Detecting latent spatial patterns in mass spectrometry brain imaging data via Bayesian mixtures

Individuazione di schemi spaziali latenti in dati di immagini cerebrali di spettrometria di massa tramite misture bayesiane

Giulia Capitoli[a], Simone Colombara[b], Alessia Cotroneo[b], Francesco De Caro[b], Riccardo Morandi[b], Chiara Schembri[b], Alfredo G. Zapiola[b], and Francesco Denti[c]

[a]Univerisity of Milan-Bicocca; giulia.capitoli@unimib.it

[b]Politecnico of Milan

[c]Università Cattolica del Sacro Cuore - Milan

Abstract

Mass spectrometry methods can record biomolecules abundance for a broad set of molecular masses given a sample of a specific biological tissue. In particular, the MALDI-MSI technique produces imaging data where, for each pixel, a mass spectrum is recorded. There is the urge to rely on suited statistical methods to model these data, fully addressing their morphologic characteristics. Here, we apply a Bayesian mixture model to segment these real biomedical images. We aim to detect groups of pixels that present similar patterns to extract interesting insights, such as anomalies that one cannot capture from the original pictures. This task is particularly challenging given the high dimensionality of the data and the spatial correlation among pixels. To account for the spatial nature of the dataset, we rely on Hidden Markov Random Fields.

I metodi di spettrometria di massa possono registrare l'abbondanza per un'ampia serie di masse molecolari in un campione di uno specifico tessuto biologico. In particolare, la tecnica MALDI-MSI produce dati di imagine in cui, per ogni pixel, viene registrato uno spettro di massa. È necessario affidarsi a metodi statistici adeguati per modellare questi dati, tenendo conto delle loro caratteristiche morfologiche. Qui applichiamo un modello di miscela bayesiano per segmentare queste immagini biomediche reali. Il nostro obiettivo è individuare gruppi di pixel che presentano modelli simili per estrarre spunti interessanti, come le anomalie che non si possono cogliere dalle immagini originali. Questo compito è particolarmente impegnativo data l'elevata dimensionalità dei dati e la correlazione spaziale tra i pixel. Per tenere conto della natura spaziale del set di dati, ci affidiamo ai campi casuali nascosti di Markov.

Keywords: Mass spectrometry, Bayesian Mixture models, Potts Model, Brain imaging.

1. Introduction

Mass spectrometry imaging (MSI) is an emerging technology capable of mapping various biomolecules within their native spatial context. This work describes an application to spatial multi-omics data ob-

tained via the state-of-the-art MALDI-MSI technology. MALDI-MSI uses a laser energy-absorbing matrix to create ions from large molecules with minimal fragmentation. In other words, this technology visualizes the distribution of molecules such as peptides, lipids, and glycans in a biological sample [10]. Here, we consider the sequential MALDI-MS imaging of lipids, glycans, and tryptic peptides on a single mouse brain formalin-fixed paraffin-embedded (FFPE) tissue section. For this work, a slice from the mouse's brain is partitioned into a grid of pixels. A low resolution depiction of the biological sample we consider is showed in panel (a) of Figure 1. Then, MALDI-MSI acquires a mass spectrum for each pixel. In particular, MALDI-MSI creates spectra pixel-by-pixel by measuring the mass-to-charge (*m/z*) values representing analytes of interest along the *x*-axis and the corresponding abundance along the *y*-axis [3]. Finally, a three-dimensional MSI dataset is obtained by arranging the mass spectra and the grid coordinates for each pixel to be investigated in further analysis. Due to experimental limitations, the MALDI-MSI spectra can contain noise affecting the statistical analysis. To mitigate this issue, a well-established biological pre-processing step has to be performed. In this contribution, we aim to segment the pixels obtained with the MALDI-MS imaging technique into biologically meaningful clusters employing Bayesian mixture models, focusing on *lipids*. Our article proceeds as follows. In the next subsection, we introduce the dataset we are considering. In Section 2, we describe our modeling approach, while in Section 3, our results are presented and discussed. Finally, in Section 4, we discuss future directions and conclude.

