Méthodes statistiques pour la comparaison de spectres de masse

Malo Hillairet
Tuteur : Guillaume Obozinski

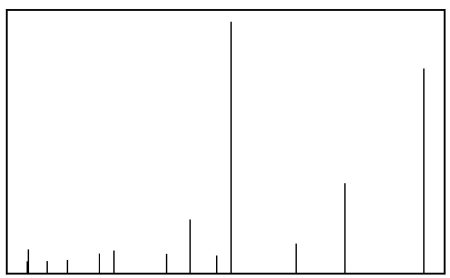
Lausanne, Suisse

Introduction

Le projet MSEI : Molecular Structure Elucidation

- Sasa Bjelić, Lilian Gasser, Eliza Harris, Guillaume Obozinski
- Identifier et classifier des molécules
- Chimie, Sciences des Données, et dans notre cas Statistiques

Données



Classification

...	...	...
Acétone	C_3H_6O	...
Acide Glutarique	$C_5H_80_4$	...
...	...	...

Plan

Cadre mathématique et notations

Modèle statistique

$$\mathcal{M} = \{\mathbb{P}_\theta, \theta \in \Theta\}$$

- θ : paramètre inconnu
- $\mathbb{P}_\theta$: distribution des données sous le paramètre θ
- X : données observées ($X \sim \mathbb{P}_\theta$)

Fonction de vraisemblance

$$L(x, \theta) = \mathbb{P}_\theta(X = x)$$

- Inférence statistique : estimation d'une quantité $g(\theta)$
- Test d'hypothèse : accepter ou rejeter une hypothèse sur θ

$$\mathcal{H}_0 = \text{“}\theta \in \Theta_0\text{”} \text{ , où } \Theta_0 \subset \Theta$$

Cadre mathématique et notations

Distribution multinomiale

- Généralise la loi binomiale à la dimension $d \geq 2$
- Paramètre $\mathbf{p} \in \Delta^{d-1} = \{\mathbf{p} \in (\mathbb{R}_+)^d \mid \sum_j p_j = 1\}$
- Si $X \sim \mathrm{Multi}(n, \mathbf{P})$ et $k_1 + \cdots + k_d = n$,

$$\mathbb{P}(X = (k_1, \ldots, k_d)) = \binom{n}{(k_1, \ldots, k_d)} \prod_{j=1}^{d} p_j^{k_j}$$

Espérance et variance

$$\mathbb{E}[X] = n \cdot \mathbf{p}$$

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{si} \quad i = j \\ -np_i \cdot p_j & \text{si} \quad i \neq j \end{cases}$$

Cadre mathématique et notations

Table de contingence

	Colonne 1	Colonne 2	...	Colonne d	Somme
Ligne 1	$n_{1,1}$	$n_{1,2}$	...	n_{1d}	n_{1+}
Ligne 2	$n_{2,1}$	$n_{2,2}$	...	n_{2d}	n_{2+}
Somme	n_{+1}	n_{+2}	...	n_{+d}	N

- n_{ij} : nombre d'observations rentrant dans la catégorie (i, j)

Modèle multinomial

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

- d : nombre de colonnes
- paramètres : $\mathbf{p_i} = (p_{i1}, \ldots, p_{id}) \in \Delta^{d-1},\ i = 1, 2$
- données : $(\mathbf{n_1}, \mathbf{n_2}) \in \mathbb{N}^d \times \mathbb{N}^d$ où $\sum_{j=1}^{d} n_{ij} = n_{i+}$

Test de vraisemblance pour le modèle multinomial

Modèle multinomial

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

- d : nombre de colonnes
- paramètres : $\mathbf{p_i} = (p_{i1}, \ldots, p_{id}) \in \Delta^{d-1},\ i = 1, 2$
- données : $(\mathbf{n_1}, \mathbf{n_2}) \in \mathbb{N}^d \times \mathbb{N}^d$ où $\sum_{j=1}^{d} n_{ij} = n_{i+}$

