entropy

MDPI

Article

A Unified Formulation of k-Means, Fuzzy c-Means and Gaussian Mixture Model by the Kolmogorov–Nagumo Average

Osamu Komori [1,*] and Shinto Eguchi [2]

[1] Department of Computer and Information Science, Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino-shi, Tokyo 180-8633, Japan
[2] The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; eguchi@ism.ac.jp
* Correspondence: komori@st.seikei.ac.jp

Abstract: Clustering is a major unsupervised learning algorithm and is widely applied in data mining and statistical data analyses. Typical examples include k-means, fuzzy c-means, and Gaussian mixture models, which are categorized into hard, soft, and model-based clusterings, respectively. We propose a new clustering, called Pareto clustering, based on the Kolmogorov–Nagumo average, which is defined by a survival function of the Pareto distribution. The proposed algorithm incorporates all the aforementioned clusterings plus maximum-entropy clustering. We introduce a probabilistic framework for the proposed method, in which the underlying distribution to give consistency is discussed. We build the minorize-maximization algorithm to estimate the parameters in Pareto clustering. We compare the performance with existing methods in simulation studies and in benchmark dataset analyses to demonstrate its highly practical utilities.

Keywords: k-means; fuzzy-c; Gaussian mixture model; Kolmogorov–Nagumo average; generalized energy function; Pareto distribution

Citation: Komori, O.; Eguchi, S. A Unified Formulation of k-Means, Fuzzy c-Means and Gaussian Mixture Model by the Kolmogorov–Nagumo Average. *Entropy* **2021**, *23*, 518. https://doi.org/10.3390/e23050518

Academic Editor: Udo Von Toussaint

Received: 22 March 2021
Accepted: 19 April 2021
Published: 24 April 2021

1. Introduction

In data analysis or data mining, there are two fundamental types of methodologies: clustering and classification [1]. Clustering, which is categorized as an exploratory paradigm, detects the underlying structure behind the data and grasps the rough image before proceeding to more intensive and comprehensive data analysis [2,3]. On the other hand, classification predicts unknown class labels of test data based on models constructed by training data with known class labels. The former is called supervised learning, while the latter is called unsupervised learning in pattern recognition [4].

Clustering algorithms fall roughly into three categories: hierarchical, partitioning, and mixture model-based algorithms [5]. In hierarchical clustering, each observation is considered as one cluster in the initial setting. Then clusters are merged recursively based on a similarity matrix defined beforehand. The resultant clusters are expressed as a dendrogram. The partition algorithm starts with a fixed number of clusters and searches for the cluster centers to minimize an objective function such as the squared distances between the centers and observations. It finds the centers simultaneously. The model-based algorithm assumes a mixture of probability distributions, which generates the observations and assigns the distributions to one of the mixture components. A Gaussian-mixture distribution-based approach is widely used in this context.

In this paper, we propose a new clustering, called Pareto clustering in the framework of quasilinear modeling [6–8]. It combines the cluster components by the Kolmogorov–Nagumo average [9] in a flexible way. We consider a generalized energy function as an objective function to estimate cluster parameters, which is an extension of the energy function proposed by [10]. The objective function consists of a survival function of the Pareto distribution, which is widely used in extreme value theory [11]. We investigate the consistency of the parameters, resulting in the underlying probability distribution of the

generalized energy function. We find that k-means [12,13] and fuzzy c-means [14] have the underlying probability distributions with singular points at the cluster centers. This fact shows a clear difference from the model-based clustering such as a Gaussian-mixture modeling. Moreover, we show that the quasilinear modeling based on the Kolmogorov–Nagumo average connects k-means, fuzzy c-means, and a Gaussian-mixture modeling using the hyperparameters of the generalized energy function. See [15,16] for the discussion of the relation between k-means and fuzzy c-means.

The paper is organized as follows. In Section 2, we introduce the generalized energy function as the objective function of the Pareto clustering and discuss the consistency of the parameters. Moreover, we show that k-means, fuzzy c-means, and a Gaussian-mixture are all derived from the generalized energy function as special cases. This fact leads to the fact that the parameters can be estimated in a unified manner by the the minorize-maximization (MM) algorithm [17], where the monotone decrease of generalized energy function is guaranteed. In Section 3, we demonstrate the performance of the Pareto clustering based on simulation studies and benchmark datasets and show its practical utilities. We summarize the results of the Pareto clustering and discuss the extensions and applications in various scientific fields.

2. Materials and Methods

2.1. Generalized Energy Function

Let T be a non-negative random variable with a probability density function $f(t)$. The survival function of T is defined as

$$S(t) = P(T > t),\ t \geq 0. \tag{1}$$

Then for d-dimensional random variables $x_1, \ldots, x_n$, we define a generalized energy function to be minimized with respect to a parameter μ for clustering

$$L_S(\mu) \quad = \quad \frac{1}{\tau}\sum_{i=1}^{n} S^{-1}\Big(\frac{1}{K}\sum_{k=1}^{K} S(\tau\|x_i - \mu_k\|^2)\Big), \tag{2}$$

where $\mu = (\mu_1, \ldots, \mu_K)$ is a set of centers and $\tau > 0$ is the shape parameter. If we take $S(t) = \exp(t)$, the function corresponds to the energy function proposed by [10], where τ can be interpreted as the temperature in physics. The formulation in (2) is called the Kolmogorov–Nagumo average [9,18] and is widely applied to bioinformatics, ecology, fisheries, etc. [6,8,19].

In Equation (2), we express an average of probabilities that x_i belongs to the kth cluster over all K clusters using $1/K\sum_{k=1}^{K} S(\|x_i - \mu_k\|^2)$, where $\|x_i - \mu_k\|^2$ is the energy of x_i associated with μ_k. Hence we view $S^{-1}\Big(1/K\sum_{k=1}^{K} S(\|x_i - \mu_k\|^2)\Big)$ as the Kolmogorov–Nagumo average of the energy of x_i with the probabilistic meanings. In effect, we take summation of the Kolmogorov-Nagumo average over the observations $\{x_1, ..., x_n\}$.

Remark 1. *The generalized energy function* (2) *has a relation with the Archimedean copula defined by*

$$1 - S\Big(\sum_{k=1}^{K} S^{-1}(1 - u_k)\Big) \tag{3}$$

for $\{u_k\}_{k=1}^{K}$ *in* $(0,1)$, *cf. [20] for an introductory discussion. In principle, the generalized energy function is a function from a vector of K cluster energy functions to a integrated energy function. The Archimedean copula is that of K marginal cumulative distribution functions to the joint cumulative distribution function. In this way, the generalized energy function expresses an interactive relation for cluster energy functions analogous with the Archimedean copula expressing the correlation among variables.*

We consider an estimator of the generalized energy function as

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}}\, L_S(\mu). \tag{4}$$

If we assume that x_i $(i = 1, \ldots, n)$ is distributed according to a probability density function $p(x, \mu^*)$, the expected generalized energy function is given by

$$\mathbb{L}_S(\mu) = \frac{1}{\tau} \int S^{-1}\Big(\frac{1}{K}\sum_{k=1}^{K} S(\tau\|x - \mu_k\|^2)\Big) p(x, \mu^*) dx. \tag{5}$$

Here we define a function for a set of cluster centers as

$$\mathcal{E}_\mu(x) = \frac{1}{K}\sum_{k=1}^{K} S(\tau\|x - \mu_k\|^2). \tag{6}$$

Thus, we note that

$$\int \mathcal{E}_\mu(x) dx = v_d \mathbb{E}(T^{\frac{d}{2}}), \tag{7}$$

where v_d is a volume constant $2\pi^{d/2}/\{\tau^{1/2} d\Gamma(d/2)\}$ because $\mathbb{E}(T) = \int_0^\infty S(t)dt$ (Appendix A). This property is a key idea in the following discussion.

Lemma 1. *Assume that the survival function $S(t)$ in (1) is convex in t. We define a function G of (μ^*, μ) as*

$$G(\mu^*, \mu) = \int S^{-1}(\mathcal{E}_\mu(x)) f(S^{-1}(\mathcal{E}_{\mu^*}(x))) dx, \tag{8}$$

Then for any μ and μ^*

$$G(\mu^*, \mu) \geq G(\mu^*, \mu^*) \tag{9}$$

with equality if and only if $\mu = \mu^*$.

