

# Model-Based Clustering and Variable Selection for Multivariate Count Data

Julien Jacques<sup>1</sup> Laboratoire ERIC, Université de Lyon  
 Thomas Brendan Murphy School of Mathematics & Statistics, University College Dublin  
 Institut d'Études Avancées, Université de Lyon

Date published: 2024-10-15 Last modified: 2024-10-15

## Abstract

Model-based clustering provides a principled way of developing clustering methods. We develop a new model-based clustering methods for count data. The method combines clustering and variable selection for improved clustering. The method is based on conditionally independent Poisson mixture models and Poisson generalized linear models. The method is demonstrated on simulated data and data from an ultra running race, where the method yields excellent clustering and variable selection performance.

*Keywords:* Count data, Model-based clustering, Variable selection

1	<b>Contents</b>	
2	<b>1 Introduction</b>	<b>2</b>
3	<b>2 Motivating Example</b>	<b>2</b>
4	<b>3 Independent Poisson Mixture</b>	<b>2</b>
5	<b>4 Variable selection</b>	<b>4</b>
6	4.1 Model setup . . . . .	5
7	4.2 Interpretation . . . . .	6
8	4.3 Stepwise selection algorithm . . . . .	6
9	4.3.1 Screening variables: Initialization . . . . .	6
10	4.3.2 Stepwise algorithm: Updating . . . . .	7
11	<b>5 Simulation study</b>	<b>7</b>
12	5.1 Illustrative example . . . . .	7
13	5.2 Scenario 1 . . . . .	9
14	5.3 Scenario 2 . . . . .	10
15	5.4 Scenario 3 . . . . .	12
16	<b>6 International Ultrarunning Association Data</b>	<b>14</b>
17	<b>7 Discussion</b>	<b>17</b>
18	<b>8 Acknowledgements</b>	<b>18</b>

<sup>1</sup>Corresponding author: [julien.jacques@univ-lyon2.fr](mailto:julien.jacques@univ-lyon2.fr)

# 1 Introduction

Multivariate count data is ubiquitous in statistical applications, as ecology (Chiquet, Mariadassou, and Robin 2021), genomic (Rau et al. 2015; Silva et al. 2019). These data arise when each observation consists of a vector of count values. Count data are often treated as continuous data and therefore modeled by a Gaussian distribution, this assumption is particularly poor when the measured counts are low. Instead, we use the reference distribution for count data which is the Poisson distribution (Agresti 2013; Inouye et al. 2017a).

When a data set is heterogeneous, clustering allows to extract homogeneous subsets from the whole data set. Many clustering methods, such as  $k$ -means (Hartigan and Wong 1979), are geometric in nature, whereas many modern clustering approaches are based on probabilistic models. In this work, we use model-based clustering which has been developed for many types of data (Bouveyron et al. 2019; McLachlan and Peel 2000; Frühwirth-Schnatter, Celeux, and Robert 2018).

Modern data are often high-dimensional, that is the number of variables is often large. Among these variables, some are useful for the task of interest, some are useless for the task of interest and some others are useful but redundant. There is a need to select only the relevant variables, and that whatever is the task. Variable selection methods are widespread for supervised learning tasks, in particular to avoid overfitting. However, variable selection methods are less well developed for unsupervised learning tasks, such as clustering. Recently, several methods have been proposed for selecting the relevant variables in model-based clustering; we refer to Fop and Murphy (2018) and McParland and Murphy (2018) for recent detailed surveys.

The goal of the present work is to provide a clustering and variable selection method for multivariate count data, which, to the best of our knowledge, has not yet been studied in depth.

## 2 Motivating Example

The International Association of Ultrarunners (IAU) 24 hour World Championships were held in Katowice, Poland from September 8th to 9th, 2012. Two hundred and sixty athletes representing twenty four countries entered the race, which was held on a course consisting of a 1.554 km looped route. An update of the number of laps covered by each athlete was recorded approximately every hour (White and Murphy 2016). Figure 1 plots the number of loops recorded each hour for the three medalists.

We can see among these three runners different strategies, the second placed runner lapped at a regular rate, the first placed runner had a fast start but slowed later, and the third placed runner also started fast but slowed more than the first place runner.

Our first goal will be, to analyze the whole data set to identify the different running strategies and to evaluate which strategies are the best ones. The second goal is to identify which variables allows to distinguish between the clusters, in order to identify which hour is essential in the management of this endurance race.

## 3 Independent Poisson Mixture

Let  $X_n = (X_{n1}, X_{n2}, \dots, X_{nM})$  be a random vector of counts for  $n = 1, 2, \dots, N$ . The goal is to clusters these  $N$  observations into  $G$  clusters. Let  $Z_n = (Z_{n1}, Z_{n2}, \dots, Z_{nG})$  be the latent cluster indicator vector, where  $Z_{ng} = 1$  if observation  $n$  belongs to cluster  $g$  and  $Z_{ng} = 0$  otherwise. We assume that  $\mathbb{P}\{Z_{ng} = 1\} = \tau_g$  for  $g = 1, 2, \dots, G$ . Let denote  $\tau = (\tau_1, \dots, \tau_G)$ . The conditionally independent Poisson mixture model (Karlis 2018, sec. 9.4.2.1) assumes that the elements of  $X_n$  are independent Poisson

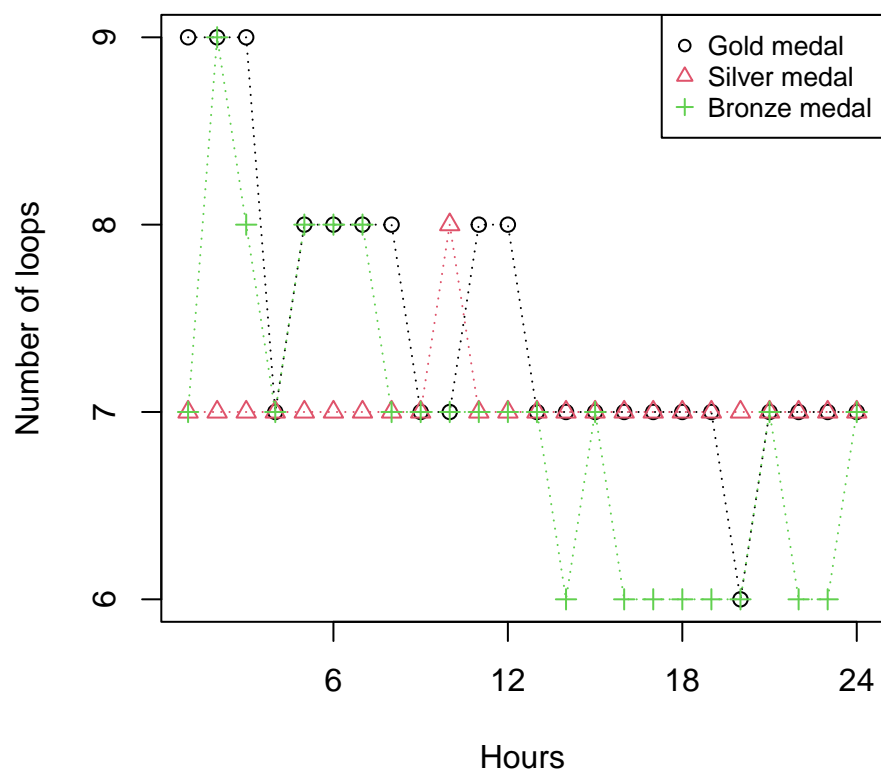


Figure 1: Number of loops per hour for the three medalists.

