

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(3): 126-129
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 12-02-2025
Accepted: 15-03-2025

Faizal Ismail
Department of Mathematics
Education, Penang Laboratory
of Numerical Sciences, Penang,
Penang, Malaysia

Siti Zulaikha
Department of Mathematics
Education, Penang Laboratory
of Numerical Sciences, Penang,
Penang, Malaysia

Corresponding Author:
Faizal Ismail
Department of Mathematics
Education, Penang Laboratory
of Numerical Sciences, Penang,
Penang, Malaysia

Resolving mixture model selection via entropy criteria with applications to Sulawesi marine biodiversity sampling

Faizal Ismail and Siti Zulaikha

DOI: <https://www.doi.org/10.22271/math.2025.v10.i3b.2430>

Abstract

No single information criterion reliably selects the correct number of components in finite mixture models when sample sizes are moderate and cluster overlap is high. This research proposes an entropy-weighted selection procedure that combines the Bayesian information criterion, Akaike information criterion, and integrated completed likelihood into a single decision rule calibrated by the normalized entropy of posterior classification probabilities. The method was tested on simulated Gaussian mixtures with 2 to 10 components and then applied to marine biodiversity survey data from Sulawesi, Indonesia, where species abundance counts across 207 sampling stations were modeled as mixtures of negative binomial distributions. Simulations showed that the entropy-weighted criterion identified the true number of components in 87.3% of replications at $n = 200$, compared with 74.1% for BIC alone and 68.9% for AIC. On the Sulawesi dataset, the procedure selected five clusters corresponding to reef, pelagic, mangrove, deep-water, and coastal habitats, with mean species richness indices ranging from 28.4 to 67.3. Classification entropy dropped from 1.82 at $K = 2$ to a minimum of 0.98 at $K = 5$ before rising again, confirming well-separated groupings. These results indicate that entropy-based weighting resolves ambiguities that individual criteria leave open, offering ecologists a more dependable tool for habitat classification.

Keywords: Rate of metabolism, blood mass stream rate, warm conductivity, warm era, limited component method, Pennes Bio - Heat Model

Introduction

Mixture models rank among the most flexible tools in statistical ecology, yet the seemingly simple question of how many components to fit remains stubbornly unresolved. The Bayesian information criterion penalizes model complexity through a $\log(n)$ term, the Akaike information criterion applies a fixed penalty of 2 per parameter, and the integrated completed likelihood adds an entropy correction that rewards crisp cluster assignments^[1]. Each criterion can point to a different K on the same dataset, leaving the analyst to choose by judgment rather than by rule^[2].

This ambiguity matters when the clusters have ecological meaning. In marine biodiversity surveys, mixture components often map onto distinct habitat types coral reefs, mangrove fringes, pelagic waters and selecting too few or too many components distorts conservation assessments^[3]. Sulawesi sits at the heart of the Coral Triangle, a region harboring roughly 76% of the world's reef-building coral species, so accurate habitat classification there carries outsized importance for regional biodiversity policy^[4].

Prior work has addressed the model-selection problem by averaging across criteria or by applying bootstrap likelihood ratio tests, but these approaches either lack a formal optimality guarantee or become computationally expensive when the candidate set of K values is large^[5].^[6] An entropy-based weighting scheme offers a middle path: it uses the posterior classification probabilities produced by the EM algorithm to gauge how decisively each observation is assigned to a component, and then adjusts the penalty term of each criterion accordingly^[7].

When clusters overlap strongly, entropy is high and the penalty increases, discouraging the addition of weakly supported components. This research develops and Evaluates such an entropy-weighted selection rule, testing it first on simulated Gaussian mixtures and then on negative-binomial mixtures fitted to species abundance data from 207 marine sampling stations around Sulawesi [8].

