Statistical methods for comparing mass spectra

Malo Hillairet
Supervisor : Guillaume Obozinski

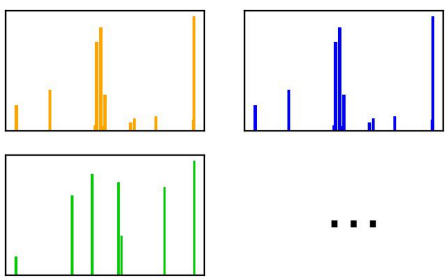
Lausanne, Switzerland

Classifying molecules from mass spectrometry data

Mass spectrometry

- MSEI project : Sasa Bjelić, Lilian Gasser, Eliza Harris, Guillaume Obozinski
- Measures molecular mass with great precision (10 ppm)
- Molecular mass $\Rightarrow$ Sum formula (ex : m/z $= 146.12 \Rightarrow C_6H_{10}O_4$)
- Fragmentation spectra $\Rightarrow$ Tell apart molecules of same formula

MS data



Classification

...	...	...
Acetone	C_3H_6O	...
Glutaric acid	C_5H_804	...
...	...	...

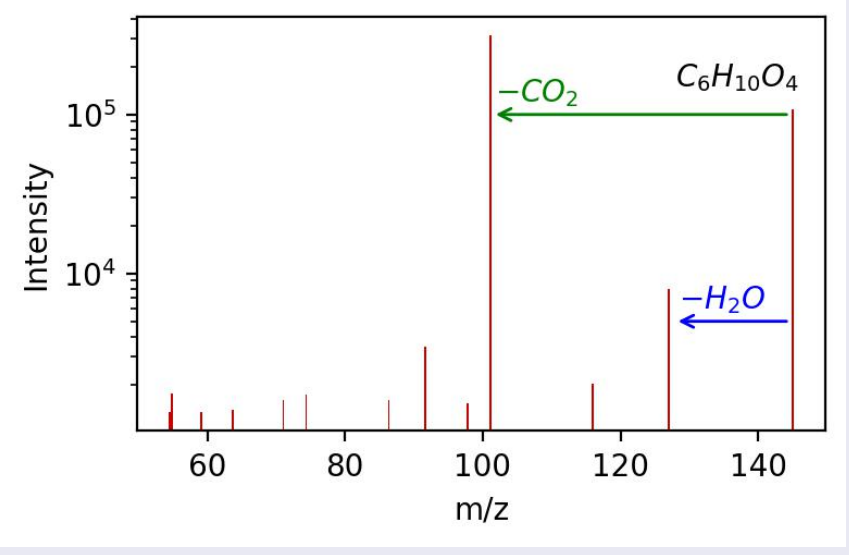
The data

MS1

- 1 molecule = 1 peak
- Measure of the molecular weight (m/z) $\Rightarrow$ Sum formula
- Example : peak at 146.12 $\Rightarrow$ $C_6H_{10}O_4$

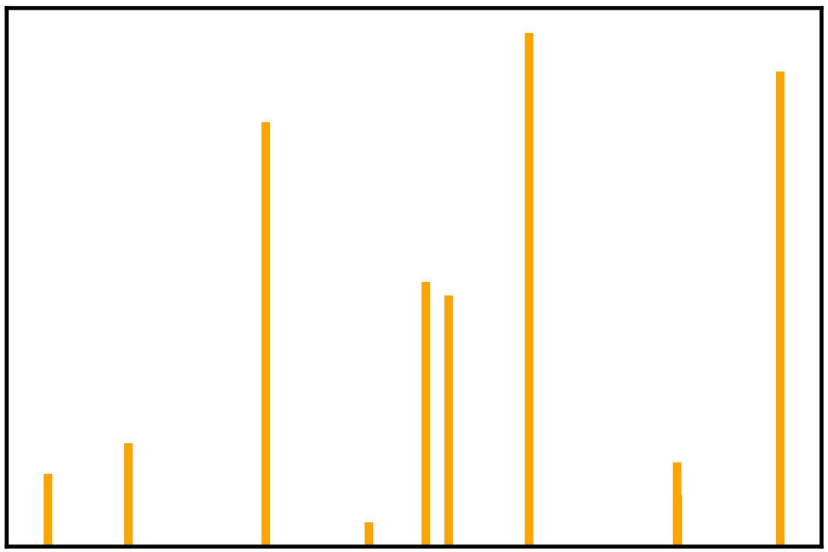
MS2

- Measure the m/z of **fragments**
- Intensity : amount of the given fragment
- Example : database search gives 3-methylglutaric acid