distributed random variables, conditional on  $Z_n$ . That is,

$$Z_n \sim \text{Multinomial}(1, \tau)$$

$$X_{nm}|(Z_{ng} = 1) \sim \text{Poisson}(\lambda_{gm}), \text{ for } m = 1, 2, \dots, M.$$

Alternative modelling frameworks exist, either to introduce some dependence between variables or to normalize the variables. We refer the interested reader to (Karlis 2018; Bouveyron et al. 2019, chap. 6) for more details.

Denoting the model parameters by  $\theta = (\tau, \lambda)$  where  $\lambda = (\lambda_{gm})_{1 \leq g \leq G, 1 \leq m \leq M}$ , and where  $X = (x_n)_{1 \leq n \leq N}$  denotes the observations, the observed likelihood is

$$L(\theta) = \sum_{n=1}^N \sum_{g=1}^G \tau_g \prod_{m=1}^M \phi(x_{nm}, \lambda_{gm}),$$

where  $\phi(x, \lambda) = \exp(-\lambda)\lambda^x/x!$ , the Poisson probability mass function.

Due to form of the mixture distribution, there are no closed form for the maximum likelihood estimators, and an iterative EM algorithm needs to be used (Dempster, Laird, and Rubin 1977) to maximize the likelihood. The EM algorithm consists, starting from an initial value  $\theta^{(0)}$  for the model parameter, and alternates the two following steps until convergence of the likelihood.

At the  $q$ th iteration of the EM algorithm, the E-step consists of computing for all  $1 \leq n \leq N$  and  $1 \leq g \leq G$ :

$$t_{ng}^{(q)} = \frac{\tau_g^{(q)} \prod_{m=1}^M \phi(x_{nm}, \lambda_{gm})}{\sum_{h=1}^G \tau_h^{(q)} \prod_{m=1}^M \phi(x_{nm}, \lambda_{hm})}.$$

In the M-step, the model parameters are updated as follows:

$$\tau_g^{(q+1)} = \frac{\sum_{n=1}^N t_{ng}^{(q)}}{N} \quad \text{and} \quad \lambda_{gm}^{(q+1)} = \frac{\sum_{n=1}^N t_{ng}^{(q)} x_{nm}}{\sum_{n=1}^N t_{ng}^{(q)}}.$$

The EM algorithm steps are iterated until convergence, where convergence is determined when  $\log L(\theta^{(q+1)}) - \log L(\theta^{(q)}) < \epsilon$ .

The number of clusters  $G$  is selected using the Bayesian information criterion (BIC) (Schwarz 1978),

$$BIC = 2 \log L(\hat{\theta}) - \{(G-1) + GM\} \log(N),$$

where  $\hat{\theta}$  is the maximum likelihood estimate of the model parameters; models with higher BIC are preferred to models with lower BIC.

## 4 Variable selection

We develop a model-based clustering method with variable selection for multivariate count data. The method follows the approach of (Raftery and Dean 2006; Maugis, Celeux, and Martin-Magniette 2009) for continuous data and (Dean and Raftery 2010; Fop, Smart, and Murphy 2017) for categorical data. It consists in a stepwise model comparison approach where variables are added and removed from a set of clustering variables.

## 4.1 Model setup

The clustering and variable selection approach is based around partitioning  $X_n = (X_n^C, X_n^P, X_n^O)$  into three parts:

- $X_n^C$ : The current clustering variables,
- $X_n^P$ : The proposed variable to add to the clustering variables,
- $X_n^O$ : The other variables.

For simplicity of notation,  $C$  will be used to denote the set of indices of the current clustering variables,  $P$  the indices of the proposed variable and  $O$  the indices of the other one. Then  $(C, P, O)$  is a partition of  $\{1, \dots, M\}$ .

The decision on whether to add the proposed variable to the clustering variables is based on comparing two models:

$M_1$  (Clustering Model), which assumes that the proposed variable is useful for clustering:

$$(X_n^C, X_n^P) \sim \sum_{g=1}^G \tau_g \prod_{m \in \{C, P\}} \text{Poisson}(\lambda_{gm}).$$

$M_2$  (Non-Clustering Model) which assumes that the proposed variable is not useful for clustering, but is potentially linked to the clustering variables through a Poisson GLM, that is,

$$X_n^C \sim \sum_{g=1}^G \tau_g \prod_{m \in C} \text{Poisson}(\lambda_{gm})$$

$$X_n^P | (X_n^C = x_n^C, Z_{ng} = 1) \sim \text{PoissonGLM}(x_n^{(C)}),$$

where Poisson GLM states that

$$\log E[X_n^P | X_n^C = x_n^C, Z_{ng} = 1] = \alpha + \beta^\top x_n^C.$$

In order to avoid non significant terms in the Poisson GLM model, a standard stepwise variable selection approach (using BIC as the variable selection criterion) is considered. Thus, the proposed variable  $X_n^P$  will be dependent on only a subset  $X_n^R$  of the clustering variables  $X_n^C$ .

The clustering and non-clustering models are represented as graphical models in Figure 2.

