이학 석사 학위 논문

Finite mixture models and model-based clustering

유한 혼합 모형과 모델 기반 군집화

2017년 2월

서울대학교 대학원

통계학과

김경민

Contents

1. Introduction----------3

2. Inference in finite mixture models

2.1 Estimation in finite mixture models ----------4

2.2 Challenges in implementation

2.2.1 Unbounded likelihood functions----------5

2.2.2 Initialization of the EM Algorithm----------5

2.3 Model selection

2.3.1 Choosing the optimal number of components----------6

2.3.2 Variable selection----------7

3. Some recent applications

3.1 Magnitude magnetic resonance imaging data----------8

3.2 Finite mixtures models in surveys----------8

4. Some additional topics and challenges

4.1. Hierarchical model-based clustering and cluster merging----------10

4.2 Non-parametric approaches to mixture modeling and model-based clustering----------10

4.3 Semi-supervised clustering----------11

4.4 Constrained clustering----------11

4.5 Diagnostics ----------12

4.6 Robust and skewed mixture models----------12

4.7 Dependent data----------13

5. Conclusion----------14

References----------15

Abstract

Kyoung-Min Kim
The Department of Statistics
The Graduate School
Seoul National University

Finite mixture models aims at identifying clusters of individuals who show similar patterns. The method is having been used in a variety of fields, especially in medicine to explain the idea of heterogeneity of treatment effects on population. The number of mixture components is typically not known and has to be chosen.

To solve this problem, EM algorithm-based approaches is considered. We will review details of mixture models and model-based clustering. Furthermore, we will provide an overview of several challenges that have been only partially resolved.

Note: Writing this paper, I mainly refer to "Finite mixture models and model-based clustering"(Melnykov, 2010)

Keyword : EM algorithm, model selection, variable selection, diagnostics, two-dimensional gel electrophoresis data, magnitude magnetic resonance images.

Student Number : 2014-20298

1. Introduction

Let $X_1, X_2, ..., X_n$ be independent, identically distributed p-dimensional observations from a distribution with probability density function

f(X;Θ)= $\sum_{k=1}^{K} \pi_k f_k(X, \theta_k)$

,where π_k represents the probability that i-th observation, X_i belongs to the k-th subpopulation with density $f_k(X, \theta_k)$ and K represents the total number of components with $\sum_{k=1}^{K} \pi_k = 1$, $\Theta = (\Pi', \theta_1', \theta_2', ..., \theta_K')$ and $\Pi = (\pi_1, \pi_2, .., \pi_K)'$.

Usually, it is assumed that the functional form of $f_k(\ ,\ )$ is completely known. In many cases, mixture models with Gaussian components are widely used and t-distribution components are used when it is observed heavy-tailed distribution.

Finite mixture modeling is typically aimed at inference on the parameters, while model-based clustering is associated with identifying groups of homogeneous observations according to some pre-specified rules.

Therefore, model-based clustering needs an additional steps ,assigning each observations to different groups according to some pre-specified rules, and Bayes rule is commonly used.
That is, for each observation $X_i, i = 1, 2, .., n$, we choose a group index with the highest value's among $\pi_k f_k(X_i, \theta_k), k = 1, 2, .., K$.

2. Inference in finite mixture models

2.1 Estimation in finite mixture models

The EM algorithm is the primary tools for finding ML-estimate in finite mixture models and model-based clustering.

E-step(of s-th iteration)

$$\pi_{ik}^{(s)} = prob\left[X_i \in k-th\ cluster | X_i ; \Theta^{(s-1)}\right] = \frac{\pi_k^{(s-1)} f_k(X_i;\ \theta_k^{(s-1)})}{\sum_{j=1}^{K} \pi_j^{(s-1)} f_j(X_i;\ \theta_j^{(s-1)})}$$

M-step

maximizes the expected conditional complete log-likelihood, historeically denoted as Q-function, with respect to the parameter vector Θ

Stopping criterion

In many case, use Aitken's rule:

$\left| l_A^{(s+1)} - l_A^{(s)} \right| < \epsilon$,where ε is the tolerance level and

