Politecnico di Torino

Politecnico di Torino

Master's Degree in Mathematical Engineering

A.Y. 2024/2025

Graduation Session November 2025

Copula Graphical Modeling of Proteomic Data in Thyroid Lesions

Supervisors:
Prof. Enrico Bibbona
Prof. Giulia Capitoli

Candidate:
Marta Nesteruk

Abstract

Thyroid nodules are a common occurrence in the general population, yet preoperative assessment of malignancy often remains inconclusive. This uncertainty can lead to surgical decisions that later may prove unnecessary. Identifying potential molecular biomarkers could refine diagnosis and support more informed care. In this setting, proteomic profiling with Matrix Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) has shown promise for characterizing tissue at the molecular level.

This thesis presents a comparative statistical study of MALDI-MSI proteomic data from thyroid biopsies classified into five diagnostic categories: Follicular Adenoma, Hürthle Cell Adenoma, Papillary Thyroid Carcinoma (PTC), Follicular Variant of Papillary Thyroid Carcinoma (FVPTC), and Noninvasive Follicular Thyroid Neoplasm with Papillary-like Nuclear Features (NIFTP), a recently defined and diagnostically challenging entity.

The analysis compares univariate and multivariate statistical approaches to explore molecular differences between diagnostic groups. Elastic Net regression and sparse Partial Least Squares Discriminant Analysis (sPLS-DA) are employed as supervised methods for feature selection and discrimination among diagnostic groups. In contrast, a Copula Graphical Model for Heterogeneous Data [1] is applied in this proteomic setting to characterize conditional dependencies among molecules and to explore how network structure may inform feature relevance across diagnostic categories.

The thesis compares the molecules identified by different methods, examining their recurrence across approaches to highlight patterns that may be of interest for future clinical studies.

Acknowledgements

I would like to thank Prof. Bibbona and Prof. Capitoli for their guidance and support throughout the development of this thesis.
To my mum and dad, thank you for giving me the chance to be where I am today.
To my family, for your unconditional love.
To my friends, for your constant support.
To Leonardo, for believing in me even when I couldn't.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Clinical Context and Diagnostic Challenges in Thyroid Nodules

Thyroid nodules are a common clinical finding, with a rising incidence in recent decades, partly due to the increased use of high-resolution [2]. While the majority of these nodules are benign, around 5-15% are malignant, making accurate diagnosis crucial [3].

The standard diagnostic pathway involves an initial ultrasound (US) evaluation, followed by a Fine Needle Aspiration (FNA) biopsy for suspicious nodules. While FNA is a safe and cost-effective method, it yields an "indeterminate for malignancy" result in approximately 20-30% of cases. This diagnostic uncertainty often leads to surgery (thyroidectomy) to obtain a definitive histological diagnosis. However, post-operative analysis reveals that up to 80% of these resected indeterminate nodules are actually benign [4]. This results in a significant rate of overtreatment, exposing patients to unnecessary surgical risks, potential lifelong hormone replacement therapy and a reduced quality of life, while also increasing healthcare costs. The identification of novel biomarkers could improve the pre-operative characterization of thyroid lesions, thus improving patient care.

The spectrum of thyroid neoplasms comprises a diverse range of pathologies, including benign tumors like Follicular Adenoma (FA) and Hurthle Adenoma, as well as established malignancies. Among the cancers, Papillary Thyroid Carcinoma (PTC) is the most prevalent, with the Follicular Variant of PTC (FVPTC) representing one of its key subtypes. Recently, the Non-Invasive Follicular Thyroid Neoplasm (NIFTP) was introduced to the classification to de-escalate therapy for indolent tumors previously considered a non-invasive form of FVPTC [5].

A significant diagnostic challenge, however, is concentrated within the specific subgroup of follicular-patterned neoplasms: the benign FA, the malignant FVPTC

and the low-risk NIFTP. These entities present a distinct clinical problem because their frequently overlapping cytomorphological features on fine-needle aspiration (FNA) make pre-operative distinction exceptionally difficult. This diagnostic ambiguity is intensified by the NIFTP entity itself, which is morphologically heterogeneous and can be indistinguishable from benign FA at one end of the spectrum and invasive FVPTC at the other.

MALDI-MSI analyzes thin tissue sections and generates thousands of mass spectra, each corresponding to a specific spatial coordinate on the sample. This technique creates a detailed molecular map, showing the spatial distribution of molecules directly within the tissue. By integrating this molecular data with traditional morphological information from pathology, MALDI-MSI can identify unique proteomic "fingerprints" associated with different tissue types, such as benign versus malignant cells. Preliminary studies have demonstrated that MALDI-MSI can differentiate thyroid lesions, suggesting it may be useful for discovering biomarkers to aid in diagnostically challenging cases [6].

1.2 Study Dataset, Objectives and Thesis Outline

This thesis utilizes a dataset derived from MALDI-MSI analysis of 332 thyroid nodule biopsies collected from 183 patients. The data represents the average proteomic profiles for each biopsy, containing the mean intensity of 70 molecules identified by their mass-to-charge ratio (m/z). Each sample corresponds to one of the five diagnostic categories discussed previously (FA, Hurthle Adenoma, PTC, FVPTC and NIFTP).

The main objective of this thesis is to evaluate the ability of Copula Graphical Models for Heterogeneous Mixed Data to capture and visualize the dependence structure among molecular features measured by MALDI-MSI and to compare the molecules chosen by this method with the ones deemed relevant by the traditional statistical classification models. To achieve this goal, the thesis is organized as follows:

- Chapter 2 presents the exploratory data analysis, including preprocessing, transformation and dimensionality reduction, aimed at understanding the distributional properties of the data.
- Chapter 3 describes the classical statistical models used as benchmarks: Elastic Net regression and sparse Partial Least Squares Discriminant Analysis (sPLS-DA).
- Chapter 4 provides a detailed explanation of the Copula Graphical Model for Heterogeneous Mixed Data, including its theoretical foundations and the algorithm used for graph construction.

- Chapter 5 reports and discusses the results obtained from both classical and graphical approaches, highlighting their differences in feature selection and interpretability.
- Chapter 6 concludes the thesis by summarizing the main findings, discussing limitations and suggesting directions for future research.

Chapter 2

Exploratory Data Analysis

This chapter details the exploratory data analysis (EDA) performed on the thyroid nodule proteomic dataset introduced in Chapter 1. The EDA aims to characterize the inherent structure, distributional properties and potential challenges within the data before proceeding to the application of statistical models in subsequent chapters. Key steps include initial data inspection, analysis of raw data distributions, evaluation of data transformations, correlation analysis and dimensionality reduction techniques for visualization.

The study cohort was composed of 183 patients who underwent thyroid surgery for different thyroid tumors at the IRCCS Fondazione San Gerardo dei Tintori, Monza, Italy, between 2021–2024. The study protocol was reviewed and approved by the local ethical committee (Comitato Etico Brianza, via Pergolesi, 33, 20900 Monza, Italy; approval number FINAL-TIR PU 3581/21; approval date: January 14, 2021). All subjects enrolled in the study signed an informed consent. It consists of 332 tissue core samples, each described by the mean abundance of 70 molecular features derived from MALDI-MSI analyses of thyroid biopsies. These samples cover five distinct diagnoses:

- two benign: Follicular Adenoma (FA) and Hurthle Adenoma;
- two cancerous: Papillary Thyroid Carcinoma (PTC) and Follicular Variant of PTC (FVPTC);
- Non-Invasive Follicular Thyroid Neoplasm with Papillary-like Nuclear Features (NIFTP).

The data were provided in anonymized form for statistical analysis; the author was not involved in patient recruitment or sample collection. The molecular profiles had already undergone standard preprocessing and quality control procedures typical for MALDI-MSI data. As received, the dataset contained no missing values or evident outliers.

Each observation corresponds to the averaged proteomic profile of a single tissue core and includes identifiers (`Core_ID`, `Patient_ID`), the histological `Diagnosis` and 70 variables representing the mean intensity for each molecular feature. These intensity values provide a quantitative measure of the abundance of each detected molecule across thousands of pixels within that tissue core: higher values indicating a greater average molecular presence in the sample.

2.1 Descriptive Statistics and Raw Data Characteristics

Before applying any transformations or modeling, an initial characterization of the dataset was performed to explore its composition and summarize the properties of the raw molecular measurements.

2.1.1 Sample Distribution

The dataset comprises 332 tissue core samples distributed across five distinct histological diagnoses. The distribution of cores across diagnostic classes is shown in Figure 2.1.

A class imbalance is evident within the dataset: NIFTP constitutes the majority

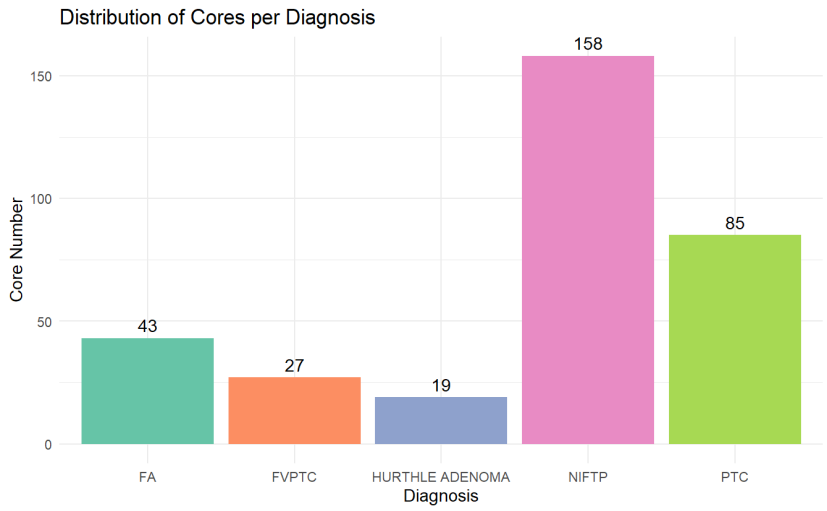


Figure 2.1: Distribution of the 332 tissue cores across the five diagnostic categories.

category (158 cores, 47.6%), whereas Hurthle Adenoma (19 cores, 5.7%) and FVPTC (27 cores, 8.1%) are less represented. This imbalance should be considered

in subsequent modeling phases, as it may influence classifier performance and interpretability.

The dataset was not designed to mirror the real-world distribution of thyroid tumor types. In clinical practice, Papillary Thyroid Carcinoma (PTC) is by far the most frequent diagnosis, while entities such as NIFTP and Hurthle Adenoma are relatively rare [7]. The dataset composition was therefore adjusted to include a sufficient number of cases from each histological category for meaningful statistical analysis.

2.1.2 Patient Contribution

The 332 samples were derived from 183 patients. The number of tissue cores contributed by each patient varies, as summarized in Table 2.1.

Cores per patient	Number of patients	Percentage
1	42	22.94 %
2	138	75.41 %
3	1	0.55 %
4	1	0.55 %
7	1	0.55 %

Table 2.1: Distribution of cores per patient.

The majority of patients contributed one or two cores, while only three patients provided three or more cores.

Although multiple tissue cores were often obtained from the same patient, these samples were treated as statistically independent observations. This assumption is commonly adopted in MALDI-MSI proteomic studies because of the limited number of available patients and the high experimental cost of data acquisition. The low incidence of certain thyroid neoplasms further constrains patient recruitment, making it necessary to increase the sample size by including multiple cores per subject. Each tissue core is analyzed independently, as it can capture intra-patient heterogeneity at the molecular level. This approach allows a more robust characterization of the proteomic variability across diagnostic groups, despite the potential within-patient correlations [8].

2.1.3 Raw Molecular Data Distributions

The raw molecular intensity data were examined using density and frequency plots for a subset of representative molecules (Figure 2.2), and boxplots displaying the distributions of all 70 molecular features across the five diagnostic groups (Figure

2.3). Together, these representations provide an overview of the distributional characteristics of the mass spectrometry data and of the variability across diagnostic classes.

The density and frequency plots (Figure 2.2) show that most molecular features exhibit a pronounced right-skewness, with the bulk of observations concentrated at lower intensities with a long tail extending towards higher values. This indicates that, for many features, only a small subset of samples expresses particularly high abundance levels. Such skewness is characteristic of MALDI-MSI data.

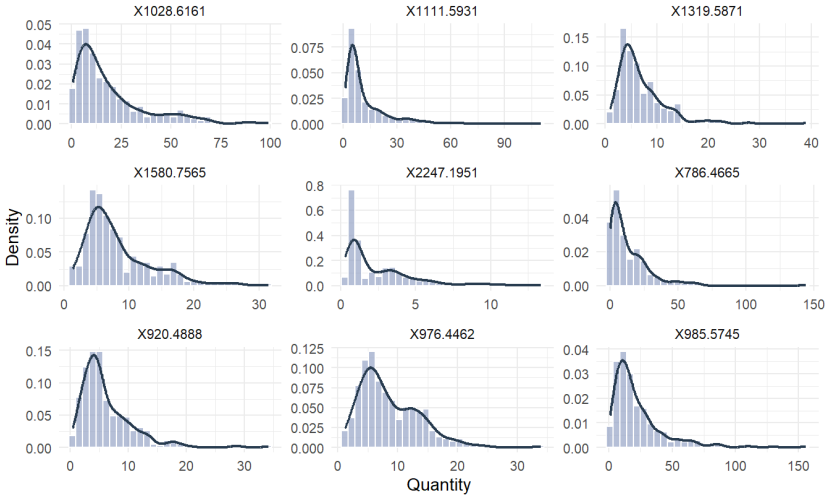
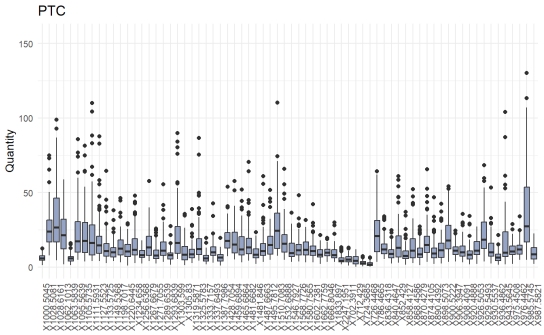
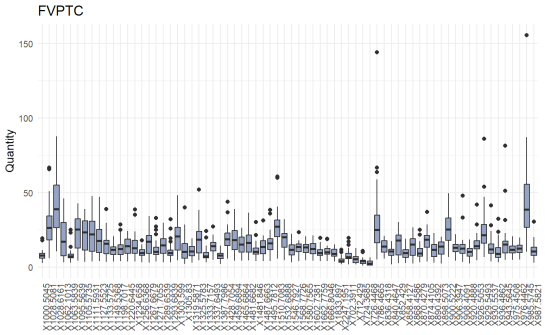


Figure 2.2: Distribution of raw abundance values for selected molecular features: X1319.5871, X1580.7565, X976.4462, X920.4888, X2247.1951 selected randomly, and X985.5745, X1028.6161, X786.4665, X1111.5931 selected for their high variance.

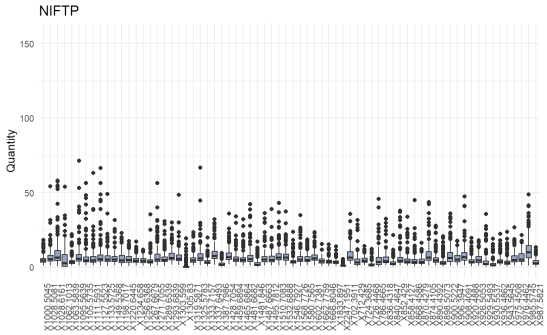
The boxplots in Figure 2.3 confirm these characteristics while illustrating both inter- and intra-group variability. Within each diagnostic category, the distributions of the 70 molecular features differ considerably in scale and dispersion. In PTC and FVPTC samples, the boxes tend to be more extended and positioned at higher intensity levels, accompanied by numerous high-value outliers indicating greater variability and overall higher molecular abundance. In contrast, NIFTP samples display more compact distributions with lower median intensities, although several outliers remain present. Hurthle Adenoma and Follicular Adenoma samples, representing benign diagnoses, are characterized by narrow boxes concentrated near zero and comparatively few outliers, reflecting more homogeneous and lower-intensity molecular profiles.