Proof. We observe that the function $S^{-1}(t)$ is a decreasing function of t given as

$$\frac{\partial S^{-1}(t)}{\partial t} = -\frac{1}{f(S^{-1}(t))} \tag{10}$$

because $S(S^{-1}(t)) = t$ and $(\partial/\partial t)S(t) = -f(t)$. Similarly,

$$\frac{\partial^2 S^{-1}(t)}{\partial t^2} = -\frac{f'(S^{-1}(t))}{\{f(S^{-1}(t))\}^3}, \tag{11}$$

which is positive for all $t \geq 0$ because $(\partial^2/\partial t^2)S(t) = -f'(t) > 0$ from the convexity assumption for $S(t)$. Therefore, $S^{-1}(t)$ is also convex in $t \in (0, 1)$. This leads to

$$\begin{aligned}
G(\mu^*, \mu) - G(\mu^*, \mu^*) &= \int \{S^{-1}(\mathcal{E}_\mu(x)) - S^{-1}(\mathcal{E}_{\mu^*})\} f(S^{-1}(\mathcal{E}_{\mu^*}(x))) dx && (12)\\
&\geq \int \{\mathcal{E}_\mu(x) - \mathcal{E}_{\mu^*}(x)\} \frac{\partial S^{-1}(\mathcal{E}_{\mu^*}(x))}{\partial t} f(S^{-1}(\mathcal{E}_{\mu^*}(x)) dx && (13)\\
&= -\int \{\mathcal{E}_\mu(x) - \mathcal{E}_{\mu^*}(x)\} dx && (14)\\
&= 0. && (15)
\end{aligned}$$

Here Equality in (13) holds if and only if $\mu = \mu^*$ from the convexity for S^{-1}. The Equality (14) is shown by

$$-f(S^{-1}(t))\frac{\partial S^{-1}(t)}{\partial t} = 1 \tag{16}$$

for any $t \geq 0$ as seen in (10). Equality (15) holds due to (7). □

Theorem 1. *If the $p(x, \mu^*)$ has a form such as*

$$p(x, \mu^*) = Z(\mu^*) f(S^{-1}(\mathcal{E}_{\mu^*}(x))), \tag{17}$$

where $Z(\mu^*) > 0$ is a normalizing constant. Then we have

$$\mathbb{L}_S(\mu) \geq \mathbb{L}_S(\mu^*). \tag{18}$$

Proof. Note that

$$\mathbb{L}_S(\mu) - \mathbb{L}_S(\mu^*) \quad = \quad \frac{Z(\mu^*)}{\tau}\{G(\mu^*, \mu) - G(\mu^*, \mu^*)\},$$

which concludes (18) from Lemma 1. □

We note that $\hat{\mu}$ is asymptotically consistent to true parameter μ^* if the probability density function has the form in (17).

2.1.1. Pareto Distribution

Let us consider a generalized Pareto distribution, where the survival function and its inverse function are defined by

$$S(t) = (1 + \beta t)^{-\frac{1}{\beta}} \quad \text{and} \quad S^{-1}(t) = \frac{t^{-\beta} - 1}{\beta},$$

where $\beta > 0$ denotes the shape parameter. Then the generalized energy function is

$$L_{\tau,\beta}(\mu) \quad = \quad \frac{1}{\tau\beta}\sum_{i=1}^{n}\left[\left\{\sum_{k=1}^{K}\frac{1}{K}\{1 + \tau\beta\|x_i - \mu_k\|^2\}^{-\frac{1}{\beta}}\right\}^{-\beta} - 1\right]. \tag{19}$$

If we consider $\beta \to 0$, then

$$\lim_{\beta\to 0} L_{\tau,\beta}(\mu) = -\frac{1}{\tau}\sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K}\frac{1}{K}\exp\left(-\tau\|x_i - \mu_k\|^2\right)\right\}, \tag{20}$$

which is reduced to the energy function proposed by [10]. Hence, we can understand that Rose's clustering (maximum-entropy clustering) is generated by a survival distribution function of an exponential distribution. Then we have

$$\lim_{\tau\to\infty} L_{\tau,\beta}(\mu) = \lim_{\tau\to\infty}\frac{1}{\tau\beta}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\frac{1}{K}\{1 + \tau\beta\|x_i - \mu_k\|^2\}^{-\frac{1}{\beta}}\right\}^{-\beta} \tag{21}$$

$$= \sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\frac{1}{K}\{\|x_i - \mu_k\|^2\}^{-\frac{1}{\beta}}\right\}^{-\beta} \tag{22}$$

The gradient with respect to μ_k is given by

$$\frac{2}{\beta}\sum_{i=1}^{n}\frac{1}{K}\left[\frac{\{\|x_i - \mu_k\|^2\}^{-\frac{1}{\beta}}}{\sum_{\ell=1}^{K}\pi_\ell\{\|x_i - \mu_\ell\|^2\}^{-\frac{1}{\beta}}}\right]^{1+\beta}(x_i - \mu_k), \tag{23}$$

which exactly leads to the estimation equations of fuzzy c-means if we take $\beta = m - 1$ [14]. Furthermore, we have

$$\lim_{\tau\to\infty,\beta\to 0} L_{\tau,\beta}(\mu) \quad = \quad \sum_{i=1}^{n} \min_{1\le k\le K} \|x_i - \mu_k\|^2, \tag{24}$$

which is the loss function of k-means. The corresponding survival function is $\lim_{\beta\to 0}(1+\beta t)^{-\frac{1}{\beta}} = \exp(-t)$. Note that the loss function is directly derived from (2) as

$$\lim_{\tau\to\infty} \frac{1}{\tau} \sum_{i=1}^{n} S^{-1}\Big(\frac{1}{K}\sum_{k=1}^{K} S(\tau\|x_i - \mu_k\|^2)\Big) = \sum_{i=1}^{n} \min_{1\le k\le K} \|x_i - \mu_k\|^2. \tag{25}$$

In addition, we have

$$\lim_{\tau\to 0} \frac{1}{\tau} \sum_{i=1}^{n} S^{-1}\Big(\frac{1}{K}\sum_{k=1}^{K} S(\tau\|x_i - \mu_k\|^2)\Big) = \sum_{i=1}^{n} \frac{1}{K}\sum_{k=1}^{K} \|x_i - \mu_k\|^2 \tag{26}$$

because $S(0) = 1$.

2.1.2. Fréchet Distribution

Next, we consider Fréchet distribution with the survival function defined as

$$S(t) = 1 - \exp(-t^{\gamma}). \tag{27}$$

where $\gamma < 0$ is the shape parameter. The generalized energy function is given by

$$L_{\gamma,\tau}(\mu) \quad = \sum_{i=1}^{n} \Big(-\frac{1}{\tau^{\gamma}} \log\Big[\frac{1}{K}\sum_{k=1}^{K} \exp(-\tau^{\gamma}\|x_i - \mu_k\|^{2\gamma})\Big]\Big)^{\frac{1}{\gamma}}. \tag{28}$$

We find that

$$\begin{aligned} \lim_{\tau\to 0} L_{\gamma,\tau}(\mu) \quad &= \quad \sum_{i=1}^{n} \lim_{\tau\to 0} \Big(-\frac{1}{\tau^{\gamma}} \log\Big[\frac{1}{K}\sum_{k=1}^{K} \exp(-\tau^{\gamma}\|x_i - \mu_k\|^{2\gamma})\Big]\Big)^{\frac{1}{\gamma}} \\ &= \quad \sum_{i=1}^{n} \Big(\min_{1\le k\le K} \|x_i - \mu_k\|^{2\gamma}\Big)^{\frac{1}{\gamma}} \\ &= \quad \sum_{i=1}^{n} \min_{1\le k\le K} \|x_i - \mu_k\|^2. \end{aligned} \tag{29}$$

Hence, this energy function is reduced to that of the *K*-means algorithm as shown in the Pareto distribution case. The estimating equation is given by

$$\frac{\partial}{\partial \mu_k} L_{\gamma,\tau}(\mu) \quad = \quad \sum_{i=1}^{n} \omega_k(x_i, \tau, \gamma)(\mu_k - x_i) = 0, \tag{30}$$

where

$$\begin{aligned} \omega_k(x_i, \tau, \gamma) \quad &= \quad \Big(-\frac{1}{\tau^{\gamma}} \log\Big[\frac{1}{K}\sum_{\ell=1}^{K} \exp(-\tau^{\gamma}\|x_i - \mu_{\ell}\|^{2\gamma})\Big]\Big)^{\frac{1}{\gamma}-1} \\ &\quad \times \frac{\exp(-\tau^{\gamma}\|x_i - \mu_k\|^{2\gamma})}{\sum_{\ell=1}^{K} \exp(-\tau^{\gamma}\|x_i - \mu_{\ell}\|^{2\gamma})} \|x_i - \mu_k\|^{2\gamma-2}. \end{aligned} \tag{31}$$

