MixAll (PMS + log transform (no other variables))

PMS

23 May, 2023

Assumptions of the Alogrithm

The clusterDiagGaussian() model assumes that the data is generated from a mixture of Gaussian distributions. It assumes independence and diagonal covariance thus meaning no correlation between variables. Each component follows a Gaussian distribution with estimated mean and standard deviation parameters. The model represents a mixture of K components, allowing for equal or different standard deviations within each component.

How it works

The MixAll model is basically a mixture model. Mixture models assume data is generated from a combination of probability distributions. Parameter estimation is achieved by maximizing the observed log-likelihood or integrated log-likelihood for data with missing values. Estimation algorithms like EM, SEM, and CEM are used and the default is EM which is highlighted below, involving steps such as imputation, conditional probability calculation, and parameter updates. The EM algorithm iteratively performs these steps until convergence.

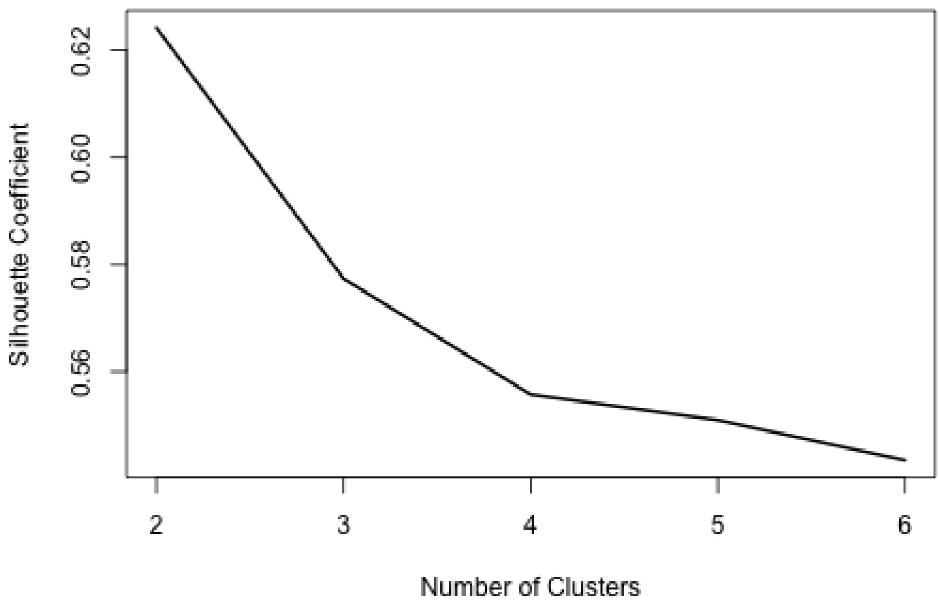
1. **I step:** Impute the missing values x_i^m using the current MAP value provided by the current parameter θ^{m-1}.
2. **E step:** Compute the current conditional probabilities t_{ik}^m for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ using the current parameter θ^{m-1}.
3. **M step:** Update the maximum likelihood estimate θ^m of θ using the conditional probabilities t_{ik}^m as conditional mixing weights, aiming to maximize the log-likelihood function, where $t^m = (t_{ik}^m, i = 1, \ldots, n, k = 1, \ldots, K)$.
4. **Parameter update:** The updated expression of mixture proportions p_k^m for $k = 1, \ldots, K$ are computed. Detailed formulas for updating the parameters λ_k and α depend on the component parameterization.

Note that there are one of two strategies that can be used as a function call: clusterFastStrategy() and clusterSemiSEMStrategy(). When using the clusterFastStrategy(), result is not guaranteed if the model is quite difficult to estimate (overlapping class for examples). If there are lots of missing values its suggested that the fff is used as it uses a MonteCarlo estimator to estimate unbiased estimators. In our case the fast strategy was used as the other would take way too long and we dont have the computing power especially for all 10 measures and trying numerous different number of clusters…

