation. The second PC is then defined as the linear combination of the original variables that accounts for the greatest amount of the remaining variation subject of being orthogonal (uncorrelated) to the first component. Subsequent components are defined likewise for the other PCA dimensions. The user must therefore keep in mind how much information is explained by the first PCs as these are used to graphically represent the PCA outputs.

Remark 1. We explain the difference and communality between a variance-covariance matrix and a correlation matrix in Appendix A.1.

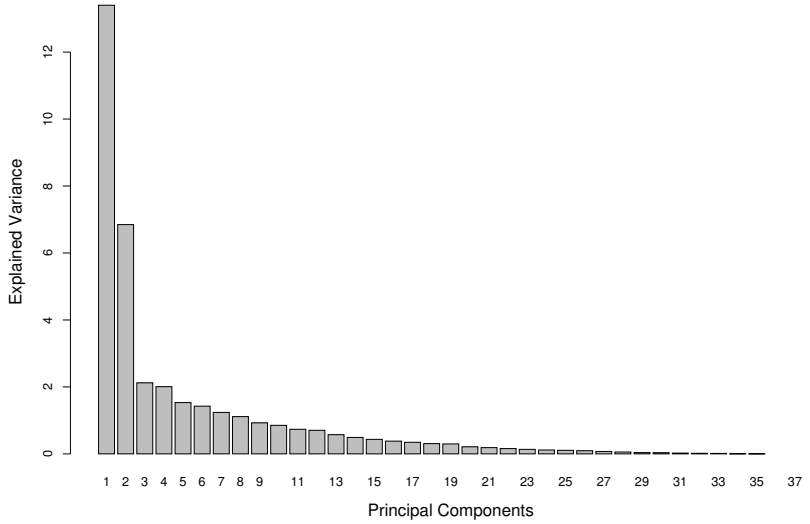


Figure 1.2: Principal Component Analysis of the yeast data set: barplot of the explained variance on each PC. This output is useful to choose the number of PCs to retain in the PCA analysis

1.3.2 Choosing the PCA dimension

We can obtain as many dimensions (i.e. number of PCs) as the number of variables. However, the goal is to reduce the complexity of the data and therefore summarize the data in fewer underlying dimension.
Fig. 1.2 displays the barplot of the eigenvalues associated with each PC . One criterion to select the number of PCs to retain in the analysis is to find the spot where the smooth decrease of the eigenvalues appears to level off to the right of the plot (i.e. when the 'elbow' appears). These eigenvalues correspond to the amount of variance explained by the components. Another criterion is the clarity of the final configuration (see next Section 1.3.3). All of this is highly subjective and the reader must keep in mind that visualisation becomes difficult above 3 dimensions.
Fig. 1.2 suggests that two PCs might be satisfactory to visualise most information from the yeast data. A usual output that can be obtained using statistical softwares is the cumulative percentage of explained variance (or information): in the yeast data, two PCs explained 54.72% of the total variance, and three PCs explained 60.45% of the total variance.

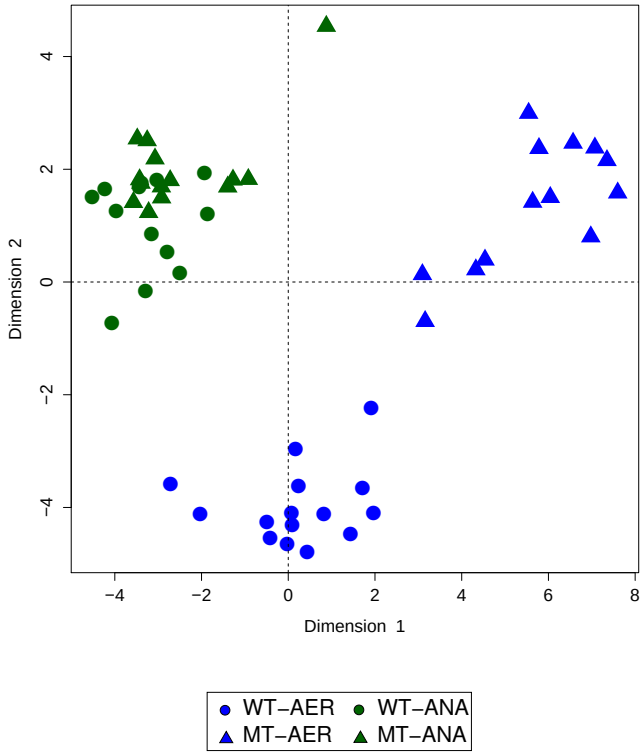


Figure 1.3: Principal Component Analysis of the yeast study and representation of the samples on the first two principal components (denoted 'Dimension 1' and 'Dimension 2'). Each dot represents a sample. Anaerobic and aerobic conditions are separated on the first PC.

