Supplementary material: Stronger findings from mass spectral data through multi-peak modeling

Tommi Suvitaival*[1], Simon Rogers†[2] and Samuel Kaski‡[1,3]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University
[2]School of Computing Science, University of Glasgow
[3]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki

March 18, 2014

Abstract

This document is the Supplementary Material document for the publication titled *Stronger findings from mass spectral data through multi-peak modeling* [6].

An R implementation and data are available at http://research.ics.aalto.fi/mi/software/peakANOVA/.

1 Software versions

- MZmine 2.9.1 [5] modified to export peak shape correlations
  - The modified version is available at http://research.ics.aalto.fi/mi/software/peakANOVA/
- R 2.15.1
- PeakANOVA package
  - Available at http://research.ics.aalto.fi/mi/software/peakANOVA/

2 Simulated data

2.1 Data generation

- Data was generated from Model 1 with same parameters as with which the learning was done

*tommi.suvitaival@aalto.fi
†simon.rogers@glasgow.ac.uk
‡samuel.kaski@aalto.fi

- Example illustration of an distribution of the peak shape similarity values: Figure 1

2.2 Inference of the clustering

- Burn-in: 1,000 Gibbs samples
- Sampling after burn-in: 1,000 Gibbs samples

2.3 Inference of covariate effects

- Burn-in: 10,000 Gibbs samples
- Sampling after burn-in: 10,000 Gibbs samples
- Thinning: Every 10th Gibbs sample saved

2.4 Significance of covariate effects

- Inferred covariate effect was deemed significantly positive/negative if at least 95 % of the probability mass of the posterior distribution was above/below zero

2.5 Test of difference between the approaches

- Posterior mean of the change computed for each cluster/peak from the Gibbs samples
- Squared error to the ground-truth change (0, 0.5, -1, or 2) is computed for each cluster/peak by both the approaches
- Statistical significance of the difference between the mean squared errors (MSE) of the two approaches is tested by one-sided paired t-test
  - Null hypothesis: No difference between the MSEs
  - Alternative hypothesis: The MSE of Model 1 is smaller
  - False discovery rate of the test was controlled by the Benjamini-Hochberg step-up procedure [1] at level 0.01

3 Benchmark data set with known changes of concentration

3.1 Data

- Data and the experiment are described in the publication [3]

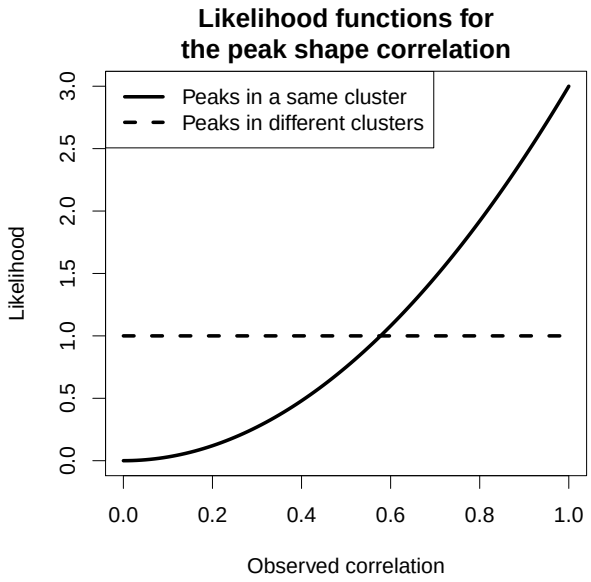

Figure 1: Probability density of the peak shape correlations. Peaks in same cluster have a high probability for high values of their mutual correlation (solid line). Peaks in different clusters are expected to have a low or entirely unobserved value of peak shape correlation (dashed line).

3.1.1 Raw mass spectrometry data

- Used to extract the peak intensities and peak shape correlations
- Available at
  http://cri.fmach.eu/Research/Computational-Biology/
  Biostatistics-and-Data-Management/download/data/
  Spiked-Apple-Data. [Accessed 11.06.2013.]