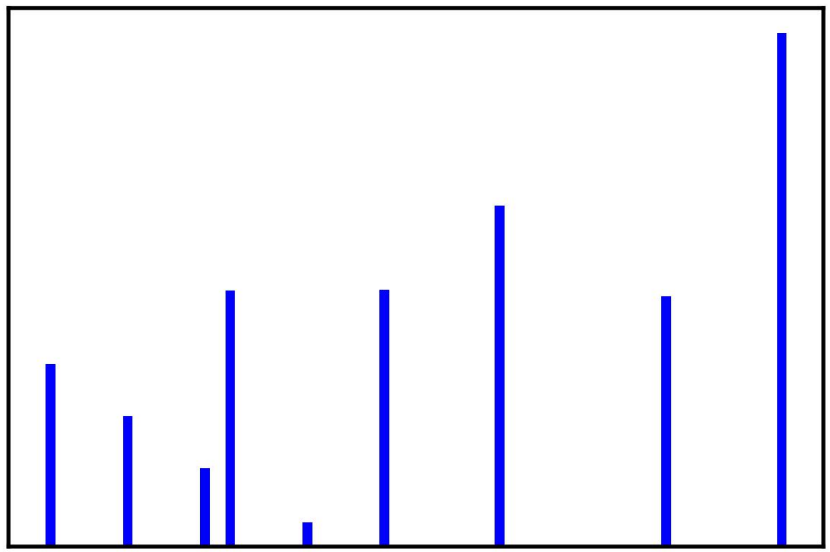


The problem

Fragmentation spectra from molecules of same m/z



Spectrum 1 from molecule 1



Spectrum 2 from molecule 2

- Method to answer "Are molecule 1 and molecule 2 the same ?"
- Take **intensities** into account

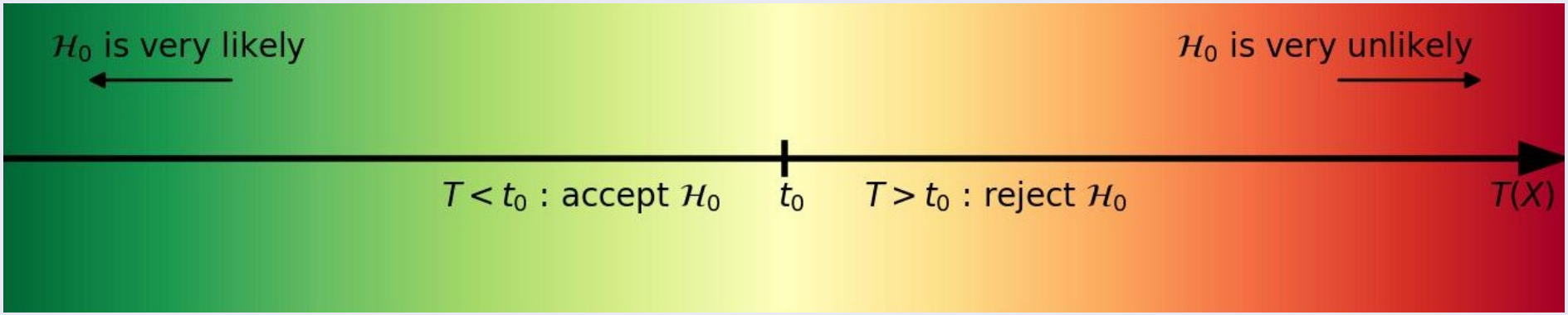
Statistical hypothesis testing

Given two fragmentation spectra, we test the hypothesis

$\mathcal{H}_0$: "the spectra correspond to the same molecule"

The test procedure

- X : the **data** (here, the two spectra)
- t_0 : the **test threshold**, a numerical quantity determined in advance
- $T(X)$: the **test statistic** depending on X only



The likelihood ratio method

The likelihood function

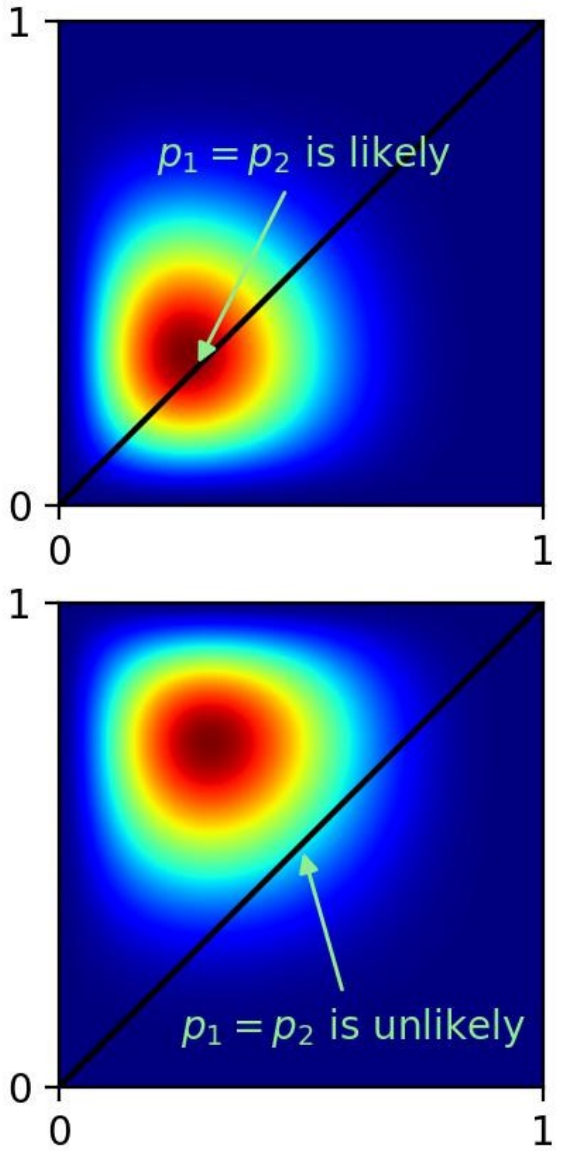
Assuming $X \sim$ Distribution(θ), the **likelihood function** $L(X, \theta)$ writes

$$L(X, \theta) = Pr(X|\theta)$$

(Probabilistic point of view $\rightarrow$ Statistical point of view)