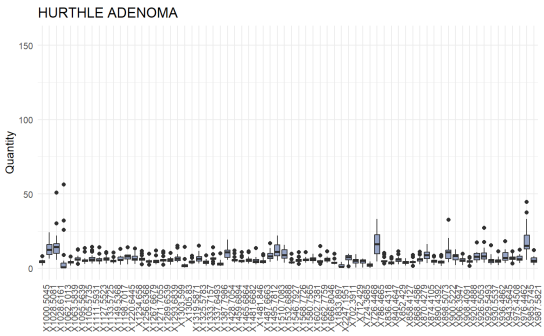
(a) Boxplot of the 70 molecules for PTC



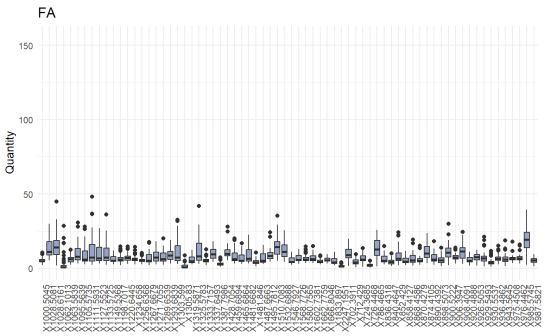
(b) Boxplot of the 70 molecules for FVPTC



(c) Boxplot of the 70 molecules for NIFTP



(d) Boxplot of the 70 molecules for Hurthle Adenoma



(e) Boxplot of the 70 molecules for FA

Figure 2.3: Boxplots showing the distribution of raw abundance for 70 molecules across diagnostic classes:(a) and (b) represent malignant diagnoses, (c) represents the NIFTP classification, (d) and (e) represent benign diagnoses.

2.2 Data Transformation and Normalization

To address the issues of wide dynamic range, strong right-skewness and heteroscedasticity typical of mass spectrometry-derived values, the dataset was first subjected to a logarithmic transformation and subsequent z-score normalization, which together reduce the influence of extreme values, stabilize variances and ensure all molecular features are comparable in scale [9].

2.2.1 Logarithmic Transformation

A logarithmic transformation ($\log(x+1)$) was applied to all 70 molecular intensity features. This transformation compressed the range of the data, reducing the dominance of extreme high-intensity values while preserving the relative differences among samples.

As illustrated in Figure 2.4, the heavy right tails observed in the raw data were substantially reduced, although moderate asymmetry persisted for several molecules. Quantitatively, the median skewness across molecular features decreased from 2.09 to 0.26, confirming a notable improvement in distributional symmetry.

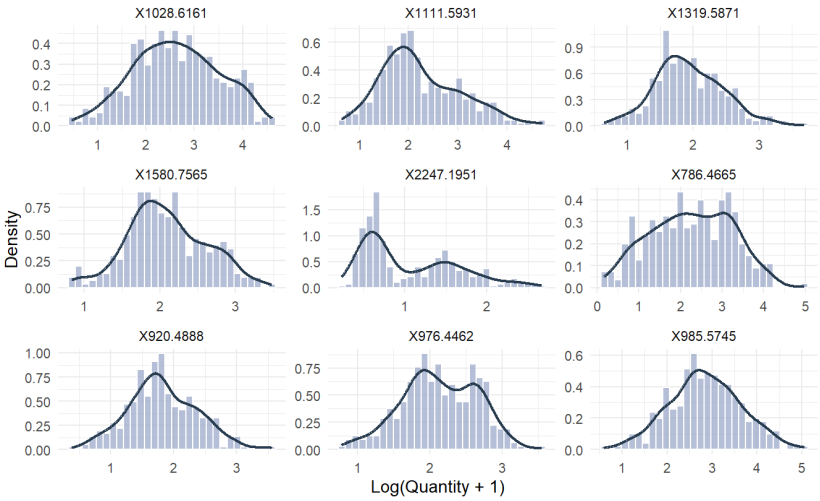


Figure 2.4: Density and frequency distributions of selected molecular features (same as in Figure 2.2) after logarithmic transformation.

2.2.2 Normalization by Z-Scoring

Following the logarithmic transformation, each molecular feature was standardized to zero mean and unit variance across all samples. For each molecule $j \in 1, \ldots, p$,

the standardized value z_{ij} for sample i was computed as:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_j)^2}$.

This z-score normalization emphasizes relative abundance patterns rather than absolute intensity levels and prevents features with larger dynamic ranges from dominating subsequent analyses.

Because most patients contributed only one or two tissue cores, feature-wise rather than per-patient standardization was adopted.

The resulting dataset retained the structure of the log-transformed data but with features on a consistent numerical scale, providing a robust foundation for correlation analysis and dimensionality reduction in the subsequent sections.

2.3 Correlation and Covariance Analysis

Pairwise Pearson correlations among the 70 molecular features were computed to assess potential redundancy.

The resulting correlation coefficients ranged from -0.45 to 0.99, with a median correlation of approximately 0.65.

Only about 10% of feature pairs exhibited weak correlation ($|r| < 0.4$), while nearly half showed strong correlation ($|r| > 0.7$). The high degree of correlation among several features points to partial redundancy, which may arise from closely related or co-detected molecular signals in the spectra.

Such correlation structures justify the subsequent use of dimensionality reduction methods (PCA, t-SNE, UMAP) to summarize the underlying relationships and minimize information redundancy.

To further explore potential associations between molecular features and diagnostic categories, one-hot encoding of the categorical variable `Diagnosis` was performed.

The resulting binary columns were combined with the log-transformed scaled molecular intensities to compute a covariance matrix capturing molecule-diagnosis relationships. The submatrix representing correlations between diagnoses and molecular features is visualized in Figure 2.5.

This analysis revealed that only a limited subset of molecules exhibited moderate correlations ($|r| > 0.5$) with any single diagnostic class. To examine this in detail, a list of molecular features showing strong correlation ($|r| > 0.5$) with at least one class was extracted. Notably, no molecules exhibited strong associations with Follicular Adenoma (FA), Follicular Variant Papillary Thyroid Carcinoma

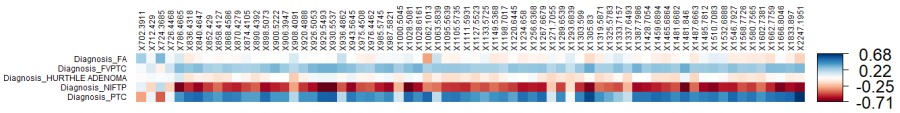


Figure 2.5: Heatmap showing correlations between molecular features and diagnostic classes (log-transformed data).

(FVPTC), or Hurthle Adenoma. In contrast, several features were moderately to strongly correlated with the NIFTP and PTC classes (36 molecules and 23 molecules, respectively), indicating partially overlapping molecular signatures between these two groups. This general lack of strong, unique molecular-to-diagnosis correlations reinforces the hypothesis that diagnostic information is not driven by single dominant biomarkers but rather by distributed, multivariate patterns among multiple features.

2.4 Volcano Plot Analysis

To perform differential-abundance analysis across diagnostic classes, `limma` (Linear Models for Microarray and Omics Data) framework was used [10], which fits a separate linear model for each molecular feature while borrowing information across features through empirical Bayes shrinkage of the residual variances.

Let Y_{ij} denote the transformed abundance of molecule j in sample i, and let X be the design matrix encoding the K diagnostic groups by means of one-hot indicator columns. For each molecule j, the coefficient vector $\beta_j = (\beta_{j,1}, \ldots, \beta_{j,K})^\top$ is estimated by fitting the linear model $Y_{\cdot,j} = X\beta_j + \epsilon_j$.

The estimated effect size for molecule j in the comparison between diagnostic groups g_1 and g_2 is $\Delta_j = \hat{\beta}_{j,g_2} - \hat{\beta}_{j,g_1}$ which corresponds to the $\log_2$ fold change on the transformed scale. Furthermore, for each j, a moderated t-statistic testing

$H_{0,j} : \Delta_j = 0$ was performed and its associated p-value was subsequently adjusted for multiple testing to yield a false-discovery-rate–controlled value q_j.

The results are summarized through volcano plots, which represent each molecule j as a point in $\mathbb{R}^2$ with coordinates $(x_j, y_j) = (\Delta_j, -\log_{10}(q_j))$.

The horizontal axis therefore reflects the signed effect size, while the vertical axis corresponds to the significance level of the test. The upper left and upper right regions of the plot contain the molecules for which simultaneously $|\Delta_j|$ is large and q_j is small, that is, features that are both strongly different between groups and statistically robust. In this work, a molecule is declared significantly deregulated in the contrast g_1 vs g_2 whenever $|\Delta_j| > 0.6$ and $q_j < 0.05$.

Positive values of Δ_j indicate relative up-regulation in group g_2, whereas negative values indicate up-regulation in group g_1. This geometric representation thus encodes the combined information of effect magnitude and statistical evidence in a single object, allowing rapid identification of molecular features driving the separation between diagnostic categories.

Across all pairwise contrasts, the volcano plots reveal a pattern that mirrors the underlying biological classification of the lesions. The comparison between the two benign entities (FA and Hurthle Adenoma) shows no molecule exceeding the predefined significance thresholds, with all features tightly concentrated around zero effect size (Figure 2.6(b)). This reflects the expected similarity between FA and Hurthle Adenoma, both of which are non-malignant. A similarly coherent picture emerges for the two malignant forms (Figure 2.6(a)): in the contrast between classical PTC and its follicular variant (FVPTC), only a single feature reaches statistical significance and the remaining molecules display minimal differences, again consistent with the known biological closeness of these two carcinoma subtypes. By contrast, the comparisons involving PTC and NIFTP or FA reveal a different structure. When PTC is contrasted with NIFTP (the low-risk but still neoplastic lesion) a large number of molecules exceed both thresholds, almost all of them up-regulated in PTC. This pronounced asymmetry indicates a clear metabolic shift associated with malignant transformation relative to NIFTP. A similarly strong signal is observed in the PTC versus FA contrast, where numerous molecules show substantially higher abundance in PTC, while only a few exhibit the opposite behavior. Taken together, these results delineate a clear gradient in molecular divergence: minimal differences between lesions of the same biological class (benign-benign or malignant-malignant) and extensive, statistically robust metabolic alterations when contrasting benign or low-risk lesions with fully malignant PTC.

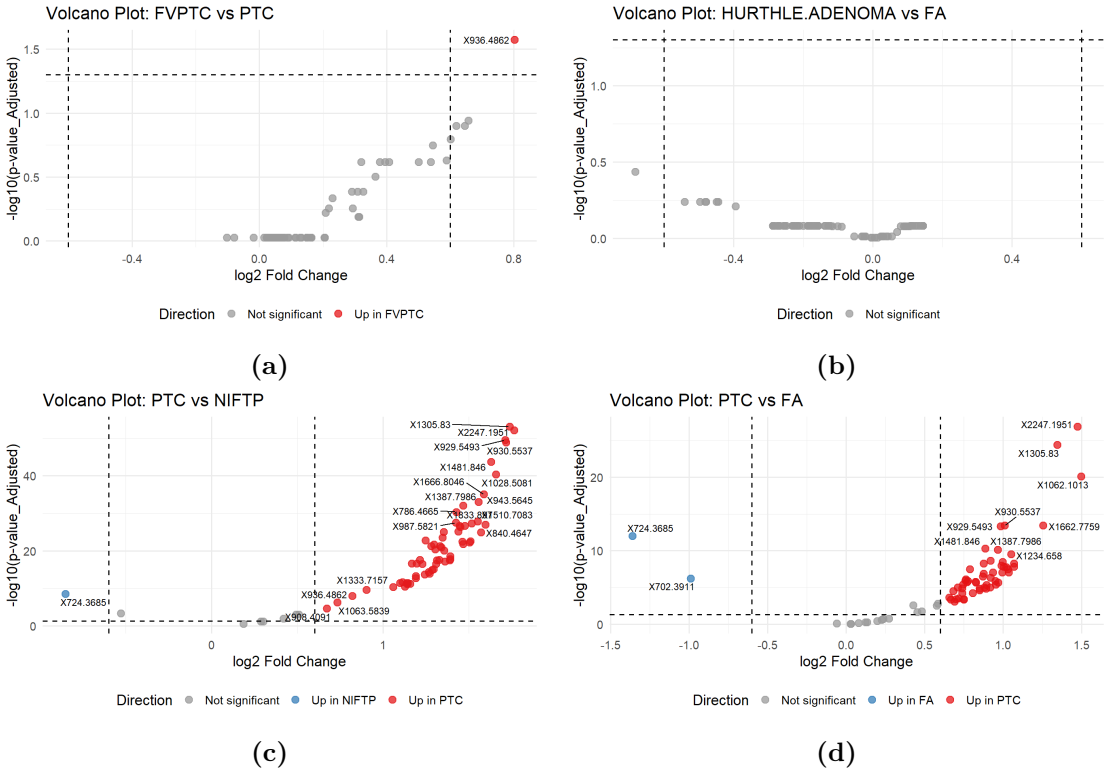


Figure 2.6: Volcano plots illustrating differential molecular abundance between diagnostic categories: (a) FVPTC vs PTC, (b) Hurthle Adenoma vs FA,(c) PTC vs NIFTP and (d) PTC vs FA. Red and blue points indicate significantly up- and down-regulated molecules, respectively.

2.5 Dimensionality Reduction for Visualization

To visualize the global data structure and assess potential diagnostic separability, three dimensionality reduction methods were applied to the transformed dataset: Principal Component Analysis (PCA) [11], t-Distributed Stochastic Neighbor Embedding (t-SNE) [12] and Uniform Manifold Approximation and Projection (UMAP) [13].

2.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was applied to the log-transformed and standardized molecular features. The first three principal components explained 69.8%, 11.0% and 7.4% of the total variance, respectively.

As shown in Figure 2.7 (a)-(b), no distinct clustering by diagnosis is evident, although partial trends can be observed, especially along the first principal component. In particular, NIFTP samples tend to occupy higher PC1 values, whereas PTC and FVPTC samples cluster more toward negative PC1 scores.

FA and Hurthle Adenoma are distributed between these two extremes, with substantial overlap across groups.

Overall, the PCA results suggest that while PC1 captures a systematic gradient across diagnostic categories, the global proteomic profiles of these groups remain largely overlapping.

2.5.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE was applied to the log-transformed and standardized molecular features to explore potential nonlinear structures in the data. The algorithm was run in three dimensions and the projections along the first two and the first and third t-SNE components are shown in Figure 2.7 (c)-(d).

Compared with PCA, t-SNE reveals more localized groupings, although substantial overlap across diagnostic categories persists. PTC and FVPTC samples tend to cluster toward the right side of the t-SNE1 axis, forming several compact regions. In contrast, NIFTP samples occupy a broader area, primarily on the left portions of the embedding. FA and Hurthle Adenoma form a cluster visible in 2.7(d), but an overall overlapping is still present.

2.5.3 Manifold Approximation and Projection (UMAP)

UMAP was applied in three dimensions to explore nonlinear relationships in the data. The projections along UMAP1–UMAP2 and UMAP1–UMAP3 are shown in Figure 2.7 (e)-(f).

