McMASTER UNIVERSITY

GRADUATE PROGRAM IN STATISTICS

STATISTICS SEMINAR

Speaker: Dr. Paul McNicholas
Department of Mathematics & Statistics
University of Guelph

Title: Model-Based Clustering: An Overview
Day: Tuesday, October 23, 2007

Time: 3:30 - 4:30 PM

Place: HH/217 - Deloitte Colloquium Room
(refreshments in HH/216 at 3:00 PM)

SUMMARY

In recent years model-based clustering has appeared in the statistics literature with increased frequency. Typically data are clustered using some assumed mixture modeling structure and the parameters associated with these models are usually estimated using some variant of the EM algorithm.

Model-based clustering techniques based on the Gaussian mixture model has received particular attention. An eigenvalue decomposition of the group covariance matrices for the Gaussian mixture model can be exploited to give a wide range of covariance structures. The resulting family of models (MCLUST) is reviewed, along with the related variable selection technique.

A family of parsimonious Gaussian mixture models (PGMMs) that use a latent Gaussian model which is closely related to the factor analysis model is also introduced. These models provide a unified modeling framework which includes the mixture of

probabilistic principal component analyzers and mixture of factor of analyzers models as special cases.

The MCLUST, variable selection and PGMM techniques are then applied to data on chemical and physical properties of Italian wines and data on biological measurements on crabs, where the models give good clustering results. Clustering performance across the techniques is compared using the Rand and adjusted Rand indices.

Finally, work-in-progress on the creation of a family of mixture models for longitudinal data is outlined and demonstrated on data on the weights of rats on three different diets. These models give good clustering performance and have excellent potential for further development.

REFERENCES

- Fraley, C. & Raftery, A. E. (1998) Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.
- McLachlan, G. J. & Peel, D. (2000) *Finite Mixture Models*, John Wiley & Sons, New York.
- McNicholas, P. D. & Murphy, T. B. (2007) Parsimonious Gaussian mixture models, *Statistics and Computing* (Conditionally accepted).
  Available at `www.uoguelph.ca/ pmcnicho/mcnicholas_murphy07.pdf`.
- Raftery, A. E. & Dean, N. (2006), Variable selection for model-based clustering, *Journal of the American Statistical Association*, **101**, 168–178.

ABOUT THE SPEAKER

Paul McNicholas received his education at Trinity College Dublin, where he received a B.A. in Mathematics (1999), an M.Sc. in High Performance Computing (2007) and a PhD in Statistics (2007). Since July 2007, he has been an assistant professor of statistics at the University of Guelph. Dr. McNicholas' main research focus to date has been on families of mixture models and model-based clustering techniques. He has also done some work on data-mining techniques, in particular association rules. Dr. McNicholas is also interested in computational statistics and high performance computing.

MORE SEMINAR INFORMATION

A list of recent and upcoming seminars is available at
`http://www.math.mcmaster.ca/canty/seminars`

For further information please contact Angelo Canty at 905-525-9140 ext. 27079, email: `cantya@mcmaster.ca`.