




# pivmet: an R package proposing pivotal methods for consensus clustering and mixture modelling

Leonardo Egidi <sup>1\*</sup>, Roberta Pappada <sup>1\*</sup>, Francesco Pauli <sup>1</sup>, and Nicola Torelli <sup>1</sup>

<sup>1</sup> Department of Economics, Business, Mathematics, and Statistics 'Bruno de Finetti', University of Trieste ¶ Corresponding author \* These authors contributed equally.

DOI: [10.21105/joss.06461](https://doi.org/10.21105/joss.06461)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Sehrish Kanwal](#)  

## Reviewers:

- [@adriancorrendo](#)
- [@larryshamalama](#)

Submitted: 06 February 2024

Published: 11 June 2024

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

## Summary

We introduce the R package `pivmet`, a software that performs different pivotal methods for identifying, extracting, and using the so-called pivotal units that are chosen from a partition of data points to represent the groups to which they belong. Such units turn out to be very useful in both unsupervised and supervised learning frameworks such as clustering, classification and mixture modelling.

More specifically, applications of pivotal methods include, among the others: a Markov-Chain Monte Carlo (MCMC) relabelling procedure to deal with the well-known label-switching problem occurring during Bayesian estimation of mixture models (Egidi et al., 2018; Frühwirth-Schnatter, 2001; Richardson & Green, 1997; Stephens, 2000); model-based clustering through sparse finite mixture models (SFMM) (Frühwirth-Schnatter & Malsiner-Walli, 2019; Malsiner-Walli et al., 2016); consensus clustering (Strehl & Ghosh, 2002), which may allow to improve classical clustering techniques—e.g. the classical  $k$ -means—via a careful seeding; and Dirichlet process mixture models (DPMM) in Bayesian nonparametrics (Escobar & West, 1995; Ferguson, 1973; Neal, 2000).

## Installation

The stable version of the package can be installed from the [Comprehensive R Archive Network \(CRAN\)](https://cran.r-project.org/):

```
install.packages("pivmet")  
library(pivmet)
```

However, before installing the package, the user should make sure to download the JAGS program at <https://sourceforge.net/projects/mcmc-jags/>.

## Statement of need

In the modern *big-data* and *machine learning* age, summarizing some essential information from a dataset is often relevant and can help simplifying the data pre-processing steps. The advantage of identifying representative units of a group—hereafter *pivotal units* or *pivots*—chosen in such a way that they are as far as possible from units in the other groups and/or as similar as possible to the units in the same group, is that they may convey relevant information about the group they belong to while saving wasteful operations.

Despite the lack of a strict theoretical framework behind their characterization, the pivots may be beneficial in many machine learning frameworks, such as clustering, classification, and mixture modelling when the interest is in deriving reliable estimates in mixture models

and/or finding a partition of the data points. The theoretical framework concerning the pivotal methods implemented in the pivmet package is provided in (Egidi et al., 2018).

The pivmet package for R is available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/package=pivmet> (Egidi et al., 2023) and implements various pivotal selection criteria to deal with, but not limited to: (i) mixture model Bayesian estimation—either via the JAGS software (Plummer, 2022) using Gibbs sampling or the Stan (Stan Development Team, 2022) software performing Hamiltonian Monte Carlo (HMC)—to tackle the so-called *label switching* problem; (ii) consensus clustering, where a variant of the  $k$ -means algorithm is available; (iii) Dirichlet Process Mixture Models (DPPM).

As far as we know from reviewing the scientific and statistical literature, the pivmet package is the only software designed to search pivotal units in a modern machine learning framework. However, since its large applicability, it exhibits some deep connections with some existing R packages commonly used for Bayesian statistics and clustering. In particular, the pivmet package:

- extends the bayesmix package (Gruen, 2015), which allows to fit univariate Gaussian mixtures, by allowing for sparse Gaussian univariate/multivariate mixtures;
- is naturally connected with the label.switching package (Papastamoulis, 2016) that offers many methods to fix label switching in Bayesian mixture models;
- in terms of computational MCMC methods, depends on the rstan package (Stan Development Team, 2022) to perform Hamiltonian Monte Carlo sampling and on the rjags package (Plummer, 2022) to perform Gibbs sampling;
- extends the classical kmeans function by allowing for a robust initial seeding.

Compared to the aforementioned packages, the pivmet package offers a novel way to retrieve pivotal units. Moreover, it contains functions to exploit the pivotal units to efficiently estimate univariate and multivariate Gaussian mixtures, by relying on pre-compiled JAGS/Stan models, and to perform a robustified version of the  $k$ -means algorithm.

