How Many Clusters Are Best? Investigating Model Selection in Robust Clustering

Louis Tran* Cristina Tortora †

Abstract

In model-based clustering, different density functions are used to model sub-populations in the data. When data are characterized by outliers, robust distributions such as the Student-t (T) or the contaminated normal (CN) distribution, and their extensions for directional tail behavior, multiple scaled (MS) T and CN, can be used. Model-based clustering methods take the number of clusters as an input parameter, and many indices exist to choose the number of clusters. In this paper, we use simulated and real data sets to compare different indices to select the number of clusters when using mixtures of T, CN, MST, and MSCN distributions. The effectiveness of each index is determined by the number of successes in selecting the right number of sub-populations in the data.

Key Words: Model based clustering, Number of clusters, Outliers, Multiple scaled distributions

1. Introduction

In model-based clustering, the population is assumed to be a mixture of subpopulations modeled by a density function. For example, when using Gaussian mixture models, also known as multivariate normal (MN) mixtures, we assume that each sub-population follows a Gaussian distribution with different parameters. A random vector $\mathbf{X}$ follows a (parametric) finite mixture distribution if its probability density function can be written as

$$f(\boldsymbol{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x} \mid \boldsymbol{\theta}_g), \tag{1}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, is the gth mixing proportion, $f_g(\boldsymbol{x} \mid \boldsymbol{\theta}_g)$ is the gth component density, and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the vector of parameters, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$. The component densities $f_1(\boldsymbol{x} \mid \boldsymbol{\theta}_1), \ldots, f_G(\boldsymbol{x} \mid \boldsymbol{\theta}_G)$ are usually taken to be of the same type. Although the MN distribution is often used in cluster analysis, it may be limiting when the clusters are characterized by mild outliers (see, e.g., Bock, 2002, Gallegos and Ritter, 2009, and Ritter, 2015). The outliers may affect the estimation of the component means and covariance matrices and the recovery of the underlying clustering structure. A common solution is to model the component with robust distributions, such as Multivariate Student-t (T) mixtures (Peel and McLachlan, 2000). The T mixtures produce robust parameter estimates but do not automatically detect outliers, an *a posteriori* procedure (i.e., a process taking place once the model is fitted) to catch outlying points with T mixtures is illustrated by McLachlan and Peel (2000). To overcome this problem, Punzo and McNicholas (2016) introduced Multivariate Contaminated Normal (MCN) mixtures that at the same time detect outliers and produce robust parameter estimates. Recently, the T and CN distributions have been extended to accommodate different tail behaviors across principal components. The models are referred to as multiple scaled distributions, i.e., MST (Forbes and Wraith, 2014) and MSCN (Punzo and Tortora, 2019) respectively. One of the challenges in cluster analysis is to estimate the number of

*San José State University, San José (CA), USA
†San José State University, San José (CA), USA

sub-populations. Many indices have been proposed in the statistical literature. However, different indices can be a better fit for different distributions. This study aims to perform a comparative study of ten popular model selection indices when clustering using a mixture of T, CN, MST, and MSCN distributions on real and simulated data sets. The effectiveness of each index is determined by the number of successes in selecting the right number of sub-populations in the data.

2. Background

2.1 Multivariate Student-t Distribution

The Multivariate Student-t(T) distribution provides robust parameter estimates that are less sensitive to outliers when compared to the MN distribution. A a p-variate random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is said to follow the T distribution with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and degree of freedom υ if its joint probability density function is in the form:

$$f\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \upsilon\right) = \frac{\Gamma\left(\frac{\upsilon+p}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\upsilon)^{\frac{p}{2}}\,\Gamma\left(\frac{\upsilon}{2}\right)\left\{1+\frac{\delta(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})}{\upsilon}\right\}^{\frac{\upsilon+p}{2}}}, \tag{2}$$

where $\delta\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Mixtures of T distributions can be obtained using the density function in equation (2) in equation (1). Parameter estimates can be obtained using the Expectation-Maximization (EM) algorithm, for details see Peel and McLachlan (2000). At convergence of the EM algorithm, cluster membership can be determined.

2.2 Multivariate Contaminated Normal Distribution

The Multivariate Contaminated Normal (MCN) Distribution proposed by Tukey (1960) has the advantage of automatically detecting outliers in the data and having robust parameter estimates. The distribution can be seen as a mixture of two components; one of the components, with a larger probability α, is assumed to have a normal distribution and represent the 'good' observations in the data. With a smaller probability $1-\alpha$, the other component represents the 'bad' observations or outliers. As an advantage, when the MCN distribution is fitted to the data, the observations are automatically detected to either be in the 'good' or 'bad' component. The 'bad' observations will then have reduced weight in estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Formally, a p-variate random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is said to follow the multivariate contaminated normal (MCN) distribution with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, proportion of good points $\alpha \in (0,1)$, and degree of contamination $\eta > 1$ if its joint probability density function (pdf) is given by

$$f_{\text{MCN}}\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \eta\right) = \alpha f_{\text{MN}}\left(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) + (1-\alpha)\, f_{\text{MN}}\left(\boldsymbol{x}; \boldsymbol{\mu}, \eta\boldsymbol{\Sigma}\right), \tag{3}$$

where $f_{\text{MN}}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the pdf of a d-variate random vector having the multivariate normal (MN) distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The EM algorithm can be used for parameter estimates in model based clustering, see Punzo and McNicholas (2016) for details.