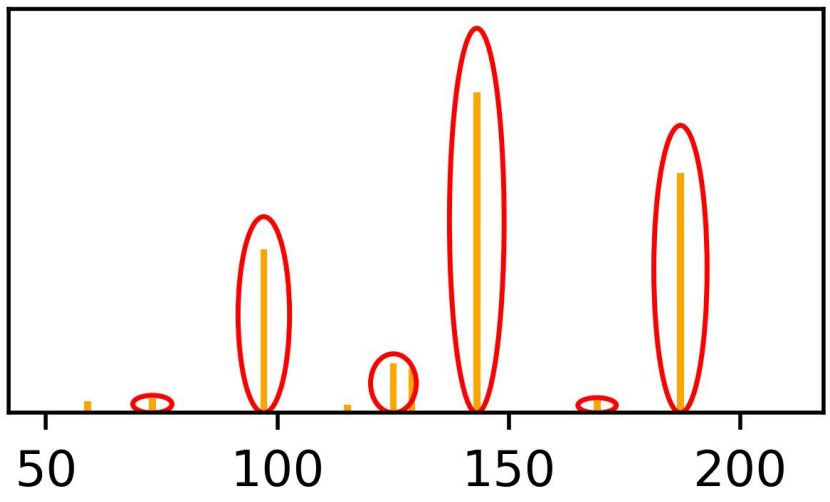
Example : test of $p_1 = p_2$

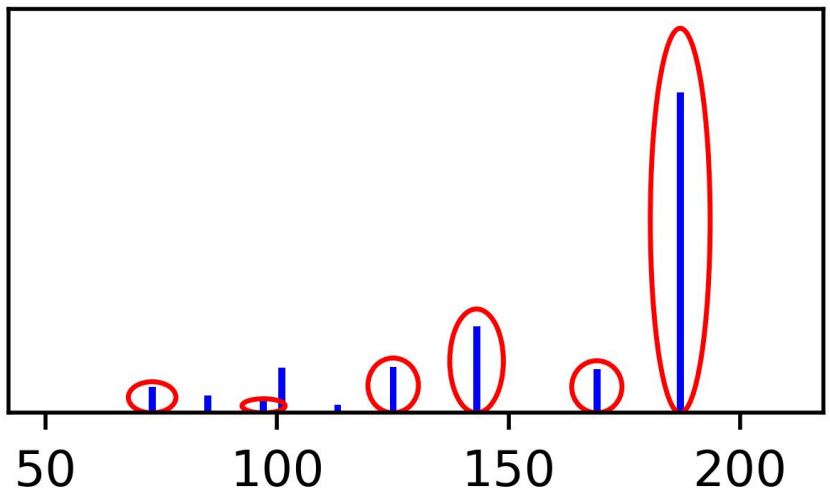
$X = (X_1, X_2)$ where $X_i \sim \mathrm{Bin}(n_i, p_i)$,
p_1 on the x-axis
p_2 on the y-axis



Conversion into contingency tables

1) Match the peaks of same m/z





2) Store the matching peaks into a table

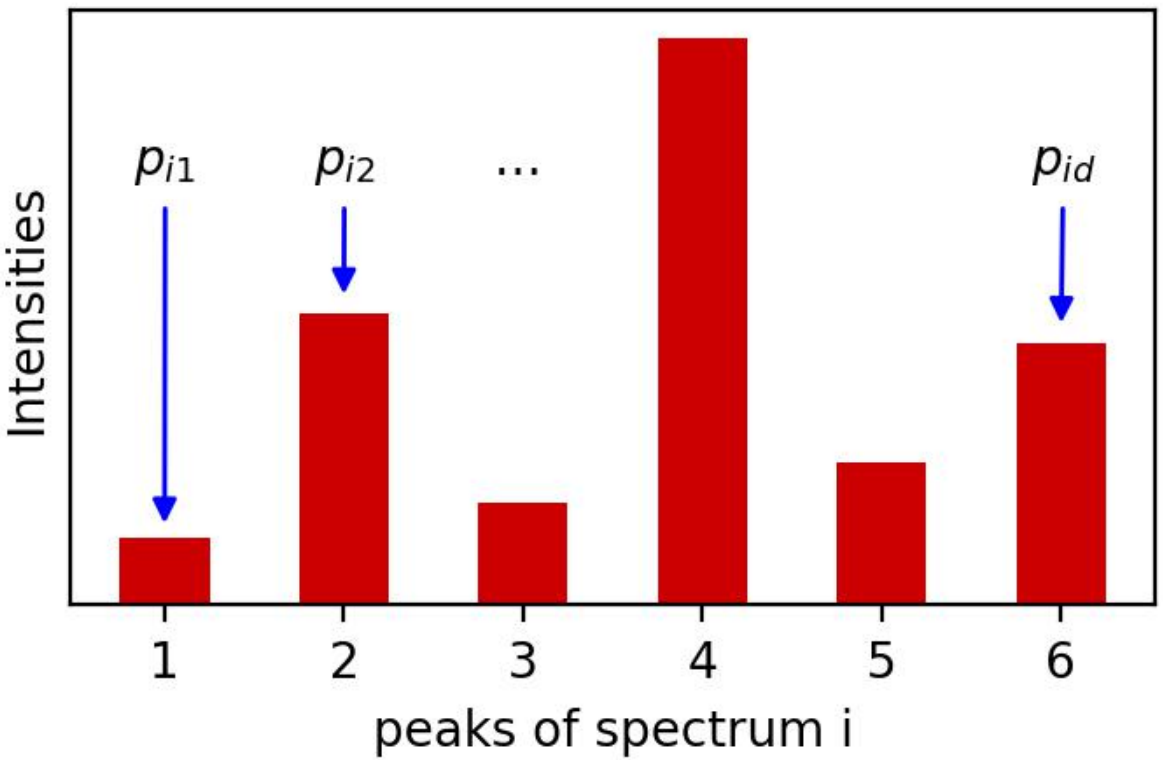
Spectrum	Intensity 1	Intensity 2	...	Intensity d	Sum
1	n_{11}	n_{12}	...	n_{1d}	n_{1+}
2	n_{21}	n_{22}	...	n_{2d}	n_{2+}

