Statistical methods for comparing mass spectra

Malo Hillairet

Contents

This document is a report of my internship at SDSC Lausanne, which took place from May 16 to August 12 2022. Its purpose was to use statistical methods for analysing mass spectrometry data. My tutor at the SDSC was Deputy Chief Data Scientist Guillaume Obozinski. I also interacted with Data Scientists Eliza Harris and Lilian Gasser, who for several years have been carrying the MSEI project to which my work is related to. This project focuses on applying data science to mass spectrometry.A mass spectrometer is a device capable of separating molecules by weight and measuring that weight with extreme precision. Mass spectrometry is used to identify and classify molecules in samples containing hundreds or even thousands of chemical compounds. However, the measurements

are very sensitive to the experimental conditions, and there are other problems, such as noise, that make it difficult to identify molecules using mass spectrometry. The aspect of mass spectrometry that justifies the use of data science and statistics is fragmentation. Indeed, the molecules are randomly fragmented into smaller pieces over the course of the experiment. This results in a lot of data being generated in the form of fragmentation spectra. A fragmentation spectrum is the result of measuring fragments that come from numerous identical molecules. A mass spectrum consists of the data of several molecular weights and the intensities indicating the amount of fragments of the corresponding weight. Since fragmentation is random but repeated for a large number of molecules, it is useful to consider statistical methods for studying mass spectra. My work at SDSC has been to elaborate and implement a statistical test method to determine whether or not two fragmentation spectra represent the same molecule. This report contains an overview of the main mathematical tools that I have been brought to use, as well as a presentation of different models that we considered for studying mass spectra and the confrontation of these models to real data. I tried to not only define the models, but also to explain some of the choices we have made and how we came from considerations on the data to building a mathematical formalism that we hoped would be appropriate for dealing with the problem.

I warmly thank Guillaume for taking time to discuss with me about the project and making me feel part of the SDSC. I also thank Eliza and Lili, whom I had the chance to meet in Zürich. They encouraged me and also took time to look at what I was working on and answer some of my questions. Finally, I thank the members of SDSC Lausanne for welcoming me, I will keep a good memory of this period thanks to them.

1 The SDSC and MSEI

The SDSC (Swiss Data Science Center) is a research organisation based in Switzerland and linked to the EPFL in Lausanne and the ETH in Zürich. There are three teams at the SDSC, academic, industry and renku, each one having members in both cities. The industry team works with companies and provide developing, applied research and consulting services. The renku team develops an online platform similar to GitHub with the aim to generalize the use of data and to make easier for scientists to share their work. Finally, the academic team collaborates with academic research programs, providing expertise in data science. The main fields in which the academic team is working are biology and computer science, but there are also projects in language for instance, or chemistry like the one I have been working on for this internship. The MSEI project has indeed been led by Saša Bjelić from ETH Zürich's chemistry departement. The aim of this project is to use data science methods to help identify molecules from mass spectrometry data. The MSEI (Molecular Structure Elucidation) project has been started in 2019 and carried at the SDSC by Eliza Harris, Lilian Gasser and Guillaume Obozinski. One of the objectives of the project was to classify spectra coming from an experiment realized with a mass spectrometer. What the mass spectrometer is able to do very well is to measure the mass of the molecules, which provides a first level of classification (refer to annex B for further details). The precision is such that the only case where it isn't possible to distinguish two molecules is when they have the same sum formula (i.e. the same number of carbons, hydrogens, oxygens...). In order to do better, the spectrometer breaks molecules into fragments and measures their weights again. The result is called a fragmentation spectrum, or MS2, and is associated to the initial compound that has been fragmented. The problem is that many of the compounds are unknowns (there are hundreds of compounds for each of the three samples) and the spectra are sometimes difficult to distinguish. Moreover, fragmentation is very random and depends a lot on the experimental condition, so that there can be great discrepancies between spectra issued from the same molecule. In the frame of the MSEI project, the SDSC team has faced the problem of determining whether or not two MS2 spectra come from the same molecule. After collecting the data, they implemented a method of tests relying solely on the presence or not of peaks in the compared spectra. This method has been able to classify with satisfying precision about 80 to 90 % of the compounds. The aim of the internship is to come up with a new statistical method for testing by using the peak intensities and not only the presence of

peaks.

At the beginning of the internship, I consulted references [1, 3, 4, 5, 6, 7, 8, 9, 13], which were given to me by Eliza, to learn about mass spectrometry and the current state of methods for comparing spectra. It is worth noting that the best performing algorithms make use of intense machine learning techniques. Annex B gives more details on mass spectrometry.

2 Mathematical background and notations

2.1 Statistical models

When studying a random phenomenon, a first step is to choose an appropriate statistical model. The model is a set of probability distributions on the space of possible outcomes of the random phenomenon and the statistical study is conducted by making the assumption that the phenomenon follows one of those distribution. The goal of the statistical study is then to identify which distributions seem to correspond the most to the behaviour of the random phenomenon from the observations that are made of it.

Example : the Bernoulli model. *As an example, let's say we have a coin that could be biased, and we would like to model the toss of this coin. The outcome of a coin toss are "Heads" and "Tails". By coding "Heads" by 1 and "Tails" by 0, and calling p the probability of "Heads", the outcome of a coin toss follows a Bernoulli distribution of parameter p. Therefore, a good choice of model would be the set of Bernoulli distributions of parameter* $p \in [0, 1]$.

In this example, the model is called parametrized by the variable p because the probability distributions of the model can be characterized by this number p. The model is still called parametrized when there are multiple parameters or parameters that live in a finite-dimensional vector space (it can even be a differentiable manifold). When studying a random phenomenon under a statistical model, the goal is to determine information on the true distribution that the phenomenon follows by observing outcomes of the phenomenon. In the example of the biased coin, it would be to get information on the parameter p by playing heads and tails a large number of times, and inferring information on p given the observations.

The practical necessities when looking for a statistical model are that it is appropriate for the random phenomenon we want to study, and that it is convenient to work with. The appropriateness of a model has to do with prior knowledge of the phenomenon, obtained through a theoretical study or a previous statistical study for instance. The more appropriate the model is, the more accurate will be the outcome of the statistical study. The convenience of the model only relies on its mathematical properties: roughly, a model is considered convenient when it involves a limited number of parameters and relies on well-known mathematical random variables. The more convenient the model is, the easier the mathematics of the model are. Convenience is related to simplicity, and going for simpler mathematical ideas means leaving aside physical aspects of the phenomenon that could make the model more appropriate. Conversely, trying to make the model appropriate leads to adding more and more complexity to the mathematics of the model. Because of that, it can be difficult to have a model that is both appropriate and convenient when dealing with complex phenomena.

In the next sections, I will sometimes introduce a model by a formula of the form

$$\mathcal{M} = \{\text{Set of probability distributions, indexed by a parameter}\}.$$

A sometimes more convenient way to introduce a model is by using one or multiple random variables related to the outcome of the event, and specifying the law of the random variable under the possible parameters. The definition of the model would be formulated in the following way : "Let X be the

outcome of that random event. We model the behaviour of X by a random variable of probability distribution $\mathbb{P}_\theta$, where θ is an unknown parameter and belongs to the space Θ."

Example : the Bernoulli model. *In the example of the coin, one toss corresponds to observing the outcome of a random variable of parameter p, and in the frame of a statistical study we will do multiple tosses, corresponding to multiple independent Bernoulli variables. Setting a number of tosses n, the model then writes*

$$\mathcal{M} = \{\mathrm{Ber}(p)^n, p \in [0,1]\},$$

or can be described by the parameter set being $\{p \in [0,1]\}$ *and the observation data being a sequence of n random variables* $X_1, \ldots, X_n$*, independent and identically distributed along the Bernoullli distribution of parameter p. This model is often called Bernoulli model.*

2.2 The likelihood function

An important tool when studying a parametrized statistical model is the likelihood function, which is often denoted L (sometimes with indexes), and takes as input a parameter coupled with a possible issue of the associated random variable. The formal definition involves a measure μ that is said to dominate the model, which means that all the probability distributions of the model are absolutely continuous with respect to μ. In practice, μ is often a common measure on the outcome space of the model like the counting measure if the outcome space is discrete or Lebesgue's measure if the outcome space is $\mathbb{R}^d$ or a subset of $\mathbb{R}^d$ for some positive integer d.

Definition 2.1. Let Ω and Θ be measurable spaces. Let $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ be a statistical model having Ω as outcome space. Let μ be a measure on Ω that dominates the model. The function $L : \Omega \times \Theta \to \mathbb{R}_+$ defined by

$$L(x, \theta) = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x)$$

is called likelihood function associated to the model, where $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$ is the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to μ.

Note that statistical reasoning assumes knowledge of x in this expression, so the likelihood function is in practice considered as being from the space of parameters to $\mathbb{R}_+$. In the case of a discrete model, the likelihood function at (x, θ) is just the probability mass of x under the distribution associated to θ, and for a continuous distribution the likelihood function at (x, θ) is the evaluation at x of probability density function of the distribution associated to θ. It is called likelihood function because it indicates how likely it is that the true parameter is θ given that the observed data is x. The range of values of the likelihood function depends strongly on the model and also the dominating measure μ. What is meaningful is comparing the values of the likelihood function evaluated at different parameters for the same observation x. Indeed, it provides information on how more likely a parameter is compared to the other ones.

Example : the Bernoulli model. *Let n be a positive integer and let* $\mathcal{M}$ *be the Bernoulli model with n trials, i.e. defined by* $\mathcal{M} = \{\mathrm{Ber}(p)^n, p \in [0,1]\}$*. For* $p \in [0,1]$*, let* $\mathbb{P}_p$ *be the probability distribution associated to a sequence* $X_1, \ldots, X_n$ *of independent random variables following Bernoulli distributions of parameter p. The Radon-Nikodym derivative of* $\mathbb{P}_p$ *with respect to the counting measure on* $\{0,1\}^n$ *is its probability mass function. Thus, for all* $x \in \{0,1\}^n$*,*

$$L(x, p) = \prod_{k=1}^{n} \mathbb{P}_p(X_k = x_k) = \prod_{k=1}^{n} \left(p^{x_k}(1-p)^{1-x_k}\right).$$

It follows that the likelihood function of the Bernoulli model satisfies

$$\forall p \in [0,1],\ \forall x \in \{0,1\}^n,\ L(x, p) = p^{\sum_{k=1}^n x_k}(1-p)^{(n-\sum_{k=1}^n x_k)}.$$

2.3 Hypothesis testing

The goal of a statistical study should not necessarily be to determine the parameter that best fits the observations. It can also be less precise. In experimental science, for example, it is common to conduct an experiment to test the validity of a hypothesis. In this case, if a statistical study is conducted, the result should be the acceptance or rejection of the hypothesis, depending on the outcome of the experiment. In statistics, there is a formal framework for dealing with this type of problem called "hypothesis testing". It requires a statistical model, as described earlier, and two statements called the "null hypothesis" and the "alternative hypothesis," where the validity of the null hypothesis is tested and the alternative hypothesis corresponds to the possibilities that may arise if the null hypothesis is invalid.

Example : the Bernoulli model. *In the example of the coin toss, we could for instance check the validity of the statement "the coin isn't biased". Within the Bernoulli model* $\mathcal{M} = \{\mathrm{Ber}(p)^n,\, p \in [0,1]\}$*, the null hypothesis would then be "*$p = \frac{1}{2}$*" and, without further information on the coin, a reasonable choice of alternative hypothesis would be "*$p \neq \frac{1}{2}$*".*

The two hypotheses can be formulated in many different ways, but in the case of parametric models, each of them corresponds to a subspace of the parameter space. It is important to note that the two hypotheses do not play the same role. They differ in the interpretation given to them and in the way the test is consequently designed. It is useful to think of the null hypothesis as the one we would assume by default if no observation were made. This is similar to the concept of presumption of innocence. A statistical study is like a trial in which confirmation of the null hypothesis plays the role of acquitting the defendant. If there is insufficient evidence against the defendant, he should be presumed innocent. The statistical data play the role of evidence used to judge whether the null hypothesis should be invalidated. Therefore, the alternative hypothesis should be considered only when the data show poor agreement with the null hypothesis. Sometimes it is not obvious to make a correct choice of hypotheses. It should be remembered that the alternative hypothesis and the null hypothesis are not necessarily complementary, and the null hypothesis is usually more restrictive.

2.4 Type I and Type II error

A standard way to conduct a test is to compute a test statistic, which is a function of the observations. The test statistic should indicate how much evidence the data provide against the null hypothesis. It follows that a useful way to distinguish between the null and alternative hypotheses is to set a threshold and reject the null hypothesis only if the test statistic is above that threshold. As always in statistics, there is a small chance that the data are not representative of reality, which would cause the test to fail. Two types of error can be assessed. What is called type I error, or false positive, is the probability that the null hypothesis is rejected when it should be accepted, and the type II error, also called false negative, is the probability that the null hypothesis is accepted when it should be rejected. Such a probability can only be defined theoretically for an individual distribution of the model. When a hypothesis encompasses multiple parameters, the worst case scenario is assumed and the error is as the supremum of the errors associated to each individual parameter. In terms of test statistic, let T be the test statistic and t_0 the threshold above which the null hypothesis is rejected. Then theoretical type I error writes as

$$\alpha = \sup_{\theta / \mathcal{H}_0} \mathbb{P}_\theta(T > t_0)\,,$$

where the supremum is taken over all the parameters θ satisfying $\mathcal{H}_0$, and theoretical type II error writes as

$$\beta = \sup_{\theta / \mathcal{H}_1} \mathbb{P}_\theta(T < t_0)\,,$$

where the supremum is taken over all the parameters θ satisfying $\mathcal{H}_1$.

The evaluation of the performance of a test relies on the values of these errors. Prescribing what is satisfactory as a Type I error and a Type II error strongly depends on the situation. The common procedure that is used in most medical tests, surveys etc... is to design the test to limit the type I error to 5 %. In this case, the test can also be said to have a level of confidence of 95 %. Type II error can then be evaluated to obtain further information on the test performance, which can be more or less relevant depending on the model and the test hypotheses. In the next example, type II error does not matter much because the test is only designed to highlight evidence against the null hypothesis.

Example : the Bernoulli model. *In the model of independent Bernoulli variables $X_1, \dots, X_n$ of parameter p, consider testing the null hypothesis "$p = \frac{1}{2}$" versus the alternative hypothesis "$p \neq \frac{1}{2}$" with a level of confidence of 95 %. The further $\frac{1}{n}\sum_{k=1}^{n} X_k$ is from $\frac{1}{2}$, the stronger is the evidence against the null hypothesis, which justifies the choice of the quantity $T = |\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{2}|$ as a test statistic. We would like to determine a value t_0 for which the decision to reject the null hypothesis when $T > t_0$ implies a type I error of 5 %. As the type I error α is the probability of rejecting the null hypothesis when it should be accepted, it writes as*

$$\alpha = \mathbb{P}_{\frac{1}{2}}(T > t_0)\,,$$

where $\mathbb{P}_{\frac{1}{2}}$ is the probability distribution of the model for $p = \frac{1}{2}$, i.e. if the null hypothesis is satisfied. It follows from the central limit theorem that $\sqrt{4n}(T - \frac{1}{2})$ converges in distribution to a centered gaussian distribution of variance 1. It is well-known that a gaussian random variable has approximately 95 % chance to fall within 1.96 standard deviations from the mean. Thus, choosing the value $t_0 = \frac{1.96}{\sqrt{4n}}$ makes α approximately equal to 5 % if n is large enough for the approximation of the central limit theorem to be valid. The alternative hypothesis contains parameters arbitrarily close to $\frac{1}{2}$, so the probability of accepting the null hypothesis can be made arbitrarily close to the probability of accepting the null hypothesis when the parameter is $\frac{1}{2}$, which is 95 %. Thus, theoretical type II error is 95 %.

Note that, in presence of real data and multiple outcomes of the test, empirical errors can be evaluated, which correspond to the empirical proportions of cases where the null hypothesis should have been accepted/ rejected but was not.

