Model based clustering with t mixture and different number of replicates

Yonghyun Kwon

5/10/2020

t mixtures models

We consider clustering a dataset y consisiting of $N = \sum_{i=1}^n J_i$ examples into K clusters, where $y_i = (y_{i1}, \cdots, y_{iJ_i})$ are genereated from the same cluseter. z_{ik} is the latent indicator variable, which represents the indicator whethjer case $i1, \cdots, iJ_i$ belongs to class k.

$$z_{ik} = I(\text{case } i \text{ belongs to class } k)$$

This problem can be modeled as the following t mixture model.

$$\begin{gathered} y_{i1}, \cdots, y_{iJ_i} \mid z_{ik} = 1 \overset{iid}{\sim} t_p(\mu_k, \Sigma_k, \nu_k) \\ z_i \overset{iid}{\sim} multinomial(1, \pi_1, \cdots, \pi_K) \end{gathered}$$

This model can be rewritten using a latent characteristic weight u_{ijk}.

$$\begin{gathered} y_{ij} \mid z_{ik} = 1,\ u_{ijk} \overset{iid}{\sim} N\left(\mu_k, \frac{\Sigma_k}{u_{ijk}}\right) \\ u_{ijk} \mid z_{ik} = 1 \overset{iid}{\sim} \gamma\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right) \\ z_i \overset{iid}{\sim} multinomial(1, \pi_1, \cdots, \pi_K) \end{gathered}$$

E-step

Since direct EM-algorithm is not applicable in this problem setup, we follow one variant of EM-algorithm called ECM(Expectation Conditional Maximization) algorithm.(Meng and Rubin, 1993) Moreover, we employ AECM(alternating expectation-conditional maximization) suggested from Andrews and McNicholas, 2010 to permit different specification of complet data at each stage. At E-step, each indicator variables z_{ik} and u_{ijk} are updated using conditional expectations.

$$\hat{z}_{ik} = E[z_{ik} \mid y_i] = \frac{\pi_k t_p(y_{i1}; \mu_k, \Sigma_k, \nu_k) \cdots t_p(y_{iJ_i}; \mu_k, \Sigma_k, \nu_k)}{\sum_{k'=1}^K \pi_{k'} t_p(y_{i1}; \mu_{k'}, \Sigma_{k'}, \nu_{k'}) \cdots t_p(y_{iJ_i}; \mu_{k'}, \Sigma_{k'}, \nu_{k'})}$$

The condtional expectation is from the joint distribution of (y_i, z_{ik})

$$f(y_i, z_{ik}) = \pi_k t_p(y_{i1}; \mu_k, \Sigma_k, \nu_k) \cdots t_p(y_{iJ_i}; \mu_k, \Sigma_k, \nu_k)$$

The characteristic weights u_{ijk} are updated using

$$\hat{u}_{ijk} = E[u_{ijk} \mid y_{ij}, z_{ik} = 1] = \frac{\nu_k + p}{\nu_k + \delta(y_{ij}; \mu_k; \Sigma_k)}$$

where $\delta(y_{ij};\mu_k;\Sigma_k)$ denotes the squared Mahalanobis distance between y_{ij} and μ_k given by

$$\delta(y_{ij};\mu_k;\Sigma_k) = (y_{ij}-\mu_k)^T\Sigma_k^{-1}(y_{ij}-\mu_k)$$

and this conditional expectation can be derived from the posterior distribution of u_{ijk}

$$u_{ijk} \mid y_{ij}, z_{ik} = 1 \sim \gamma\left(\frac{\nu_k + p}{2}, \frac{\nu_k + \delta(y_{ij};\mu_k;\Sigma_k)}{2}\right)$$

CM-1 step

For CM-1 step, the complete log likelihood is given by

$$l(\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left[z_{ik}\log\pi_k + z_{ik}\sum_{j=1}^{J_i}\log\gamma\left(u_{ijk};\frac{\nu_k}{2},\frac{\nu_k}{2}\right) + z_{ik}\sum_{j=1}^{J_i}\log N\left(y_{ij};\mu_k,\frac{\Sigma_k}{u_{ijk}}\right)\right]$$

replacing z_{ik} with $\hat{z}_{ik}$ and u_{ijk} with $\hat{u}_{ijk}$, we get $Q(\Theta \mid \Theta^*)$ for M-step and one can derive the updating rule for π_k and μ_k

$$\pi_k \leftarrow \frac{\sum_{i=1}^{n}\hat{z}_{ik}}{\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}}, \quad \mu_k \leftarrow \frac{\sum_{i=1}^{n}\sum_{j=1}^{J_i}\hat{z}_{ik}\hat{u}_{ijk}y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{J_i}\hat{z}_{ik}\hat{u}_{ijk}}$$

To update ν_k, we differentiate the conditional expectation of the log-likelihood $l(\Theta)$ with respect to ν_k.

$$l(\nu_k) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\sum_{j=1}^{J_i}\left[-\log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}\log\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}(\log u_{ijk} - u_{ijk})\right] + const$$

Write $u_{ijk} \mid y_{ij} \sim \gamma(\alpha = \frac{\nu_k+p}{2}, \beta = \frac{\nu_k+\delta(y_{ij};\mu_k;\Sigma_k)}{2})$. Then the conditional expectation of $\log u_{ijk} \mid y_{ij}$ is

$$\begin{aligned} E[\log u_{ijk} \mid y_{ij}] &= -\log\beta + \psi(\alpha) \\ &= \log\frac{\alpha}{\beta} - \log\alpha + \psi(\alpha) \\ &= \log E[u_{ijk} \mid y_{ij}] - \log\frac{\nu_k+p}{2} + \psi\left(\frac{\nu_k+p}{2}\right) \end{aligned}$$

where ψ is digamma function.

Hence, the corresponding ν_k maximizes the conditional log likelihood function

$$\begin{aligned} Q(\nu_k \mid \nu_k^*)/n_k = -\log\Gamma\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}\log\left(\frac{\nu_k}{2}\right) + \\ \frac{\nu_k}{2n_k}\sum_{i=1}^{n} z_{ik}\sum_{j=1}^{J_i}\left[\log\hat{u}_{ijk} - \hat{u}_{ijk} - \log\frac{\nu_k^{old}+p}{2} + \psi\left(\frac{\nu_k^{old}+p}{2}\right)\right] \end{aligned}$$

where $n_k = \sum_{i=1}^{n}\hat{z}_{ik}J_i$. Differentiating with respect to ν_k, followed by multiplying 2, we get

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\frac{\nu_k}{2} + 1 + \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}\sum_{j=1}^{J_i}[\log\hat{u}_{ijk} - \hat{u}_{ijk}] + \psi\left(\frac{\nu_k^{old}+p}{2}\right) - \log\frac{\nu_k^{old}+p}{2} = 0$$

Hence, we can update ν_k by finding a solution of the non-linear equation above. Instead, one can use a novel closed-form approximation for ν_k discussed in Andrews et al., 2018.

$$v_k = \frac{-\exp(\kappa_k) + 2\exp(\kappa_k)\left[\exp\left(\psi(\frac{v_k^{old}}{2})\right) - \left(\frac{v_k^{old}}{2} - \frac{1}{2}\right)\right]}{1 - \exp(\kappa_k)} \tag{1}$$

where

$$\kappa_k = -1 - \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}\sum_{j=1}^{J_i}[\log \hat{u}_{ijk} - \hat{u}_{ijk}] - \psi\left(\frac{v_k^{old}+p}{2}\right) + \log\left(\frac{v_k^{old}+p}{2}\right)$$

If ν_k is same accross all the clusters(that is, constraints on the degree of freedom accross all the clusters, $\nu_k = \nu$), the update rule becomes

$$-\psi\left(\frac{\nu}{2}\right) + \log\frac{\nu}{2} + 1 + \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik}\sum_{j=1}^{J_i}[\log \hat{u}_{ijk} - \hat{u}_{ijk}] + \psi\left(\frac{\nu^{old}+p}{2}\right) - \log\frac{\nu^{old}+p}{2} = 0$$

where $N = \sum_{k=1}^{K} n_k$.

In this case, ν_k^{old} and κ_k in (1) are replaced by ν^{old} and

$$\kappa = -1 - \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik}\sum_{j=1}^{J_i}[\log \hat{u}_{ijk} - \hat{u}_{ijk}] - \psi\left(\frac{v^{old}+p}{2}\right) + \log\left(\frac{v^{old}+p}{2}\right)$$

CM-2 step

When estimating Σ_k, we can impose eigen decomposition on $\Sigma_k = \lambda_k D_k A_k D_k^T$ where D_k is the matrix of eigenvectors and A_k is the diagonal matrix with determinent 1. We can impose a constraint to these individual matrices. For instance, supppose $D_k = D$ and $A_k = A$ so that $\Sigma_k = \lambda_k DAD^T$.(VEE case) The complete log-likelihood is

$$l(\Sigma_k) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}J_i p\log(\lambda_k) - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\sum_{j=1}^{J_i}\frac{u_{ijk}}{\lambda_k}(y_{ij}-\mu_k)^T(DAD^T)^{-1}(y_{ij}-\mu_k)$$

Substituting the conditional expectation, Q function becomes

$$-\frac{1}{2}p\sum_{k=1}^{K} n_k\log\lambda_k - \frac{1}{2}\sum_{k=1}^{K}\frac{n_k}{\lambda_k}tr(S_kC^{-1})$$

where $C = DAD^T$ and

$$S_k = \frac{1}{n_k}\sum_{i=1}^{n}\hat{z}_{ik}\sum_{j=1}^{J_i}\hat{u}_{ijk}(y_{ij}-\mu_k)(y_{ij}-\mu_k)^T$$

Differentiating with respect to λ_k, estimated λ_k is

$$\lambda_k = \frac{tr(S_kC^{-1})}{p}$$

Applying Theorem A.1 from Celeux and Govaert (1995), C is updated as

$$C = \frac{\sum_{k=1}^{K}\frac{n_k}{\lambda_k}S_k}{\mid \sum_{k=1}^{K}\frac{n_k}{\lambda_k}S_k \mid^{1/p}}$$

For EVV case, refer to Celeux and Govaert, 1995.

More complicated cases such as EVE or VVE has been introduced in Andrews and McNicholas, 2012 and Browne and McNicholas, 2014.

Initialization and convergence assessment

Note that EM algorithm tends to converge to a local maximum of mle, which is generally not global. Therefore, we need to specify the initial value carefully and we use K-means algorithm for initialization. Also, to assess whether algorithm converges, we stop the iteration when $l(\Theta^{(t+1)}) - l(\Theta^{(t)}) < \varepsilon$ for small enough ε.

Model selection

Although we have assumed that the number of clusters is known, it is more likely that the number of clusters is unknown in practical application. To determine the number of clusters, one can use the Bayesian information criterion(BIC)

$$\text{BIC} = -2l(\hat{\Theta}) + m \log n$$

where $l(\Theta)$ is the maximized log likelihood, m is t he number of free parameters, and n is the sample size. One may experiment with different values of K and choose the one which attains the smallest BIC.

Performance assessment

To measure the class agreement, one can use the following popular measure: accuracy = (number of pairwise agreements) / (number of pairs), which takes a value on [0,1] and 1 indicates perfect agreement. Since there is multiple clusters, one can think of confusion matrix for each class to express accuracy.

Naive approach

One naive approach is to use kmeans(or EM algorithm) based only on the averages of each replicates. That is, k-means algorithm or model based clustering can be applied to each $\bar{y}_1, \cdots \bar{y}_n$. We will compare this naive method with the proposed one using all the data.

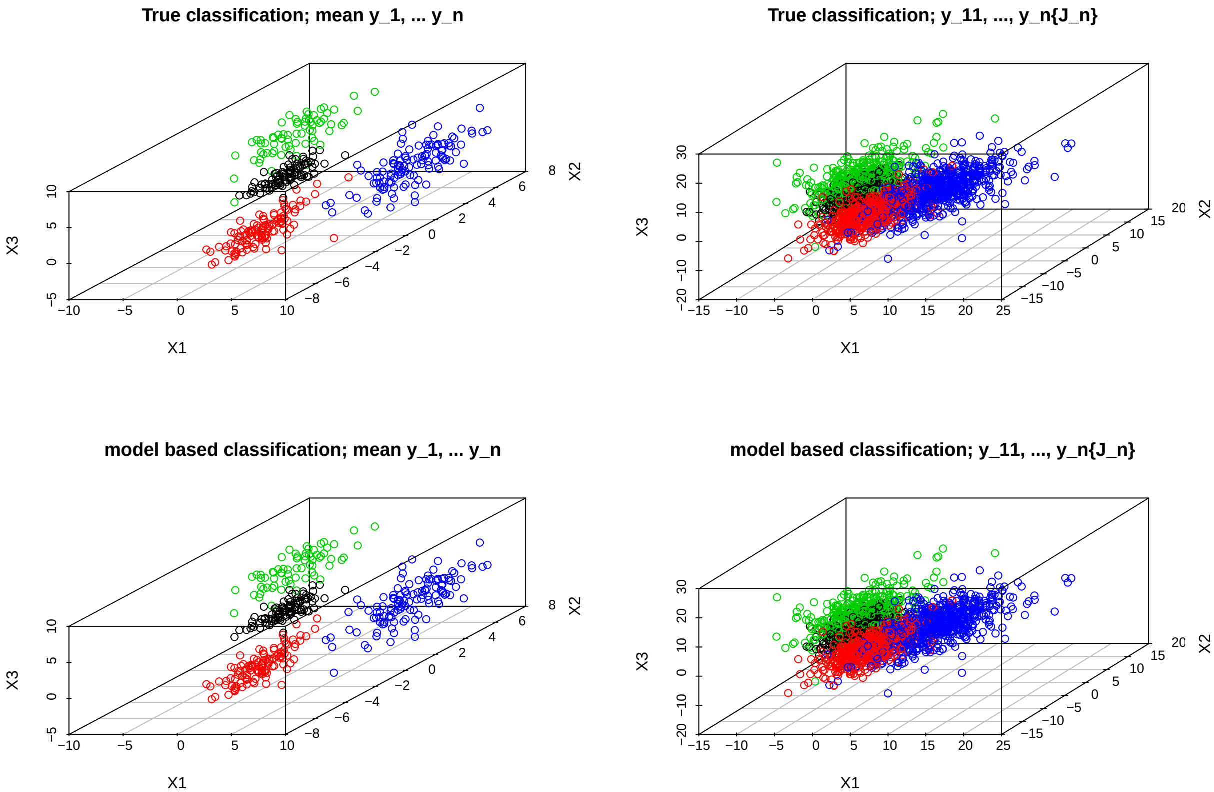
Simulation

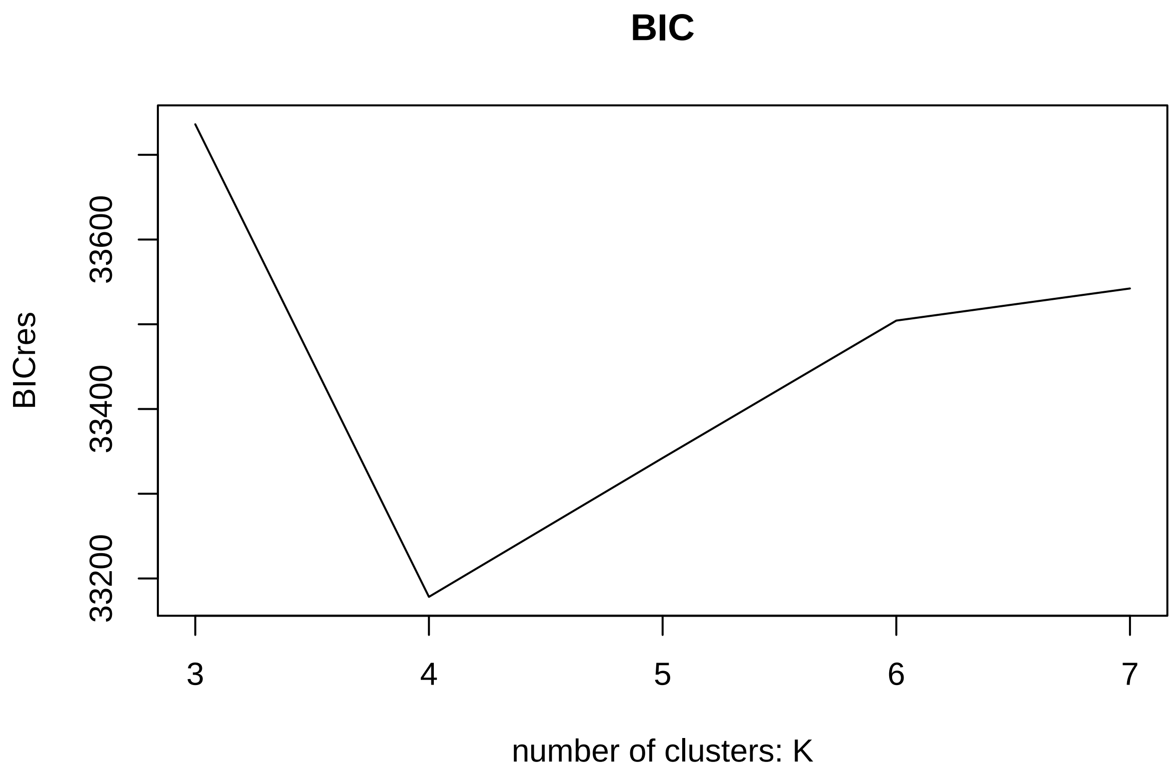
Simulated data

In order to evaluate the suggested methodology through simulation experiments, we first generate four-clusters($K = 4$) trivariate data($p = 3$) from a Normal distribution and investigate the preposed method with $n = 400$ sample size. Different number of replicates were simulated by generating $J = 10$ replicates for each element $i = 1, \cdots, n$ of and then dropping such replicates out with probability 0.5. There exists in turn $j = 1, \cdots, J_i$ replicates for each element i.

As presented in the following 3 dimensional scatter plot, the proposed model based clustering algorithm performs well with respect to classifcation accuracy. In this example, due to small within variance of each cluster of data, both proposed approach and naive one gives acceptable misclassification rate.(0.01, and 0.0275 respectively). While using group means for each elements $i = 1, \cdots, n$(naive approach) allows fast and accurate clustering, we can see that using EM algorithm for all replicates $i = 1, \cdots, n, j = i, \cdots, J_i$(proposed approach) gives better classification accuracy.

In terms of the number of clusters, we can see that $K = 4$ attains the smallest BIC, which is equal to the true parameter K. Hence, as shown in this example, we can select the number of clusters based on BIC.





Discussion

We only discussed one example of constrained covariance matrix(VEE case). However, it can be extended to a general cases such as EVV or EVE and we can apply the suggested approach to general constrained covariance matrices.