Le problème

À partir des données, **accepter** $\mathcal{H}_0$ ou la rejeter pour $\mathcal{H}_1$, où

$$\begin{array}{rcl} \mathcal{H}_0 & : & \text{“}\mathbf{p_1} = \mathbf{p_2}\text{”} \\ \mathcal{H}_1 & : & \text{“}\mathbf{p_1} \neq \mathbf{p_2}\text{”} \end{array}$$

Test de vraisemblance pour le modèle multinomial

Statistique de log-vraisemblance

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

- $L(\mathbf{n}, \mathbf{p})$: fonction de vraisemblance de $\mathcal{M}_{MN}$
- $\mathbf{n}$: données $(\mathbf{n_1}, \mathbf{n_2})$
- $\mathbf{p}^{(\alpha)}$: **estimateur du maximum de vraisemblance** sous $\mathcal{H}_\alpha$

Interprétation

- T_{LL} **grande** $\Leftrightarrow$ **faible** pertinence de $\mathcal{H}_0$ en comparaison à $\mathcal{H}_1$
- divergence de Kullback-Leibler :

$$T_{LL} = 2 \left(n_{1+} \cdot \mathrm{KL}(\mathbf{p_1}^{(1)}, \mathbf{p_1}^{(0)}) + n_{2+} \cdot \mathrm{KL}(\mathbf{p_2}^{(1)}, \mathbf{p_2}^{(0)}) \right)$$

Surdispersion et distribution Dirichlet-multinomiale

Distribution composée

Variable aléatoire X définie par

- La loi d'une variable aléatoire p
- La loi conditionnelle de X sachant p

Expression si X est discrète :

$$\mathbb{P}(X = x) = \int_p \mathbb{P}(X = x|p)\mathrm{d}\,\mathbb{P}(p)$$

Décomposition de la variance

$$\mathrm{Var}(X) = \mathbb{E}_p(\mathrm{Var}(X|p)) + \mathrm{Var}_p(\mathbb{E}[X|p])$$

Surdispersion et distribution Dirichlet-multinomiale

Distribution de Dirichlet

- Famille de distributions indexée par $\boldsymbol{\theta} \in \Delta^{d-1}$ et $\varphi > 0$ (choix)
- Généralise les lois Beta (cas $d = 2$)
- Définie sur le simplexe Δ^{d-1}, de densité

$$f_{\mathrm{Dir}(\boldsymbol{\theta},\varphi)}(\mathbf{p}) = \frac{1}{\mathrm{B}(\varphi^{-1}\boldsymbol{\theta})} \prod_{j=1}^{d} p_j^{(\varphi^{-1}\theta_j - 1)}$$

où B est la fonction beta multivariée (coefficient de normalisation).

- A priori conjugué pour le modèle multinomial

Espérance et variance

- $\mathbb{E}[\mathbf{p}] = \boldsymbol{\theta}$
- $\mathrm{Cov}(p_i, p_j) = (\delta_{ij}\theta_i - \theta_i\theta_j)\frac{1}{1+\varphi^{-1}}$

Surdispersion et distribution Dirichlet-multinomiale

Distribution Dirichlet-multinomiale

- Paramètres $n \in \mathbb{N}^*, \theta \in \Delta^{d-1}, \varphi > 0$
- $X \sim \mathrm{DMN}(n; (\boldsymbol{\theta}, \varphi))$ si

$$\begin{aligned} \mathbf{p} &\sim \mathrm{Dir}(\boldsymbol{\theta}, \varphi) \\ X|p &\sim \mathrm{Multi}(n, \mathbf{p}) \end{aligned}$$

- Distribution multinomiale surdispersée

Analogie avec la distribution multinomiale

Distribution	Paramètres	Espérance	Coefficients de covariance
Multi	$N, \mathbf{p}$	$N\mathbf{p}$	$N(\delta_{i,j}p_i - p_i p_j)$
DMN	$N, \boldsymbol{\theta}, \varphi$	$N\boldsymbol{\theta}$	$N^2(\delta_{i,j}\theta_i - \theta_i\theta_j)\frac{1+(N\varphi)^{-1}}{1+\varphi^{-1}}$

