
# Working Group on Statistical Learning Seminar

**Title:** Application of Compositional Models for Glycan HPLC Data

**Speaker:** Marie Galligan

**Date:** Fri 11th March 2011 at 1:00PM

**Location:** Statistics Seminar Room- L550 Library building

**Abstract:** Compositional data consist of positive real variables which sum to a constant. The sample space for a p dimensional vector of such data is a p-1 simplex. This constraint requires careful consideration in the statistical analysis of compositional data. We are currently modelling the data in a simplex sample space, using a generalization of the Dirichlet distribution, known as the Nested Dirichlet distribution. However, finding the model structure from this data proves to be computationally intense, which has encouraged us to seek a method for reducing the dimension of this data before model fitting. As we wish dierentiate between disease groups in our data, this approach could also be useful for directly identifying disease biomarkers. We are using a simple variable selection algorithm, similar to that used in Murphy et al. (2010). We are specifying the beta distribution as the marginal densitiy for each compositional variable. We base our decision to add or remove a variable to/from the model by looking at the difference in BICs for that variable when modelled with and without group information.