3 Likelihood ratio test in a multinomial model

3.1 From Poisson variables to contingency tables

A single spectrum consists in the data of a certain number of peaks, each peak being represented by a couple of non-dimensional quantities called m/z and intensity. The name m/z stands for mass-to-charge ratio and is the ratio of the molecular mass normalized by the mass of a nucleon, by the absolute value of the charge number of the ion. For instance, 3-Methylglutaric acid has sum formula $C_6H_{10}O_4$, and the corresponding ion that has undergone the mass spectrometry experiment is $C_6H_{10}0_4^-$. Its molecular mass is about 146, and its charge number is -1, so its m/z is approximately 146. The intensity associated to a m/z value roughly corresponds to the amount of molecules of that m/z that have been measured. The following graphic shows a representation of a fragmentation spectrum corresponding to 3-Methylglutaric acid. The choice of bar diagram is standard for mass spectrometry data and is due to the fact that intensity corresponds to a count of molecular fragments. As we can see, there is a peak at a m/z value around 146, which should correspond to the molecules of 3-Methylglutaric acid that have not been dissociated throughout the fragmentation phase. It is quite common for a molecule to lose an H_2O or a CO_2 throughout the fragmentation. On the spectrum, the corresponding peaks are at m/z near 128 (for $C_6H_{10}O_4 - H_2O$) and near 102 (for $C_6H_{10}O_4 - CO_2$). The first model that we tried was considering that the peak mz are fixed and modelling their intensities by independent Poisson variable. This is reasonable because, given the randomness of fragmentation, obtaining a fragment of a particular molecular weight can be considered as a rare event, and Poisson variables are good for

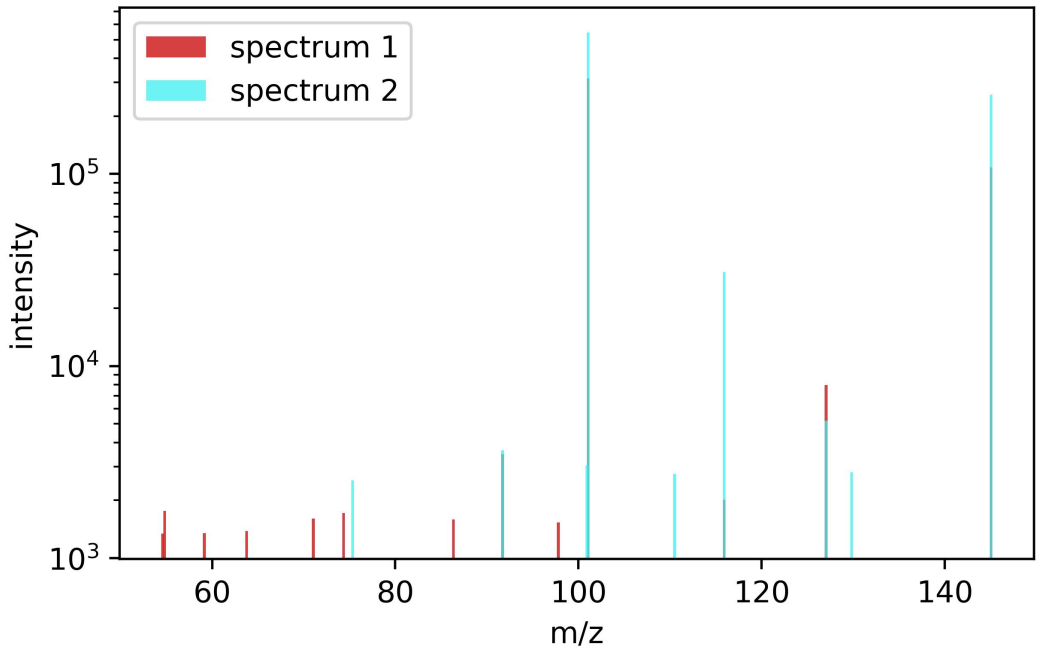


Figure 2.1: Superposition of two fragmentation spectra corresponding to 3-Methylglutaric acid

modelling rare events. However, as we can see on figure 2.1, some peaks appear in one spectrum but not in the other one. Most probably, the majority of the latter come from noise and therefore should not be included in the d peaks of which the intensities follow Poisson variables and are characteristic of the molecule. The strategy that I adopted to implement this model nonetheless was to, given two spectra to be compared, select all peaks having m/z values appearing in both spectra and discard the other peaks. A problem that I could have with this method is that noise peaks are sometimes matched with real peaks, but it in fact happens very rarely thanks to the spectrometer precision on the m/z measure (I mention in annex C some experiments I have done with the data, that make me confident about this claim). On the other hand, some true peaks, thus holding relevant information, could be mistakenly discarded. It makes the test results less significant when only a few peaks remain.

Since two spectra are compared, there is one d-dimensional vector of independent Poisson variables for each spectrum. Also, the initial number of molecules that undergo fragmentation is unknown. This is represented by a quantity n^* as a factor in the parameters of the Poisson variables. There are two spectra, so two such quantities n_1^* and n_2^*, and the model writes

$$\mathcal{M}_1 = \left\{ \prod_{j=1}^{d} \mathcal{P}(n_1^* \times \theta_{1,j}) \otimes \prod_{j=1}^{d} \mathcal{P}(n_2^* \times \theta_{2,j}),\ \theta \in \Theta \right\},$$

where Θ is the space of parameters $\Theta = \{(\theta_{1,1}, \ldots, \theta_{1,d}, \theta_{2,1}, \ldots, \theta_{2,d}) \in (\mathbb{R}_+^*)^{2d}\}$. Here, the outcome of the Poisson variable of parameter $n_i^* \theta_{i,j}$ is the intensity of the j^{th} peak in spectrum i. This intensity is denoted by $n_{i,j}$ in the sequel. It is to note that the n_i^* contain no information on the nature of the molecules, and therefore should not be considered as parameters, nor as random variables in the model. It is however important to take into account that these quantities are unknown, and this will play a role in our choice of test hypotheses.

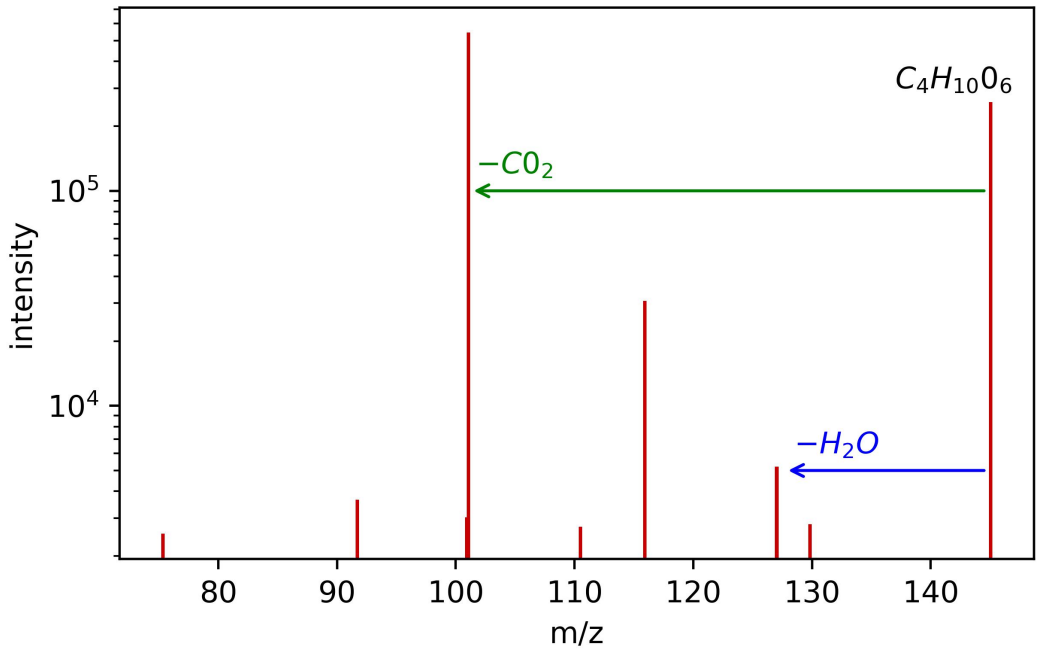


Figure 3.1: A fragmentation spectrum corresponding to 3-Methylglutaric acid

In model $\mathcal{M}_1$, the parameters characterizing molecule 1 are the $\theta_{1,j}$, and those characterizing molecule 2 are the $\theta_{2,j}$. Thus, the hypothesis that the molecules match should be "$\forall j \in \{1,\dots,d\}$, $\theta_{1,j} = \theta_{2,j}$" . Conversely, the hypothesis that the molecules are different would be "$\exists j \in \{1,\dots,d\}$, $\theta_{1,j} \neq \theta_{2,j}$". The latter hypothesis is a lot wider than the other one. The set of parameters $(\theta_{i,j})$ corresponding to the hypothesis of the molecules matching is a thin subspace of dimension d of the $2d$-dimensional space Θ, whereas the set corresponding to distinct molecules is everything else. It would therefore be a bad choice to take the hypothesis of no match as null hypothesis, because no statistical data can really provide evidence against it. The null hypothesis should then be the one of matching. However, remember that the n_i^* are unknown, so that the vectors $(\theta_{1,j})_{1\leqslant j\leqslant d}$ and $(\theta_{2,j})_{1\leqslant j\leqslant d}$ can only be estimated up to multiplication by a positive number. As a consequence, it is not possible to do better than taking as null hypothesis $\mathcal{H}_0$: "$\exists \alpha > 0$, $(\theta_{1,j})_{1\leqslant j\leqslant d} = \alpha \cdot (\theta_{2,j})_{1\leqslant j\leqslant d}$", and as alternative hypothesis the converse statement. As a subspace of the space of parameters, $\mathcal{H}_0$ can be represented as

$$\Theta_0 = \{\theta \in \Theta / \forall j \in \{1,\dots,d-1\},\, \theta_{1,j}\theta_{2,j+1} = \theta_{1,j+1}\theta_{2,j}\}\,.$$

Indeed, the vectors $(\theta_{1,j})_{1\leqslant j\leqslant d}$ and $(\theta_{2,j})_{1\leqslant j\leqslant d}$ are collinear if and only if for all $1 \leqslant j \leqslant d-1$, the vectors $\begin{pmatrix}\theta_{1,j}\\ \theta_{1,j+1}\end{pmatrix}$ and $\begin{pmatrix}\theta_{2,j}\\ \theta_{2,j+1}\end{pmatrix}$ are collinear, which is characterized by the determinant $\theta_{1,j}\theta_{2,j+1} - \theta_{1,j+1}\theta_{2,j}$ of these two vectors being equal to 0.

The data consisting in the intensities of the $2d$ peaks can be arranged into what is called a contingency table and is a common tool in statistics. Indeed, our model the intensities can be classified in the following table, where the intensity of the j^{th} peak of spectrum i appears in the (i,j) cell of the contingency table, $n_{i,+}$ is the sum of the intensities in row i, $n_{+,j}$ is the sum of the intensities that are in column j, and N is the total sum of all the intensities in the table.

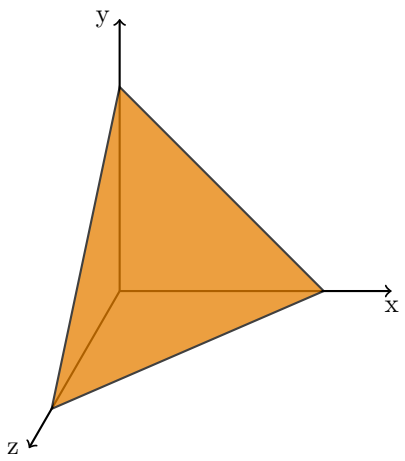


Figure 3.2: The 2-dimensional simplex Δ^2 in $\mathbb{R}^3$

	Peak 1	Peak 2	...	Peak d	Row sums
Spectrum 1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,d}$	$n_{1,+}$
Spectrum 2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,d}$	$n_{2,+}$
Column sums	$n_{+,1}$	$n_{+,2}$	...	$n_{+,d}$	N

There is a lot of documentation on tests and analysis of contingency tables when the data follows a multinomial law (see [11, 12]). Thanks to the following lemma, which is a property of the Poisson distribution, it is the case for the contingency table of peak intensities under the model $\mathcal{M}_1$. The proof of this lemma as well as the other proofs of this section are given in annex A.

Lemma 3.1. *Let d be a positive integer and $N_1, \dots, N_d$ be independent random variables following Poisson distributions of respective parameters $\lambda_1, \dots, \lambda_d$. Let $N = N_1 + N_2 + \dots + N_d$. The distribution of the random vector $(N_1, \dots, N_d)$ given the sum N is multinomial with number of trials N and probability weights $(\frac{\lambda_i}{\lambda_1 + \dots + \lambda_d})_{1 \leqslant i \leqslant d}$.*

Applying lemma 3.1 to the each of the rows of the table $(n_{i,j})_{\substack{1 \leqslant i \leqslant 2 \\ 1 \leqslant j \leqslant d}}$, conditionally to $n_{1,+}$ and $n_{2,+}$, the row vectors of intensities $(n_{1,j})_{1 \leqslant j \leqslant d}$ and $(n_{2,j})_{1 \leqslant j \leqslant d}$ follow multinomial random variables of respective parameters $(n_{1,+}, (p_{1,j})_{1 \leqslant j \leqslant d})$ and $(n_{2,+}, (p_{2,j})_{1 \leqslant j \leqslant d})$, where $p_{i,j} = \frac{\theta_{i,j}}{\sum_{j=1}^{d} \theta_{i,j}}$ for all $1 \leqslant i \leqslant 2$, $1 \leqslant j \leqslant d$. This justifies the choice of a new model

$$\mathcal{M}_{\text{MN}} = \left\{ \text{Multi}(n_{1,+}, \mathbf{p_1}) \otimes \text{Multi}(n_{2,+}, \mathbf{p_2}),\ (\mathbf{p_1}, \mathbf{p_2}) \in \Delta^{d-1} \times \Delta^{d-1} \right\}, \tag{3.1}$$

where Δ^{d-1} is the $d-1$ dimensional simplex, that is, the set of d-uples of positive real numbers summing up to 1. Figure 3.2 shows a representation of the 2-dimensional simplex. A new expression for the hypothesis of matching is

$$\mathcal{H}_0 : \text{“}\mathbf{p_1} = \mathbf{p_2}\text{”}. \tag{3.2}$$

In the vocabulary of contingency tables, this corresponds to a hypothesis of independence for two categorical variables, one represented by the column of the table and the other one by its rows. In the sequel, E denotes the set $\Delta^{d-1} \times \Delta^{d-1}$ as space parametrizing $\mathcal{M}_{\text{MN}}$, and E_0 denotes the subset $\{\mathbf{p_1} = \mathbf{p}_2\}$ corresponding to the subspace of parameters satisfying $\mathcal{H}_0$.

3.2 Constructing a likelihood ratio test

The likelihood function gives information, given data observations and a parameter, on how well the distribution associated to this parameter fits the observations. Then, the most likely parameters are

those at which the likelihood function is the greatest. When dealing with hypothesis testing, this reasoning has led to the method of likelihood ratio test.

Definition 3.1. Let $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ be a statistical model admitting a maximum likelihood function L. Let Θ_0 and Θ_1 be two subsets of Θ defining a null hypothesis $\mathcal{H}_0$: "$\theta \in \Theta_0$" and an alternative hypothesis $\mathcal{H}_1$: "$\theta \in \Theta_1$". Given observation data x, the likelihood statistic T_{LL} associated to the test of $\mathcal{H}_0$ against $\mathcal{H}_1$ is defined as

$$T_{\mathrm{LL}} = 2 \log \frac{\max_{\theta \in \Theta_1} L(x, \theta)}{\max_{\theta \in \Theta_0} L(x, \theta)} . \tag{3.3}$$

The logic behind the definition is that the likelihood ratio $\frac{\max_{\theta \in \Theta_1} L(x,\theta)}{\max_{\theta \in \Theta_0} L(x,\theta)}$ indicates how more likely it is for the alternative hypothesis to be true compared to the null hypothesis. Taking double the logarithm of this quantity is justified by the result that it approaches a χ-squared distribution in the limit of a large amount of data. This property is called Wilk's theorem, there is a demonstration in [15]. The likelihood statistic is in fact closely linked to Pearson's χ^2 statistic (see [2]). The χ^2 tests are quite useful when dealing with contingency table, but the main advantages are that the test statistic is fast and easy to compute and that the computation of p-values from it is straightforward. However, the χ^2 approximation did not give significant results with the data I worked with, probably because the model does not describe fragmentation well enough. Therefore, I decided to only use the statistic derived from the likelihood.