d = nb of matching peaks in each spectrum = nb of columns of the table

Multinomial model

Definition

- $(n_{11}, n_{12}, \ldots, n_{1d}) \sim MN(n_{1+}; (p_{11}, p_{12}, \ldots, p_{1d}))$
- $(n_{21}, n_{22}, \ldots, n_{2d}) \sim MN(n_{2+}; (p_{21}, p_{22}, \ldots, p_{2d}))$
- test of $\mathcal{H}_0 =$ "$(p_{11}, p_{12}, \ldots, p_{1d})) = (p_{21}, p_{22}, \ldots, p_{2d}))$"



Comments

- Conditional with respect to n_{i+}
- $n_{i+} \sim 10^6$
- $d \in [1, 20]$ most of the time

Dirichlet-multinomial model

Definition

- $(p_{i1}, \ldots, p_{id}) \sim \mathrm{Dir}((\theta_{i1}, \ldots, \theta_{id}); \varphi)$
- $(n_{11}, n_{12}, \ldots, n_{1d})|\mathbf{p_i} \sim MN(n_{i+}; (p_{i1}, p_{i2}, \ldots, p_{id}))$
- test of $\mathcal{H}_0 =$ "$(\theta_{11}, \theta_{12}, \ldots, \theta_{1d})) = (\theta_{21}, \theta_{22}, \ldots, \theta_{2d}))$"

Comments

- Overdispersed multinomial model
- φ controls the variance of the n_{ij}
- $\boldsymbol{\theta_i}$ in DMN $\equiv \mathbf{p_i}$ in MN

Comparison between the models

Distribution	Parameters	Expectation	Covariance matrix
MN	$n_{i+}, \mathbf{p_i}$	$n_{i+} \cdot \mathbf{p_i}$	$(n_{i+} \cdot (\delta_{j,k} p_{ij} - p_{ij} p_{ik}))_{j,k}$
DMN	$n_{i+}, \boldsymbol{\theta_i}, \varphi$	$n_{i+} \cdot \boldsymbol{\theta_i}$	$n_{i+}^2 (\delta_{j,k} \theta_{ij} - \theta_{ij} \theta_{ik}) \frac{1+(n_{i+}\varphi)^{-1}}{1+\varphi^{-1}}$

Likelihood statistic

Multinomial model (MN)

- Generalisation of binomial distributions
- **Data** $= (\mathbf{n_1}, \mathbf{n_2})$ where $\mathbf{n_i} = (n_{i1}, \ldots, n_{id})$
- **Parameters** $= (\mathbf{p_1}, \mathbf{p_2})$, $\mathbf{p_i} = (p_{i1}, \ldots, p_{id})$ of sum 1