2.3 Multiple Scaled Distributions

Although the T and CN distributions have many advantages, they also have one main limitation: the shape's flexibility is limited. For the CN distribution, when an observation is

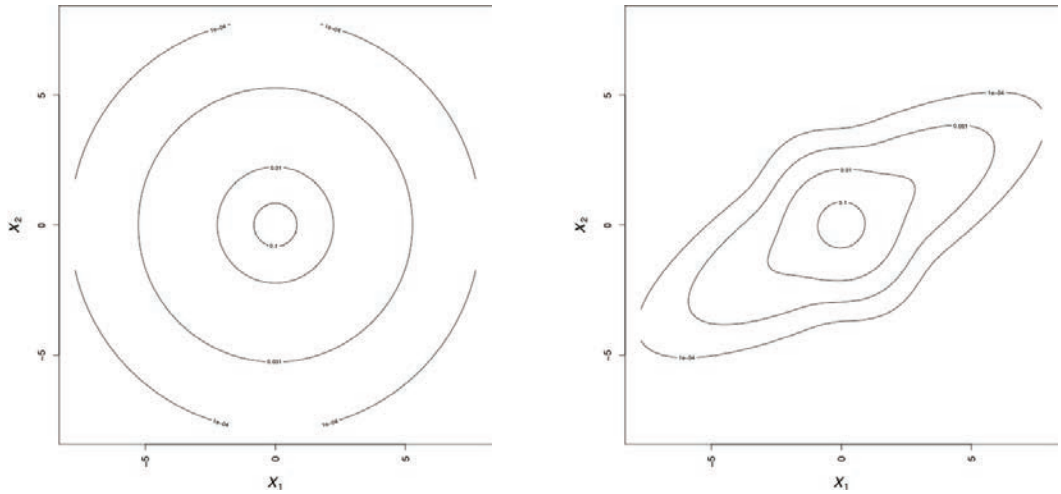

Figure 1: Example of contour plots obtained using the CN and MSCN distributions respectively with tail parameters $\alpha = 0.7$ and $\eta = 10$ for the CN, $\boldsymbol{\alpha} = (0.7, 0.9)$ and $\boldsymbol{\eta} = (10, 2)$ for the MSCN.

labeled 'bad', it is considered a 'bad' observation globally, even if it is an outlier for some variables. Similarly, for the T distribution, the degrees of freedom regulate the tail behavior globally. The multiple scaled distributions, including multiple scaled contaminated normal (MSCN), and multiple scaled Student-t (MST) distributions, are proposed to exceed these limitations. To define the MSCN and MST let first notice that both the T and CN distributions are normal-scale mixture. The distribution of a p-dimensional random variable $\boldsymbol{X}$ is said to be a normal scale-mixture if its density can be written in the form

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \int_0^\infty \phi_p(\boldsymbol{x} \mid \boldsymbol{\mu}, w\boldsymbol{\Sigma})\, h(w \mid \boldsymbol{\theta})\, dw, \tag{4}$$

where $\phi_p(\boldsymbol{x} \mid \boldsymbol{\mu}, w\boldsymbol{\Sigma})$ is the density of a p-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$, and covariance matrix $w\boldsymbol{\Sigma}$, and $h(w \mid \boldsymbol{\theta})$ is the density of a univariate random variable $W > 0$ that has the role of a weight function (see Barndorff-Nielsen et al., 1982; Gneiting, 1997). This weight function can take on many forms, some of which lead to density representations for well-known non-Gaussian distributions, e.g., if $h(w \mid \boldsymbol{\theta})$ is the density of an inverse-gamma random variable with parameters $(\nu/2, \nu/2)$, then (4) is a representation of the T distribution with ν degrees of freedom. The multiple scaled approach decomposes the scale matrix $\boldsymbol{\Sigma}$ using the eigendecomposition $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Lambda}'$ where $\boldsymbol{\Lambda}$ is an orthogonal matrix whose columns are the normalized eigenvectors, and $\boldsymbol{\Gamma}$ is the diagonal matrix of the eigenvalues. Moreover, the CN introduces p Bernoulli random variables indicating whether a point is good or bad separately for each principal component of the space spanned by the columns of $\boldsymbol{\Gamma}$. The result is a distribution with the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, controlling the proportion of good points and the degree of contamination, respectively, separately for each principal component. Similarly, for the T distribution, the MST uses p univariate Gamma distributions, obtaining different degrees of freedom per principal component. The MST and MSCN mixtures allow for directional robustness. Figure 1 shows an example of contour plots obtained using the CN and MSCN distributions, respectively. For both distributions $\boldsymbol{\mu} = (0, 0)$, $\boldsymbol{\Sigma} = 0.75\boldsymbol{I}$, for the CN $\alpha = 0.7$ and $\eta = 10$, while for the MSCN $\boldsymbol{\alpha} = (0.7, 0.9)$ and $\boldsymbol{\eta} = (10, 2)$. Parameter estimates can be achieved using an extension of the EM algorithm, see Forbes and Wraith (2014) and Punzo and Tortora (2019) for details.

3. Information Criteria for Model Selection

Since parameter estimation for the model-based clustering methods presented in Section 2 is obtained using a maximum likelihood approach, the indices evaluated to choose the number of components are all based on the maximum likelihood approach. With n sample size and G number of clusters, each with some unknown weight π_1,...,π_G, the likelihood function is defined as (Fraley, 1998):

$$L(\psi) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}_i|\boldsymbol{\theta}_g),$$

where $f_g(\boldsymbol{x}_i|\boldsymbol{\theta}_g)$ is the probability density function of the components.
In the following we will define as

- $LogL(\psi)$ the maximum log likelihood value obtained at convergence of the algorithm.
- d is the number of parameters in the model.
- n is the sample size.
- $EN(\tau) = -\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_{ij} log \tau_{ij}$ is the entropy of the fuzzy classification matrix of elements τ_{ij}.
- $LogL_c(\psi) = LogL(\psi) + EN(\tau)$ as the log-likelihood function for the completed data. (Hathaway, 1986)
- $\psi(.)$ is the digamma or psi function.

The criteria used in our study are provided below:

- Akaike's Information Criterion (AIC) (Akaike (1998)):

$$AIC = -2LogL(\psi) + 2d$$

- Akaike's Information Criterion (AIC3) (Bozdogan (1993)):

$$AIC = -2LogL(\psi) + 3d$$

- Consistent Akaike's Information Criterion (CAIC) (Bozdogan (1987)):

$$AIC = -2LogL(\psi) + d(log(n) + 1)$$

- When sample size n is small relative to d, a version of AIC called AICc is used (Hurvich and Tsai (1989)):

$$AICc = -2LogL(\psi) + 2dn/(n - d - 1)$$

- Bayesian Information Criterion (BIC) (Schwarz et al. (1978)):

$$BIC = -2LogL(\psi) + dlog(n)$$

- Classification Likelihood Criterion (CLC) (Biernacki et al. (2000)):

$$CLC = -2LogL(\psi) + 2EN(\tau)$$

- Approximate Weight of Evidence (AWE) (Banfield and Raftery (1993)):

$$AWE = -2LogL_c + 2d\left(\frac{3}{2} + log(n)\right)$$

- Kullback Information Criterion (KIC) (Cavanaugh (1999)):

$$KIC = -2LogL(\psi) + 3(d+1)$$

- Bias correction of the Kullback Information Criterion (KICc) (Seghouane and Bekara (2004)):

$$KICc = -2LogL(\psi) + \frac{2(d+1)n}{n-d-2} - n\psi\left(\frac{n-d}{2}\right) + nlog\left(\frac{n}{2}\right)$$