1.1 Data description and statistical preprocessing

The raw dataset produced by the MALDI-MSI technology undergoes some initial preprocessing steps to filter out the noise. The steps are baseline correction, smoothing, normalization, spectra alignment, peak detection and extraction. In addition, the preprocessing needs to reduce the intra-sample variability and correct for analytical and instrumental variability following sample preparation to ensure accurate *m/z* localization. For more details, see, for example, [4]. The resulting dataset for our statistical analysis has the following characteristics. The lipid image of the biological tissue is analyzed with a raster of 50 μm resulting in a total of about 18,000 laser shots (defining the pixels). In other words, we obtain thousands of spectra, each with its pixel coordinates (s_1,s_2). Recall that, for each pixel, the abundance of the lipids as a function of different values of molecular masses (*m/z*) are recorded, resulting in potentially hundreds of variables. Therefore, given the data's large dimensionality, as the first step, we perform pixel-wise functional principal component analysis [9]. We then consider only the first functional principal component loadings (fPCl) for each pixel. We report such distributions of fPCls in panel (b) of Figure 1. The shape of the distribution suggests using Gaussian mixtures to segment the picture of the image.

2. Gaussian Mixtures and Potts Models

Here, we describe the model-based clustering approach we adopt for the pre-processed data presented in the previous section. We estimate clustering solution via mixture models. Consider the vector $\boldsymbol{y} = (y_1, \ldots, y_n)$, where $y_i \in \mathbb{R}$ denotes the value of the first fPCl assumed by i-th pixel. Moreover, denote with $\mathcal{N}_i$ the set of all the neighboring (adjacent) pixels to the i-th pixel. Assuming the different pixels are fully exchangeable (i.e., ignoring any spatial relation), one can specify the following basic, univariate Gaussian mixture likelihood with a fixed number K of components:

$$p(y_i \mid \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k, \sigma_k^2), \qquad i = 1, \ldots, n, \tag{1}$$

where $\phi(m, s^2)$ denotes a Normal density with mean m and variance s^2 and $\boldsymbol{w} = (w_1, \ldots, w_K)$ is the collection of mixture weights. To complete our model specification in a Bayesian setting, one can adopt a Dirichlet prior for the mixture weights, and a Normal and Gamma priors for the mixture components' means and variances, respectively. In formulas, $\boldsymbol{w} \sim Dir_K(\boldsymbol{\lambda})$ and for $1, \ldots, K$, we assume $\mu_k \sim Normal(m, s^2)$ and $\sigma_k^2 \sim InvGamma(a, b)$. As customary with mixtures, we can augment