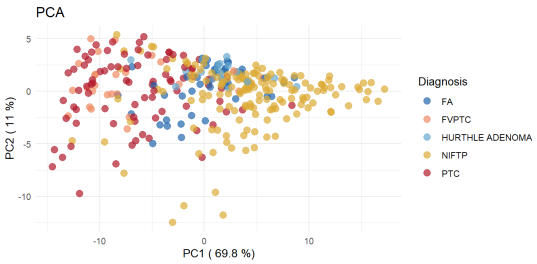
Compared with PCA and t-SNE, UMAP reveals more compact and visually distinct groupings. PTC and FVPTC samples once again form a relatively cohesive cluster at higher UMAP1 values, consistently visible in both projections. NIFTP samples are more dispersed, extending along a curved structure primarily on the right side of the embedding, while FA and Hurthle Adenoma samples appear in more localized regions toward lower UMAP1 values.

Despite these more defined structures, the diagnostic classes are not perfectly separated: several samples still lie in transition zones between clusters, indicating gradual molecular variation rather than strictly discrete boundaries.

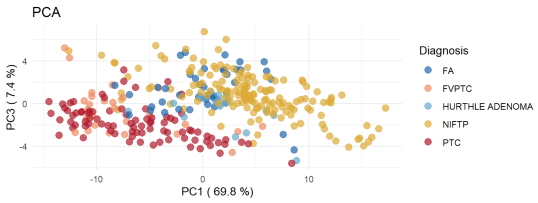
2.6 Discussion and Implications

The dataset exhibits properties that complicate standard analysis: right-skewed intensities, high feature correlation and substantial class overlap in low-dimensional views.

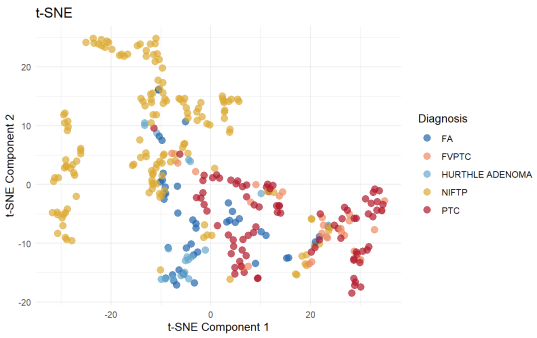
From a methodological standpoint, these properties make strong separation by simple linear or marginal tests unlikely. Instead, our approach requires multivariate models capable of handling high collinearity while pooling many weak signals. We will therefore employ a regularized linear model (Elastic Net Regression), a supervised projection method (sPLS-DA) as well as the Copula Graphical Model to capture the complex, higher-order dependencies in the data. In the next chapters (Chs. 3 and 4) we introduce these models and compare them in Chapter 5.



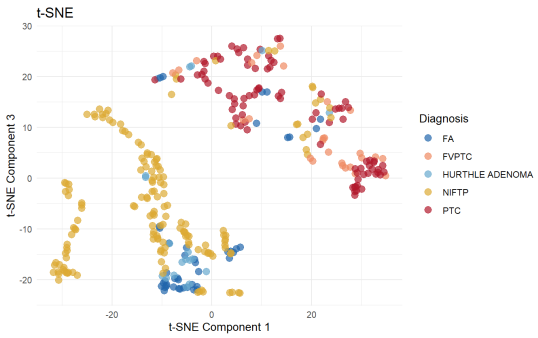
(a) PCA: Component 1 vs Component 2



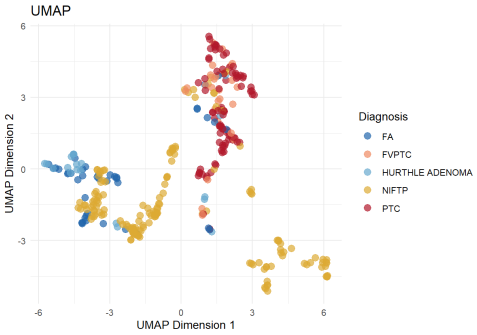
(b) PCA: Component 1 vs Component 2



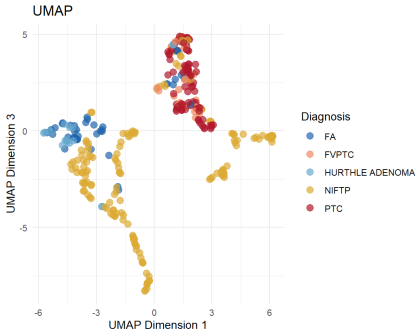
(c) t-SNE: Component 1 vs Component 2



(d) t-SNE: Component 1 vs Component 3



(e) UMAP: Component 1 vs Component 2



(f) UMAP: Component 1 vs Component 3

Figure 2.7: Two-dimensional projections of the z-scored dataset using (a)-(b) PCA, (c)-(d) t-SNE and (e)-(f) UMAP, colored by diagnostic category.

Chapter 3

Classical Statistical Models

This chapter introduces two established supervised methods adept at handling proteomic data characteristics: the Elastic Net and sparse Partial Least Squares Discriminant Analysis (sPLS-DA).

3.1 Elastic Net Regularization

Regression models aim to describe the relationship between a response variable and a set of predictors. When the response is categorical, however, the standard linear regression framework is not suitable, as it can yield predictions outside the [0,1] interval. Logistic regression provides a natural extension by modeling the probability that an observation belongs to a certain class as a function of its predictors.

In the binary case, where $Y_i \in \{0,1\}$, the logistic model assumes that the log-odds of the positive outcome are a linear function of the predictors:

$$\log\left(\frac{\mathbb{P}(Y_i = 1 \mid X_i)}{1 - \mathbb{P}(Y_i = 1 \mid X_i)}\right) = \beta_0 + X_i^\top \beta,$$

which can be rewritten as:

$$\mathbb{P}(Y_i = 1 \mid X_i) = \frac{\exp(\beta_0 + X_i^\top \beta)}{1 + \exp(\beta_0 + X_i^\top \beta)},$$

where $X_i \in \mathbb{R}^p$ is the predictor vector for sample i, β_0 is the intercept, and $\beta \in \mathbb{R}^p$ is the vector of coefficients. The parameters are estimated by maximizing the log-likelihood of the observed data.

For problems with more than two classes, logistic regression can be extended to the multinomial setting. For K possible classes, the probability that observation i

belongs to class k is modeled as:

$$\mathbb{P}(Y_i = k \mid X_i) = \frac{\exp(\beta_{0k} + X_i^\top \beta_k)}{\sum_{l=1}^{K} \exp(\beta_{0l} + X_i^\top \beta_l)}, \tag{3.1}$$

where each class k has its own intercept β_{0k} and coefficient vector β_k.

While this model can be estimated by maximum likelihood, in high-dimensional problems or when predictors are highly correlated, it can lead to overfitting and unstable coefficient estimates. To mitigate these issues, regularization techniques introduce a penalty on the magnitude of the coefficients, improving both generalization and interpretability. Two classical forms of regularization are Ridge and Lasso regression.

Ridge [14] regression applies an ℓ_2-norm penalty ($||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$), which shrinks all coefficients toward zero but rarely makes them exactly zero. It performs well when predictors are strongly correlated, as it distributes their influence more evenly.

Lasso [15], in contrast, applies an ℓ_1-norm penalty ($||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$) which encourages sparsity by forcing some coefficients to zero. This property enables variable selection, but Lasso may behave inconsistently when groups of variables are highly correlated, often retaining only one variable from such groups.

The Elastic Net [16] combines the strengths of both approaches by including a convex combination of the ℓ_1 and ℓ_2 penalties. It introduces a mixing parameter $\alpha \in [0,1]$ that controls the relative contribution of each term and a regularization strength $\lambda \geq 0$ that determines the overall degree of shrinkage.

For multinomial logistic regression, the coefficients are estimated by maximizing the penalized log-likelihood:

$$\max_{\beta_0,\beta} \left[\frac{1}{n} \sum_{i=1}^{n} \log(\mathbb{P}(Y_i = y_i \mid X_i)) - \lambda \sum_{k=1}^{K} \left(\frac{1-\alpha}{2} ||\beta_k||_2^2 + \alpha ||\beta_k||_1 \right) \right]. \tag{3.2}$$

By adjusting α and λ, the Elastic Net balances between Ridge-like stability and Lasso-like sparsity, making it particularly suitable for high-dimensional, correlated data such as proteomic measurements. This dual regularization allows it to serve as a baseline classifier and as a tool for feature selection.

3.2 Sparse Partial Least Squares Discriminant Analysis (sPLS-DA)

Partial Least Squares (PLS) [17] regression is a supervised dimension reduction technique that models the relationship between a set of predictors and response variables through a small number of latent components capturing their shared

variation. Unlike Principal Component Analysis (PCA), which seeks components that explain maximum variance within the predictors alone, PLS extracts components that maximize the covariance between the predictors X and the response Y, ensuring that the derived latent variables are directly relevant for prediction.

Formally, given a predictor matrix $X \in \mathbb{R}^{n\times p}$ and a response matrix $Y \in \mathbb{R}^{n\times q}$, PLS seeks low-rank approximations of the form:

$$X = TP^\top + E, \qquad Y = TQ^\top + F, \tag{3.3}$$

where $T \in \mathbb{R}^{n\times A}$ is the matrix of latent scores (components), $P \in \mathbb{R}^{p\times A}$ and $Q \in \mathbb{R}^{q\times A}$ are the corresponding loading matrices, and E and F are residual matrices. The columns t_h of T are constructed iteratively as linear combinations of the predictors, $t_h = Xw_h$, with weight vectors w_h, so as to maximize the covariance between t_h and Y, subject to orthogonality constraints among the extracted components. Using the same score matrix T in the decompositions of X and Y enforces a shared latent structure that captures the part of the variation in X that is most predictive of Y.

For classification problems, the response variable is categorical and can be encoded as a binary indicator matrix Y of dimension $n \times K$, where K is the number of classes and $Y_{ik} = 1$ if observation i belongs to class k. Applying PLS regression to (X, Y) in this form yields Partial Least Squares Discriminant Analysis (PLS-DA) [18], a supervised extension of PLS designed for class discrimination. The method constructs latent components that summarize the variation in X that is most relevant for distinguishing among the classes encoded in Y.

Using the decompositions in (3.3), the PLS regression coefficients can be written as:

$$B = W\left(P^\top W\right)^{-1} Q^\top, \tag{3.4}$$

so that the fitted values are

$$\hat{Y} = XB. \tag{3.5}$$

Here, $W \in \mathbb{R}^{p\times A}$ contains the weight vectors defining the components, $P \in \mathbb{R}^{p\times A}$ contains the loadings of X, and, in the classification setting where $Y \in \mathbb{R}^{n\times K}$, $Q \in \mathbb{R}^{K\times A}$ contains the loadings of Y, relating the latent scores T to the class indicators through $Y \approx TQ^\top$. The matrix $\hat{Y} \in \mathbb{R}^{n\times K}$ contains continuous scores that approximate the class indicator matrix Y. Classification is then performed by assigning each observation to the class (column) with the largest predicted score in its corresponding row.

While PLS-DA is effective for high-dimensional and highly correlated predictors, it uses dense linear combinations of all variables to form its latent components. As a consequence, it does not inherently perform feature selection and may obscure which predictors are most relevant for class discrimination. In applications such as proteomics or metabolomics, where the number of measured features far exceeds

the number of samples and interpretability is essential for biomarker discovery, this lack of sparsity can be limiting.

To address this issue, sparse Partial Least Squares Discriminant Analysis (sPLS-DA) [19] introduces an additional sparsity constraint on the weight vectors used to form the latent components. This constraint forces only a subset of predictors to contribute to each component, thereby combining supervised dimension reduction with variable selection. One way to define the first sparse component is via the optimization problem:

$$\max_{w} \left\| \Sigma_{XY}^{\top} w \right\|_2 \quad \text{subject to} \quad \|w\|_2 = 1,\ \|w\|_1 \leq \tau_x, \tag{3.6}$$

where $\Sigma_{XY} = \frac{1}{n} X^{\top} Y$ is the empirical cross-covariance matrix between predictors and responses and $\tau_x > 0$ is a sparsity parameter controlling the number of nonzero elements in w. Smaller values of τ_x produce sparser solutions, allowing only the most discriminative variables to contribute to the component. This formulation is equivalent to a penalized maximization problem with an ℓ_1-norm penalty on the predictor weights.

After the first component is computed, the data matrices are deflated by subtracting the variation explained by the component, preparing them for the extraction of subsequent sparse components. Concretely, given the score vector t_h and loadings

$$p_h = \frac{X^{(h)\top} t_h}{t_h^{\top} t_h}, \qquad q_h = \frac{Y^{(h)\top} t_h}{t_h^{\top} t_h},$$

the deflated matrices are obtained as:

$$X^{(h+1)} = X^{(h)} - t_h p_h^{\top}, \qquad Y^{(h+1)} = Y^{(h)} - t_h q_h^{\top}.$$

Repeating this procedure yields a sequence of sparse weight vectors and corresponding latent scores that capture the most predictive structure relating X and Y while maintaining interpretability through variable selection.

The introduction of sparsity has two main advantages. First, it directly identifies a limited subset of variables with the strongest discriminative power, thereby facilitating the discovery of potential biomarkers. Second, it improves generalization by reducing model complexity and mitigating overfitting in high-dimensional settings. As in PLS-DA, the resulting sPLS-DA model can be used to predict class membership for new observations using the latent components and their associated class scores.

Chapter 4

A Graphical Modeling Approach for Heterogeneous Proteomic Data

4.1 Foundations of Graphical Modeling

Graphical models are a powerful tool for analyzing complex biological data, such as the mass spectrometry results in this thesis. In these models, individual biological units (e.g., molecules) are represented as nodes and the statistical relationships between them are visualized as edges. Valuable insights into underlying biological phenomena can be gained by analyzing a single network's topology as well as by comparing how these structures change across different biological conditions.

We will focus on undirected graphical models: in this framework a multivariate random variable $X = (X_1, \ldots, X_p)$ is represented as a graph $\mathcal{G}(V, E)$, where V is the set of nodes (the p components of X) and $E \subset V \times V$ is the set of edges representing pairwise relationships.

The symmetric relationships modeled by an undirected graph are a reasonable assumption for our data. While alternative models like Directed Acyclic Graphs (DAGs) represent causality, their underlying assumptions are difficult to verify in this context. Hence, it is more appropriate to model conditional associations without implying directionality.

A common approach within the family of undirected graphical models is the Gaussian Graphical Model (GGM): a statistical learning technique used to infer conditional independence relationships from data assumed to follow a multivariate normal distribution. Therefore, $X = (X_1, \ldots, X_p) \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_p) \in \mathbb{R}^p$ is the vector of means and $\Sigma \in \mathbb{R}^{p \times p}$ is the covariance matrix.

The precision matrix is defined as $\Theta = \Sigma^{-1}$.

The conditional independence comes into play when defining the edge set E:

- the edge $(i, j) \notin E$ if and only if $\theta_{ij} = 0$;
- if the edge $(i, j) \in E$ then it is proportional to the partial correlation between X_i and X_j.

The partial correlation is defined as:

$$\rho_{X_i,X_j|X_{-i,-j}} = \frac{\mathrm{Cov}[X_i, X_j|X_{-i,-j}]}{\sqrt{\mathrm{Var}[X_i|X_{-i,-j}]}\sqrt{\mathrm{Var}[X_j|X_{-i,-j}]}}, \tag{4.1}$$

where $X_{-i,-j} = V \setminus \{X_i, X_j\}$. It follows that:

$$\rho_{X_i,X_j|X_{-i,-j}} = \frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}, \tag{4.2}$$

which clearly links the task of estimating the precision matrix Θ with defining the graph structure. This estimation is conventionally achieved through Maximum Likelihood Estimation (MLE), a well-established method for Gaussian data.

