Working Group on Statistical Learning Seminar

Title: Analyzing Longitudinal Metabolomic Data

Speaker: Gift Nyamundanda

Date: Fri 6th May 2011 at 1:00PM

Location: Statistics Seminar Room- L550 Library building

Abstract: Metabolomics is the term used to describe the study of small molecules or metabolites present in biological samples. Data sets from metabolomic studies are typically high-dimensional and complex. In a longitudinal metabolomic study, multiple metabolites are measured from subjects at multiple time points. Typically the number of samples n in such studies is much less than the number of variables p, $n \ll p$.

Traditional principal component analysis (PCA) is currently the most widely used statistical technique for analyzing metabolomic data. However, the application of PCA to longitudinal metabolomic studies is limited by the fact that it assumes independence of the repeated measurements and it is not based on a statistical model. Probabilistic principal component analysis (PPCA) detailed in the article by Nyamundanda $et\ al$ 2010, addresses some of the limitations of PCA. Here, we propose an extension of PPCA called dynamic PPCA which allows us to use PPCA to model metabolomic data, while taking into account the correlation due to repeated measurements. Dynamic PPCA reduces the dimension of the data by defining the p-dimensional observation $\underline{\mathbf{x}}_{im}$, i.e. the metabolomic spectrum for sample i at time point m, as a linear

transformation of the lower q-dimensional latent variable $\underline{\mathbf{u}}_{im}$:

$$\underline{\mathbf{x}}_{im} = \mathbf{W}_m \underline{\mathbf{u}}_{im} + \underline{\mu}_m + \underline{\epsilon}_{im}$$

where $\mathbf{W}_m$ and $\underline{\mu}_m$ are a $p \times q$ loadings matrix and the mean of the data at time point m respectively and $\underline{\epsilon}_{im}$ is a multivariate Gaussian noise process for sample i at time point m, i.e. $p(\underline{\epsilon}_{im}) = MVN_p(\underline{\mathbf{0}}, \sigma_m^2 \mathbf{I})$ and $\mathbf{I}$ denotes the identity matrix. The dynamic PPCA model corrects for the correlation in repeated measurements by assuming that $\log(\sigma_m^2)$ has a stationary autoregressive model of order 1, centered around mean ν with persistence parameter ϕ:

$$\log(\sigma_m^2) = \nu + \phi(\log(\sigma_{m-1}^2) - \nu) + r_{im}.$$

The innovations r_{im} are assumed to be independent $r_{im} \sim N(0, \upsilon^2)$.

This model allows us to observe the change in position of subjects in the latent principal subspace and to identify the spectral regions responsible for the structure in the data at each time point. The usefulness and applicability of dynamic PPCA is demonstrated on a longitudinal metabolomic study of urine samples of animals taken over 15 days.