## Overview and main functions

The package architecture strongly relies on three main functions:

- The function `piv_MCMC()` is used to fit a Bayesian Gaussian mixture model with underlying Gibbs sampling or Hamiltonian Monte Carlo algorithm. The user can specify distinct prior distributions with the argument `priors` and the selected pivotal criterion via the argument `piv.criterion`.
- The function `piv_rel()` takes in input the model fit returned by `piv_MCMC` and implements the relabelling step as outlined by (Egidi et al., 2018).
- The function `piv_KMeans()` performs a robust consensus clustering based on distinct  $k$ -means partitions. The user can specify some options, such as the number of consensus partitions.

## Example 1: relabelling for dealing with label switching

The Fishery dataset in the bayesmix (Gruen, 2015) package has been previously used by Titterton et al. (1985) and Papastamoulis (2016). It consists of 256 snapper length measurements—see left plot of Figure 1 for the data histogram, along with an estimated kernel density. Analogously to some previous works, we assume a Gaussian mixture model with  $k = 5$  groups, where  $\mu_j$ ,  $\sigma_j$  and  $\eta_j$  are respectively the mean, the standard deviation and the weight of group  $j = 1, \dots, k$ . We fit our model by simulating 15000 samples from

the posterior distribution of  $(z, \mu, \sigma, \eta)$ , by selecting the default argument `software="rjags"`; for univariate mixtures, the MCMC Gibbs sampling is returned by the function `JAGSrun` in the package `bayesmix`. Alternatively, one could fit the model according to HMC sampling and with underlying Stan ecosystem by typing `software="rstan"`. By default, the burn-in period is set equal to half of the total number of MCMC iterations. Here below we include the relevant R code.

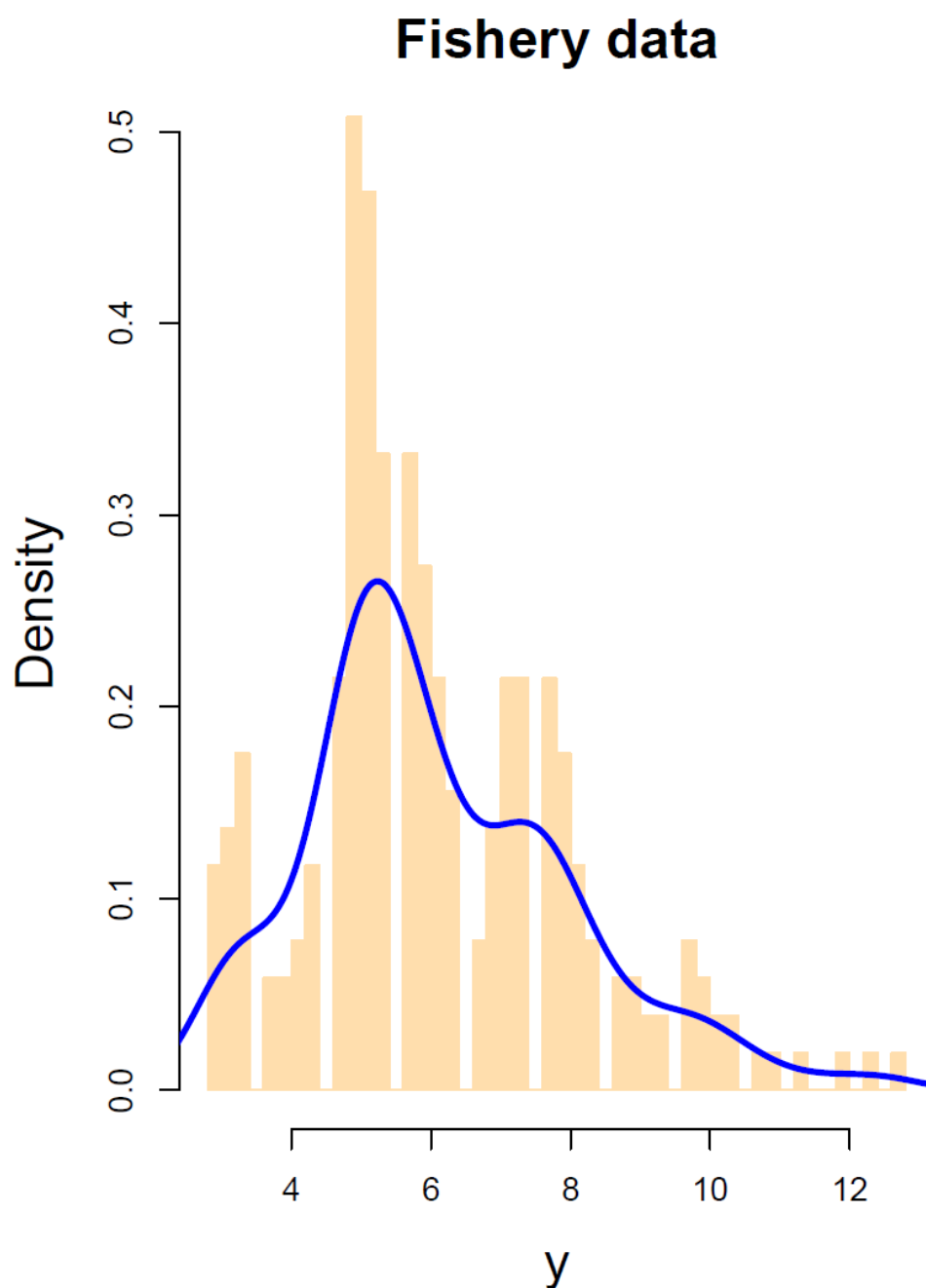
```
# required packages
library(bayesmix)
set.seed(100)

# load data
data(fish)
y <- fish[,1]
k <- 5
nMC <- 15000

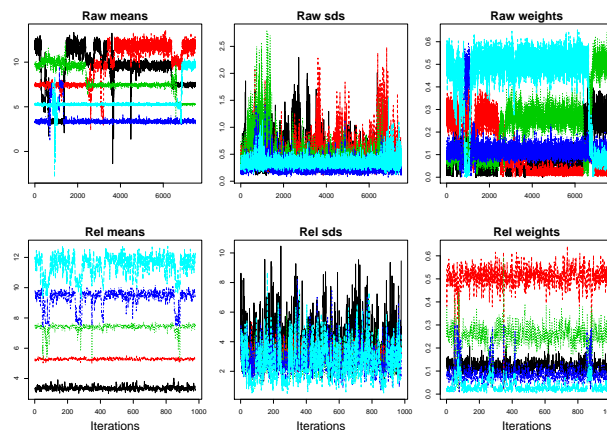
# fit the mixture model for univariate data and find the pivots
res <- piv_MCMC(y = y, k = k, nMC = nMC, burn = 7500, software = "rjags")

# relabel the chains: figure
rel <- piv_rel(mcmc=res)
piv_plot(y = y, mcmc = res, rel_est = rel, type="chains")

# use Stan
res_stan <- piv_MCMC(y = y, k = k, nMC = 5000, burn = 2500, software = "rstan")
cat(res_stan$model)
```