The standard Gaussian Graphical Model [20] relies on two strong assumptions that are often violated by complex, real-world datasets like the one in this thesis. The first is the Gaussian assumption, as observational biological data are typically of a mixed-type, consisting of a combination of continuous (often skewed), discrete and categorical variables. The second is the i.i.d (independent and identically distributed) assumption (needed for MLE), which is unrealistic when data originates from distinct subgroups, such as the different diagnoses in this study. While various models have been developed to address non-normal data, they often still assume homogeneity. Separately, methods were proposed to handle heterogeneous data, but these typically require normality.

To solve both problems simultaneously, Hermes, van der Heerwaarden and Behrouzi proposed the Copula Graphical Model for Heterogeneous Mixed Data [1]. This model provides the methodological foundation for the analysis conducted in this thesis.

4.2 Copula Transformation for Mixed Data

We consider the data for each of the K diagnosis groups independently ($K = 5$): let $X^{(k)} = (X_1^{(k)}, \ldots, X_p^{(k)})$ be the data matrix for group k, where $X_j^{(k)}$ is a column of length n_k (n_k is not necessarily equal to $n_{k'}$ for $k \neq k'$). The data are assumed to be i.i.d within each group, but not across groups.

The joint c.d.f (cumulative distribution function) for the group k is given by:

$$F^{(k)}(x_1^{(k)}, \ldots, x_p^{(k)}) = \mathbb{P}(X_1^{(k)} \leq x_1^{(k)}, \ldots, X_p^{(k)} \leq x_p^{(k)}) \tag{4.3}$$

The previous exploratory analysis makes it clear that the data do not follow a multivariate normal distribution. This violation of the normality assumption prevents the direct application of standard Gaussian graphical models for estimating the conditional dependence structure via the precision matrix $\Theta^{(k)}$. To address this challenge, we employ the theory of copulas.

Definition 4.2.1 (Copula) *An n-dimensional copula is a function* $C : [0,1]^n \to [0,1]$ *that satisfies:*

1. $\forall \underline{u} \in [0,1]^n \; C(\underline{u}) = 0$ *if* $\exists i$ *such that* $u_i = 0$*;*
2. $C(\underline{u}) = u_i$ *if* $\forall k \neq i \; u_k = 1$*;*
3. C *is n-non-decreasing: i.e for each* $B = \prod_{i=1}^{n} [x_i, y_i] \subset [0,1]^n$ *the* C*-volume of* B *is non-negative.*

A central result in Copula theory is Sklar's theorem, which establishes the link between a multivariate distribution and its univariate marginals.

Theorem 4.2.1 (Sklar's theorem) *Let* F *be a joint cumulative distribution function with marginals* $G_1, \ldots, G_n$*. Then there exists a copula* C *such that for all* $x_1, \ldots, x_n \in \mathbb{R} : F(x_1, \ldots, x_n) = C(G_1(x_1), \ldots, G_n(x_n))$*. Furthermore, if the marginals* $G_1, \ldots, G_n$ *are continuous, then the copula* C *is unique.*
Conversely, if C *is a copula and* $G_1, \ldots, G_n$ *are c.d.f., then the function* F *defined above is a joint c.d.f. with marginals* $G_1 \ldots, G_n$*.*

Sklar's theorem guarantees that we can separate the behavior of the individual marginals (G_i) from the dependence structure (C). This framework allows us to separate the modeling of the dependence structure from the modeling of the individual marginal distributions.

It is a common choice in Copula Graphical Models to assume that the Gaussian copula describes the dependency structure. It is motivated by the closed-form of its density and because the dependence is entirely parameterized by the correlation matrix Σ.

This assumption is equivalent to stating that our observed data $X_j^{(k)}$ arise from a monotonic transformation of unobserved (latent) variables $Z_j^{(k)}$, where the vector $Z^{(k)}$ follows a multivariate normal distribution, $Z^{(k)} \sim N_p(0, \Sigma^{(k)})$ and $\Sigma^{(k)}$ is the correlation matrix for group k. The formal relationship is:

$$X_j^{(k)} = {F_j^{(k)}}^{-1}(\Phi(Z_j^{(k)})), \tag{4.4}$$

where $\Phi()$ is the standard gaussian c.d.f. and $F_j^{(k)^{-1}}()$ is the quantile function (generalized inverse of the c.d.f.). The joint c.d.f. for group k can thus be written using the Gaussian copula as:

$$F\left(x_1^{(k)}, \ldots, x_p^{(k)}\right) = C\left(F_1^{(k)}(x_1^{(k)}), \ldots, F_p^{(k)}(x_p^{(k)})\right) = \\ = \Phi_{\Sigma^{(k)}}\left(\Phi^{-1}(F_1^{(k)}(x_1^{(k)})), \ldots, \Phi^{-1}(F_p^{(k)}(x_p^{(k)}))\right)$$

where $\Phi_{\Sigma^{(k)}}()$ is a c.d.f. of a multivariate normal distribution.

Usually, the marginal distributions $F_j^{(k)}(\cdot)$ are unknown. In principle, if they were known for all j, we could transform the observed data to the latent Gaussian scale via

$$z_{ij}^{(k)} = \Phi^{-1}\left(F_j^{(k)}(x_{ij}^{(k)})\right)$$

and then estimate the correlation matrix $\Sigma^{(k)}$ (and the corresponding precision matrix $\Theta^{(k)}$) by standard methods. However, specifying and estimating appropriate parametric forms for all the marginals is both complicated and prone to error. Instead, we exploit a key property of the model: the transformation from $Z^{(k)}$ to $X^{(k)}$ is monotone in each coordinate, so it preserves the ordering of the observations.

Since both $\Phi(\cdot)$ and $F_j^{(k)^{-1}}(\cdot)$ are non-decreasing functions, we have, for $i \neq i'$,

$$x_{ij}^{(k)} < x_{i'j}^{(k)} \;\Rightarrow\; z_{ij}^{(k)} < z_{i'j}^{(k)}, \qquad z_{ij}^{(k)} < z_{i'j}^{(k)} \;\Rightarrow\; x_{ij}^{(k)} < x_{i'j}^{(k)}.$$

Hence, the ranks of $X_j^{(k)}$ and $Z_j^{(k)}$ coincide. This means that observing $X_j^{(k)}$ imposes a set of inequality constraints on the corresponding latent vector $Z_j^{(k)}$, but does not identify its exact values.

Formally, for each variable j and group k, the latent vector $z_j^{(k)}$ must lie in the set:

$$z_j^{(k)} \in D(x_j^{(k)}) = \left\{z_j^{(k)} \in \mathbb{R}^{n_k} : L_{ij}(x_{ij}^{(k)}) < z_{ij}^{(k)} < U_{ij}(x_{ij}^{(k)}),\ i = 1, \ldots, n_k\right\}, \quad (4.5)$$

where $L_{ij}(x_{ij}^{(k)}) = \max\left\{z_{i'j}^{(k)} : x_{i'j}^{(k)} < x_{ij}^{(k)}\right\}$ is the largest latent value among those corresponding to observations strictly smaller than $x_{ij}^{(k)}$, and
$U_{ij}(x_{ij}^{(k)}) = \min\left\{z_{i'j}^{(k)} : x_{ij}^{(k)} < x_{i'j}^{(k)}\right\}$ is the smallest latent value among those corresponding to observations strictly larger than $x_{ij}^{(k)}$. Collecting all variables and groups, we define:

$$D(x) = \left\{z \in \mathbb{R}^{\left(\sum_{k=1}^{K} n_k\right) \times p} : z_j^{(k)} \in D(x_j^{(k)})\ \ \forall j, k\right\}.$$

Thus, observing $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(K)})^\top$ is equivalent, under the Gaussian Copula model, to observing that the latent vector $\boldsymbol{Z} = (Z^{(1)}, \ldots, Z^{(K)})^\top$ lies in the region $D(\boldsymbol{X})$, $(\boldsymbol{Z} \in D(\boldsymbol{X}))$, which encodes only the rank information.

Let $\boldsymbol{F} = \{F_1^{(1)}, \ldots, F_p^{(1)}, \ldots, F_1^{(K)}, \ldots, F_p^{(K)}\}$ denote the collection of marginal distributions and $\boldsymbol{\Theta} = \{\Theta^{(1)}, \ldots, \Theta^{(K)}\}$ the collection of precision matrices. The probability of the observed data can be written as

$$\mathbb{P}(\boldsymbol{X} \mid \boldsymbol{\Theta}, \boldsymbol{F}) = \mathbb{P}\Big(\boldsymbol{Z} \in D(\boldsymbol{X}) \mid \boldsymbol{\Theta}, \boldsymbol{F}\Big)\, \mathbb{P}\Big(\boldsymbol{X} \mid D(\boldsymbol{X}), \boldsymbol{\Theta}, \boldsymbol{F}\Big).$$

The first factor,

$$\mathbb{P}\Big(\boldsymbol{Z} \in D(\boldsymbol{X}) \mid \boldsymbol{\Theta}, \boldsymbol{F}\Big),$$

is the probability that the latent Gaussian vector $\boldsymbol{Z}$ satisfies the rank constraints implied by $\boldsymbol{X}$. Because $\boldsymbol{Z}$ follows a multivariate normal distribution determined solely by the copula parameters (the correlation matrices and hence $\boldsymbol{\Theta}$), and because $D(\boldsymbol{X})$ depends only on the ordering of the observations (which is invariant to monotone marginal transformations), this probability does not depend on the marginals $\boldsymbol{F}$. We can therefore write it as

$$\mathbb{P}\Big(\boldsymbol{Z} \in D(\boldsymbol{X}) \mid \boldsymbol{\Theta}, \boldsymbol{F}\Big) = \mathbb{P}\Big(\boldsymbol{Z} \in D(\boldsymbol{X}) \mid \boldsymbol{\Theta}\Big),$$

which yields the Extended Rank Likelihood:

$$\mathbb{P}\Big(\boldsymbol{Z} \in D(\boldsymbol{X}) \mid \boldsymbol{\Theta}\Big) = \int_{D(\boldsymbol{X})} \mathbb{P}(\boldsymbol{Z} \mid \boldsymbol{\Theta})\, d\boldsymbol{Z}. \tag{4.6}$$

This likelihood depends only on the association parameters $\boldsymbol{\Theta}$ (through the Gaussian copula) and is free of the marginal distributions $\boldsymbol{F}$. It thus provides a convenient semiparametric framework to estimate the conditional dependence structure for mixed (continuous and discrete) data without specifying or estimating the marginals.

4.3 Model Estimation via the EM (Expectation-Maximization) Algorithm

Maximizing the Extended Rank Likelihood (4.6) is analytically intractable because its integral form lacks a closed-form expression. Consequently, standard optimization that involves taking derivatives of the log-likelihood is not directly applicable. This is because the process reveals a circular dependency: estimating the model parameters ($\boldsymbol{\Theta}$) requires knowing the values of the latent variables ($\mathbf{Z}$), while the latent variables can only be estimated if the parameters are known. This interdependence makes a direct analytical solution impossible and necessitates an iterative approach.
The Expectation-Maximization (EM) algorithm is commonly used for this purpose, as it can numerically solve these two sets of equations by alternating between estimating the latent variables (E-step) and updating the parameters (M-step) until convergence.

4.3.1 The E-Step: Computing the Expected Log-Likelihood

The aim of the E-step is to compute the expected sufficient statistics of the latent variables Z, given the observed data and the current parameter estimates $\hat{\boldsymbol{\Theta}}^{(m)}$. These expectations capture the information that the unobserved latent variables would contribute to the complete data log-likelihood. To express this formally, we introduce the Q function, defined as the expected value of the complete data log-likelihood conditional on the observed data x (and therefore the set of rank constraints D) and on the current parameter estimates $\hat{\boldsymbol{\Theta}}^{(m)}$:

$$Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(m)}) = \mathbb{E}\left[\log\left[L(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{X})\right]|x,\hat{\boldsymbol{\Theta}}^{(m)},D\right], \tag{4.7}$$

where

$$L(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{X}) = \prod_{k=1}^{K}\prod_{i=1}^{n_k}\phi_p(z_i^{(k)}|\boldsymbol{\Theta}^{(k)})\mathbb{1}_{(z_i^{(k)}\in D(x_i^{(k)}))}$$

is the complete data likelihood.
The complete data log-likelihood is derived from the multivariate normal probability density function ϕ_p. For a single observation $z_i^{(k)}$ we get:

$$\log(\phi_p(z_i^{(k)}|\boldsymbol{\Theta}^{(k)})) = -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log(\det(\boldsymbol{\Theta}^{(k)})) - \frac{1}{2}(\boldsymbol{z}_i^{(k)})^T\boldsymbol{\Theta}^{(k)}\boldsymbol{z}_i^{(k)}. \tag{4.8}$$

Summing over all observations in all groups, the complete-data log-likelihood is:

$$\log(L(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{X})) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\frac{1}{2}\log(\det(\boldsymbol{\Theta}^{(k)})) - \frac{1}{2}(\boldsymbol{z}_i^{(k)})^T\boldsymbol{\Theta}^{(k)}\boldsymbol{z}_i^{(k)}\right) + C \tag{4.9}$$

where C is a constant that does not depend on $\boldsymbol{\Theta}$. Taking the expectation of this expression to get the Q function yields:

$$\begin{aligned} &Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(m)}) = \\ &\quad= \frac{1}{2}\sum_{k=1}^{K} n_k \log\left(\det(\boldsymbol{\Theta}^{(k)})\right) - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\mathbb{E}\left[(z_i^{(k)})^\top\hat{\boldsymbol{\Theta}}^{(k)}z_i^{(k)}|x_i^{(k)},\hat{\boldsymbol{\Theta}}^{(m)},D\right] + C \end{aligned} \tag{4.10}$$

Using the cyclic property of the trace, where $v^\top Mv = \mathrm{Tr}(Mvv^\top)$, we can rewrite the expectation term:

$$\mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\Theta}^{(k)}\boldsymbol{z}_i^{(k)}(\boldsymbol{z}_i^{(k)})^T\right)|x_i^{(k)},\hat{\boldsymbol{\Theta}}^{(m)},D\right] = \mathrm{Tr}\left(\boldsymbol{\Theta}^{(k)}\mathbb{E}\left[\boldsymbol{z}_i^{(k)}(\boldsymbol{z}_i^{(k)})^T|x_i^{(k)},\hat{\boldsymbol{\Theta}}^{(m)},D\right]\right)$$

Substituting this back into the Q function gives:

$$\begin{aligned} &Q(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}^{(m)}) = \\ &\quad= \frac{1}{2}\sum_{k=1}^{K}\left[n_k\log\left(\det(\hat{\boldsymbol{\Theta}}^{(k)})\right) - \mathrm{Tr}\left(\hat{\boldsymbol{\Theta}}^{(k)}\sum_{i=1}^{n_k}\mathbb{E}\left[z_i^{(k)}(z_i^{(k)})^\top|x_i^{(k)},\hat{\boldsymbol{\Theta}}^{(m)},D\right]\right)\right] + C \end{aligned} \tag{4.11}$$

By defining the estimated correlation matrix as:

$$\overline{R}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbb{E}\left[Z_i^{(k)} Z_i^{(k)\top} | x_i^{(k)}, \hat{\mathbf{\Theta}}^{(m)}, D \right] \tag{4.12}$$

we arrive at the final expression for the Q function, after dropping the constants:

$$Q(\mathbf{\Theta} | \hat{\mathbf{\Theta}}^{(m)}) \propto \frac{1}{2} \sum_{k=1}^{K} n_k \left[\log\left(\det(\hat{\mathbf{\Theta}}^{(k)}) \right) - \mathrm{Tr}\left(\hat{\mathbf{\Theta}}^{(k)} \overline{R}^{(k)} \right) \right]. \tag{4.13}$$

The calculation of $\overline{R}^{(k)}$ is central to the E-step. Since the true values of the latent variables Z are unknown, their direct contribution to the model's log-likelihood cannot be computed. Instead of assigning a single value to each latent variable, the E-step computes their expected contribution by averaging the sufficient statistics of the complete data over the posterior distribution of Z. The resulting matrix, $\overline{R}^{(k)}$, therefore serves as the probabilistically-weighted input required for updating the parameters in the subsequent M-step.

