Latent representation in metabolomics

Evariste NJOMGUE[1] and Silvia BOTTINI[2,3]

[1] Student at Université Côte d'Azur, France
evariste.njomgue-fotso@etu.univ-cotedazur.fr
[2] Operational director of the Medical Data Laboratory, MSI, UCA, France
silvia.bottini@univ-cotedazur.fr

Abstract. Modern metabolomics experiments yield high-dimensional datasets with hundreds to thousands of measured metabolites in large human studies with thousands of participants. Such datasets are routinely generated to profile the molecular phenotype of disease and identify the underlying pathological mechanisms of action. Extracting systemic effects from high-dimensional datasets requires dimensionality reduction approaches to untangle the high number of metabolites into the processes in which they participate. Latent representation also called latent space, in contrast to observed variables, are information that are not measurable. As example, blood pressure, pH or body temperature are measurable, while pain, stress or happiness are not directly measurable; they are hidden. In this report, the aim is to review how the use of latent space modeling has been improving these kind of studies. We will summarize the main techniques to infer latent representation from observed features, their applications in metabolomics data, which are still the open challenges and limitations.

Keywords: Metabolomics · Latent feature space · Dimension reduction.

1 Introduction

Biological systems such as cell cultures, tissues, organs, and entire organisms produce, transform, and consume small molecules (less than 1500 Da). The Da (Daltons) is the unit of measurement of the molecular mass of a given molecule. These small molecules or metabolites perform functions or tasks necessary for cell growth, defense, inhibition, and stimulation. Specifically, metabolites are substances that assist cells to communicate messages and transmit signals and may vary according to their genetic makeup and environmental stimuli [1]. Thus, metabolites are highly responsive to different stimuli. This high sensibility of metabolites and all information they carry is one of the reasons for the great interest in their study. Furthermore, metabolites disturbances have been associated with many human diseases, including cancer, diabetes, and cardiovascular disease. As a result, the identification of the metabolites produced by a biological system is a powerful approach in discovering and understanding an organism's functions and responses. Thus, metabolomics is one of the several emerging fields of biomedical research ending in the suffix "-omics," alongside

genomics, transcriptomics, and proteomics [1]. Just as genomics is the study of DNA and genetic information within a cell, and transcriptomics is the study of RNA and differences in mRNA expression; metabolomics is the study of substrates and products of metabolism, which are influenced by both genetic and environmental factors (Fig. 1).

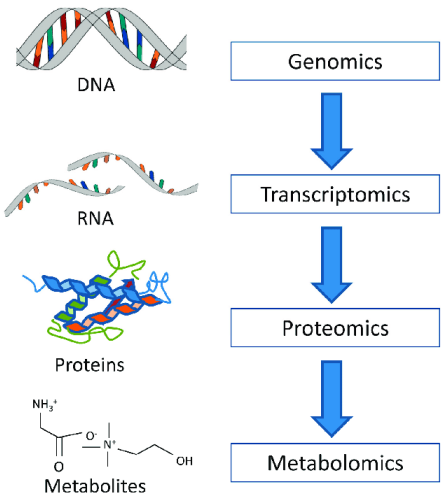

Fig. 1: Overview of the four major "omics" fields, from genomics to metabolomics. Figure from [15].

Metabolomics provides a link between metabolic pathways and the upstream genome that governs them. Thus, revealing the interactions between metabolites and others biological layers (genes, RNA, proteins) on one hand would help to improve the knowledge of genes/proteins function, and on the other hand, can be used to understand, diagnose and treat several diseases. Metabolomics is becoming a powerful tool for identifying diagnostic biomarkers of diseases, elucidating the pathological mechanisms, discovering novel drug targets, predicting drug responses, interpreting the mechanisms of drug action, as well as enabling precision treatment of patients.

The drawback of metabolomics, as long as other omics, is the curse of dimensionality presenting very large "n" (features) and very small "p" (patients/samples). For instance, the Human Metabolome Database contains 220,945 metabolite entries including both water-soluble and lipid soluble metabolites [3]. When looking specifically at human urine, it contains more than 2650 metabolite species. Thus, a techniques to reduce dimension or infer the latent space from quantified features is needed.