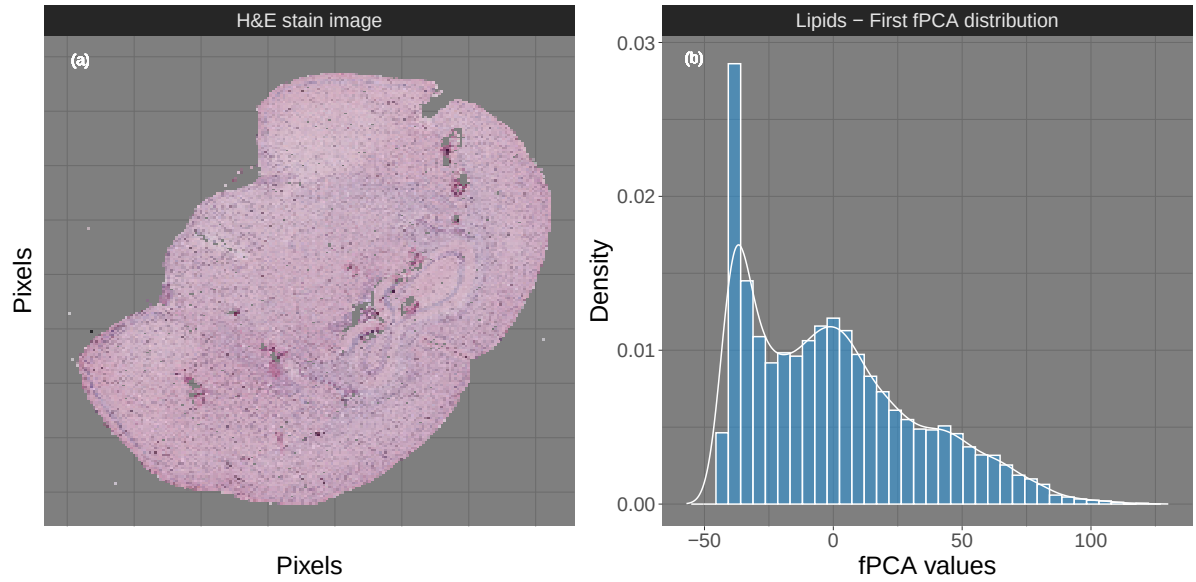

Figure 1: Panel (a): Hematoxylin and Eosin (H&E) staining image (low resolution). The reported mouse brain slice underwent digital scanning. Panel (b): The distribution of the first fPCl considered for our analysis. The shape suggests using a mixture of normal distributions to model the data.

the likelihood specification (1) adding a set of n latent membership labels $\{z_i\}_{i=1}^n$. Each membership labels has a discrete distribution with support over $1, \dots, K$, and $z_i = k$ implies that the i-th observation has been assigned to the k-th cluster. The augmented likelihood can be expressed as, for $i = 1, \dots, n$:

$$y_i \mid z_i, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \sim Normal(\mu_{z_i}, \sigma^2_{z_i}), \qquad p(z_i|\boldsymbol{w}) = \sum_{k=1}^{K} w_k \, \delta_k(z_i), \tag{2}$$

where $\delta_i(j)$ is the delta function, equal to 1 when $i = j$ and 0 otherwise. Note how y_i only depends on its correspondent latent membership z_i. It becomes evident how the addition of the latent variables considerably simplifies inference, usually carried out via MCMC techniques. We apply model (2) to set a benchmarking result. Despite the Bayesian Gaussian mixture model (GMM) presents a fully probabilistic extension to the classical K-means algorithm, the above specification completely disregard the available spatial information. Intuitively, especially when performing image segmentation, we expect that neighboring pixels should have an higher probability of belonging to the same cluster.
To introduce such spatial relation in the model, we can rely on Hidden Markov Random Fields. In particular, following [1; 2] we can intervene on the distribution of the membership labels, introducing dependence of each pixel i on the sets of its neighbors $\mathcal{N}_i$. We assume a Gibbs distribution for the vector $\boldsymbol{z}$ which can be specified via the following conditional probability statement:

$$p(z_i|\boldsymbol{z}_{-i}, \beta) \propto \exp(\beta \sum_{j \in \mathcal{N}_j} \delta_{z_i}(z_l)) \tag{3}$$

where $\mathbf{z_{-i}}$ denotes all the variables in $\boldsymbol{z}$ without the i-th, and β is the *inverse temperature parameter*, which defines the strength of the spatial connection. Tuning and simulating the β parameter is challenging due to the doubly-intractable nature of its full conditional distributions, More importantly, the Potts model undergoes a phase transition, switching from a disordered to an ordered state. The phase transition happens as β exceeds a threshold β^*, also called *critical value*. The critical value for a regular 2D lattice is given by $\beta = \log(1 + \sqrt{K})$ [6; 8]. For simplicity, in this application, we assume a fixed value of inverse temperature, setting $\beta = \beta^*$. To fit both the GMM and the Hidden Potts model (HPM), we rely on the R package `bayesImageS` [5].