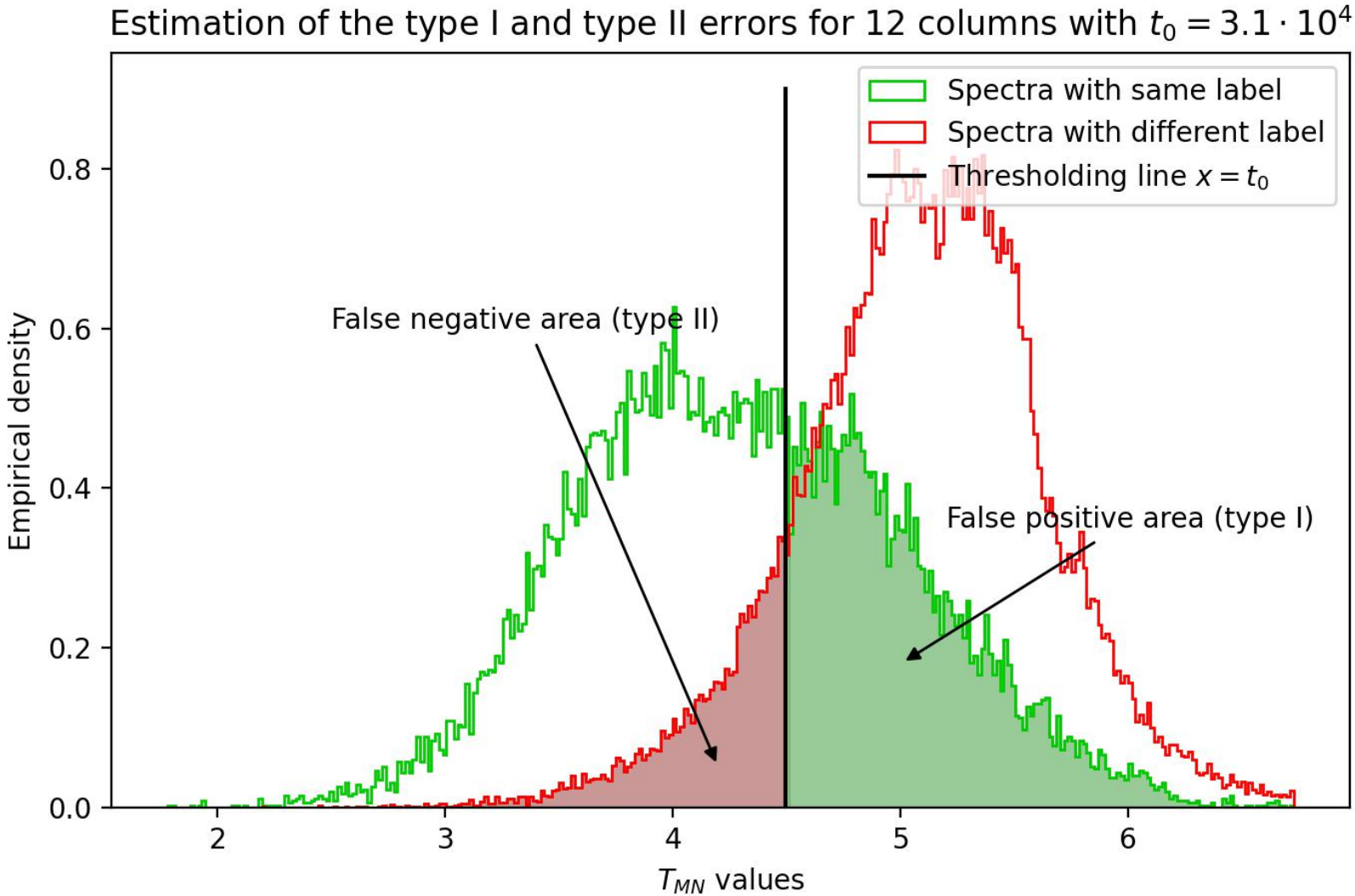
$$T_{MN} = 2 \log \left(\frac{L(\mathbf{n}, \mathbf{p}^{(1)})}{L(\mathbf{n}, \mathbf{p}^{(0)})} \right) \quad \text{with} \quad p_{ij}^{(1)} = \frac{n_{ij}}{n_{i+}} \quad , \quad p_{ij}^{(0)} = \frac{n_{+j}}{N}$$

Dirichlet-multinomial model (DMN)