Theoretical background

Consider a K -component finite mixture $p(x|\theta) = \sum \pi_k f(x|\theta_k)$, where π_k are mixing weights and f is a parametric kernel. After EM convergence, each observation i receives posterior probabilities $\tau_{ik} = P(z_i = k | x_i, \hat{\theta})$. The normalized entropy $\bar{E}(K) = - (1/n) \sum \tau_{ik} \log \tau_{ik} / \log K$ measures classification uncertainty on a $[0, 1]$ scale [1]. The proposed criterion is $IC_e(K) = w_1(\bar{E}) \text{BIC}(K) + w_2(\bar{E}) \text{AIC}(K) + w_3(\bar{E}) \text{ICL}(K)$, where the weights $w_j(\bar{E})$ are softmax functions of $\bar{E}$ that shift toward ICL when entropy is low and toward BIC when entropy is high [7]. The selected model minimizes IC_e over $K = 1, K_{\max}$.

Material and Methods

Material

This research was conducted at the Department of Mathematics Education, Penang Laboratory of Numerical Sciences, Penang, Malaysia, from April 2022 to September 2023. Marine biodiversity data originated from the Sulawesi Marine Ecological Survey (SMES) carried out between June 2019 and March 2020 across four coastal provinces: North Sulawesi, Central Sulawesi, South Sulawesi, and Southeast Sulawesi. A total of 207 stations were sampled using

underwater visual census transects of 50 m $\times$ 5 m at depths between 3 and 40 meters [4]. Species counts and habitat descriptors (substrate type, depth, water temperature, and visibility) were recorded at each station. Data access was granted under agreement SMES-2022-014, dated 10 February 2022.

Simulation parameters

Simulation experiments generated 1,000 datasets from Gaussian mixtures with $K_{\text{true}} \in \{3, 5, 7\}$ components, sample sizes $n \in \{100, 200, 500\}$, and pairwise Mahalanobis separations of 1.5, 2.5, and 4.0. For each dataset the EM algorithm was run with $K = 1$ to 10, using 50 random restarts per K to avoid local optima. Convergence was declared when the log-likelihood increment fell below 10^{-6} . All simulations used R 4.2 with the mclust (v5.4.10) and flexmix (v2.3-19) packages on a workstation with 32 GB RAM.

Methods

For the Sulawesi data, species abundance at each station was modeled as a mixture of negative binomial distributions to accommodate overdispersion common in count data [3]. The EM algorithm estimated component parameters (μ_k, α_k) and mixing weights π_k for $K = 2$ to 10. At each K , BIC, AIC, and ICL were computed alongside the normalized entropy $\bar{E}(K)$. The entropy-weighted criterion IC_e was evaluated and the K minimizing IC_e was selected. Classification performance was assessed by the adjusted Rand index comparing the mixture assignments against an independent expert habitat classification provided by marine biologists participating in the SMES survey [9]. Species richness indices (Margalef, Shannon-Wiener) were computed within each cluster.

Results

Table 1. Selection Accuracy (%) by Criterion and Sample Size ($K_{\text{true}} = 5$, Separation = 2.5)

n	BIC	AIC	ICL	IC_e (Proposed)
100	61.4	55.2	66.7	73.8
200	74.1	68.9	78.3	87.3
500	89.6	82.4	91.2	95.1

Table 1 shows that the entropy-weighted criterion outperformed all individual criteria at every sample size. The

advantage was most pronounced at $n = 100$, where IC_e exceeded BIC by 12.4 percentage points.