1.3.3 Graphical outputs

PCA is an extremely valuable visualisation tool to explore a dataset. It can reveal the discriminatory structure, as well as experimental bias in the data. Two types of graphical outputs can be obtained in PCA:

- *Sample representation* can be obtained by plotting the principal components to observe the similarities between the samples which account for most variation, but also to give a 'meaning' to the PCs. For example, in Figure 1.3, the first PC tends to discriminate anaerobic vs. aerobic conditions, whereas the second PC tends to discriminate the wild type aerobic vs. the other conditions. Remark that as noted above, only 2 PCs might be satisfactory enough to summarize most of the information from these data.

- *Variable representation* with correlation circle plots. They provide an insightful way to visualise the contribution of each variable to the definition of each component, as well as the correlation between variables (see also González *et al.* (2012)).

- A biplot allows to graphically display *both samples and variables*. Samples are displayed as dots while variables are usually displayed as vectors. If the data are centered and scaled, the cosine angle between the variable vector and the PC indicates the correlation coefficient

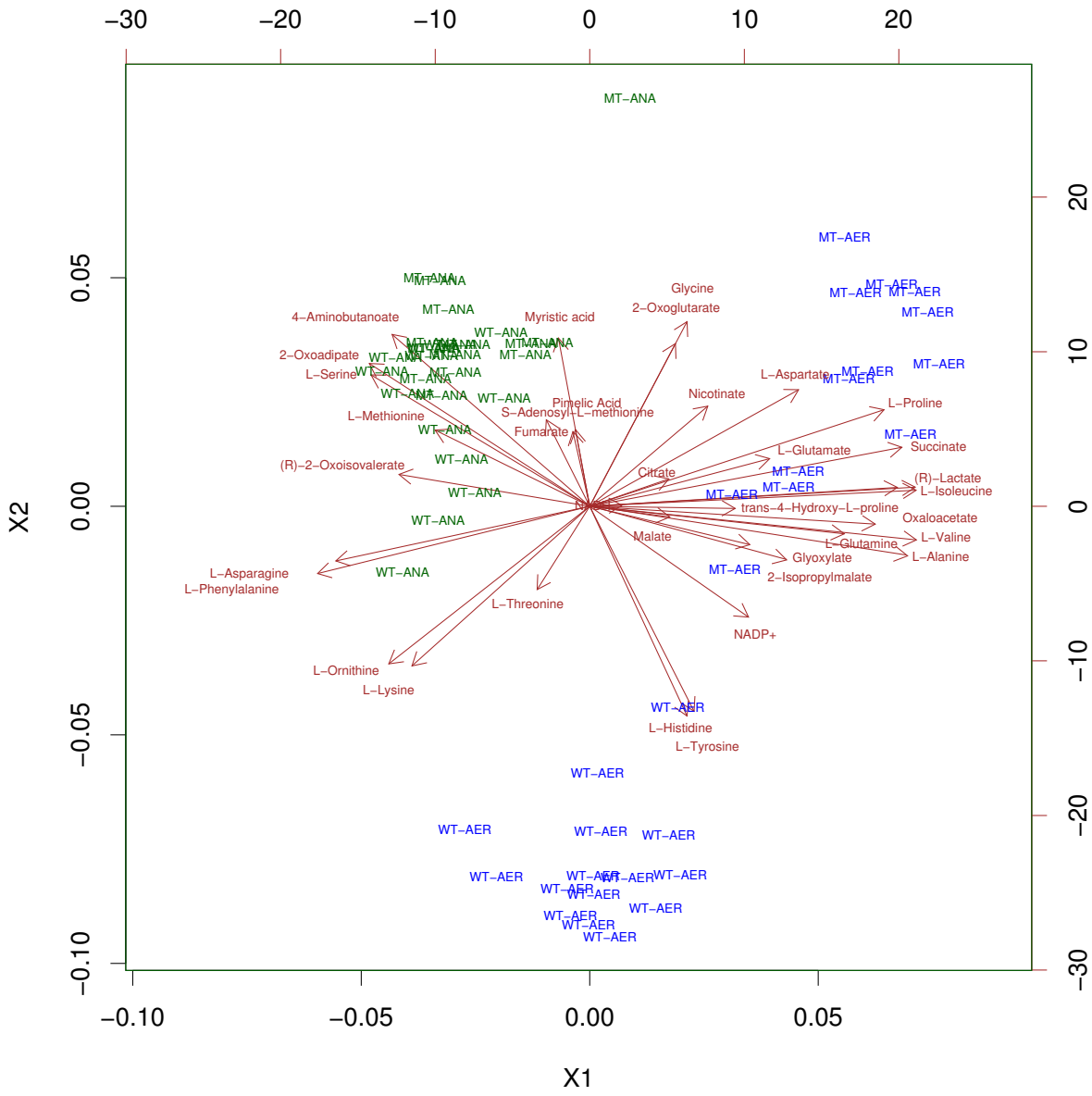


Figure 1.4: Biplot from the PCA analysis on the yeast data, simultaneous representation of the samples (dots) and variables (vectors) on the first 2 PCs (x-axis: PC1 and y-axis: PC2). Clusters of metabolites correlated with the biological conditions can be identified.

between the variable and the PC. This is therefore a useful way to give a meaning to each PC. For example, Fig. 1.4 allows to identify which metabolites are highly correlated (negatively or positively) to the first and second PCs. The metabolites represented with long arrows and highly correlated with the principal components are the ones that explain most of the variation between the different conditions. For example the group of metabolites pointing towards MT-AER are highly expressed in this group, but under expressed in the anaerobic group.

