Preprint:

https://doi.org/10.48550/arXiv.2501.01840

The Spiked Mixture Model: A New Clustering Approach for Imaging Mass Spectrometry Offers Advanced Resilience to Noise

Paul-Louis Delacour[1], Sander Wahls[2], Lauren N. Emmerson[3,4], Madeline E. Colley[3,5], Jeffrey M. Spraggins[3,5,6,7], Lukasz G. Migas[1], Raf Van de Plas[1,3,6]

TU Delft
Van de Plas Lab

The Spiked Mixture Model (SMM)

$$\mathbf{y} = \begin{cases} \alpha\mathbf{x_1} + \varepsilon \text{ with probability } \pi_1 \\ \alpha\mathbf{x_2} + \varepsilon \text{ with probability } \pi_2 \\ \cdots \\ \alpha\mathbf{x_K} + \varepsilon \text{ with probability } \pi_K \end{cases}$$

$$\alpha \sim \mathcal{N}(0,1),\ \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I_d})$$
$$\textstyle\sum_{k=1}^{K} \pi_k = 1,\ \mathbf{x}_1, \ldots, \mathbf{x}_K \in \mathcal{R}^d$$

Imaging mass spectrometry (IMS):

- IMS is a molecular imaging technique that combines spatial mapping with mass spectral analysis.
- IMS offers detailed chemical maps of organic tissue samples, measuring the spatial distributions of hundreds of molecular species concurrently.
- Each pixel records a full mass spectrum or m/z profile, and prior labeling of compounds is not required.

Challenges:

1. IMS experiments can acquire vast amounts of spatially resolved data.
2. IMS data can be noisy.

Modeling IMS data using the Spiked Mixture Model:

- Every observation or pixel $\mathbf{y}$ in an IMS dataset is modeled as a randomly scaled representative of a biological signature, subpopulation, or spike $\mathbf{x_K}$.
- α is the random scaling factor of observation $\mathbf{y}$:
- ε is the random noise of observation $\mathbf{y}$:
- $\mathbf{z}$ is a latent categorical variable indicating which spike $\mathbf{x_K}$ is underlying $\mathbf{y}$:

$$\alpha \sim \mathcal{N}(0,1),\ \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I_d})$$
$$\mathbf{z} \sim \mathrm{cat}(\pi)$$

Recovering signals underlying noisy IMS measurements (and clustering them):

- Given noisy observations $\{\mathbf{y}_1, \ldots \mathbf{y}_N\}$, find $\mathbf{K}$ underlying subpopulations $\mathbf{x}_1, \ldots, \mathbf{x_K}$

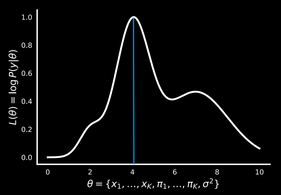
i.e., find parameters of interest: $\theta = \{\mathbf{x}_1, \ldots, \mathbf{x_K}, \pi_1, \ldots, \pi_K, \sigma^2\}$

that deliver a maximum likelihood estimate (MLE) for the given data: $\theta_{\mathrm{MLE}} = \arg\max_{\theta} \sum_{i=1}^{N} \log\left(\mathbf{p}(\mathbf{y_i} \mid \theta)\right)$

Custom EM algorithm for SMM-based recovery and clustering:

- MLE is intractable in this case.
- We developed an Expectation-Maximization (EM) approach, customized for SMM, to find a candidate MLE that:
  - is guaranteed to converge to a local maximum; and
  - does not require matrix inversion of large matrices.

Our custom EM algorithm is described in https://doi.org/10.48550/arXiv.2501.01840.



Motivation – Synthetic Experiment Reveals Higher Noise Tolerance

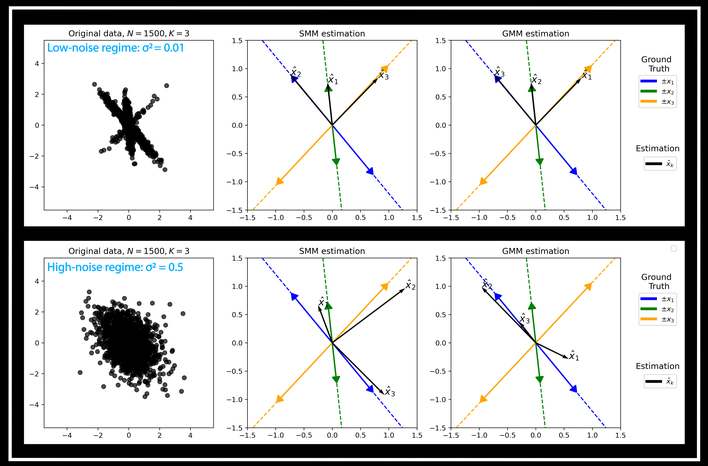
Goal: Compare SMM's performance to the more classical Gaussian Mixture Model (GMM)-based recovery.