This report is organized in two parts. In the first part from chapter 2 to 3, we recall techniques to infer latent representations (the ones we will focus on),

then review the usage of latent spaces in literature on metabolomics data. In the second part from chapter 4 to 6, we share our experience on latent space inference applied on a metabolomics data, then discuss on our results. Our main contribution is to provide some practical applications of those techniques on an example of metabolomics data.

2 Main latent representation techniques

In the following we have chosen to split latent representation techniques into two categories: linear and non-linear. Linear techniques suppose that there is a linear relationship between the observed variables and the latent space. Under this assumption, the latent space can then be inferred from observed variables. Non-linear techniques are other tools which assume that the relationship between the latent space and observed variables is not linear. However some non-linear techniques can, under some constraints, help to infer linear relationship while linear techniques can only infer linear relationships. For example we can setup a neural network (NN) architecture to infer a linear relationship.

2.1 Latent representation: Linear techniques

Factor Analysis (FA) models are used in order to describe a larger number of observed variables by a smaller number of unobserved variables, the factors, whereby all correlation between observed variables is explained by common factors [4]. One main prerequisites to use FA is the fact that observed variables must be highly correlated. It's an advantage but also a drawback of this technique. When observed variables are highly correlated the inferred latent space by FA is very significant. However, when observed data are poorly correlated this technique is useless. FA model is defined by $X = Z * W$ where W is the loading matrix which defines the feature importance for each sample, Z the latent space and X the input data: with $Z \in \mathbb{R}^{d \times n}$ and $X \in \mathbb{R}^{p \times n}$

Principal Components Analysis (PCA) is another statistical technique for reducing the dimension of a dataset. It seeks a projection that preserves as much information in the data as possible. The main issue with PCA is when data are correlated. Indeed, it will summarize it by uncorrelated axes (principal components or principal axes).

Mixture of Probabilistic PCA (MPPCA) assumes that the mixture incorporates an internal and group specific dimension reduction through the Probabilistic PCA (PPCA). The MPPCA is defined by the following model:

- $Y \sim M(1; \pi) \Longrightarrow P(Y = k) = \pi_k$
- $Z|Y = k \sim N(Z; \mu_k, I_d)$ where $\mu_k \in \mathbb{R}^d$ with $d << p$
- $\epsilon|Y = k \sim N(\epsilon; 0, \sigma_k^2 I_p)$

- $X = U_k^t Z + \epsilon$ if $Y = k$.

U_k is the projection matrix of size $p \times d$. Z is the latent representation of the data in $\mathbb{R}^d$ and ϵ is the additive noise in $\mathbb{R}^p$. It's then possible to write the marginal distribution of X: $P(X) = \sum_{k=1}^{K} \pi_k * N(x; \mu_k, U_k^t U_k + \sigma^2 I_p)$. With $P(X)$ we can recognize the gaussian mixture form and X defined in the last previous equation recalls the PPCA model conditioned on Y

2.2 Latent representation: Non-linear techniques

In this section, Neural Network techniques to infer latent space are out of our scope. Those techniques are already applied on metabolomics as we will see in the next section. We will mainly focus on non-linear statistical techniques to infer the latent representation of our data.

Supersized Gaussian process Latent Variable Modeling (Supervized GP LVM) . The GP LVMm is a non-parametric models to infer a latent space from the data. It interprets PCA as a particular Gaussian process prior on a mapping from a latent space to the observed data-space [5]. By relaxing the constraint on the covariance function, GP LVM assumes that the observed data is generated from a lower dimensional data Z. The generation process is given by $Z = f(X) + \epsilon$ where ϵ is the noise with a gaussian distribution and f a non-linear function with a GP prior. This technique is very useful when data are very complex and difficult to interpret. The Supersized GP LVM framework uses the latent variables to connect observations and their corresponding labels.

High dimensional discriminant analysis (HDDA) is the generalisation of MPPCA by relaxing the constraint on d and by allowing other constraints on model parameters (especially on the shape of the variance of the data). So the second equation of MPPCA is updated as $Z|Y = k \sim N(Z; \mu_k, I_{d_k})$ where $\mu_k \in \mathbb{R}^{d_k}$. Thus each component can have it's own dimension. The main avantage of MPPCA as HDDA relates to the fact that the gaussian mixture model is flexible and can be adapted to high dimensional data. However it's not easy to visualize because of the different sub-spaces.