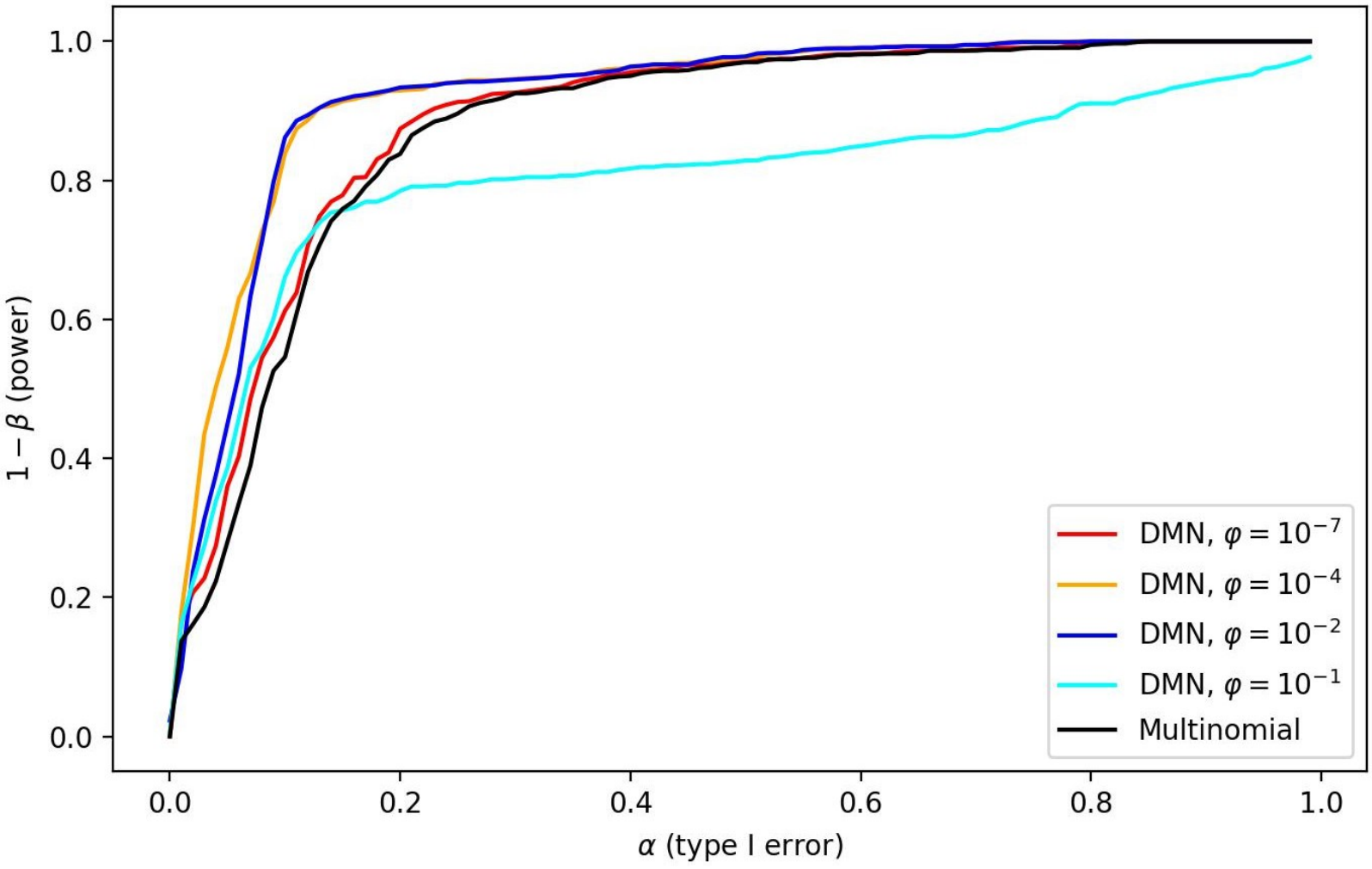
- Additional parameter φ accounting for **overdispersion**
- **Data** $= (\mathbf{n_1}, \mathbf{n_2})$
- **Parameters** $= (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$, $\boldsymbol{\theta_i} = (\theta_{i1}, \ldots, \theta_{id})$ of sum 1

$$T_{DMN} = 2 \log \left(\frac{L(\mathbf{n}, \boldsymbol{\theta}^{(1)})}{L(\mathbf{n}, \boldsymbol{\theta}^{(0)})} \right) \quad \text{with} \quad \theta_{ij}^{(1)} = \frac{n_{ij}}{n_{i+}} \quad , \quad \theta_{ij}^{(0)} = \frac{n_{+j}}{N}$$