Synthetic dataset:

- Number of subpopulations K = 3, so the goal is to recover underlying signals x_1, x_2, and x_3.
- Number of noisy measurements N = 1500, with noise variance σ^2, and two m/z bins per signal (d = 2 for visualization).
- Two noise levels explored:
  - Low-noise regime: $\sigma^2 = 0.01$ (top panel)
  - High-noise regime: $\sigma^2 = 0.5$ (bottom panel)

Observations:

- In the low-noise regime:
  - the signal recovery performance of SMM and GMM appear to be similar.
- In the high-noise regime:
  - the accuracy of both methods decreases, but GMM only finds one signal, x_1, and produces a third estimate, which is a mixture of two ground truth signals.
  - on the other hand, SMM delivers three clearly separate directions without the 'collapse' of estimates we that see in the GMM case.



Conclusions:

Under low-noise conditions, SMM performs comparably to GMM, but in high-noise regimes, where robust computational methods are most needed, SMM substantially outperforms GMM.

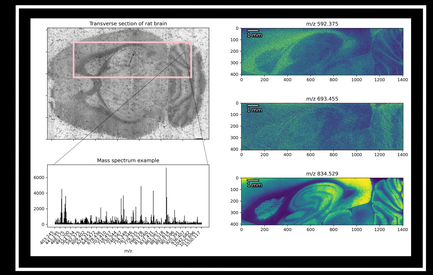
Overall, the SMM offers a means of making algorithms aware of structures and perturbations that naturally arise in MS measurements, enabling advanced noise resilience and aiding discovery of biological patterns that might lie hidden in the noise.

Results – Clustering of Rat Brain IMS dataset

Goal: Use SMM to recover underlying molecular signatures from real-world noisy IMS measurements of a rat brain tissue section, and to implicitly cluster the mass spectra, i.e., segment the tissue section into areas of similar chemical composition. Furthermore, we compare SMM's clustering results to ones given by traditional methods such as GMM and k-means clustering.

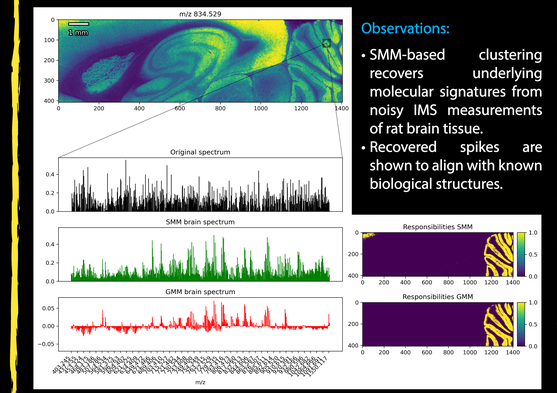
Rat brain IMS dataset:

A transverse rat brain section was measured using a timsTOF FleX (Bruker Daltonics) in QTOF mode across m/z 400-2,000, using a 10-µm pixel size and yielding 572,832 individual spectra. Spectral alignment was performed to correct for drift along the m/z domain. After alignment and calibration, an average mass spectrum based on all pixels in the dataset was computed. The average spectrum was peak-picked, 843 peaks were detected, and their corresponding ion intensities were retrieved.



Comparison of clustering results by SMM, GMM, k-Means Clustering:

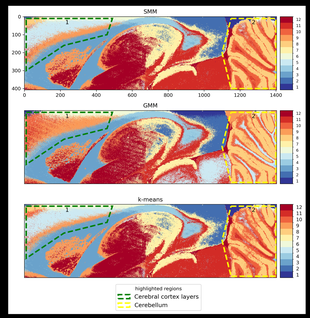
Spectral perspective



Observations:

- SMM-based clustering recovers underlying molecular signatures from noisy IMS measurements of rat brain tissue.
- Recovered spikes are shown to align with known biological structures.

Spatial perspective



Observations:

- Estimated responsibility variables segment the tissue according to molecular content. SMM retrieves histological patterns missed by other methods.
- SMM found subdivisions of cerebral cortex layers (e.g., molecular, granular, pyramidal, and multiform) that are known to align parallel to the surface of the brain, but that were missed by GMM.
- SMM delivered sharper delineation of anatomical structures (e.g., white matter, molecular layer, and granule cell layer) and exhibited less susceptibility to noise than k-means clustering.

1 Delft Center for Systems and Control, Delft University of Technology, Delft, Netherlands
2 Karlsruhe Institute of Technology, Institute of Industrial Information Technology, Karlsruhe, Germany
3 Mass Spectrometry Research Center, Vanderbilt University, Nashville, TN, USA
4 Chemical Physical Biology Program, Vanderbilt University, Nashville, TN, USA
5 Department of Cell and Developmental Biology, Vanderbilt University, Nashville, TN, USA
6 Department of Biochemistry, Vanderbilt University, Nashville, TN, USA
7 Department of Chemistry, Vanderbilt University, Nashville, TN, USA

Research was supported by the National Institutes of Health (NIH)'s NIDDK (U54DK134302 and U01DK133766), NEI (U54EY032442), NIAID (R01AI138581 and R01AI145992),NIA (R01AG078803), NCI (U01CA294527), and by the Chan Zuckerberg Initiative DAF (2021-240339 and 2022-309518). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors declare no competing financial interest.

TU Delft
VANDERBILT UNIVERSITY