When we assume the unbiasedness for the estimating function in (30), that is

$$\mathbb{E}\{\omega_k(X, \tau, \gamma)(\mu_k - X)\} = 0, \tag{32}$$

the underlying distribution has a density function proportional to

$$\Big(-\frac{1}{\tau^\gamma} \log M(x,\mu) \Big)^{\frac{\gamma}{1-\gamma}} M(x,\mu). \tag{33}$$

where

$$M(x,\mu) = \frac{1}{K} \sum_{\ell=1}^{K} \exp(-\tau^\gamma \|x-\mu_\ell\|^{2\gamma}). \tag{34}$$

We confirm that

$$\omega_k(x_i,\tau,\gamma) = \begin{cases} 1 & \text{if } \|x_i-\mu_k\|^2 = \min_{1\le\ell\le K} \|x_i-\mu_\ell\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

as τ goes to 0. Then we consider the limit of τ to ∞, which provides

$$\begin{aligned} \lim_{\tau\to\infty} L_{\gamma,\tau}(\mu) &= \textstyle\sum_{i=1}^n \lim_{\tau\to\infty} \Big(-\frac{1}{\tau^\gamma} \log \Big[\frac{1}{K} \sum_{k=1}^K \exp(-\tau^\gamma \|x_i-\mu_k\|^{2\gamma}) \Big] \Big)^{\frac{1}{\gamma}} \\ &= \textstyle\sum_{i=1}^n \Big(\frac{1}{K} \sum_{k=1}^K \|x_i-\mu_k\|^{2\gamma} \Big)^{\frac{1}{\gamma}} \end{aligned} \tag{36}$$

which is equal to (22). This also leads to Fuzzy c-means if we take as $\gamma = 1/(1-m)$ [14].

2.2. Estimation of Variances and Mixing Proportions in Clusters

In stead of the Euclidean distance $\|x_i-\mu_k\|^2$, we consider $\|x_i-\mu_k\|^2_{\Sigma_k^{-1}} = (x_i-\mu_k)^\top \Sigma_k^{-1}(x_i-\mu_k)$ to incorporate the variance structure around μ_k. Bezdek et al. [14] considered a common variance structure $\Sigma_k = \sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})^\top$, where $\bar{x} = 1/n \sum_{i=1}^n x_i$ for $k = 1,\dots,K$. On the other hand, we estimate distinct Σ_k for each μ_k.

For this purpose, we modify the generalized energy function in (2) to allow for a variances $\Sigma_1,\dots,\Sigma_K$ and mixing proportions $\pi_1,\dots,\pi_K$ ($\sum_{k=1}^K \pi_k = 1$ and $\pi_k \ge 0$ for $k = 1,\dots,K$) as

$$L_S(\theta) \quad = \quad \frac{1}{\tau} \sum_{i=1}^n S^{-1}\Big(\sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} S(\tau \|x_i-\mu_k\|^2_{\Sigma_k^{-1}}) \Big), \tag{37}$$

where $\theta = (\mu_k, \Sigma_k, \pi_k)_{k=1}^K$. We assume that $S(t)$ is convex so that the domain of $S^{-1}(t)$ can be extended from $[0,1]$ to $[0,\infty)$ to allow for $|\Sigma_k|^{-\frac{1}{2}}$. The estimator of this modified generalized energy function is given as

$$\hat{\theta} = \operatorname*{argmin}_\theta L_S(\theta), \tag{38}$$

The expected generalized energy function is given by

$$\mathbb{L}_S(\theta) = \frac{1}{\tau} \int S^{-1}\Big(\sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} S(\tau \|x-\mu_k\|^2_{\Sigma_k^{-1}}) \Big) p(x,\theta^*) dx, \tag{39}$$

where $p(x,\theta^*)$ is the underlying probability density function.

For a cumulative distribution function $F(t) = 1-S(t)$ we have $L_S(\theta) = L_F(\theta)$ if and only if $\sum_{k=1}^K \pi_k |\Sigma_k|^{-1/2} = 1$. On the other hand, it always holds that $L_S(\mu) = L_F(\mu)$ for the original generalized energy function in (2).

Similarly to (6), we define

$$\mathcal{E}_\theta(x) = \sum_{k=1}^K \pi_k |\Sigma_k|^{-\frac{1}{2}} S(\tau \|x-\mu_k\|^2_{\Sigma_k^{-1}}), \tag{40}$$

and we notice that $\int \mathcal{E}_\theta(x)dx$ is also independent of μ as

$$\int \mathcal{E}_\theta(x)dx = v_d \mathbb{E}(T^{\frac{d}{2}}). \tag{41}$$

Lemma 2. *Assume that the survival function $S(t)$ in (1) is convex in t. We define a function G of (μ^*, μ) as*

$$G(\theta^*, \theta) = \int S^{-1}(\mathcal{E}_\theta(x)) f(S^{-1}(\mathcal{E}_{\theta^*}(x)))dx. \tag{42}$$

Then for any θ and θ^*

$$G(\theta^*, \theta) \geq G(\theta^*, \theta^*) \tag{43}$$

with equality if and only if $\theta = \theta^*$.

Proof. It is obvious from Lemma 1 and the fact that $\int \mathcal{E}_\theta(x)dx$ is independent of μ. □

From Lemma 2, we can easily show the following theorem regarding $\mathbb{L}_S(\theta)$.

Theorem 2. *If the $p(x, \theta^*)$ has a form such as*

$$p(x, \theta^*) = Z(\theta^*) f(S^{-1}(\mathcal{E}_{\theta^*}(x))), \tag{44}$$

where $Z(\theta^*) > 0$ is a normalizing constant. Then we have

$$\mathbb{L}_S(\theta) \geq \mathbb{L}_S(\theta^*). \tag{45}$$

For the Pareto distribution, we have from (37)

$$\begin{aligned} L_{\tau,\beta}(\theta) &= \frac{1}{\tau\beta} \sum_{i=1}^{n} \left[\left\{ \sum_{k=1}^{K} \pi_k |\Sigma_k|^{-\frac{1}{2}} \{1 + \tau\beta \|x_i - \mu_k\|^2_{\Sigma_k^{-1}}\}^{-\frac{1}{\beta}} \right\}^{-\beta} - 1 \right] & (46) \\ &= \frac{1}{\tau} \sum_{i=1}^{n} \phi\left(\sum_{k=1}^{K} \pi_k w(x_i, \mu_k, \Sigma_k) \right), & (47) \end{aligned}$$

where

$$\begin{aligned} w(x_i, \mu_k, \Sigma_k) &= |\Sigma_k|^{-\frac{1}{2}} \{1 + \tau\beta \|x_i - \mu_k\|^2_{\Sigma_k^{-1}}\}^{-\frac{1}{\beta}} & (48) \\ \phi(t) &= \frac{t^{-\beta} - 1}{\beta}. & (49) \end{aligned}$$

From (44), the underlying probability density function is

$$p_{\tau,\beta}(\theta^*) = Z_{\tau,\beta}(\theta^*) \left\{ \sum_{k=1}^{K} \pi_k^* w(x_i, \mu_k^*, \Sigma_k^*) \right\}^{1+\beta}, \tag{50}$$

where $Z_{\tau,\beta}(\theta^*)$ is a normalizing constant. When $\beta \to 0$, we have

$$\lim_{\beta \to 0} L_{\tau,\beta}(\theta) = -\frac{1}{\tau} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k |\Sigma_k|^{-\frac{1}{2}} \exp(-\tau \|x_i - \mu_k\|^2_{\Sigma_k^{-1}}) \right\}, \tag{51}$$

which is the negative log likelihood function of the normal mixture distributions apart from a constant term $(2\pi)^{-d/2}$ when $\tau = 1/2$.

Similarly, we have the estimation equation of fuzzy c-means allowing for the Mahalanobis distance when $\tau \to \infty$. Moreover, we have k-means with the use of the Maha-

lanobis distance when $\tau \to \infty$ and $\beta \to 0$. For the other extreme cases, we observe that both $\lim_{\beta\to\infty} L_{\tau,\beta}(\theta)$ and $\lim_{\tau\to 0} L_{\tau,\beta}(\theta)$ diverge or converge to 0 depending on the values of π_k and Σ_k $(k = 1, \ldots, K)$. Hence we choose large values for τ and small values for β in the subsequent data analysis.