Non-linear Probabilistic PCA is the non-linear version of PPCA.

3 Literature review: use of latent representation in metabolomics data analysis

Latent space representations in metabolomics have been applied to classify individuals based on their phenotype (i.e. healty/disease). Following, some examples from the literature that use these methodology.

Nyamundanda et al. [7] had successfully used Probabilistic Principal Component Analysis (PPCA) to identify metabolites which were responsive to pentylenetetrazole (treatment used in the study). They also used a Mixture of PPCA (MPPCA) to simultaneously cluster and reduce the dimension of metabolites data. They have demonstrated that the application of those techniques help in the identification of disease phenotypes or treatment responsive phenotypes.

Gomari1 et al. [13] trained a VAE model on 217 metabolite measurements in 4644 blood samples from the TwinsUK study. They analysed the features importance with Shapley Additive Global Importance (SAGE) technique at different levels such as metabolites, sub-pathways, and super-pathways. They showed that VAE latent dimensions capture a complex mix of functions related to subpathways, thus capturing major metabolic processes in the dataset. For instance, one latent dimension captures essential catabolic processes, such as ketone bodies, fatty acid metabolism, and the TCA cycle. In addition, they also found that the VAE latent dimensions were generalizable to other datasets. In fact they show substantial disease associations in unseen Type 2 Diabetes.

Chardin et al. [10] proposed a supervised Auto Encoder (SAE) which provided an accurate localization of the patients in the latent space, and an efficient confidence score. The confidence score is the probability of the diagnosis, which gives an additional information to the clinician about the classification. The SAE was tested on three different metabolomic datasets : the “LUNG” , “BREAST”, and “BRAIN” datasets. Those datasets are respectively about Lung Cancer, breast tumor and glial tumors. The SAE also included a feature selection step, which was able to identify on those datasets, the metabolites known to be biologically relevant.

Metabolomics have been applied also in studies beyond human pathologies. Here some examples with a special focus on latent space applications. Hamzehzarghani et al. [8] used factor analysis to profile metabolic of spikelets of wheat cultivars, Roblin and Sumai3, respectively, susceptible and resistant to fusarium head blight. By combining factors and factor loading they were able to identify metabolites involved in pathogen-stress and their metabolic pathways of synthesis.

Date et al [9] described an improved DNN-based analytical approach that incorporated an importance estimation for each variable using a mean decrease accuracy (MDA) calculation. The performance of the DNN-MDA approach was evaluated using a data set of metabolic profiles derived from yellowfin goby that lived in various rivers throughout Japan.

Finally, GAN, HDDA, non-linear PPCA or GP LVM seems to be less used with metabolomics data. However, some examples of applications is described on other omics. Oleksii Prykhodko et al [11] had used GAN to generate random drug-like. They showed promising results to discover new molecules. Similarly, Joseph Mellor et al [14], used the regression version of gaussian process to provide a tool which gives a probability estimate that an enzyme catalyzes a reaction. They experimentally validated these estimates by applying their model to

metabolite in Escherichia coli, in order to search for the enzymes catalyzing its associated reactions.

As mentioned above Neural Networks (except GAN) are already used in metabolomics but HDDA or GP LVM seems to be less used. For our experimentation we will then focus on non-linear techniques described in chapter 2.2 then compare them with linear techniques. As we are in a classification problem, GAN is out of our scope as it's not relevant for our purpose.

4 Materials and Methods

4.1 Dataset

The metabolomics dataset used in this study is taken from (MTBLS28 [12]). This dataset includes 2945 metabolites concerning urine samples from 469 Lung Cancer patients and 536 controls. Metabolites measured in this dataset are molecules smaller than 1500 Da. Our goal is to find out the latent space which best classifies diseased vs healthy individuals by looking at the hidden patterns from metabolomics measurements. The dataset is splitted in training (80%) and testing set (20%).

4.2 Data preprocessing

Regarding data preprocessing, there is no missing values in this dataset to deals with. It's a quite common characteristics of metabolomics data. However this data set is subject of outliers. For this experimentation, outliers have been kept as is. Discussions with experts are still in progress to find out the appropriate solution to deal with those outliers.