4.3.2 E-Step Implementation: Gibbs Sampling

Direct computation of $\mathbb{E}\left[Z_i^{(k)} Z_i^{(k)\top} | x_i^{(k)}, \hat{\mathbf{\Theta}}^{(m)}, D \right]$, although possible, is computationally expensive. To overcome this, a Gibbs sampler is employed to approximate this value. The key insight is that the observed data X imposes constraints on the possible values of the latent data Z, as seen in 4.5. Hence for each observation $x_i^{(k)}$, the corresponding latent vector $z_i^{(k)}$ is not drawn from a standard multivariate normal distribution, but from a truncated multivariate normal distribution (TN). The Gibbs sampling procedure proceeds as follows for each group k:

- For each observation $i \in \{1, \dots n_k\}$, lower and upper truncation bounds are determined for the latent vector $z_i^{(k)}$ based on the observed data $x_i^{(k)}$;
- N samples of the latent vector are drawn from a truncated multivariate normal distribution, $TN(0, \Sigma^{(k)}, \text{lower bounds}, \text{upper bounds})$, where $\Sigma^{(k)}$ is the covariance matrix from the previous M-step;
- The required expectation $\mathbb{E}\left[Z_i^{(k)} (Z_i^{(k)})^T | x_i^{(k)}, \hat{\mathbf{\Theta}}^{(m)}, D \right]$ is then approximated by taking the sample mean of the outer products of these N simulated vectors;
- The individual estimates are averaged across all n_k observations in the group to compute the estimated correlation matrix $\overline{R}^{(k)}$.

4.3.3 The M-Step: Penalized Maximization

The M-step updates the parameter estimates by maximizing the Q function derived in the E-step. This maximization is not straightforward for two main reasons, which are addressed by introducing penalty terms controlled by tuning parameters λ_1 and λ_2.

In graphical modeling, we are often interested in the zero elements of the precision matrix Θ, as these correspond to pairs of variables that are conditionally independent. However, standard maximum likelihood estimation generally produces a dense matrix where no elements are exactly zero, making it difficult to identify these relationships. To encourage a sparse solution where irrelevant connections are set to zero, an l_1 Lasso penalty is applied to the off-diagonal elements of the precision matrices. This penalty is controlled by the tuning parameter λ_1. On the other hand, our dataset is composed of observations from distinct but related classes. Therefore, it is reasonable to expect that their underlying network structures will share some similarities. Estimating each graph independently would fail to leverage this shared information. To borrow strength across the classes, a second fused penalty is introduced. This penalty, controlled by the tuning parameter λ_2, encourages similarity by penalizing the differences between corresponding elements in each pair of precision matrices.
Combining the Q function with these two penalties, the M-step consists of solving the following optimization problem:

$$\hat{\boldsymbol{\Theta}}^{(m+1)}_{\lambda_1,\lambda_2} = \arg\max_{\boldsymbol{\Theta}} \left\{ \frac{1}{2} \sum_{k=1}^{K} n_k \Big[\log\det(\hat{\boldsymbol{\Theta}}^{(k)}) - \mathrm{Tr}\Big(\hat{\boldsymbol{\Theta}}^{(k)} \overline{R}^{(k)}\Big) \Big] \right. \\ \left. - \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta^{(k)}_{ij}| - \lambda_2 \sum_{k<k'} \sum_{i,j} |\theta^{(k)}_{i,j} - \theta^{(k')}_{i,j}| \right\} \quad (4.14)$$

This problem is solved using the Alternating Directions Method of Multipliers (ADMM). The ADMM [21] algorithm works by introducing a set of auxiliary variables with the constraint that $\Theta^{(k)} = A^{(k)}$. This allows the problem 4.14 to be split into iterative steps:

- Update Θ: The precision matrices are updated by solving a problem related to the log-likelihood, which has a direct solution involving an eigendecomposition;
- Update A: The auxiliary variables are updated by applying the Lasso and fused penalties, which is computationally fast;
- Update Dual Variables: dual variables are updated to help enforce the constraint that Θ and A converge to the same solution.

These steps are repeated until the algorithm converges to the optimal parameter estimates for the current M-step.

Note that this algorithm is guaranteed to converge to a local maximum, but not necessarily the global maximum.

4.4 Model Selection

To determine the optimal pair of (λ_1, λ_2), a data-driven approach was implemented. A two-dimensional grid of candidate values was defined for both parameters, with λ_1 and λ_2 each ranging from 0 to 1 in increments of 0.1. The model was fitted for every possible (λ_1, λ_2) combination within this grid, generating a collection of 121 candidate models.

The final model was chosen by identifying the parameter combination that minimizes the Extended Bayesian Information Criterion (EBIC), a model selection criterion particularly well-suited for graph identification in high-dimensional settings. The EBIC balances model fit against model complexity, penalizing dense graphs. The EBIC for heterogeneous data is defined as:

$$\begin{aligned}\mathrm{EBIC}(\lambda_1, \lambda_2) = \sum_{k=1}^{K} \Big[n_k \operatorname{Tr}(S^{(k)} \hat{\Theta}^{(k)}_{\lambda_1,\lambda_2}) - n_k \log(\det(\hat{\Theta}^{(k)}_{\lambda_1,\lambda_2})) \\ + \log(n_k) v^{(k)}_{\lambda_1,\lambda_2} + 4\gamma \log(p) v^{(k)}_{\lambda_1,\lambda_2} \Big] \quad (4.15)\end{aligned}$$

where $v^{(k)}_{\lambda_1,\lambda_2}$ represents the number of non-zero edges in the k-th graph and $\gamma = 0.5$ is a hyperparameter penalizing model complexity.

This procedure ensures that the selected model is not only well-fitted to the data but is also suitably sparse and interpretable, avoiding overfitting.

Algorithm 1 HeteroMixGM [22] Estimation via EM and Gibbs Sampling

Input:
Observed data $\boldsymbol{X} = \{X^{(1)}, \ldots, X^{(K)}\}$,
Penalties λ_1, λ_2, ▷ Evaluated on a grid
Threshold ϵ. ▷ 0.001
Output:
Estimated precision matrices $\hat{\boldsymbol{\Theta}} = \{\hat{\Theta}^{(1)}, \ldots, \hat{\Theta}^{(K)}\}$.

function HeteroMixGM($X, \lambda_1, \lambda_2, \epsilon$)
  Pre-computation: For all groups and variables estimate empirical marginal CDFs $\hat{F}_j^{(k)}(x) = \frac{1}{n_k+1} \sum_{i=1}^{n_k} \mathbb{1}\left(X_{ij}^{(k)} \leq x\right)$.
  Initialize: $m \leftarrow 0$,
  $\hat{\Theta}^{(k,0)} \leftarrow I_p$ for all $k \in \{1, \ldots, K\}$.
  repeat
    $m \leftarrow m + 1$.
    E-Step:
    for $k \leftarrow 1$ to K **do**
      $\Sigma^{(k)} \leftarrow (\hat{\Theta}^{(k,m-1)})^{-1}$.
      Compute $L_i^{(k)}, U_i^{(k)}$ for each observation i from $X^{(k)}$ and $\hat{F}^{(k)}$.
      Draw samples $Z_{i,s}^{(k)} \sim TN_p(0, \Sigma^{(k)}, L_i^{(k)}, U_i^{(k)})$ for $s \in \{1, \ldots, N\}$.
      $\tilde{R}^{(k)} \leftarrow \frac{1}{n_k N} \sum_{i=1}^{n_k} \sum_{s=1}^{N} Z_{i,s}^{(k)} (Z_{i,s}^{(k)})^\top$.
    end for
    M-Step:
    Update $\hat{\Theta}^{(m)}$ by solving the Fused Graphical Lasso objective:

$$\hat{\Theta}^{(m)} \leftarrow \arg\max_{\Theta} \left\{ \sum_{k=1}^{K} \frac{n_k}{2} \left[\log(\det(\Theta^{(k)})) - \operatorname{tr}(\Theta^{(k)} \tilde{R}^{(k)}) \right] - \lambda_1 \sum_{k=1}^{K} ||\Theta^{(k)}||_1 - \lambda_2 \sum_{k<k'} ||\Theta^{(k)} - \Theta^{(k')}||_1 \right\}.$$

  until $||\hat{\Theta}^{(m)} - \hat{\Theta}^{(m-1)}||_F < \epsilon$
  return $\hat{\Theta}^{(m)}$
end function

Chapter 5

Results and Discussion

This chapter presents and interprets the results obtained from the statistical analyses applied to the MALDI-MSI proteomic dataset of thyroid lesions. The aim is to evaluate how the proposed models in Chapters 3 and 4 capture diagnostic information and to assess the meaning of the extracted molecular patterns.

5.1 Elastic Net Results

The Elastic Net model was achieved by firstly splitting the data into training and test subsets using a stratified 75/25 partition to preserve the class proportions. Within the training set, the preprocessing pipeline removed zero-variance predictors, applied a $\log(x + 1)$ transformation to stabilize variances and standardized all features. These same transformations, estimated from the training set, were later applied to the test data to ensure consistent scaling.

To address the substantial imbalance across diagnostic groups, SMOTE (Synthetic Minority Oversampling Technique) was incorporated before model fitting.

SMOTE generates additional minority-class samples by interpolating between neighboring observations: for a minority sample x_i and one of its k nearest neighbours x_j, a synthetic point is created as $x_{new} = x_i + \delta(x_j - x_i)$, $\delta \sim \mathcal{U}(0,1)$. This enlarges the minority regions of feature space and prevents the classifier from being dominated by the majority classes, while avoiding simple duplication of observations.

After preprocessing, the Elastic Net model was tuned within the training set using five-fold stratified cross-validation. The regularization parameters λ and α were selected by maximizing the macro-averaged ROC AUC. The model was then refitted on the full (SMOTE-balanced) training set using the optimal parameter values and finally evaluated on the test set.

The Elastic Net classifier achieved an overall accuracy of 0.765 on the test set,

with a balanced accuracy of 0.801. The macro-averaged ROC AUC reached 0.959, reflecting strong separation of classes in terms of predicted probabilities even when some categorical assignments were incorrect.

		Truth				
		FA	**FVPTC**	**HA**	**NIFTP**	**PTC**
Prediction	**FA**	8	1	1	0	0
	FVPTC	0	2	0	0	7
	HA	3	0	3	0	3
	NIFTP	1	0	0	39	0
	PTC	0	3	0	1	13

Table 5.1: Confusion Matrix for Elastic Net model.

The confusion matrix in Table 5.1 summarizes the model's performance across the five diagnostic categories. Despite the use of SMOTE to rebalance the training data, the test set still reflects the original class imbalance and this is visible in the distribution of errors. The highest accuracy was observed for NIFTP and PTC, which are also the best-represented classes. NIFTP was identified almost perfectly, with 39 of 40 cases correctly classified and only a single FA sample misassigned to this category. PTC also showed strong performance, with 13 of 17 cases correctly recognized; the remaining samples were mainly predicted as FVPTC.

Performance on benign categories was more heterogeneous. Follicular Adenoma (FA) achieved moderate recognition, with 8 correct predictions and the remaining cases mainly assigned to Hurthle Adenoma (HA) or NIFTP, which is consistent with the known overlap among more benign lesions. Hurthle Adenoma (HA), despite its very small sample size, was handled reasonably well: 3 of 4 cases were correctly identified and the single misclassification occurred within the benign spectrum (HA predicted as FA), which is clinically less problematic than cross-boundary errors. The most problematic group was FVPTC: although SMOTE increased its representation during training, in the test set only 2 of 6 samples were correctly identified, while most were predicted as PTC, an outcome that reflects both its true scarcity and its molecular similarity to classical PTC.

These class-specific limitations illustrate that SMOTE mitigates but does not eliminate the intrinsic difficulty of distinguishing rare or heterogeneous categories. However, if the focus is simplified to the clinically most relevant distinction, benign (FA+HA), NIFTP and malignant (PTC+FVPTC), the model performs substantially better. The resulting three-class confusion matrix yields recalls of 0.94 for benign lesions, 0.98 for NIFTP and 0.86 for malignant disease, corresponding to a balanced accuracy of approximately 92.5%. This indicates that, despite the persistent challenges in resolving the finer subtypes, the broader diagnostic

separation between benign, low risk and malignant lesions is captured effectively.

		Truth		
		Benign	**NIFTP**	**Malignant**
Prediction	**Benign**	15	0	4
	NIFTP	1	39	0
	Malignant	0	1	25

Table 5.2: Aggregated Confusion Matrix for Elastic Net model.

A key advantage of the Elastic Net is its embedded feature selection, which shrinks uninformative coefficients to zero. In this analysis, the model retained a moderate subset of the seventy molecular predictors for each class, with 24 non-zero coefficients for FA, 32 for FVPTC, 32 for Hurthle Adenoma, 31 for NIFTP and 30 for PTC. Thus, all classes share a partially overlapping molecular basis, but with distinct patterns of coefficient signs and amplitudes. The corresponding coefficient patterns are shown in Figure 5.1, where each column corresponds to a diagnostic class and each row to a molecule; red and blue indicate positive and negative coefficients, respectively. Several of the largest-magnitude coefficients (dark red or blue) appear in the Hurthle Adenoma and NIFTP columns, while PTC and FA display predominantly moderate intensities and FVPTC is visually the least intense. Across many molecules the signs are aligned across classes, indicating shared trends in expression, whereas a smaller subset of features shows clear sign reversals, which are likely to drive the main discriminative boundaries between benign, borderline and malignant groups.

These coefficient patterns should be interpreted in light of the class-specific SMOTE upsampling applied during training. Because NIFTP was not oversampled and PTC required only minimal augmentation, their coefficients primarily reflect the structure of the original data. In contrast, FVPTC and Hurthle Adenoma were heavily upsampled, so many of their coefficients were estimated from a mixture of real and synthetic observations. Since SMOTE generates interpolated samples and reduces within-class variability, the Elastic Net may retain a larger number of moderate coefficients for these classes rather than a few dominant ones. As a result, the magnitude and distribution of the class-specific coefficients reflect not only the intrinsic heterogeneity of each group but also the varying degree of artificial balancing introduced during training; the heatmap should therefore be interpreted with this smoothing effect in mind.