2.3. Estimating Algorithm

The direct optimization of $L_{\tau,\beta}(\theta)$ in (46) is difficult due to the mixture structure. Thus, we employ the idea of expectation and maximization (EM) algorithm [21] and the minorize-maximization (MM) algorithm [17] similar to [19]. Our proposed clustering method (Pareto clustering) is as follows in Algorithm 1.

Algorithm 1: Pareto clustering

1. Set initial values $(\mu_k^{(0)}, \Sigma_k^{(0)}, \pi_k^{(0)})$ for $k = 1, \ldots, K$.
2. Repeat the following steps for $t = 0, \ldots, T-1$ and $k = 1, \ldots, K$ until convergence.
3. $$q_k^{(t)}(x_i) = \frac{\pi_k^{(t)} w(x_i, \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{\ell=1}^K \pi_k^{(t)} w(x_i, \mu_k^{(t)}, \Sigma_k^{(t)})} \tag{52}$$
$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^n \left\{q_k^{(t)}(x_i)\right\}^{1+\beta} x_i}{\sum_{i=1}^n \left\{q_k^{(t)}(x_i)\right\}^{1+\beta}} \tag{53}$$
$$\Sigma_k^{(t+1)} = \frac{\tau(2-d\beta)\sum_{i=1}^n \left\{q_k^{(t)}(x_i)\right\}^{1+\beta}(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top}{\sum_{i=1}^n \left\{q_k^{(t)}(x_i)\right\}^{1+\beta}} \tag{54}$$
$$\pi_k^{(t+1)} = \frac{\left\{\sum_{i=1}^n \left\{q_k^{(t)}(x_i)\right\}^{1+\beta}\left\{w(x_i, \mu_k^{(t+1)}, \Sigma_k^{(t+1)})\right\}^{-\beta}\right\}^{\frac{1}{1+\beta}}}{\sum_{\ell=1}^K \left\{\sum_{i=1}^n \left\{q_\ell^{(t)}(x_i)\right\}^{1+\beta}\left\{w(x_i, \mu_\ell^{(t+1)}, \Sigma_\ell^{(t+1)})\right\}^{-\beta}\right\}^{\frac{1}{1+\beta}}} \tag{55}$$
4. Output $(\hat{\mu}_k, \hat{\Sigma}_k, \hat{\pi}_k) = (\mu_k^{(T)}, \Sigma_k^{(T)}, \pi_k^{(T)})$ for $k = 1, \ldots, K$.

The initial values $(\mu_k^{(0)}, \Sigma_k^{(0)}, \pi_k^{(0)})$ are determined by the hierarchical clustering in a similar way to the algorithm by [22]. The derivation of the estimating algorithm is as follows. First, we have

$$\begin{aligned} L_{\tau,\beta}(\theta) &= \frac{1}{\tau}\sum_{i=1}^n \phi\Big(\sum_{k=1}^K \pi_k w(x_i, \mu_k, \Sigma_k)\Big) && (56)\\ &= \frac{1}{\tau}\sum_{i=1}^n \phi\Big(\sum_{k=1}^K q_k(x_i)\frac{\pi_k w(x_i, \mu_k, \Sigma_k)}{q_k(x_i)}\Big) && (57)\\ &\le \frac{1}{\tau}\sum_{i=1}^n \sum_{k=1}^K q_k(x_i)\phi\Big(\frac{\pi_k w(x_i, \mu_k, \Sigma_k)}{q_k(x_i)}\Big) && (58)\end{aligned}$$

where $q_k(x_i)$ is a positive weight such as $\sum_{k=1}^K q_k(x_i) = 1$ and $\phi(t)$ is the convex function defined in (49). The equality holds if and only if

$$\frac{\pi_1 w(x_i, \mu_1, \Sigma_1)}{q_1(x_i)} = \cdots = \frac{\pi_K w(x_i, \mu_K, \Sigma_K)}{q_K(x_i)}, \tag{59}$$

which is equivalent to

$$q_k(x_i) = \frac{\pi_k w(x_i, \mu_k, \Sigma_k)}{\sum_{k=1}^K \pi_k w(x_i, \mu_k, \Sigma_k)}, \ (k = 1, \ldots, K). \tag{60}$$

Based on $q_k^{(t)}(x_i)$ in (52), we define

$$Q(\theta|\theta^{(t)}) = \frac{1}{\tau}\sum_{i=1}^{n}\sum_{k=1}^{K} q_k^{(t)}(x_i)\left\{\left(\frac{\pi_k w(x_i,\mu_k,\Sigma_k)}{q_k^{(t)}(x_i)}\right)^{-\beta} - 1\right\}\Big/\beta \tag{61}$$

$$= \frac{1}{\tau\beta}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\{q_k^{(t)}(x_i)\}^{1+\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}\{1+\tau\beta(x_i-\mu_k)^\top\Sigma_k^{-1}(x_i-\mu_k)\} - q_k^{(t)}(x_i)\right] \tag{62}$$

$$= \frac{1}{\tau\beta}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\{q_k^{(t)}(x_i)\}^{1+\beta}\pi_k^{-\beta}\{|V_k^{-1}|^{\frac{\beta}{d\beta-2}} + \tau\beta(x_i-\mu_k)^\top V_k^{-1}(x_i-\mu_k)\} - q_k^{(t)}(x_i)\right], \tag{63}$$

where $V_k^{-1} = |\Sigma_k|^{\beta/2}\Sigma_k^{-1}$. Then we have

$$\frac{\partial}{\partial\mu_k}Q(\theta|\theta^{(t)}) = -2\pi_k^{-\beta}V_k^{-1}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}(x_i-\mu_k) \tag{64}$$

$$= 0 \tag{65}$$

which means that

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}x_i}{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}}. \tag{66}$$

Similarly, we have

$$\frac{\partial}{\partial V_k^{-1}}Q(\theta|\theta^{(t)})\Big|_{\mu_k=\mu_k^{(t+1)}} = \frac{1}{\tau\beta}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\pi_k^{-\beta}\Big\{\frac{\beta}{d\beta-2}|V_k^{-1}|^{\frac{\beta}{d\beta-2}}V_k + \tau\beta(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^\top\Big\} \tag{67}$$

$$= \frac{1}{\tau\beta}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\pi_k^{-\beta}\Big\{\frac{\beta}{d\beta-2}\Sigma_k + \tau\beta(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^\top\Big\} \tag{68}$$

$$= 0, \tag{69}$$

which means that

$$\Sigma_k^{(t+1)} = \frac{\tau(2-d\beta)\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^\top}{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}}. \tag{70}$$

Next we consider

$$R(\theta|\theta^{(t)}) = Q(\theta|\theta^{(t)}) + \lambda\Big(\sum_{k=1}^{K}\pi_k - 1\Big), \tag{71}$$

where λ is a Lagrange multiplier. Then

$$\frac{\partial}{\partial\pi_k}R(\theta|\theta^{(t)})\Big|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)}} = -\frac{1}{\tau}\pi_k^{-\beta-1}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}w(x_i,\mu_k^{(t+1)},\Sigma_k^{(t+1)})^{-\beta} + \lambda \tag{72}$$

$$= 0, \tag{73}$$

which means

$$\pi_k^{(t+1)} = \frac{\{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}w(x_i,\mu_k^{(t+1)},\Sigma_k^{(t+1)})^{-\beta}\}^{\frac{1}{1+\beta}}}{\sum_{\ell=1}^{K}\{\sum_{i=1}^{n}\{q_\ell^{(t)}(x_i)\}^{1+\beta}w(x_i,\mu_\ell^{(t+1)},\Sigma_\ell^{(t+1)})^{-\beta}\}^{\frac{1}{1+\beta}}}. \tag{74}$$

Remark 2. *The generalized energy function in (46) is monotonically decreasing in the estimating algorithm. That is, we have*

$$L_{\tau,\beta}(\theta^{(t+1)}) \leq Q(\theta^{(t+1)}|\theta^{(t)}) \leq Q(\theta^{(t)}|\theta^{(t)}) = L_{\tau,\beta}(\theta^{(t)}). \tag{75}$$

See Appendix B for more details.