In model $\mathcal{M}_{\mathrm{MN}}$, for each $i \in \{1, 2\}$, the row $\mathbf{n_i} = (n_{i,1}, \dots, n_{i,2})$ follows a multinomial distribution of parameters $(n_{i,+}, \mathbf{p_i})$, and the two rows are independent. Thus, the likelihood function writes

$$L(\mathbf{n_1}, \mathbf{n_2}, \mathbf{p_1}, \mathbf{p_2}) = \prod_{i=1}^{2} \left[\binom{n_{i,+}}{\mathbf{n}_i} \prod_{j=1}^{d} p_{i,j}^{n_{i,j}} \right] .$$

It is easy to show that the logarithm of the likelihood function is concave, so the computation of the likelihood statistic basically involves making its gradient vanish as function on E for $\mathcal{H}_1$ and as function on E_0 for $\mathcal{H}_0$. The full proof is given in annex A.

Proposition 3.1. The likelihood statistic for the test of hypothesis defined by 3.2 in model $\mathcal{M}_{\mathrm{MN}}$ is given by

$$T_{\mathrm{LL}} = 2 \sum_{j=1}^{d} n_{i,j} \log \left(\frac{n_{i,j} N}{n_{1,+} n_{2,+}} \right) . \tag{3.4}$$

Qualitatively, we can see that this expression makes sense. Indeed, under the alternative hypothesis, the probability weights $p_{1,1}, p_{1,2}, \dots, p_{1,d}, p_{2,1}, \dots, p_{2,d}$ that fit best the data are given by $\hat{p}_{i,j}^{(1)} = \frac{n_{i,j}}{n_{i,+}}$. In the vocabulary of mass spectra, $\hat{p}_{i,j}^{(1)}$ is the intensity of peak j in spectrum i normalized by the sum of intensities of spectrum i. Under the null hypothesis $\mathbf{p_1}$ and $\mathbf{p_2}$ have to be equal, so this estimation becomes $p_{i,j}^{(0)} = \frac{n_{1,j}+n_{2,j}}{n_{1,+}+n_{2,+}} = \frac{n_{2,+}}{N}$, and the likelihood ratio involved in the likelihood statistic is simply the ratio of the likelihood function when evaluated with those parameters. Then, the likelihood statistic writes as

$$T_{\mathrm{LL}} = 2 \sum_{i=1}^{2} \left(n_{i,+} \cdot \sum_{j=1}^{d} \hat{p}_{i,j}^{(1)} \log \left(\frac{\hat{p}_{i,j}^{(1)}}{\hat{p}_{i,j}^{(0)}} \right) \right) . \tag{3.5}$$

Given $i \in \{1, 2\}$, the sum over j in (3.5) can be expressed as the Kullback-Leibler divergence of the distributions of categorical variables on $\{1, \dots, d\}$ of respective probability weights $(\hat{p}_{i,j}^{(1)})_{1 \leqslant j \leqslant d}$ and $\hat{p}_{i,j}^{(0)})_{1 \leqslant j \leqslant d}$. As Kullback-Leibler divergence is a distance on the set of distributions on a given space, the greater the discrepancy is between $\hat{\mathbf{p}}^{(1)}$ and $\hat{\mathbf{p}}^{(0)}$, the greater the likelihood statistic will be.

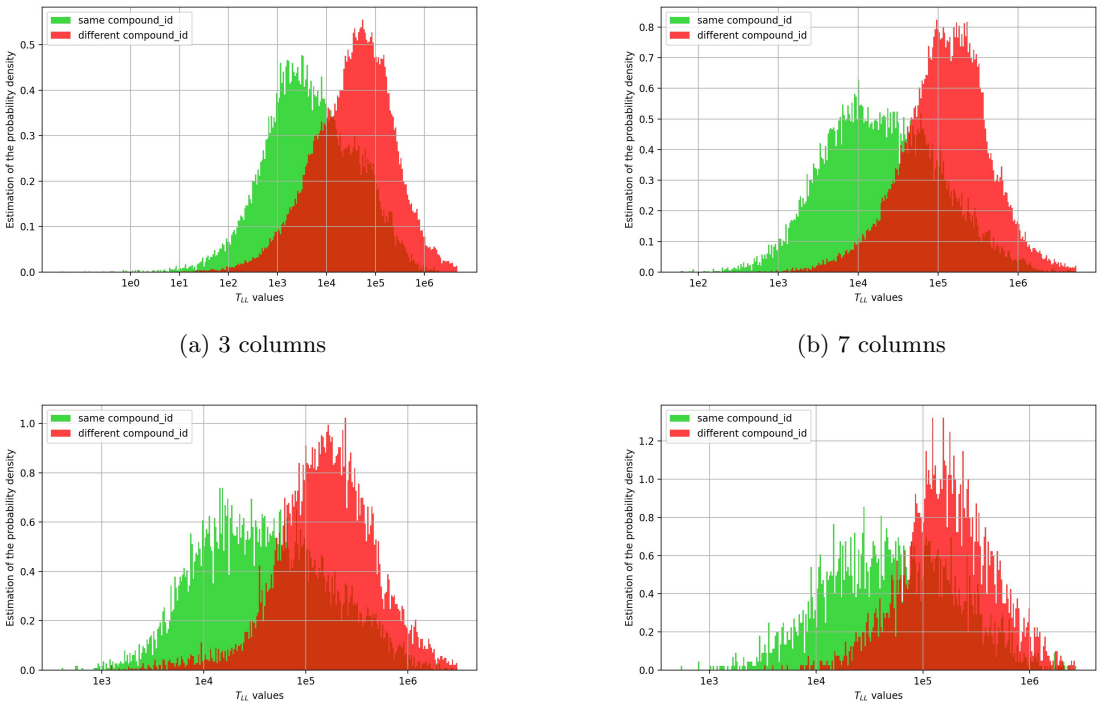
3.3 First results and interpretation

There are 4 samples available, called DiAcids, Fallopia, Ruthenium and Ruthenium2, and Fallopia is the largest one with more than 800 compound labels and almost 20 000 fragmentation spectra. I ended up evaluating my models almost only on the Fallopia sample, especially because there was that much data and this is more convenient for a statistical study. The labels I used are the ones provided by a commercial algorithm included to the mass spectrometer. I call these labels "compound id" because that is how they are referenced in the csv files containing the data. Bear in mind that this algorithm is likely to have made mistakes, but hopefully not too many, and labels are necessary to have a chance to estimate the proportion of false positives and false negatives. The commercial algorithm includes a search in a database, so that some compound ids are assigned to a known compound. When it is the case, there is a match quality (in short, mq) indicating of the confidence with which the spectrum has been assigned to the database compound. It means that compound id with high match quality are unlikely to have been mislabelled. Guillaume warned me about using only high mq compounds though, because he wanted me to test my models on all the compounds and not only those able to be identified accurately by previous methods. Only fragmentation spectra coming from molecules having the same molecular weight can be compared. Thanks to the spectrometer precision, this implies in most cases for the two initial molecules to have the same sum formula. For instance, there is no need to compare the MS2 spectra of one molecule having $C_6H_{10}O_4$ and $C_5H_{10}O_5$ as respective sum formulas because we already know that these molecules are not the same. Around two thirds of the compounds can be compared to at least an other one, in most cases to exactly one but groups of 3 to 5 compound ids sharing the same sum formula are not rare either. Thus, I split the tables into two groups, the first one containing tables computed from spectra having the same compound id and the other one from spectra having different compound id. I made the choice to discard the compounds that could be compared to no other for computing empirical distributions of the log-likelihood statistic in the different groups. My main worry was that the molecules sharing sharing the sum formula of no other could be more likely to have particular physical or geometric properties skewing the distribution of the likelihood statistic.

The following graphics show empirical distributions of the logarithm of the likelihood statistic. Taking the logarithm is only intended to make the graphics readable (this is a way to put a logarithmic scale on the x-axis, but it also changes the shape of the distribution). The data has been separated with respect with the number of columns in the tables. Indeed, the number of columns plays the role of a number of degrees of freedom in the multinomial model, so comparing values of the likelihood statistic obtained from tables having different number of columns would make little sense.

We observe that the likelihood statistic is overall lower for tables corresponding to two spectra labelled by the same compound id (the green histogram on figure 3.3). This is satisfying, it indicates that the multinomial model is relevant to solve the problem of distinguishing molecules from MS2 spectra. Ideally, the red histogram would be almost completely to the right of the green histogram. This would imply that there is a clear threshold able to distinguish between spectra from identical molecules and spectra from different molecules. Here the curves overlap so that the empirical errors of type I and II might not be as small as we would like.

Now, a threshold has to be chosen for applying the test. The distribution of the likelihood statistic for the multinomial model with d peaks converges to a χ-squared distribution with $d-1$ degrees of freedom when the number of trials N converges to $+\infty$ (this is discussed in [11, 12, 15]). This number ranges from 10^3 to 10^7 in the data, which is large enough for this estimation to hold really well if the data were to exactly fit the multinomial model (10^3 largely overrides 30, which is often given as a condition for the central limit theorem to be valid with good approximation). In fact, the likelihood statistic asymptotically approaches Pearson's χ-squared statistic, which is frequently used for tests involving contingency tables like this one (see articles). It is standard for those kinds of tests to set a type I error α and use as a threshold the $(1-\alpha)$-quantile of the χ-squared distribution with the appropriate number of degrees of freedom. It makes the theoretical type I error equal to α. However, this would be a bad idea here because the χ-squared distribution has a mean equal to its number



(a) 3 columns

(b) 7 columns

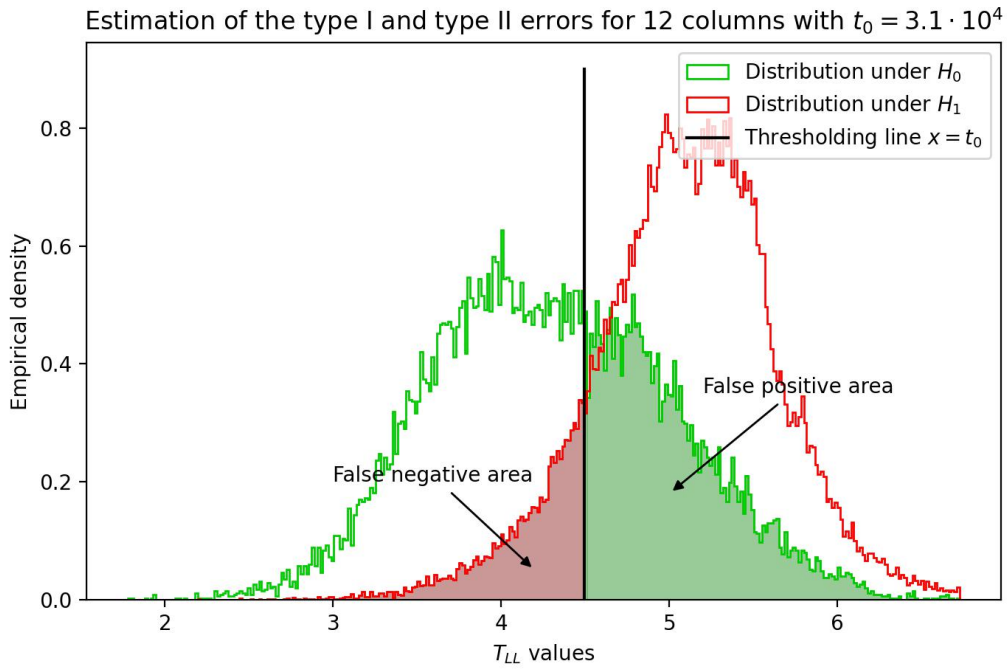
(c) 12 columns

(d) 16 columns

Figure 3.3: Repartition of the likelihood statistic for tables with same compound id vs different compound id for different number of columns.

of degrees of freedom, whereas most of the likelihood statistic values obtained from mass spectra are bigger than 100 for any number of columns, whereas the number of degrees of freedom is $d-1$ when the number of columns is d. This is a way in which the multinomial model differ from the real data. We thus have no choice but to estimate a threshold from the data. Furthermore, a straightforward way of still analysing type I and type II errors is to estimate those empirically for any choice of threshold, which can be summarized with a ROC curve, as illustrated on figure 3.4. Note that all of the Fallopia spectra have been used for plotting these graphics, including the spectra labelled with low match quality and the spectra that haven't been assigned to any compound at all. We can see that for a particular threshold t_0, the estimation of type I error α is 0.39, and the estimation of type II error is 0.13, so that the power $1-\beta$ is 0.87. The ROC curve is the parametrized curve described by $(x, y) = (\alpha(t_0), 1-\beta(t_0))$ obtained by making the threshold t_0 vary over $\mathbb{R}$. It gives information on the performance of the test: a straight line going from $(0,0)$ to $(1,1)$ indicates that the test statistic makes no difference between the null hypothesis and the alternative hypothesis ; and the more the ROC curve is bent towards the point $(0,1)$, the better the test performs. Another advantage of the ROC curve is that it makes it easy to read the power associated to a fixed type I error, and the type I error associated to a fixed power.

A disadvantage of using ROC curves to evaluate performance of the test is that one ROC curve corresponds to only one number of columns. Moreover, in practice we would like to decide of a power or type I error to reach beforehand and set the threshold accordingly. Eliza and Lili told me that the emphasis should be put on power. Indeed, what chemists need is a classifier for the molecules, so that they really don't want different molecules to be identically labelled. An usual choice would then be to go for 95 % power, but we can see on the ROC curve that it would result in a huge type I error (more