Modèle Dirichlet-multinomial et statistique de test

Le modèle Dirichlet-multinomial

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

- paramètres : $\theta_i = (\theta_{i1}, \ldots, \theta_{id}) \in \Delta^{d-1}$, $i = 1, 2$ (seulement !)
- données $(\mathbf{n_1}, \mathbf{n_2})$

Hypothèses du test

Comme dans le modèle multinomial, avec $\boldsymbol{\theta}$ au lieu de $\mathbf{p}$

$$\begin{array}{rcl} \mathcal{H}_0 & : & \text{“}\boldsymbol{\theta_1} = \boldsymbol{\theta_2}\text{”} \\ \mathcal{H}_1 & : & \text{“}\boldsymbol{\theta_1} \neq \boldsymbol{\theta_2}\text{”} \end{array}$$

Modèle Dirichlet-multinomial et statistique de test

Statistique de test

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

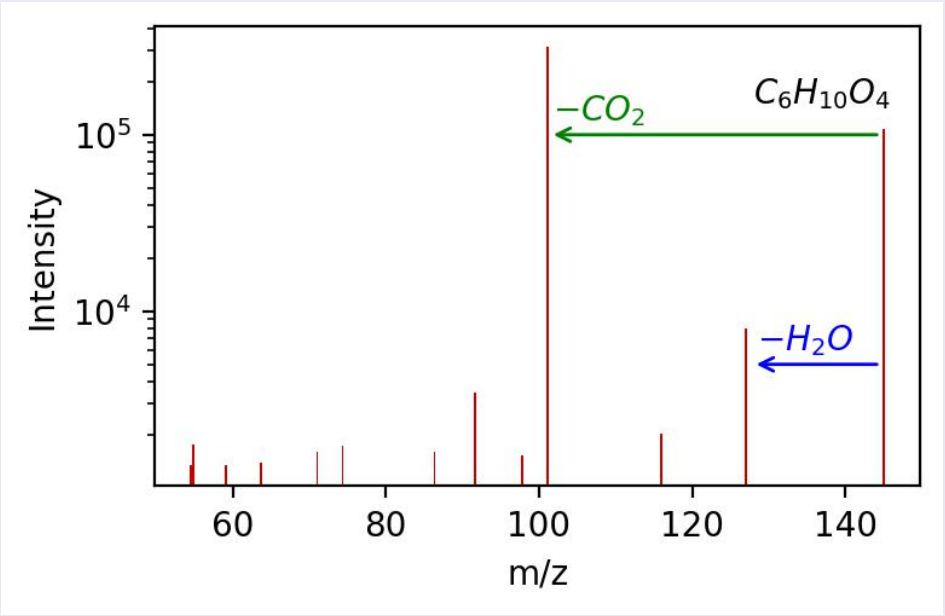
Remarque importante

Construction par **analogie** avec T_{LL}, mais ce **n'est pas** la statistique de log-vraisemblance de $\mathcal{M}_{DMN}$.

- T_{DMN} dépend aussi de φ, à **ajuster**
- Sous-estime la log-vraisemblance
- Peut être négative

