CRAN Task View: Cluster Analysis & Finite Mixture Models

Maintainer: Friedrich Leisch and Bettina Gruen
Contact: Bettina.Gruen at jku.at
Version: 2019-08-28
URL: https://CRAN.R-project.org/view=Cluster

This CRAN Task View contains a list of packages that can be used for finding groups in data and modeling unobserved cross-sectional heterogeneity. Many packages provide functionality for more than one of the topics listed below, the section headings are mainly meant as quick starting points rather than an ultimate categorization. Except for packages stats and cluster (which ship with base R and hence are part of every R installation), each package is listed only once.

Most of the packages listed in this CRAN Task View, but not all are distributed under the GPL. Please have a look at the DESCRIPTION file of each package to check under which license it is distributed.

Hierarchical Clustering:

- Functions `hclust()` from package stats and `agnes()` from cluster are the primary functions for agglomerative hierarchical clustering, function `diana()` can be used for divisive hierarchical clustering. Faster alternatives to `hclust()` are provided by the packages fastcluster and flashClust.
- Function `dendrogram()` from stats and associated methods can be used for improved visualization for cluster dendrograms.
- The dendextend package provides functions for easy visualization (coloring labels and branches, etc.), manipulation (rotating, pruning, etc.) and comparison of dendrograms (tangelgrams with heuristics for optimal branch rotations, and tree correlation measures with bootstrap and permutation tests for significance).
- Package dynamicTreeCut contains methods for detection of clusters in hierarchical clustering dendrograms.
- Package genie implements a fast hierarchical clustering algorithm with a linkage criterion which is a variant of the single linkage method combining it with the Gini inequality measure to robustify the linkage method while retaining computational efficiency to allow for the use of larger data sets.
- hybridHclust implements hybrid hierarchical clustering via mutual clusters.
- Package idendr0 allows to interactively explore hierarchical clustering dendrograms and the clustered data. The data can be visualized (and interacted with) in a built-in heat map, but also in GGobi dynamic interactive graphics (provided by rggobi), or base R plots.
- Package isopam uses an algorithm which is based on the classification of ordination scores from isometric feature mapping. The classification is performed either as a hierarchical, divisive method or as non-hierarchical partitioning.
- The package protoclust implements a form of hierarchical clustering that associates a prototypical element with each interior node of the dendrogram. Using the package's `plot()` function, one can produce dendrograms that are prototype-labeled and are therefore easier to interpret.
- pvclust is a package for assessing the uncertainty in hierarchical cluster analysis. It provides approximately unbiased p-values as well as bootstrap p-values.

Partitioning Clustering:

- Function `kmeans()` from package stats provides several algorithms for computing partitions with respect to Euclidean distance.
- Function `pam()` from package cluster implements partitioning around medoids and can work with arbitrary distances. Function `clara()` is a wrapper to `pam()` for larger data sets. Silhouette plots and

spanning ellipses can be used for visualization.
- Package apcluster implements Frey's and Dueck's Affinity Propagation clustering. The algorithms in the package are analogous to the Matlab code published by Frey and Dueck.
- Package ClusterR implements k-means, mini-batch-kmeans, k-medoids, affinity propagation clustering and Gaussian mixture models with the option to plot, validate, predict (new data) and estimate the optimal number of clusters. The package takes advantage of RcppArmadillo to speed up the computationally intensive parts of the functions.
- Package clusterSim allows to search for the optimal clustering procedure for a given dataset.
- Package clustMixType implements Huang's k-prototypes extension of k-means for mixed type data.
- Package evclust implements various clustering algorithms that produce a credal partition, i.e., a set of Dempster-Shafer mass functions representing the membership of objects to clusters.
- Package flexclust provides k-centroid cluster algorithms for arbitrary distance measures, hard competitive learning, neural gas and QT clustering. Neighborhood graphs and image plots of partitions are available for visualization. Some of this functionality is also provided by package cclust.
- Package kernlab provides a weighted kernel version of the k-means algorithm by `kkmeans` and spectral clustering by `specc`.
- Package kml provides k-means clustering specifically for longitudinal (joint) data.
- Package skmeans allows spherical k-Means Clustering, i.e. k-means clustering with cosine similarity. It features several methods, including a genetic and a simple fixed-point algorithm and an interface to the CLUTO vcluster program for clustering high-dimensional datasets.
- Package Spectrum implements a self-tuning spectral clustering method for single or multi-view data and uses either the eigengap or multimodality gap heuristics to determine the number of clusters. The method is sufficiently flexible to cluster a wide range of Gaussian and non-Gaussian structures with automatic selection of K.
- Package trimcluster provides trimmed k-means clustering. Package tclust also allows for trimmed k-means clustering. In addition using this package other covariance structures can also be specified for the clusters.