Remark 2. Note that using a smaller number of preselected metabolites (i.e. removing noise) may improve the quality of such graphical outputs.

1.4 Multidimensional Scaling (MDS, not covered)

MDS is a more general approach to dimension reduction (Kruskal and Wish, 1978; Borg and Groenen, 2005). Similar to PCA, it allows for the visualisation of multivariate data. MDS projects the data onto a lower dimensional space using a proximity measure, i.e. a distance or a similarity measure between the samples. Similarly to PCA, the aim is to find a lower-dimensional representation of the data, while preserving the pairwise distances as well as possible. MDS is much more flexible than PCA as it allows different types of distance matrices: Euclidean distance or mutual information are mostly used, while the covariance or correlation distance will proceed similar to PCA.
Non linear projections can also be used, in particular with high signal-to-noise ratio data, and are usually more efficient. In that case, a nonmetric, monotone transformation of the observed input data (the distances) is applied in an attempt to reproduce the general rank-ordering of distances between the objects in the analysis.
The criteria to specify the number of dimensions to retain are similar to PCA. MDS methods are applicable to a wide variety of data as the distance measures can be obtained in any number of ways. It does not impose any restriction on the underlying data distribution, and as long as the rank-ordering of distances (or similarities) in the matrix is meaningful, MDS can be used.

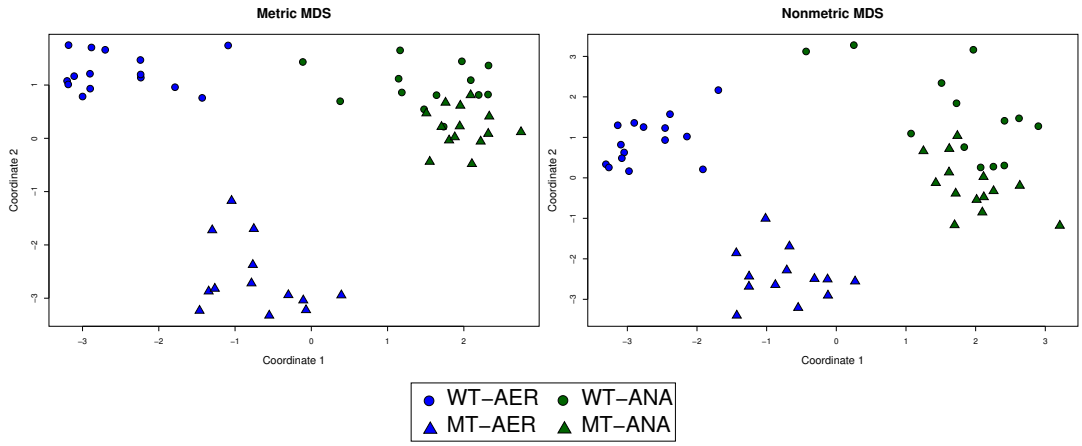


Figure 1.5: Multidimensional scaling of the yeast data set and representation of the samples on the first two MDS coordinates for a linear and non linear MDS. Each dot represents a sample.

