PhD projects R3-5 and R3-8

Statistical Procedures and Numerical Algorithms for Analyzing MALDI Data
Two-Sample Testing for the Comparison of MALDI Spectra

J. von Schroeder, V. Vutov and T. Dickhaus

RTG 2224

Statistical Analysis of MALDI Data

- Spectra obtained using Matrix-Assisted Laser Desorption/Ionization (MALDI) imaging
- Observations $x \in \mathbb{R}^d$ with d between 2.000 and 100.000
- Each entry x_i is proportional to the number of ions detected at a specific m/z (mass-to-charge) ratio.

Statistical Analysis of MALDI Data - Modelling

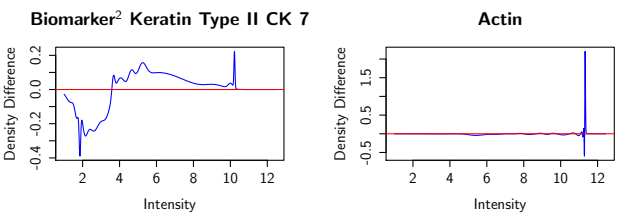
- Total number of ions is large[1] ($\approx 10^9$)
- Number of ions correlated with concentration of analyte[1]

Marginal approximation of the log-likelihood

$$\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} p_{j,k} \cdot \underbrace{\phi\left(x_j^{(i)} \,|a_i\mu_{j,k}, a_i\sigma_{j,k}^2\right)}_{\textbf{Gaussian approximation}} \right) + \underbrace{\log R(\sigma_j)}_{\textbf{Regularization}} \tag{1}$$

$x^{(1)}, \ldots, x^{(n)}$ independent observations, $a_i := \left\|x^{(i)}\right\|_1$ and **penalty** R is an inverse gamma prior on the variance

- Observations not identically distributed
- ML estimates for $p_j, \mu_j, \sigma_j \in \mathbb{R}^K$ can be obtained by applying Expectation-Maximization to (1).
- **Vectorized implementation** on the GPU



Statistical Analysis of MALDI Data - Feature Selection

Goal: Find m/z values that differ between samples
↪ Potential biomarkers

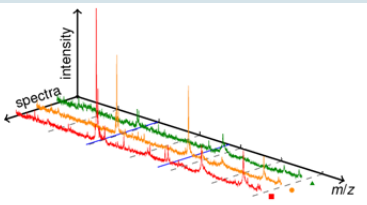
Developed Approach[3] (FDR-controlling Feature Selection)

1. Normalize variance trend
2. Subsampling bootstrap of test statistic[4] (or of AUC)
3. FDR control (BY) to determine number of features

Supervised learning (cross-validated accuracy reported)

Method/Data set	Leuschner et al. (2018)	Task ADSQ (Spot) of Behrmann et al. (2017)
IsotopeNet[5]		0.845
Flog_int (60 Features)[6]	0.899	
FDR-controlling FS[3] + RF	**0.926**	**0.866**

2-sample testing for the comparison of MALDI spectra



- Compute effective numbers of tests for two-sample or k-sample comparison problems in the context of MALDI.[7]
- Multiplicity- and dependency-adjustment method (MADAM) for control of the family-wise error rate [8]
- Supervised classification
- Feature Selection
- Arbitrary dependence structure
- Copula-based methods
- Resampling-based methods
- Multiple type I error rates
  - Family-wise error rate [9]
  - False discovery rate [10]
  - False discovery exceedance
- Likelihood-Ratio test based on (1)

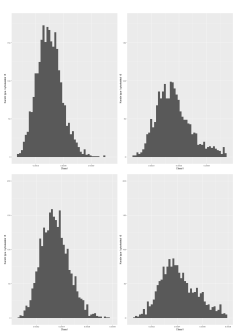
Elements of the project

An effective number of test of order i, given by
$M_{eff}^{(i)} \equiv M_{eff}^{(i)}(\alpha_{loc}, \mathbf{T}) = 1 + \epsilon(i) + \sum_{j=i\vee 2}^{M} \kappa_j^{(i)}$

The $MADAM_i$ transforms the values of the test statistics $T_1, T_2, \cdots, T_M$ into one of the following multiplicity- and dependency-adjusted p-values.

$$p_{\Sigma,j}^{MADAM_i}(x) = b^{(i)}(\mathbb{P}, t_j)$$
$$p_{\prod,j}^{MADAM_i}(x) = 1 - \beta^{(i)}(\mathbb{P}, t_j)$$

where the order of *i* can be considered as a tuning parameter of the MADAM approach.



Do the distributions between classes differ?

[1] Bae et al., (2012). Degree of Ionization in MALDI of Peptides: Thermal Explanation for the Gas-Phase Ion Formation. J. Am. Soc. Mass Spectrom., 23(8), 1326–1335. doi:10.1007/s13361-012-0406-y
[2] Kriegsmann et al., (2016). Reliable Entity Subtyping in Non-small Cell Lung Cancer by Matrix-assisted Laser Desorption/Ionization Imaging Mass Spectrometry on Formalin-fixed Paraffin-embedded Tissue Specimens. MCP, 15(10), 3081–3089. doi:10.1074/mcp.M115.057513
[3] von Schroeder, J. (2020). Stable Feature Selection for MALDI Imaging Mass Spectrometry Data. Manuscript in preparation.
[4] Chatterjee, S. (2020). A New Coefficient of Correlation. J. Am. Stat. Assoc., 1–26. doi:10.1080/01621459.2020.1758115
[5] Leuschner et al., (2018). Supervised Non-Negative Matrix Factorization Methods for MALDI Imaging Applications. Bioinformatics, 35(11), 1940–1947. doi:10.1093/bioinformatics/bty909
[6] Behrmann et al., (2017). Deep Learning for Tumor Classification in Imaging Mass Spectrometry. Bioinformatics, 34(7), 1215–1223. doi:10.1093/bioinformatics/btx724
[7] Dickhaus and Stange, (2013). Multiple Point Hypothesis Test Problems and Effective Numbers of Tests for Control of the FWER. Calcutta Stat. Assoc. Bull., 65(1–4), 123-144. doi:10.1177/0008068320130108
[8] Stange et al., (2016). Multiplicity- and Dependency-Adjusted P-Values for Control of the Family-Wise Error Rate. Statistics & Probability Letters, 111, 32-40. doi:10.1016/j.spl.2016.01.005
[9] Dickhaus, T., (2014). Simultaneous Statistical Inference. Springer Berlin Heidelberg. doi:10.1007/978-3-642-45182-9.
[10] von Schroeder and Dickhaus, (2020). Efficient Calculation of the Joint Distribution of Order Statistics. CSDA, 144, 106899. doi:10.1016/j.csda.2019.106899

DFG Universität Bremen