Bayesian functional models with wavelets for disease identification in Mass Spectroscopy Proteomics.

Philip Brown*[1], Jeffrey Morris[2], Kevin Coombes[2], Keith Baggerly[2]

[1] University of Kent
[2] U Texas MD Anderson Cancer Center

1 Introduction

This work applies Bayesian modelling to wavelet transformed Mass Spectroscopy data obtained from time of flight instruments. It is an extended application of Morris *et al.* (2003), see also Morris and Carroll (2004), Brown *et al.* (2001). The framework of modelling is quite general and allows fitting fixed and random effects to data that comprises functions of time of flight, and in particular mass to charge m/z data. In this talk we will provide

- the modelling framework and introduce wavelets
- describe the mass spectroscopy instruments used (MALDI and SELDI)
- illustrate with applications to a pancreatic cancer study, Koomen *et al* (2005), and an organ- cell line experiment
- give the form of the fitted models
- describe the identification of peaks
- apply methods for discrimination of diseased and normal.

The methodology is very flexible, allowing functions of arbitrary form and smoothness and the full range of fixed effects and between curve covariance structures for a mixed models framework. Since it is based on wavelets it is particularly suited to the spikey data which characterises mass spectroscopy, although the approach can be generalised to any other orthogonal basis functions.

We use a linear multivariate mixed effects model:

$$Y = XB + ZU + E. \tag{1}$$

Here Y, E are $N \times q$ random matrices, rows of Y being mass spectra at q mass to charge (m/z) ratios; X and Z are $N \times p$ and $N \times m$ model matrices for the fixed effects and random effects, respectively. Thus rows of $B,\ (p \times q)$ are the fixed effects curves. The random effects U and the errors E generate the covariance structure of a multivariate Normal model.

The Bayesian fitting process after wavelet transform embodies non-linear shrinkage as a consequence of an empirical Bayes fitted 'slab and spike' prior.

MCMC samples allow rich inference possibilities, including both pointwise and joint Bayesian inference and prediction. In identifying differentially expressed peaks we may use a cutpoint that controls the expected Bayesian False Discovery Rate (Newton *et al*, 2004).

References

Brown, P. J., Fearn, T. and Vannucci, M. (2001). Bayesian Wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist Assoc.,* **96**, 398-40.

Koomen, J. M., Shih, L.N., Coombes, K. R. Li, D. Xiao, L-C., Fidler, I. J., Abbruzzese, J. L. and Kobayashi, R. (2005) Plasma Protein Profiling for diagnosis of Pancreatic Cancer reveals the presence of host response proteins. *Clinical Cancer Research,* **11**, 1110-1118.

Morris, J. S. Vannucci, M., Brown, P. J. and Carroll, R. J. (2003). Wavelet-based nonparametric modelling of hierarchical functions in colon carcinogenesis. Special Chosen Applications of Case Studies Read Paper with Discussion, receiving the Mitchell Prize, Joint Statistical Meetings, San Francisco, *J Amer. Statist. Ass.,* **98**, 573-597.

Morris J. S. and Carroll, R. J. (2004) Wavelet-Based functional mixed models. *Technical Report Biostatistics, MD Anderson Cancer Center*.

Newton, M.A., Noueiry,A., Sarkar, D. and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155-176.