Remark 3. *The estimating algorithm of fuzzy c-means by [14] is given as*

$$u_{ik}^{(t)} = \left[\sum_{\ell=1}^{K}\left(\frac{\|x_i-\mu_k^{(t)}\|^2}{\|x_i-\mu_\ell^{(t)}\|^2}\right)^{\frac{1}{m-1}}\right]^{-m}, \tag{76}$$

$$\mu_k^{(t)} = \frac{\sum_{i=1}^n u_{ik}^{(t)} x_i}{\sum_{i=1}^n u_{ik}^{(t)}}, \tag{77}$$

where $\{u_{ik}^{(t)}\}^{1/m}$ *is called the membership function of* x_i *in cluster* k *at the iteration step* t*. These are special cases of (52) and (53) with* $\tau \to \infty$, $\Sigma_k = I$, $\pi_k = 1/K$ *and* $\beta = m-1$*. Hence we observe that the original algorithm of fuzzy c-means can be interpreted as the EM algorithm.*

Remark 4. *In analogy with the membership function of fuzzy c-means by [14], we define* $q_k^{(t)}(x_i)$ *in (52) as a membership function of* x_i *in cluster* k *at the iteration step* t *in Pareto clustering. Hence we estimate cluster* C_k *as*

$$C_k = \{x_i | q_k^{(T)}(x_i) \geq q_\ell^{(T)}(x_i), \ell = 1,\ldots,K,\ i = 1,\ldots,n\}, \tag{78}$$

where $\cup_{k=1}^K C_k = \{x_1,\ldots,x_n\}$.

Remark 5. *In high-dimensional setting* $p \gg 1$*, we consider the ridge regularization for* $\Sigma_k^{(t+1)}$ *as in [23]*

$$\Sigma_k^{(t+1)}(\alpha) = \alpha\Sigma_k^{(t+1)} + (1-\alpha)\hat{\sigma}_k^2 I,$$

where $\alpha = 0.95$ *and* σ_k^2 *is the scalar variance estimated to be the maximum value of the diagonal elements of* $\Sigma_k^{(t+1)}$*. Moreover, we take* $\beta \ll 1$ *to make* $\Sigma_k^{(t+1)}$ *positive definite.*

2.4. Evaluation of Clustering Methods

We compare the performances of k-means, fuzzy c-means, Gaussian mixture modeling (Gaussian), partitioning around medoids (PAM), and Pareto clustering. To implement these methods, we use the `kmeans` function in the `stat` package [24], the `cmeans` function in the `e1071` package [14], the `Mclust` in the `mclust` package [22] and the `pam` function in the `cluster` package [25] in the statistical software `R`, where the default settings are used for each function. In Pareto clustering, $\tau = 0.5$ and $\beta = 1$ are used as the default settings. We assume that the number of clusters K is known and compare the performances as in [26].

2.4.1. Metrics

Cluster C_k $(k = 1,\ldots,K)$ estimated by a clustering method is evaluated by a predefined reference class set D_ℓ $(\ell = 1,\ldots,L)$ such as

$$\text{Precision}(C_k, D_\ell) = \frac{|C_k \cap D_\ell|}{|C_k|}$$

$$\text{Recall}(C_k, D_\ell) = \frac{|C_k \cap D_\ell|}{|D_\ell|},$$

where $\text{Recall}(C_k, D_\ell) = \text{Precision}(D_\ell, C_k)$. $\text{Precision}(C_k, D_\ell)$ counts data points in cluster C_k belonging to class ℓ. Hence $\max_\ell \text{Precision}(C_k, D_\ell)$ represents the purity of cluster C_k regarding the classes. By taking the weighted average, we have

$$\text{Purity} = \sum_{k=1}^{K} \frac{|C_k|}{n} \max_\ell \text{Precision}(C_k, D_\ell),$$

where n is the sample size. $\text{Recall}(C_k, D_\ell)$ counts data points in a class set D_ℓ estimated to be in cluster C_k. Precision and recall correspond to the positive predictive value and sensitivity, respectively [27].

A metric combining precision and recall is proposed by [28] such as

$$\text{F-value} = \sum_{k=1}^{K} \frac{|D_k|}{n} \max_\ell F(D_k, C_\ell),$$

where

$$F(D_k, C_\ell) = \frac{2 \times \text{Recall}(D_k, C_\ell) \times \text{Precision}(D_k, C_\ell)}{\text{Recall}(D_k, C_\ell) + \text{Precision}(D_k, C_\ell)},$$

which is the harmonic mean of $\text{Precision}(D_k, C_\ell)$ and $\text{Recall}(D_k, C_\ell)$, and is called the F-measure [29].

The cluster level similarity between the estimated center $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_K)$ and the reference value (ground truth) of center $\mu^* = (\mu_1^*, \ldots, \mu_K^*)$ is the centroid index (CI) proposed by [26] as

$$\text{CI}(\hat{\mu}, \mu^*) = \max(\text{CI}'(\hat{\mu}, \mu^*), \text{CI}'(\mu^*, \hat{\mu})),$$

where

$$\begin{aligned} \text{CI}'(\hat{\mu}, \mu^*) &= \sum_{k=1}^{K} \text{orphan}(\mu_k^*) \\ \text{orphan}(\mu_k^*) &= \begin{cases} 1 & q_\ell \neq k \ \forall \ell \\ 0 & \text{otherwise} \end{cases} \\ q_\ell &= \underset{1 \leq k \leq K}{\text{argmin}} \|\hat{\mu}_\ell - \mu_k^*\|^2. \end{aligned}$$

Here, q_ℓ indicates the index of the element of the reference center μ^* that is the nearest to $\hat{\mu}_\ell$. The function $\text{orphan}(\mu_k^*)$ indicates whether μ_k^* is an isolated element (orphan) or not, which is *not* nearest to any elements of $\hat{\mu}$. Hence $\text{CI}'(\hat{\mu}, \mu^*)$ indicates the dissimilarity between $\hat{\mu}$ and μ^*. Due to the asymmetry of $\text{CI}'(\hat{\mu}, \mu^*)$ with respect to $\hat{\mu}$ and μ^*, we take the maximum of $\text{CI}'(\hat{\mu}, \mu^*)$ and $\text{CI}'(\mu^*, \hat{\mu})$. Hence $\text{CI}(\hat{\mu}, \mu^*)$ measures how many clusters are differently located among $\hat{\mu}$ and μ^*.

Another metric to measure the similarity between $\hat{\mu}$ and μ^* is defined as the mean squared error (MSE) over the number of clusters K, which is given as

$$\text{MSE} = \frac{1}{K} \sum_{k=1}^{K} ||\hat{\mu}_k - \mu_k^*||^2.$$

Differently from Purity and F-value, MSE can be calculated based on only estimated and reference centers $\hat{\mu}$ and μ^*. This property is useful in a situation where the reference class sets $D_1, \ldots, D_K$ are difficult to determine but μ^* is easily identified. We use MSE in the simulation studies to evaluate the accuracy of $\hat{\mu}$ and Purity and F-value in the analysis of benchmark datasets.

2.4.2. Simulation Studies

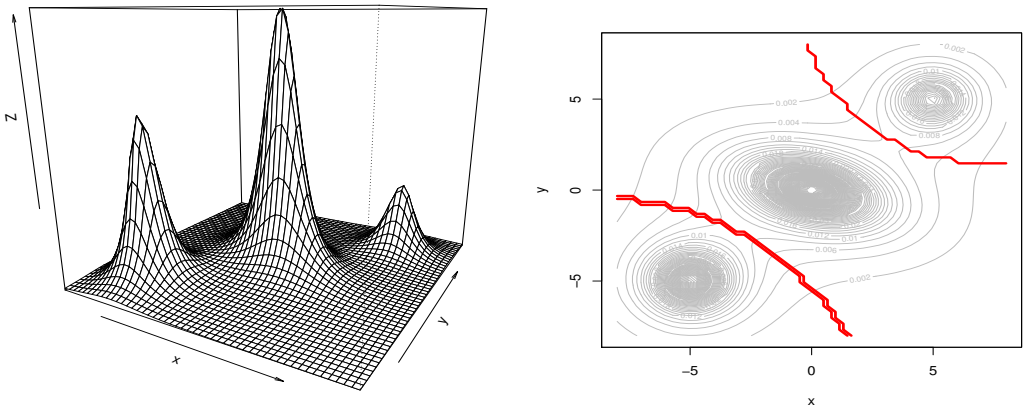
We generate samples according to the density function $p_{\tau,\beta}(\theta^*)$ in (50) using the Metropolis-Hastings algorithm [30,31] as

$$x \sim Z_{\tau,\beta}(\theta^*)\left\{\sum_{k=1}^{K}\pi_k^* w(x_i,\mu_k^*,\Sigma_k^*)\right\}^{1+\beta}, \tag{79}$$