- Integrated Completed Likelihood (ICL) (approximation) (Biernacki and Govaert (1997)):

$$ICL = BIC - EN(\tau)$$

4. Results

All the results were obtained using the software `R` (R Core Team, 2016), the mixture of T distributions is available with the package `teigen` (Andrews et al., 2018), the CN with the package `ContaminatedMixt` (Punzo et al., 2016), for the MST and MSCN we used the code provided by the authors of Punzo and Tortora (2019).

4.1 Results on Real Data

In this section, we compare the information criteria's performance on some popular real data sets described in Table 1.

Data set	Sample Size	Number of Variables	Number of Clusters	Reference
Iris	150	4	3	native in R
Faithful	272	2	2	native in R
Bankruptcy	66	2	2	`MixGHD` package (Tortora et al., 2021)
Wine	178	27	3	`pgmm` package (McNicholas et al., 2011)
Liver	345	6	2	UCI ML repository
Ecoli	336	7	4	UCI ML repository
Ruspini	75	2	4	`cluster` package (Maechler et al., 2021)

Table 1: Description of real data sets

Table 2 presents the result of the information criteria obtained clustering the Iris data set using the CN method. The star indicates the smallest value per index. The number in boldface is the correct number of clusters. The value Inf means that the index was too big to be calculated and, therefore, could not be used for model selection. The tables for the other methods and data sets are in Appendix.

Table 3 shows if the index picks the correct number of clusters per method per data set. The value NA is used when the method did not run on the data set. Many indices pick the correct number of clusters for 4 out of the 7 data sets when using the CN distribution, with BIC, ICL, and CAIC being the best picking the correct number of clusters 5 times out of 7. The BIC, ICL, and CAIC pick the correct number of clusters for 5 out of the 7 data sets for the T distribution. When using the MSCN, the BIC and ICL pick the correct number of clusters for 4 out of the 7, the corresponding indices for the MST are the KIC and AIC3.

	CN Iris			
	2	**3**	4	5
AIC	494.71	460.37*	463.38	470.34
BIC	594.06*	610.90	665.09	723.23
ICL	594.06*	614.11	672.31	736.43
KIC	530.71	513.37*	533.38	557.34
KICc	555.96*	578.82	671.82	826.21
AWE	858.41*	1011.43	1201.81	1396.13
AIC3	527.71	510.37*	530.38	554.34
CAIC	627.06*	660.90	732.09	807.23
AICc	514.05	511.89*	574.50	690.03
CLC	428.72	370.12	347.11	331.72*

Table 2: Values of all the indices on the Iris data set using CN

4.2 Simulation Design

We generated simulated data sets from three different distributions, the Student-t distribution (R package `mvtnorm` Genz et al. (2019)), contaminated normal distribution (R package `ConiaminatedMixt` Punzo et al. (2016)), and generalized hyperbolic distribution (GHD) (R package `MixGHD` Tortora et al. (2021)). For each type of distribution, we created 2, 3, and 4 clusters with 200 observations each and 10 variables, Table 4 shows a summary of the simulation set up. For each scenario we simulated 5 different data sets, and report the percentage of times each information criteria recommend the right number of clusters.

For all the distributions, the vectors $\boldsymbol{\mu}$ were pseudo randomly generated from a uniform distribution, $\boldsymbol{\mu}_1$ between 0 and 1, $\boldsymbol{\mu}_2$ between 3 and 4, $\boldsymbol{\mu}_3$ the first 5 variables between 6 and 7 and the second five between 0 and 1, $\boldsymbol{\mu}_4$ between 6 and 7. The scale matrices were randomly generated using the function `genPositiveDefMat`, from the R package `clustergeneration` (Qiu and Joe., 2020). For the Student-t distribution, the degrees of freedom were randomly generated between 5 and 20. For the contaminated normal distribution, α_g was randomly generated between 0.6 and 0.9, while η_g was between 2 and 10. For the GHD distribution, the skewness was randomly generated between 0.6 and 0.9, the concentration parameter, ω, was set to 1, and the index, λ to 0.5.

5. Simulation Results

Figure 2 and table 5 show the percentage of times each information criteria recommend the right number of clusters overall. Figure 3 and tables 6, 7, and 8 in appendix show the percentages for each distribution used; figure 4 and tables 9, 10, and 11 in appendix for each number of clusters. Each method need to be analyzed separately. The indices that work the best for the MSCN and MST distributions are the BIC, ICL, KIC, AIC3, and CAIC. Although these indices are not always the best, their performances are close to the best ones. Similarly, for the Student-t, the best indices are the BIC, ICL, and CAIC. While for the CN, the best indices are BIC, ICL, CAIC, and AICc. We also noticed that the percentage of times each index picks the correct number of clusters decreases as the number of clusters increases for all methods.

	CN									
	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
Iris	**T**	F	F	**T**	F	F	**T**	F	**T**	F
Faithful	F	**T**	**T**	**T**	**T**	**T**	**T**	**T**	F	**T**
Bankruptcy	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	F
Wine	**T**	F	F	F	F	F	F	F	F	F
Liver	F	**T**	**T**	F	**T**	**T**	F	**T**	**T**	F
Ecoli	F	**T**	**T**	F	**T**	F	F	**T**	F	F
Ruspini	F	**T**	**T**	**T**	F	F	**T**	**T**	**T**	F
Total successes	3	5	5	4	4	3	4	5	4	1
	T									
	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
Iris	F	F	F	**T**	F	F	**T**	F	**T**	F
Faithful	F	**T**	**T**	F	**T**	**T**	F	**T**	F	F
Bankruptcy	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**
Wine	F	F	F	F	F	F	F	F	F	F
Liver	F	**T**	**T**	F	F	**T**	F	**T**	F	F
Ecoli	F	**T**	**T**	F	**T**	F	F	**T**	F	F
Ruspini	F	**T**	**T**	**T**	**T**	F	**T**	**T**	**T**	F
Total successes	1	5	5	3	4	3	3	5	3	1
	MSCN									
	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
Iris	**T**	F	F	F	F	F	F	F	F	F
Faithful	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	F
Bankruptcy	F	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	F
Wine	F	F	F	F	F	F	F	F	F	F
Liver	F	**T**	**T**	F	**T**	**T**	F	**T**	**T**	F
Ecoli	F	F	F	F	F	F	F	F	F	F
Ruspini	**T**	**T**	**T**	**T**	F	F	**T**	F	F	F
Total successes	3	4	4	3	3	3	3	3	3	0
	MST									
	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
Iris	**T**	F	F	**T**	F	F	**T**	F	NA	F
Faithful	F	**T**	**T**	**T**	**T**	**T**	**T**	**T**	NA	F
Bankruptcy	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	NA	F
Wine	F	F	F	F	NA	F	F	F	**T**	F
Liver	F	F	F	F	F	F	F	F	NA	F
Ecoli	**T**	**T**	**T**	**T**	**T**	**T**	**T**	**T**	NA	**T**
Ruspini	F	F	F	F	F	F	F	F	NA	F
Total successes	3	3	3	4	3	3	4	3	1	1

Table 3: The information criteria results in determining the correct number of clusters using CN, T, MSCN, MST

6. Conclusion

This study compared the accuracy of ten popular information criteria in model-based cluster analysis using four robust distributions. Together with the well known multivariate Student-t distribution, the paper studies the contaminated normal distribution and their multiple

Distribution	Sample Size	Number of Variables	Number of Clusters
Student-t	200	10	2/3/4
Contaminated Normal	200	10	2/3/4
GHD	200	10	2/3/4

Table 4: Description of simulated data sets

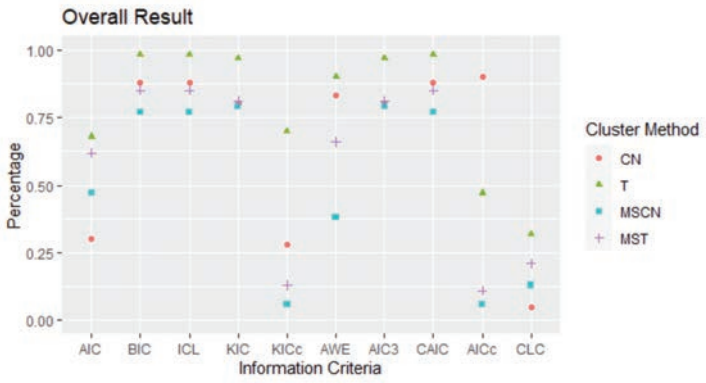

Figure 2: Overall accuracy plot

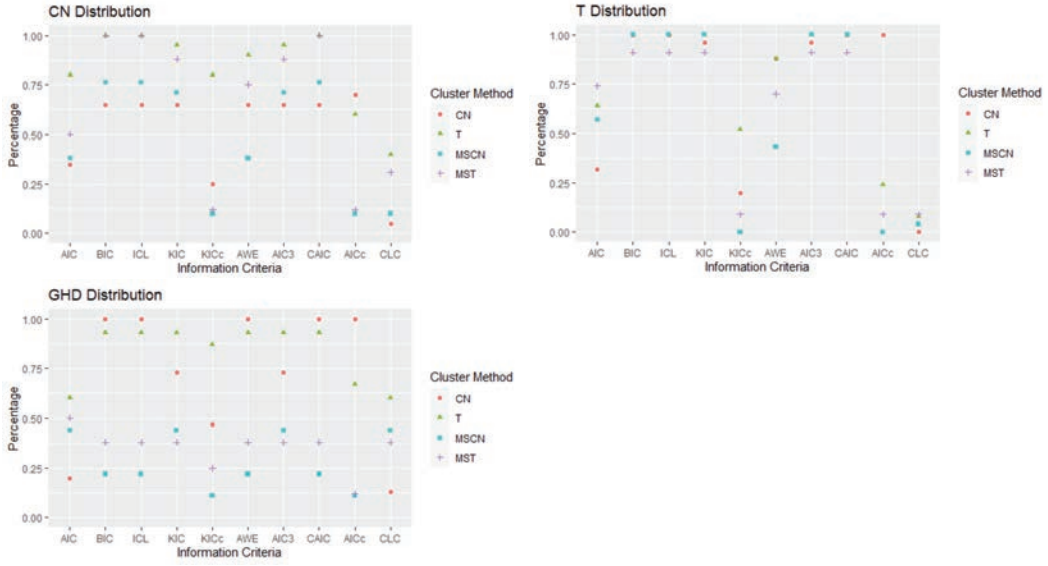

Figure 3: Accuracy plot in different distributions

scaled versions, i.e., multiple scaled Student-t and multiple scaled contaminated normal distributions. BIC, ICL, and CAIC choose the correct number of clusters more frequently when we cluster data using multivariate Student-t distribution. The correspondingly best information criteria while using Contaminated Normal distribution are BIC, ICL, CAIC, and AICc. When clustering using multiple scaled contaminated normal and multiple scaled Student-t distributions, BIC, ICL, KIC, AIC3, and CAIC can be used to obtain a better number of clusters. Since the mentioned indices have similar performance we recommend to use all the selected indices per each model and compare the results when performing cluster analysis. The effect of the number of clusters on the accuracy of the information criteria is also investigated. Interestingly, the more clusters in the data, the fewer times the