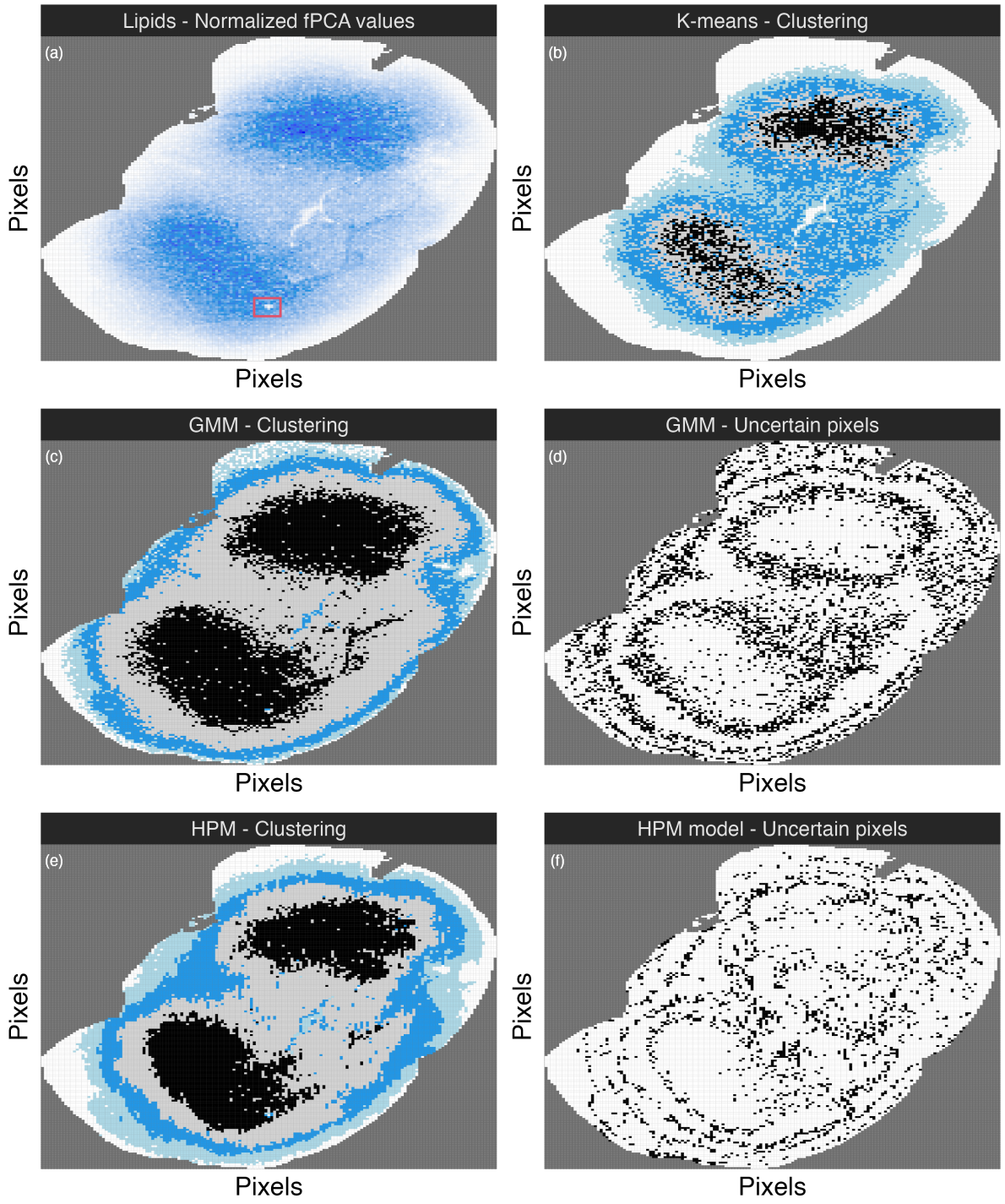

Figure 2: Each panel presents a pixellated image of the considered brain slice colored according different results. (a) Intensities of the (normalized) first fPCl; (b) K-means clustering results; (c) GMM modal clustering results and (d) corresponding most uncertain pixels (e) HPM modal clustering results and (f) corresponding most uncertain pixels.

3. MALDI-MSI Bayesian Segmentation Results

All the results described in this section are summarized in Figure 2. Panel (a) shows the fPCl values across the pixels. Spatial patterns are apparent, highlighting the hippocampal formation (light blue), the white (high values in the blue scale), and gray matter (white pixel). Biologically, we expect to ob-

