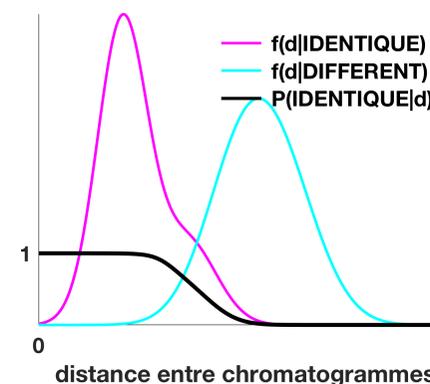
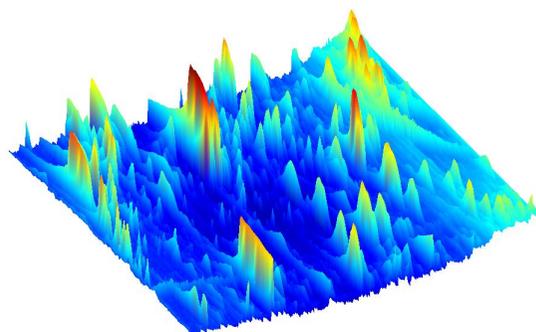


GCxGC-MS et approche bayésienne pour l'identification d'individus par leur empreinte olfactive



I. Rivals¹, V. Cuzuel², R. Leconte¹, G. Cognon², D. Thiébaud³, C. Sauleau², J. Vial³

¹ Équipe de Statistique Appliquée, ESPCI Paris

² Institut de Recherche Criminelle de la Gendarmerie Nationale, Cergy Pontoise

³ Laboratoire Sciences Analytiques, Bioanalytiques et Miniaturisation, ESPCI Paris

Paris, 30-31 janvier 2018,
Chimiométrie XIX

Introduction

- **But** : identifier une personne à son odeur par une méthode objective.
- **Odorat canin** : la trace olfactive contient une information suffisante.
- **Problème** : on ignore les bases chimiques sur lesquelles se fonde la reconnaissance canine.

→ **Projet de développer une stratégie complète :**

- prélèvement de l'odeur direct au niveau des mains
- l'analyse chimique des échantillons (GCxGC-MS)
- *le traitement statistique de leur composition :*



Question : **étant donné un chromatogramme de référence, celui de l'odeur prélevée sur la scène de crime, le chromatogramme de l'échantillon prélevé sur un suspect permet-il d'affirmer que cette odeur est bien la sienne, ou pas ?**

→ **approche bayésienne**

Approche bayésienne de la question

H_0 : les deux chromatogrammes correspondent au même individu

H_1 : les deux chromatogrammes correspondent des individus distincts

Si l'on note D les données observées (les deux chromatogrammes), la formule de Bayes fournit :

$$P(H_0 | D) = \frac{f(D | H_0)P(H_0)}{f(D | H_0)P(H_0) + f(D | H_1)P(H_1)}$$

Démarche adoptée :