Quelques résultats

Un spectre de fragmentation



- m/z : masse moléculaire
- Intensité : quantité de fragments à ce m/z

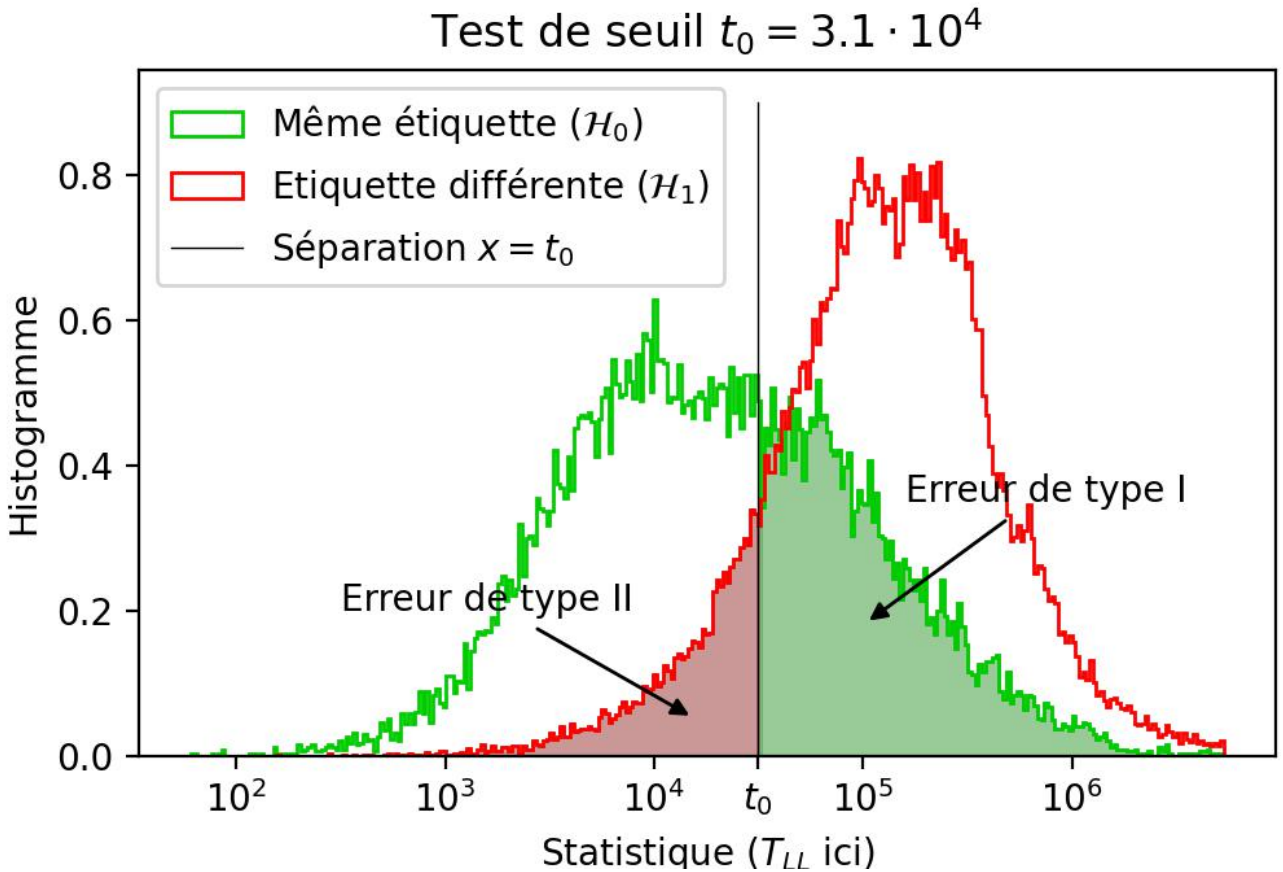
Table de contingence

Spectre	Intensité 1	Intensité 2	...	Intensité d	Somme
1	n_{11}	n_{12}	...	n_{1d}	n_{1+}
2	n_{21}	n_{22}	...	n_{2d}	n_{2+}

