UNIVERSITY OF AMSTERDAM

*New Methods for Modelling and Data Analysis in Gas Chromatography: a Bayesian View.*

A. Barcaru

Andrei Barcaru

# New methods for modelling and data analysis in Gas Chromatography: a Bayesian view

## Summary

One of the major aspects in gas chromatography data analysis is the link of the evidences from the data with the previous knowledge or “believes” upon the information being extracted. Taking into account the errors provided by the instrumentation the answer itself should have a sort of degree of “possibility”. This is the essence of the Bayesian approach. The answer provided by this framework is probabilistic and may seem to be sometimes difficult to operate with. However, a probabilistic answer is often preferred to a deterministic one when the cost function is high or when the uncertainty upon the estimated parameter, hypothesis under test or value of interest is high. The aim of this thesis is to bring the benefits of the Bayesian statistics into the field of gas chromatography.

In the second chapter is presented a computational physical model of GC×GC, capable accurately to predict retention times in both dimensions. Once fitted, the model is used to make predictions, which are always subject to error. Hence, the prediction can result rather in a probability distribution of (predicted) retention times than in a fixed (most likely) value. Implementing the Bayesian paradigm with a flat prior, the confidence intervals are approximately the same as the credible intervals. In other words, the confidence interval in this case actually give, with a good approximation, the 95% probability to have the predicted value within the outlined limits. Overfitting is one of the major concerns when fitting unknown parameters. In the second chapter is described the application of the k-fold cross-validation technique to avoid the problem of overfitting and to assess the errors of the predictions. Another technique of error assessment proposed in the same chapter is the use of error propagation using Jacobians. The robustness of any optimization algorithm is considerably improved if the predictions are regarded as intervals rather than precise values.

In chapter three a novel peak tracking method based on Bayesian statistics is proposed. The method described in this chapter consist of a probabilistic assignment of peaks between two GCxGC-FID peak tables of the same sample taken in different conditions. That is to say, the result of the algorithm is a list of possible candidates ranked by the posterior probability of matching. The algorithm proved to be fast and accurate (78% of the selected peaks were ranked with maximum posterior value).

The forth chapter describes an algorithm aimed to highlight in a probabilistic way the differences between two GCxGC-TOFMS data sets. One sample being considered the “reference sample” and the other, with potential adducts, is the “query sample”. The comparison is based on the evaluation of the Jensen-Shannon divergence between subsets of data from both chromatograms (i.e. moving windows) which eliminates the misalignment problem. The probabilistic answer is further provided by the use of Bayesian factor. In the same chapter was proven that this approach is a versatile tool in gcxgc-ms data analysis, especially when the differences are embedded inside a complex matrix. The algorithm was tested on diesel samples.

In the fifth chapter a complete, probabilistic deconvolution algorithm for high resolution GC-Orbitrap data is proposed. The novelty of the method is the fact that the problem of the number of components is tackled in a Bayesian probabilistic way. The algorithm first estimates the retention times, peak with and the peak height for each mass-channel (i.e. for each ion). The statistical dependency between m/z channels was implied by including penalties in the objective function. Further, the Expectation Maximization algorithm is used to cluster the compounds in the “retention time – peak with” space. Bayesian Information Criterion (i.e. BIC) was used as Occam’s razor for the probabilistic assessment of the number of components. The result of the algorithm is a list of possible components with a ranking value associated to each compound in the list.