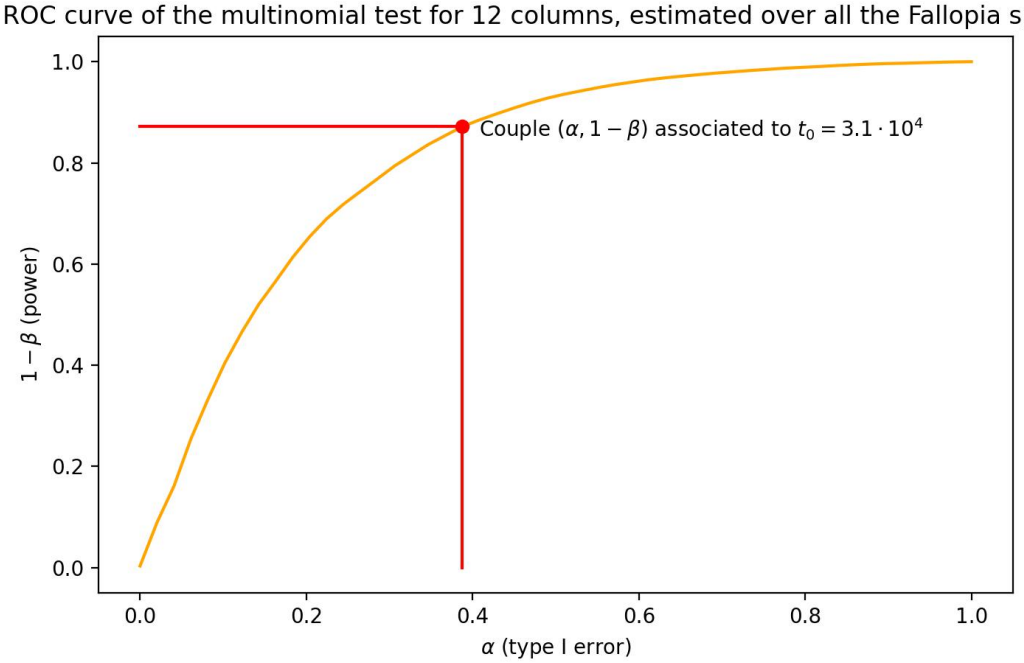


Figure 3.4: ROC curve of the test in the multinomial model for 12 columns

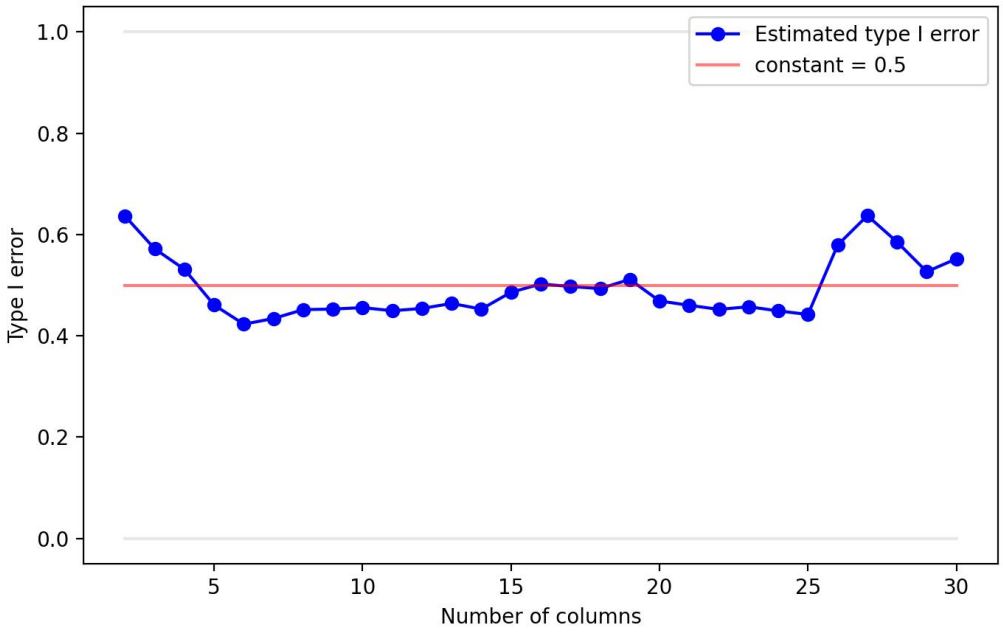


Figure 3.5: Type I error associated to 90 % power with respect to the number of columns

than 50 %) which is really not satisfying. Thus, I decided to go for a power of 90 %. Now, I only need to find the value of t_0 making the empirical proportion of false negative 10 % and evaluate the type I error associated to this threshold for every number of columns. The result is showed on figure 3.5. This way of evaluating the test is pushed further in annex D, with the aim to add confidence intervals and deal with correlation within the data. As we can see on figure 3.5, the type I error obtained by going for 90 % power is just below 0.5 for most number of columns, which is still not very satisfying. There are very few tables with more than 30 columns so it is not very relevant to plot the results for that much columns.

4 Adding overdispersion to the multinomial model

4.1 The Dirichlet-multinomial model

We have seen that the multinomial model was able to make a difference appear between the "same compounds" group and the "different compounds" group. However, it was a rather simple model and the distributions associated to the null and alternative hypothesis still overlap quite much for all number of columns. Furthermore, the observations corresponding to the null hypothesis are far from the values of a χ^2 distribution as it would be the case for contingency tables really following multinomial distribution with entries of order 10^3 and more. It means that the multinomial model is not describing fragmentation accurately enough. In fact, it looks like the variance of the data is greater in reality than in the multinomial model. It means that introducing overdispersion to the model could be a good idea. This can be done by replacing the multinomial distribution by a compound distribution. In the new model, the random variable modelling fragmentation is still assumed to follow a multinomial distribution but this time the parameter $\mathbf{p}$ of the multinomial is also a random variable. In this new model, there is a new set of parameters describing the distribution of $\mathbf{p}$. This is said to introduce

overdispersion because the observations $\mathbf{n}$ behave like a multinomial variable with greater variance. Articles [10, 16] give a more detailed description of the Dirichlet-multinomial model.

In order to give an idea of how the Dirichlet multinomial distribution is constructed mathematically, I expand a bit on the following example of the compound distribution obtained by making the parameter p of a binomial variable follow a Beta distribution. This is by the way the case $d = 2$ of the Dirichlet-multinomial distribution which we will use as a model for fragmentation spectra in the sequel. Let $N \geqslant 2$ be an integer. Let α, β be positive real numbers and let p be a random variable following a beta distribution of parameters α and β. The expression $X|p \sim \mathrm{Bin}(N, p)$ defines a random variable X with values in $\{0, 1, \ldots, N\}$. Then, X satisfies

$$\begin{aligned}
\mathbb{P}(X = k) &= \int_0^1 \mathbb{P}(X = k|p = t) \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \, \mathrm{d}t \\
&= \int_0^1 \binom{n}{k} t^k (1-t)^{n-k} \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \, \mathrm{d}t \\
&= \binom{n}{k} \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 t^{k+\alpha-1}(1-t)^{n-k+\alpha-1} \, \mathrm{d}t \\
&= \binom{n}{k} \frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha, \beta)} \, .
\end{aligned}$$

The fact that we were able to integrate the conditional likelihood function $\mathbb{P}(X = k|p = t)$ against the probability density of the beta distribution is related to the beta distribution being a conjugate prior to the binomial distribution. There are other ways of writing the probability mass function of X, for instance replacing the beta functions by products of gamma simplifying some terms. We will dive further into that kind of computations later, when implementing the computation of the likelihood function. Recall that the binomial random variable of parameters (N, p) has expectation Np and variance $Np(1-p)$. The expectation of the beta-binomial distribution of parameters (α, β) is

Proposition 4.1. Let α, β be positive real numbers and $N \geqslant 2$. Let X be a random variable following a beta-binomial distribution of parameters (α, β). Then

$$\mathbb{E}[X] = N \frac{\alpha}{\alpha + \beta}$$

and

$$\mathrm{Var}(X) = N \frac{\alpha}{\alpha+\beta}\left(1 - \frac{\alpha}{\alpha+\beta}\right)\frac{N+\alpha+\beta}{1+\alpha+\beta} \, .$$

For $p \in [0, 1]$, a binomial distribution of parameters (N, p) has expectation Np and variance $Np(1-p)$. Now, let α, β be positive real numbers satisfying $\frac{\alpha}{\alpha+\beta} = p$. According to proposition 4.1, the beta-binomial distribution of parameters (N, α, β) has then expectation Np and variance $Np(1-p)\frac{N+\alpha+\beta}{1+\alpha+\beta}$. It looks a lot like the mean and variance of the binomial distribution of parameters (N, p), but there is an additional factor in the expression of the variance. This factor is always greater than 1, so that the beta-binomial distribution is more dispersed than the regular binomial distribution. Note that, with fixed N, the beta-binomial family is parametrized by the two variables α and β whereas the binomial family is only parametrized by the number p. A new parameter is the cost for introducing overdispersion to the model. The family of Dirichlet distributions defined below is a generalization of the Beta distributions to higher-dimensional simplexes.

Definition 4.1. Let $d \geqslant 2$ be an integer and let $\alpha_1, \ldots, \alpha_d$ be positive real numbers. The Dirichlet distribution of parameters $(\alpha_1, \ldots, \alpha_d)$ is a probability measure subordinate to Lebesgue's measure on the simplex Δ^{d-1}, of probability density function p_{Dir} given by

$$\forall \mathbf{x} = (x_1, \ldots, x_d) \in \Delta^{d-1}, \quad p_{\mathrm{Dir}(\boldsymbol{\alpha})}(x) = \frac{1}{\mathrm{B}(\alpha_1, \ldots, \alpha_d)} \prod_{j=1}^{d} x_j^{\alpha_j - 1} \, ,$$

where B is the multivariate beta function defined by

$$\mathrm{B}(\alpha_1, \dots, \alpha_d) = \int_{\Delta^{d-1}} \prod_{j=1}^{d} t_j^{\alpha_j} \, \mathrm{d}t = \frac{1}{\Gamma(\alpha_1 + \cdots + \alpha_d)} \prod_{j=1}^{d} \Gamma(\alpha_j) \, .$$

It is convenient for qualitative interpretation to re-parametrize the Dirichlet distribution by letting $\boldsymbol{\theta} \in \Delta^{d-1}$ and $\varphi > 0$ be such that $\boldsymbol{\alpha} = \frac{1}{\varphi}\boldsymbol{\theta}$ and $\sum_j \theta_j = 1$ (so that $\frac{1}{\varphi} = \sum_j \alpha_j$). I found this parametrization in [10], where φ is called "overdispersion parameter". It is easier to evaluate the influence of the multinomial behavior and the overdispersion separately with this choice of parameters. Indeed, $\boldsymbol{\theta}$ is the expectation of the Dirichlet distribution of parameter $\boldsymbol{\alpha}$ and φ is related to its variance, letting φ go to 0 making the Dirichlet distribution of parameter $(\boldsymbol{\theta}, \varphi)$ converge to the Dirac mass at $\boldsymbol{\theta}$, and letting φ go to $+\infty$ making the Dirichlet distribution approach a uniform distribution on the simplex.

A Dirichlet-multinomial distribution with d degrees of freedom is the compound distribution obtained by picking a parameter $p \in \Delta^{d-1}$ of under a Dirichlet distribution and evaluating observing a draw of a multinomial distribution having p as probability weights. The case $d = 2$ reduces to the beta-binomial distribution described above. Thus, the Dirichlet-multinomial distributions family is parametrized by the number of trials $N \in \mathbb{N}$ and the parameter of the Dirichlet distribution $(\boldsymbol{\theta}, \varphi)$. The Dirichlet-multinomial distribution of parameters N, $\boldsymbol{\theta}$ and φ will be denoted by $\mathrm{DMN}(N, \boldsymbol{\theta}, \varphi)$ in the sequel. In the new model for describing fragmentation spectra, each row $(n_{i,j})_{1 \leqslant j \leqslant d})$ of the contingency table follows a Dirichlet-multinomial distribution instead of a multinomial distribution. This model writes

$$\mathcal{M}_{\mathrm{DMN}} = \{\mathrm{DMN}(n_{1,+}, \boldsymbol{\theta_1}, \varphi) \otimes \mathrm{DMN}(n_{2,+}, \boldsymbol{\theta_2}, \varphi),\ (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \in \Delta^{d-1} \times \Delta^{d-1}\} \tag{4.1}$$

It was a choice not to include φ in the parameters of the model. It means that we assume φ to be the same for all molecules, or at least for all the molecules when considering the same number of peaks d. The parameter φ is called overdispersion parameter and the greater φ is, the greater the variance of the multinomial distribution is. There were multiple problems with including φ as a parameter in the model, the biggest being the difficulty to estimate φ for computing an approximation of the maximum likelihood.

Proposition 4.2. The likelihood function of the Dirichlet-multinomial model satisfies the formula

$$L(\mathbf{n}; (N, \boldsymbol{\theta}, \varphi)) = \frac{\Gamma(\varphi^{-1})\Gamma(N+1)}{\Gamma(\varphi^{-1} + N)} \prod_{j=1}^{d} \frac{\Gamma(\varphi^{-1}\theta_j + n_j)}{\Gamma(\varphi^{-1}\theta_j)\Gamma(n_j + 1)} \, . \tag{4.2}$$

where $\boldsymbol{\theta}, \varphi$ are the parameter variables and $\mathbf{n} \in \mathbb{N}^d$ is the outcome and satisfies $\sum_{j=1}^{d} n_j = N$.

The computation leading to expression 4.2 is detailed in annex A. If $\mathbf{n}$ is a random variable following a Dirichlet-multinomial distribution of parameters $(N, \boldsymbol{\theta}, \varphi)$, then its expectation and covariance matrix are given by the expressions

$$\forall j \in \{1, \dots, d\},\ \mathbb{E}[n_j] = N\theta_j \, ,$$

$$\forall i, j \in \{1, \dots, d\}^2,\ \mathbb{E}[n_j] = N(\delta_{i,j}\theta_i - \theta_i\theta_j)\frac{1 + \frac{1}{N\varphi}}{1 + \frac{1}{\varphi}} \, .$$

As in the beta-binomial expression of the variance, the factor $\frac{1+\frac{1}{N\varphi}}{1+\frac{1}{\varphi}}$, which is greater than 1, describes how much the data is overdispersed in comparison with the simple multinomial distribution.

4.2 Formulation of a test statistic

The likelihood function of the Dirichlet-multinomial distribution is not as easy to maximize as the multinomial likelihood function. In fact, there are multiple ways to express the likelihood function, which all could be relevant. for instance, it is possible to eliminate the use of the Gamma function by using the identity $\Gamma(x+1) = x\Gamma(x)$:

$$L(n;(N,\boldsymbol{\theta},\varphi)) = \frac{N\cdot N!}{\prod_{k=0}^{N-1}(\varphi^{-1}+k)}\prod_{j=1}^{d}\frac{\prod_{k=1}^{n_j}(\varphi^{-1}\theta_j+k)}{n_j\cdot n_j!}$$

I worked a lot with this expression at first in the hope of getting polynomial expressions that would be convenient for determining a maximum likelihood. Also, when taking the logarithm, the products turn into sums of logarithms of affine functions of the θ_j. However, the numbers of terms in those sum are the n_j, which are meant to be spectra intensities. Because these spectra intensities can go up to 10^7, my computer would take days to massively evaluate the likelihood function with these expressions. In fact, the determination of a maximum likelihood estimator would imply the use of advanced optimization techniques, such as Newton-Raphson method. There are efficient built-in functions for this purpose in Python's library scipy, but I decided to circumvent the problem of computing the exact maximum likelihood estimator by replacing it with a moment estimator. I would hardly have been able to go forward with the project without doing this, and it is in fact inspired by expression (3.5) of the multinomial statistic.