Figures 1.5 display the classical MDS (metric, Euclidean distance) and non linear MDS. The nonmetric MDS seems to cluster the samples better than the classical MDS. These graphical outputs are similar to those obtained with PCA.

Remark 3. The axes are inversed to those obtained with PCA. The orientation of the axes is arbitrary and does not affect the interpretation of the outputs.

1.5 Supervised analysis

The term 'supervised learning' comes from the Machine Learning field. The aim is to find a relationship between the data (the input) and an output. This output can be discrete, indicating for example the type of treatment undergone from each sample (e.g. k treatments or k classes), in that case we are in a classification framework. The outcome can be continuous for a regression framework. This type of analysis is different from what was presented above (unsupervised analysis) where no explicit

output was given. In this section we will mainly focus on classification problems. Two aspects are important:

- **descriptive** aim: which are the variables that best separate the k classes,
- **predictive** aim: predicting the class of a new samples given its variables values

1.5.1 Motivating example

We saw in the previous sections that all the exploratory approaches were able to clearly separate the anaerobic condition from the wild-type and mutant aerobic conditions. However, the two genotypes in the anaerobic condition seem difficult to separate. By labelling the 4 classes, the supervised classification will 'push' the analysis towards the identification of the features (the metabolites) that will help discriminating each class.

1.5.2 Cross-validation to assess prediction performance

Cross-validation (CV) is a model validation technique used in statistical and machine learning to assess whether the results of an analysis can be generalised to an independent data set. It consists in dividing the data set into M subsets (or folds), fitting the model on $M - 1$ subsets and evaluating the prediction performance on the left-out subset. This process is iterated until each subset is left out once; the prediction performance are then averaged. In the `mixOmics` methods presented here (PLS-DA, sPLS-DA), prediction performance refers to either an overall misclassification error rate or a balanced error rate calculated on the left-out samples.

We define *stratified CV* when there is approximately the same proportion of each class in each of the folds. Repeated cross-validation implies that the whole CV process is repeated a number of times `nrepeat` to reduce variability across the different subset partitions. In the case of Leave-One-Out CV (`validation = 'loo'`), each sample is left out once ($s = N$) and therefore `nrepeat` is set to 1.

1.5.3 Linear methods for classification

We now focus on a classification problem using linear methods. Discriminant analysis can be seen as an extension of a multiple regression problem with a discrete outcome. In particular, we might be interested in the following aims:

- seek for the linear combinations of variables that best separate the classes and obtain a graphical representation that best illustrate the separation,
- improve the prediction accuracy with a simpler model that removes noisy variables: parsimony is especially important when the number of variables is large.

1.5.4 LDA

Linear Discriminant Analysis (Fisher, 1936) makes the assumption that the groups are linearly separable, i.e. the groups can be separated by a linear combination of features that describe the objects. Similar to PCA, LDA is a dimension reduction technique that looks for linear combinations of variables which best explain the data. However, unlike PCA that does not take into account any difference in the classes, LDA maximizes the ratio of between-class variance to the within-class variance to guaranty maximal separability.

1.5.5 PLS-DA

Despite very good classification results, LDA has numerical limitations. In particular, for large data sets with too many correlated predictors, LDA uses too many parameters that are estimated with a high variance. Another option is to use Partial Least Squares Discriminant Analysis (PLS-DA, Barker and Rayens (2003) which also seeks for linear combination of variables in a classification framework. Unlike LDA, PLS-DA does not encounter numerical issues with large data sets. Using PLS-DA, the classification error obtained was 15.05% (5-fold cross-validation repeated 50 times). Although not directly comparable to the performance obtained by other methods (e.g. random forests) as the validation method can be different (bootstrap vs. cross-validation), PLS-DA enables to visualise the samples, similarly to a PCA output. Figures 1.6 shows that a slight improvement is obtained compared to the previous tested approaches to better separate the genotypes with the anaerobic conditions but is still not completely satisfactory.

Remark 4. Note: to obtain a parsimonious models, *sparse models* have been introduced in these type of approaches to perform variable selection. It led to sparse LDA approaches, as well as sparse PLS-DA.