Error types and choice of t_0



Adjusting φ in the DMN model



ROC curves : it works best when $\varphi \in [10^{-5}; 10^{-2}]$

Evaluating performance of the model

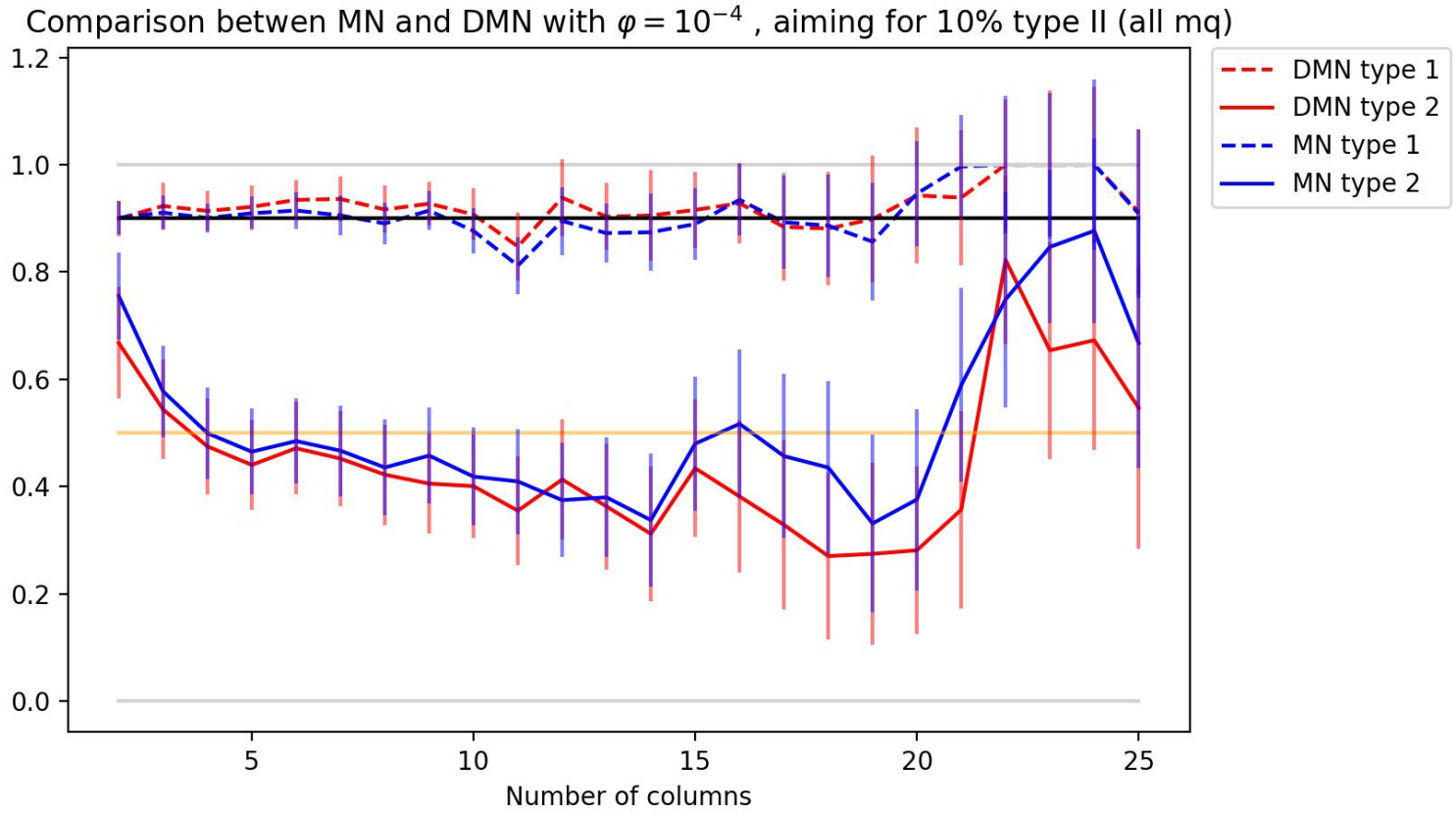
Issues

- Estimation of t_0
- Take into account **match quality**
- **Bias** : 1000 spectra of same label outweigh 5 spectra of same label

Solutions

1. Use all of the database for both t_0 and error estimation
   $\rightarrow$ can handle match quality
   $\rightarrow$ risk of bias
2. **Split** the data in : 40 % to determine t_0, 10 % to compute errors
   $\rightarrow$ less bias
   $\rightarrow$ confidence intervals
   $\rightarrow$ not enough remaining data to be picky about mq

Comparison between the models



- DMN seems to perform a bit better, especially for $15 \leq d \leq 20$
- Not enough data for significant results for $d > 20$

Conclusions

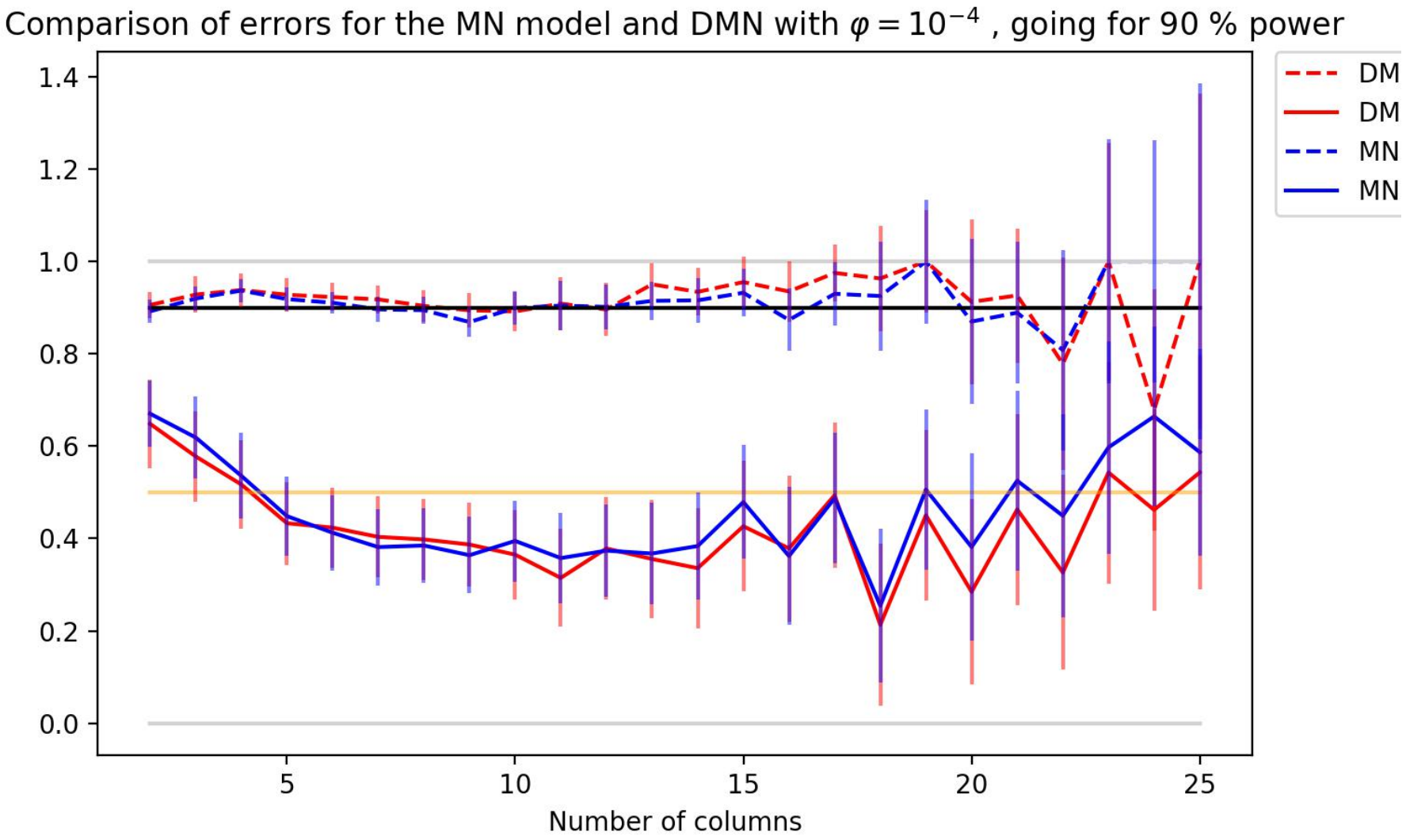
- Intensities can definitely be used to analyze mass spectrometry data
- Modelling peak intensities with overdispersion
- Need to make a trade-off between types of errors