Let $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ be defined in function of N and the data $\mathbf{n} = (n_{i,j})_{\substack{1\leqslant i\leqslant 2\\ 1\leqslant j\leqslant d}}$ by

$$\hat{\boldsymbol{\theta}}^{(1)}_{i,j} = \frac{n_{i,j}}{n_{i,+}}\,,$$

$$\hat{\boldsymbol{\theta}}^{(1)}_{i,j} = \frac{n_{i,j}+n_{2,j}}{n_{1,+}+n_{2,+}}\,;$$

I defined the DMN statistic T_{DMN} by

$$T_{\mathrm{DMN}} = \log\left(\frac{L\left(\mathbf{n_1};(n_{1,+},\hat{\boldsymbol{\theta}}^{(1)}_{\mathbf{1}},\varphi)\right)L\left(\mathbf{n_2};(n_{2,+},\hat{\boldsymbol{\theta}}^{(1)}_{\mathbf{2}},\varphi)\right)}{L\left(\mathbf{n_1};(n_{1,+},\hat{\boldsymbol{\theta}}^{(0)}_{\mathbf{1}},\varphi)\right)L\left(\mathbf{n_2};(n_{2,+},\hat{\boldsymbol{\theta}}^{(0)}_{\mathbf{2}},\varphi)\right)}\right)\,, \tag{4.3}$$

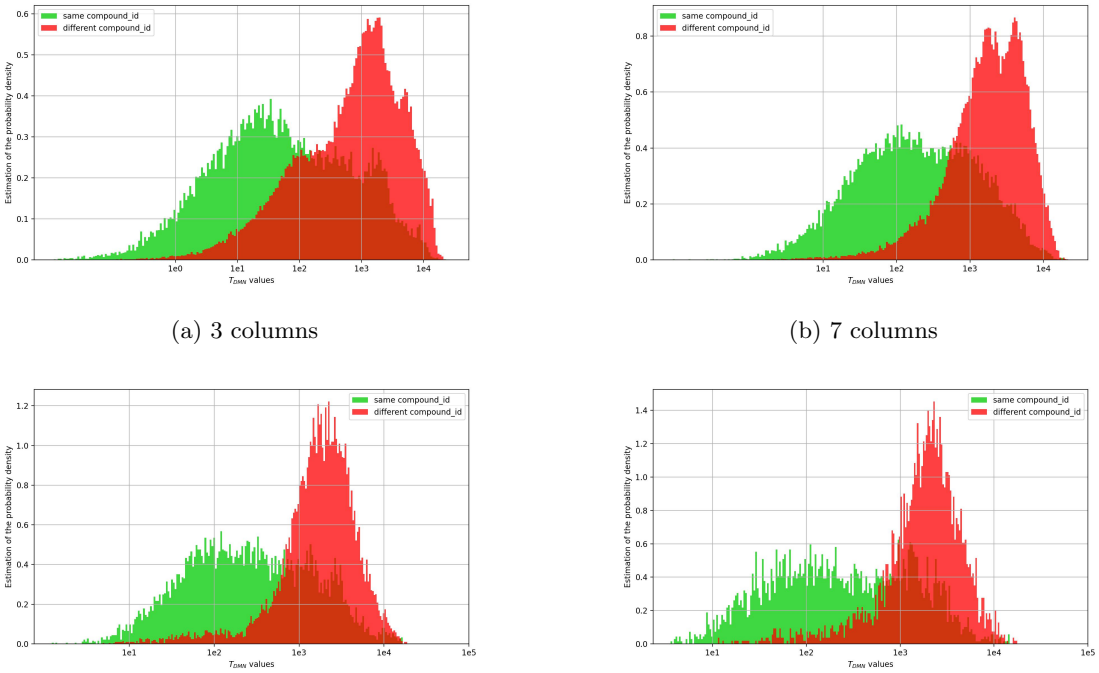
where L is the likelihood function of model $\mathcal{M}_{\mathrm{DMN}}$ given by expression 4.2. For the implementation, it was convenient to express the likelihood with the beta function as follows

$$L(\mathbf{n};(n,\boldsymbol{\theta},\varphi)) = \frac{N\mathrm{B}(N,\varphi^{-1})}{\prod_{j=1}^{d}(n_j\mathrm{B}(n_j,\varphi^{-1}\theta_j))}\,.$$

When plugging this formula into 4.3, the factors only depending on $\mathbf{n_1}, \mathbf{n_2}, N$ and φ simplify and there remains

$$T_{\mathrm{DMN}} = -\sum_{i=1}^{2}\sum_{j=1}^{d}\mathrm{B}(n_{i,j},\varphi^{-1}\hat{\theta}^{(1)}_{i,j}) + \sum_{i=1}^{2}\sum_{j=1}^{d}\mathrm{B}(n_{i,j},\varphi^{-1}\hat{\theta}^{(0)}_{i,j})\,. \tag{4.4}$$

Remember that φ has been left to be chosen. As expression (4.1) defining the DMN model suggests, the same value of φ should be used for all the tables having the same number of columns. In fact, it is a choice I made in the beginning and I also implemented a way to estimate a new φ for each table. Since the results were less convincing in the latter case, I sticked to the original decision of using the same φ for every table. But then, probably some values of φ would yield better results than others. For instance, letting φ go to zero makes the DMN statistic converge towards the multinomial statistic,



(a) 3 columns

(b) 7 columns

(c) 12 columns

(d) 16 columns

Figure 4.1: Repartition of the DMN statistic for tables with same compound id vs different compound id for different number of columns and $\varphi = 10^{-4}$.

so the same results as for the multinomial model are expected for small values of φ. On the other hand, letting φ go to $+\infty$ makes the DMN distribution of parameters $(N, \boldsymbol{\theta}, \varphi)$ go to a multinomial distribution with probability weight chosen uniformly on the simplex for all θ, so that the likelihood function is almost the same when evaluated at $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(0)}$, and the DMN statistic would make almost no distinction between the null and the alternative hypothesis. Thus, we hope there are intermediate values of φ doing better than the multinomial model.

4.3 Results and comparison with the multinomial model

The results that are presented here follow the scheme of the presentation of the multinomial results given in 3.3. When evaluating the models for different values of φ, three ranges of values appeared, which correspond to the three cases I mentioned at the end of 4.2:

Multinomial regime	Intermediate regime	Uniform regime
$\varphi \leqslant 10^{-8}$	$10^{-8} < \varphi \leqslant 10^{-2}$	$\varphi > 10^{-2}$

In the multinomial regime, the results are the same as for the multinomial model and in the uniform regime, the DMN statistic makes little difference between the null and alternative hypothesis, as if the prior Dirichlet distributions introduced in the new model were uniform for both rows of the contingency table. In the intermediate regime, the shapes of the DMN statistic distributions and the type I and II error basically the same for φ comprised between 10^{-5} and 10^{-2}, and I decided to plot the distributions for the particular value $\varphi = 10^{-4}$. Note that the range of values of the DMN statistic differ depending on the value of φ, it ranges from slightly negative values to 10^2 for $\varphi = 10^{-2}$, and it can also take negative values for smaller values of φ, but the upper bound and the values around which the DMN

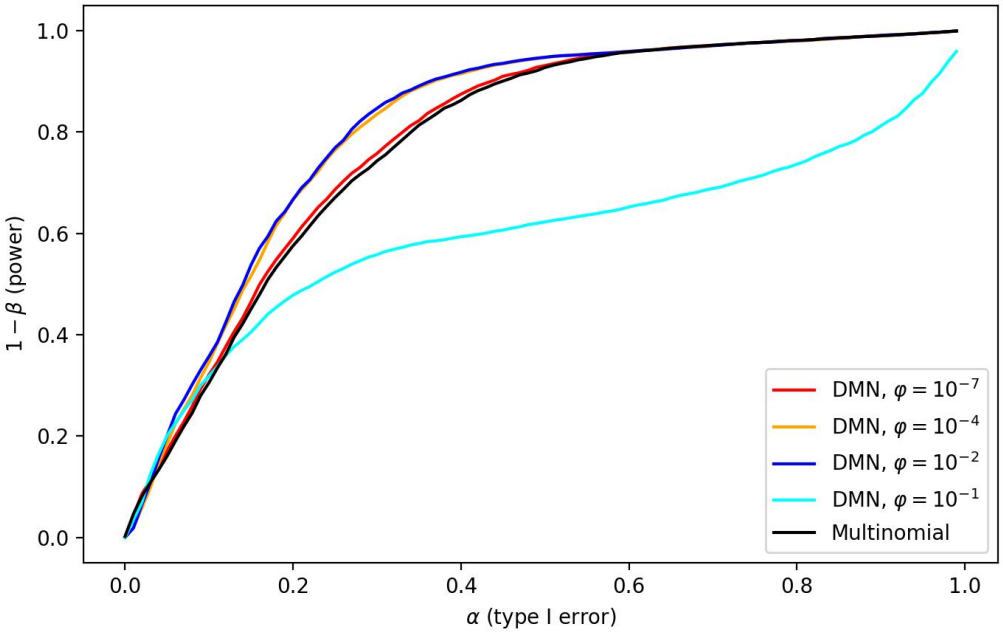


Figure 4.2: ROC curves comparison for 12 columns

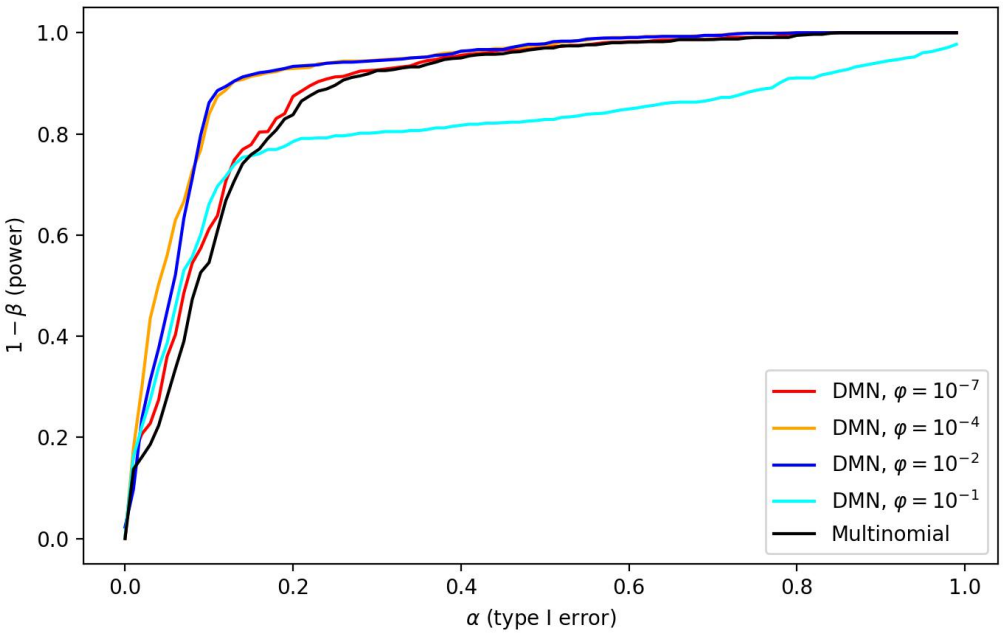


Figure 4.3: ROC curves comparison for 12 columns and match quality > 0.85

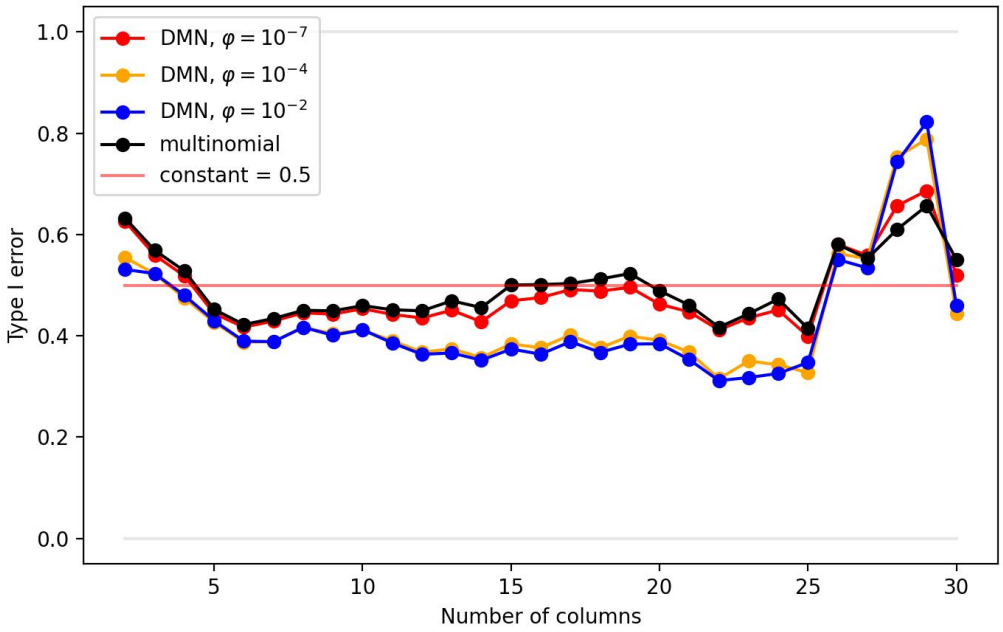


Figure 4.4: Type I error associated to 90 % power with respect to the number of columns

statistic is concentrated increase as φ decreases. It could be envisioned to learn φ from the data in order to get the best possible values of type I and II errors.

Next are, on the same graphic, the ROC curves for tables having 12 columns, for different values of φ and compared to the ROC curve of the multinomial model. We can see that $\varphi = 10^{-7}$ gives results very similar to the multinomial model and that the ROC curves for $\varphi = 10^{-2}$ and $\varphi = 10^{-4}$ almost overlap, while $\varphi = 10^{-1}$ performs clearly worse in terms of error of type II, which is the main concern in the problem of classifying spectra.

In 3.3, I did not take into account the possible mistakes that the commercial algorithm could have made when assigning a compound id to the spectra. There are two interesting ways of testing the results of the model by taking this match quality into account. The first one is to only consider the compound id that have a very high match quality. From the beginning Eliza recommended me to use compounds with a match quality of 0.85 if I wanted to do that. The other interesting choice would be to only use the compound ids that have been assigned a referenced molecule. When plotting the ROC curves with those choices and other number of columns, we get that considering compound ids with high match quality yields significantly better results (see figure 4.3), and only slightly better with compound ids assigned to a referenced molecule. Finally, I included the curves of the estimation of the type I error associated to a power of 90 % with respect to the number of columns.

5 Discussions

5.1 On the evaluation of the performance

The results of the test presented in 3.3 and 4.3 are not significantly better than the previous method used in the MSEI project for classifying molecules. However, it is satisfying to observe that the Dirichlet-multinomial model is doing better than the multinomial model. It means that it was worth

spending time on developing this new model (which was more challenging than the multinomial model). I have to mention that the estimations of t_0 and α in 3.3 and 4.3 could be made more rigorously. I describe an attempt at doing so in annex D. I chose not to include this in the main text because the results are basically the same as those in 3.3 and 4.3. Also, I did not have the occasion to discuss the content of annex D with Guillaume. I can however explain the reasons why I carried on with this approach. First, the data from all the database has been used. It is in general more rigorous when evaluating the performance of a model to split the data in two sets, one of which will be used to estimate the threshold for the test and the other one on which the type I and II errors induced by that threshold are estimated. This introduces independence between the estimation of the threshold and the computation of the errors. Also, it is sometimes good to avoid taking into account part of the database for the same reason, because estimating the threshold on the first group introduces a bias on the evaluation of the errors on the second group due to the fact that it is precisely all the other elements that have been sampled. Second, the model is imperfect because the spectra come from different molecules, and probably there is a correlation within data coming from spectra labelled by the same molecule. Thus, treating all the tables independently could be a problem because the data coming from molecules represented by numerous spectra will dominate everywhere. I discussed a lot with Lili near the end of the internship in order to overcome those two problems. She told me of certain ways of splitting the data into different sets in order to reduce bias and correlation in the estimations. Then, I set up a generalization of the Bernoulli model that would be appropriate to study type I error. In particular, it makes it possible to add confidence intervals to figures 3.5 and 4.4.

5.2 Towards better models

Near the end of the internship, we thought about ways to improve the Dirichlet-multinomial model. The idea is to introduce a prior assumption about the parameters $\boldsymbol{\theta}$ that are the mean of the Dirichlet-multinomial distributions. Indeed, in the case the spectra come from different molecules, it is likely for those molecules to be similar to each other. Moreover, there are often a lot of spectra for the same compound and we could imagine infer information on how much these $\boldsymbol{\theta}$ are close from each other within a family of molecules having the same sum formula. We used Dirichlet distributions again because we already encountered these and we know they can be used to adjust the concentration of a variable taking values in the simplex. The idea would be to, when comparing spectra 1 and spectra 2, to imagine that the parameters $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are drawn at random from Dirichlet distributions of the same parameter. In the case the spectra come from the same molecule, which is the null hypothesis, the draws of $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are correlated so that they are equal. In the alternative hypothesis, $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ are independent random variables. These hypotheses thus write

$$\begin{aligned}
&\text{“}H_1 : (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \sim \mathrm{Dir}(\boldsymbol{\alpha}) \otimes \mathrm{Dir}(\boldsymbol{\alpha}) \\
&\qquad (\mathbf{p_1}, \mathbf{p_2}) \,|\, \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \sim \mathrm{Dir}(\boldsymbol{\theta_1}, \varphi) \\
&\qquad (\mathbf{n_1}, \mathbf{n_2}) \,|\, \mathbf{p_1}, \mathbf{p_2} \sim \mathrm{Multi}(n_{1,+}, \mathbf{p_1}) \otimes \mathrm{Multi}(n_{2,+}, \mathbf{p_2})\text{”}
\end{aligned}$$

and

$$\begin{aligned}
&\text{“}H_0 : (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \sim \mathrm{Dir}(\boldsymbol{\alpha}) \cdot (1,1) \\
&\qquad (\mathbf{p_1}, \mathbf{p_2}) \,|\, \boldsymbol{\theta_1}, \boldsymbol{\theta_2} \sim \mathrm{Dir}(\boldsymbol{\theta_1}, \varphi) \\
&\qquad (\mathbf{n_1}, \mathbf{n_2}) \,|\, \mathbf{p_1}, \mathbf{p_2} \sim \mathrm{Multi}(n_{1,+}, \mathbf{p_1}) \otimes \mathrm{Multi}(n_{2,+}, \mathbf{p_2})\text{”}\,,
\end{aligned}$$

where the parameters are $\boldsymbol{\alpha} \in (\mathbb{R}_+^*)^d$ and $\varphi > 0$. The parameter $\boldsymbol{\alpha}$ hides an overdispersion parameter which again could be adjusted or estimated. Here, the Dirichlet distribution is not a conjugate prior to the Dirichlet-multinomial distribution, so the likelihood function is under integral form. Guillaume suggested me to use a Monte-Carlo method with importance sampling for estimating the integrals appearing in the likelihood ratio for those models, but I did not have the time to implement it. Guillaume also talked to me about generalizations of the Dirichlet-multinomial distribution involving even

more parameters.

