




# 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.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

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

Editor: [Sehrish Kanwal](#)  

## Reviewers:

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

Submitted: 06 February 2024

Published: unpublished

## License

Authors of papers retain copyright, and release the work under a Creative Commons Attribution 4.0 International License ([CC 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 of a dataset that are chosen to represent the groups of data points to which they belong. These algorithms turn out to be very useful in many unsupervised and supervised learning frameworks such as clustering, classification and mixture modelling.

More specifically, applications of pivotal methods could cover, among the others: a Markov-Chain Monte Carlo (MCMC) relabelling procedure to deal with the well-known label-switching problem (Egidi et al., 2018; Frühwirth-Schnatter, 2001; Richardson & Green, 1997; Stephens, 2000) occurring during Bayesian estimation of mixture models; 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) (Escobar & West, 1995; Ferguson, 1973; Neal, 2000) in Bayesian nonparametrics.

## Installation

The stable version of the package can be installed from the [Comprehensive R Archive Network \(CRAN\)](#):

```
{r, eval = FALSE} install.packages("pivmet") library(pivmet)
```

## Statement of need

In the modern *big-data* and *machine learning* age, summarizing some essential information from a data pattern 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*—somehow chosen to be as far as possible from units in the other groups and 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 to derive reliable estimates and/or a better grouping partition.

A deep and theoretical detail around the package's supported pivotal methods is provided in (Egidi et al., 2018).

The `pivmet` package (Egidi et al., 2023) for R, available from the Comprehensive R Archive Network (CRAN) at <http://CRAN.R-project.org/package=pivmet>, implements various pivotal

40 selection criteria to deal with, but not limited to: (i) mixture model Bayesian estimation—either  
 41 via the JAGS software (Plummer, 2022) using Gibbs sampling or the Stan (Stan Development  
 42 Team, 2022) software performing Hamiltonian Monte Carlo (HMC)—to tackle the so-called  
 43 *label switching* problem; (ii) consensus clustering, where a variant of the  $k$ -means algorithm is  
 44 available; (iii) Dirichlet Process Mixture Models (DPPM).

## 45 Overview and main functions

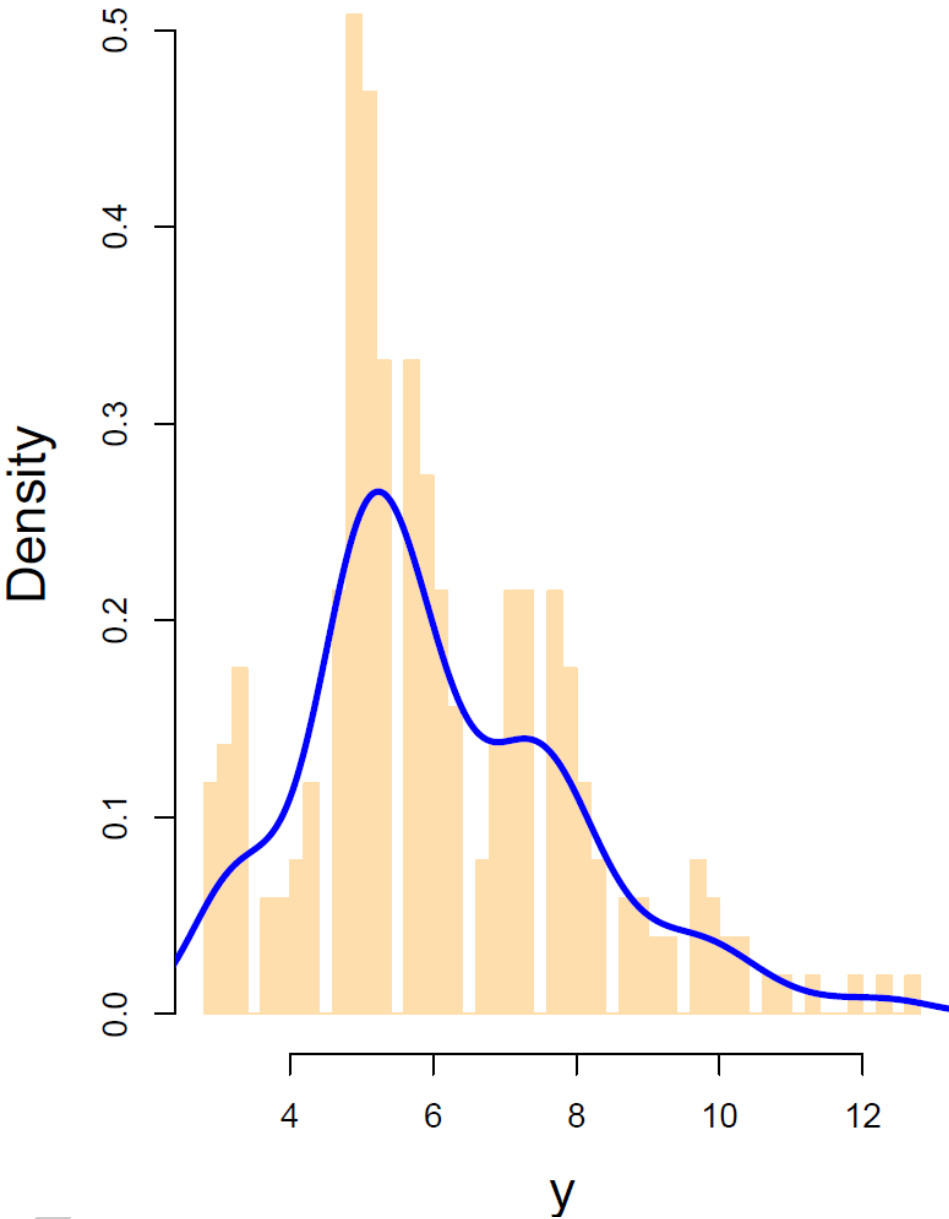
46 The package architecture strongly relies on three main functions:

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

## 56 Example 1: relabelling for label switching

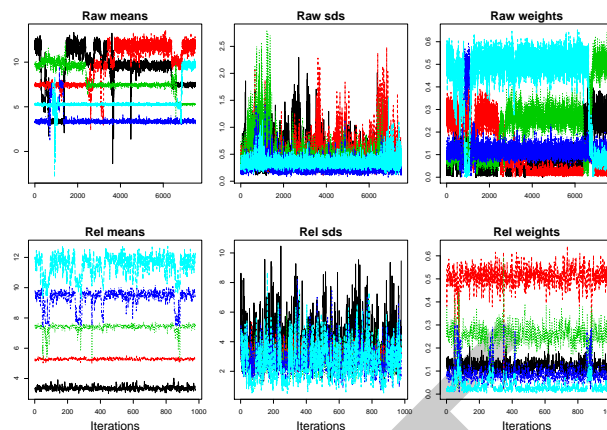
57 The Fishery dataset in the `bayesmix` (Gruen, 2015) package has been previously used by  
 58 Titterton et al. (1985) and Papastamoulis (2016). It consists of 256 snapper length  
 59 measurements—see left plot of Figure ?? for the data histogram, along with an estimated  
 60 kernel density. Analogously to some previous works, we assume a Gaussian mixture model  
 61 with  $k = 5$  groups, where  $\mu_j$ ,  $\sigma_j$  and  $\eta_j$  are the mean, the standard deviation and the  
 62 weight of group  $j$ , respectively. We fit our model by simulating 15000 samples from the  
 63 posterior distribution of  $(z, , , ,)$ , by selecting the default argument `software="rjags"`; for  
 64 univariate mixtures, the MCMC Gibbs sampling is returned by the function `JAGSrun` in the  
 65 package `bayesmix`. Alternatively, one could fit the model according to HMC sampling and  
 66 with underlying Stan ecosystem by typing `software="rstan"`. By default, the burn-in period  
 67 is set equal to half of the total number of MCMC iterations.