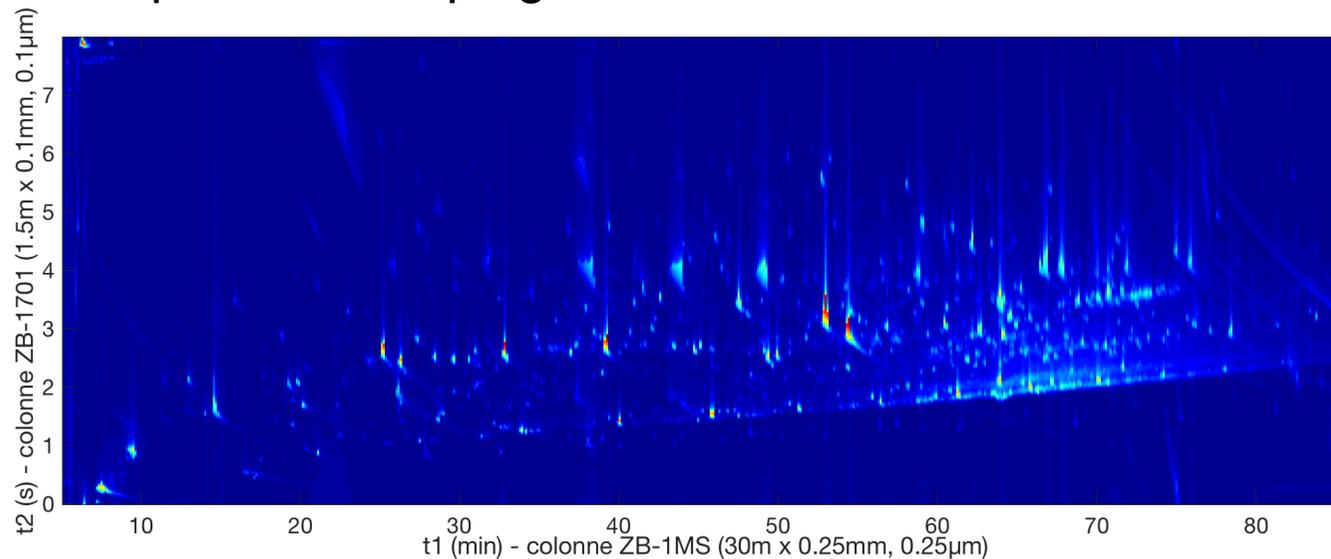
- Définition d'une **distance d entre chromatogrammes** ($D \equiv d$)
- Panel de chromatogrammes d'individus (prélevés plusieurs fois) divisé en **ensembles de calibration et de test indépendants**
- Ensemble de calibration \rightarrow estimation des **distributions de d** pour des couples de chromatogrammes du même individu $f(d|H_0)$ et d'individus distincts $f(d|H_1)$
- Ensemble de test \rightarrow estimation de la **performance** (AUC, sensibilité, spécificité)

De l'odeur des individus aux chromatogrammes

Panel de 119 sujets

	sexe		âge (années)			phototype		
	♂	♀	10-23	24-36	37-81	1	2	3
total	61	58	39	39	41	25	79	15

- 4 prélèvements directs au niveau des mains/sujet, avec des Sorb-star®
- Thermodésorption* et couplage GC×GC-MS**



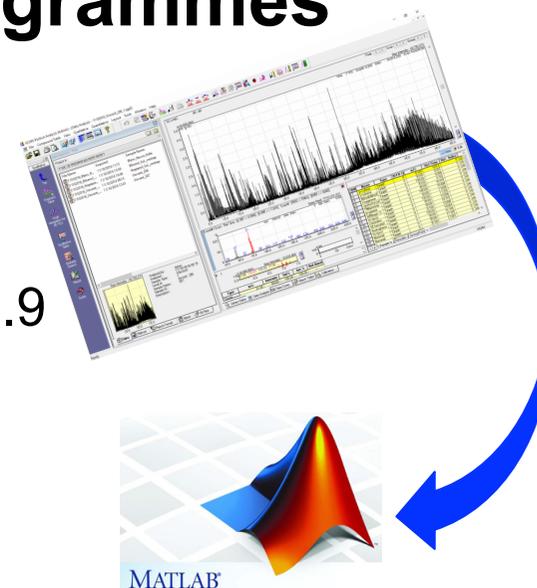
* Cuzuel et al., **Sampling method development and optimization in view of human hand odor analysis by thermal desorption coupled with gas chromatography and mass spectrometry**, 2017, Anal. Bioanal. Chem.

** Cuzuel et al., **Human odor and forensics. Optimization of a comprehensive gas chromatography method based on orthogonality: how not to choose between criteria.**, 2017, Journal of Chromatography A

Prétraitement des chromatogrammes

Acquisition et conversion

- GC Real Time Analysis 4.20 (Shimadzu)
→ fichiers mzXML
- Matlab R2017b + Bioinformatics toolbox version 4.9
→ intensités à chaque instant ($f = 50$ Hz),
pour chaque canal de masse (29-250 m/z)



Bibliothèque de composés personnalisée

- Traitement manuel de 25 chromatogrammes obtenus sur 3 sujets (♂ & ♀)
→ listes de plusieurs centaines de pics → bibliothèque de composés caractérisés par t_1 , t_2 , index de rétention linéaire, spectre de masse, nom du composé si trouvé dans la bibliothèque NIST
- Ajout de composés caractéristiques de l'odeur de la main humaine*
→ bibliothèque de 600 composés de référence identifiés ou non

un chromatogramme



le vecteur des intensités des pics correspondant aux 600 composés

* Cuzuel et al., Origin, analytical characterization and use of human odor in forensics, 2017, *J. Forensic Sci.*

