ON CLUSTER ANALYSIS

~

A BAYESIAN AND MODEL-BASED APPROACH

Jessica Franzén

Licentiate Dissertation
Department of Statistics
Stockholm University
2006

Licentiate Dissertation
Department of Statistics
Stockholm University
S-106 91 Stockholm

ISBN 91-7155-261-8

Abstract

Cluster analysis is the automated search for homogenous and cohesive groups in a given data set. Traditional cluster analysis is based on deterministic methods which use measures between objects and objects and centroids to create well separated groups. Despite considerable research, there is little guidance how to handle practical questions such as how many clusters there are and how to handle outliers objects. A model-based approach to cluster analysis is presented. As opposed to the mechanical classification used in deterministic clustering, we regard observations as outcomes of different distributions. A finite mixture model is used, where each probability distribution corresponds to a cluster. This approach opens up for new possibilities. The model is capable to handle groups of different sizes, shapes, and directions by allowing for different distributions and parametrization among clusters. In reality, clusters do seldom appear as well separated. The method handles overlapping groups, by taking into account cluster membership probabilities in these areas. In many data sets there are objects not suitable for classification. A special approach of this thesis is to create a deviant cluster of larger variance, consisting of these outlier objects. Bayesian inference via Gibbs sampling is used to estimate distribution parameters and proportions between clusters. The method is tested on simulated and real data sets and shows promising results. Model selection by an approximation of Bayes factors is applied, with the purpose of selecting the number of clusters and to decide if a deviant group is to prefer in the model.

Keywords: Clustering, Classification, Mixture distribution, MCMC, Gibbs sampler, BIC, Deviant group

The support from the Bank of Sweden Tercentenary Foundation (Grant no 2000-5063) is gratefully acknowledged.

Acknowledgements

There are quite a few people I want to thank who have contributed to my thesis and been a support in different ways during the time writing it.
First of all, I am most grateful to my supervisor Professor Daniel Thorburn. Your ideas, guidance and what it seems, unlimited knowledge have been invaluable to me. After numerous hours of discussions this thesis finally took form.
I am very grateful to Professor Lars R. Bergman at the Department of Psychology, not only for providing me with the data material but also with the time you spent discussing ideas and coming with valuable inputs.
University Lecturer Karin Dahmström has been very helpful proofreading my thesis. Thank you for your interest.
Thank you to all my colleagues, former and present, at the department. It has been encouraging and inspiring to have you all around. A special thank to Håkan, for providing me with computers with extended memory along the way. Daniel, Mattias, and Ellinor in the Scooby-Doo Gang for cooperation and sharing experiences during the journey we all started at the same time. A special thanks to my dear roommate and friend Ellinor for endless discussions on high and low levels, relevant and irrelevant for this thesis.
Åsa - Thank you for sharing ups and downs, discussing statistics, taking walks, and drinking wine with me. You are a true friend.
My whole family has been a big support. Thank you for always being there, loving me, and coming with cheerful comments along the way, even though you don't always understand what I'm doing. Sometimes I don't get it myself.
Knut - First of all, thank you for valuable and inspiring discussions on such high levels and also for your proofreading. Most of all, thank you for your never ending love and support. I love you.

Stockholm, May 2006

Jessica Franzén

1 Introduction

Cluster analysis is a grouping of objects on the basis of (dis)similarities between them. Most clustering, done in practise, is based on traditional *deterministic* methods. These methods are developed for situations with homogenous and well separated groups. One widely used deterministic method involves hierarchical clustering. It starts with as many clusters as there are observations, and the number of clusters is decreased one by one, at each step. Two groups are merged at each stage, according to some optimization criteria Commonly used criteria for merging are cluster measures, such as smallest dissimilarity (single-linkage), average dissimilarity (average linkage), or maximum dissimilarity (complete linkage); see Oh and Raftery (2003). Another commonly used deterministic method is nonhierarchical clustering, which is based on iterative relocation. Objects are relocated between a predetermined number of groups until there is no further improvement according to some criteria used. All deterministic methods have in common that they use measures between objects, and objects and centroids, to create well separated and homogenous clusters. There is a vast literature on traditional deterministic clustering methods, see for instance Sharma (1996), Jain and Dubes (1988), and Everitt et al. (2001).

Deterministic clustering is well suited for cohesive and well separated groups, but it is not constructed for situations with clusters of different sizes, shapes, direction, and overlapping clusters. There is little guidance of how to handle practical questions such as how many clusters data should be divided into, and how to handle outlier objects. Moreover, these methods are not based on standard principles of statistical inference. They do not take into consideration measurement error in the dissimilarities, and they do not provide an assessment of clustering uncertainties (Oh and Raftery (2003)).