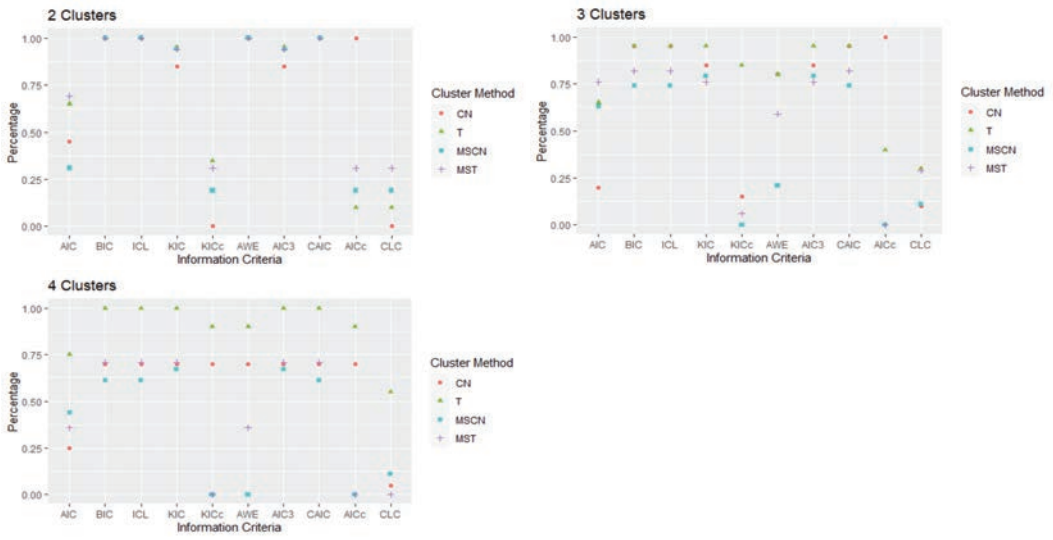

Figure 4: Accuracy plot in different numbers of cluster

information criteria obtain the correct number of clusters.

References

Hans-Hermann Bock. Clustering methods: From classical models to new approaches. *Statistics in Transition*, 5(5):725–758, 2002.

María Teresa Gallegos and Gunter Ritter. Trimmed ML estimation of contaminated mixtures. *Sankhyā: The Indian Journal of Statistics A*, 71(2):164–220, 2009.

Gunter Ritter. *Robust Cluster Analysis and Variable Selection*, volume 137 of *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. CRC Press, 2015.

David Peel and Geoffrey J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

A. Punzo and Paul D. McNicholas. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, 58(6):1506–1537, 2016.

Florence Forbes and Darren Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014.

Antonio Punzo and Cristina Tortora. Multiple scaled contaminated normal distribution and its application in clustering. *Statistical Modelling*, page 1471082X19890935, 2019.

John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

Ole Barndorff-Nielsen, John Kent, and Micheal Sørensen. Normal variance-mean mixtures and z distributions. *International Statistical Review / Revue Internationale de Statistique*, 50(2):145–159, 1982.

Ole Barndorff-Nielsen, John Kent, and Micheal Sørensen. Normal variance-mean mixtures and z distributions. *International Statistical Review / Revue Internationale de Statistique*, 50(2):145–159, 1982.

Tilmann Gneiting. Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation*, 59(4):375–384, 1997.

Chris Fraley. Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Hamparsum Bozdogan. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In *Information and classification*, pages 40–54. Springer, 1993.

Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

Joseph E Cavanaugh. A large-sample model selection criterion based on kullback's symmetric divergence. *Statistics & Probability Letters*, 42(4):333–343, 1999.

A-K Seghouane and Maiza Bekara. A small sample model selection criterion based on kullback's symmetric divergence. *IEEE transactions on signal processing*, 52(12):3314–3323, 2004.

Christophe Biernacki and Gérard Govaert. Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, pages 451–457, 1997.

R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*, 2016.

Jeffrey L Andrews, Jaymeson R Wickins, Nicholas M Boers, and Paul D McNicholas. teigen: An r package for model-based clustering and classification via the multivariate t distribution. *Journal of Statistical Software*, 83(7):1–32, 2018.

Antonio Punzo, Angelo Mazza, and Paul D McNicholas. Contaminatedmixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *arXiv preprint arXiv:1606.03766*, 2016.

Cristina Tortora, Ryan P Browne, Aisha ElSherbiny, Brian C Franczak, and Paul D McNicholas. Model-based clustering, classification, and discriminant analysis using the generalized hyperbolic distribution: Mixghd r package. *Journal of Statistical Software*, 98(1):1–24, 2021.

Paul D. McNicholas, K. Raju Jampani, Aaron F. McDaid, T. Brendan Murphy, and Larry Banks. *pgmm: Parsimonious Gaussian Mixture Models*, 2011. R package version 1.0.

Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2021. URL `https://CRAN.R-project.org/package=cluster`. R package version 2.1.2 — For new features, see the 'Changelog' file (in the package source).

Alan Genz, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl, and Torsten Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2019. R package version 1.0-10.

Weiliang Qiu and Harry Joe. *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*, 2020. URL `https://CRAN.R-project.org/package=clusterGeneration`. R package version 1.3.7.

Appendix: Simulation results

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.30	0.88	0.88	0.80	0.28	0.83	0.80	0.88	0.90*	0.05
T	0.68	0.98*	0.98*	0.97	0.70	0.90	0.97	0.98*	0.47	0.32
MSCN	0.47	0.77	0.77	0.79*	0.06	0.38	0.79*	0.77	0.06	0.13
MST	0.62	0.85*	0.85*	0.81	0.13	0.66	0.81	0.85*	0.11	0.21

Table 5: Overall percentage of times the information criteria correctly specify the number of cluster

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.32	1.00*	1.00*	0.96	0.20	0.88	0.96	1.00*	1.00*	0.00
T	0.64	1.00*	1.00*	1.00*	0.52	0.88	1.00*	1.00*	0.24	0.08
MSCN	0.57	1.00*	1.00*	1.00*	0.00	0.43	1.00*	1.00*	0.00	0.04
MST	0.74	0.91*	0.91*	0.91*	0.09	0.70	0.91*	0.91*	0.09	0.09

Table 6: Percentage of times the information criteria correctly specify the number of cluster for T

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.35	0.65	0.65	0.65	0.25	0.65	0.65	0.65	0.70*	0.05
T	0.80	1.00*	1.00*	0.95	0.80	0.90	0.95	1.00*	0.60	0.40
MSCN	0.38	0.76*	0.76*	0.71	0.10	0.38	0.71	0.76*	0.10	0.10
MST	0.50	1.00*	1.00*	0.88	0.12	0.75	0.88	1.00*	0.12	0.31

Table 7: Percentage of times the information criteria correctly specify the number of cluster for CN