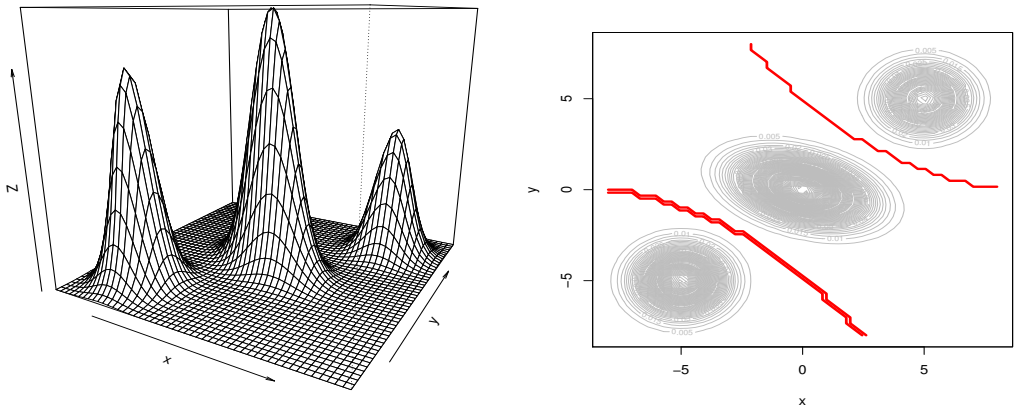
where $\mu_1^* = (0,0)^\top$, $\mu_2^* = (5,5)$, $\mu_3^* = (-5,-5)^\top$, $\pi_1^* = 0.5$, $\pi_2^* = 0.2$, $\pi_3^* = 0.3$ and

$$\Sigma_1^* = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \Sigma_2^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_3^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Figure 1 illustrates the perspective plots and contour plots for $(\tau,\beta) = (0.5,1), (0.5,0)$, (10, 1). The shape of $p_{\tau,\beta}(\theta^*)$ varies according to the values of τ and β. The Gaussian mixture distribution corresponds with $\tau = 0.5$ and $\beta = 0$ in panel (b). When $\beta = 1$, the variance of each component increases and the contours connect with each other. On the other hand, for a large value of $\tau = 10$, the distribution shows high peaks around the centers. This indicates that $p_{\tau,\beta}(\theta^*)$, including fuzzy c-means when $\tau \to \infty$, has a quite different shape from he Gaussian mixture distribution. Other versions of the shapes are also illustrated in Appendix C. The performance of each method is evaluated by MSE based on 100 simulated samples with sample size $n = 3000$.

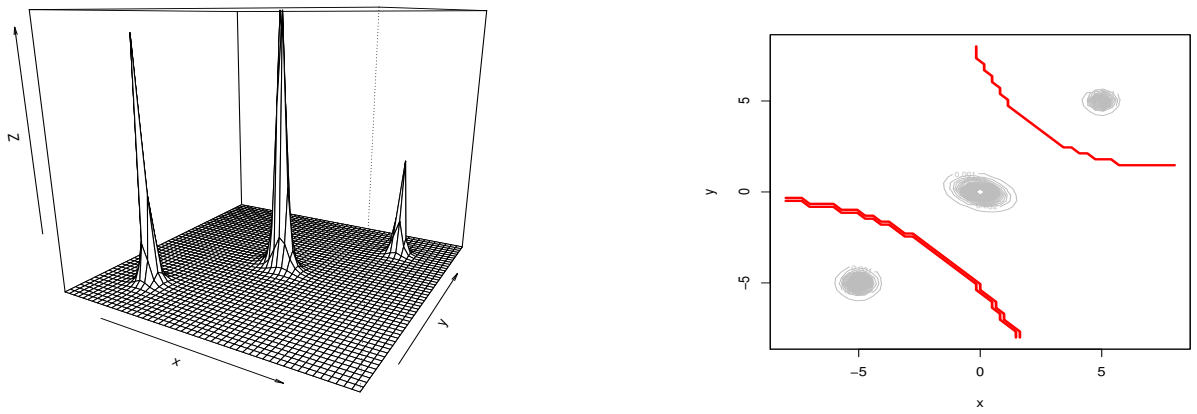


(**a**) $\tau = 0.5$, $\beta = 1$



(**b**) $\tau = 0.5$, $\beta = 0$ (Gaussian mixture distribution)

Figure 1. *Cont.*



(**c**) $\tau = 10$, $\beta = 1$

Figure 1. Perspective plots (left panels) and contour plots for $p_{\tau,\beta}(\theta^*)$ with boundaries marked in red for (**a**) $\tau = 0.5$, $\beta = 1$, (**b**) $\tau = 0.5$, $\beta = 0$ and (**c**) $\tau = 10$, $\beta = 1$.

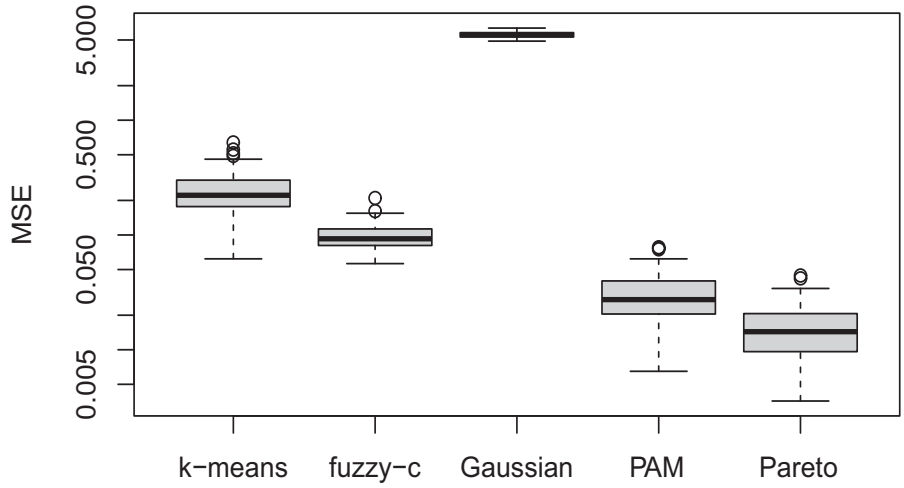
2.4.3. Benchmark Data Analysis

The performance of our proposed method is evaluated using benchmark datasets prepared by [32]. It includes a variety of datasets with low and high cluster overlap, various sample sizes, low and high dimensionalities and unbalanced cluster sizes. Hence, these datasets are suitable for clarifying the statistical performance of the clustering methods. In this setting, we compare the performance of k-means, fuzzy c-means, Gaussian, PAM, and Pareto clustering as well as the variants of Pareto clustering with several values of (τ, β) as explained in Table S1. The characteristics of the benchmark datasets such as the sample sizes, the number of clusters, and dimensionality are summarized in Table S2.

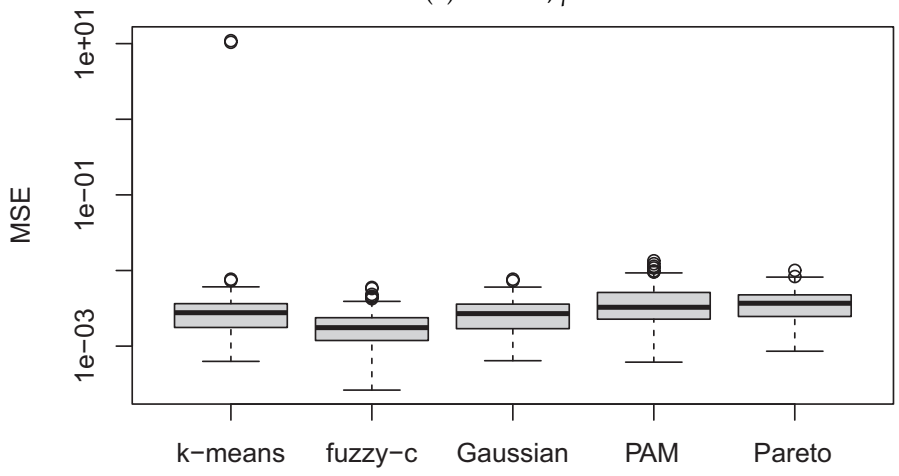
3. Results

Figure 2 illustrates the results of MSE in the simulation studies. Pareto clustering provides the best performance in panel (a), where the samples are generated by the underlying distribution $p_{0.5,1}(\theta^*)$ of Pareto clustering. The shape of the distribution is similar to Gaussian mixture; however, the variance of each component becomes larger and the contour lines are connected to each other as in panel (a) of Figure 1. On the other hand, in panel (c), the variance of each component becomes smaller and contour lines are completely separated. In the both cases, the performances of the Gaussian mixture are clearly degenerated. In the case of panel (b) in which the data are generated from the Gaussian mixture, the performances are comparable to each other, suggesting that k-means, fuzzy c-means, PAM, and Pareto clustering are robust to the underlying distributions to some extent.