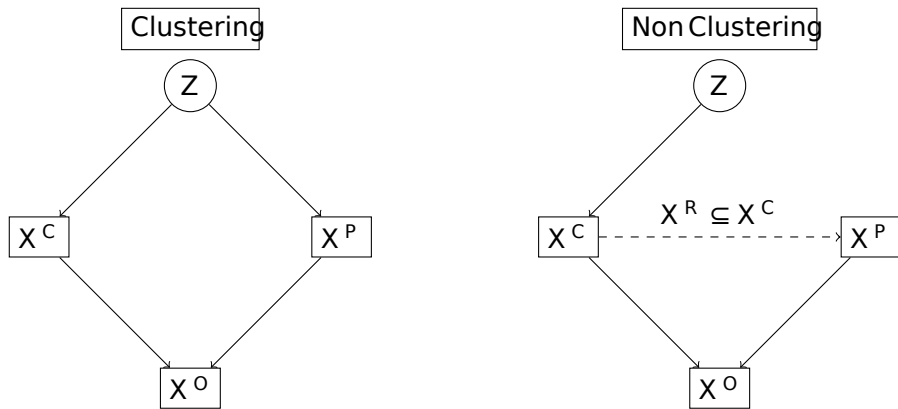


Figure 2: Graphical model representations of the clustering and non-clustering models.

Thus, there is two reasons for which  $M_2$  can be preferred to  $M_1$ : either  $X_n^P$  does not contain information about the latent clustering variable at all (ie.  $X_n^R = \emptyset$ ), or  $X_n^P$  does not add further useful information about the clustering given the information already contained in the current clustering variables. In the first situation, we say that  $X_n^P$  is an irrelevant variable, because it contains no clustering information. In the second situation, we say that  $X_n^P$  is a redundant variable because it contains no extra information about the clustering beyond the current clustering variables ( $X_n^C$ ).

Additionally, both models assume the same form for the conditional distribution for  $X_n^O | (X_n^C, X_n^P)$  and whose form doesn't need to be explicitly specified because it doesn't affect the model choice.

Variable  $P$  is added to  $C$  if the clustering model ( $M_1$ ) is preferred to the non-clustering model ( $M_2$ ). In order to compare  $M_1$  and  $M_2$ , following (Dean and Raftery 2010), we consider the Bayes Factor:

$$B_{1,2} = \frac{p(X|M_1)}{p(X|M_2)}$$

which is asymptotically approximated (Fop, Smart, and Murphy 2017; Kass and Raftery 1995) using the difference of the BIC criteria for both models:

$$2 \log B_{1,2} \simeq BIC_{M_1} - BIC_{M_2}.$$

The same modelling framework can be used for removing variables from the current set of clustering variables.

## 4.2 Interpretation

Comparing  $M_1$  and  $M_2$  is equivalent to comparing the following  $X_n^P | (X_n^C = x_n^C)$  structures.

The  $M_1$  (Clustering Model) assumes that,

$$X_n^P | (X_n^C = x_n^C) \sim \sum_{g=1}^G \mathbb{P}\{Z_{ng} = 1 | X_n^C = x_n^C\} \text{Poisson}(\lambda_{gm}),$$

where

$$\mathbb{P}\{Z_{ng} = 1 | X_n^C = x_n^C\} = \frac{\tau_g \prod_{m=1}^M \phi(x_{nm}, \lambda_{gm})}{\sum_{h=1}^G \tau_h \prod_{m=1}^M \phi(x_{nm}, \lambda_{hm})}.$$

Whereas, the  $M_2$  (Non-Clustering Model) assumes that,

$$X_n^P | (X_n^C = x_n^C) = \text{PoissonGLM}(x_n^C).$$

The method contrasts which of conditional model structures is better describing the distribution of the proposed variable  $X^P$ . The clustering model ( $M_1$ ) uses a mixture model, with covariate dependent weights, for the conditional model whereas the non-clustering model ( $M_2$ ) is a Poisson generalized linear model. The model selection criterion chooses the model that best models this conditional distribution.

## 4.3 Stepwise selection algorithm

### 4.3.1 Screening variables: Initialization

We start with an initial choice of  $C$  by first screening each individual variable by fitting a mixture of univariate Poisson distributions (eg. Everitt and Hand 1981, chap. 4.3),

$$X_{nm} \sim \sum_{g=1}^G \tau_g \text{Poisson}(\lambda_{gm}), \text{ for } G = 1, 2, \dots, G_{max}.$$

The initial set of variables is set to be those variables where the any model with  $G > 1$  is preferred to the  $G = 1$  model.

#### 4.3.2 Stepwise algorithm: Updating

We consider a stepwise algorithm which alternates between adding and removing steps. In the removal step, all the variables in  $X^C$  are examined in turn to be removed from the set. In the adding step, all the variables in  $X^O$  are examined in turn to be added to the clustering set.

The algorithm also performs the selection of the number  $G$  of clusters finding at each stage the optimal combination of clustering variables and number of clusters. The procedure stops when no change has been made to the set  $X^C$  after consecutive exclusion and inclusion steps.

With the present stepwise selection algorithm, it can occur that during the process, we get back on a solution (a set of clustering variable) already explored. Since our algorithm is not stochastic, we fall into an infinite cycle. In this situation the algorithm is stopped, and the best solution according to BIC among the solution of the cycle is kept.

## 5 Simulation study

In this section, we evaluate the proposed variable selection method through three different simulation scenarios. We start with an illustrative example in which, using a data set simulated according to the proposed model, we show how to perform the variable selection.

Then, simulation studies are performed to evaluate the behaviour of the proposed selection method, when the data are simulated according to the proposed model (Section 5.2) and when the model assumptions are violated. In Section 5.3, the link between  $X^R$  and  $X^C$  is no longer a Poisson GLM but a linear model. In Section 5.4, the clustering variables are no longer conditionally independent.

### 5.1 Illustrative example

In the first simulation setting we consider 10 Poisson random variables. Variables  $X_1$ ,  $X_2$ ,  $X_3$  and  $X_4$  are the clustering variables, distributed according to a mixture of  $G = 3$  independent Poisson mixture distributions with mixing proportions 0.4, 0.3, 0.3. Variables  $X_5$ ,  $X_6$  and  $X_7$  are redundant variables, each one generated dependent on the clustering variables. These three variables are linked to the four first ones through a Poisson GLM. The last three variables,  $X_8$ ,  $X_9$  and  $X_{10}$  are irrelevant variables not related to the previous ones.

Below is the result obtained for one data set of size  $N = 400$ . The evaluation criteria is the selected features (true one are  $X_1$  to  $X_4$ ) and the Adjusted Rand Index (Rand 1971; Hubert and Arable 1985) obtained with the selected variables in comparison to those obtained with the full set of variables and with the true clustering variables.

The independent Poisson mixture model was fitted to the simulated data with  $N = 400$  rows and  $P = 10$  columns. Models with  $G = 1$  to  $G = 2$  were fitted using the EM algorithm.

The values of BIC for the independent Poisson mixture model are plotted in Figure 3.

The model with the highest BIC has  $G = 2$  components and the resulting estimates of  $\tau$  and  $\lambda$  are given as:

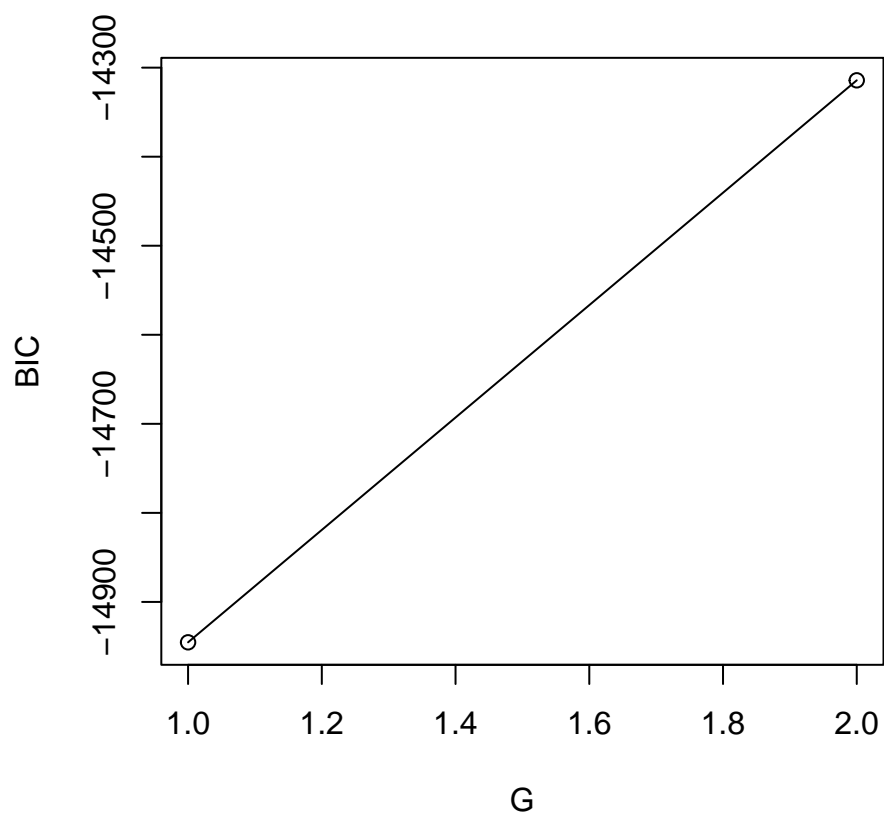


Figure 3: Bayesian Information Criterion (BIC) for the independent Poisson mixture model.



Table 1: Estimates of the mixing proportions and component parameters.

	$\tau_g$	$\lambda_{g1}$	$\lambda_{g2}$	$\lambda_{g3}$	$\lambda_{g4}$	$\lambda_{g5}$	$\lambda_{g6}$	$\lambda_{g7}$	$\lambda_{g8}$	$\lambda_{g9}$	$\lambda_{g10}$
$g = 1$	0.4	3.72	3.68	3.52	4.47	2.33	1.75	3.60	3.98	1.93	1.08
$g = 2$	0.6	1.37	1.38	1.12	2.19	1.35	1.11	1.47	3.88	2.03	0.98

170 Let start by initializing the stepwise algorithm.

```
fit_screen <- poissonmix_screen(x, G = 1:Gmax)
jchosen <- fit_screen$jchosen
```

171 The variables selected by the screening procedure are {1, 2, 3, 4, 6, 7}.

172 Now, we execute the stepwise selection algorithm:

```
fit <- poissonmix_varssel(x, jchosen=jchosen, G = 1:Gmax)
```

```
173 [1] "Initial Selected Variables: 1,2,3,4,6,7"
174 [1] "Iteration: 1"
175 [1] "Add Variable: NONE 10 BIC Difference: -5.4"
176 [1] "Remove Variable: 6 BIC Difference: 84.1"
177 [1] "Current Selected Variables: 1,2,3,4,7"
178 [1] "Iteration: 2"
179 [1] "Add Variable: NONE 10 BIC Difference: -5.4"
180 [1] "Remove Variable: 7 BIC Difference: 58.3"
181 [1] "Current Selected Variables: 1,2,3,4"
182 [1] "Iteration: 3"
183 [1] "Add Variable: NONE 10 BIC Difference: -5.7"
184 [1] "Remove Variable: 3 BIC Difference: 33.1"
185 [1] "Current Selected Variables: 1,2,4"
186 [1] "Iteration: 4"
187 [1] "Add Variable: NONE 9 BIC Difference: -5.6"
188 [1] "Remove Variable: NONE 2 BIC Difference: -30.1"
189 [1] "Current Selected Variables: 1,2,4"
```

190 The final chosen variables are {1, 2, 4}.

191 Finally, the ARI obtained with the selected variables, which turn out to be the true clustering variable,  
192 is 0.404 whereas it is 0.398 with all the variables.

## 193 5.2 Scenario 1

194 The section give the results for 4 simulated data sets for simulation setting number 1, which is similar  
195 to the previous illustrative example.

196 Table 2 shows the number of times, among the 4 simulated data sets, that each variable is selected. The  
197 model selection procedure perform perfectly, selecting each time only the true clustering variables.  
198 Figure 4 shows the ARI obtained on 4 simulations with the selected variables, with all the variables  
199 and with the true clustering variables.

Table 2: Number of selection for each variable, simulation setting number 1.

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$
Number of selections	100	100	100	100	0	0	0	0	0	0

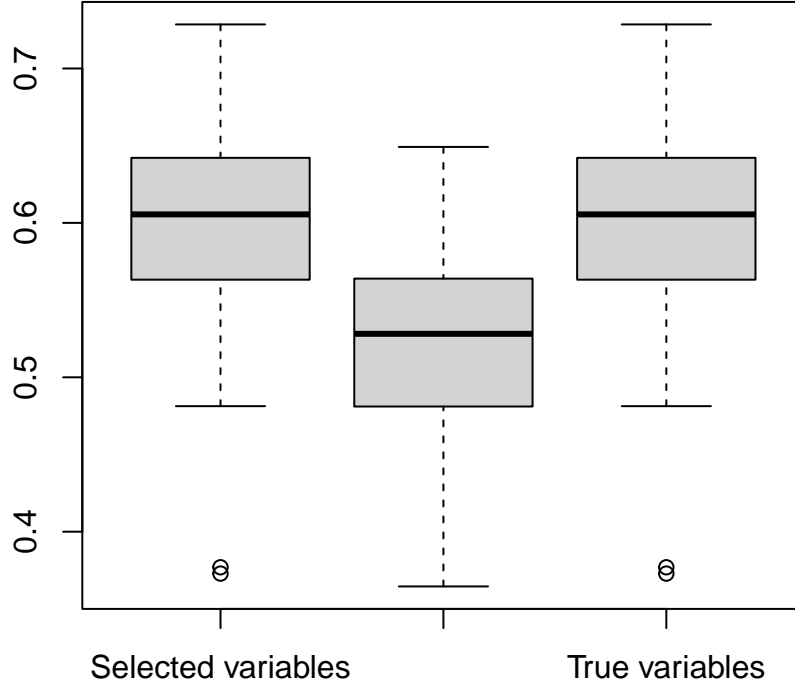


Figure 4: Adjusted Rand Index in function of the variables used for clustering, simulation setting number 1.