Découpage de la base de données en ensembles de calibration et de test

- Répartition ~ 2/3 calibration, 1/3 test (nb chromatogrammes / nb sujets)

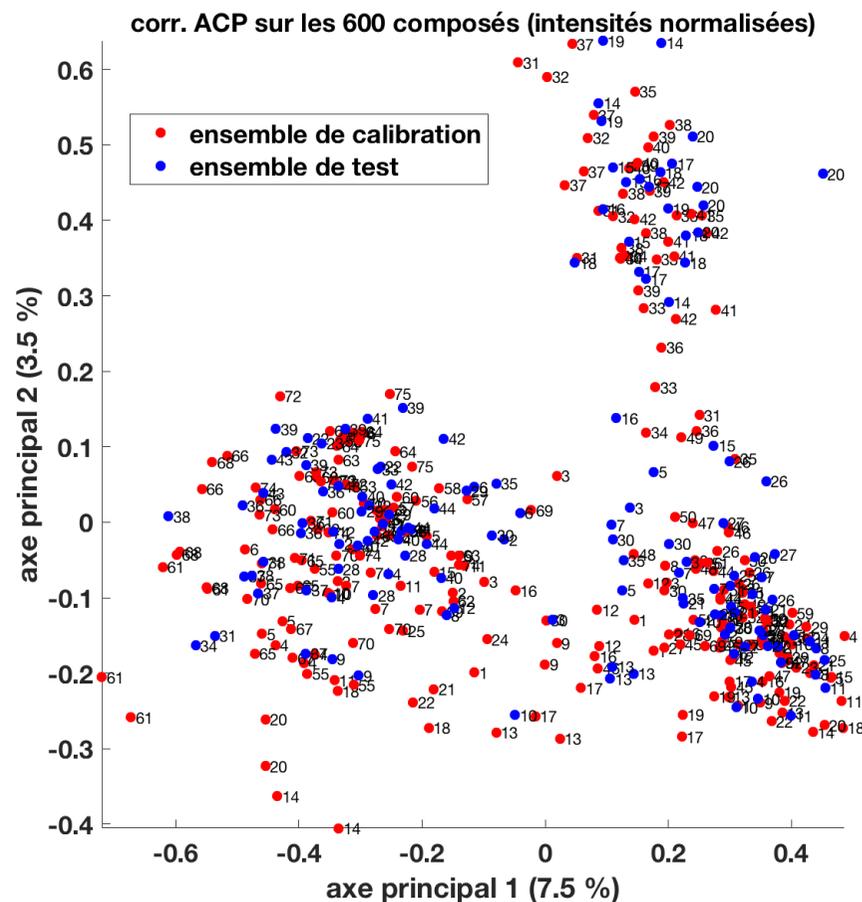
	calibration	test
total	260 / 75	139 / 44
♂	130 / 36	77 / 25
♀	130 / 39	62 / 19