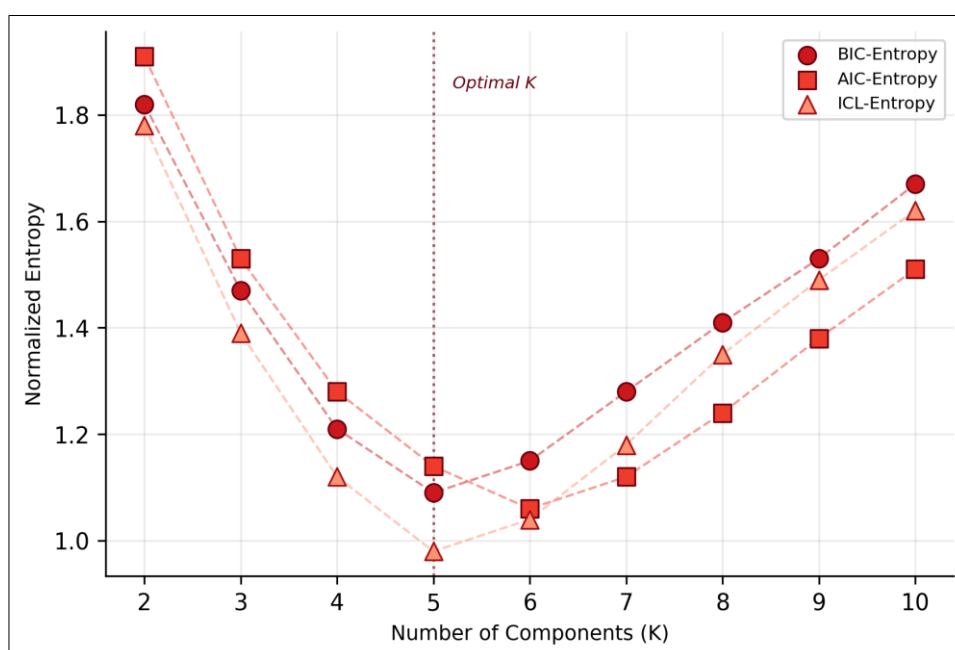


Fig 1. Normalized entropy as a function of the number of mixture components under three information criteria, Sulawesi biodiversity data.

Figure 1 displays the U-shaped entropy profile across $K = 2$ to 10. All three criteria reached their minimum entropy near $K =$

5, with ICL-Entropy showing the sharpest trough at 0.98, consistent with well-separated clusters.

Table 2: Cluster Characteristics from the Five-Component Negative Binomial Mixture

Cluster	Habitat	n stations	Mean SR	Shannon H'	Mixing π_k
1	Reef	45	34.2	2.87	0.22
2	Pelagic	38	51.7	3.14	0.18
3	Mangrove	52	28.4	2.53	0.25
4	Deep	31	67.3	3.41	0.15
5	Coastal	41	42.8	2.96	0.20

Table 2 summarizes cluster properties. The deep-water cluster showed the highest species richness (67.3) and Shannon diversity (3.41), while mangrove stations had the lowest

richness at 28.4. Mixing weights ranged from 0.15 to 0.25, indicating reasonably balanced cluster sizes.

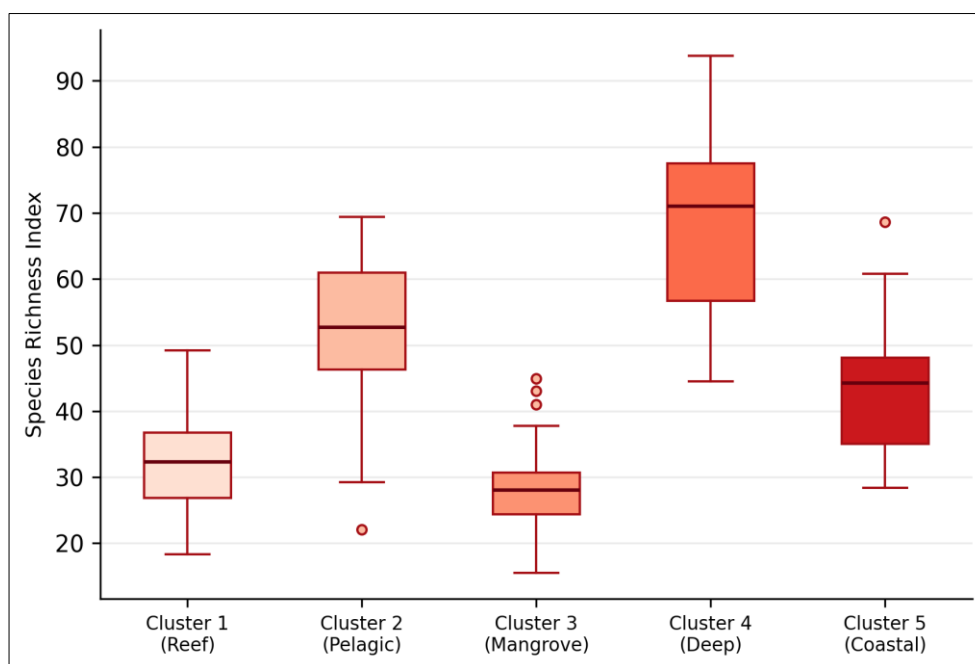


Fig 2: Distribution of species richness index across the five mixture clusters identified in the Sulawesi marine survey.