Model-Based Clustering:

- ML estimation:
  - For semi- or partially supervised problems, where for a part of the observations labels are given with certainty or with some probability, package bgmm provides belief-based and soft-label mixture modeling for mixtures of Gaussians with the EM algorithm.
  - EMCluster provides EM algorithms and several efficient initialization methods for model-based clustering of finite mixture Gaussian distribution with unstructured dispersion in unsupervised as well as semi-supervised learning situation.
  - Packages funHDDC and funFEM implement model-based functional data analysis. The funFEM package implements the funFEM algorithm which allows to cluster time series or, more generally, functional data. It is based on a discriminative functional mixture model which allows the clustering of the data in a unique and discriminative functional subspace. This model presents the advantage to be parsimonious and can therefore handle long time series. The funHDDC package implements the funHDDC algorithm which allows the clustering of functional data within group-specific functional subspaces. The funHDDC algorithm is based on a functional mixture model which models and clusters the data into group-specific functional subspaces. The approach allows afterward meaningful interpretations by looking at the group-specific functional curves.
  - Package GLDEX fits mixtures of generalized lambda distributions and for grouped conditional data package mixdist can be used.
  - Package GMCM fits Gaussian mixture copula models for unsupervised clustering and meta-analysis.

  - Package HDclassif provides function `hddc` to fit Gaussian mixture model to high-dimensional data where it is assumed that the data lives in a lower dimension than the original space.
  - Package teigen allows to fit multivariate t-distribution mixture models (with eigen-decomposed covariance structure) from a clustering or classification point of view.
  - Package mclust fits mixtures of Gaussians using the EM algorithm. It allows fine control of volume and shape of covariance matrices and agglomerative hierarchical clustering based on maximum likelihood. It provides comprehensive strategies using hierarchical clustering, EM and the Bayesian Information Criterion (BIC) for clustering, density estimation, and discriminant analysis. Package Rmixmod provides tools for fitting mixture models of multivariate Gaussian or multinomial components to a given data set with either a clustering, a density estimation or a discriminant analysis point of view. Package mclust as well as packages mixture and Rmixmod provide all 14 possible variance-covariance structures based on the eigenvalue decomposition.
  - Package MetabolAnalyze fits mixtures of probabilistic principal component analysis with the EM algorithm.
  - For grouped conditional data package mixdist can be used.
  - Package MixAll provides EM estimation of diagonal Gaussian, gamma, Poisson and categorical mixtures combined based on the conditional independence assumption using different EM variants and allowing for missing observations. The package accesses the clustering part of the Statistical ToolKit STK++ .
  - mixtools provides fitting with the EM algorithm for parametric and non-parametric (multivariate) mixtures. Parametric mixtures include mixtures of multinomials, multivariate normals, normals with repeated measures, Poisson regressions and Gaussian regressions (with random effects). Non-parametric mixtures include the univariate semi-parametric case where symmetry is imposed for identifiability and multivariate non-parametric mixtures with conditional independent assumption. In addition fitting mixtures of Gaussian regressions with the Metropolis-Hastings algorithm is available.
  - Fitting finite mixtures of uni- and multivariate scale mixtures of skew-normal distributions with the EM algorithm is provided by package mixsmsn.
  - Package MoEClust fits parsimonious finite multivariate Gaussian mixtures of experts models via the EM algorithm. Covariates may influence the mixing proportions and/or component densities and all 14 constrained covariance parameterizations from package mclust are implemented.