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.20	1.00*	1.00*	0.73	0.47	1.00*	0.73	1.00*	1.00*	0.13
T	0.60	0.93*	0.93*	0.93*	0.87	0.93*	0.93*	0.93*	0.67	0.60
MSCN	0.44*	0.22	0.22	0.44*	0.11	0.22	0.44*	0.22	0.11	0.44*
MST	0.50*	0.38	0.38	0.38	0.25	0.38	0.38	0.38	0.12	0.38

Table 8: Percentage of times the information criteria correctly specify the number of cluster for GHD

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.45	1.00*	1.00*	0.85	0.00	1.00*	0.85	1.00*	1.00*	0.00
T	0.65	1.00*	1.00*	0.95	0.35	1.00*	0.95	1.00*	0.10	0.10
MSCN	0.31	1.00*	1.00*	0.94	0.19	1.00*	0.94	1.00*	0.19	0.19
MST	0.69	1.00*	1.00*	0.94	0.31	1.00*	0.94	1.00*	0.31	0.31

Table 9: Percentage of times the information criteria correctly specify the number of cluster for 2 clusters

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.20	0.95	0.95	0.85	0.15	0.80	0.85	0.95	1.00*	0.10
T	0.65	0.95*	0.95*	0.95*	0.85	0.80	0.95*	0.95*	0.40	0.30
MSCN	0.63	0.74	0.74	0.79*	0.00	0.21	0.79*	0.74	0.00	0.11
MST	0.76	0.82*	0.82*	0.76	0.06	0.59	0.76	0.82*	0.00	0.29

Table 10: Percentage of times the information criteria correctly specify the number of cluster for 3 clusters

	AIC	BIC	ICL	KIC	KICc	AWE	AIC3	CAIC	AICc	CLC
CN	0.25	0.70*	0.70*	0.70*	0.70*	0.70*	0.70*	0.70*	0.70*	0.05
T	0.75	1.00*	1.00*	1.00*	0.90	0.90	1.00*	1.00*	0.90	0.55
MSCN	0.44	0.61	0.61	0.67*	0.00	0.00	0.67*	0.61	0.00	0.11
MST	0.36	0.71*	0.71*	0.71*	0.00	0.36	0.71*	0.71*	0.00	0.00

Table 11: Percentage of times the information criteria correctly specify the number of cluster for 4 clusters

Appendix: Real data results

	2	**3**	4	5
AIC	494.71	460.37*	463.38	470.34
BIC	594.06*	610.90	665.09	723.23
ICL	594.06*	614.11	672.31	736.43
KIC	530.71	513.37*	533.38	557.34
KICc	555.96*	578.82	671.82	826.21
AWE	858.41*	1011.43	1201.81	1396.13
AIC3	527.71	510.37*	530.38	554.34
CAIC	627.06*	660.90	732.09	807.23
AICc	514.05	511.89*	574.50	690.03
CLC	428.72	370.12	347.11	331.72*

Table 12: CN Iris

	2	**3**	4	5
AIC	705.46	670.11	669.41*	676.77
BIC	798.79*	811.61	859.08	914.61
ICL	798.80*	815.10	867.11	930.49
KIC	739.46	720.11*	735.41	758.77
KICc	761.51*	776.54	852.73	980.91
AWE	1047.14*	1198.37	1384.65	1580.97
AIC3	736.46	717.11*	732.41	755.77
CAIC	829.79*	858.61	922.08	993.61
AICc	722.28	714.35*	763.17	857.34
CLC	643.45	565.86	522.50	485.24*

Table 13: T Iris

	2	**3**	4	5
AIC	511.43	489.07*	506.65	497.73
BIC	646.91*	693.79	780.61	840.94
ICL	646.90*	697.01	787.61	844.32
KIC	559.42*	560.06	600.64	614.73
KICc	610.34*	704.24	951.37	1502.71
AWE	1007.38*	1241.26	1514.99	1757.05
AIC3	556.42*	557.06	597.64	611.73
CAIC	691.90*	761.79	871.61	954.94
AICc	551.23*	604.92	795.33	1246.87
CLC	421.42	350.32	319.23	266.83*

Table 14: MSCN Iris

	2	**3**	4	5
AIC	499.64	469.14*	479.86	486.40
BIC	611.03*	637.74	705.66	769.40
ICL	611.04*	639.12	709.28	774.38
KIC	539.64	528.14*	557.86	583.40
KICc	572.16*	614.72	748.17	976.66
AWE	907.43*	1089.09	1313.70	1532.36
AIC3	536.64	525.14*	554.86	580.40
CAIC	648.03*	693.74	780.66	863.40
AICc	Inf	Inf	Inf	Inf
CLC	425.64	354.39	322.62	288.43*

Table 15: MST Iris

	2	3	4	5
AIC	2293.03	2298.53	2307.41	2321.17
BIC	2361.54	2403.10	2448.04	2497.85
ICL	2362.02	2426.46	2529.88	2640.07
KIC	2315.03	2330.53	2349.41	2373.17
KICc	2319.15	2340.03	2366.88	2401.49
AWE	2525.49	2669.63	2839.67	3003.14
AIC3	2312.03	2327.53	2346.41	2370.17
CAIC	2380.54	2432.10	2487.04	2546.85
AICc	2296.05	2305.72	2320.86	2343.24
CLC	2254.59	2223.56	2173.41	2139.56

Table 16: MSCN Faithful

	2	3	4	5
AIC	2286.81	2284.94	2291.89	2301.55
BIC	2340.90	2367.88	2403.67	2442.18
ICL	2341.12	2381.08	2431.88	2488.50
KIC	2304.81	2310.94	2325.89	2343.55
KICc	2307.44	2316.92	2336.77	2361.02
AWE	2470.43	2592.21	2726.87	2870.45
AIC3	2301.81	2307.94	2322.89	2340.55
CAIC	2355.90	2390.88	2434.67	2481.18
AICc	Inf	Inf	Inf	Inf
CLC	2256.37	2212.54	2173.47	2130.91

Table 17: MST Faithful

	2	3	4	5
AIC	2290.53	2284.43	2291.37	2295.19
BIC	2344.61	2367.36	2403.15	2435.82
ICL	2345.13	2400.75	2474.26	2558.48
KIC	2308.53	2310.43	2325.37	2337.19
KICc	2311.15	2316.41	2336.25	2354.66
AWE	2473.70	2565.29	2669.93	2771.45
AIC3	2305.53	2307.43	2322.37	2334.19
CAIC	2359.61	2390.36	2434.15	2474.82
AICc	2292.40	2288.88	2299.64	2308.64
CLC	2261.92	2332.68	2398.14	2456.40