- **ACP :**

- bonne répartition calibration/test

- fidélité limitée des mesures

- Intensités **normalisées** ($\sum = 1$) ou **binarisées** (présence/absence) 6



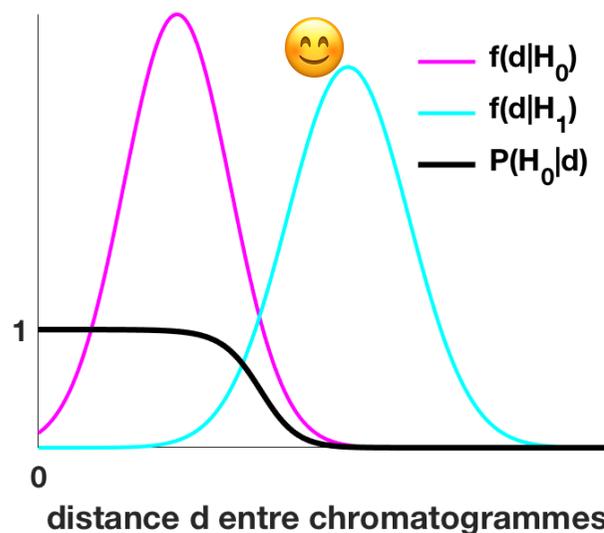
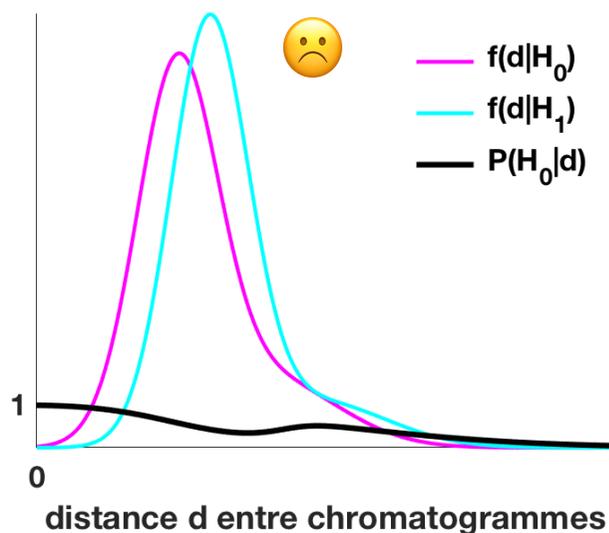
Ingrédients de l'approche bayésienne

Soient deux chromatogrammes (600-vecteurs) :

- H_0 l'hypothèse qu'ils proviennent du même individu
- H_1 l'hypothèse qu'ils proviennent de deux individus différents
- d une distance entre les deux chromatogrammes

$$\text{alors : } P(H_0 | d) = \frac{f(d | H_0)P(H_0)}{f(d | H_0)P(H_0) + f(d | H_1)P(H_1)}$$

- Probabilités *a priori* : $P(H_0) = P(H_1) = 0.5$
- Vraisemblances : $f(d|H_0)$ et $f(d|H_1)$: exemples fictifs



Choix d'une distance entre chromatogrammes et estimation des vraisemblances

Options :

- pour d , 3 distances entre les deux 600-vecteurs d'intensités :
 - a) **distance euclidienne**
 - b) **1 – coefficient de corrélation de Pearson**
 - c) **1 – coefficient de corrélation de Spearman**

- intensités **normalisées** / **binarisées** (b=c)

Ensemble de calibration (260 chromatogrammes / 75 sujets)

- 341 couples de chromatogrammes sous H_0 (du même sujet)
- 33 329 couples de chromatogrammes sous H_1 (de sujet différents)
- **histogrammes** de valeurs de d sous H_0 et H_1

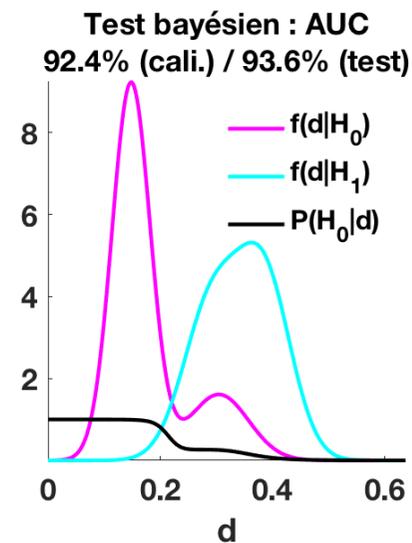
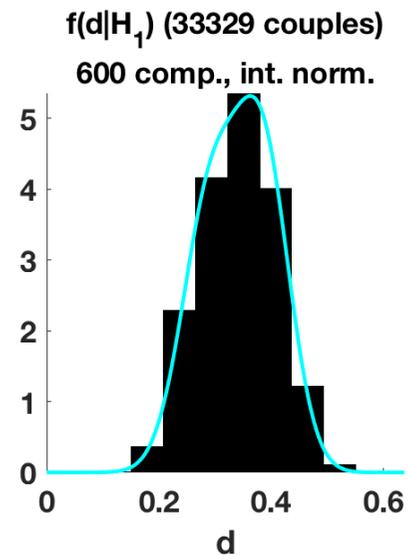
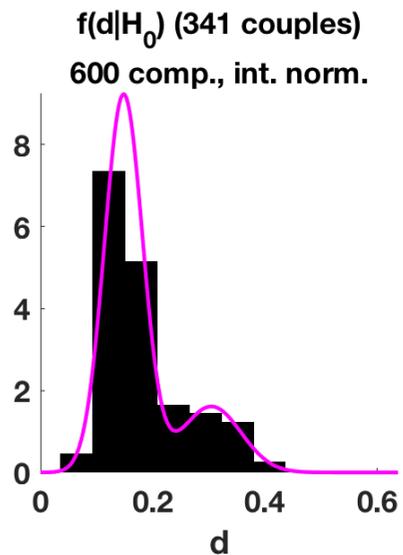
Ajustement de chaque histogramme par un **mélange de gaussiennes**