$l_A^{(s)}$ is the Aitken accelerated estimate of the limiting value such that

$$l_A^{(s+1)} = l_A^{(s)} + \frac{l_A^{(s+1)} - l_A^{(s)}}{1 - \frac{l_A^{(s+1)} - l_A^{(s)}}{l_A^{(s)} - l_A^{(s-1)}}}$$

2.2 Challenges in implementation

2.2.1 Unbounded likelihood functions

Contrary to Gaussian mixtures with homogeneous components, Gaussian mixtures with heterogeneous dispersions may have unbounded log likelihood function. Because covariance matrix being estimated could be singular as a consequence of degraded components that have only one observation, or several nearly-identical observations.

There are several suggested methods to address this problem:

In the case of univariate normal components, Hathaway(1985) suggested introducing additional constraints:

$\sigma_i^2 \sigma_j^{-2} \geq c > 0$ for any i and j .

In the case of multivariate normal components, McIntyre and Blashfield (1980) suggested restrictions $|\Sigma_i|^{-1}|\Sigma_j| \geq c > 0$ for any i and j .

Here, c has to be pre-specified but it is unclear how to choose a reasonable value.

Regardless of mixture distribution of components, we can use a penalized log likelihood function that contains a penalty term (Chen and Li ,2008)

2.2.2 Initialization of the EM Algorithm

Choosing the proper starting point is important, because the log likelihood could have numerous local maxima. Many different methods have been suggested but no strategy works uniformly well.

A model-based hierarchical clustering approach was proposed (Banfield and Raftery ,1993).

however, its application is limited in larger data sets.

Maitra (2009) suggested approach based on finding the most separated

local modes, which is very time-consuming for severely multi-dimensional data sets.

In practice, try different strategies and choose the solution with the highest log likelihood value.

2.3 Model selection

2.3.1 Choosing the optimal number of components

Most methods for estimating the number of components are divided into two categories. The first group of methods is parsimony-based and the second group of methods relies on testing procedures.

parsimony-based approaches:
choose the K minimizing the negative log likelihood function augmented by some penalty function to reflect its complexity

"Various information-based criteria such as An Information Criterion (AIC) (Akaike, 1973), Bayes Information Criterion (BIC) (Schwarz,1978) and their modifications such as quadratic AIC/BIC (Ray and Lindsay,2008), the Integrated Classification Likelihood criterion (ICL)(Biernacki et al.,2000)., Normalized Entropy Criterion (NEC) (Biernacki et al., 1999), Minimum Information Ratio criterion (MIR) (Windham and Cutler, 1992), and Laplace-Empirical Criterion(LEC) (McLachlan and Peel ,2000) fall into this category."

However, it is impossible to have an exact criteria figure to evaluate a model performance improvements because it may depend on n(sample size) and p(number of parameters)(Kass and Raftery,1995)

testing-based approaches:
Most testing-based approaches use a likelihood ratio test (LRT) or some

derivation thereof. However, direct application of LRT is impossible due to boundary problem.

To solve this problem, several methods have been suggested.

(Aitkin and Rubin,1985) suggested moving parameter vectors into the interior of parameter space.

(Feng and McCulloch,1996) recommeded bootstraping the LRT statistic.

(Maitra and Melnykov, 2010a) proposed a likelihood based testing procedures.

"To keep derivations of the null distribution of the LRT statistic tractable, we introduced an additional assumption stating that a fit of the (simpler) model under the null hypothesis H_0 implies that the alternative (and more complex) model under H_a also fits the data adequately. Under H_a however, only the alternative model provides a good fit."

2.3.2 Variable selection

In many multivariate data sets, some of the variables are highly correlated with others, so that they do not carry much additional information. The elimination of such variables can improve model performance. Raftery and Dean(2006) introduced a greedy variable selection algorithm based on Bayes factors. But it did not allow irrelevant variables to be independent of clustering variables, potentially leading to erroneous model choices.

(Pan and Shen, 2006) proposed new approaches using the regularized log-likelihood function penalized by the term $-\lambda\sum_{k=1}^{K}\sum_{j=1}^{p}|\mu_{kj}|$ where μ_{kj} is the j-th coordinate of the k-th mean vector.