serve different structures highlighting regions of the gray matter and the cerebral cortex. Moreover, the Thalamus and Hypothalamus are expected to be distinct within the inter-brain formations. Led by this biological information, we set the number of mixture components K equal to 5.
To benchmark our mixture models' results, we also fit a K-means algorithm on the first fPCl values. The resulting partition is reported in panel (b). The partition is fuzzy in the central parts of the slices, and no clear pattern is evident.
We then fit both the GMM and the HPM. No label switching problem was detected: therefore, we estimate the clustering membership for each pixel as $\hat{c}_i = \arg\max_k f_{i,k}$, where $f_{i,k}$ is the proportion of iterations in which the MCMC allocated pixel i in cluster k. In panels (c) and (e), we report the results of the GMM and the HPM, respectively. The addition of spatial information helps the model detect clearer boundaries across the different regions of the picture.
Moreover, we compute a Gini-Simpson index for each pixel as $g_i = 1 - \sum_{k=1}^{K} f_{i,k}^2$, summarizing the uncertainty of the model in the allocation of a specific pixel. Indeed, a large value of g_i implies that, across all the MCMC iterations, the pixels have been assigned to multiple clusters with similar probability, making $\hat{c}_i$ an unreliable estimate. In panels (d) and (f), we flag as uncertain (in black) all the pixels for which $g_i > 0.4$.
These last two panels showed a discrete separation between the inter-brain formations (hippocampal white matter regions) and the external ones (gray matter). Since uncertain pixels correspond to cells that fall in the separation border between clusters, the separation enhances as we add spatial information. In fact, the number of pixels flagged as uncertain decreases from panels (d) to (f). The presence of uncertain pixels has a biological explanation. The laser diameter used by the MALDI-MSI in this analysis was 50 μm, while the average dimension of a single cell is $\approx 10\mu m$. Therefore, pixels at the border between two brain regions may contain cells of multiple natures that hinder their classifications.
Finally, the histological sample shows some anomalies (see, for example, the pixels highlighted by the red rectangle in panel (a) of Figure 2). These pixels correspond to pieces of tissue that detached from their original position and moved during sample preparation and analysis. The GMM correctly classifies these pixels by their histological information. While retaining these anomalies, we can appreciate how the GMM eliminates impurities better than the K-means. Instead, the HPM manages to smooth out this noise, making the results of the clusters more homogeneous.

4. Conclusions

The results presented in this contribution are promising, and many future directions can be pursued. First, one can enhance the modeling aspect by allowing the estimation of a stochastic β using, for example, the pseudo-likelihood approach or the exchange algorithm [7].
Second, one could extend the likelihood specification to handle multivariate measurements. This extension would allow the joint modeling of multiple functional principal components, increasing the amount of information preserved after our dimensionality reduction step.
Alternatively, these findings also encourage the development of statistical methods that can be used to jointly model lipids, peptides, and glycans, uncovering hidden molecular patterns resulting from the relationship between these multiple molecular levels. Integrating the three molecular data sets could improve the separation across distinct histopathological regions of interest of the mouse brain sections.
Lastly, one could focus on the original image of the biological tissue, modeling its RGB encoding to detect anomalies in its morphological characteristics patterns.
We plan to explore these research avenues in the future.

References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.
[2] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series*

B (Methodological), 48:259–302, 1986.

[3] K.J. Boggio, E. Obasuyi, K. Sugino, S.B. Nelson, N.Y.R. Agar, and J.N. Agar. Recent advances in single-cell maldi mass spectrometry imaging and potential clinical impact. *Expert Rev Proteomics*, 8(5):591–604, 2011. doi:10.1586/epr.11.53.

[4] V. Denti, G. Capitoli, I. Piga, F. Clerici, L. Pagani, L. Criscuolo, G. Bindi, L. Principi, C. Chinello, G. Paglia, F. Magni, and A. Smith. Spatial Multiomics of Lipids, N-Glycans, and Tryptic Peptides on a Single FFPE Tissue Section. *Journal of Proteome Research*, 21(11):2798–2809, 2022.

[5] M.T. Moores, D. Feng, and K. Mengersen. bayesImageS: Bayesian Methods for Image Segmentation using a Potts Model. *R package (v0.6-1)*, 2021.

[6] M.T. Moores, G.K. Nicholls, A.N. Pettitt, and K. Mengersen. Scalable bayesian inference for the inverse temperature of a hidden potts model. *Bayesian Analysis*, 15(1):1–27, 2020.

[7] I. Murray, Z. Ghahramani, and D.J.C. MacKay. MCMC for doubly-intractable distributions. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, pages 359–366, 2006.

[8] R. B. Potts. Some generalized order-disorder transformations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(1):106–109, 1952.

[9] Silverman B.W. Ramsay J.O. *Functional data analysis*. Springer New York, NY, New York, 2nd edition, 2005.

[10] T.C. Rohner, D. Staab, and M. Stoeckli. Maldi mass spectrometric imaging of biological tissue sections. *Mech Ageing Dev*, 126(1):177–185, 2005. doi:10.1016/j.mad.2004.09.032.