→ $f(d|H_0)$ et $f(d|H_1)$

Résultats d'identification

(%AUC calibration / %AUC test)

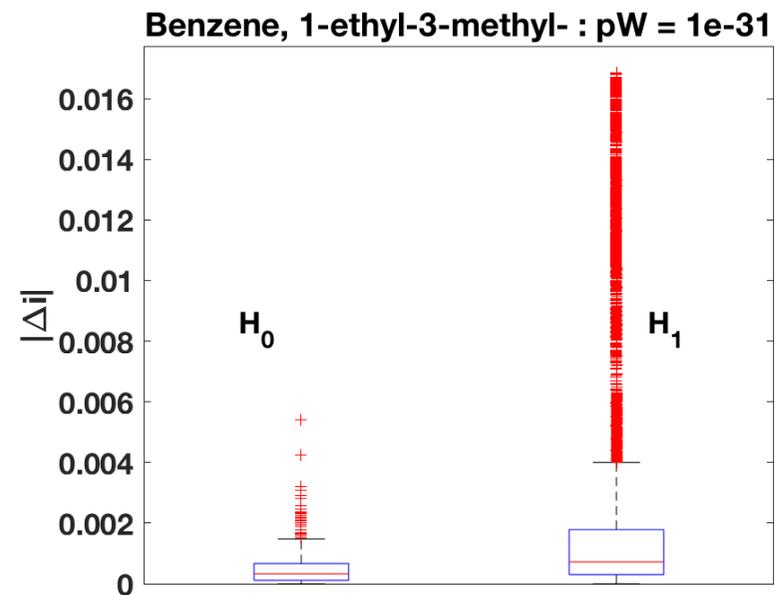
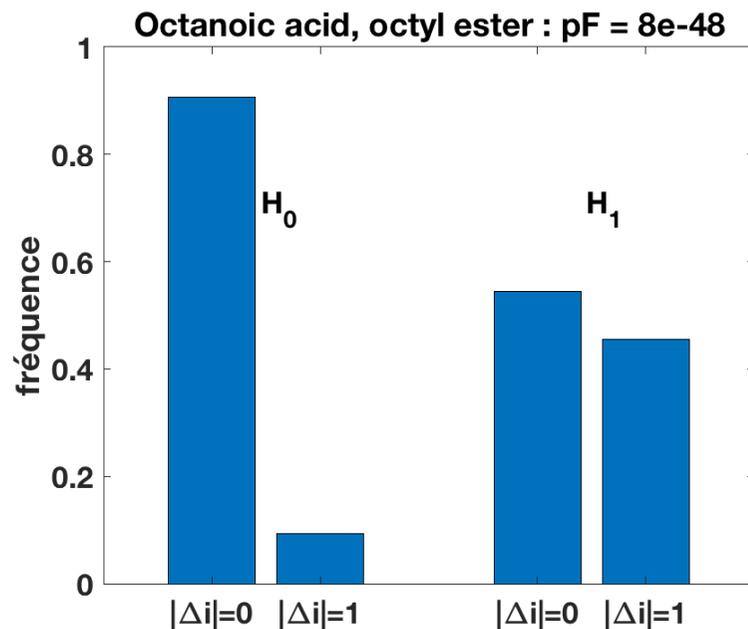
distance intensités	euclidienne	$1 - \rho_{\text{Pearson}}$	$1 - \rho_{\text{Spearman}}$
normalisées	62.7% / 64.6%	74.6% / 74.7%	92.4% / 93.6%
binarisées	88.4% / 91.6%	89.6% / 91.7%	



N.B. Pour l'ensemble de test, il y a 173 / 9 418 couples sous H_0 / H_1 .