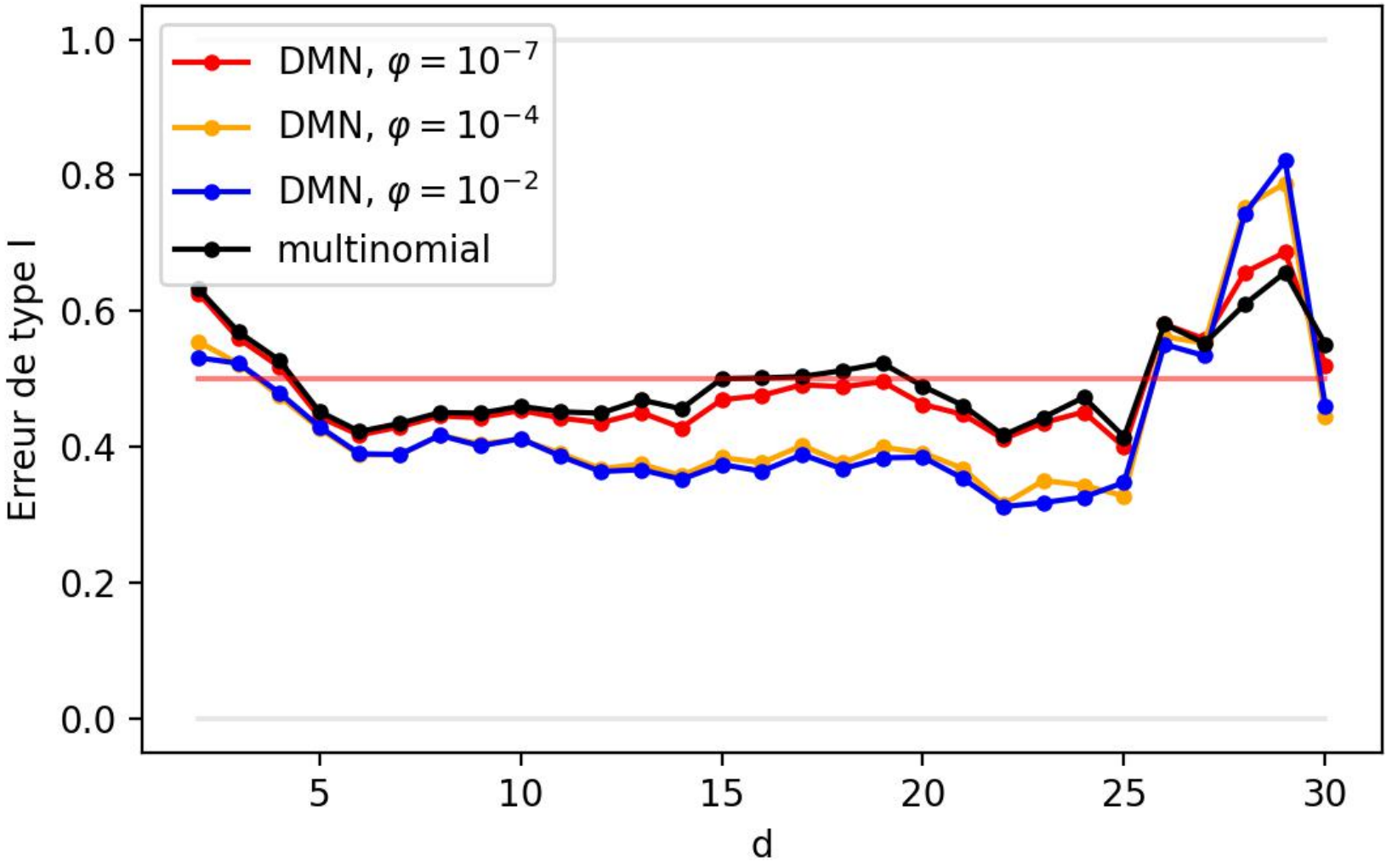
Quelques résultats

Test de puissance 90 %

rouge : apprentissage de t_0
vert : estimation de l'erreur de type 1



Quelques résultats



Conclusion et perspectives

- La vraisemblance : un outil pour construire des tests
- Asymétrie de $\mathcal{H}_0$ et $\mathcal{H}_1$
- Distribution composée : modélise une surdispersion

- Meilleurs modèles (bayésiens, plus de paramètres, ...)
- Machine learning
- Autres données

Merci pour votre attention

- Böcker, S. (2017). Searching molecular structure databases using tandem MS data: are we there yet?
- Cochran, W. G. (1952). The χ^2 Test of Goodness of Fit.
- Lydersen, S., Fagerland, M. W. & P. Laake (2009). Recommended tests for association in 2×2 tables.
- 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.
- Wilks, S. S. (1938). The Large-sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses