ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Ewa Witek**

ON AN IMPROVEMENT OF THE MODEL-BASED CLUSTERING METHOD

Abstract. An improvement of the model-based clustering (MBC) method in the case when EM algorithm fails as a result of singularities is the basic aim of this paper. Replacement of the maximum likelihood (MLE) estimator by a maximum a posteriori (MAP) estimator, also found by the EM algorithm is proposed. Models with different number of components are compared using a modified version of BIC, where the likelihood is evaluated at the MAP instead of MLE. A highly dispersed proper conjugate prior is shown to avoid singularities, but when these are not present it gives similar results to the standard method of MBC.

Key words: Model-based clustering (MBC), Gaussian mixture models, EM algorithm, MLE, MAP, BIC, conjugate prior.

I. MODEL-BASED CLUSTERING

In model-based clustering, individual clusters are described by multivariate normal distributions, where the class labels, parameters and proportions are unknown. The data $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{im}]^T$ are assumed to be generated by a mixture with density:

$$f(|\mathbf{x}) = \prod_{i=1}^{n} \sum_{s=1}^{u} \tau_s f_s(\mathbf{x}_i|\Theta_s), \tag{1}$$

where $f_s(\mathbf{x}_i|\Theta_s)$ is a probability distribution with parameters Θ_s, and τ_s is the probability of belonging to the sth component. The parameters of the model are usually estimated by maximum likelihood using the Expectation-Maximization (EM) algorithm (Dempster et al. [1997]). Each EM iteration consist of two steps

* Ph.D student, Department of Statistics, The Karol Adamiecki University of Economics, Katowice

– an E-step and an M-step. Given an initial guess for the cluster means $\boldsymbol{\mu}_s$, covariances $\boldsymbol{\Sigma}_s$ and proportions τ_s, the E-step calculates the conditional probability that object i belongs to the sth component:

$$\hat{\mathbf{z}}_{is} = \frac{\hat{\tau}_s f_s(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_s)}{\sum_{r=1}^{u} \hat{\tau}_r f_r(\mathbf{x}_i | \hat{\boldsymbol{\mu}}_r, \hat{\boldsymbol{\Sigma}}_r)} \tag{2}$$

The maximization step (M-step) consists of estimating the parameters from the data and the conditional probabilities $\hat{\mathbf{z}}_{is}$. The E- and M-steps iterate until convergence. Finally, each object is classified in the class in which it has the highest conditional or posterior probability. The results of the EM are highly dependent on the initial values, model-based hierarchical clustering can be a good solution (Banfield and Raftery [1993]; Dasgupta and Raftery [1998])

In order to select the optimal clustering, model several measures have been proposed (McLachlan and Peel [2000]). In several applications, the BIC approximation to the Bayes factor (Schwarz [1978]) has performed quite well (Dasgupta and Raftery [1998], Fraley and Raftery [1998], [2002]). The BIC has the form:

$$BIC_s = 2 \log p(\mathbf{x} | \hat{\boldsymbol{\Theta}}_s, M_s) - v_s \log(n), \tag{3}$$

where $\log p(\mathbf{x} | \hat{\boldsymbol{\Theta}}_s, M_s)$ is the maximized loglikelihood for the model and data, v_s is the number of parameters to be estimated in the model M_s and n is the number of observations in the data.

The strategy for model selection has been found to be effective in mixture estimation and clustering is given bellow:

1. Determine a maximum number of clusters, u, (as small as possible) and a set of mixture models to consider.

2. Estimate parameters via EM for each parameterization and each number of components up to u. The conditional probabilities corresponding to a classification from model-based hierarchical clustering are good choices for initial values.

3. Compute the BIC for the mixture model using the optimal parameters from EM for 2,...,u clusters. This results with a matrix of BIC values corresponding to each possible combination of parameterization and number of clusters.

4. Plot all of the BIC values. A decisive first local maximum indicates strong evidence for a model (parameterization and number of clusters).

For a review of model-based clustering, see Fraley and Raftery (2002).

II. LIMITATIONS OF EM ALGORITHM

The EM algorithm for clustering has a number of limitations. First, the rate of convergence can be very slow. This does not appear to be a problem in practice for well-separated mixtures when started with reasonable values. Second, the number of conditional probabilities associated with each observations is equal to number of components in the mixture, so that the EM algorithm for clustering may not be practical for models with very large numbers of components. Finally, EM breaks down when the covariance matrix corresponding to one or more components becomes ill-conditioned (singular or nearly singular). In general it cannot proceed if clusters contain only a few observations or if the observations they contain are very nearly collinear. If EM for a model having a certain number of components is applied to a mixture in which there are actually fewer groups, then it may fail due to ill-conditioning.

III. BAYEASIAN REGULARIZATION FOR MULTIVARIATE NORMAL MIXTURES

Fraley and Raftery (2005) proposed a replacement of the MLE by the maximum a posteriori (MAP) estimate from a Bayesian analysis to eliminate convergence failures of the EM algorithm. They proposed a prior distribution on the parameters that eliminates failure due to singularity, while having little effect on stable results obtainable without prior. The Bayesian predictive density for the data is assumed to be of the form

$$L_{mix}(\mathbf{x}|\tau_s,\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s) = \mathrm{P}(\tau_s,\boldsymbol{\mu}_s,\Sigma_s|\boldsymbol{\Theta}),$$

Where L_{mix} is the mixture likelihood:

$$L_{mix}(\mathbf{x}|\tau_s,\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s) = \prod_{i=1}^{n}\sum_{s=1}^{u}\tau_s\phi(\mathbf{x}_i|\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s)$$

$$= \prod_{i=1}^{n}\sum_{s=1}^{u}\tau_s\left|2\pi\boldsymbol{\Sigma}_s\right|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_s)^T\boldsymbol{\Sigma}_s^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_s)\}, \qquad (4)$$

and P is a prior distribution on the parameters τ_s, $\boldsymbol{\mu}_s$ and $\boldsymbol{\Sigma}_s$. Fraley and Raftery (2005) proposed to find a posteriori mode or MAP (maximum a posteri-

ori) rather than a maximum likelihood estimate for the mixture parameters. They used BIC for model selection, but in a modified form- the first term on the right-hand side of (3), equal to twice the maximized log-likelihood is replaced by twice the log-likelihood evaluated at the MAP or posterior mode.

For multivariate data, a normal prior on the mean (conditional on the covariance matrix) has a form:

$$\boldsymbol{\mu} \sim |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{\kappa_p}{2} tr\left[(\boldsymbol{\mu}-\boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_p)\right]\right\}, \tag{5}$$

and an inverse Wishart prior on the covariance matrix:

$$\boldsymbol{\Sigma} \sim |\boldsymbol{\Sigma}|^{-\frac{\nu_p+m+1}{2}} \exp\left\{-\frac{1}{2} tr\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_p^{-1}\right]\right\}. \tag{6}$$

Hyperparameters $\boldsymbol{\mu}_p$, κ_p, ν_p, are called *mean*, *shrinkage* and *degrees of freedom* respectively, of the prior distribution. The hyperparameter $\boldsymbol{\Lambda}_p$, which is a matrix, is called the scale of the inverse Wishart prior. The prior defined in this way is called *conjugate prior* for a multivariate normal distribution and an inverse Wishart distribution. Under this prior, the posterior means of the mean vector and covariance matrix are:

$$\hat{\boldsymbol{\mu}} = \frac{n\bar{\mathbf{x}} + \kappa_p \boldsymbol{\mu}_p}{\kappa_p + n}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Lambda}_p^{-1} + (\frac{\kappa_\rho n}{\kappa_\rho + n})(\bar{\mathbf{x}}-\boldsymbol{\mu}_p)(\bar{\mathbf{x}}-\boldsymbol{\mu}_p)^T + \Sigma_{i=1}^n (\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T}{\tilde{\nu}_p + m + 2}. \tag{7}$$

The normal inverted Wishart prior and its conjugacy to the multivariate normal are discussed in e.g. Gelman et al. (1995) and Schafer (1997).

Fraley i Raftery (2005) proposed the following choices for the *prior* hyperparemeters $(\kappa_p, \nu_p, \boldsymbol{\Lambda}_p, \varsigma_p^2)$ for multivariate mixtures:

μ_p is the mean of the data,

$\kappa_p = 0{,}01.$

The posterior mean $\frac{n_s\bar{\mathbf{x}}_s + \kappa_p\boldsymbol{\mu}_p}{\kappa_p + n_s}$ can be viewed as adding κ_p observations with value $\boldsymbol{\mu}_p$ to each group in the data. The value was determined by experimentations. Values close to and bigger than 1 caused large perturbations in the cases where there were no missing BIC values without prior. $\kappa_p = 0{,}01$ resulted in BIC curves that appeared to be smooth extensions to their counterparts without the prior.

$$\nu_p = m + 2 \tag{8}$$

The marginal prior distribution for $\boldsymbol{\mu}$ is multivariate t centered $\boldsymbol{\mu}_p$ with $\nu_p - m + 1$ degrees of freedom. The mean of this distribution is $\boldsymbol{\mu}_p$ provided that $\nu_p > m$, and it has a finite covariance matrix provided $\nu_p > m + 1$ (Schafer [1997]).

$\varsigma_p^2 = \frac{tr(S)/m}{u^{2/m}}$ (for spherical and diagonal models). The average of the diagonal elements of the empirical covariance matrix of the data- **S** divided by the number of components to the $2/m$ power.

$\boldsymbol{\Lambda}_P = \frac{\mathbf{S}}{u^{2/m}}$ (for ellipsoidal models) the empirical covariance matrix of the data divided by the square of the number of component to the $1/m$.

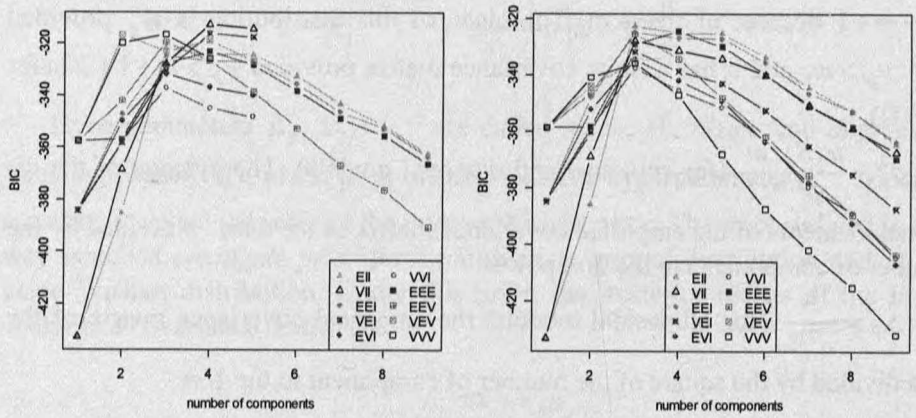
IV. EXAMPLE

The data was generated by cluster.Gen function (cluster.Sim package of **R**). Three elongated clusters contain two-dimensional data. The number of observations in each classes is: 13, 10, 13. The observations are independently drawn from bivariate normal distribution with means (0;0), (1.5;7), (3;14) and covariance matrices: $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & -0{,}9 \\ -0{,}9 & 1 \end{bmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1{,}5 & 0 \\ 0 & 1{,}5 \end{bmatrix}$, $\boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0{,}5 \\ 0{,}5 & 1 \end{bmatrix}$.

Functions of mclust package of **R** were implemented to Bayesian regularization for mixture models

For the analyzed dataset the model and classification chosen according BIC without prior chooses four component VII model with four components, when the known number of components is three. The standard BIC values based on the MLE are not available for six models (VII, VEI, EVI, VVI, VEV, VVV)

with five or more mixture components. For those number of components models fail to converge without the prior because one of the covariances becomes singular as the EM iterations progress, as shown in Figure 1a). The hierarchical clustering result based on the unconstrained model used for initialization assings a single observation to one of the groups in those cases. The Bayesian regularization allows identification of a group with a single member while allowing the covariance matrix to vary between clusters, which is not possible without the prior. The BIC with the prior peaks the 3 groups classification for EII model. The EII model with three components is chosen according to BIC with prior. In this case failures due to singularity for almost all models are eliminated and the right number of clusters is selected.



a) without prior b) with prior

Figure 1. BIC values

Source: Own research.

IV. CONCLUSIONS

We have shown an improvement of the model-based clustering for avoiding the singularities that can arise in estimation using EM algorithm. The method involves a proper conjugate prior and uses the EM algorithm to find the MAP estimator. For model selection it uses a version of BIC that is modified by replacing the maximized likelihood by the likelihood evaluated at the MAP.

REFERENCES

Banfield J.D., Raftery A.E. (1993), *Model-based Gaussian and non-Gaussian clustering*, „Biometrics", 49, 803–821.

Biernacki C, Celeux G., Govaert G., Langrognet F. (2006), *Model-based cluster and disriminant analysis with the MIXMOD software*, „Computational Statistics and Data Analysis", 51, 587–600.

Dasgupta A., Raftery A.E. (1998), *Detecting features in spatial point processes with clutter via model-based clustering*, „Journal of the American Statistical Association", 93, 294–302.

Dempster A.P., Laird N.M., Rubin D.B. (1977), *Maximum likelihood for incomplete data via the EM algorithm (with discussion)*, „Journal of the Royal Statistical Society", ser. B, 39, 1–38.

Fraley C., Raftery A.E. (1998), *How many clusters? Which clustering method? Answers via model-based cluster analysis*, „The Computer Journal", 41, 577–588.

Fraley C., Raftery A.E. (2002), *Model-based clustering, discriminant analysis, and density estimation*, „Journal of the American Statistical Association", 97, 611–631.

Fraley C., Raftery A.E. (2005), *Bayesian regularization for normal mixture estimation and model-based clustering*, Technical Report 486, Department of Statistics, University of Washington

Fraley C., Raftery A.E. (2006), MCLUST *Version 3: An R package for normal mixture modeling and model- based clustering*, 1–50.

Gelman A., Carlin J.B., Stern H.S, Rubin D.B. (1995), *Bayesian data analysis*, Chapman and Hall, London.

McLachlan G.J., Peel D. (2000), *Finite mixture models*, Wiley, New York.

Schafer J.L. (1997), *Analysis of incomplete multivariate data by simulation*, Chapman and Hall, London.

Schwarz G. (1978), *Estimating the dimension of a model*, „The Annals of Statistics", 6, 461–464.

Ewa Witek

O PEWNEJ MODYFIKACJI W METODZIE TAKSONOMII OPARTEJ NA MODELACH MIESZANYCH

W artykule przedstawiona została modyfikacja metody taksonomii opartej na modelach mieszanych, w przypadku gdy niemożliwym staje się oszacowanie parametrów modelu za pomocą algorytmu EM. Gdy liczba obiektów przypisanych do klasy jest mniejsza niż liczba zmiennych opisujących te obiekty, niemożliwym staje się oszacowanie parametrów modelu. By uniknąć tej sytuacji estymatory największej wiarygodności zastępowane są estymatorami o największym prawdopodobieństwie a posteriori. Wybór modelu o najlepszej parametryzacji i stosownej liczbie klas dokonywany jest wówczas za pomocą zmodyfikowanej statystyki BIC.