A Novel Bayesian Approach for Uncovering Potential Spectroscopic Counterparts for Clinical Variables in ^{1}H NMR Metabonomic Applications

Aki Vehtari[1*], Ville-Petteri Mäkinen[1], Pasi Soininen[2], Petri Ingman[3], Sanna Mäkelä[4], Markku Savolainen[4], Minna Hannuksela[4], Kimmo Kaski[1], and Mika Ala-Korpela[1*]

[1] Laboratory of Computational Engineering, Systems Biology and Bioinformation Technology, Helsinki University of Technology, P.O. Box 9203, FI-02015 HUT, Finland; [2] Department of Chemistry, University of Kuopio; [3] Department of Chemistry, University of Turku; [4] Department of Internal Medicine, University of Oulu, Finland.
{*Aki.Vehtari, *Mika.Ala-Korpela}@hut.fi

Abstract. Metabonomic approaches based on spectroscopic data are in their infancy in biomedicine. A key challenge in clinical metabonomics is uncovering and understanding the relations between the multidimensional spectroscopic data and the clinical measures currently used for disease risk assessment and diagnostics. A novel Bayesian approach for revealing clinically relevant signals is presented here for a real ^{1}H NMR metabonomics data set. The results are not only mathematically superior but also biochemically fully coherent.

Keywords: Bayes; Metabonomics; NMR; Spectroscopy; Clinical.

1 Background

Genomics is increasingly complemented by *metabonomics* – the quantitative measurement of the time-related multiparametric metabolic responses of multicellular systems to pathophysiological stimuli or genetic modification [1]. To this end, mass and nuclear magnetic resonance (NMR) spectroscopy have become the two key technologies. An appealing feature of NMR spectroscopy for metabonomic applications is its specific yet non-selective nature: proton (^{1}H) NMR can efficiently obtain information on a large number of metabolites in biological samples like human serum. The abundance of protons and the inherently narrow as well as heterogeneous chemical shift range of ^{1}H NMR leads to highly informative spectra that contain heavily overlapping resonances. Recently, a call for applying ^{1}H NMR metabonomics to facilitate disease risk assessment and clinical diagnostics has emerged [1-3]. A key issue in bringing metabonomics for clinical use will be to bridge the gap between biochemistry – as revealed by ^{1}H NMR spectroscopy – and the relevant measures of current clinical practice.

Biomedical research relies heavily on the statistical analysis of empirical findings and extrapolation from limited sample sets to the general population. Conventionally, hypothesis testing according to analytically derived formulations is used. Typical modelling assumptions include normally distributed independent variables and linear dependence between explanatory variables and outcomes. For multidimensional data sets, such as ^{1}H NMR spectra, construction of a single null hypothesis is not possible (multiple testing issue) and a different approach is required.

Here the challenge is to uncover clinically relevant ^{1}H NMR signals. We know that one metabolite can manifest several peaks, and that the signal intensities are both biochemically and (patho)physiologically related. Furthermore, the data sets are extensive but redundant: one measurement yields tens of thousands of data points, but the effective dimensionality is much less (yet high compared to the number of samples) due to a smaller number of NMR-visible compounds from which the spectral resonances arise. The benefits of the Bayesian approach to tackle this particular problem include the incorporation of model selection as a part of inference and accurate reliability estimates for non-linear models. Interestingly, by allowing the selection of input variables to be a target of modelling, we can reduce the effect of prior assumptions compared to conventional statistics.

2 ^{1}H NMR Metabonomics and Clinical Data

The ^{1}H NMR experiments of the serum samples were targeted at two different molecular windows – lipoprotein lipids and low-molecular-weight metabolites. These data were recorded at the physiological temperature of 310 K on a Bruker AVANCE spectrometer operating at 500.13 MHz using a double tube system facilitating absolute metabolite quantification. Prior to Fourier transformation, the measured free induction decays were multiplied by an exponential window function with a line-broadening of ≤ 1.0 Hz. The main aliphatic region from 0.4 to 3.3 ppm, including 18,093 data points, was used in the Bayesian spectral analysis.

Two clinical variables, very low density lipoprotein triglycerides (VLDL-TG) and high density lipoprotein cholesterol (HDL-C), were chosen for preliminary analysis due to their use in atherosclerosis risk evaluation. Clinical and ^{1}H NMR data were available on 75 and 67 individuals in the case of VLDL-TG and HDL-C, respectively. The biochemical assays for these lipid variables and the serum ^{1}H NMR spectra are physically independent. Thus, by modelling the quantitative relation between the ^{1}H NMR metabonomics data and the clinical variables, we can make statistical estimates of the two lipid fractions from serum spectra alone.

3 Bayesian Rationale, Computation and Results

The rapid increase in computing resources and new algorithms has made Bayesian inference an alternative to hypothesis testing. We begin with a limited prior knowledge of a phenomenon and after making observations, our knowledge is increased. Recalling that we can only study limited sample sets of the general

population, the uncertainty of our beliefs about the true nature of the phenomenon is described mathematically by probability distributions, thus replacing the concept of statistical significance. The technical challenge is to compute these posterior distributions, since the statistical models cannot be presented by closed analytical formulas in complex problems. On the other hand, once the algorithms are in place and the distributions sampled, the inference is straightforward. Note that no statistical testing is needed, hence the null hypothesis is no longer required and the multiple testing issue is eliminated altogether.