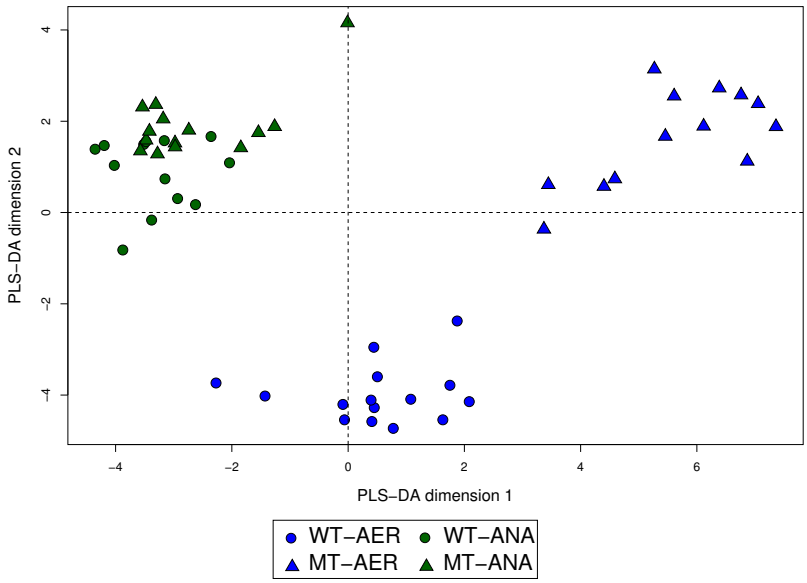


Figure 1.6: Partial Least Squares Discriminant Analysis on the yeast data set and sample representation on the first two PLS-DA dimensions. Each dot represents a sample.

1.5.6 sparse PLS-DA

sPLS-DA is a variant of PLS-DA that enables feature selection (Lê Cao *et al.*, 2011). In this method, loading vectors, which indicate the importance of each variable in the linear combination or PLS-DA components are shrunk towards zero using LASSO penalisations. The result is a small subset of features that define each PLS-DA component and the identification of a molecular signature that discriminate the groups of samples.

1.5.7 Parameter tuning in PLS-DA and sPLS-DA

For those methods, the parameters we need to choose are:

1. The number of components that reduce the dimension of the data (at least $K - 1$, where K is the number of classes to discriminate),

2. For sPLS-DA only, the number of variables to select on each component.

Those parameters can be set ad-hoc – depending on the complexity of the data and the number of samples, or using cross-validation, as detailed below, using the `mixOmics` package..

Number of components. Using cross-validation, we assess the performance of the method for a 'sufficiently' large number of components using the function `perf`. Performance is measured via overall misclassification error rate and Balanced Error Rate (BER). BER is appropriate in case of an unbalanced number of samples per class as it calculates the average proportion of wrongly classified samples in each class, weighted by the number of samples in each class. Therefore, BER is less biased towards majority classes during the performance assessment. The choice of the parameters is made according to the best prediction accuracy, i.e. the lowest overall error rate or lowest BER. Different prediction distances can be used, see Rohart *et al.* (2017b) Supplemental material for more details.

Number of variables to select (sPLS-DA). The performance of the model is assessed for each value of number of variables `keepX` to select provided as a grid by the user from the first component to the last component, one component at a time. The grid needs to be carefully chosen to achieve a trade-off between resolution and computational time. Firstly, one should consider the minimum and maximum values of the selection size that can be handled practically for follow-up analyses (e.g. wet-lab experiments may require a small signature, gene ontology a large signature). Secondly, one should consider the computational aspect, as the tuning function performs repeated cross-validation. A coarse tuning grid can be assessed first to evaluate the likely boundaries of the `keepX` values before setting a finer grid.
The `tune` function returns the set of `keepX` values that achieve the best predictive performance for each of the components in the model.

1.6 Summary

Data exploration or data mining for high-throughput biological datasets often requires dimension reduction to visualise and better understand the characteristics of the data.

Principal Component Analysis (PCA) is a multivariate dimension reduction technique and provides a good way to visualise the data, using sample and variable plots together (biplot). The principal components are built so that their variance is maximised. A component is a linear combination of the observed variables, whereas a component is an artificial variable. PCA is an unsupervised method that does not take into account any information about known sample groups during the analysis.

Classification, or supervised analysis aims to separate known groups of samples.

Projection to Latent Structure Discriminant Analysis (PLS-DA) is a supervised multivariate methd that maximises the covariance between the data and the sample group information. PLS-DA components also reduce the dimension of the data. In addition, sparse PLS-DA enables to perform variable selection and identify the key molecules or variables that drive the separation between sample groups, if such separation exist in the data.