4.3 Feature selection

A technique to select the features based on biological knowledge was implemented. To compute those features we have used the Kolmogorov–Smirnov test, by comparing the distribution of both populations (diseased vs healthy) in each feature. The P Value was set to $6e^{-7}$ to only keep 202 features. The Kolmogorov–Smirnov statistic quantifies the distance between the empirical distribution of two samples.

4.4 Methods used to infer the latent space

The Table. 1 lists all techniques we have tested and the corresponding implementation (R package) used. The feature selection, when present, is performed as explained in the paragraph 5.2.

Table 1: Setting of latent spaces techniques tested

Latent space technique	Classifier	Feature Selection	Prior/Hyper-parameters	R Package
PCA	Xgboost	No	Random search	stats and xgboost
FA	Xgboost	Yes	Random search	EFAtools and xgboost
Mixture of Probabilistic PCA (MPPCA)	Statistical classification from the inferred latent space	Yes	Default	HDclassif
High dimensional discriminant analysis (HDDA)	Statistical classification from the inferred latent space	Yes	Default	HDclassif
Supervised GP-LVM	Statistical classification from latent space	Yes	Gaussian Radial Basis kernel	kernlab
Non-linear Probabilistic PCA (PPCA)	Xgboost	No	Polynomial kernel (degree=3) for the prior	kernlab

4.5 Classification

As highlighted in Table. 1 some techniques only infer a latent space without any classifier. So we have added a classifier to reach our objective of diseased vs healthy classification in the latent space. In those cases, we have choose Xgboost as classifier. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning algorithm. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. Talking about hyper-parameters of Xgboost, they was initialized by a random search using cross validation on the training set. A 10-fold cross validation repeated twice was used for this purpose. For others techniques default hyper-parameters was used and we have defined a prior for the GP-LVM. This prior for the GP-LVM is not based on any biological knowledge of those type of metabolites and we will also discuss that point later.

4.6 Performance evaluation

For metrics, we have computed the accuracy and AUC score. As the dataset is slightly imbalanced, we will focus on the AUC ROC (Area Under The Curve Receiver Operating Characteristics) score to compare the performance of different models. As recall, the ROC curve summarizes the performance of a binary classification model on the positive class (True positive rate against False positive rate); the cancer patients in this case. The performance metrics are computed on the test set. To compare those metrics, we have compute their 95% confidence interval using the bootstrap technique with replacement (1000 iterations). The principle of bootstrapping is to assume that our sampling population (the LUNG dataset in our case) is a faithful representation of the data. We can then simulate multiple data sets by bootstrap replicates (Efron and Tibshirani (1993)). The distribution of the bootstrap estimates is an ideal approximation of the sampling distribution of our metrics, and can be used to quantify the uncertainty in our estimate, as well as to perform metrics inference.

4.7 Data availability statement

The R code is available on github.

5 Results

5.1 latent space performances

As shown in Table. 2 the best model performance (confidence interval) is got by the Xgboost classifier with the feature selection based on biological knowledge. When we compare this performance with our model reference (Xgboost with all features and no latent space technique), we can see that it's manly the lower bound of confidence interval (CI) which is improved; the upper bound is quite the same. This improvement got by the preprocessing is also confirmed by others models when it's compared with the version without the preprocessing. The Table. 2 also shows that PCA (both linear and nonlinear) gave the worst results.

Table 2: Classification results from latent spaces

Latent space model	Classifier	Pre-processing	Accuracy 95% CI	ROC AUC 95% CI
No	Xgboost: our reference	No	[0.714, 0.797]	[0.793, 0.872]

Continued on next page

Table 2: Classification results from latent spaces (Continued)