As previously menthioned, model based clustering problem with different number of replicates can be interpreted as the corresponding clustering problem with missing data(i). In Rccp code, we stored a data using this idea and replaced the unobserved with NA in a dataset. However, when the number of replicates has high variablity, such as one or two elements in some groups, while a thousands elements in the other, we need to come up with a useful datatype to store different number of replicates.

Reference

J. L. Andrews, J. R. Wickins, N. M. Boers, and P. D. McNicholas, "teigen: An R package for model-based clustering and classification via the multivariate t distribution," Journal of Statistical Software, vol. 83, no. 1, pp. 1–32, 2018.

J. L. Andrews and P. D. McNicholas, "Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions," Statistics and Computing, vol. 22, no. 5, pp. 1021–1029, Sep 2012.

J. L. Andrews, P. D. McNicholas, and S. Subedi, "Model-based classification via mixtures of multivariate t distributions," Computational Statistics and Data Analysis, vol. 55, no. 1, pp. 520 – 529, 2011

Browne RP, McNicholas PD (2014). "Estimating Common Principal Components in High Dimensions." Advances in Data Analysis and Classification, 8(2), 217–226. doi:10.1007/s11634-013-0139-1.

Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. 28, 781–793 (1995)

Emily Goren, Robust Efficient Model-based Clustering with Partial Records.