Bayesian analysis of SELDI-TOF data

Kelly Handley*, William J. Browne and Ian L. Dryden

University of Nottingham

1 Introduction

The development of surface-enhanced laser desorption/ionisation time-of-flight (SELDI-TOF) mass spectrometry has enabled the analysis of complex protein samples over a large range of molecular weights. Recent work has found that SELDI-TOF has the potential to identify some biomarkers for disease which could enable earlier diagnoses for patients.

In this study our aim is to use mass spectra to differentiate between drug-treated breast cancer cell-lines and non-treated controls as in Dryden *et al.* (2005), and Ball *et al.* (2002) who use neural networks. The mass spectra for a sample in this study consists of around 14,000 datapoints. Each datapoint comprises a relative intensity of proteins at a particular mass over charge (*m/z* value). The *m/z* value is calculated by dividing the protein mass by the number of charges induced by ionisation. We consider *m/z* values between 2kDa and 30kDa.

We propose a method for the discovery of potential biomarkers via the use of a Bayesian hierarchical model for the mass spectra. Possible biomarkers are those *m/z* values which exhibit significant differences between the groups.

2 The model

The dataset consists of 144 mass spectra separated into 6 groups. The groups are ADC (MCF7/ ADR control), ADT (MCF7/ADR Taxol treated), MCC (MCF7 control), MCT (MCF7 Taxol treated), TDC (T47D control) and TDT (T47D Taxol treated). For each group 24 observations are available. These consist of readings from two independent experiments and, within each experiment, three separate samples were taken at 24 hour intervals for four days. For each spectrum, the m/z values below 2,000 Daltons have been removed as they are mostly considered to be background noise.

From initial investigations it was noted that the cell lines consist of a sequence of peaks of varying heights. A possible modelling approach is therefore to fit a series of Gaussian peaks to the data with locations, heights and variances to be estimated. This approach can be implemented using Markov Chain Monte Carlo (MCMC) methods (e.g. Gilks *et al.*, 1996) to construct samples from the joint posterior distribution of the unknown parameters, namely the locations, heights and variances of the peaks.

The model used for each datapoint y_{is} on the spectrum s is distributed as $y_{is} \sim N(\theta_{is}, \tau^{-1})$ where

$$\theta_{is} = \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\xi\mu_j^2)^{-1}(x_i - \mu_j)^2\right) \tag{2.1}$$

where the index $i = 1, \ldots, p$ represents the position on the cell line, x_i is the i^{th} *m/z* value, $s = 1, \ldots, n$ is the spectra number and $j = 1, \ldots, k$ is the peak number.

The parameters μ_j and h_{js} represent respectively the mean and height of the peaks in the model, and ξ is a constant of proportionality that models the fact that the variance of the peaks increases with the mean. The prior distributions are taken to be:

$\mu_j \sim N(0, V)$, $h_{js} \sim Uniform(c, d)$, $\tau \sim Gamma(a, b)$ and $\xi \sim Uniform(e, f)$.

The model we are fitting has common means and a common constant of proportionality for each of the spectra. The heights of each peak are allowed to differ, both across groups and across days within the same group. The posterior distribution under consideration is:

$$\begin{aligned} posterior &\propto likelihood \times prior \\ &\propto \prod_{i=1}^{p}\prod_{s=1}^{n} \tau^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}\left[y_{is} - \sum_{j=1}^{k} h_{js}(\xi\mu_j^2)^{-\frac{1}{2}} \exp\left(-(\xi\mu_j^2)^{-1}(x_i - \mu_j)^2\right)\right]^2\right) \\ &\times \quad \frac{b^a \tau^{a-1} e^{-b\tau}}{\Gamma(a)} \\ &\times \quad \prod_{j=1}^{k} V^{-1/2} e^{\frac{-\mu_j^2}{2V}} \end{aligned}$$

The μ_j, h_{js} and ξ parameters are updated using Metropolis-Hastings steps and the τ parameter is updated by a Gibbs step.

3 Results

For simplicity, a small section of the data (1,001 datapoints) from m/z values 6,794.3 to 8,397.2 Daltons was initially studied. Ten peaks were fitted to each scan in each of the six groups on all four days. Using the above model and updating methods and running for 5,000 iterations resulted in, for example, the fitted models for the ADC group shown in figure 1. Examples of the trace plots obtained are shown in figure 2 - one for each of the four types of parameter.

The lighter curves are the original cell lines - there are six such curves in each group resulting from three samples taken from each of two experiments. The darker lines show the maximum a posteriori estimates of θ_{is} in equation (2.1).

In each of the drug groups AD, MC, and TD there are differences in the heights of particular peaks between the control and the Taxol treated cell lines and differences between the controls.

The model seems to fit the spectra quite well and most of the important peaks are picked out. However, some of the fitted heights are not quite correct for the larger peaks. This is probably because of the relatively small number of iterations carried out (5,000) and there would be an improvement in the fit with a larger run. In two cases the same peak is picked out by two separate parameters - *m/z* values around 6,920 and 8,110.