**Figure 1:** Histograms of the Fishery data. The blue line represents the estimated kernel density.



**Figure 2:** Fishery dataset: traceplots of the parameters  $(\mu, \sigma, \eta)$  obtained via the `rjags` option for the `piv_MCMC` function (Gibbs sampling, 15000 MCMC iterations). Top row: Raw MCMC outputs. Bottom row: relabelled MCMC samples.

**Figure 2** displays the traceplots for the parameters  $(\mu, \sigma, \eta)$ . From the first row showing the raw MCMC outputs as given by the Gibbs sampling, we note that label switching clearly occurred. Our algorithm is able to fix label-switching and reorder the means  $\mu_j$  and the weights  $\eta_j$ , for  $j = 1, \dots, k$ , as emerged from the second row of the plot.

## Example 2: consensus clustering

As widely known, one of the drawbacks of the  $k$ -means algorithm is represented by its inefficiency in distinguishing between groups of unbalanced sizes. The recent literature on clustering methods has explored some approaches to combine several partitions via a consensus clustering, which may improve the solution obtained from a single run of a clustering algorithm. Here, we consider a consensus clustering technique based on  $k$ -means and pivotal methods used for a careful initial pivotal seeding.

For illustration purposes, we simulate three bivariate Gaussian distributions with 20, 100 and 500 observations, respectively—see **Figure 3**. The plots refer to the pivotal criteria MUS, maxsumint, and maxsumdiff; moreover, we consider Partitioning Around Medoids (PAM) method via the `pam` function of the `cluster` package and agglomerative hierarchical clustering (`agnes`), with average, single, and complete linkage. Group centers and pivots are marked via asterisks and triangles symbols, respectively. As can be seen, pivotal  $k$ -means methods are able to satisfactorily detect the true data partition and outperform the alternative approaches in most of the cases. Here below we include the relevant R code.

```
library(mvtnorm)

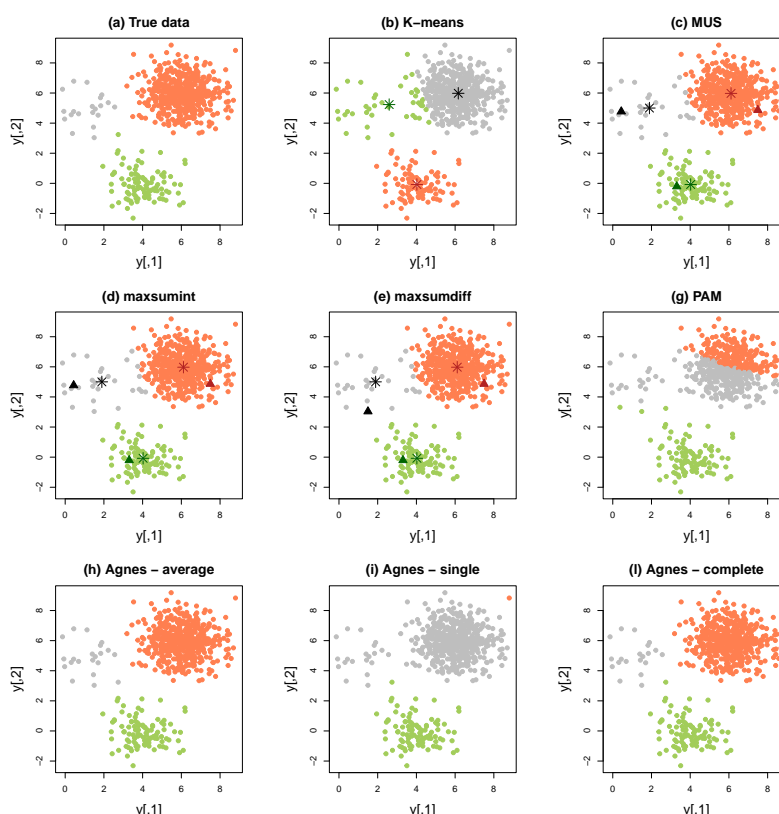
# simulate data
set.seed(123)
n=620
k=3
n1=20
n2=100
n3=500
x=matrix(NA, n,2)
gruppovero=c( rep(1,n1), rep(2, n2), rep(3, n3))

x[1:n1,]=rmvnorm(n1, c(1,5), sigma=diag(2))
```

```
x[(n1+1):(n1+n2),]=rmvnorm(n2, c(4,0), sigma=diag(2))
x[(n1+n2+1):(n1+n2+n3),]=rmvnorm(n3, c(6,6), sigma=diag(2))
```

```
kmeans_res <- kmeans(x, centers=k)
```

```
res <- piv_KMeans(x, k, alg.type = "hclust",
  piv.criterion ="maxsumdiff",
  prec_par=n1)
```



**Figure 3:** Consensus clustering via the `piv_KMeans` function assuming three bivariate Gaussian distributions and three groups with 20, 100 and 500 observations, respectively.

## Conclusion

The `pivmet` package proposes various methods for identifying pivotal units in datasets with a grouping structure and uses them for improving inferential conclusions and clustering partitions. The package suits well for both supervised and unsupervised problems, by providing a valid alternative to existing functions for similar applications, and keeping low the computational effort. It is of future interest to include additional functions that may allow to deal with the estimation of the number of components in the data when this information is latent or unknown and provide more graphical tools to diagnose pivotal selection.

## Reproducibility

The R code required to generate the examples is available at <https://github.com/LeoEgidi/pivmet/tree/master/paper/rcode>.

## Acknowledgements

The authors thank Ioannis Ntzoufras and Dimitris Karlis from Athens University of Economics and Business (AUEB) for their valuable suggestions about the package structure.

## References

- Egidi, L., Pappadà, R., Pauli, F., & Torelli, N. (2018). Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, 28(4), 957–969.
- Egidi, L., Pappadà, R., Pauli, F., & Torelli, N. (2023). *Pivmet: Pivotal methods for bayesian relabelling and k-means clustering*. <https://CRAN.R-project.org/package=pivmet>
- Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), 577–588.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 209–230.
- Frühwirth-Schnatter, S. (2001). Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453), 194–209.
- Frühwirth-Schnatter, S., & Malsiner-Walli, G. (2019). From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13(1), 33–64.
- Gruen, B. (2015). *Bayesmix: Bayesian mixture models with JAGS*. <https://CRAN.R-project.org/package=bayesmix>
- Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1-2), 303–324.
- Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2), 249–265.
- Papastamoulis, P. (2016). label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets*, 69(1), 1–24.
- Plummer, M. (2022). *Rjags: Bayesian graphical models using MCMC*. <https://CRAN.R-project.org/package=rjags>
- Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59(4), 731–792.
- Stan Development Team. (2022). *RStan: The R interface to Stan*. <http://mc-stan.org/>
- Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, 62(4), 795–809.
- Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, 3, 583–617.
- Titterton, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley, New York.