Perspectives

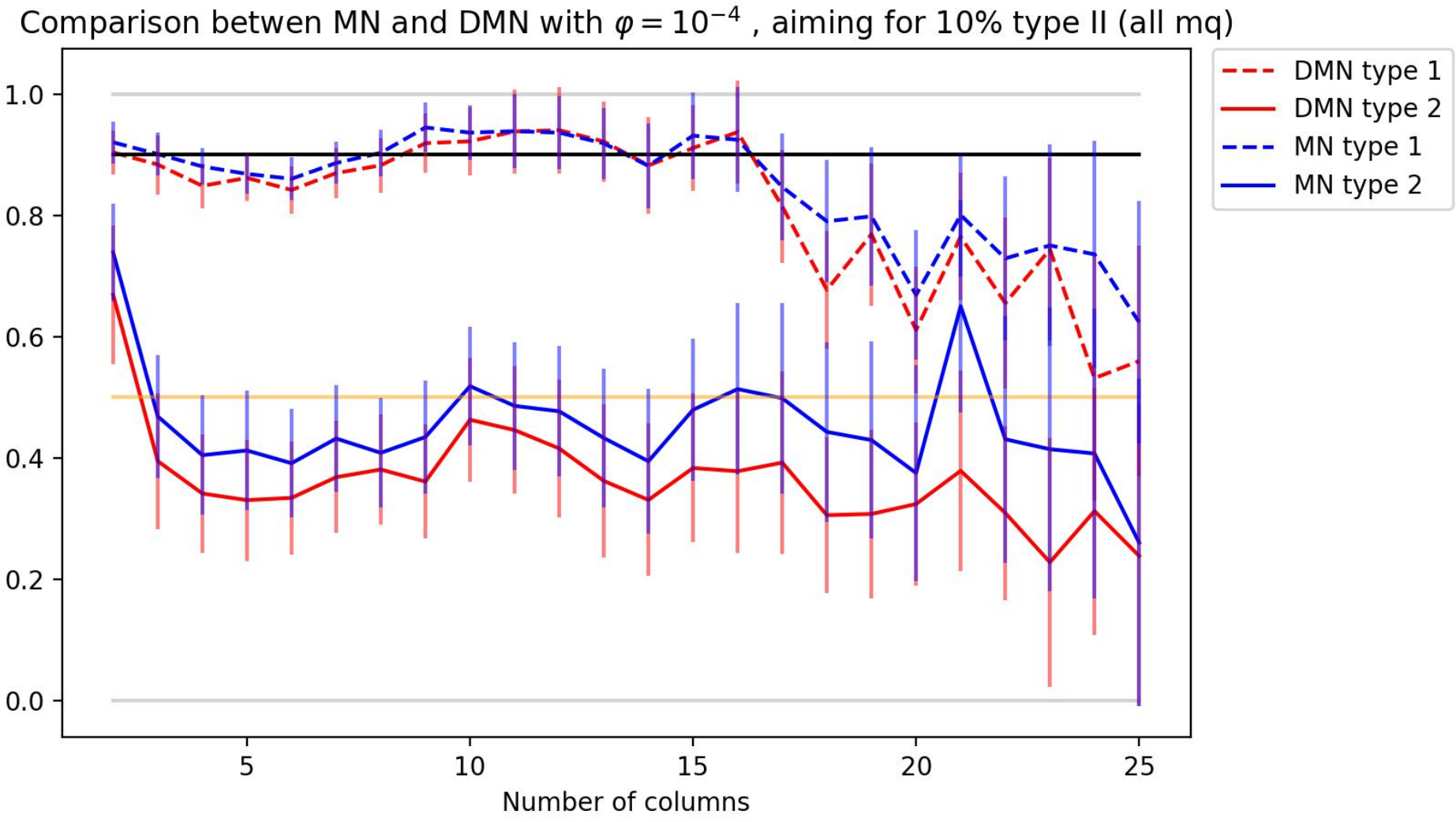
- New models : bayesian approach, more sophisticated overdispersion
- Using both similarity index and intensities
- Comparing more than 2 spectra at a time
- Machine learning

Thank you for your attention

Additional graphics (other splitting of the data)



Additional graphics (other splitting of the data)



Additional graphics (compounds with positive match quality)