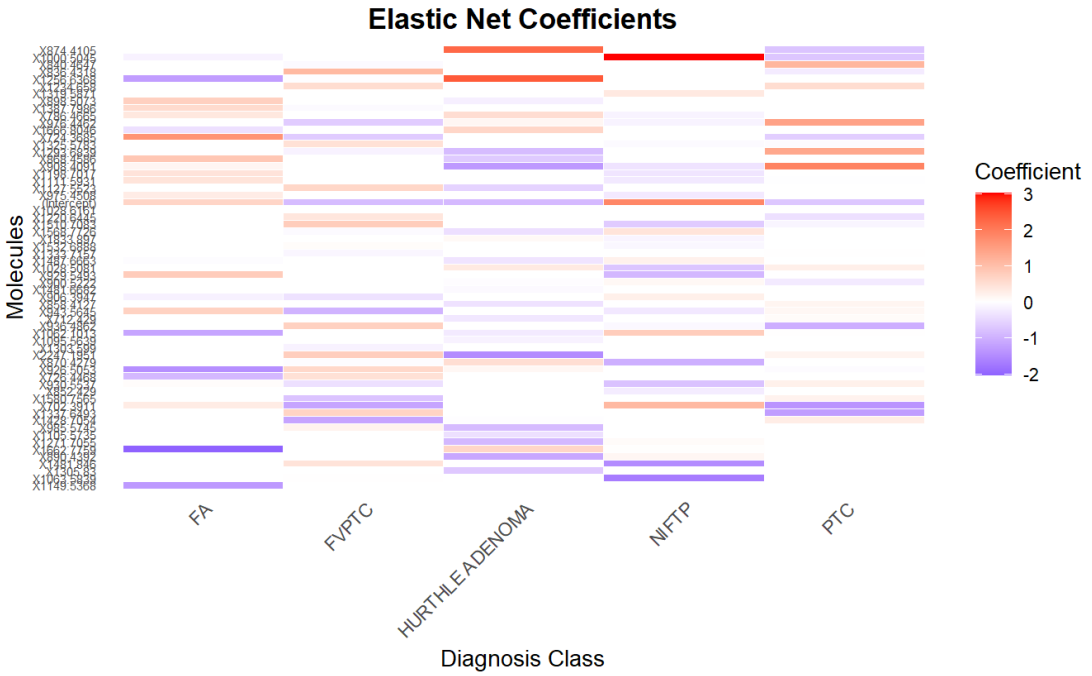


Figure 5.1: Heatmap of non-zero Elastic Net coefficients across diagnostic classes.

5.2 sPLS-DA Results

For the sparse Partial Least Squares Discriminant Analysis (sPLS-DA), the same data partitioning strategy used for the Elastic Net was adopted: the dataset was split into a training set (75%) and a test set (25%) using a stratified split to preserve the original class proportions. Within the training set, all predictors were transformed using a $\log(x+1)$ transformation to reduce skewness and stabilize the variance and were then standardized to zero mean and unit variance. The corresponding transformation parameters, estimated solely from the training data, were subsequently applied to the test set to ensure consistent scaling.

Class imbalance was handled analogously to the previous case by incorporating SMOTE (Synthetic Minority Oversampling Technique) within the preprocessing pipeline, restricted to the training data only. The test set was not oversampled and underwent only the deterministic log and scaling transformations.

Model tuning was performed through a grid search on the training set, exploring up to four latent components together with a wide range of *keepX* values. Repeated 5-fold cross-validation was used, with the Balanced Error Rate (BER) as the selection criterion. The optimal model consisted of four components, with *keepX* values equal to $(3, 5, 16, 5)$ for components 1 to 4. The corresponding mean cross-validated BER values were 0.621, 0.454, 0.386 and 0.360, showing a clear

improvement in class discrimination as additional components were included.

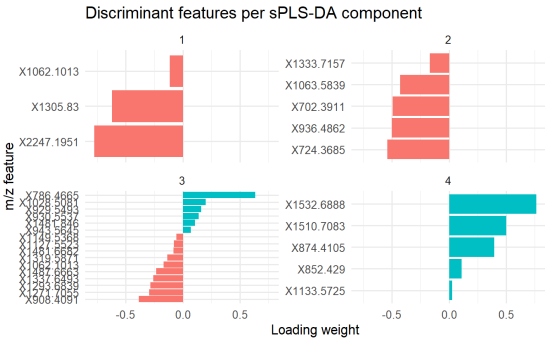


Figure 5.2: Molecular Features chosen for each latent component.

The final sPLS-DA model was refitted on the full SMOTE-balanced training set using these optimal hyperparameters and subsequently evaluated on the held-out test set. Class predictions were obtained using the maximum-distance decision rule, which assigns each observation to the class associated with the largest value in its predicted dummy-response vector. The classifier achieved a balanced accuracy of 0.817 on the test data.

		Truth				
		FA	**FVPTC**	**HA**	**NIFTP**	**PTC**
Prediction	**FA**	4	0	0	5	0
	FVPTC	1	4	0	1	5
	HA	6	0	4	1	1
	NIFTP	1	0	0	31	0
	PTC	0	2	0	2	17

Table 5.3: Confusion Matrix for the sPLS-DA model.

The confusion matrix in Table 5.3 provides a detailed view of the predictive behaviour of the sPLS-DA model across the five diagnostic categories. Compared with the Elastic Net, the most evident deterioration concerns NIFTP: only 31 cases were correctly recognised, while several were reassigned to FA, HA, or PTC. This indicates that the latent components extracted by sPLS-DA do not isolate this low risk category as cleanly as the Elastic Net solution.

Misclassification among malignant categories is broadly similar to the Elastic Net: PTC and FVPTC are often confused, with PTC samples predicted as FVPTC and vice versa, reflecting their biological proximity.

Benign categories remain difficult to separate. FA is again the weakest group, with only 4 correctly classified cases and many predictions drifting towards HA or

NIFTP. Hurthle Adenoma shows slightly better recognition (4 correct cases) but still receives misassignments from FA and even one PTC sample.

Overall, sPLS-DA captures the broad diagnostic structure but shows greater ambiguity across class boundaries, especially for NIFTP and FA and a higher degree of confusion between PTC and FVPTC. These patterns confirm that the Elastic Net classifier provides more reliable and precise discrimination, particularly for underrepresented or borderline categories.

To assess the classifier at a more clinically meaningful level, the five categories were again collapsed into three groups: benign (FA + HA), low risk (NIFTP) and malignant (PTC + FVPTC). The resulting confusion matrix is reported in Table 5.4.

	Truth		
Prediction	**Benign**	**NIFTP**	**Malignant**
Benign	14	6	1
NIFTP	1	31	0
Malignant	1	3	28

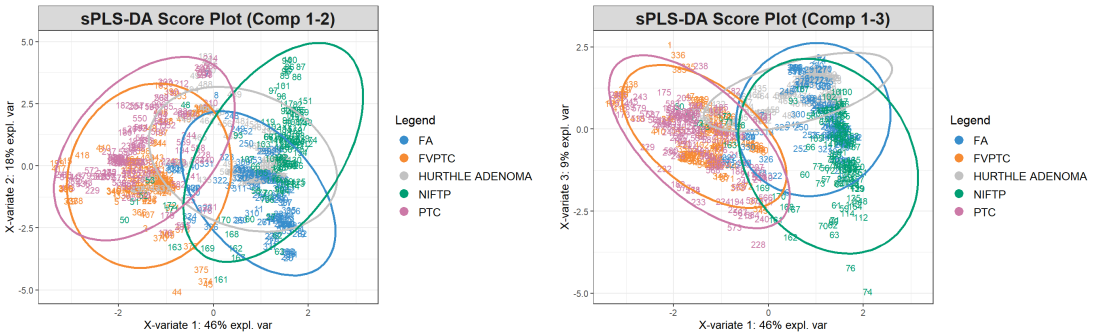
Table 5.4: Aggregated Confusion Matrix for the sPLS-DA model.

Aggregation simplifies the task but does not remove the underlying difficulties. Benign lesions are only moderately well recovered, with several NIFTP assigned to benign samples. NIFTP appears more accurate in this coarser setting, yet this partly reflects the merging of its main competitors (FA, HA, FVPTC) into broader groups rather than a genuinely sharper separation in the latent space. Malignant lesions are classified more reliably, but the strong PTC-FVPTC overlap is now absorbed into a single malignant label.

The three-group recalls correspond to a balanced accuracy of 0.87. While this indicates reasonable performance, it remains clearly below the value obtained in the same setting with the Elastic Net (about 92.5%), confirming that sPLS-DA is less effective for this dataset, especially for borderline and heterogeneous categories such as NIFTP.

The sPLS-DA score plots in Figures 5.3a and 5.3b show the distribution of samples along the first three latent components, together with class-wise confidence ellipses.

The first component explains a substantial proportion of the variance (46%), while the second and third components account for 18% and 9%, respectively. Nevertheless, no projection yields a clear separation of all five classes. In both the (Comp 1–2) and (Comp 1–3) planes, PTC and FVPTC occupy largely overlapping regions, consistent with the strong mutual confusion observed in the confusion



(a) sPLS-DA score plot for Components 1 and 2

(b) sPLS-DA score plot for Components 1 and 3

Figure 5.3: sPLS-DA score plot.

matrices. NIFTP, FA and HA also show extensive overlap: the NIFTP cluster intrudes into the benign region and FA points are widely scattered between HA and NIFTP. These plots therefore visually confirm the numerical results, indicating that the latent space learned by sPLS-DA captures some global structure but does not provide sharply distinct clusters for the diagnostically challenging categories, in line with its lower performance compared with the Elastic Net model.

5.3 Graphical Model Results

The Copula Graphical Model for heterogeneous Mixed Data was implemented following the algorithmic outline (1) described in Chapter 4.

The input consisted of the raw intensity values from the MALDI-MSI dataset, grouped into the five diagnostic categories described in Chapter 2. No preprocessing was applied prior to model fitting, as the Copula Graphical Model estimates dependence structures through rank-based likelihood and therefore accepts all types of data.

The grid search for penalty parameters yielded $\lambda_1 = 0.1$ and $\lambda_2 = 0$, thus the final solution promotes moderate within-class sparsity while not imposing cross-class similarity. The EBIC preference for $\lambda_2 = 0$ indicates that enforcing a global similarity constraint across all diagnostic classes did not improve the model's fit–complexity trade-off, which is consistent with the presence of heterogeneous dependence structures among the five groups. Nevertheless, this result does not preclude the existence of partial similarities between certain diagnostic categories, rather, it indicates that a single, fully fused structure would oversimplify the underlying biological variability.

Each of the resulting five graphs comprises 70 nodes (molecules), with edges

representing non-zero partial correlations derived from the precision matrices $\Theta^{(k)}$. The resulting edge counts and densities were:

- PTC: 464, 19.2%;
- FVPTC: 458, 19.0%;
- NIFTP: 505, 20.9%;
- Hurthle Adenoma: 453, 18.8%;
- FA: 485, 20.1%;

Thus, all five networks are moderately sparse, with NIFTP being the most connected and Hurthle Adenoma and FVPTC slightly less connected.

Pairwise comparisons confirm substantial but incomplete overlap in network structure. Numbers of shared edges between diagnostic pairs are reported in Table 5.5.

Comparison	**Shared**	**Only In G1**	**Only In G2**
FA vs FVPTC	231	256	224
FA vs Hurthle Adenoma	234	253	221
FA vs NIFTP	264	223	241
FA vs PTC	269	218	193
FVPTC vs Hurthle Adenoma	180	275	275
FVPTC vs NIFTP	261	194	244
FVPTC vs PTC	283	172	179
Hurthle Adenoma vs NIFTP	199	256	306
Hurthle Adenoma vs PTC	205	250	257
NIFTP vs PTC	280	225	182

Table 5.5: Summary of Pairwise Graph Edge Comparisons

The largest overlaps occurred for FVPTC vs PTC (283 shared) and NIFTP vs PTC (280 shared), followed by FA vs NIFTP (264 shared) and FVPTC vs NIFTP (261 shared). This could be interpreted as a sign of similarity between the follicular entities (FA, NIFTP and FVPTC) as well as the relationship between PTC and FVPTC (as the latter is a subtype of the former). In contrast, FVPTC vs Hurthle Adenoma (180 shared) and Hurthle Adenoma vs NIFTP (199 shared) showed the lowest overlap. Importantly, every pair retained well over a hundred unique edges, indicating that while subsets of dependencies are conserved, the global network structure remains diagnosis-specific, which is consistent with $\lambda_2 = 0$.

At the node level, centrality measures (Table 5.6) provide a complementary view of how molecular features contribute to network organization. Across the five diagnoses, the mean degree ranged from approximately 14.0 in FVPTC and Hurthle adenoma to 15.4 in NIFTP, confirming that the latter network is slightly denser and more interconnected on average. Mean betweenness centrality values were low and consistent across groups (0.014–0.015), suggesting that few nodes act as critical bridges within these networks. In contrast, the mean closeness centrality showed moderate variation (from 0.496 in PTC to 0.508 in FA) indicating small differences in overall graph compactness.

Diagnosis	**Degree**	**Betweenness**	**Closeness**
FA	14.9	0.0143	0.508
FVPTC	14.0	0.0147	0.501
HURTHLE ADENOMA	14.0	0.0145	0.506
NIFTP	15.4	0.0147	0.501
PTC	14.2	0.0151	0.496

Table 5.6: Summary of Mean Node Metrics per Diagnosis

Taken together, the results depict sparse and diagnosis-specific dependence structures with partial shared substructures across certain pairs. Figure 5.4 shows the five graphs, although, given the number of nodes and edges in each network, direct visual comparison is not informative: the resulting graphs are too dense for meaningful interpretation beyond general topology.

A more informative representation is provided by the heatmaps of the partial correlation matrices (Figure 5.5), which reveal the overall structure and strength of conditional associations. Across all diagnoses, most partial correlations are weak, with only a limited number of strong positive connections (deep red) and few negative ones (blue). Certain structural patterns appear consistently across the five matrices, suggesting the existence of a core set of molecular relationships that are preserved across thyroid lesion types, albeit with varying intensity. In addition, some pairwise similarities emerge, particularly between FVPTC and PTC, whose matrices display comparable block-like regions of positive dependence.

To summarize structural overlap across diagnoses, the binary adjacency matrices were aggregated into a single consensus matrix, where each cell represents the number of diagnostic groups in which a given edge was present (values from 0 to 5). As shown in Figure 5.6, only a small subset of connections appears consistently across all classes, while the majority are present in one or two groups. This pattern quantitatively confirms the presence of a shared but limited core network and extensive diagnosis-specific variability, consistent with the pairwise comparisons reported in Table 5.5.

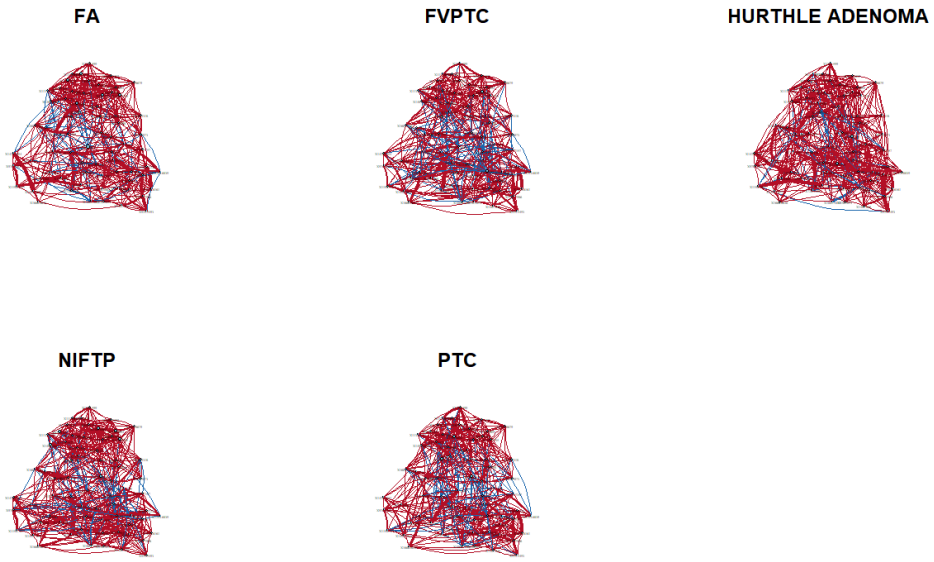


Figure 5.4: Estimated conditional dependence networks for the five diagnostic categories: each node represents a molecular feature and edges correspond to non-zero partial correlations: red edges indicate positive associations and blue edges indicate negative ones. All graphs share a common layout to facilitate visual comparison.