The mean parameters not referring to obvious peaks in a particular group generally seem to have small heights as was expected to occur.

As indicated earlier there are occasions when more than one location parameter refers to the same fitted peak. This can be rectified by using a Strauss process on the location parameters. This penalises proposals where the new location would be within a certain tolerance of any of the other current locations. Also reversible jump MCMC could be implemented into the model. Currently the number of fitted peaks has to be specified beforehand and reversible jump MCMC will remove this stipulation.

A good choice of parameter starting values could reduce the number of iterations needed in the MCMC algorithm and so a deterministic peak finding method has been developed. We have *n* spectra with *p* observations in each spectra. Let Y_{ij} be the i^{th} observation in the j^{th} spectra. To find the location of the first peak find an i_1 to maximise $\sum_{j=1}^{k} Y_{ij}$ over i and let $\mu_1 = i_1$. Place

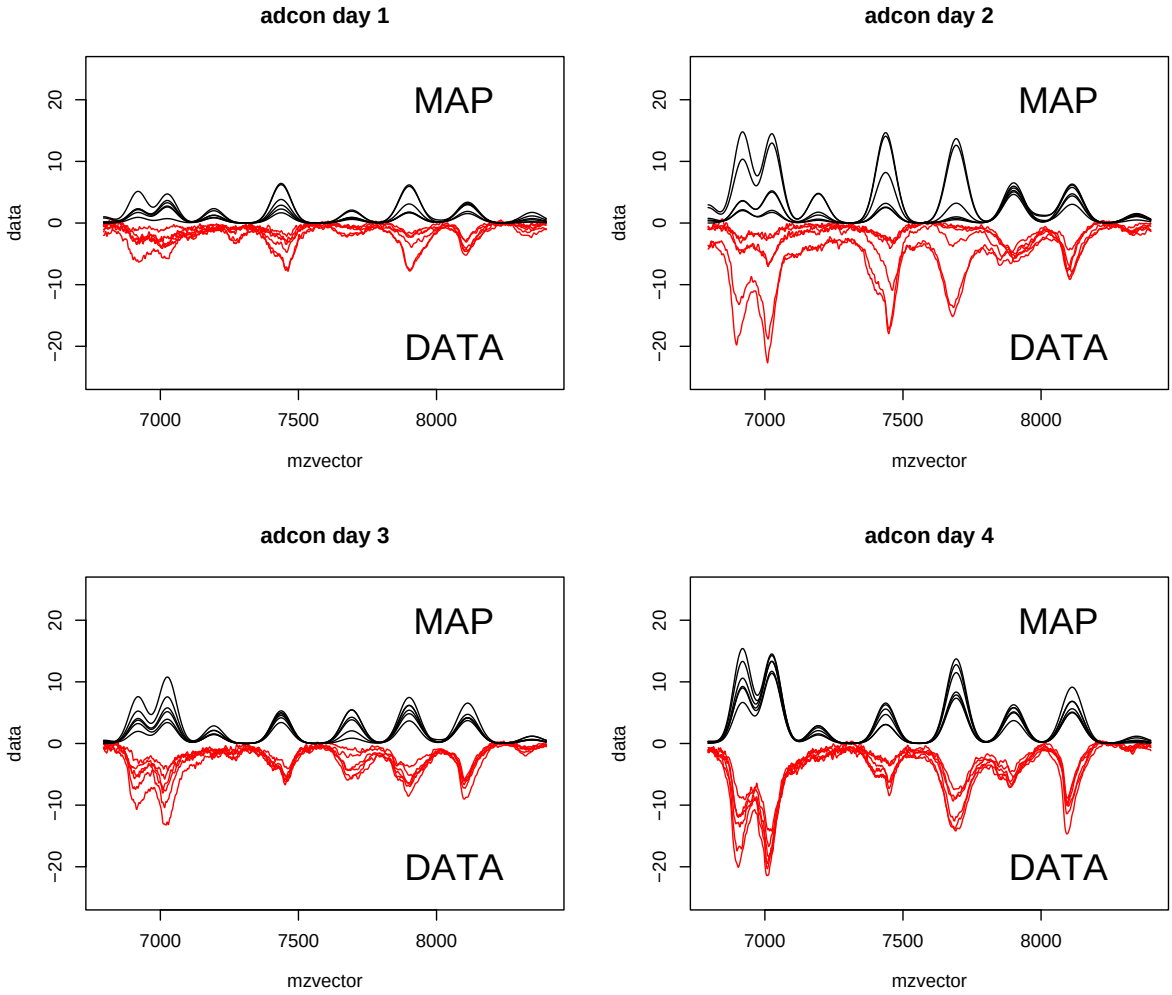

Figure 1: The MAP after 5,000 iterations and original data for all 24 spectra in the ADC group on each day separately using Metropolis-Hastings and Gibbs sampling to update the parameters.

a normal distribution of the form $c_{1j}N(\mu_1, \sigma_1^2) = c_{1j}f_1$ at this location with the value of σ_1^2 to be determined. The scaling parameter c_{1j} then needs to be calculated for each scan *j*. In order to find the location of the next peak, first form $X_{ij} = Y_{ij} - c_{1j}f_1(i)$. This effectively 'subtracts' the peak just accounted for in the data. Repeat the procedure on the X_{ij} each time subtracting contributions from peaks already fitted until the desired number of peaks is reached.