Sélection des composés

- Composés discriminants entre H_0 et H_1 : ceux dont les **différences d'intensité** $|\Delta i|$ sont **significativement plus petites sous H_0 que sous H_1**
- Quantification : **p -valeur** du **test unilatéral** de **Fisher** (intensités binarisées) ou de **Wilcoxon** (intensités normalisées) sur les $|\Delta i|$
- Exemples :



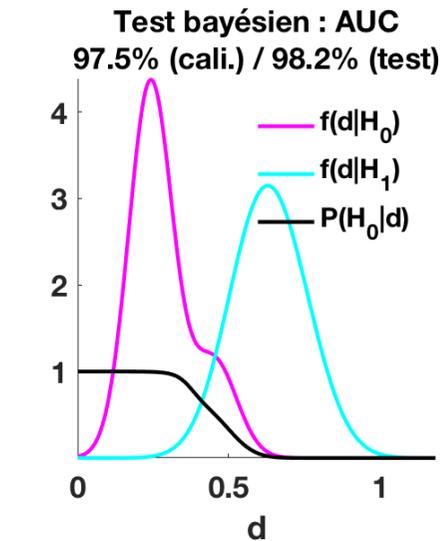
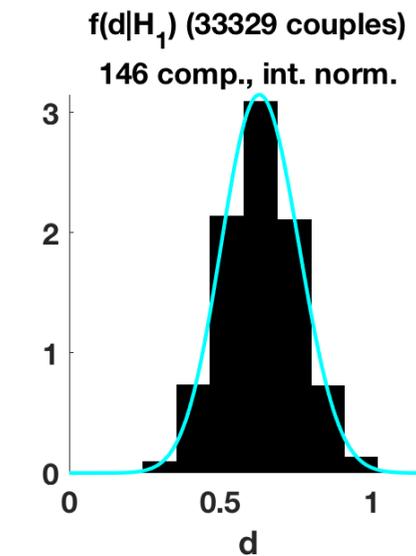
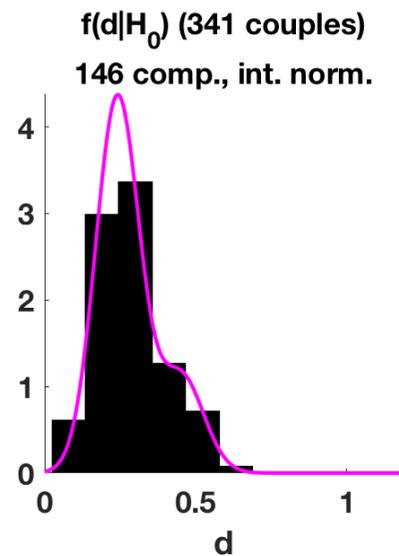
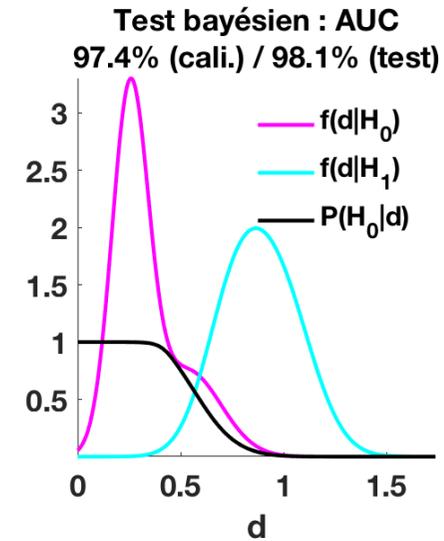
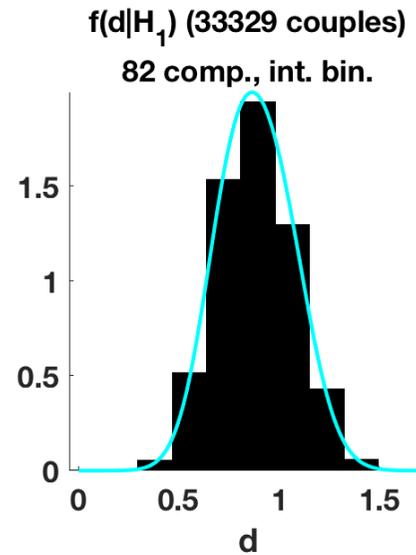
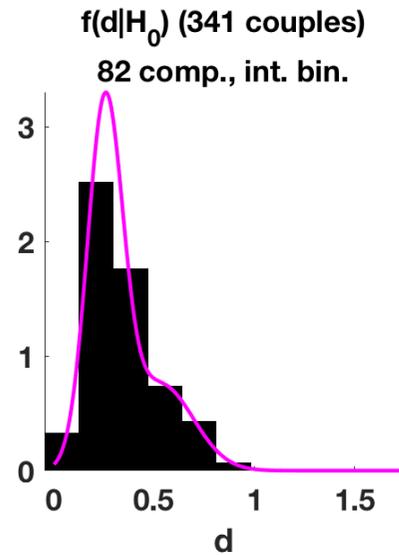
Résultats d'identification avec sélection des composés