Xie and Shen(2008) extended this approach by including a new regularization scheme that groups together multiple parameters of the same variable across clusters.

3. Some recent applications

3.1 Magnitude magnetic resonance imaging data

Data sets from MRI or MRA are typically magnitudes of complex observations, whose real and imaginary parts are both independent univariate Gaussian-distributed realization (Wang and Lei,1994).

Thus in stead of using Gaussian mixtures to segment these data sets, Chung and Noble (1999) and Maitra and Faden(2009) suggest using a mixture of Rice distributions to characterize the MR signal at each voxel. The Rice distribution is given by the density function

$f(X;\Theta)=\sum_{k=1}^{K}\pi_k f_k(X;\mu_k,\sigma^2)$, where π_k represents the proportion of voxels with signal μ_k and the noise parameter σ is assumed to be common for all k= 1, 2, . . . ,K

Maitra and Faden(2009) rovide details on computational implementation and application of EM algorithm on this literature.

3.2 Finite mixtures models in surveys

When identifying clusters of individual from respondents to multiple-choice questions in surveys in order to tailor and market products and surveys, we can use finite mixture models Maitra and Melnykov,2010a).

Suppose there are p questions and for each $j-th$ question, there are d_j responses. The respondent's choice for the j-th question can be modeled by a multinomial distribution:

$$f(X_{jr};\rho_{jr}|r=1,2,...,d_j)=n_j\prod_{r=1}^{d_j}\frac{\rho_{jr}^{x_{jr}}}{x_{jr}!},\ x_{jr}=0,1,$$

,where ρ_{jr} is the probability that r-th option has been selected while x_{jr} represents the actual choice made by a respondent.

Note that $n_j = \sum_{r=1}^{d_j} x_{jr}$ and $\sum_{r=1}^{d_j} \rho_{jr} = 1$.

If we assume the independence of multinomial variables for the reponses to each questions, p-responses observation from each individuals can be modeled as a finite products-of-multinomials mixture model:

$$g(X_{jr}; \pi_k, \rho_{kjr} | k = 1,2,..,K, j = 1,2,..,p, r = 1,2,.,d_j) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} n_j \prod_{r=1}^{d_j} \frac{\rho_{kjr}^{x_{jr}}}{x_{jr}!} , x_{jr} = 0,1,$$

4. Some additional topics and challenges

4.1. Hierarchical model-based clustering and cluster merging

When all mixture components are well-separated, one-to-one correspondence between every component in a fitted mixture components and one cluster holds. However, in other cases, it may well be that one group is better modeled using several mixture components. Thus, by merging components, clustered partition of the data sets can be represented by several mixture components. For this purpose, model-based hierarchical clustering has been suggested(Goldberger and Roweis, 2004).

Goldberger and Roweis(2004) suggested using a single Gaussian components replacing each group of of Gaussian components in a Gaussian mixture models.

Hennig(2010) also discussed hierarchical merging methods using concepts of unimodality and misclassification.

Baudry(2010) suggested two-stage approaches with regard to choosing the number of clusters based on merging. At the first stage, using BIC, find the number of Gaussian components.

At the second stage, using ICL, eliminate unnecessary components and merge them hierarchically.

4.2 Non-parametric approaches to mixture modeling and model-based clustering

To resolve the problem addressed in 4.1, we can use non-parametric mixture modeling. we assume that the observations are from a mixture of densities: $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$

The basic idea is to associate clusters with a local maximum, or mode.

For this procedure, Li and Lindsay(2007) suggested EM-type non-parametric algorithm called Modal EM. The suggested algorithm is then extended to hierarchical clustering.

Minnotte and Scott(1993) proposed different approach using visualizing tool called a mode tree.

4.3 Semi-supervised clustering

In many situations, we can not obtain information about which classes of some observations belong to which group. In this case, we need adaptations to the EM algorithm:

M step is as before. But E step needs to be changed.

Posterior probabilities for labeled data do not need to be updated.

The other probabilities corresponding to unlabeled data are computed as usual.

Discussion, so far, assume that all the classes in the entire data sets are represented in the classes represented in labeled data so that K is known and model selection is not an issue.