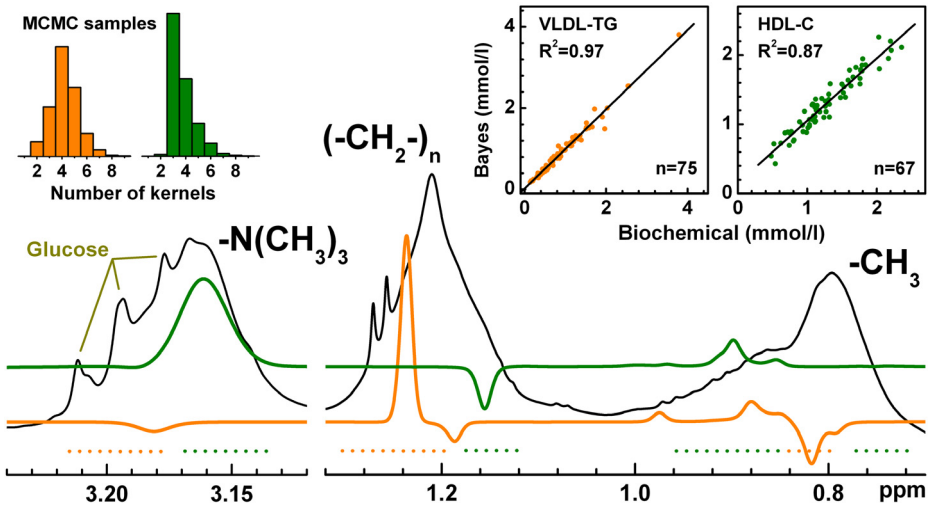

Fig. 1. Illustration of the main kernels for VLDL-TG (*orange lines)*, HDL-C (*olive lines)*, the related regions of the experimental ^{1}H NMR spectra (*black lines)* and the biochemically known spectral regions in which significant contributions originate from VLDL (*orange dashed lines*) and HDL (*olive dashed lines*) particles. The number of kernels (*left insets*) and the predictive performance of the Bayesian model (*right insets*) are also shown. Note an excellent coincidence of the main kernels and the biochemically related spectral regions.

As previously mentioned, the selection of relevant spectral regions is a key issue in metabonomics. One could proceed with an assumption that all the data points are independent. However, with thousands of variables this would be difficult and time-consuming [4]. Here, the spectroscopic fact that adjacent data points are strongly correlated is not ignored but an unknown and non-constant correlation length is allowed. Local correlation is captured by a kernel that represents the weighted sum of intensities in a region: Gaussian kernels truncated at 3σ's were used with their number unknown and inferred from the data. The minimum width of the kernels was constrained to fulfil known molecular characteristics in the NMR spectra. Also, the mean level of each spectrum was used as an additional covariate. Based on the application specific knowledge, it was reasonable to assume that a linear model of the target variables and kernel outputs was appropriate. Student's t-distribution was preferred as a more robust residual model over the Gaussian distribution [4].

The posterior inference was made by Markov chain Monte Carlo (MCMC) [4]. A useful property of our model specification is that marginal likelihoods, obtained by

analytically integrating over the linear model weights, can be used to significantly improve the sampling quality [5]. Kernel locations and widths, and the degree of freedom for the residual model were inferred by slice sampling. The number of kernels was sampled by reversible jump MCMC in which the proposal distributions for new parameters were the corresponding prior distributions. The rest of the model parameters were updated using Gibbs' sampling with conditional distributions [4].

Ten independent chains of 10,000 iterations each were run. From each chain, the first 2,000 iterations were discarded, and from the rest every 20^{th} iteration was saved. Independent chains were used to estimate an approximate convergence and a predictive replicate approach was used to validate the results. Here, the test data set was generated from the predictive distribution of the model and subsequently used to estimate the predictive performance. In a preliminary phase, a 10-fold-cross-validation was used as a more robust strategy to check that the predictive replicate approach produced meaningful results. The predictive R^{2}'s were 0.97 and 0.87 for the VLDL-TG and HDL-C, respectively (see Fig. 1). These values represent excellent quantitative correspondence in this extensively studied complex application [3]. Fig. 1 also shows the marginal posterior distribution for the number of kernels. It is good to note that the prior on linear model weights does influence these distributions. When integrating over the unknown number of kernels, the predictions are not sensitive to this prior selection, but if we were to select any single value for the number of kernels to construct a simpler model, the selection should be based on predictive criteria [5].

As a conclusion, the measures related to and obtained from the Bayesian model indicate excellent mathematical performance with a relatively small number of kernels. In addition, it is rather surprising to note that the most important kernels for the two clinical measures are exactly on the spectral regions known to be the best representatives for these particular lipoprotein lipids [3]. Hence, it is most probable that the Bayesian methodology will have a crucial role in paving the way for metabonomics on the clinical arena.

Acknowledgments. This work has been supported by the Academy of Finland.

References

1. Nicholson, J.K., Wilson, I.D.: Understanding 'global' systems biology: metabonomics and the continuum of metabolism. Nat. Rev. Drug Disc. 2 (2003) 668-676.
2. Clayton, T.A., Lindon, J.C., Cloarec, O., Antti, H., Charuel, C., Hanton, G., Provost, J.P., Le Net, J.L., Baker, D., Walley, R.J., Everett, J.R., Nicholson, J.K.: Pharmaco-metabonomic phenotyping and personalized drug treatment. Nature 440 (2006) 1073-1077.
3. Ala-Korpela, M., Lankinen, N., Salminen, A., Suna, T., Soininen, P., Laatikainen, R., Ingman, P., Jauhiainen, M., Taskinen, M.-R., Héberger, K., Kaski, K.: The inherent accuracy of ^{1}H NMR spectroscopy to quantify plasma lipoproteins is subclass dependent. Atherosclerosis (2006) in press.
4. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.R.: Bayesian Data Analysis, Second edition, 2003, Chapman & Hall.
5. O'Hagan, A., Forster, J.: Kendalls´s Advanced Theory of Statistics, Volume 2B, Bayesian Inference, Second edition, 2004, Arnold.