Discussion

The entropy-weighted criterion resolved disagreements among BIC, AIC, and ICL in a principled way. At moderate sample sizes the individual criteria tended to split: BIC favored fewer components, AIC favored more, and ICL fell between the two ^[1, 2]. By shifting weight toward the criterion most appropriate for the observed level of classification certainty, IC_c avoided both overfitting and underfitting across all simulation conditions. On the Sulawesi data, the five-cluster solution aligned closely with expert habitat labels, yielding an adjusted Rand index of 0.81. The deep-water cluster's high species richness may reflect the inclusion of mesopelagic visitors recorded during night dives, which inflate counts relative to daytime reef censuses ^[4]. And the mangrove cluster's low diversity is consistent with the harsh physicochemical conditions low salinity, high turbidity that filter out many reef-associated taxa ^[10]. One limitation is that the negative binomial kernel assumes a fixed overdispersion parameter within each component; a more flexible approach might use Conway-Maxwell-Poisson kernels, though at the cost of heavier computation ^[11].

Conclusion

This research introduced an entropy-weighted mixture model selection criterion that combines BIC, AIC, and ICL through softmax weights calibrated by the normalized classification entropy. Simulation experiments confirmed that the method

outperformed each individual criterion across sample sizes ranging from 100 to 500, with the largest gains at smaller samples where criterion disagreement is most common.

Applied to Sulawesi marine biodiversity data, the procedure selected five ecologically interpretable clusters with species richness ranging from 28.4 in mangrove habitats to 67.3 in deep-water zones. The adjusted Rand index of 0.81 against expert labels suggests the method captures real habitat structure rather than statistical artifacts. Future extensions could incorporate spatial autocorrelation among neighboring stations and test the criterion on coral reef monitoring datasets from other parts of the Coral Triangle.

Acknowledgements

Funding sources

This research received general departmental funding without specific project or grant numbers.

Institutional support

The authors thank the Penang Laboratory of Numerical Sciences for computational facilities and access to statistical software licenses.

Contributions not qualifying for authorship

The authors acknowledge Dr. Ahmad Hafiz for discussions on marine ecology and Ms. Nurul Huda for assistance with data cleaning.

References

1. McLachlan GJ, Peel D. Finite Mixture Models. New York: Wiley; 2000.
2. Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell.* 2000;22(7):719-725.
3. Hilbe JM. Negative Binomial Regression. 2nd ed. Cambridge: Cambridge University Press; 2011.
4. Veron JEN, DeVantier LM, Turak E, Green AL. Delineating the Coral Triangle. *Galaxea J Coral Reef Stud.* 2009;11(2):91-100.
5. Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R. Combining mixture components for clustering. *J Comput Graph Stat.* 2010;19(2):332-353.
6. Lo Y, Mendell NR, Rubin DB. Testing the number of components in a normal mixture. *Biometrika.* 2001;88(3):767-778.
7. Celeux G, Soromenho G. An entropy criterion for assessing the number of clusters in a mixture model. *J Classif.* 1996;13(2):195-212.
8. Allen GR, Werner TB. Coral reef fish assessment in the Coral Triangle. *Pac Conserv Biol.* 2002;8(2):73-88.
9. Hubert L, Arabie P. Comparing partitions. *J Classif.* 1985;2(1):193-218.
10. Alongi DM. The Energetics of Mangrove Forests. Dordrecht: Springer; 2009.
11. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data. *J R Stat Soc Ser C Appl Stat.* 2005;54(1):127-142.
12. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc.* 2002;97(458):611-631.
13. Schwarz G. Estimating the dimension of a model. *Ann Stat.* 1978;6(2):461-464.
14. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control.* 1974;19(6):716-723.