In conclusion of this report, here are a few personal impressions. For sure, I have learnt a lot about statistical hypothesis testing and correlated topics. It took me a long time, almost three weeks, to familiarize with the data and the coding environment. The project was very rich and I feel like so much more could have been done. I feel it even more after having written the report and having realized that I left a lot of questions unanswered, including questions I would now know how to deal with. I am nonetheless satisfied by how things turned out. Although the final results cannot be called breathtaking, I am satisfied with how introducing mathematical ideas was able to make the statistical model a bit better. I will remember these three months at the SDSC as a great experience.

References

[1] Böcker, S. (2017). Searching molecular structure databases using tandem MS data: are we there yet? *Current Opinion in Chemical Biology, 36,* 1-6. 10.1016/j.cbpa.2016.12.010

[2] The χ^2 Test of Goodness of Fit. *Ann. Math. Statist., 23 (3),* 315-345. 10.1214/aoms/1177729380

[3] De Vijlder, T. et al. (2018). A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass Spectrometry Reviews, 37,* 607-629. 10.1002/mas.21551

[4] Dührkop, K. et al. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *PNAS, 41,* 12580-12585. 10.1073/pnas.1509788112

[5] Dührkop, K. et al. (2021). Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology, 39,* 462-471. 10.1038/s41587-020-0740-8

[6] Hill, D. W. et al. (2008). Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra. *Analytical Chemistry, 80,* 5574-5582. 10.1021/ac800548g

[7] Hufsky, F., Scheubert, K. & Böcker, S. (2014). New kids on the block: novel informatics methods for natural product discovery. *Natural Products Reports, 31,* 807-817. 10.1039/c3np70101h

[8] Hufsky, F. & Böcker, S. (2017). Mining Molecular Structure Databases: Identification of Small Molecules based on Fragmentation Mass Spectra. *Mass Spectrometry Reviews, 36,* 624-633. 10.1002/mas.21489

[9] Kind, T. et al. (2018). Identidification of small molecules using accurate MS/MS search. *Mass Spectrometry Reviews, 37,* 513-532. 10.1002/mas.21535

[10] La Rosa, P. S. et al (2012) Hypothesis Testing and Power Calculations for Taxonomic Human-Based Human Microbiome Data. *PloS ONE 7 (12),* e52078 10.1371/journal.pone.0052078

[11] Lydersen, S., Fagerland, M. W. & P. Laake (2009). Recommended tests for association in 2×2 tables. *Statistics in Medicine, 28,* 1159-1175. 10.1002/sim.3531

[12] Mehrotra, D. V., Chan, I. S. F. & Berger, R. L. (2003) A Cautionary Note on Exact Unconditional Inference for a Difference between Two Independent Binomial Proportions. *Biometrics, 59,* 441-450. 10.1111/1541-0420.00051

[13] Peironcely, J. E. et al. (2013). Automated Pipeline for de Novo Matabolite Identification Using Mass Spectrometry-Based Metabolomics. *Analytical Chemistry, 85,* 3576-3583. 10.1021/ac303218u

[14] Ruttkies, C., Neumann S. & Posch, S. (2019). Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics, 20,* 376. 10.1186/s12859-019-2954-7

[15] Wilks, S. S. (1938) The Large-sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses *Ann. Math. Statist., 9 (1),* 60-62. 10.1016/S0167-7152(00)00171-1

[16] Wilson J. R. (1989) Chi-square Tests for Overdisperion with Multiparameter Estimates. *Appl. Statist., 38 (3),* 441-453

6 Annex A : Proofs and computations.

Proof. Proof of lemma 3.1.
Let d be a positive integer, let $\lambda_1, \dots, \lambda_d$ be positive real numbers, let $(N_1, N_2 \dots, N_d) \sim \mathcal{P}(\lambda_1) \otimes \mathcal{P}(\lambda_2) \otimes \cdots \otimes \mathcal{P}(\lambda_d)$ and $N = N_1 + N_2 + \ldots + N_d$.
Let k be an integer and $k_1, \dots, k_d$ be d integers adding up to k. The variable N is the sum of the independent Poisson variables $N_1, \dots, N_d$, thus follows a Poisson distribution of parameter $\lambda_1 + \ldots + \lambda_d$, that is denoted by λ in the sequel. Proposition 3.1 derives from the following computation :

$$\begin{aligned}
\mathbb{P}(N_1 = k_1, \dots, N_d = k_d | N = k) &= \frac{\prod_{i=1}^{d} \mathbb{P}(N_i = k_i)}{\mathbb{P}(N = k)} \\
&= e^{\lambda} \frac{k!}{\lambda^k} \prod_{i=1}^{d} e^{-\lambda_i} \frac{\lambda_i^{k_i}}{k_i!} \\
&= \underline{(} e^{\lambda} \prod_{i=1}^{d} e^{-\lambda_i} \underline{)(} \frac{k!}{k_1!\, k_2! \cdots k_d!} \underline{)(} \lambda^{-\sum_{i=1}^{d} k_i} \prod_{i=1}^{d} \lambda_i^{k_i} \underline{)} \\
&= \binom{k}{k_1, k_2, \dots, k_d} \prod_{i=1}^{d} \left(\frac{\lambda_i}{\lambda} \right)^{k_i} .
\end{aligned}$$

□

It is often the case with convenient models (i.e. relying on simple combinations of well-known random variables) that the likelihood function is concave and differentiable in the parametric variables. Thus, it admits a maximum which is attained at parameters making the gradient vanish, and this maximum is unique if the likelihood function is strictly convex. The parameters in which the maximum is attained are called maximum likelihood estimators of the parameters, and the maximum itself is called maximum likelihood. Furthermore, the logarithm of the likelihood function is also concave and vanishes at the same points, so finding the maximum likelihood estimators can be done by differentiating the log-likelihood function instead. A standard method to determine the maximum likelihood in a model is then to compute the maximum likelihood estimators by using differential calculus, and evaluating the likelihood function at the maximum likelihood estimator. In the multinomial model, the parameter variable lives in a simplex so the following lemma is helpful for finding the maximum likelihood estimator. Here I introduce the simplex deprived of its boundary, $\mathring{\Delta}^{d-1} = \Delta^{d-1} \backslash \partial\Delta^{d-1}$, which is a bit more convenient to work with because a function on the simplex can be differentiable everywhere in $\mathring{\Delta}^{d-1}$ whereas there is a problem on the boundary in Δ^{d-1}.

Lemma 6.1. *Let d be a positive integer. Let f be a real-valued function from $\mathring{\Delta}^{d-1}$ to $\mathbb{R}$, of class C^1 on $\mathring{\Delta}^{d-1}$. Let $\tilde{f}$ be an extension of class C^1 of f to an open set of $\mathbb{R}^d$ that contains $\mathring{\Delta}^{d-1}$. Then the gradient of f vanishes at a point $x \in \mathring{\Delta}^{d-1}$ if and only if all the coordinates of $\nabla \tilde{f}(x)$ are equal, i.e.*

$$\frac{\partial \tilde{f}}{\partial x_1}(x) = \frac{\partial \tilde{f}}{\partial x_2}(x) = \cdots = \frac{\partial \tilde{f}}{\partial x_d}(x) \, .$$

Proof. Proof of lemma 6.1.
As subset of the affine hyperplane $\{\sum_{j=1}^d x_j = 1\}$, the set $\mathring{\Delta}^{d-1}$ is a submanifold of $\mathbb{R}^d$. The tangent space of this submanifold is the hyperplane $\{\sum_{j=1}^d = 0\}$ at all $p \in \mathring{\Delta}^{d-1}$. Therefore, for all $p \in \mathring{\Delta}^{d-1}$, $\nabla f(p)$ is the orthogonal projection of $\nabla \tilde{f}$. It follows that $\nabla f(p)$ vanishes if and only if $\nabla \tilde{f}(p)$ is orthogonal to $\sum_{j=1}^d x_j = 0\}$. This is equivalent to being colinear to the vector $(1, 1, \ldots, 1) \in \mathbb{R}^d$, and this concludes the proof because a vector is colinear to $(1, \ldots, 1)$ if and only if all of its coordinates are equal.

A way to make the proof without talking about submanifolds would be to parametrize $\mathring{\Delta}^{d-1}$ by the first $d-1$ coordinates and compute the gradient of the corresponding function from an open subset of $\mathbb{R}^{d-1}$ to $\mathbb{R}$ using the gradient of $\tilde{f}$. □

A standard method to estimate a parameter given the data is to consider the value of this parameter which maximizes the likelihood function. This is called the maximum likelihood estimator of the model.

Definition 6.1. Let $\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ admitting a likelihood function L. Given observation data x, if the function $L(x, \cdot)$ admits a unique maximum on Θ at a parameter value $\hat{\theta}$, then $\hat{\theta}$ is said to be the maximum likelihood estimator of θ. In a mathematical expression, the maximum likelihood estimator of $\hat{\theta}$ of the parameter θ writes

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} L(x, \theta)\,.$$

Maximum likelihood estimators can be useful to compute the log-likelihood statistic because a maximum likelihood is just the evaluation of the likelihood function at the maximum likelihood estimator. Therefore, getting to the quantities $\max_{\theta \in \Theta_0} L(x, \theta)$ and $\max_{\theta \in \Theta_1}$ defining the log-likelihood statistic can be done by determining the maximum likelihood estimators in each of the submodels induced by the null and alternative hypotheses. The computation of the likelihood statistic in model $\mathcal{M}_{\text{MN}}$ essentially relies on proposition 6.1 below, which give the expression of the maximum likelihood estimator in the simple multinomial model. Before stating the proposition, let's introduce some notations. Let $d \geqslant 2$ and $n_+ \geqslant 1$ be integers and let $\mathcal{M}$ be the model defined by

$$\mathcal{M} = \{\text{Multi}(n_+, \mathbf{p}), \quad \mathbf{p} \in \Delta^{d-1}\}\,. \tag{6.1}$$

Let's denote by $\mathbf{n} = (n_1, \ldots, n_d)$ the random variable associated to this model. Note that $\mathbf{n}$ is here implicitly defined as a vector of non-negative integers adding up to n_+.

Proposition 6.1. Given data $\mathbf{n} \in \mathbb{N}^d$ satisfying $\sum_{j=1}^d n_j = n_+$, there exists a maximum likelihood estimator of $\mathbf{p}$ within the model $\mathcal{M}$. It is unique in Δ^{d-1} and, by denoting it $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_d)$, it satisfies

$$\forall j \in \{1, 2, \ldots, d\}, \quad p_j = \frac{n_j}{n_+}\,. \tag{6.2}$$

Proof. Proof of proposition 6.1.
The likelihood function of model $\mathcal{M}$ is given by

$$L(\mathbf{n}, \mathbf{p}) = \binom{n_+}{\mathbf{n}} \prod_{j=1}^d p_j^{n_j}\,. \tag{6.3}$$

where the factors in the form of 0^0 are being considered equal to 1. This can only happen if one of the n_j is equal to zero. Let $\mathbf{n}$ be a vector of non-negative integers summing to n_+. Since the likelihood function is continuous on the simplex, which is compact, it admits a maximum.

Let's first assume that none of the n_j are zero. Then, the likelihood function is positive on $\mathring{\Delta}^{d-1}$ and vanishes on the boundary of the simplex. Thus, any maximum of the likelihood function is reached

in $\mathring{\Delta}^{d-1}$. By taking the logarithm of the likelihood function and shifting it by $\log(\binom{n_+}{\mathbf{n}})$, we can see that maximizing the function $\mathbf{p} \mapsto L(\mathbf{n}, \mathbf{p})$ is equivalent to maximizing over $\mathbf{p} \in \mathring{\Delta}^{d-1}$ the quantity

$$\mathcal{L}(\mathbf{p}) = \sum_{j=1}^{d} n_j \log(p_j) \,.$$

The function $\mathcal{L}$ is of class C^1 on $\mathring{\Delta}^{d-1}$ and extends to a neighborhood of $\mathring{\Delta}^{d-1}$ to a function $\tilde{\mathcal{L}}$ which is of class C^1 and satisfies

$$\forall j \in \{1, \dots, d\}, \forall \mathbf{p} \in \mathring{\Delta}^{d-1}, \, \frac{\partial \tilde{\mathcal{L}}}{\partial p_j}(\mathbf{p}) = \frac{n_j}{p_j} \,.$$

Applying lemma 6.1, it follows that the $\nabla \mathcal{L}$ vanishes at a point $\mathbf{p}$ if and only if there is a real number a such that

$$\forall j \in \{1, \dots, d\}, \, n_j = a p_j \,.$$

Since $\mathbf{n}$ has sum n_+ and $\mathbf{p}$ has sum 1, a must be equal to n_+. Therefore, the vector $\hat{\mathbf{p}}$ defined by

$$\forall j \in \{1, \dots, d\}, \, \hat{p}_j = \frac{n_j}{n_+}$$

is the unique element of $\mathring{\Delta}^{d-1}$ at which$\nabla \mathcal{L}$ vanishes. The function $\mathcal{L}$ is a positively-weighted sum of logarithms of the coordinates of $\mathbf{p}$, so it is strictly concave and differentiable. Consequently, it admits a unique maximum at which $\nabla \mathcal{L}$ is equal to zero. We deduce that $\hat{\mathbf{p}}$ is the unique point at which $\mathcal{L}$ reaches a maximum, i.e. it is the maximum likelihood estimator in model $\mathcal{M}$.

Let's now deal with the case where some of the n_j are zero. Assume there is $j_0 \in \{1, \dots, d\}$ such that $n_{j_0} = 0$.

Let $\mathbf{p} \in \Delta^{d-1}$. Consider the vector $\bar{\mathbf{p}}$ defined

$$\bar{p}_{j_0} = \delta_{j, j_0} \frac{p_j}{1 - p_{j_0}} \,.$$

Then $\bar{\mathbf{p}}$ belongs to Δ^{d-1} and

$$\frac{L(\mathbf{n}, \bar{\mathbf{p}})}{L(\mathbf{n}, \mathbf{p})} = \prod_{j \neq j_0} \left(\frac{\bar{p}_j}{p_j} \right)^{n_j} = \prod_{j \neq j_0} \left(\frac{1}{1 - p_{j_0}} \right)^{n_j} > 1 \,.$$

Thus, the likelihood function reaches its maximum on $\{\mathbf{p} \in \Delta^{d-1} / p_{j_0} = 0\}$. By using the same argument for all the indexes j such that $n_j = 0$, the problem reduces to maximizing $L(\mathbf{n}, \cdot)$ on the smaller simplex simplex $\Delta^J = \{\mathbf{p} \in \Delta^{d-1} / \forall j \notin J, \, p_j = 0\}$ where $J = \{j \in \{1, \dots, d\} / n_j > 0\}$. As the likelihood function writes

$$\mathrm{L}(\mathbf{n}, \mathbf{p}) = \prod_{j \in J} p_j^{n_j} \,,$$

this reduces to the case where none of the n_j are zero, and we deduce that there exist a unique maximum $\hat{p}$, satisfying

$$\forall j \in J, \, \hat{p}_j = \frac{n_j}{n_+}$$

and

$$\forall j \notin J, \, \hat{p}_j = 0 \,.$$

Since $n_j = 0$ for $j \notin J$, it follows that $\hat{p}$ satisfies expression (6.2) and the proof is complete.

