Working Group on Statistical Learning Seminar

Title: Latent Variable Models for Network and Nodal Attribute Data

Speaker: Isabella Gollini (UCD)

Date: Fri 8th April 2011 at 1:00PM

Location: Statistics Seminar Room- L550 Library building

Abstract: Recently there has been a growing interest in the modelling of network data. Network data consists of a set of nodes and a list of edge connections between the nodes. In many cases, nodal attributes are also available. It can therefore be useful to build a model that jointly summarises both the network and the nodal attributes.

Latent space models for network data were introduced by Hoff et al. (2002) under the basic assumption that each node has an unknown position in a D-dimensional Euclidean latent space: generally the smaller is the distance between two nodes in the latent space, the greater the probability of them being connected. In the latent space model, the posterior distribution can not be calculated analytically, so we use a variational Bayes approach to estimate the model: we introduce a variational posterior that depends on variational parameters, than we minimize the Kullback-Leibler divergence to the true posterior. In this approximation, the expected log-likelihood is approximated by using Jensen's inequality.

Factor analysis is a common latent variable model for continuous variables in which it is assumed that a D-dimensional continuous latent variable underlies the behaviour

of the response variables within each observation. The factor analysis model can be estimated by using an EM algorithm.

We combine these models in a latent variable model that merges the information given by the network and the nodal attributes. The probability of a node being connected with other nodes and the behaviour of nodal attributes are explained by the same latent variable.

To estimate the joint model we propose an EM algorithm: we use the parameters obtained fitting the two models independently to find the joint posterior distribution. Then these results are used to update the estimate in the factor analysis and latent space models. We continue this algorithm until convergence is attained.

This model is demonstrated on the analysis of a Saccharomyces cerevisiae (yeast) genes dataset, where the estimated latent positions give a clearer visualization of the data than either the latent space model or the factor analysis model alone.