5.3.1 Network Rewiring Analysis

The main goal of this thesis is to investigate whether the graphs constructed through the Copula Graphical Model can provide meaningful insight into which molecular features become activated or deactivated across different types of thyroid lesions. In other words, the aim is to determine whether changes in the estimated conditional dependence structures reflect biologically relevant differences in molecular interactions between diagnostic groups. To achieve this, a dedicated procedure was developed to systematically compare the estimated networks and identify the molecules most involved in such structural variations.

The starting point of the analysis is the set of partial correlation matrices $\boldsymbol{\rho}^{(k)}$, one for each diagnostic group k. Each element $\rho_{ij}^{(k)}$ quantifies the conditional association between molecules i and j within group k, after controlling for the remaining variables. To assess how these conditional relationships differ between diagnoses, a pairwise comparison was performed across all possible combinations of diagnostic groups. For each pair (k, k'), an edge-wise difference matrix $D^{(k,k')}$ was

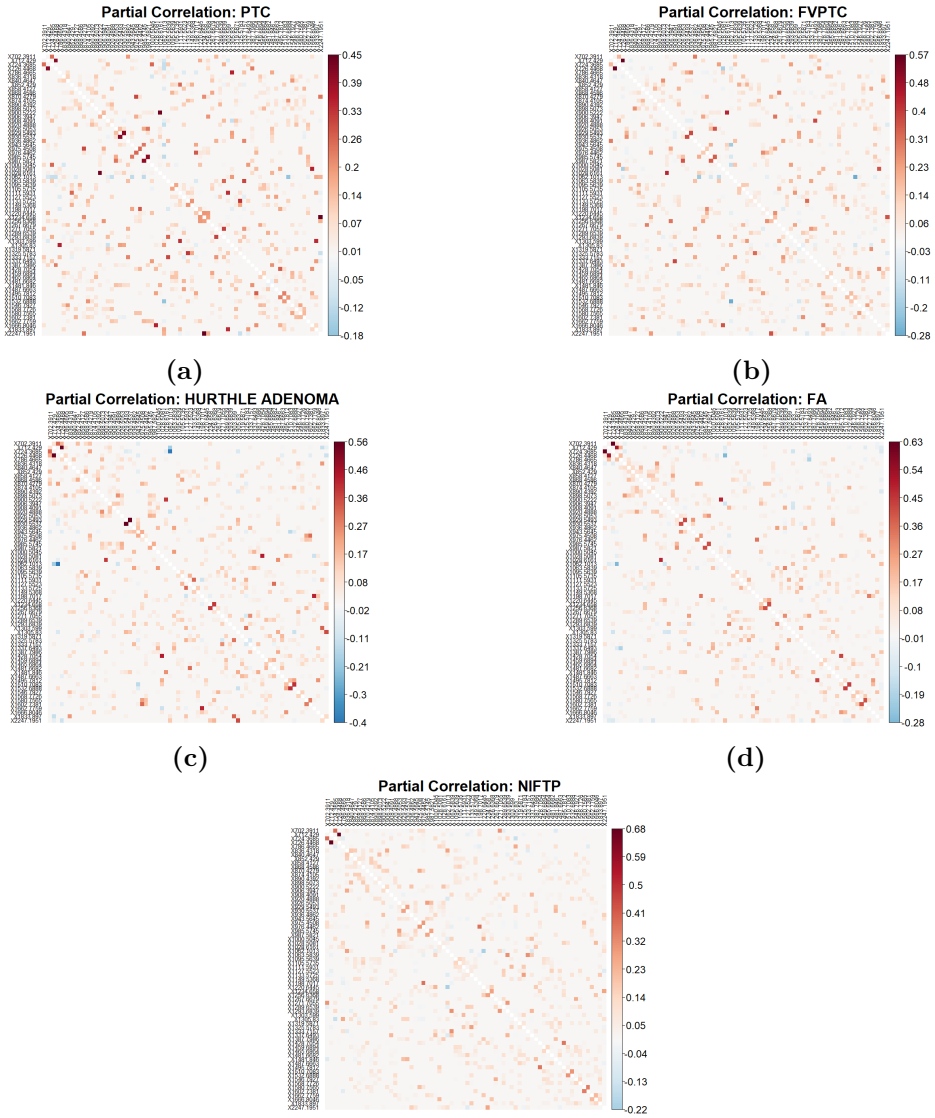


Figure 5.5: Heatmaps of the partial correlation matrices estimated for each diagnostic class. Positive partial correlations are shown in red and negative in blue.

defined as:

$$D_{ij}^{(k,k')} = |\rho_{ij}^{(k)} - \rho_{ij}^{(k')}| \tag{5.1}$$

which measures the magnitude of change in conditional dependence between molecules i and j when moving from diagnosis k to k'. This matrix captures not only the appearance or disappearance of edges but also substantial changes in

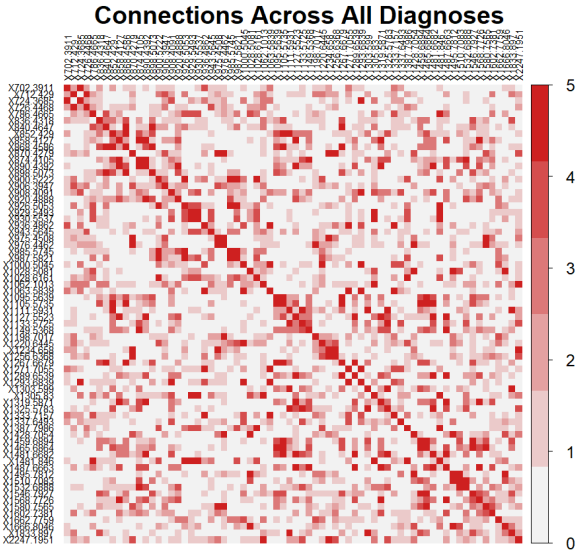


Figure 5.6: Consensus Matrix: derived by summing the adjacency matrices of the 5 graphs.

the strength or direction of associations. In this way, $D^{(k,k')}$ serves as a quantitative representation of network rewiring, describing how the underlying molecular connectivity reorganizes between diagnostic conditions.

Because small numerical fluctuations are expected even in stable relationships, a data-driven threshold was introduced to isolate the most pronounced structural changes. Specifically, for each matrix $D^{(k,k')}$, the distribution of its off-diagonal values was examined and a high quantile q was chosen as the cutoff point.

Denoting by $t_q^{(k,k')}$ the empirical q-th quantile of $\{D_{ij}^{(k,k')} : i < j\}$, edges with $D_{ij}^{(k,k')} > t_q^{(k,k')}$ were classified as rewired edges, corresponding to substantial modifications in conditional dependencies. This adaptive approach allows the cutoff to vary across pairs, accommodating differences in network sparsity and overall correlation variability.

Once the rewired edges were identified, the analysis proceeded to a node-level interpretation. For each molecule i, the number of rewired edges connected to it was counted:

$$n_i^{(k,k')} = \sum_{j \neq i} \mathbb{I}\{D_{ij}^{(k,k')} > t_q^{(k,k')}\}. \tag{5.2}$$

However, molecules differ in their overall connectivity, so highly connected nodes are a priori more likely to accumulate rewired edges. To account for this, a degree-normalized rewiring rate was defined.

First, for each pair (k, k'), a binary adjacency matrix was constructed for each

group by marking nonzero partial correlations,

$$A_{ij}^{(k)} = \mathbb{I}\{\rho_{ij}^{(k)} \neq 0\},\ A_{ij}^{(k')} = \mathbb{I}\{\rho_{ij}^{(k')} \neq 0\}.$$

The degree of molecule i in the union network was then given by:

$$\deg_i^{(k,k')} = \sum_{j \neq i} \mathbb{I}\{A_{ij}^{(k)} = 1 \text{ or } A_{ij}^{(k')} = 1\}$$

i.e. the number of distinct neighbors connected to i in at least one of the two diagnoses.

The degree-normalized rewiring rate for molecule i between diagnoses k and k' was then defined as

$$r_i^{(k,k')} = \frac{n_i^{(k,k')}}{\deg_i^{(k,k')}}. \tag{5.3}$$

This quantity represents the proportion of molecule i's interactions that undergo substantial rewiring when moving from one diagnostic group to the other and it takes values in [0,1].

Molecules were declared significantly rewired if their rewiring rate exceeded a pre-specified threshold $r_i^{(k,k')} \geq \tau$.

A sensitivity analysis was carried out by varying both the edge-level quantile q and the rewiring-rate threshold τ over a grid of values; for each combination (q, τ), the number of significantly rewired molecules was recorded for the diagnostic pair FA vs PTC (5.7).

Altogether, this framework provides an interpretable way to detect and quantify differences in network architecture, highlighting molecular features that may underlie or reflect the biological heterogeneity of thyroid lesions.

5.4 Application Example

To illustrate how the different modeling strategies can be combined in a diagnosis-specific manner, we focused on a pair of groups, namely Follicular Adenoma (FA) versus Papillary Thyroid Carcinoma (PTC), and derived for each method a set of discriminant molecular features.

For the univariate differential analysis underlying the Volcano plots, molecules were deemed significant if they exhibited a sufficiently large effect size and strong statistical evidence, so if they appeared as up or downregulated on the FA vs. PTC Volcano plot.

For the Elastic Net, let $\beta_{k,j}$ denote the coefficient associated with molecular feature j in class k; for the FA vs. PTC comparison we defined a feature-wise

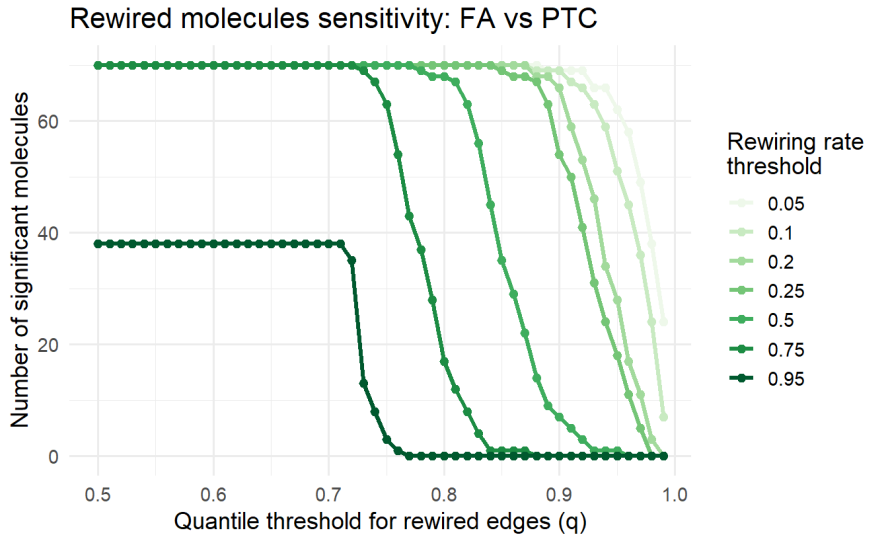


Figure 5.7: Graphical representation of the number of molecules chosen as "rewired" for different edge and rate thresholds.

contrast score $s_j^{(\mathrm{EN})} = \beta_{\mathrm{PTC},j} - \beta_{\mathrm{FA},j}$ and selected as Elastic Net markers those molecules with $|s_j^{(\mathrm{EN})}|$ above an empirical quantile of the non-zero scores.

For the sPLS-DA model a pairwise FA vs. PTC score was derived by exploiting the structure of the latent components. sPLS–DA represents each sample through a sequence of latent components (or score vectors) $T_{\cdot,c}$, where

$$T_{i,c} = \sum_{j=1}^{p} p_{j,c}\, X_{i,j},$$

and $p_{j,c}$ denotes the loading of molecule j on component c. These components are explicitly constructed to maximise discrimination among the diagnostic groups, meaning that observations from different classes exhibit different average scores on each component. For each component c, the mean score of the FA and PTC groups was computed as

$$\mu_{\mathrm{FA},c} = \frac{1}{n_{\mathrm{FA}}} \sum_{i \in \mathrm{FA}} T_{i,c}, \qquad \mu_{\mathrm{PTC},c} = \frac{1}{n_{\mathrm{PTC}}} \sum_{i \in \mathrm{PTC}} T_{i,c},$$

and their difference,

$$\Delta_c = \mu_{\mathrm{PTC},c} - \mu_{\mathrm{FA},c},$$

quantifies how strongly component c separates PTC from FA, including the direction of separation. A molecule contributes to this pairwise discrimination if it has a large

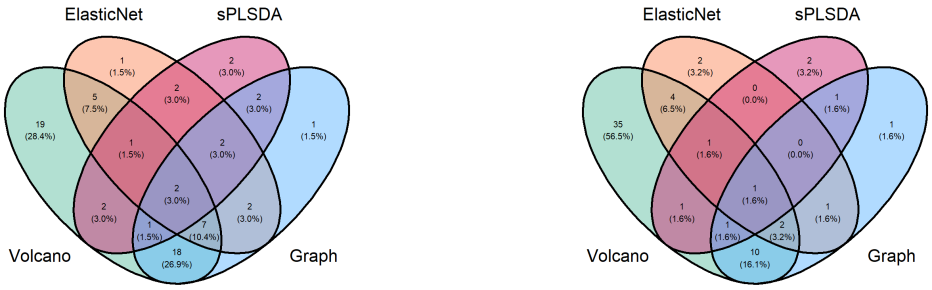
loading on components for which the FA and PTC centroids differ substantially. To formalise this idea, a feature-level contrast score was defined as:

$$s_j^{(\mathrm{sPLS-DA})} = \sum_{c=1}^{4} p_{j,c}\,\Delta_c,$$

where the sum extends over the four latent components selected by the tuning procedure. Large positive values of $s_j^{(\mathrm{sPLS-DA})}$ indicate molecules associated with components shifting towards PTC, whereas large negative values correspond to features more characteristic of FA. Molecules with $|s_j^{(\mathrm{sPLS})}|$ above an empirical quantile were retained as the sPLS–DA markers for the FA vs. PTC comparison.

Finally, for the graphical model, we used the degree-normalized rewiring analysis described above: for each pair of diagnoses (k, k') we constructed the edge-wise difference matrix $D^{(k,k')}$ and defined rewired edges as those with $D_{ij}^{(k,k')}$ above the q-th empirical quantile; for each molecule i we then computed the number of rewired edges $n_i^{(k,k')}$ incident to it, the union degree $\deg_i^{(k,k')}$ across the two networks, and the corresponding rewiring rate $r_i^{(k,k')} = n_i^{(k,k')} / \deg_i^{(k,k')}$, declaring molecule i rewired if $r_i^{(k,k')} \geq \tau$.

The four resulting sets of molecules for the FA vs. PTC contrast (Volcano, Elastic Net, sPLS-DA and graphical rewiring) were then compared by means of a Venn diagram, in order to highlight features consistently selected across multiple approaches as well as method-specific signals that may point to complementary aspects of the underlying biological differences.



(a) Elastic Net quantile: 50, sPLS-DA quantile: 50, edge quantile: 80, node threshold: 0.5.

(b) Elastic Net quantile: 75, sPLS-DA quantile: 75, edge quantile: 85, node threshold: 0.75.