Since in this simulation setting, the selected variables are always the true clustering variables (Table 2), the first and third boxplot are similar. In addition, as expected the ARI is better with the selected variables than when using all the variables.

Since the variability of ARI among the simulations could hide the difference observed for each simulation, Figure 5 plots the histogram of the difference of ARI with the selected variables and with all the variables. We can see that the difference is almost always positive.

### 5.3 Scenario 2

The second scenario is similar to the first one, except for variables  $X_5$  and  $X_6$  which are still redundant but linked to the true clustering variables through a linear and quadratic term in an identity link

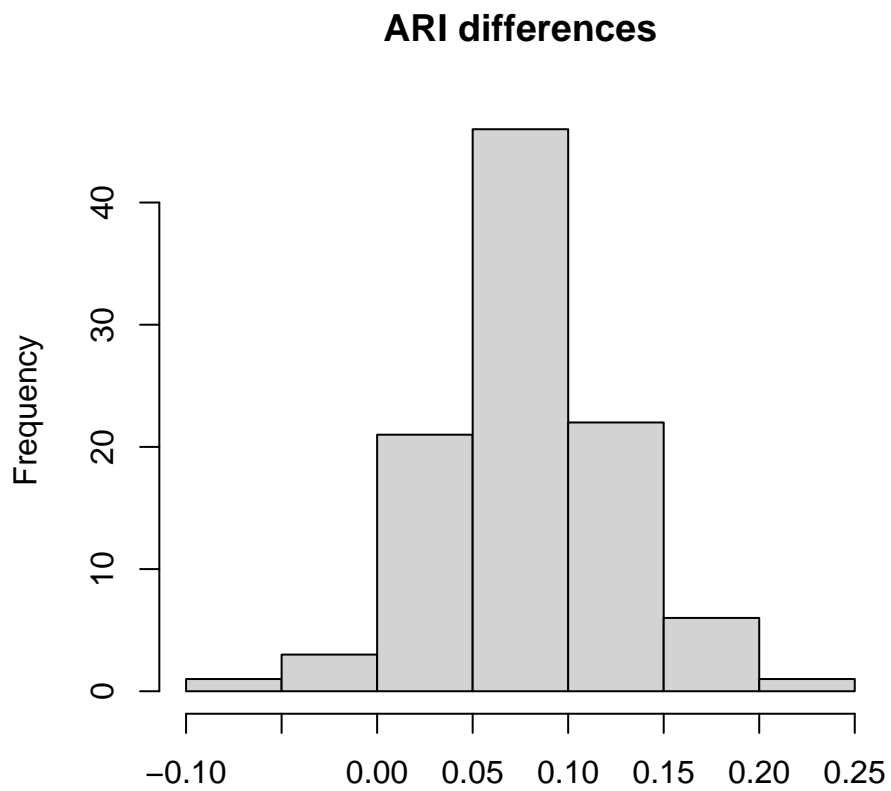


Figure 5: Distribution of the ARI differences for the model with the selected variables and the model with all variables, simulation setting number 1.

function, respectively, and not a Poisson GLM with logarithm link function. Thus, the data are simulated from a model which does not satisfy assumptions of model  $M_2^*$ .

Due to the fact the link between the redundant and the true clustering variables is not a standard Poisson GLM, the variable selection is perturbed and variables  $X_5$  is sometimes selected (Table 3). But the ARI results (Figure 6 and Figure 7) are still better with the selected variables than with all the variables.

Table 3: Number of selection for each variable, simulation setting number 2.

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$
Number of selections	97	100	90	98	44	0	0	0	0	0

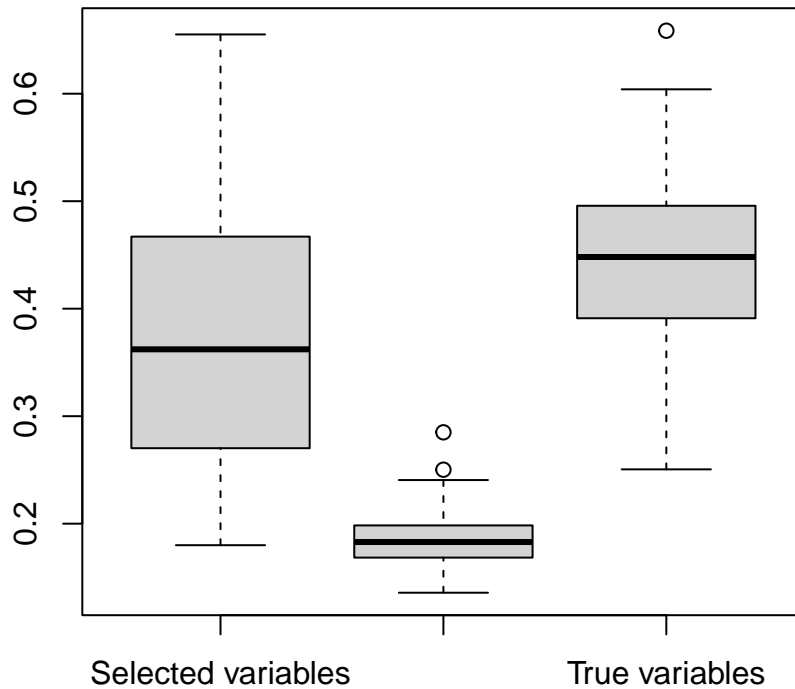


Figure 6: Adjusted Rand Index in function of the variables used for clustering, simulation setting number 2.

## 5.4 Scenario 3

The third scenario is similar to the second one, but some dependence between the clustering variables  $X_1$  and  $X_2$  is introduced, in order to create some redundancy among the true clustering variables.