However, if assumption does not hold, several issues arise.

initialization in the EM algorithm:

One option is to consider only unlabeled information ,ignoring the labeled observation.

But, by considering both labeled and unlabeled data, we can improve performance.

model selection:

" Chen et al. (2010) have advocated using the quantitation map for choosing the model at desired significance and have shown excellent performance on a range of simulation and classification data sets."

4.4 Constrained clustering

Consider, for instance, the example of two-dimensional gel electrophoresis, where there are a given number of proteins and an equivalent number of

protein spots. Our aim is to assign each observed spot to the corresponding protein. In this situation, this brings in a constraint that no two spots can be assigned to the same protein and we need modification in the E-step of EM algorithm (Morris and Gutstein ,2008)

4.5 Diagnostics

Influential and outlying observations impact performances of many model-based clustering algorithms. In general, the general method to identify influential observations is not known. For the case of identifying out-liers, two approaches have been suggested.

McLachlan and Basford (1988) proposed what they called an atypicallity measure.

Wang et al.(1997) suggested using a modified likelihood ratio test comparing two models.

parametric or non-parametric bootstrap is recommended for assessing the null distribution of the obtained test statistic. Modified bootstrap also can be used for large data sets, because it is computationally less demanding.

4.6 Robust and skewed mixture models

If outliers can dramatically affect all estimates, it is important to develop mixture models that are robust to outliers.

Peel and McLachlan (2000) proposed using a mixture of multivariate t-distributions instead of multivariate Gaussians.

When the data represents non-Gaussian patterns, Azzalini (2005) suggested using a skewed-normally distributed mixture components.

Azzalini (1985) introduced the density of the univariate skew normal distributions:

$$\psi(x;\mu,\sigma,\lambda) = \frac{2}{\sigma}\varphi(\frac{x-\mu}{\sigma})\Phi(\lambda\frac{x-\mu}{\sigma}),$$

where $\phi()$ is the probability density function of a standard univariate normal distribution, while $\Phi()$ is the corresponding cumulative distribution

function.

The parameters μ and σ here have meaning similar to their counterparts for the normal distribution while λ represents the skewness parameter.

Azzalini and Dalla(1996) introduced multivariate skew normal distributions.

4.7 Dependent data

Suppose that we have n observations $Y=(Y_1, Y_2, .., Y_n)'$ consisting of univariate normally distributed observations following an autoregressive AR(1) model, and that there are K components with mean μ_k, $k=1,2,...,K$ and common variance σ^2.

Let $X=\begin{pmatrix} I_{11} .. I_{1K} \\ I_{21} .. I_{2K} \\ : \ : \ : \\ I_{n1} .. I_{nK} \end{pmatrix}$,where every I_{ik} represents the indicator function I($Y_i \in k-$th cluster).

If the class memberships of observations are known, it can be written in the form $Y \sim MVN(X\beta, \sigma^2 R(\rho)), where\ R(\rho) is\ a\ correlation\ structure$.
Although expressions for the M-step of the EM algorithm can be derived, expressions for the EM iterations are more complicated and involve taking derivatives of $R^{-1}(\rho)$ with respect to ρ, for which closed-form expressions may not be available. As a result, while the EM algorithm can be set up and used for parameter estimation in the same way as usual, estimation becomes far more difficult. Estimation of variance σ^2 is also difficult, in the case of dependent observations.

5. Conclusion

This paper provides an overview of mixture models with specific reference to model-based clustering.

Two applications involving finite mixture distributions are presented and some additional topics such as semi-supervised clustering, constrained-clustering, diagnostics and dependent observations are presented and unresolved challenges outlined.

The fields has attracted a lot of interest, but there are still many problems and issues that remain unresolved ,as we have seen.