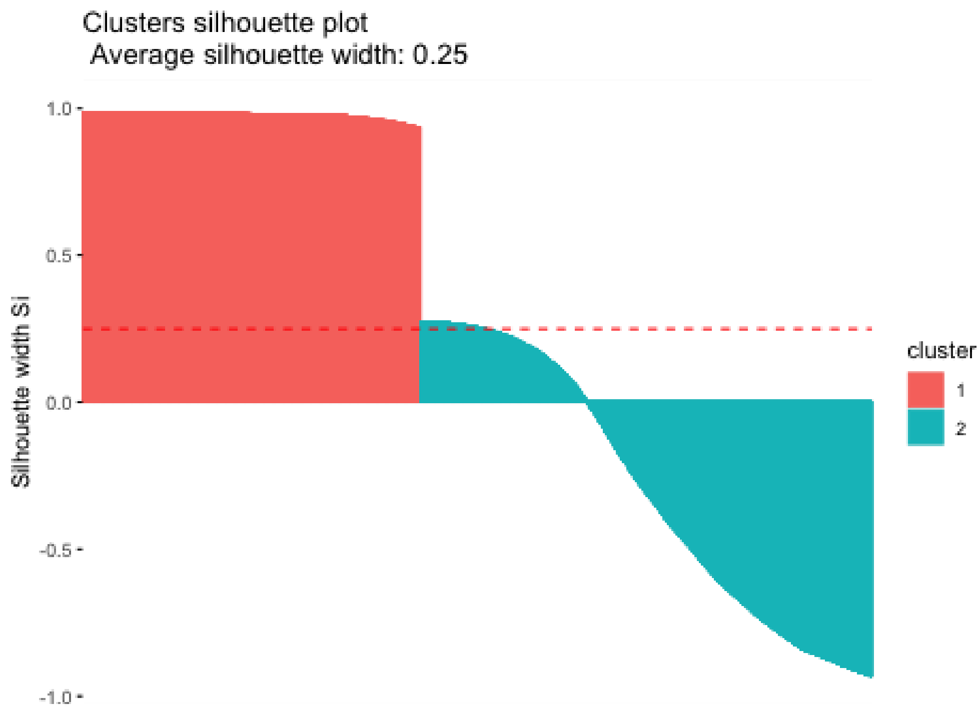
More information can be found here

Amenities

Employment




```
## [1] "Maximum silhouete coefficient: 0.62416982312449 For 2 clusters."
##   cluster size ave.sil.width
## 1       1 5410          0.98
## 2       2 7243         -0.29
```




```
## [1] "Cluster profiles:"
## [1] "Num of DBs:"
##  Cluster 1 Cluster 2
##       6309      8381
##
##
##
##  DB Population:
##  Cluster 1 Cluster 2
##       71.9        73
##
##
##
##  CSD Population:
##  Cluster 1 Cluster 2
##   241592.9    228831
##
##
##
##  CMA Type:
##   Cluster 1 Cluster 2
##        2750      3688
## B      2618      3497
## D       721       897
## K       220       299
##
##
##
##  Index of Remoteness:
```

```
##  Cluster 1 Cluster 2
##      0.227     0.229
##
##
##
##  Provinces:
##                      Cluster 1 Cluster 2
## Alberta                    194       264
## BritishColumbia            312       411
## NewBrunswick                48        74
## NorthwestTerritories         3         4
## NovaScotia                 204       246
## Ontario                    971      1269
## Quebec                     369       489
## Saskatchewan                38        33
## NA's                      4170      5591
##
##
##
##  Amenity dense:
##   Cluster 1 Cluster 2
## 0      5697      7597
## 1       468       616
## 2        69        78
## F        75        90
##
##
##
##  PMS_prox_idx_emp :
##  Cluster 1 Cluster 2
##    0.02555   0.02518
##
##
##
##  PMS_prox_idx_pharma :
##  Cluster 1 Cluster 2
##     0.0458    0.0443
##
##
##
##  PMS_prox_idx_childcare :
##  Cluster 1 Cluster 2
##    0.07592   0.07695
##
##
##
##  PMS_prox_idx_health :
##  Cluster 1 Cluster 2
##    0.01391   0.01368
##
##
##
##  PMS_prox_idx_grocery :
##  Cluster 1 Cluster 2
```

```
##    0.07203   0.07004
##
##
##
##  PMS_prox_idx_educpri :
##  Cluster 1 Cluster 2
##    0.11806   0.11607
##
##
##
##  PMS_prox_idx_educsec :
##  Cluster 1 Cluster 2
##    0.10284   0.10397
##
##
##
##  PMS_prox_idx_lib :
##  Cluster 1 Cluster 2
##    0.11602   0.11146
##
##
##
##  PMS_prox_idx_parks :
##  Cluster 1 Cluster 2
##    0.07016   0.06712
##
##
##
##  PMS_prox_idx_transit :
##  Cluster 1 Cluster 2
##     0.0178   0.01817
```

---

text