Figure 5.8: Comparison between Venn diagrams for different sets of thresholds

To illustrate the sensitivity of feature selection to user-defined cut-offs, we present two threshold configurations (Figure 5.8). The looser setting (Figure 5.8a) produces broader marker sets for all methods and reveals a richer pattern of overlaps. In contrast, the stricter configuration (Figure 5.8b) yields far fewer retained molecules,

leaving only the strongest and most stable signals. Comparing the two diagrams shows that both the number of selected features and the extent of cross-method agreement depend heavily on the chosen thresholds: when the cut-offs are relaxed, several molecules are jointly selected across methods; when they are tightened, many of these intersections vanish. This demonstrates that apparent concordance between methods is not inherent, but largely driven by the thresholding scheme, which is a crucial consideration in small, high-dimensional biomedical datasets.

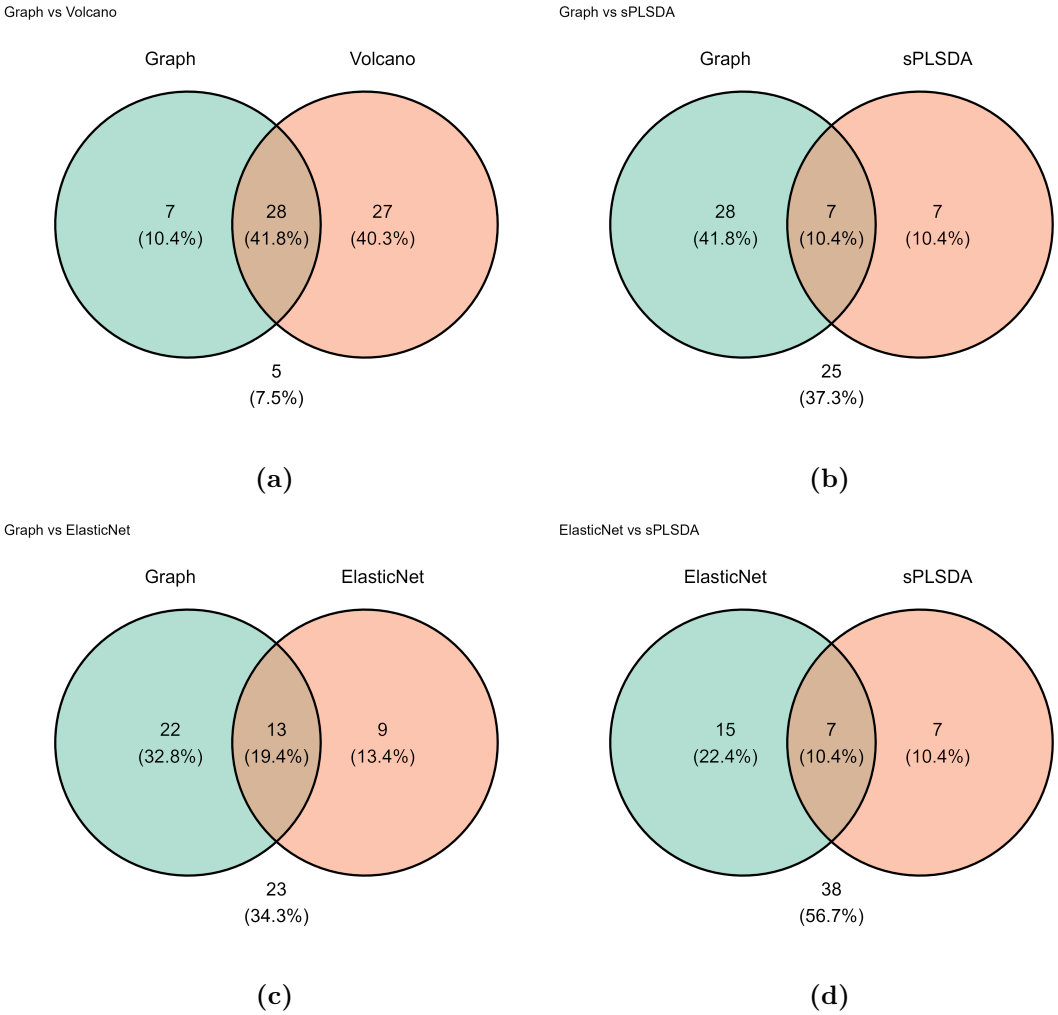


Figure 5.9: Pairwise Venn Diagrams.

Focusing on the looser threshold configuration (Figure 5.9), the intersections among the four methods become substantially larger, allowing a clearer view of the shared and method-specific signals. When comparing the graphical rewiring markers with the three abundance-based approaches (Volcano, Elastic Net and

sPLS-DA), no full overlap emerges. The graphical model shares 13 molecules with Elastic Net, 28 with the Volcano analysis and 7 with sPLS-DA, confirming that each method extracts a different aspect of the FA vs. PTC comparison. This behavior is expected: the graphical model detects changes in conditional dependence structure, whereas Elastic Net and sPLS-DA identify molecules that directly drive class separation in the abundance space. It is worth noting that the intersection between Elastic Net and sPLS-DA, two multivariate sparse classifiers, is also limited (7 shared molecules).

Despite these differences, a small core of recurrent signals emerges under the looser thresholds. In particular, two molecules (X724.3685 and X1062.1013) are consistently selected by all four methods, making them especially strong candidates for robust FA vs. PTC biomarkers. Several additional molecules (including X976.4462, X1256.6368, X1337.6493 and X2247.1951) appear in three out of four methods and therefore represent promising features deserving further biological exploration. Overall, the looser threshold configuration highlights that the four modeling strategies are complementary rather than redundant, with each contributing different insights into the underlying metabolic differences between FA and PTC.

These findings must be interpreted with caution. The dataset is relatively small, as is typical for biomedical metabolomics and methods like Elastic Net and sPLS-DA can be sensitive to sample size, oversampling and regularization choices. Moreover, several thresholds used across the analysis (fold-change cutoffs, empirical quantiles, rewiring-rate criteria) are somewhat arbitrary and influence the size and composition of each feature set. For these reasons, the molecules identified here should be regarded as preliminary statistical candidates. Their biological plausibility must be assessed by domain experts and the most promising markers should be further explored and independently validated in larger cohorts or functional studies.

Chapter 6

Conclusions

This thesis analyzed MALDI-MSI proteomic profiles of thyroid lesions using two complementary statistical perspectives. The aim was to understand how molecular features differ across five diagnostic groups and to investigate whether modeling relationships between molecules provides information that classical classification approaches cannot capture.

The first perspective used supervised classification models: Elastic Net and sPLS-DA. These methods take molecular abundances as input and aim to predict the diagnosis of each sample. Their output is a set of features whose levels differ between classes, either directly (Elastic Net coefficients) or through latent components (sPLS-DA loadings). These models therefore highlight which molecules distinguish one diagnosis from another, focusing on differences in abundance.

The second perspective, based on the Copula Graphical Model, serves a different purpose. Instead of predicting classes, it reconstructs a network of conditional dependencies among molecules within each diagnostic group. Each graph describes how molecules are connected after adjusting for all others, capturing their coordinated behaviour rather than their individual levels. This shift in focus, from differences in abundance to differences in molecular relationships, is an important conceptual change and represents the most innovative contribution of this thesis.

The introduced rewiring analysis provides a practical and interpretable way to work with these networks. For each pair of diagnoses, partial correlation matrices were compared to detect edges whose strength changed substantially. For each molecule, a degree-normalised rewiring rate was then computed to quantify the proportion of its connections that were altered. Molecules with high rewiring rates were identified as graph-driven markers, meaning that they are relevant not because their abundance changes, but because their interaction patterns shift across diagnostic conditions.

This perspective is particularly meaningful in a biological context. Cellular behavior is governed by networks of interacting molecules rather than isolated

entities. Rarely does a single molecule determine a biological state on its own; instead, physiological and pathological processes arise from the balance of many coordinated interactions. When this balance is perturbed, the structure of the network can reorganize, triggering new functional states or disease transitions. The rewiring analysis therefore captures changes that abundance-based models are not meant to detect and offers access to a complementary layer of biological information.

The supervised models and the graphical model produced distinct but partially overlapping sets of relevant molecules. For one comparison (FA vs. PTC) and under one specific choice of thresholds, some molecules were selected by multiple approaches, while others were method-specific. This indicates that abundance-based models and the graphical rewiring approach highlight complementary aspects of molecular differences rather than converging on a single type of signal. The Elastic Net achieved the strongest predictive performance and identified clear discriminative markers, while sPLS-DA offered interpretable latent structures but lower accuracy. The Copula Graphical Model produced diagnosis-specific networks with both shared and distinct structures, and the rewiring analysis revealed molecules whose interaction patterns change most across lesions.

Several limitations must be acknowledged. The sample size is modest relative to the number of features and class imbalance introduces challenges that oversampling and rank-based inference only partially mitigate. Thresholds used in the rewiring analysis are data-dependent, and MALDI-MSI does not directly identify the chemical nature of the molecules, requiring further biological validation.

Future work could expand the dataset, apply stability-selection techniques; the diagnostic structure may also be reconsidered: the five-class framework used is biologically meaningful but statistically challenging given the available sample size. A more clinically aligned three-class approach, distinguishing benign lesions, NIFTP and malignant lesions, could provide a more stable modeling framework and clearer biological interpretations. Overall, the integration of network-level modeling and rewiring analysis represents a promising avenue for extracting biologically meaningful patterns from MALDI-MSI data.

Bibliography

[1] Sjoerd Hermes, Joost van Heerwaarden, and Pariya Behrouzi. «Copula Graphical Models for Heterogeneous Mixed Data». In: *Journal of Computational and Graphical Statistics* 33.3 (2024), pp. 991–1005. DOI: 10.1080/10618600.2023.2289545. eprint: https://doi.org/10.1080/10618600.2023.2289545. URL: https://doi.org/10.1080/10618600.2023.2289545 (cit. on pp. i, 22).

[2] Jose P Zevallos, Christine M Hartman, Jack R Kramer, Erich M Sturgis, and Elizabeth Y Chiao. «Increased thyroid cancer incidence corresponds to increased use of thyroid ultrasound and fine-needle aspiration: a study of the Veterans Affairs health care system». In: *Cancer* 121.5 (2015), pp. 741–746. DOI: 10.1002/cncr.29122 (cit. on p. 1).

[3] Stavroula A Paschou, Andromachi Vryonidou, and Dimitrios G Goulis. «Thyroid nodules: A guide to assessment, treatment and follow-up». In: *Maturitas* 96 (2017), pp. 1–9. DOI: 10.1016/j.maturitas.2016.11.002 (cit. on p. 1).

[4] Massimo Bongiovanni, Achille Spitale, William C Faquin, Luca Mazzucchelli, and Zubair W Baloch. «Comparison of 5-tiered and 6-tiered diagnostic systems for the reporting of thyroid cytopathology». In: *Cancer Cytopathology* 120.2 (2012), pp. 117–125. DOI: 10.1002/cncy.20195 (cit. on p. 1).

[5] Yuri E Nikiforov, Raja R Seethala, Giovanni Tallini, Zubair W Baloch, Ronald Ghossein, K Hc, Ricardo V Lloy, M Matias-Pm, and P L. «Nomenclature revision for encapsulated follicular variant of papillary thyroid carcinoma: a paradigm shift to reduce overtreatment of indolent tumors». In: *JAMA Oncology* 2.8 (2016), pp. 1023–1029. DOI: 10.1001/jamaoncol.2016.0386 (cit. on p. 1).

[6] Jeremy L Norris and Richard M Caprioli. «Imaging mass spectrometry: a new tool for pathology in a molecular age». In: *Proteomics Clinical Applications* 7.11-12 (2013), pp. 733–738. DOI: 10.1002/prca.201300055 (cit. on p. 2).

[7] Arif Shaukat. «The rising trend in papillary thyroid carcinoma. True increase or over diagnosis?» In: *Saudi Medical Journal* 39 (May 2018), pp. 531–531. DOI: 10.15537/smj.2018.5.22592 (cit. on p. 6).

[8] Yize Li et al. «Histopathologic and proteogenomic heterogeneity reveals features of clear cell renal cell carcinoma aggressiveness». In: *Cancer Cell* 41.1 (2023), 139–163.e17. DOI: 10.1016/j.ccell.2022.12.001 (cit. on p. 6).

[9] Robert A Van den Berg, Huub CJ Hoefsloot, Johan A Westerhuis, Age K Smilde, and Mariet J Van der Werf. «Centering, scaling, and transformations: improving the biological information content of metabolomics data». In: *BMC Genomics* 7.1 (2006), p. 142. DOI: 10.1186/1471-2164-7-142 (cit. on p. 9).

[10] Gordon K Smyth. «Linear models and empirical Bayes methods for assessing differential expression in microarray experiments». In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), Article3. DOI: 10.2202/1544-6115.1027 (cit. on p. 11).

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. 2nd. The MIT Press, 2018. ISBN: 0262039400 (cit. on p. 14).

[12] Laurens van der Maaten and Geoffrey Hinton. «Visualizing Data using t-SNE». In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html (cit. on p. 14).

[13] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. «UMAP: Uniform Manifold Approximation and Projection». In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861. URL: https://doi.org/10.21105/joss.00861 (cit. on p. 14).

[14] Arthur E. Hoerl and Robert W. Kennard. «Ridge Regression: Biased Estimation for Nonorthogonal Problems». In: *Technometrics* 42.1 (2000), pp. 80–86. ISSN: 00401706. URL: http://www.jstor.org/stable/1271436 (visited on 11/14/2025) (cit. on p. 18).

[15] Robert Tibshirani. «Regression Shrinkage and Selection Via the Lasso». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Dec. 2018), pp. 267–288. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x. eprint: https://academic.oup.com/jrsssb/article-pdf/58/1/267/49098631/jrsssb_58_1_267.pdf (cit. on p. 18).

[16] Hui Zou and Trevor Hastie. «Regularization and Variable Selection Via the Elastic Net». In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67.2 (Mar. 2005), pp. 301–320. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2005.00503.x (cit. on p. 18).

[17] Paul Geladi and Bruce R Kowalski. «Partial least-squares regression: a tutorial». In: *Analytica Chimica Acta* 185 (1986), pp. 1–17. DOI: 10.1016/0003-2670(86)80028-9 (cit. on p. 18).

[18] Matthew Barker and William Rayens. «Partial least squares for discrimination». In: *Journal of Chemometrics* 17.3 (2003), pp. 166–173. DOI: 10.1002/cem.785 (cit. on p. 19).

[19] Kim-Anh Lê Cao, Simon Boitard, and Philippe Besse. «Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems». In: *BMC Bioinformatics* 12.1 (2011), p. 253. DOI: 10.1186/1471-2105-12-253. URL: https://doi.org/10.1186/1471-2105-12-253 (cit. on p. 20).

[20] Michael Altenbuchinger, Antoine Weihs, John Quackenbush, Hans Jörgen Grabe, and Helena U Zacharias. «Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools». In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863.6 (2020), p. 194418. DOI: 10.1016/j.bbagrm.2019.194418 (cit. on p. 22).

[21] Patrick Danaher, Pei Wang, and Daniela M Witten. «The Joint Graphical Lasso for inverse covariance estimation across multiple classes». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.2 (2014), pp. 373–397. DOI: 10.1111/rssb.12033 (cit. on p. 28).

[22] Sjoerd Hermes, Joost van Heerwaarden, and Pariya Behrouzi. *heteromixgm: Copula Graphical Models for Heterogeneous Mixed Data*. R package version 2.0.2. 2024. URL: https://CRAN.R-project.org/package=heteromixgm (cit. on p. 30).