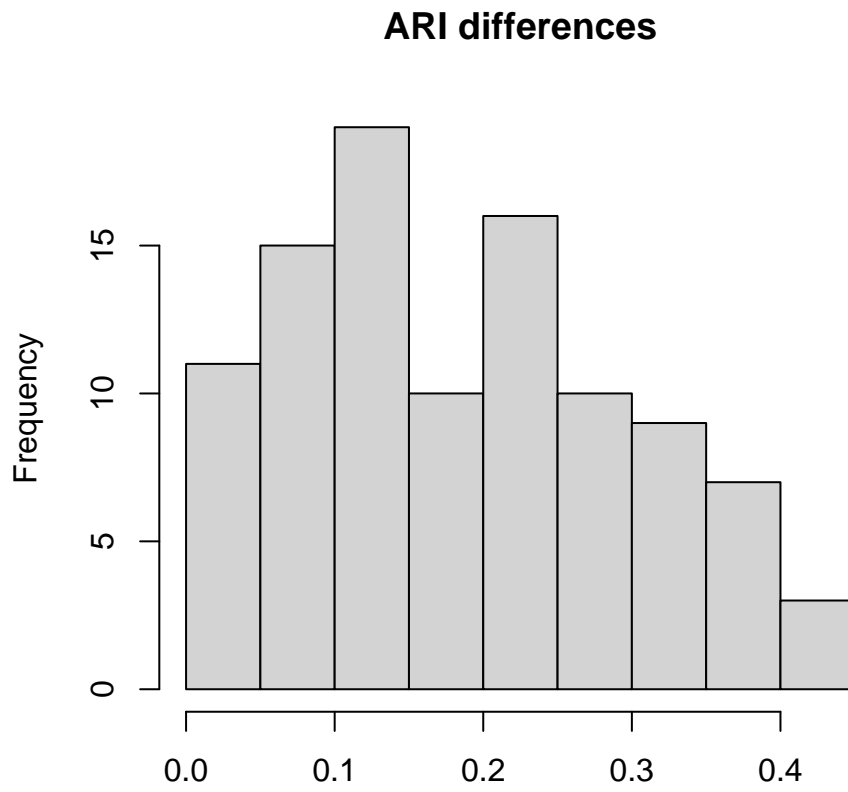


Figure 7: Distribution of the ARI differences for the model with the selected variables and the model with all variables, simulation setting number S.

218 The results is that the dependency between  $X_1$  and  $X_2$  perturb the variable selection, and only one  
 219 of them is selected (and even sometimes none of them). Redundant variables  $X_5$  and  $X_6$ , which are  
 220 linked to the clustering variables but with a linear link, are also sometimes selected (Table 4). But  
 221 from an ARI point of view, the selected variables lead to a better ARI than using all the variables  
 222 (Figure 8, Figure 9).

Table 4: Number of selection for each variable, simulation setting number 3.

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$X_{10}$
Number of selections	48	35	89	88	65	34	3	0	0	0

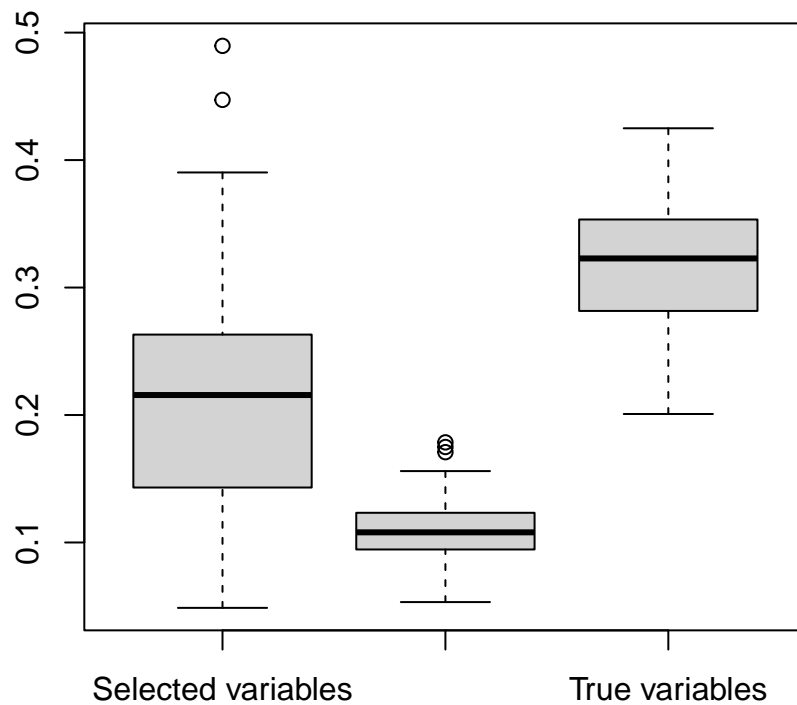


Figure 8: Adjusted Rand Index in function of the variables used for clustering, simulation setting number 3.

## 223 6 International Ultrarunning Association Data

224 We apply the proposed procedure to the data from the 2012 International Ultrarunning Association  
 225 World 24H Championships.

226 We start by initializing the stepwise algorithm, and find the variables selected by the screening  
 227 procedure:

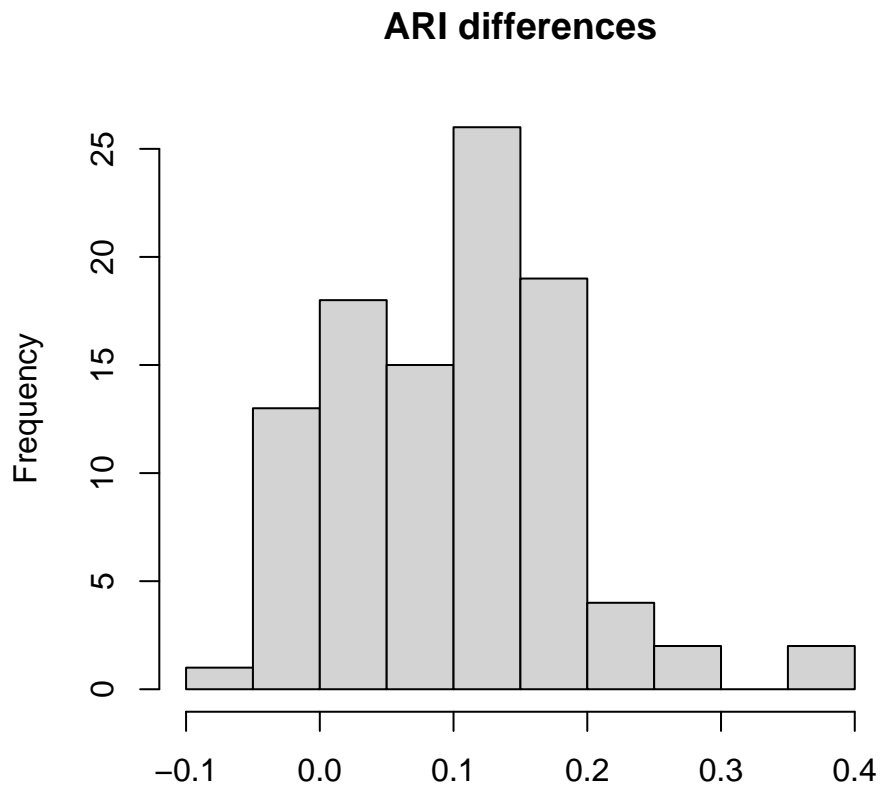


Figure 9: Distribution of the ARI differences for the model with the selected variables and the model with all variables, simulation setting number 3.

```
fit_screen <- poissonmix_screen(x, G = 1:Gmax)
jchosen <- fit_screen$jchosen
jchosen
```