□

As the random variable described by the multinomial model $\mathcal{M}_{\text{MN}}$ introduced in 3.1 is the product of two independent multinomial random variable, Proposition 3.1 is deduced easily from proposition 6.1.

Proof. Proof of proposition 3.1.
Let $\mathbf{n_1}, \mathbf{n_2}$ be observation data for model $\mathcal{M}_{\text{MN}}$. The likelihood function in model $\mathcal{M}_{\text{MN}}$ writes

$$L(\mathbf{n_1}, \mathbf{n_2}, \mathbf{p_1}, \mathbf{p_2}) = \tilde{L}(\mathbf{n_1}; (n_{1,+}, \mathbf{p_1})) \cdot \tilde{L}(\mathbf{n_2}; (n_{2,+}, \mathbf{p_2}))$$

where $\tilde{L}$ is the likelihood function of the simple multinomial model defined by (6.1). By proposition 6.1, for $i \in \{1, 2\}$, $\mathbf{p} \mapsto \tilde{L}(\mathbf{n_i}; (n_{i,+}, \mathbf{p}))$ reaches a maximum at $\hat{\mathbf{p}}_\mathbf{i}^{(1)} = (\hat{p}_{i,j}^{(1)})_{1 \leqslant j \leqslant d} \in \Delta^{d-1}$ defined by

$$\forall j \in \{1, \dots, d\},\ \hat{p}_{i,j}^{(1)} = \frac{n_{i,j}}{n_{i,+}}\,.$$

The subset of $\Delta^{d-1} \times \Delta^{d-1}$ corresponding to the alternative hypothesis in model $\mathcal{M}_{\text{MN}}$ is dense, so the maximum likelihood under the alternative hypothesis is the value of the likelihood function at point $(\hat{\mathbf{p}}_\mathbf{1}^{(1)}, \hat{\mathbf{p}}_\mathbf{2}^{(1)})$. In order to get the maximum likelihood under the null hypothesis, we will use proposition 6.1 again, but this time by restricting model $\mathcal{M}_{\text{MN}}$ to the subset of parameters corresponding to the null hypothesis. In fact, if k_1, k_2 are positive integers and $\mathbf{p} \in \Delta^{d-1}$, then the sum of independent multinomial distributions of parameters $(k_1, \mathbf{p})$ and $(k_2, \mathbf{p})$ is a multinomial distribution of parameters $(k_1 + k_2, \mathbf{p})$. It follows that there is a constant $C(\mathbf{n_1}, \mathbf{n_2})$ such that for all $\mathbf{p} \in \Delta^{d-1}$,

$$L(\mathbf{n_1}, \mathbf{p}, \mathbf{n_2}, \mathbf{p}) = C(\mathbf{n_1}, \mathbf{n_2}) \tilde{L}(\mathbf{n_1} + \mathbf{n_2}; (n_{1,+} + n_{2,+}, \mathbf{p}))\,.$$

By proposition 6.1, the likelihood function reaches a maximum under $\mathcal{H}_0$ for $\mathbf{p_1} = \mathbf{p_2} = \hat{\mathbf{p}}^{(0)}$ such that

$$\hat{p}_{i,j}^{(0)} = \frac{n_{1,j} + n_{2,j}}{n_{1,+} + n_{2,+}}\,.$$

Plugging $(\hat{p}_{i,j}^{(1)})_{\substack{1 \leqslant i \leqslant 2 \\ 1 \leqslant j \leqslant d}}$ and $(\hat{p}^{(0)})_{1 \leqslant j \leqslant d}$ into the likelihood function and taking twice the logarithm of the ratio as in expression (3.3) yields (3.5), and consequently (3.4), which concludes the proof. □

Lastly, I included the computation justifying the expression of the likelihood function of the Dirichlet-multinomial model and a result similar to proposition 6.1 describing maximum likelihood estimators in the Dirichlet-multinomial model.

Proof. Proof of formula (4.2). In this proof, a bayesian notation is adopted so that the expression $p_{\mathcal{L}}(x|\theta)$ denotes evaluation at x of the probability mass function or density of a variable following the distribution $\mathcal{L}$ of parameter θ. The Dirichlet-multinomial distribution probability masses satisfy by definition

$$p_{DMN}(\mathbf{x}|N, \alpha) = \int_{\theta \in \Delta^{d-1}} p_{\text{Multi}}(\mathbf{x}|N, \theta) p_{Dir}(\theta|\alpha)\, \mathrm{d}\theta\,.$$

It follows that

$$p(\mathbf{x}|N, \alpha) = \int_{\theta \in \Delta^{d-1}} \binom{N}{\mathbf{x}} \prod_{i=1}^{d} (\theta_i^{x_i}) \times \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{d} (\theta_i^{\alpha_i - 1})\, \mathrm{d}\theta\,,$$

where B is the multivariate beta function, so that $\mathrm{B}(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$. Thus,

$$p(\mathbf{x}|N, \alpha) = \binom{N}{\mathbf{x}} \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int_{\theta \in \Delta^{d-1}} \prod_{i=1}^{d} (\theta_i^{x_i + \alpha_i - 1})\, \mathrm{d}\theta\,,$$

and we recognize the density of the Dirichlet distribution of parameters $(x_i+\alpha_i)_{1\leqslant i\leqslant d}$ in the inteegral, up to the normalizing factor $B(\mathbf{x}+\alpha)$. Finally,

$$p(\mathbf{x}|N,\alpha)=\binom{N}{\mathbf{x}}\frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\frac{\prod_i\Gamma(x_i+\alpha_i)}{\Gamma(\sum_i(x_i+\alpha_i))}\,.$$

Replacing the multiomial coefficient using the gamma expression of factorials, and replacing $\sum_i x_i$by N yields

$$p(\mathbf{x}|N,\alpha)=\frac{\Gamma(N+1)\Gamma(\sum_i\alpha_i)}{\Gamma(N+\sum_i\alpha_i)}\frac{\prod_i\Gamma(x_i+\alpha_i)}{\prod_i\Gamma(x_i+1)\prod_i\Gamma(\alpha_i)}\,.$$

hence the result. □

Proposition 6.2. The simple Dirichlet-multinomial model admits a maximum likelihood estimator $(\hat{\theta}_1,\dots,\hat{\theta}_d)\in\Delta^{d-1}$, which satisfies

$$\forall j\in\{1,\dots,d-1\}, f(\hat{\theta}_j,X_j)=f_\varphi(\hat{\theta}_{j+1},X_{j+1})\,,$$

where the function f is defined by

$$f:\begin{array}{ccc}\mathbb{R}\times\mathbb{N} & \to & \mathbb{R}\\ (\theta,n) & \mapsto & \sum_{k=0}^{n-1}\frac{1}{\theta+k\varphi}\end{array}$$

Proof. Let L be the likelihood function, of which the expression is

$$L(X,\theta)=\frac{\Gamma(N+1)}{\prod_j\Gamma(X_j)}\frac{\Gamma(\frac{1}{\varphi})}{\Gamma(N+\frac{1}{\varphi})}\prod_{j=1}^{d}\left[\frac{\theta_j}{\varphi}(\frac{\theta_j}{\varphi}+1)\dots(\frac{\theta_j}{\varphi}+X_j-1)\right]\,. \tag{6.4}$$

This expression vanishes if one of the θ_j vanishes while $X_j>0$. Let's first assume that none of the X_j vanish. Then L vanishes on $\{\theta\in\Delta^{d-1}/\exists j\in\{1,\dots,d\},\,\theta_j=0\}$, which is also the boundary of Δ^{d-1} as a submanifold of $\mathbb{R}^d$. Let $\mathring{\Delta}^{d-1}=\Delta^{d-1}\backslash\partial\Delta^{d-1}$ be the simplex deprived of its boundary. The function L is continuous on the compact Δ^{d-1} so it admits a maximum on this set, and L is positive on $\mathring{\Delta}^{d-1}$ and vanishes on $\partial\Delta^{d-1}$, so that the maximum belongs to $\mathring{\Delta}^{d-1}$. In the right-hand side of formula (6.4), only the product indexed on j depends on θ and by taking the logarithm and taking $\frac{1}{\varphi^N}$ out of the expression, we see that maximizing L over θ is the same as maximizing the function Λ defined as follows on $\mathring{\Delta}^{d-1}$:

$$\Lambda(\theta)=\sum_{j=1}^{d}\sum_{k=0}^{X_j-1}\log(\theta_j+k\varphi)\,.$$

Thus, for all $j\in\{1,\dots,d\}$,

$$\frac{\partial\Lambda}{\partial\theta_j}=\sum_{k=0}^{X_j-1}\frac{1}{\theta_j+k\varphi}\,.$$

Let's call $\hat{\theta}$ a value in which Λ reaches a maximum. By proposition 6.1, we get that $\nabla\Lambda(\hat{\theta})$ is collinear to $(1,...,1)$, which is equivalent to be orthogonal to the space spanned by $(e_j-e_{j+1})_{1\leqslant j\leqslant d-1}$, where $(e_1,\dots,e_d)$ is the canonical basis of $\mathbb{R}^d$. Thus

$$\forall 1\leqslant j\leqslant d-1, \langle\nabla\Lambda(\hat{\theta}),e_j-e_{j+1}\rangle=0$$

so that

$$\forall 1\leqslant j\leqslant d-1, f_\varphi(\hat{\theta}_j,X_j)=f_\varphi(\hat{\theta}_j,X_{j+1})\,. \tag{6.5}$$

Let $t\in\mathbb{R}_+^*$. For all $n\in\mathbb{N}$, the function $f_\varphi(\cdot,n)$ is decreasing, $f_\varphi(0^+)=\infty$ and $f_\varphi(+\infty)=0^+$. Therefore, $f_\varphi(\cdot,n)$ is a bijection from $(0,+\infty)$ to $(0,+\infty)$ of which we call $g_\varphi(\cdot,n)$ the inverse, so

that $g_\varphi(t,n)$ is the unique solution of the equation $f_\varphi(x,n)=t$ for the variable x. Now, $g_\varphi(t)=\sum_{j=1}^d g_\varphi(\cdot,X_j)$ is the sum of d decreasing functions and is therefore decreasing, $g_\varphi(0^+)=+\infty$ and $g_\varphi(+\infty)=0^+$. Hence, there exists a unique $\hat{t}\in\mathbb{R}_+^*$ such that $g_\varphi(\hat{t})=1$, and by setting $\hat{\theta}_j=g_\varphi(\hat{t},X_j)$, $\hat{\theta}$ is the unique element of $\mathring{\Delta}^{d-1}$ satisfying condition (6.5), and it is therefore the maximum likelihood estimator for the Dirichlet-multinomial model.
The case where one or more of the X_j is equal to 0 remains to be dealt with. In that case the likelihood doesn't depend on the θ_j for the indexes j such that $X_j=0$. Let J be the set of indexes such that $X_j\neq 0$. Then,

$$L(X,\theta)=\frac{\Gamma(N+1)}{\prod_j\Gamma(X_j)}\frac{\Gamma(\frac{1}{\varphi})}{\Gamma(N+\frac{1}{\varphi})}\prod_{j\in J}\left[\frac{\theta_j}{\varphi}(\frac{\theta_j}{\varphi}+1)\dots(\frac{\theta_j}{\varphi}+X_j-1)\right].$$

For all $j\in\{1,\dots,d\}\backslash J$, let's assign a value in $[0,1)$ to θ_j. Thus, $(\theta_j)_{j\in J}$ belongs to the simplex of dimension $\mathrm{Card}(J)-1$ defined by

$$\Delta^J(1-\sum_{j\notin J}\theta_j)=\{(\theta_j)_{j\in J}\in\mathbb{R}^J/\sum_{j\in J}\theta_j=1-\sum_{j\notin J}\theta_j\}.$$

By the argument for the case of all X_j being positive, there is a $(\hat{\theta}_j)_{j\in J}$ in $\mathring{\Delta}^J(1-\sum_{j\notin J}\theta_j)$ that realizes a maximum of $L(X,\cdot)$ on this set, the values of $\theta_j, j\notin J$ being fixed. This vector satisfies

$$\forall j\in J,\ \hat{\theta}_j=g_\varphi(\tilde{t},X_j)$$

where $\tilde{t}$ is the unique positive real number such that $\sum_{j\in J}g_\varphi(\tilde{t},X_j)=1-\sum_{j\notin J}\theta_j$, with the same notations as above. By monotony of $g_\varphi(\cdot,n)$ for all $n\geqslant 0$, it follows that the bigger $1-\sum_{j\notin J}$ is, the bigger the $\hat{\theta}_j, j\in J$ are, and $L(X,\cdot)$ is a polynomial with non-negative coefficients so it is non-decreasing in every $\theta,\theta\notin J$. We deduce that if there is a maximum, it must be reached at a point satisfying $\forall j\notin J,\theta_j=0$. As a conclusion, there is still a unique maximum likelihood estimator, defined by the formula of the proposition for the indexes j such that $X_j\neq 0$ and satisfying $\theta_j=0$ when $X_j=0$. □

7 Annex B : More details on mass spectrometry

In biology, the field of metabolomics aims to study cellular behaviors by identifying the chemical compounds, called metabolites, that are produced by the operation of the cell. A single sample can contain thousands of different and potentially unknown metabolites. Thus, advanced molecular analysis techniques such as mass spectrometry are needed in order to run metabolomic studies.

A mass spectrometer uses electromagnetic interactions technology to measure with extreme precision mass-to-charge ratios of ionized chemical compounds. The precision is such that it is possible to infer from the measures the sum formulas of the initial molecules that were introduced in the machine. Since the sum formula is not enough to identify a molecule, mass spectrometers are used in combination with chromatography and often feature molecular fragmentation techniques to get further information by measuring mass-to-charge ratios of fragments of the same molecule. This method of fragmentation is called tandem mass spectrometry, or MS^n if it is performed many times in a row. This generates a large amount of data, and algorithms are needed to analyze the data and identify molecules from the obtained spectra. This is not an easy task because the mechanisms of fragmentation are very random, not completely understood and depend on experimental parameters that are virtually impossible to model. Thus, the algorithms that perform the best make large use of machine learning methods and are still unable to identify a lot of metabolites, especially those which aren't referenced in any database.

Mass spectrometry technology is always coupled with chromatography to separate the compounds beforehand. Chromatography consists in making the initial sample migrate through a solvent, thus

separating the compounds having different migration times, also called retention times. This way, among the thousands of compounds that are in the initial sample, up to a few dozens go through the spectrometer at the same time. Without this treatment, the concentration of each compound would bee too low for the results to be significant. After that, the molecules are ionized in order to make them able to react magnetic interaction, which is at the core of the measuring process. Most of the spectrometers use ESI (ElectroSpray Ionization) as ionization technique, which in short consists in putting neutral molecules into a charged solvent and making the solvent evaporate to transfer the charge to the molecules. It is to note that the outcome of ionization depends on the power of the ionization, the concentration of the ionized compound and the compound itself. It is in fact very difficult to describe the behaviour of ionization and the effect it has on the outcome of the measure, especially peak intensity. An other step of the experiment that is to be mentioned is fragmentation. This step consists in isolating molecules having the same mass-to-charge ratio at the first measurement, and randomly breaking bonds in those molecules to get a new set of molecular fragments which will again be measured. Several techniques are used in practice for this step, depending on the spectrometer. A common one is CID (Collision-Induced Dissociation), where the molecules to be fragmented are accelerated to make them collide with each other. The outcome of fragmentation is random, but it is performed on a large number of molecules, which justifies the use of statistical tools to analyse spectra. The spectrum corresponding to the first measurement is called MS1, and each MS1 gives birth to several MS2 spectra, each of those being the fragmentation spectrum corresponding to one peak of the MS1. In sum, mass spectrometry allows to identify molecules relying on their retention times in the chromatography, their mass-to-charge ratio, from which the sum formula can be derived, and their fragmentation spectra, which contain information on the molecular structure.

8 Annex C : Summary of what has been implemented

The language and interface I used were Python and Jupyter notebooks. The libraries that I needed the most were pandas for dealing with the data, numpy for array manipulation and matplotlib for plotting graphics. The following list is an inventory of the programs I engaged in.