  - Package movMF fits finite mixtures of von Mises-Fisher distributions with the EM algorithm.
  - mritc provides tools for classification using normal mixture models and (higher resolution) hidden Markov normal mixture models fitted by various methods.
  - prabclus clusters a presence-absence matrix object by calculating an MDS from the distances, and applying maximum likelihood Gaussian mixtures clustering to the MDS points.
  - Package psychomix estimates mixtures of the dichotomous Rasch model (via conditional ML) and the Bradley-Terry model. Package mixRasch estimates mixture Rasch models, including the dichotomous Rasch model, the rating scale model, and the partial credit model with joint maximum likelihood estimation.
  - Package pmclust allows to use unsupervised model-based clustering for high dimensional (ultra) large data. The package uses pbdMPI to perform a parallel version of the EM algorithm for mixtures of Gaussians.
  - Package rebmix implements the REBMIX algorithm to fit mixtures of conditionally independent normal, lognormal, Weibull, gamma, binomial, Poisson, Dirac or von Mises component densities as well as mixtures of multivariate normal component densities with unrestricted variance-covariance matrices.
- Bayesian estimation:
  - Bayesian estimation of finite mixtures of multivariate Gaussians is possible using package bayesm. The package provides functionality for sampling from such a mixture as well as estimating the model using Gibbs sampling. Additional functionality for analyzing the MCMC

chains is available for averaging the moments over MCMC draws, for determining the marginal densities, for clustering observations and for plotting the uni- and bivariate marginal densities.
    - Package bayesmix provides Bayesian estimation using JAGS.
    - Package bclust allows Bayesian clustering using a spike-and-slab hierarchical model and is suitable for clustering high-dimensional data.
    - Package Bmix provides Bayesian Sampling for stick-breaking mixtures.
    - Package bmixture provides Bayesian estimation of finite mixtures of univariate Gamma and normal distributions.
    - Package dpmixsim fits Dirichlet process mixture models using conjugate models with normal structure. Package profdpm determines the maximum posterior estimate for product partition models where the Dirichlet process mixture is a specific case in the class.
    - Package GSM fits mixtures of gamma distributions.
    - Package IMIFA fits Infinite Mixtures of Infinite Factor Analyzers and a flexible suite of related models for clustering high-dimensional data. The number of clusters and/or number of cluster-specific latent factors can be non-parametrically inferred, without recourse to model selection criteria.
    - Package mcclust implements methods for processing a sample of (hard) clusterings, e.g. the MCMC output of a Bayesian clustering model. Among them are methods that find a single best clustering to represent the sample, which are based on the posterior similarity matrix or a relabeling algorithm.
    - Package mixAK contains a mixture of statistical methods including the MCMC methods to analyze normal mixtures with possibly censored data.
    - Package NPflow fits Dirichlet process mixtures of multivariate normal, skew normal or skew t-distributions. The package was developed oriented towards flow-cytometry data preprocessing applications.
    - Package PReMiuM is a package for profile regression, which is a Dirichlet process Bayesian clustering where the response is linked non-parametrically to the covariate profile.
    - Package rjags provides an interface to the JAGS MCMC library which includes a module for mixture modelling.
- Other estimation methods:
    - Package AdMit allows to fit an adaptive mixture of Student-t distributions to approximate a target density through its kernel function.
    - Package CEC uses cross-entropy clustering to automatically remove unnecessary clusters, while at the same time allowing the simultaneous use of various types of Gaussian mixture models.
    - Circular and orthogonal regression clustering using redescending M-estimators is provided by package edci.