3.1.2 Pre-processed and annotated intensity data

- Used to extract the peak locations, annotations and ground-truth perturbations
- Intensity data not used
- Available in the R-package `BioMark`
  - Version 0.4.1 used
  - Available at
    http://cran.r-project.org/web/packages/BioMark/

3.2 Pre-processing in MZmine

3.2.1 Peak detection

- Peak list file:

  - m/z and RT values of the detected peaks from the R-package `BioMark`
  - `apple-peak_list-positive.csv`
  - `apple-peak_list-negative.csv`
- Field separator: `,`
- Ignore first line: `No`
- Intensity tolerance: `30.0` %
- m/z tolerance: `0.0010` m/z or `5.0` ppm
- Retention time tolerance: `1.0` *absolute (min)*

3.2.2 Peak alignment

- m/z tolerance: `0.0020` m/z or `10.0` ppm
- Weight for m/z: `70`
- Retention time tolerance `2.0` *absolute (min)*
- Weight for RT: `50`
- Require same charge state: `No`
- Require same ID: `Yes`
- Compare isotope pattern: `No`

3.2.3 Peak shape correlations

- Modified version of MZmine 2 used
- Time window: `0.05` (min)

3.3 Clustering

- Parameters of the beta prior distribution for the observed shape correlation values
  - Within a cluster: $a_{\text{in}} = 2$, $b_{\text{in}} = 1$
  - Between clusters: $a_{\text{out}} = 1$, $b_{\text{out}} = 1$
- Prior probability of missing value in the shape correlation data
  - Within a cluster: $p_0^{\text{in}} = 0.25$
  - Between clusters: $p_0^{\text{out}} = 0.99$
- Least-squares clustering is picked from the set of Gibbs samples as proposed by [2] (Eq. 10.10) and used for the analysis that follows
- If cluster k contains annotated peaks from compound c, it is annotated as the compound c

3.4 Covariate effects

3.4.1 Pre-processing

- Log-transformation of the intensity data (natural logarithm)
- Mean of the control sample group subtracted from the intensity data of each peak (control group-based centering)
- Peaks with a standard deviation of zero (constant observed value) throughout the samples removed from the data

3.4.2 Analysis

- Model 1: Posterior mean of the covariate effects computed from the Gibbs samples
- Single-peak approach:
  - Difference of means between the treatment and control group is computed for each cluster and treatment group
  - For annotated clusters, each annotated peak is used
  - For non-annotated clusters, the strongest peak by the mean of the control group (before subtracting the mean at the pre-processing) is used
- Comparison of performance between the approaches
  - Squared error to the ground truth change (0, 0.2, 0.4 or 1.0) is computed for each cluster/peak by both the approaches
  - Statistical significance of the difference between the mean squared errors (MSE) of the two approaches is tested by one-sided paired t-test
    - Null-hypothesis: No difference between the MSEs
    - Alternative hypothesis: The MSE of Model 1 is smaller

4 Lipidomic data from a gene silencing study

4.1 Data

- Ultra performance liquid chromatography-mass spectrometry (UPLC)
- Ion mode: negative
- Experimental setup
  - Silenced genes:
    - ACACA (acetyl-CoA carboxylase α)
    - ELOVL1 (elongation of very long chain fatty acid-like 1)
    - FASN (fatty acid synthase)
    - INSIG1 (insulin-induced gene 1)

    - SCAP (sterol regulatory element-binding protein cleavage-activating protein)
    - SCD (stearoyl-CoA desaturase)
    - THRSP (thyroid hormone-responsive protein)
    - (Ineffective control silencing)
  - Time points:
    - 48 hours
    - 72 hours
  - Replicates in each treatment-time point category: 2
  - Total number of samples: 32
- The raw data is available at http://research.ics.aalto.fi/mi/software/peakANOVA/
- Data and the experiment described in more detail in the publication [4]

4.2 Pre-processing in MZmine 2

4.2.1 Peak detection

- Default settings

4.2.2 Peak alignment

- Default settings

4.2.3 Peak shape correlations

- Modified version of MZmine 2 used
- Time window: `0.05` (min)

4.3 Normalization

Intensities of a sample were normalized by

- intensities of standard compounds in the sample (standard compound normalization)
- amount of protein contained in the sample (tissue normalization)

4.4 Clustering

- Parameters of the beta prior distribution for the observed shape correlation values
  - Within a cluster: $a_{\text{in}} = 3$, $b_{\text{in}} = 1$
  - Between clusters: $a_{\text{out}} = 1$, $b_{\text{out}} = 1$
- Prior probability of missing value in the shape correlation data