1. Extraction of the data and conversion from peaks mz and intensity values in the form of csv files to contingency tables as explained in 3.1.
2. Computation of the multinomial statistic T_{LL} given by (3.4) from the contingency tables.
3. Histograms plots showed in figure 3.3, meant to observe how well the multinomial model is able to separate the null and alternative hypotheses.
4. Histogram plots showing the distribution of all peak intensities, and separation depending on whether the peak is part of one or more contingency tables. This was to evaluate whether or not peaks due to noise were likely appear often in the contingency tables. I concluded from this experiment that there is a negligible amount of noise peaks appearing in the contingency tables.
5. In the MSEI project, an other test statistic called similarity index, has been used before. I combined it with the multinomial statistic into a two-dimensional statistic. The multinomial statistic gives information on how well the intensities of peaks of identical mz correspond to each other and the similarity index indicates how well the mz lists correspond to each other. I considered using a linear separator for splitting the plane into two regions and accept or reject the null hypothesis depending on which the two-dimensional statistic lies in. I did not go forward with this approach to the end because it did not seem to perform significantly better than using only the multinomial statistic.
6. Combination of the study of noise with histograms and the similarity index. By using a bayesian score, it is possible to skew the similarity index in order to make the peaks due to noise matter less. Once again, this did not lead to promising results, so I did not include it in the final report.

7. Different programs for computing the likelihood function of the Dirichlet-multinomial model and also the maximum likelihood estimator. After running these programs on small portions of the data, I concluded that using the built-in log-beta function of Python's scipy library was the most convenient way of computing the likelihood function. Furthermore, using the exact maximum likelihood a lot of time, which motivated the design of the DMN statistic defined in (4.3).

8. Computation of the DMN statistic with given φ, using formula (4.3), from contingency tables.

9. Estimation of a threshold t_0 for the multinomial statistic and the DMN statistic with different values of φ. Estimation of the resulting type I or type II error. This led to the ROC curves from figures 3.4, 4.2 and 4.3, as well as the curves showing type I error in function of the number of column for a threshold corresponding to 90 % power from figures 3.5 and 4.4.

10. Splitting of the data set into smaller sets for reducing bias and correlation in the estimation of t_0 and the test errors. It consists in picking at random a "learning group" consisting in around 40 % of the data, that is used to estimate a threshold t_0, and then use an other 10 % of the data to estimate the resulting type I and II errors.

11. Use a generalization of the Bernoulli model for taking into account potential correlation inherent to the data. Indeed, we can imagine that the test statistic presents a positive correlation within spectra having the same compound id. This led to graphics similar to figures 3.5 and 4.4. This does not make the results significantly different, but the more rigorous approach made me gain confidence in the results that I obtained at the end of 4.1.

9 Annex D: Improvements on the method for evaluating the performance of the models

The problem is a following: we have a certain number of contingency tables (each one corresponding to the comparison of two spectra), and let's say that we would like the test to have a type I error of $\alpha \in (0,1)$. We would like to estimate a threshold t_0 for the test statistic that gives approximately a type I error of α from a learning set. The learning set is composed of tables of couples of spectra having same compound id, each table giving rise to a test statistic X, and we denote the observations of this statistic by $(X_i)_{1\leqslant i\leqslant N}$. We want an estimator of t_α which satisfies $\mathbb{P}(X > t_\alpha) = \alpha$. A way to do this is by sorting the variables X, let's denote $(X^{(i)})_{1\leqslant i\leqslant N}$ the permutation of $(X_i)_{1\leqslant i\leqslant N}$ such that $(X^{(i)})$ is non-decreasing. Then, we set $\hat{t}_\alpha = X^{\lfloor N(1-\alpha)\rfloor}$ (which is the $(1-\alpha)$-quantile of the statistical series $(X_i)_{1\leqslant i\leqslant N}$.In fact, it is a lot more convenient to estimate $\mathbb{P}(X > t_0)$ given t_0: it is the expectation of the Bernoulli variable $\mathbf{1}_{X>t_0}$, and the maximum likelihood estimator is the average $\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{X_i>t_0}$. The estimator $\hat{t}_\alpha$ of which the construction is given above is the value that makes the Bernoulli estimator of $\mathbb{P}(X_i > \hat{t}_\alpha)$ equal to α. This is the least sophisticated way to estimate t_α and does not take into account the possible correlations between the different tables. For instance, there can be a lot of tables computed from spectra having a certain compound id, thus this particular compound id will weigh more in the computation of the estimator.

What follows is an approach at estimating t_α by taking into account the correlation between tables issued from the same compound id. We denote this time by $X_1^i, \ldots, X_{n_i}^i$ the random variables associated to tables from the same compound id, where $i \in \{1, \ldots, k\}$ are the indexes of the compound ids. We still assume that all the X_n^i are independent, but they are identically distributed only within the same compound id (indexed by i). Thus, there is for each i a t_i that is the $(1-\alpha)$-quantile of the distribution of X^i. Now let's assume that t_i is chosen at random and parametrizes the distribution of X^i (which makes X^i a compound distribution, conditioned by the random value of t_i). So here, we come back to saying that the X_n^i are identically distributed random variables, but they are no longer considered independent because for each i there could be a different t_i. Here, we want to estimate a value t such that $\mathbb{P}(X > t) = \alpha$ (where X is a random variable having the same distribution as the

common distribution of the X_k^i). We can first reverse the problem: let $t \in \mathbb{R}$ be fixed and say we want an estimator of the Bernoulli variable $\mathbf{1}_{X>t}$. Not knowing anything about the distribution of t_i, we can still write

$$\mathbb{P}(X > t) = \int \mathbb{P}(X > t|t_i)p(t_i)\,\mathrm{d}\mu(t_i)\,.$$

where $\mathrm{d}\mu$ is the probability distribution of t_i. In the design of the model, t_i is sampled (in the form of a random determination through the spectra coming from the same compound id), and it makes sense to use a Monte Carlo-type estimator of this integral, that is

$$\hat{\mathbb{P}}(X > t) = \frac{1}{k}\sum_{i=1}^{k} \hat{\mathbb{P}}(X > t|t_i)\,.$$

However, Monte-Carlo estimations can be made better by introducing weighting the terms that are summed in order to reduce the variance of the estimator. This is closely related to the notion of importance sampling, which is for computing integrals faster with the Monte-Carlo method. Putting weights in the previous estimator writes as

$$\hat{p} = \frac{1}{\sum_i w_i}\sum_{i=1}^{k} w_i\hat{\mathbb{P}}(X > t|t_i)\,,$$

where the w_i should be chosen to minimize the total variance of the estimator $\hat{\mathbb{P}}(X > t)$. The estimators that we have of $\mathbb{P}(X > t|t_i)$ write as averages of n_i independent Bernoulli variables of parameter p_i where p_i is the real $\mathbb{P}(X > t|t_i)$. By using the law of total variance, it is possible to express the variance of the weighted Monte-Carlo estimator $\hat{p}$ in the form of

$$\mathrm{Var}(p) = \frac{1}{(\sum_{i=1}^{k} w_i n_i)^2}\sum_{i=1}^{k}(w_i^2(a \cdot n_i + b \cdot n_i^2))$$

where $a = \mathbb{E}[p_i](1 - \mathbb{E}[p_i])$ and $b = \mathrm{Var}(p_i)$. By a computation similar to the one leading to lemma 6.1, minimizing the sum while letting $\sum_{i=1}^{k} w_i n_i$ induces the choice of weights

$$w_i = \frac{1}{a \cdot n_i + b \cdot n_i^2}\,.$$

Since a and b can be expressed in function of the moments of p_i, they can be estimated by a method of moments. This method can easily be adapted for estimating type II error given a threshold, and we even have access to an expression of the variance of the estimator, which makes it easy to compute confidence interval with the central limit theorem. However, we still have not answered the question of estimating a threshold. If we know the weights w_i, we can do the same as before, ordering the values X_i^k of the statistic into $X^{(1)}, \ldots, X^{(N)}$, but this time they are associated to weights $w^{(i)}$, and the value $\hat{t}_\alpha$ that makes the weighted Monte-Carlo estimator equal to α satisfies $\hat{t}_\alpha = X^{(i_0)}$ where

$$\sum_{i=1}^{i_0} w^{(i)} <= \alpha < \sum_{i=1}^{i_0+1} w^{(i)}\,.$$

When implementing this method, the estimators of a and b depend on t, and $\hat{t}_\alpha$ depend on a and b, so I started by settings all the weights w_i to 1, estimated a first approximation of t with the initial Bernoulli threshold estimator, and then estimated a and b from this value of t, in the fashion of a fixed point method. I tried doing more iterations, and it was converging in all the cases I looked at, towards a limit that was really close to the value obtained after the first iteration, so I stuck to one iteration. I then kept the estimated values of a and b to plug them in the computations of type I and type II error in the 10 % of the database I picked for this purpose. As a result, we have the following curves:

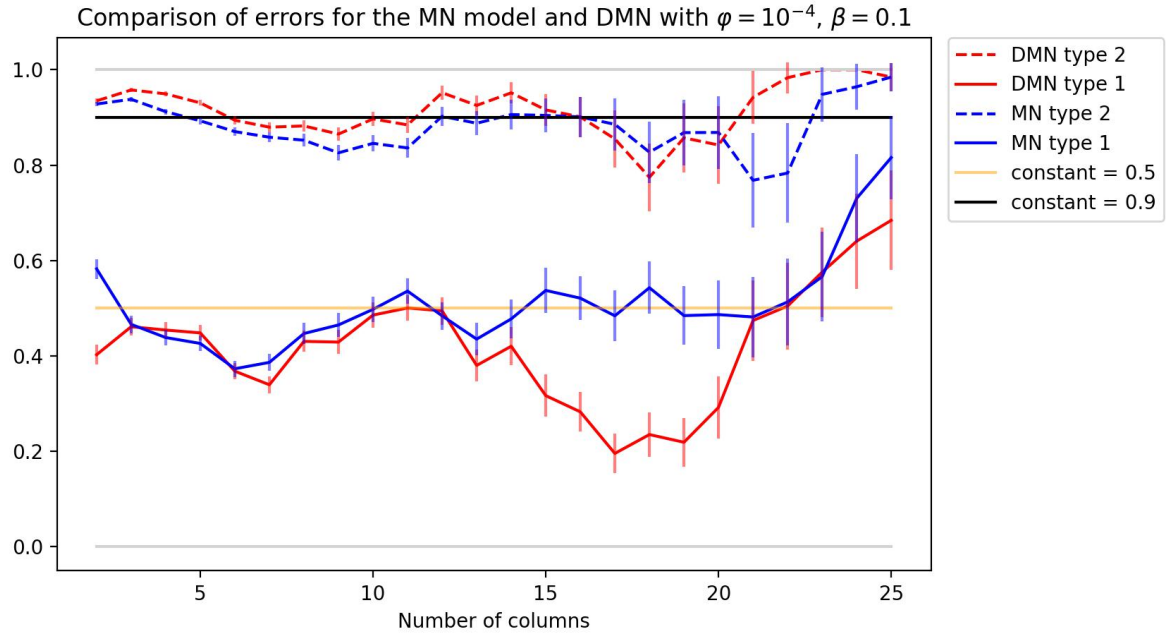


Figure 9.1: With only $\varphi = 10^{-4}$ for a first combination of subset of the database, chosen at random.

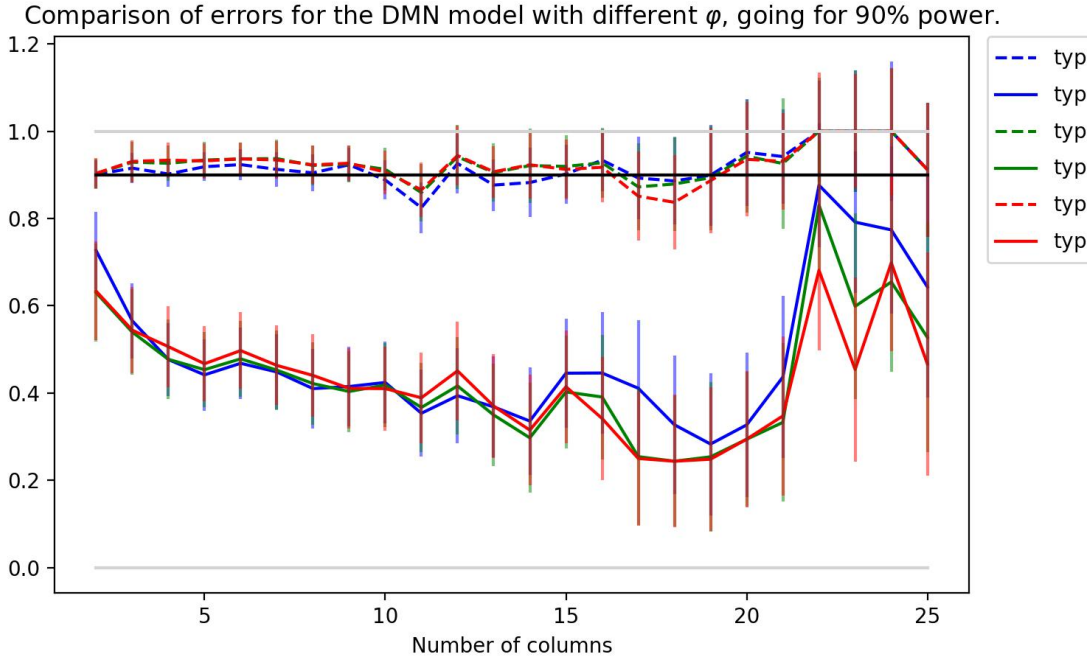


Figure 9.2: With different values of φ for the same combination of sets.

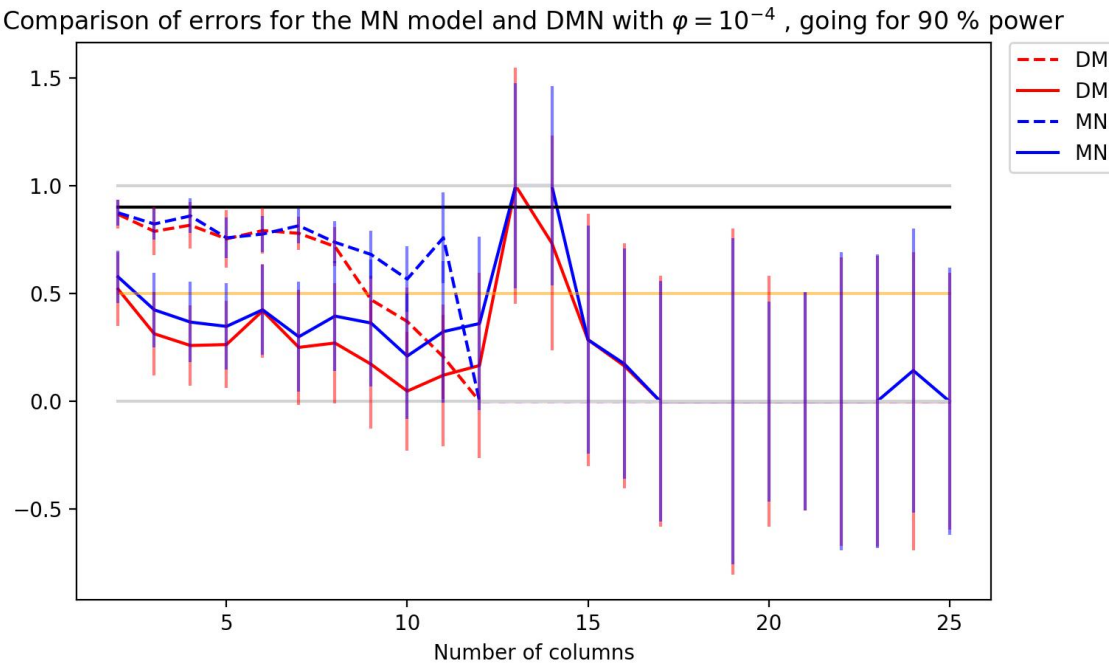


Figure 9.3: The same as figure 9.1 the first plot for a different combination of sets.