Other Cluster Algorithms:

- Package ADPclust allows to cluster high dimensional data based on a two dimensional decision plot. This density-distance plot plots for each data point the local density against the shortest distance to all observations with a higher local density value. The cluster centroids of this non-iterative procedure can be selected using an interactive or automatic selection mode.
- Package amap provides alternative implementations of k-means and agglomerative hierarchical clustering.
- Package biclust provides several algorithms to find biclusters in two-dimensional data.
- Package cba implements clustering techniques for business analytics like "rock" and "proximus".
- Package CHsharp clusters 3-dimensional data into their local modes based on a convergent form of Choi and Hall's (1999) data sharpening method.
- Package clue implements ensemble methods for both hierarchical and partitioning cluster methods.
- Package CoClust implements a cluster algorithm that is based on copula functions and therefore allows

to group observations according to the multivariate dependence structure of the generating process without any assumptions on the margins.
- Fuzzy clustering and bagged clustering are available in package e1071. Further and more extensive tools for fuzzy clustering are available in package fclust.
- Package compHclust provides complimentary hierarchical clustering which was especially designed for microarray data to uncover structures present in the data that arise from 'weak' genes.
- Package dbscan provides a fast reimplementation of the DBSCAN (density-based spatial clustering of applications with noise) algorithm using a kd-tree.
- Package FactoClass performs a combination of factorial methods and cluster analysis.
- The hopach algorithm is a hybrid between hierarchical methods and PAM and builds a tree by recursively partitioning a data set.
- For graphs and networks model-based clustering approaches are implemented in latentnet.
- Package optpart contains a set of algorithms for creating partitions and coverings of objects largely based on operations on similarity relations (or matrices).
- Package pdfCluster provides tools to perform cluster analysis via kernel density estimation. Clusters are associated to the maximally connected components with estimated density above a threshold. In addition a tree structure associated with the connected components is obtained.
- Package prcr implements the 2-step cluster analysis where first hierarchical clustering is performed to determine the initial partition for the subsequent k-means clustering procedure.
- Package randomLCA provides the fitting of latent class models which optionally also include a random effect. Package poLCA allows for polytomous variable latent class analysis and regression. BayesLCA allows to fit Bayesian LCA models employing the EM algorithm, Gibbs sampling or variational Bayes methods.
- Package RPMM fits recursively partitioned mixture models for Beta and Gaussian Mixtures. This is a model-based clustering algorithm that returns a hierarchy of classes, similar to hierarchical clustering, but also similar to finite mixture models.
- Self-organizing maps are available in package som. Package somspace uses self-organizing maps and complex networks to classify time series in space.
- Several packages provide cluster algorithms which have been developed for bioinformatics applications. These packages include FunCluster for profiling microarray expression data and ORIClust for order-restricted information-based clustering.

Cluster-wise Regression:

- Multigroup mixtures of latent Markov models on mixed categorical and continuous data (including time series) can be fitted using depmix or depmixS4. The parameters are optimized using a general purpose optimization routine given linear and nonlinear constraints on the parameters.
- Package flexmix implements an user-extensible framework for EM-estimation of mixtures of regression models, including mixtures of (generalized) linear models.
- Package fpc provides fixed-point methods both for model-based clustering and linear regression. A collection of asymmetric projection methods can be used to plot various aspects of a clustering.
- Package lcmm fits a latent class linear mixed model which is also known as growth mixture model or heterogeneous linear mixed model using a maximum likelihood method.
- Package mixreg fits mixtures of one-variable regressions and provides the bootstrap test for the number of components.
- mixPHM fits mixtures of proportional hazard models with the EM algorithm.
- Package gamlss.mx fits finite mixtures of gamlss family distributions.

Additional Functionality:

- Mixtures of univariate normal distributions can be printed and plotted using package nor1mix.
- Package clusterfly allows to visualize the results of clustering algorithms.