PCA: 0.9 cumulative variance explained by 400 features	Xgboost	No	[0.620, 0.703]	[0.675, 0.762]
Preprocessing	**Xgboost**	**Yes**	**[0.736, 0.807]**	**[0.815, 0.878]**
FA - dim: 17	Xgboost	Yes	[0.700, 0.769]	[0.775, 0.839]
FA - dim: 44	Xgboost	Yes	[0.691, 0.764]	[0.766, 0.837]
MPPCA	Statistical	No	[0.608, 0.686]	[0.634, 0.719]
MPPCA	Statistical	Yes	[0.634, 0.728]	[0.688, 0.793]
HDDA	Statistical	No	[0.571, 0.685]	[0.614, 0.734]
HDDA	Statistical	Yes	[0.570, 0.715]	[0.664, 0.780]
GP-LVM	Statistical	No	[0.674, 0.759]	[0.754, 0.834]
GP-LVM	Statistical	Yes	[0.722, 0.788]	[0.800, 0.865]
Non-linear PPCA - dim: 50	Xgboost	No	[0.581, 0.687]	[0.615, 0.743]

Factor Analysis was run with two different numbers of factors (17 and 44). Those numbers was suggested by most of the factor retention criteria. When we combine the preprocessing and Factor Analysis with 17 factors, we can reduce from 2956 features to 17 features for the latent space with a not too bad model performance. This result gives us the hint that there is something we can learn from the correlation on those features; this point will be discussed later. When we look at performance results got for MPPCA (both constraint dimension and without the constraint), they are between PCA and Factor Analysis. To end this section on results, when we look at the performance of the GP-LVM with the preprocessing, it's better than Factor Analysis: this means that there are some non-linearities which must be considered.

5.2 Features importance

For feature importance analysis, we will only focus on 4 models: our reference model, the best one and both Factor Analysis, as features analysis is a drawback of other models.

The top 3 features importance for our reference and best model are respectively (MZ 264.12 / MZ 358.11 / mz 441.16) and (MZ 264.12 / mz 441.16 / MZ 269.12). We can notice that the feature "MZ 264.12" is the top feature importance of both models (reference model and the best one). This feature most likely

corresponds to creatine riboside (Mathe et al. [16]). Creatine riboside (CR) is a novel metabolite of cancer metabolism [17]. It is a urinary diagnostic biomarker of lung and liver cancer risk and prognosis. The level of CR is highly positive correlated in tumor and urine indicating that it is derived from human lung and liver cancers.

Table 3: FA 17 - Top 5 factors importance and the position of "MZ 264.12"

Top 5	Top 3 Features with highest correlation with the factor
F3	MZ 422.21 (0.74) — MZ 561.34 (0.78) — MZ 638.36 (0.79) — **MZ 264.12 (0.55)**
F2	mz 352.08 (0.83) — mz 353.08 (0.82) — MZ 354.10 (0.88)
F17	MZ 286.09 (0.66) — MZ 286.14 (0.66) — MZ 370.05 (0.48)
F5	mz 125.09 (0.85) — mz 187.09 (0.89) — MZ 211.09 (0.88)
F6	MZ 284.18 (0.90) — MZ 286.20 (0.95) — MZ 287.20 (0.95)

Table 4: FA 44 - Top 5 factors importance and the position of "MZ 264.12"

Top 5	Top 3 Features with highest correlation with the factor
F6	MZ 284.18 (0.90) — MZ 286.20 (0.97) — MZ 287.20 (0.97)
F21	mz 613.35 (0.78) — MZ 561.34 (0.76) — MZ 638.36 (0.80)
F3	MZ 422.21 (0.89) — MZ 486.25 (0.65) — MZ 584.26 (0.79)
F14	MZ 247.09 (0.85) — MZ 247.13 (0.83) — MZ 264.12 (0.55) — **MZ 264.12 (0.55)**
F1	MZ 423.00 (0.92) — MZ 437.97 (0.92) — MZ 438.97 (0.92)

In Tables 3 and 4 we highlight the top 3 features with highest correlation with the corresponding factor. We can notice that the metabolite "MZ 264.12" is correlated with the top factor F3 of the model FA 17 and with the fourth factor F14 of the model FA 44. This important metabolite ("MZ 264.12") is also in the top 3 of FA models. Altogether these findings validate our analysis.