(%AUC calibration / %AUC test)

- **Seuil π sur $-\log_{10}(p)$** du test de Fisher (intensités binarisées) ou de Wilcoxon (intensités normalisées) : valeur optimisée par **validation croisée** (K=3) sur l'**ensemble de calibration**

distance intensités	euclidienne	$1 - \rho_{\text{Pearson}}$	$1 - \rho_{\text{Spearman}}$
normalisées	$\pi = 12 / 61$ comp. 76.2% / 73.9%	$\pi = 13 / 54$ comp. 78.1% / 75.2%	$\pi = 7 / 146$ comp. 97.5% / 98.2%
binarisées	$\pi = 18 / 82$ comp. 93.1% / 94.8%	$\pi = 18 / 82$ comp. 97.4% / 98.1%	

Résultats d'identification avec sélection des composés



Discussion

Performances

intensités	AUC	sensibilité	spécificité	nb. comp.
binarisées	97.4% / 98.1 %	89.4% / 90.0%	94.9% / 92.5%	82
normalisées	97.5% / 98.2 %	89.1% / 85.9%	93.7% / 95.0%	146

- Choix d'une distance adéquate → exploitation **quantitative** des intensités des composés malgré la variabilité analytique
- **Sélection** → forte atténuation des seconds modes de $f(d|H_0)$ et $f(d|H_1)$
→ nette amélioration des résultats
- Classifieur à intensités binarisées plus **parcimonieux**
- 67 composés communs aux deux classifieurs

Limites

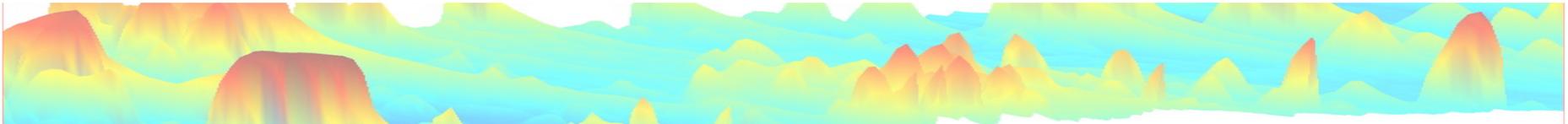
Contrairement au cas d'une **scène de crime**

- les deux prélèvements sont de même nature (prélèvements directs)
- pas de pollution par d'autres odeurs

Conclusion

Avancées

- Première combinaison d'une stratégie analytique fondée sur la GCxGC-MS et d'une approche bayésienne pour l'identification d'individus par leur empreinte olfactive à des fins forensiques
- Méthode validée sur un large panel de 119 sujets des deux sexes, de tous âges et phototypes → classifieurs bayésiens calibrés sur 75 sujets avec une AUC de 98.2% sur un ensemble indépendant de 44 sujets



Perspectives

- Examen des composés discriminants
- Normalisation sur des composés spécifiques
- Choix d'une distance plus complexe (possiblement vectorielle)
- Prélèvements directs dans des conditions plus variées
- Confrontation de prélèvements directs et indirects