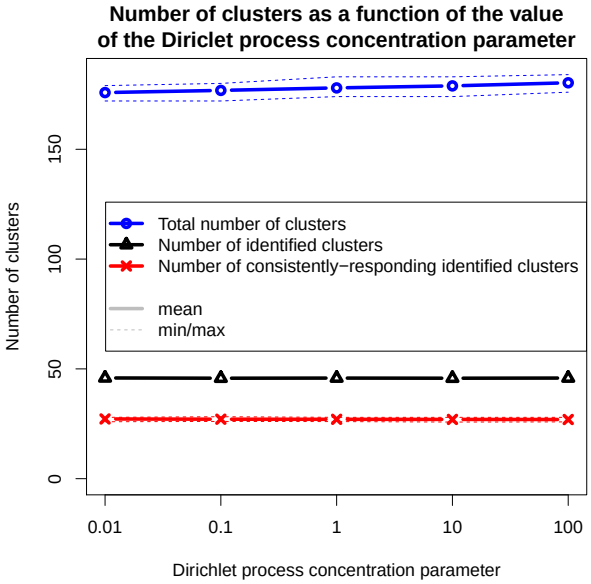

Figure 2: The clustering inferred from the lipidomic gene silencing data is insensitive to the value of the Dirichlet process concentration parameter (α_{DP}) and thus also the inferred covariate effects remain unchanged by the parameter. Total number of clusters (blue line) increases slightly as the number as the value of parameter increases on the logarithmic scale. The performance – the proportion of consistently-responding clusters (red line) among all identified clusters (red line) – remains unchanged.

  - Within a cluster: $p_0^{\mathrm{in}} = 0.01$
  - Between clusters: $p_0^{\mathrm{out}} = 0.5$

- Least-squares clustering is picked from the set of Gibbs samples as proposed by [2] (Eq. 10.10) and used for the analysis that follows

4.5 Stability of clustering

- Dirichlet process concentration parameter grid:
  $\alpha_{\mathrm{DP}} = \{0.01, 0.1, 1, 10, 100\}$
- Result: Figure 2

4.6 Covariate effects

- Pre-processing
  - Log-transformation of the intensity data (natural logarithm)
  - Mean of the control sample group subtracted from the intensity data of each peak (control group-based centering)
  - Peaks with a standard deviation of zero (constant observed value) throughout the samples removed from the data

4.7 Consistency of effects

- Families of lipids included and the number of annotated lipids with two or more peaks identified:
  - Phosphatidylcholines (PC): 7
  - Phosphatidylethanolamines (PE): 4
  - Sphingomyelins (SM): 3
- Cross-validation
  - Number of folds: 3
  - Number of randomizations of the fold assignments and artificial noise: 100

4.8 Robustness of effects

- All identified peaks included (both annotated and non-annotated)
- Number of randomizations of artificial noise: 100

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met*, 57(1):289–300, 1995.

[2] D. B. Dahl. *Bayesian Inference for Gene Expression and Proteomics*, chapter Model-based clustering for expression data via a Dirichlet process mixture model, pages 201–218. Cambridge University Press, Cambridge, 2006.

[3] P. Franceschi, D. Masuero, U. Vrhovsek, F. Mattivi, and R. Wehrens. A benchmark spike-in data set for biomarker identification in metabolomics. *J Chemometr*, 26(1-2):16–24, 2012.

[4] M. Hilvo, C. Denkert, L. Lehtinen, B. Müller, S. Brockmöller, T. Seppänen-Laakso, J. Budczies, E. Bucher, L. Yetukuri, S. Castillo, E. Berg, H. Nygren, M. Sysi-Aho, J. Griffin, O. Fiehn, S. Loibl, C. Richter-Ehrenstein, C. Radke, T. Hyötyläinen, O. Kallioniemi, K. Iljin, and M. Orešič. Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression. *Cancer Res*, 71(9):3236–3245, 2011.

[5] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1):395, 2010.

[6] T. Suvitaival, S. Rogers, and S. Kaski. Stronger findings from mass spectral data through multi-peak modeling. Submitted.