Table 18: CN Faithful

	2	3	4	5
AIC	794.75	791.10	775.91	781.89
BIC	841.62	863.22	873.27	904.49
ICL	842.38	947.11	983.28	1023.25
KIC	810.75	814.10	805.91	818.89
KICc	812.76	818.66	814.14	832.04
AWE	955.42	1191.34	1307.11	1423.09
AIC3	807.75	811.10	802.91	815.89
CAIC	854.62	883.22	900.27	938.49
AICc	796.16	794.45	782.11	791.93
CLC	766.83	595.10	520.42	487.89

Table 19: T Faithful

	2	3
AIC	265.32	257.80
BIC	306.93	321.30
ICL	315.98	327.64
KIC	287.32	289.80
KICc	309.81	352.94
AWE	449.98	534.95
AIC3	284.32	286.80
CAIC	325.93	350.30
AICc	281.85	306.13
CLC	220.88	194.65

Table 20: MSCN Bankruptcy

	2	3
AIC	258.05	256.57
BIC	290.89	306.93
ICL	294.27	310.28
KIC	276.05	282.57
KICc	289.47	317.66
AWE	405.50	479.00
AIC3	273.05	279.57
CAIC	305.89	329.93
AICc	Inf	Inf
CLC	221.28	203.86

Table 21: MST Bankruptcy

	2	3
AIC	272.16	256.68
BIC	305.00	307.05
ICL	310.59	315.09
KIC	290.16	282.68
KICc	303.58	317.78
AWE	412.85	472.41
AIC3	287.16	279.68
CAIC	320.00	330.05
AICc	281.76	282.97
CLC	257.66	232.77

Table 22: CN Bankruptcy

	2	3
AIC	248.69	245.45
BIC	277.16	289.24
ICL	284.33	303.32
KIC	264.69	268.45
KICc	274.66	293.72
AWE	389.26	469.12
AIC3	261.69	265.45
CAIC	290.16	309.24
AICc	255.69	264.11
CLC	204.06	169.35

Table 23: T Bankruptcy, random starts for 3

	2	3	4	5
AIC	10120.24	9600.27	-2166.82	4641.88
BIC	13044.30	13987.95	-1555.70	5406.74
ICL	13044.30	13987.95	-1540.75	5436.00
KIC	11042.24	10982.27	-2004.82	4843.88
KICc	7587.32	6093.92	-1669.82	5501.26
AWE	20563.36	25270.63	-119.67	7225.12
AIC3	11039.24	10979.27	-2007.82	4840.88
CAIC	13963.30	15366.95	-1396.70	5605.74
AICc	7841.32	6433.85	-1891.79	5190.84
CLC	8282.24	6842.27	-2514.72	4185.37

Table 24: MSCN wine

	2	3	4	5
AIC	4640.62	9825.45	4583.73	-1312.16
BIC	4898.14	13955.41	5102.61	-662.60
ICL	4912.38	13955.41	5128.83	-654.00
KIC	4710.62	11126.45	4721.73	-1140.16
KICc	4752.36		4937.80	-741.85
AWE	5519.14	24575.36	6348.92	849.16
AIC3	4707.62	11123.45	4718.73	-1143.16
CAIC	4965.14	15253.41	5237.61	-493.60
AICc	Inf	6817.24	Inf	Inf
CLC	4478.13	7229.45	4261.31	-1667.37

Table 25: MST wine

	2	3	4	5
AIC	10495.56	10393.34	9713.33	10050.11
BIC	13088.71	14284.66	14902.81	16537.77
ICL	13088.71	14284.66	14902.82	16537.77
KIC	11313.56	11619.34	11347.33	12092.11
KICc	8183.72	7215.93	5678.17	5164.41
AWE	19756.86	24290.98	28247.30	33220.42
AIC3	11310.56	11616.34	11344.33	12089.11
CAIC	13903.71	15507.66	16533.81	18576.77
AICc	Inf	Inf	Inf	Inf
CLC	8865.57	7947.34	6451.33	5972.11

Table 26: CN wine

	2	3	4	5
AIC	10479.29	10370.16	10232.60	11842.71
BIC	13066.08	14251.94	15409.37	18775.82
ICL	13066.20	14251.94	15409.37	18775.86
KIC	11295.29	11593.16	11862.60	14024.71
KICc	8171.70		6205.82	6666.47
AWE	19718.32	24233.75	28721.13	36604.20
AIC3	11292.29	11590.16	11859.60	14021.71
CAIC	13879.08	15471.94	17036.37	20954.82
AICc	8398.22	7513.75	6579.15	7097.23
CLC	8852.84	7930.13	6978.60	7097.23
CLC	8716.87			7484.43

Table 27: T wine, random starts for 5

	2	3	4	5
AIC	4643.30	4575.52	4598.41	4636.65
BIC	4946.94	5032.90	5209.53	5401.51
ICL	4962.54	5056.27	5232.88	5426.29
KIC	4725.30	4697.52	4760.41	4838.65
KICc	4785.40	4854.63	5095.41	5496.03
AWE	5676.78	6132.02	6662.35	7210.93
AIC3	4722.30	4694.52	4757.41	4835.65
CAIC	5025.94	5151.90	5368.53	5600.51
AICc	4691.00	4702.45	4873.44	5185.61
CLC	4454.10	4290.78	4233.71	4189.09

Table 28: MSCN liver

	2	3	4	5
AIC	4640.62	4561.08	-2261.78	-2060.82
BIC	4898.14	4949.28	-1742.90	-1411.26
ICL	4912.38	4974.65	-1727.97	-1402.55
KIC	4710.62	4665.08	-2123.78	-1888.82
KICc	4752.36	4770.82	-1907.71	-1490.51
AWE	5519.14	5893.22	-519.16	100.72
AIC3	4707.62	4662.08	-2126.78	-1891.82
CAIC	4965.14	5050.28	-1607.90	-1242.26
AICc	Inf	Inf	Inf	Inf
CLC	4478.13	4308.34	-2561.64	-2416.23