Fishery data



<sup>68</sup>  
<sup>69</sup> = 60%}

{width



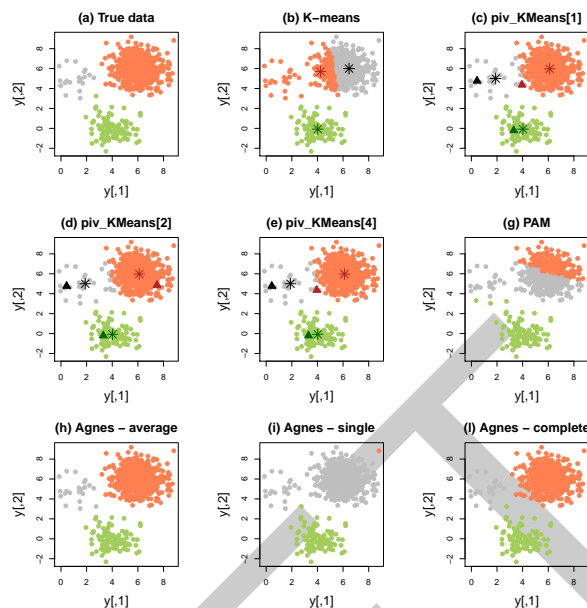
**Figure 1:** Fishery dataset: traceplots of the parameters  $(\mu, \sigma, \eta)$  obtained via the `rjags` option for the `piv_MCMC` function (Gibbs sampling, 15000 MCMC iterations). Top row: Raw MCMC outputs. Bottom row: relabelled MCMC samples.

Figure 1 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. For these reasons, the clustering scientific literature claims that a better robust clustering solution is usually obtained if more partitions are obtained, in such a way the final partition works as a sort of *consensus*. We perform here a consensus clustering technique based on single  $k$ -means configurations, where each of these has been obtained through a careful initial pivotal seeding.

For illustration purposes, we simulate three bivariate Gaussian distributions with 20, 100 and 500 observations, respectively—see Figure 2. The plots with titles ‘piv KMeans’ refer to the pivotal criteria MUS, (i) or `maxsumint`, (ii) or `maxsumdiff`, where the labels 1, 2, and 4 follow the order used in the R function; 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. The partitions from the classical  $k$ -means are obtained using multiple random seeds. Group centers and pivots are marked via asterisks and triangles symbols, respectively. As we may notice, pivotal  $k$ -means methods are able to satisfactorily detect the true data partition.



**Figure 2:** 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 using 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 aspects in the software, such as the estimation of the number of components in the data when this information is latent/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

We want to 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.

- 111 Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of*  
112 *Statistics*, 209–230.
- 113 Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic  
114 switching and mixture models. *Journal of the American Statistical Association*, 96(453),  
115 194–209.
- 116 Frühwirth-Schnatter, S., & Malsiner-Walli, G. (2019). From here to infinity: Sparse finite  
117 versus dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and*  
118 *Classification*, 13(1), 33–64.
- 119 Gruen, B. (2015). *Bayesmix: Bayesian mixture models with JAGS*. [https://CRAN.R-project.](https://CRAN.R-project.org/package=bayesmix)  
120 [org/package=bayesmix](https://CRAN.R-project.org/package=bayesmix)
- 121 Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. (2016). Model-based clustering based  
122 on sparse finite gaussian mixtures. *Statistics and Computing*, 26(1-2), 303–324.
- 123 Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models.  
124 *Journal of Computational and Graphical Statistics*, 9(2), 249–265.
- 125 Papastamoulis, P. (2016). label.switching: An R package for dealing with the label switching  
126 problem in MCMC outputs. *Journal of Statistical Software, Code Snippets*, 69(1), 1–24.
- 127 Plummer, M. (2022). *Rjags: Bayesian graphical models using MCMC*. [https://CRAN.R-project.](https://CRAN.R-project.org/package=rjags)  
128 [org/package=rjags](https://CRAN.R-project.org/package=rjags)
- 129 Richardson, S., & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown  
130 number of components (with discussion). *Journal of the Royal Statistical Society: Series*  
131 *B*, 59(4), 731–792.
- 132 Stan Development Team. (2022). *RStan: The R interface to Stan*. <http://mc-stan.org/>
- 133 Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal*  
134 *Statistical Society: Series B*, 62(4), 795–809.
- 135 Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining  
136 multiple partitions. *Journal on Machine Learning Research*, 3, 583–617.
- 137 Titterton, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture*  
138 *distributions*. Wiley, New York.