228 [1] 3 5 6 7 8 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24

229 We then execute the proposed stepwise selection algorithm:

```
fit <- poissonmix_varsel(x, jchosen = jchosen, G = 1:Gmax)
```

230 The final chosen variables found by the algorithm are:

231 [1] 9 10 11 12 14 15 16 17 18 19 20 21 22 24

232 In order to illustrate the results, we plot the cluster means according to the 24 variable mean  
 233 parameters per cluster. For each variable not in the chosen variable set, a Poisson regression model is  
 234 fitted with the chosen variables as predictors. Forward and backwards variable selection is conducted  
 235 on this regression, if the regression model has any predictor variables, then the variable is called  
 236 “redundant” and if the regression model has no predictor variables, then the variable is called  
 237 “irrelevant”.

238 Figure 10 shows the cluster mean for each variable, where the label indicates if the variable is  
 239 irrelevant for clustering (“I”), redundant (“R”) or useful (the label is then the cluster number).

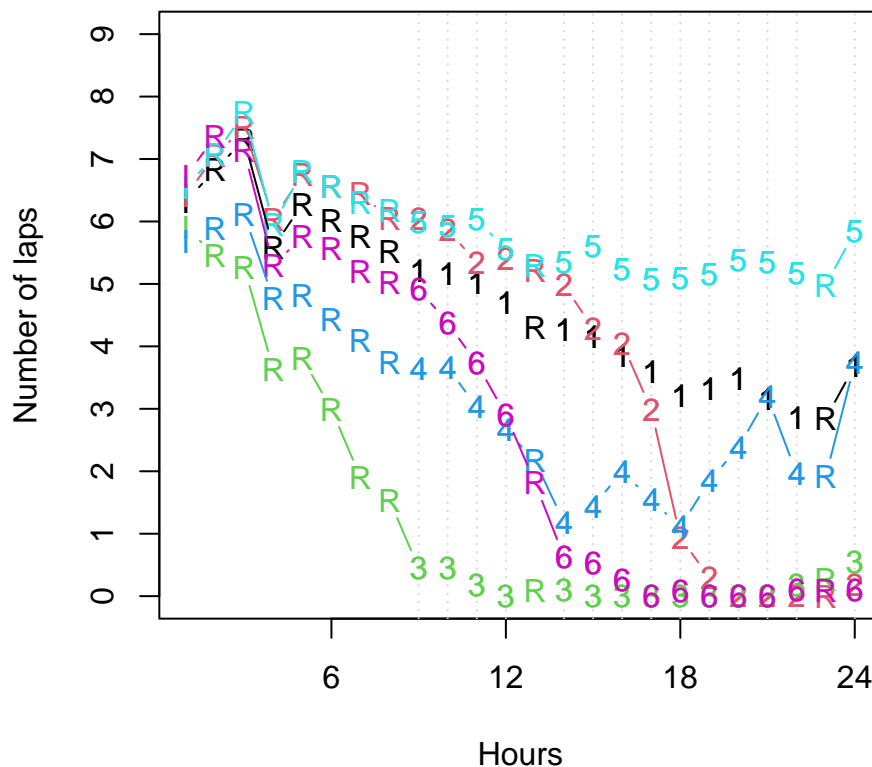


Figure 10: Cluster means and usefulness of the variables.

240 The variables discriminate the clusters pacing strategies of the runners are the number of laps covered



during the last two thirds of the race (except during the 13th and 23rd hours). The number of laps covered during the first eight hours does not provide any additional clustering information, and even no information at all for the number of laps covered during the first hour.

Figure 11 shows boxplots of the total number of loops covered by the runners in each of the clusters.

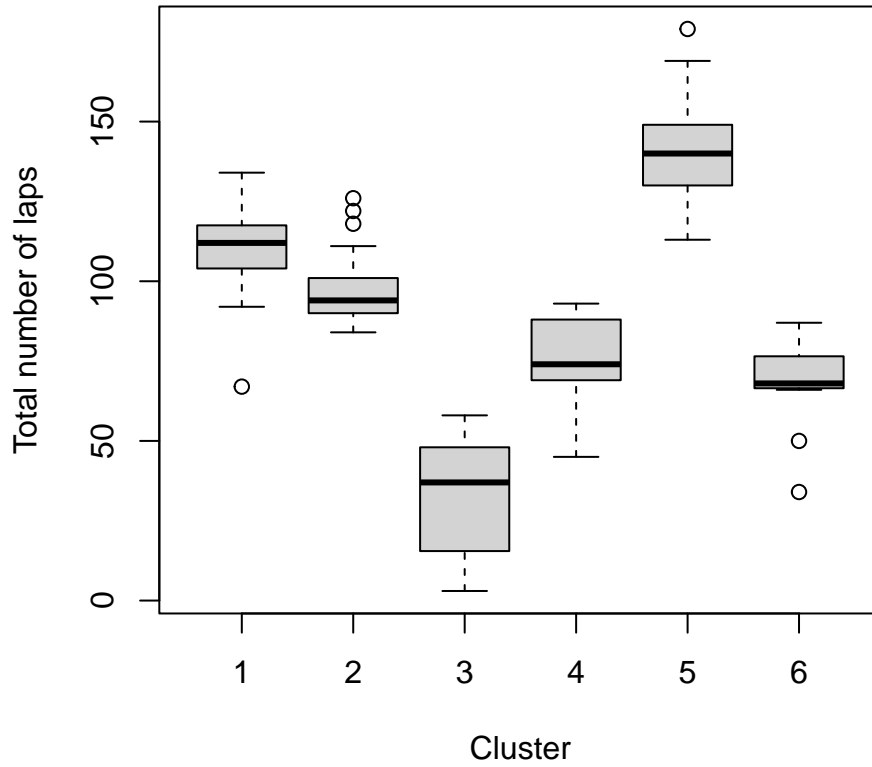


Figure 11: Number of loops covered by the runners of each clusters.