In classification methods, we often use cross-validation or subsampling that create an artificial test set either to assess the prediction performance of a method, or to choose the hyperparameters.

	Exploratory analysis	Discriminant Analysis
Type of approach	Unsupervised approach	Supervised Learning
Samples	Samples are not labelled	Samples are labelled into known categories
Purpose	To identify sources of variation and experimental biases	To separate distinct sets of samples and define classification rules
Methods	PCA, MDS	LDA, PLS-DA
After training (usage)	Not Applicable	To classify the samples and/or predict the class of new sample; to identify key biomarkers in a large dataset

Table 1.1: The table summarises the different types of methodologies covered in this Chapter.

1.7 To go further

In this chapter we focused on the analysis of one single 'omics data set. Multi-'omics assays that measure multiple 'omics on the same biological samples are becoming more common and require an integrative approach, often data-driven and hypothesis free. To know more about how we can use similar component-based multivariate methods as presented here, you can have a look at our toolkit `mixOmics`, and in particular the DIABLO (Singh *et al.*, 2017) and the MINT (Rohart *et al.*, 2017a) frameworks.

Bibliography

Barker and Rayens (2003). Partial least squares for discrimination. *J. of Chemometrics*, **17**, 166–173.

Borg and Groenen (2005). *'Modern Multidimensional Scaling: theory and applications*. Springer-Verlag.

González, I., Lê Cao, K.-A., Davis, M. J., Déjean, S., *et al.* (2012). Visualising associations between paired 'omics' data sets. *BioData mining*, **5**(1), 19.

Jolliffe, I. (2005). *Principal component analysis*. Wiley Online Library.

Kruskal and Wish (1978). *Multidimensional Scaling*. Sage Publications.

Lê Cao, K.-A., Boitard, S., and Besse, P. (2011). Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics*, **12**(1), 253.

Rohart, F., Matigian, N., Eslami, A., S, B., and Lê Cao, K.-A. (2017a). Mint: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms.

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017b). mixomics: an r package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, **13**(11).

Singh, A., Gautier, B., Shannon, C., Rohart, F., Vacher, M., S, T., and Lê Cao, K.-A. (2017). Diablo: identifying key molecular drivers from multi-omic assays, an integrative approach.

Villas-Boâs, S., Akesson, M., Moxley, J., Stephanopoulos, G., and Nielsen, J. (2005). High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem. J*, **388**, 669–677.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, **2**(1), 37–52.

Appendix A

Appendix

A.1 Correlation and variance-covariance matrices

The covariance matrix (also known as dispersion matrix or variance - covariance matrix) is fundamental to several statistical analyses, including PLS and CCA methods. It generalizes the notion of variance to multiple variables.
A close cousin to the covariance matrix is the **correlation matrix**, which is a table reporting every pairwise correlation between two numerical variables.

Example. Pearson's correlation matrix on the Iris data with the 4 variables `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`:

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

We can see that there is a high correlation between Sepal and Petal Length ($\mathbf{r} = 0.87$).

The correlation matrix is symmetric and all the correlations on the diagonal are equal to 1 as they are the correlation of each variable with itself.

A **covariance matrix** for a data matrix composed of 4 variables can be written as:

$$\begin{pmatrix} \text{Var}_1 & \text{Cov}_{1,2} & \text{Cov}_{1,3} & \text{Cov}_{1,4} \\ \text{Cov}_{2,1} & \text{Var}_2 & \text{Cov}_{2,3} & \text{Cov}_{2,4} \\ \text{Cov}_{3,1} & \text{Cov}_{3,2} & \text{Var}_3 & \text{Cov}_{3,4} \\ \text{Cov}_{4,1} & \text{Cov}_{4,2} & \text{Cov}_{4,3} & \text{Var}_4 \end{pmatrix}$$

A covariance matrix is very similar to a correlation matrix, except for two aspects:

1. The covariance between two variables is an *unstandardized* version of their correlation (remember that in the correlation coefficient we divide the covariance by the standard deviation of both variables to remove units of measurement). The covariance is therefore a correlation measured in the units of the original variables.

2. Contrary to the correlation coefficient, the covariance coefficient is not between -1 or 1. Similar to a correlation coefficient, a value of 0 indicates no linear relationship.

Note that since the covariance is in the original units of the variables, variables on scales with bigger numbers, and with wider distributions, will necessarily have bigger covariances. Below is the variance-covariance matrix of the Iris data:

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```