Table 29: MST liver

	2	3	4	5
AIC	14716.60	14630.66	14610.34	14614.45
BIC	14943.37	14972.73	15067.73	15187.14
ICL	14993.17	15042.86	15133.85	15260.30
KIC	14778.60	14722.66	14732.34	14766.45
KICc	14810.29	14801.43	14889.45	15046.79
AWE	15465.14	15759.81	16120.11	16504.83
AIC3	14775.60	14719.66	14729.34	14763.45
CAIC	15002.37	15061.73	15186.73	15336.14
AICc	14741.44	14693.48	14737.28	14843.68
CLC	14705.39	14602.85	14526.87	14474.09

Table 30: CN liver

	2	3	4	5
AIC	4651.89	4569.11	4554.24	4579.38
BIC	4870.81	4899.40	4995.92	5132.43
ICL	4897.81	4975.00	5074.22	5200.02
KIC	4711.89	4658.11	4672.24	4726.38
KICc	4741.43	4731.22	4817.31	4983.50
AWE	5437.40	5816.67	6185.65	6563.90
AIC3	4708.89	4655.11	4669.24	4723.38
CAIC	4927.81	4985.40	5110.92	5276.43
AICc	4675.01	4627.33	4671.26	4789.23
CLC	4475.22	4240.13	4151.19	4132.97

Table 31: T liver, random starts for 4 and 5

	2	3	4	5
AIC	-36624.78	283.49	263.95	294.95
BIC	-36246.89	350.69	354.33	408.51
ICL	-36246.89	350.70	354.34	408.56
KIC	-36522.78	315.49	305.95	346.95
KICc	-36418.12	366.07	419.57	591.83
AWE	-35373.99	562.91	639.72	767.17
AIC3	-36525.78	312.49	302.95	343.95
CAIC	-36147.89	379.69	393.33	457.51
AICc	-36540.88	322.15	353.09	490.95
CLC	-36822.78	225.48	185.94	196.85

Table 32: MSCN ecoli

	2	3	4	5
AIC	-37544.74	-33965.26	-40042.85	-39673.39
BIC	-37220.29	-33476.67	-39390.12	-38856.53
ICL	-37220.29	-33476.45	-39384.60	-38844.03
KIC	-37456.74	-33834.26	-39868.85	-39456.39
KICc	-37383.31	-33637.69	-39434.75	-38554.23
AWE	-36470.83	-32347.63	-37871.34	-36944.68
AIC3	-37459.74	-33837.26	-39871.85	-39459.39
CAIC	-37135.29	-33348.67	-39219.12	-38642.53
AICc	Inf	Inf	Inf	Inf
CLC	-37714.74	-34221.71	-40395.90	-40126.38

Table 33: MST Ecoli

	2	3	4	5
AIC	2967.92	2838.37	2666.20	2614.57
BIC	3139.69	3097.93	3013.55	3049.72
ICL	3147.55	3103.49	3031.54	3062.14
KIC	3015.92	2909.37	2760.20	2731.57
KICc	3034.35	2953.92	2846.12	2878.61
AWE	3536.46	3697.50	3815.91	4054.87
AIC3	3012.92	2906.37	2757.20	2728.57
CAIC	3184.69	3165.93	3104.55	3163.72
AICc	2982.20	2873.52	2734.82	2733.21
CLC	2899.39	2720.61	2527.93	2423.84

Table 34: CN Ecoli after excluding 2 features 4 and 5

	2	3	4	5
AIC	2958.24	2814.17	2653.13	2599.41
BIC	3122.38	3062.28	2985.21	3015.48
ICL	3127.08	3067.12	3000.86	3030.86
KIC	3004.24	2882.17	2743.13	2711.41
KICc	3021.01	2922.53	2820.58	2843.20
AWE	3517.46	3651.95	3793.21	4019.44
AIC3	3001.24	2879.17	2740.13	2708.41
CAIC	3165.38	3127.28	3072.21	3124.48
AICc	2971.20	2845.95	2714.87	2705.52
CLC	2856.30	2667.62	2438.22	2338.52

Table 35: T Ecoli after excluding 2 features 4 and 5

	2	3	4	5
AIC	382.65	295.19	263.95	275.44
BIC	426.68	362.40	354.33	388.99
ICL	426.69	362.40	354.34	389.45
KIC	404.65	327.19	305.95	327.44
KICc	423.46	377.77	419.57	572.31
AWE	565.72	574.61	639.72	748.47
AIC3	401.65	324.19	302.95	324.44
CAIC	445.68	391.40	393.33	437.99
AICc	396.47	333.86	353.09	471.44
CLC	344.65	237.18	185.94	176.51

Table 36: MSCN ruspini

	2	3	4	5
AIC	375.12	283.34	4622.38	256.68
BIC	409.88	336.65	5141.25	347.06
ICL	409.88	336.65	5167.27	347.65
KIC	393.12	309.34	4760.38	298.68
KICc	404.49	338.26	4976.45	412.30
AWE	519.65	504.96	6387.16	633.62
AIC3	390.12	306.34	4757.38	295.68
CAIC	424.88	359.65	5276.25	386.06
AICc	Inf	Inf	Inf	Inf
CLC	345.12	237.33	4300.35	177.50

Table 37: MST ruspini

	2	3	4	5
AIC	321.70	283.19	247.95	242.32
BIC	356.46	336.49	319.79	332.71
ICL	356.46	336.50	319.80	332.72
KIC	339.70	309.19	281.95	284.32
KICc	351.07	338.11	341.95	397.95
AWE	466.22	504.79	546.63	618.09
AIC3	336.70	306.19	278.95	281.32
CAIC	371.46	359.49	350.79	371.71
AICc	329.83	304.84	294.09	331.47
CLC	291.70	237.25	186.01	164.39

Table 38: CN ruspini

	2	3	4	5
AIC	318.01	277.51	240.27	237.21
BIC	348.13	323.86	302.84	316.00
ICL	348.13	323.88	302.86	316.76
KIC	334.01	300.51	270.27	274.21
KICc	342.50	321.57	312.62	350.96
AWE	443.26	470.35	500.56	566.32
AIC3	331.01	297.51	267.27	271.21
CAIC	361.13	343.86	329.84	350.00
AICc	323.97	293.06	272.44	296.71
CLC	292.01	237.36	186.12	167.68

Table 39: T ruspini