A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures.

William Astle, Maria De Iorio, Timothy Ebbels, Sylvia Richardson.

Nuclear Magnetic Resonance (NMR) spectroscopy is a technique for obtaining structural and quantitative information about molecules from the resonant radio frequencies of their constituent nuclei in a magnetic field. The chemical environment of a magnetic nucleus perturbs the local magnetic field and this leads to characteristic shifts in and splitting of the absorption frequencies of the nucleus. The aggregated shift and splitting signals from nuclei in multiple chemical groups can be used for molecular identification while the strength of the resonance signals can be used for molecular quantification. We present a novel model for proton NMR spectra, which allows us to identify and quantify the metabolic compounds present in complex mixtures such as biofluids.

Resonance signals combine additively in NMR spectra so our modelling is based on a Bayesian linear regression. We assume the spectral data are generated with Gaussian error from

$$\sum_j w_j S_j(\delta) + \eta(\delta),$$

sampled at regular intervals of the chemical shift δ (a parameter proportional to the frequency of exposure radiation). Here, S_j is a template function determined by the chemistry of the j^{th} metabolite, through the nuclear shifts and multiplet splittings. We aim to deconvolve a spectrum into components corresponding to individual metabolites, which will allow us to make inference about the concentration parameters w_j. We take advantage of substantial prior information about the shape of these templates for selected metabolites. NMR theory implies that each S_j is a mixture of Lorentzian curves, the position, shape and relative heights of which are determined by the number of hydrogen nuclei in the compound and their shift and splitting parameters. The shift and splitting parameters for a particular biofluid can be determined empirically and for many metabolites these data are available from online databases.

To complement our model for the spectral signal generated by compounds with well characterised NMR signatures we model the residual signal, mostly due to unidentified compounds (i.e. those for which no template S_j is available), by a flexible non-parametric component η. We aim to use the parameters w_j and the component η, representing unidentified metabolites, for supervised and unsupervised classification of individuals by their biofluid metabolite profile.

Posterior inference is performed by MCMC methods. We will demonstrate the performance of our model using simulated and real data.