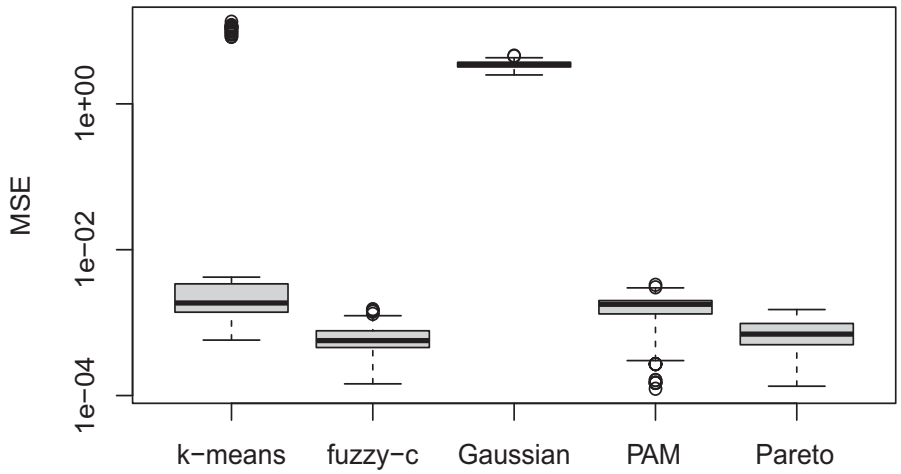
In the benchmark data analysis, metrics of Purity, F-value, and CI are evaluated in Tables S3–S5, where variance Σ_k and mixing proportion π_k $(k = 1, \ldots, K)$ in Pareto clustering are estimated. For the two-dimensional shape datasets such as Flame, Compound, D31, Aggregation, Jain, Pathbased, and Spiral in the upper rows of Table S3, existing methods such as k-means, fuzzy c-means, PAM, and Gaussian mixture outperform our proposed methods. In high-dimensional data with $d = 1024$ (Dim1024), k-means and a Gaussian mixture do not work well. Other methods achieved the maximum value (1) of Purity. For datasets with a large number of clusters, D31 ($K = 31$) and A3 ($K = 50$), PAM performs best. For datasets with large sample size of $n \geq 5000$ and a moderate number of clusters $K = 8, 15$, our proposed method performs best. As for the effect of τ, it barely affects the performance of our proposed method. On the other hand, β slightly affects the performance, resulting that the intermediates among Gaussian mixture, Pareto clustering, k-means, and fuzzy c-means, such as GP, GPKF_1, GPKF_{10}, and GPKF_{100}, show relatively good performances as a whole. We observe similar tendencies regarding the F-value (Table S4).



(**a**) $\tau = 0.5, \beta = 1$



(**b**) $\tau = 0.5, \beta = 0$



(**c**) $\tau = 10, \beta = 1$

Figure 2. Mean squared errors (MSE) on the log scale based on 100 random samples for each method. The samples are generated based on $p_{\tau,\beta}(\theta^*)$ with (**a**) $\tau = 0.5$, $\beta = 1$, (**b**) $\tau = 0.5$, $\beta = 0$ and (**c**) $\tau = 10$, $\beta = 1$.

As for the CI, the values are relatively small for all methods, suggesting that cluster locations are properly estimated. However, some methods do not work for some datasets: Gaussian mixture for A3 (CI = 18), Birch1 (CI = 34) and Birch2 (CI = 49); k-means for D31 (CI = 7), Dim1024 (CI = 4), A3 (CI = 6), Birch1 (CI = 12) and Birch2 (CI = 23); and fuzzy c-means for D31 (CI = 5), A3 (CI = 7), Birch1 (CI = 18) and Birch2 (CI = 25). On the other hand, PAM and our proposed methods show stable results. The results, where Σ_k and π_k are not estimated and fix $\Sigma_k = I$ and $\pi_k = 1/K$ in Pareto clustering, are shown in Tables S6–S8.

4. Discussion

We propose a new clustering method based on the generalized energy function derived from the Kolmogorov–Nagumo average. The survival function used in the generalized energy function plays an important role to ensure the minimum consistency of the parameters, which is shown in Lemma 1 using the property of divergence $G(\mu^*, \mu)$. We consider two examples of the survival function based on the Pareto and Fréchet distributions and show a connection among k-means, fuzzy c-means, and Gaussian mixture, leading to new methods that are intermediates among them. For the underlying distribution of our method in (50), we observe that k-means and fuzzy c-means do not have probabilistic interpretations because the corresponding underlying distributions become singular. We also propose an estimating algorithm for cluster locations, variances, and mixing proportions using the MM algorithm.

Simulation studies and benchmark data analysis show that intermediates among k-means, fuzzy c-means, and the Gaussian mixture perform well. This observation suggests that our proposed method has a wide range of applications in which k-means, fuzzy c-means, and the Gaussian mixture are used. For example, simultaneous deep learning and clustering [33] in which a deep neural network and k-means are jointly used, image segmentation using fuzzy c-means in a deep neural network [34], an application of fuzzy c-means in classification problems [35] and a parallel computation for large datasets by fuzzy c-means [36] can be investigated in the framework of the generalized energy function by the Pareto distribution.

As for the tuning parameters τ and β, we consider an approach using the leave-one-out cross validation in the Supplementary Materials in order to improve the clustering performance. The objective function in the leave-one-out cross validation is derived from an anchor loss as in [37] to estimate the optimal values of τ and β properly. The benchmark data analysis suggests that the performance is insensitive to the values of τ but is sensitive to the values of β. Hence, this approach should be useful to determine the optimal value of β.

Banerjee et al. [38] has proposed a clustering method based on Bregman divergences and clarified the relationship between the exponential families and the corresponding Bregman divergences. They separately consider hard and soft clustering; the former corresponds to k-means style clustering and the latter corresponds to mixture model clustering. In our proposed model, the tuning parameters τ and β bridge the gap between them and the performances are investigated by simulation studies and benchmark datasets. The extension of our method by replacing the squared distance $\|x_i - \mu_k\|^2_{\Sigma_k^{-1}}$ with Bregman divergence should improve its practical flexibility and utility. When β or γ divergence is used, the clustering method should be robust to contamination in observations as suggested by [39,40].

It is well known that MM algorithm and EM algorithm converge to a local optimum and the resultant clusters are sensitive to initial values [41]. One way to circumvent this difficulty is to prepare several sets of initial values and select the best one among them such as the global k-means algorithm [42]. Another approach is to combine MM algorithm and genetic algorithm (GA) to expand thoroughly the search space for the optimal solution [41,43]. The both approach can be incorporated into the Pareto clustering to make it robust to the initial values and to escape from local optimal solutions.

Supplementary Materials: The following are available at https://www.mdpi.com/article/10.3390/e23050518/s1, A: Notations of methods and characteristics of bench-mark datasets, Figure S1: Summary of clustering methods, Table S2: Sample sizes, the number of clusters and dimensions of benchmark datasets. B: Results of benchmark data with $\Sigma_k \neq I$ in Pareto clustering, Table S3: The result of Purity ($\Sigma_k = I$), Table S4: The result of F values ($\Sigma_k \neq I$), Table S5: The result of Centroid index ($\Sigma_k \neq I$). C: Results of benchmark data with $\Sigma_k = I$ in Pareto clustering, Table S6: The result of Purity ($\Sigma_k = I$), Table S7: The result of F values ($\Sigma_k = I$), Table S8: The result of Centroid index ($\Sigma_k = I$). D: Tuning of parameters τ and β. E: R code of the Pareto clustering.

Author Contributions: Conceptualization, S.E.; methodology, S.E. and O.K.; formal analysis, O.K.; writing—original draft preparation, O.K.; All authors have read and agreed to the published version of the manuscript.