6 Conclusions and perspectives

Inferring the latent space from observed features has given successful results in research. Thanks to technologies improving, those techniques found various applications in biological studies and resulted in numerous new findings in their respective fields. Here we have focused on the use of latent space to improve samples classifications based on metabolites measurements. Overall, we have seen that once a prior knowledge (the preprocessing in our case) can be added to infer the suitable latent space, all those methods performs better. On one hand this result confirm the fact the preprocessing technique chosen to reduce

the data dimension is appropriated, but on the other hand, as there is no improvement on the upper bound of the CI (best model compared to the reference one), we may suspect the fact that this technique is too restrictive; we have probably removed some features with significant differences between both population. This intuition is confirmed by the analysis done on features importance where 3 features from the top 10 of our reference model was removed.

The poor performance given by both PCA techniques (linear and non-linear) makes sense when we compare it with promising results got with Factor Analysis. Indeed Factor Analysis shows that those data are highly correlated. So it's quite normal that using PCA which summarizes data by uncorrelated components, will decrease model performance.

The GP-LVM is an elegant and powerful technique once we know the relevant prior for your data. In our case we can see that with a Gaussian kernel, without including all biological knowledge of the data, the model performance is quite correct. It's a good alternative to NN as it doesn't need a lot of data to train the model.

With Factor analysis we have seen that the dataset is highly correlated. When we look at factors F6 and F1 in Tables 3 and 4, it may be interesting to analyze those highest correlation (> 0.9 for example) to improve the measurements techniques. In fact if we can measure less correlated metabolites and focus more on significant ones, we can expect a model performance improvement. With this experimentation, we realize that those latent space techniques can be also used to improve our knowledge on metabolomics data. We can then added those prior knowledge in the model settings. Indeed an improvement in metabolites measurements combined with a better understanding on the non-linearity of the data may help to design the suitable prior for GP-LVM for example.

We can conclude that, with methods tested we have highlight one key point which may be to use those techniques to improve our knowledge on metabolomics data, then add those prior into models to enhance their performance. To go further with this experimentation, one can fine tune the models hyper-parameters or try others kernels for non-linear techniques. There are few applications of latent space on metabolomics data despite their current usage for other omics. Thus we hope our experimentation will contribute to increase practical applications with its promising results.

References

1. Yolanda Smith, AZO LIFE SCIENCES, What is Metabolomics?, 2021.
2. Schmidt et al., CA: A Cancer Journal for Clinicians, Metabolomics in cancer research and emerging applications in clinical oncology, 2021.
3. Human Metabolome Database Version 5.0, `https://hmdb.ca/`.
4. Iosifina Pournara et al., BMC Bioinformatics, Factor analysis for gene regulatory networks and transcription factor activity profiles, 2007.
5. Lawrence, Neil, Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data, 2003

6. Kramer, Mark A., Nonlinear principal component analysis using autoassociative neural networks, 1991.
7. Nyamundanda et al, Probabilistic principal component analysis for metabolomic data, 2010.
8. H. Hamzehzarghani et al, Metabolic profiling and factor analysis to discriminate quantitative resistance in wheat cultivars against fusarium head blight, 2005.
9. Yasuhiro Date et al, Application of a Deep Neural Network to Metabolomics Studies and Its Performance in Determining Important Variables, 2017.
10. David Chardin et al, Learning a confidence score and the latent space of a new supervised autoencoder for diagnosis and prognosis in clinical metabolomic studies, 2022.
11. Oleksii Prykhodko et al, A de novo molecular generation method using latent vector based generative adversarial network, 2019.
12. Cancer Research, Mathé, Ewy A. et al, Noninvasive Urinary Metabolomic Profiling Identifies Diagnostic and Prognostic Markers in Lung Cancer, 2014.
13. Daniel P. Gomari1 et al, Variational autoencoders learn transferrable representations of metabolomics data, 2022.
14. Joseph Mellor et al, Semisupervised Gaussian Process for Automated Enzyme Search, 2016.
15. Tonje Haukaas et al, Metabolic Portraits of Breast Cancer by HR MAS MR Spectroscopy of Intact Tissue Samples, 2017.
16. Mathé E, et al. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer Res. 2014;74(12):3259–70.
17. Daxesh P Patel, et al. Improved detection and precise relative quantification of the urinary cancer metabolite biomarkers - Creatine riboside, creatinine riboside, creatine and creatinine by UPLC-ESI-MS/MS: Application to the NCI-Maryland cohort population controls and lung cancer cases, 2020.