- Package clusterGeneration contains functions for generating random clusters and random covariance/correlation matrices, calculating a separation index (data and population version) for pairs of clusters or cluster distributions, and 1-D and 2-D projection plots to visualize clusters. Alternatively MixSim generates a finite mixture model with Gaussian components for prespecified levels of maximum and/or average overlaps. This model can be used to simulate data for studying the performance of cluster algorithms.
- Package clusterCrit computes various clustering validation or quality criteria and partition comparison indices.
- For cluster validation package clusterRepro tests the reproducibility of a cluster. Package clv contains popular internal and external cluster validation methods ready to use for most of the outputs produced by functions from package cluster and clValid calculates several stability measures.
- Package clustvarsel provides variable selection for Gaussian model-based clustering. Variable selection for latent class analysis for clustering multivariate categorical data is implemented in package LCAvarsel. Package VarSelLCM provides variable selection for model-based clustering of continuous, count, categorical or mixed-type data with missing values where the models used impose a conditional independence assumption given group membership.
- Functionality to compare the similarity between two cluster solutions is provided by `cluster.stats()` in package fpc.
- The stability of k-centroid clustering solutions fitted using functions from package flexclust can also be validated via `bootFlexclust()` using bootstrap methods.
- Package MOCCA provides methods to analyze cluster alternatives based on multi-objective optimization of cluster validation indices.
- Package NbClust implements 30 different indices which evaluate the cluster structure and should help to determine on a suitable number of clusters.
- Package seriation provides `dissplot()` for visualizing dissimilarity matrices using seriation and matrix shading. This also allows to inspect cluster quality by restricting objects belonging to the same cluster to be displayed in consecutive order.
- Package sigclust provides a statistical method for testing the significance of clustering results.
- Package treeClust calculates dissimilarities between data points based on their leaf memberships in regression or classification trees for each variable. It also performs the cluster analysis using the resulting dissimilarity matrix with available heuristic clustering algorithms in R.

CRAN packages:

- AdMit
- ADPclust
- amap
- apcluster
- BayesLCA
- bayesm
- bayesmix
- bclust
- bgmm
- biclust
- Bmix
- bmixture
- cba
- cclust
- CEC
- CHsharp
- clue
- cluster (core)

- clusterCrit
- clusterfly
- clusterGeneration
- ClusterR
- clusterRepro
- clusterSim
- clustMixType
- clustvarsel
- clv
- clValid
- CoClust
- compHclust
- dbscan
- dendextend
- depmix
- depmixS4
- dpmixsim
- dynamicTreeCut
- e1071
- edci
- EMCluster
- evclust
- FactoClass
- fastcluster
- fclust
- flashClust
- flexclust (core)
- flexmix (core)
- fpc
- FunCluster
- funFEM
- funHDDC
- gamlss.mx
- genie
- GLDEX
- GMCM
- GSM
- HDclassif
- hybridHclust
- idendr0
- IMIFA
- isopam
- kernlab
- kml
- latentnet
- LCAvarsel
- lcmm
- mcclust
- mclust (core)
- MetabolAnalyze
- mixAK

- MixAll
- mixdist
- mixPHM
- mixRasch
- mixreg
- MixSim
- mixsmsn
- mixtools
- mixture
- MOCCA
- MoEClust
- movMF
- mritc
- NbClust
- nor1mix
- NPflow
- optpart
- ORIClust
- pdfCluster
- pmclust
- poLCA
- prabclus
- prcr
- PReMiuM
- profdpm
- protoclust
- psychomix
- pvclust
- randomLCA
- rebmix
- rjags
- Rmixmod (core)
- RPMM
- seriation
- sigclust
- skmeans
- som
- somspace
- Spectrum
- tclust
- teigen
- treeClust
- trimcluster
- VarSelLCM

Related links:

- CRAN Task View: MachineLearning
- Bioconductor Package: hopach