Cluster 5 are clearly the most efficient runners. Looking at the running strategy in Figure 10, we can see that they start as runners of Cluster 1 and Cluster 2, but they managed to keep a constant pace on the second part of the race, unlike those of the other two clusters which faltered. Runners of Cluster 3 has covered the fewest number of laps. Indeed, looking at their running strategy, we can see that most of these runners stop after the first third of the race. Cluster 6 is relatively similar to Cluster 3, but runners manage to continue running until half of the race is completed. Finally, Cluster 4 obtains slightly better results than Cluster 6, starting more carefully, and managing to run until the end of the race, even if the pace of the last hours is not very constant.

## 7 Discussion

A method for clustering and variable selection for multivariate count data has been proposed. The method is shown to give excellent performance on both simulated and real data examples. The method selects set of relevant variables for clustering and other variables are not selected if they are irrelevant or redundate for clustering purposes.

The proposed method is shown to give interesting insights in the application domain, where some clusters members are shown to perform better overall to others and the benefits of constant (or near constant pacing) are shown.

The level of variable selection is determined by the relative performance of the two models (as shown in Section 4.2) is compared. Alternative models to the Poisson GLM model which have greater flexibility could lead to a smaller set of selected variables than the proposed method achieves. This is a topic for future research.

The proposed method could be extended to other count data distributions, including multivariate distributions without the conditional independence assumption (eg. Karlis 2018; Karlis and Meligkotsidou 2007; Inouye et al. 2017b).

The code for the proposed approach will be made available as an R package.

## 8 Acknowledgements

This work was supported by the Science Foundation Ireland Insight Research Centre (SFI/12/RC/2289\_P2) and a visit to the Collegium – Institut d’Études Avancées de Lyon.

Agresti, Alan. 2013. *Categorical Data Analysis*. Third. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Bouveyron, Charles, Gilles Celeux, T. Brendan Murphy, and Adrian E. Raftery. 2019. *Model-Based Clustering and Classification for Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. <https://doi.org/10.1017/9781108644181>.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin. 2021. “The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances.” *Frontiers in Ecology and Evolution* 9: 188. <https://doi.org/10.3389/fevo.2021.588292>.

Dean, Nema, and Adrian E. Raftery. 2010. “Latent Class Analysis Variable Selection.” *Annals of the Institute of Statistical Mathematics* 62 (1): 11–35. <https://doi.org/10.1007/s10463-009-0258-9>.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society: Series B* 39: 1–38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>.

Everitt, Brian S., and David J. Hand. 1981. *Finite Mixture Distributions*. Chapman & Hall.

Fop, Michael, and Thomas Brendan Murphy. 2018. “Variable Selection Methods for Model-Based Clustering.” *Statistics Surveys* 12: 18–65. <https://doi.org/10.1214/18-SS119>.

Fop, Michael, Keith Smart, and Thomas Brendan Murphy. 2017. “Variable Selection for Latent Class Analysis with Application to Low Back Pain Diagnosis.” *Annals of Applied Statistics*. 11: 2085–115.

Frühwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P. Robert. 2018. *Handbook of Mixture Analysis*. Chapman; Hall/CRC. <https://doi.org/10.1201/9780429055911>.

Hartigan, John A., and M. Anthony Wong. 1979. “A k-Means Clustering Algorithm.” *Applied Statistics* 28 (1): 100–108.

Hubert, Lawrence, and Phipps Arable. 1985. “Comparing Partitions.” *Journal of Classification* 2 (1): 193–218. <https://doi.org/10.1007/BF01908075>.

Inouye, David I., Eunho Yang, Genevera I. Allen, and Pradeep Ravikumar. 2017a. “A Review of Multivariate Distributions for Count Data Derived from the Poisson Distribution.” *WIREs Computational Statistics* 9 (3): e1398. <https://doi.org/https://doi.org/10.1002/wics.1398>.

———. 2017b. “A Review of Multivariate Distributions for Count Data Derived from the Poisson Distribution.” *WIREs Computational Statistics* 9 (3): e1398. <https://doi.org/10.1002/wics.1398>.

Karlis, Dimitris. 2018. “Mixture Modelling of Discrete Data.” In *Handbook of Mixture Analysis*, edited by Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert, 193–218. CRC Press.

- Karlis, Dimitris, and Loukia Meligkotsidou. 2007. "Finite Mixtures of Multivariate Poisson Distributions with Application." *Journal of Statistical Planning and Inference* 137 (6): 1942–60. <https://doi.org/10.1016/j.jspi.2006.07.001>.
- Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–95. <https://doi.org/10.1080/01621459.1995.10476572>.
- Maugis, Cathy, Gilles Celeux, and Marie-Laure Martin-Magniette. 2009. "Variable Selection in Model-Based Clustering: A General Variable Role Modeling." *Computational Statistics & Data Analysis* 53 (11): 3872–82. <https://doi.org/10.1016/j.csda.2009.04.013>.
- McLachlan, Geoffrey, and David Peel. 2000. *Finite Mixture Models*. New York: Wiley. <https://doi.org/10.1002/0471721182>.
- McParland, Damien, and Thomas Brendan Murphy. 2018. "Mixture Modelling of High-Dimensional Data." In *Handbook of Mixture Analysis*, edited by Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert, 247–80. CRC Press.
- Raftery, Adrian E., and Nema Dean. 2006. "Variable Selection for Model-Based Clustering." *Journal of the American Statistical Association* 101 (473): 168–78. <https://doi.org/10.1198/016214506000000113>.
- Rand, William M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* 66 (336): 846–50.
- Rau, Andrea, Cathy Maugis-Rabusseau, Marie-Laure Martin-Magniette, and Gilles Celeux. 2015. "Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models." *Bioinformatics* 31 (9): 1420–27. <https://doi.org/10.1093/bioinformatics/btu845>.
- Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. <https://doi.org/10.1214/aos/1176344136>.
- Silva, Anjali, Steven J. Rothstein, Paul D. McNicholas, and Sanjeena Subedi. 2019. "A Multivariate Poisson-Log Normal Mixture Model for Clustering Transcriptome Sequencing Data." *BMC Bioinformatics* 20 (1): 394. <https://doi.org/10.1186/s12859-019-2916-0>.
- White, Arthur, and Thomas Brendan Murphy. 2016. "Exponential Family Mixed Membership Models for Soft Clustering of Multivariate Data." *Advances in Data Analysis and Classification* 10: 521–40.