Model-based cluster analysis is another cast of mind developed in recent years. The idea is to base cluster analysis on a probability model. The population of interest consists of J different subpopulations, each with its own distribution. Data $(\mathbf{y}_1, ..., \mathbf{y}_n)$ are viewed as coming from a mixture model according to (1), where each distribution f_j represents a cluster.

$$f(\mathbf{y}_i \,|\boldsymbol{\theta}) = \sum_{j=1}^{J} p_j f_j(\mathbf{y}_i \,|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \qquad i = 1, ..., n \tag{1}$$

The proportions $0 < p_j < 1$ satisfy $\sum_{j=1}^{J} p_j = 1$.

The development of cluster analysis in this direction opens for understanding the true process and origin of clusters, and for suggestions of new and better methods. One is able to handle groups of different sizes, shapes, and directions. Various geometric properties are obtained through different parametrization of the distributions, or even completely different distributions among clusters. Measurement errors are an inherent part of the model, and outliers can be modeled by adding a cluster with larger variance. Finite mixture models in the context of clustering have been studied in Wolfe (1970), Edwards and Cavalli-Sforza (1965), Day (1969), Scott and Symons (1971), and Binder (1978). In recent years it has been recognized that model-based clustering can answer practical questions such as how many clusters data should be divided into, which distributions and parametrization to use, and how to handle outlier objects. McLachlan and Basford (1988), Banfield and Raftery (1993), Cheeseman and Stutz (1995), and Fraley and Raftery (1998) all have made contributions in the field.

Many recent publications have shown promise in a number of practical applications. Identification of textile flaws from images in Campbell et al. (1997), microarray images in DNA in Li et al. (2005) and Yeung et al. (2001), setting in social networks in Schweinberger and Snijders (2003), classification of astronomical data in Bensmail et al. (1997), separating species in Raftery and Dean (2004), color image quantization, or clustering of the color space in Murtagh et al. (2001), and curvilinear clustering for detecting minefield and seismic fault in Dasgupta and Raftery (1998) and Stanford and Raftery (2000).

Bayesian inference is used in this thesis. We are interested in estimating the parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ for each distribution, and the proportions between clusters in the mixture model (1). According to Bayesian methodology, our prior assumptions together with a likelihood function from the data, generate the posterior distribution. Its exact evaluation requires complex integration. One problem with, and criticism of (non-philosophical), Bayesian mixture estimation is its computational difficulties. Thanks to the availability and development of high-speed computing in recent years, the use of Bayesian inference has increased. Markov Chain Monte Carlo (MCMC) methods was introduced in Tanner and Wong (1987) and Gelfand and Smith (1990) as powerful alternatives to numerical integration (Robert (1994)). MCMC methods evaluate the posterior by drawing samples from a Markov Chain, with the true posterior as equilibrium. After a burn-in period, the draws can be treated as coming from the target distribution. It is suitable in situations where the joint distribution of the parameters of interest, say $p(\alpha, \beta, \delta)$, is difficult to calculate, but the conditional distributions $p(\alpha \,|\beta, \delta\,)$, $p(\beta \,|\alpha, \delta\,)$, and $p(\delta \,|\alpha, \beta\,)$ are possible to simulate from. Gibbs sampler is a particular MCMC algorithm working with conditional states. The Gibbs sampler was first introduced in Geman and Geman (1984) and Tanner and Wong (1987). Each iteration of the Gibbs sampler cycles through the conditional distributions of all the parameters. In each iteration step, new parameters are generated, and the conditional distributions are updated for the next iteration. This iterative procedure makes the process approach the equilibrium $p(\alpha, \beta, \delta)$.

The model-based approach brings advantages in the sense of flexibility in size and structure between clusters, and the ability to handle overlapping groups. These features are used for the special approach of this thesis - a deviant cluster among a more or less homogenous cluster structure. In many real data sets there are objects not suitable for classification. These objects are characterized by their discrepancy from all other objects in the data set. We collect these deviant observations into one cluster with its own distribution of larger variance than the other clusters. The deviant cluster can be spread over part of, or the whole sample space.

The papers in this thesis give a detailed explanation of the model-based clustering approach and its advantages. The mixture model is presented, and an overview of Bayesian inference is given. Prior and posterior distributions are reviewed. The Gibbs sampler simulation method is described in detail. An explanation of Bayes factors, as a model comparison tool, is introduced. We are able to compare models of different number of clusters by an approximation of Bayes factors. The existence of a deviant cluster can also be tested.