References

[1] Aitkin, M. and Rubin, D. (1985). Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society B 47, 67-5.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second international symposium on information theory. 267-81. MR0483125

[3] Azzalini, A. (1985). A class of distributions which includes the normal ones. Scandinavian Journal of Statistics 12, 171-78. MR0808153

[4] Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion). Scandinavian Journal of Statistics 32, 159-00. MR2188669

[5] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. Biometrika 83, 715-26. MR1440039

[6] Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803-21. MR1243494

[7] Baudry, J.-P., Raftery, A., Celeux, G., Lo, K., and Gottardo, R. G. (2010). Combining mixture components for clustering. Journal of Computational and Graphical Statistics, to appear.

[8] Biernacki, C., Celeux, G., and Gold, E. M. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 719-725.

[9] Biernacki, C., Celeux, G., and Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Pattern Recognition Letters 20, 267-72.

[10] Chen, J. and Li, P. (2008). Hypothesis testing for normal mixture models: the EM approach. submitted to Annals of Statistics.

[11] Chen, W.-C., Maitra, R., and Melnykov, V. (2010). Model-based

semi-supervised clustering. In preparation.

[12] Chung, A. C. S. and Noble, J. A. (1999). Statistical 3d vessel segmentation using a Rician distribution. In MICCAI. 82-9.

[13] Feng, Z. and McCulloch, C. (1996). Using bootstrap likelihood ratio in finite mixture models. Journal of the Royal Statistical Society B 58, 609-17.

[14] Goldberger, J. and Roweis, S. (2004). Hierarchical clustering of a mixture model. NIPS 2004.

[15] Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Statistics & Probability Letters 4, 53-6. MR0790575

[16] Hennig, C. (2010). Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification 4, 3-4.

[17] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association 90, 773-95.
V. Melnykov and R. Maitra/Finite mixture models and model-based clustering 113

[18] Li, J., Ray, S., and Lindsay, B. (2007). A nonparametric statistical approach to clustering via mode identification. The Journal of Machine Learning Research 8, 1687-723. MR2332445

[19] Li, P., Chen, J., and Marriott, P. (2008). Non-finite Fisher information and homogeneity: an EM approach. Biometrika, 1-5.

[20] Maitra, R. (2009). Initializing partition-optimization algorithms. IEEE/ACM Transactions on Computational Biology and Bioinformatics 6, 144-57.

[21] Maitra, R. and Faden, D. (2009). Noise estimation in magnitude MR datasets. IEEE Transactions on Medical Imaging 28, 10, 1615-622.
V. Melnykov and R. Maitra/Finite mixture models and model-based clustering 114

[22] Maitra, R. and Melnykov, V. (2010a). Assessing significance in finite mixture models. Tech. Rep. 10-01, Department of Statistics, Iowa State University.

[23] McIntyre, R. M. and Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research 15, 225-38.

[24] McLachlan, G. and Peel, D. (2000). Finite Mixture Models. John Wiley and Sons, Inc., New York. MR1789474

[25] McLachlan, G. J. and Basford, K. E. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York. MR0926484

[26] Minnotte, M. and Scott, D. (1993). The mode tree: a tool for visualization of nonparametric density features. Journal of Computational and Graphical Statistics 2(1), 51-8.

[27] Morris, J. S., Clark, B. N., and Gutstein, H. B. (2008). Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data. Bioinformatics 24, 529-536.

[28] Pan, W. and Shen, X. (2006). Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research 8, 1145-164.

[29] Peel, D. and McLachlan, G. (2000). Robust mixture modeling using the t-distribution. Statistics and Computing 10, 339:348.

[30] Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. Journal of the American Statistical Association 101, 168-78. MR2268036

[31] Wang, S. J., Woodward, W. A., Gray, H. L., Wiechecki, S., and Satin, S. R. (1997). A new test for outlier detection from a multivariate mixture distribution. Journal of Computational and Graphical Statistics 6, 285-99. MR1466869

[32] Wang, T. and Lei, T. (1994). Statistical analysis of MR imaging and its application in image modeling. In Proceedings of the IEEE International Conference on Image Processing and Neural Networks. Vol. 1. 866-70.

[33] Windham, M. P. and Cutler, A. (1992). Information ratios for validating mixture analyses. Journal of the American Statistical Association 87,1188-192.

[34] Xie, B., Pan, W., and Shen, X. (2008). Variable selection in penalized model-based clustering via regularization on grouped parameters. Bioinformatics 64, 921-30.