When these changes have been implemented, further inference will be carried out to determine which of the peaks are significantly different between control and drug scans within the same drug group and also between the control groups. This process will potentially identify a group of *m/z* values which can be used as biomarkers in the identification of disease.

References

Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McArdle, S., Ellis, I.O., Creaser, C. and Rees, RC (2002). An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers, *Bioinformatics*, **18**, 395-404.

Dryden, I.L., Mian, S., Browne, W.J., Handley, K., di Nisio, R. and Rees, R. (2005). Statistical

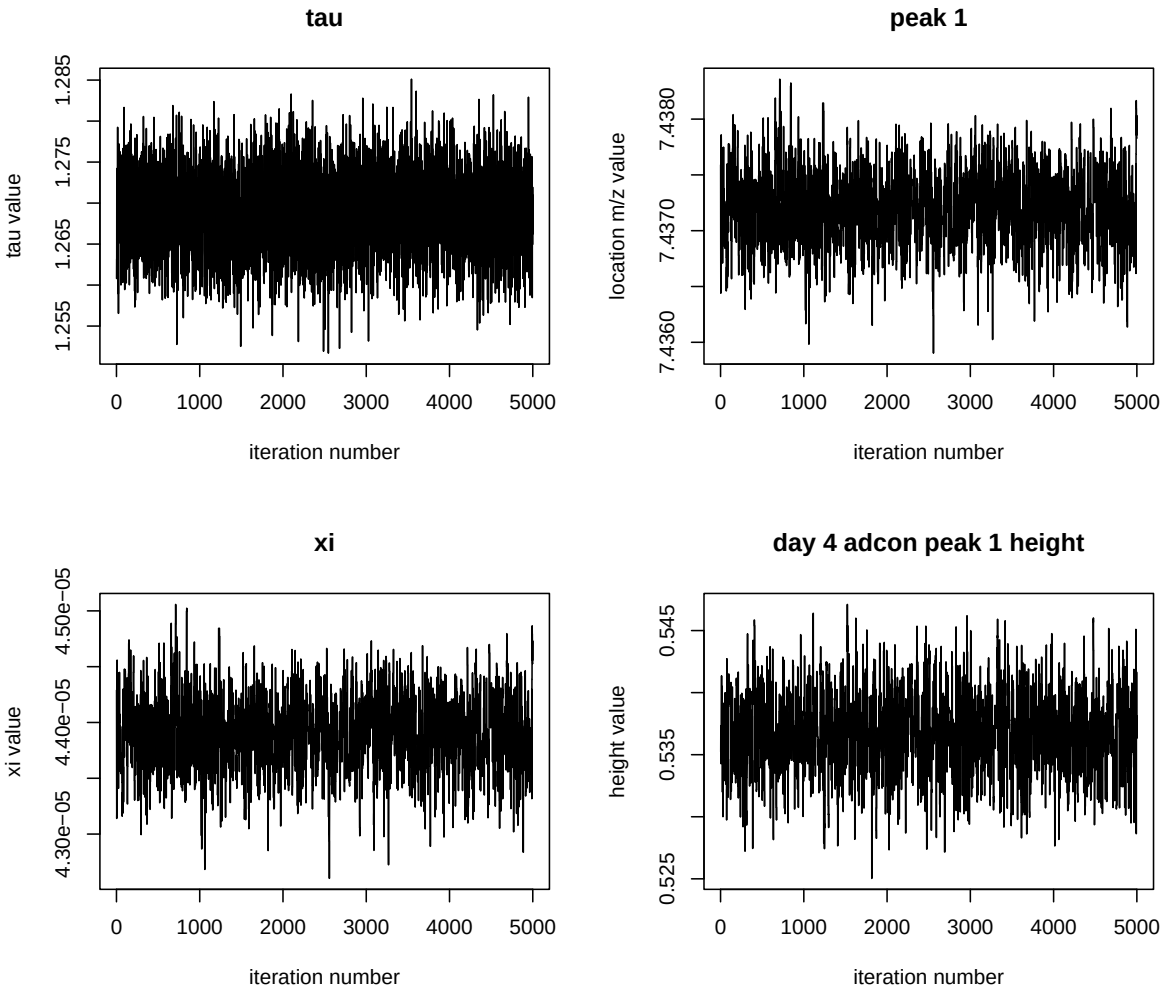

Figure 2: The trace plots obtained over 5,000 iterations for ξ and τ, and one example each of the traces for μ_j and h_{js}.

analysis of surface-enhanced laser desorption/ionization (SELDI) protein chip data from breast cancer cell lines exposed to chemotherapeutic agents, *submitted*, available as research report 05-02, Division of Statistics, School of Mathematical Sciences, University of Nottingham.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo In Practice*. London, Chapman & Hall.