The first paper in this thesis is of a more technical art. It gives a detailed explanation of the method, the convergence properties, and the statistical terms used. The method is tested on two simulated data sets with thriving outcome. A Gaussian mixture model is used to describe data. One deviant cluster of smaller size and larger variance is successfully distinguished. The second paper is also intended for readers in the behavioral science field. Complicated derivations and formulas are left out. The method is applied to data on the school performance of 935 children. It was collected by the Individual Development and Adaption (IDA) program at the Department of Psychology, Stockholm University. A longitudinal data base has been created with the purpose of studying individual development process. A selection of seven variables is used in the attempt to find a cluster structure among a group of twelve year old students. We compare models with and without a deviant cluster, and with different number of groups. The method manages to separate data into logical clusters of different sizes, shapes, and directions and moreover, identify outlier objects by placing them in a separate cluster. The best model consists of five clusters plus one cluster with “deviant“ students. The Bayes factor between this and the next best model (seven clusters plus one deviant) is 112, which can be interpreted as very strong evidence for our solution. The results from our solution are compared with those from clustering by Ward’s method, giving a promising outcome for our model-based method.

References

[1] Banfield, J. D. and Raftery, A. E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics*, 49, 3, 803-821.

[2] Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis". *Statistics and Computing*, 7, 1-10.

[3] Binder, D. A. (1978). "Bayesian Cluster Analysis", *Biometrika*, 65, 31-38.

[4] Campbell, J. G., Fraley, C., Stanford, D., Murtagh, F. and Raftery, A. E. (1997). "Model-Based Methods for Textile Fault Detection", *International Journal of Imaging Science and Technology*, 10, 339-346.

[5] Cheeseman, P. and Stutz, J. (1995). "Bayesian Classification (AutoClass): Theory and Results", in *Advances in Knowledge Discovery and Data Mining,* AAAI Press, 153-180.

[6] Day, N. E. (1969). "Estimating the Components of a Mixture of Normal Distributions", *Biometrika*, 56, 463-474.

[7] Dasgupta, A. and Raftery, A. E. (1998). "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering". *Journal of the American Statistical Association*, 93, 441, 294-302.

[8] Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). "A Method for Cluster Analysis", *Biometrics*, 21 362-375.

[9] Everitt, B. S., Landau, S and Leese, M. (2001). *Cluster Analysis*. London: Oxford University Press Inc..

[10] Fraley, C. and Raftery, A. E. (1998). "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis" *The Computer Journal*, 41, 578-588.

[11] Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association*. 85, 410, 398-409.

[12] Geman, S., Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

[13] Jain, A. K. and Dubes R. C. (1988), *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice Hall.

[14] Li, Q., Fraley, C., Bumgarner, R. E., Yeung, K. Y. and raftery, A. E. (2005). "Donuts, Scratches and Blanks: Robust Model-Based Segmentation of Microarray Images", *Technical Report no. 473*, Department of Statistics, University of Washington.

[15] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

[16] Murtagh, F., Raftery, A. E. and Starck, J.-L. (2001). "Bayesian Inference for Color Image Quantization via Model-Based Clustering Trees", *Technical Report no. 402*, Department of Statistics, University of Washington.

[17] Oh, M.-S. and Raftery, A. E. (2003). "Model-Based Clustering with Dissimilarities: A Bayesian Approach", *Technical Report no. 441*, Department of Statistics, University of Washington.

[18] Raftery, A. E. and Dean, D. (2004). "Variable Selection for Model-Based Clustering". *Technical Report no. 452*, Department of Statistics, University of Washington.

[19] Robert, C. P. (1994). "Discussion: Markov Chains for Exploring Posterior Distributions", *The Annals of Statistics*, 22, 4, 1742-1747.

[20] Schweinberger, M. and Snijders, T. A. (2003). "Settings in Social Networks: A Measurement Model", *Sociological Methodology*, 33, 307-341.

[21] Scott, A. J. and Symons, M. J. (1971). "Clustering Methods Based on Likelihood Ratio Criteria", *Biometrics*, 27, 387-397.

[22] Stanford, D. C. and Raftery, A. E. (2000). "Principal Curve Clustering with Noise", *IEEE Transaction on Pattern Analysis and Machine Analysis*, 22, 601-609.

[23] Sharma, S. (1996). *Applied Multivariate Techniques. New York:* John Wiley and Sons, Inc..

[24] Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation". *Journal of the American Statistical Association,* 82, 398, 528-550.

[25] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). "Model-Based Clustering and data transformations for gene expression data", *Bioinformatics*, 17, 102001, 977-987.

[26] Wolfe, J. H. (1970). "Pattern Clustering by Multivariate Mixture Analysis", *Multivariate Behavioral Research*, 5, 329-350.