Funding: Financial support was provided by the Japan Society for the Promotion of Science KAKENHI Grant Number JP18K11190 and and JP18H03211.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Benchmark datasets are available at http://cs.uef.fi/sipu/datasets/.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivation of the Volume Constant v_d

We define

$$\mathcal{E}_\mu(x) = \frac{1}{K}\sum_{k=1}^{K} S(\tau\|x-\mu_k\|^2). \tag{A1}$$

Then we have

$$\int S(\tau\|x-\mu_k\|^2)dx = \tau^{-\frac{1}{2}}\int S(y^\top y)dy,\ (\because y = \tau^{\frac{1}{2}}(x-\mu_k)) \tag{A2}$$

$$= \tau^{-\frac{1}{2}}\int_0^\infty S(r^2)\frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}r^{d-1}dr,\ (\text{from polar coordinate system}) \tag{A3}$$

$$= \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})\tau^{\frac{1}{2}}}\int_0^\infty S(t)t^{\frac{d-2}{2}}dt,\ (\because t = r^2) \tag{A4}$$

$$= \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})\tau^{\frac{1}{2}}}\left\{\left[\frac{2}{d}t^{\frac{d}{2}}S(t)\right]_0^\infty + \frac{2}{d}\int_0^\infty t^{\frac{d}{2}}f(t)dt\right\},\ (\because S'(t) = -f(t)) \tag{A5}$$

$$= \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})\tau^{\frac{1}{2}}d}E[T^{\frac{d}{2}}] \tag{A6}$$

The last equality holds under a condition that $\lim_{t\to\infty} t^{\frac{d}{2}}S(t) = 0$. Hence we have

$$\int \mathcal{E}_\mu(x)dx = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})\tau^{\frac{1}{2}}d}E[T^{\frac{d}{2}}]. \tag{A7}$$

When $d = 2$, we have

$$\int \mathcal{E}_\mu(x)dx = \frac{\pi}{\tau^{\frac{1}{2}}}E[T]. \tag{A8}$$

Appendix B. Monotone Decrease of the Generalized Energy Function

The Q function at iteration t in the estimating algorithm is defined as

$$Q(\theta|\theta^{(t)}) = \frac{1}{\tau\beta}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\{q_k^{(t)}(x_i)\}^{1+\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}\{1+\tau\beta(x_i-\mu_k)^\top\Sigma_k^{-1}(x_i-\mu_k)\} - q_k^{(t)}(x_i)\right], \tag{A9}$$

where

$$q_k^{(t)}(x_i) = \frac{\pi_k^{(t)}w(x_i,\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{k=1}^{K}\pi_k^{(t)}w(x_i,\mu_k^{(t)},\Sigma_k^{(t)})},\ (k=1,\ldots,K). \tag{A10}$$

The estimate of μ_k is given as

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}x_i}{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}} = \sum_{i=1}^{n}v_k^{(t)}(x_i)x_i,\ (k=1,\ldots,K), \tag{A11}$$

where

$$v_k^{(t)}(x_i) = \frac{\{q_k^{(t)}(x_i)\}^{1+\beta}}{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}}.$$

Here we observe that

$$\sum_{i=1}^{n}v_k^{(t)}(x_i)(x_i-\mu_k)^\top\Sigma_k^{-1}(x_i-\mu_k) \tag{A12}$$

$$=\sum_{i=1}^{n}v_k^{(t)}(x_i)(x_i-\mu_k^{(t+1)})^\top\Sigma_k^{-1}(x_i-\mu_k^{(t+1)}) + \sum_{i=1}^{n}v_k^{(t)}(x_i)(\mu_k^{(t+1)}-\mu_k)^\top\Sigma_k^{-1}(\mu_k^{(t+1)}-\mu_k) \tag{A13}$$

Hence we have

$$Q(\theta|\theta^{(t)}) - Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)}} \tag{A14}$$

$$=\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\{1+\tau\beta(x_i-\mu_k)^\top\Sigma_k^{-1}(x_i-\mu_k)\} \tag{A15}$$

$$-\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\{1+\tau\beta(x_i-\mu_k^{(t+1)})^\top\Sigma_k^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A16}$$

$$=\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}c_q^{(t)}\sum_{i=1}^{n}v_k^{(t)}(x_i)\{(x_i-\mu_k)^\top\Sigma_k^{-1}(x_i-\mu_k) - (x_i-\mu_k^{(t+1)})^\top\Sigma_k^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A17}$$

$$=\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}c_q^{(t)}\sum_{i=1}^{n}v_k^{(t)}(x_i)(\mu_k^{(t+1)}-\mu_k)^\top\Sigma_k^{-1}(\mu_k^{(t+1)}-\mu_k) \geq 0 \tag{A18}$$

where

$$c_q^{(t)} = \sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}. \tag{A19}$$

The estimate of Σ_k is given by

$$\Sigma_k^{(t+1)} = \tau(2-d\beta)\sum_{i=1}^{n}v_k^{(t)}(x_i)(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^\top,$$

where $d\beta < 2$. Hence we have

$$Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)}} - Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)}} \tag{A20}$$

$$=\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\{1+\tau\beta(x_i-\mu_k^{(t+1)})^\top\Sigma_k^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A21}$$

$$-\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k^{(t+1)}|^{\frac{\beta}{2}}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\{1+\tau\beta(x_i-\mu_k^{(t+1)})^\top{\Sigma_k^{(t+1)}}^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A22}$$

$$=\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}c_q^{(t)}\sum_{i=1}^{n}v_k^{(t)}(x_i)\{1+\tau\beta(x_i-\mu_k^{(t+1)})^\top\Sigma_k^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A23}$$

$$-\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k^{(t+1)}|^{\frac{\beta}{2}}c_q^{(t)}\sum_{i=1}^{n}v_k^{(t)}(x_i)\{1+\tau\beta(x_i-\mu_k^{(t+1)})^\top{\Sigma_k^{(t+1)}}^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A24}$$

$$=\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}c_q^{(t)}\left[1+\tau\beta\text{trace}\left\{\Sigma_k^{-1}\sum_{i=1}^{n}v_k^{(t)}(x_i)(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^\top\right\}\right] \tag{A25}$$

$$-\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k^{(t+1)}|^{\frac{\beta}{2}}c_q^{(t)}\left[1+\tau\beta\text{trace}\left\{{\Sigma_k^{(t+1)}}^{-1}\sum_{i=1}^{n}v_k^{(t)}(x_i)(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^\top\right\}\right] \tag{A26}$$

$$=\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k|^{\frac{\beta}{2}}c_q^{(t)}\left[1+\tau\beta\text{trace}\left\{\Sigma_k^{-1}\frac{1}{\tau(2-d\beta)}\Sigma_k^{(t+1)}\right\}\right]-\frac{1}{\tau\beta}\pi_k^{-\beta}|\Sigma_k^{(t+1)}|^{\frac{\beta}{2}}c_q^{(t)}\left[1+\frac{d\beta}{2-d\beta}\right] \tag{A27}$$

Here we notice that

$$Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)}} - Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)}} \geq 0 \tag{A28}$$

$$\Leftrightarrow \quad \frac{|\Sigma_k^{(t+1)}|^{\frac{\beta}{2}\frac{2}{2-d\beta}}}{|\Sigma_k|^{\frac{\beta}{2}}[1+\frac{\beta}{2-d\beta}\text{trace}(\Sigma_k^{-1}\Sigma_k^{(t+1)})]} \leq 1 \tag{A29}$$

$$\Leftrightarrow \quad \frac{2}{2-d\beta} \leq |\Sigma_k^{-1}\Sigma_k^{(t+1)}|^{-\frac{\beta}{2}}\left\{1+\frac{\beta}{2-d\beta}\text{trace}(\Sigma_k^{-1}\Sigma_k^{(t+1)})\right\}, \tag{A30}$$

where $\text{trace}\{\Sigma_k^{-1}\Sigma_k^{(t+1)}\} = \text{trace}\{\Sigma_k^{-1/2}\Sigma_k^{(t+1)}\Sigma_k^{-1/2}\} = \lambda_1 + \ldots + \lambda_d \geq 0$ with λ_j being the non-negative eigen value of $\Sigma_k^{-1/2}\Sigma_k^{(t+1)}\Sigma_k^{-1/2}$. Here we have

$$|\Sigma_k^{-1}\Sigma_k^{(t+1)}|^{-\frac{\beta}{2}}\left\{1+\frac{\beta}{2-d\beta}\text{trace}(\Sigma_k^{-1}\Sigma_k^{(t+1)})\right\} \tag{A31}$$

$$= \quad (\lambda_1\cdots\lambda_d)^{-\frac{\beta}{2}}\left\{1+\frac{\beta}{2-d\beta}(\lambda_1+\cdots+\lambda_d)\right\} \tag{A32}$$

$$\geq \quad \left(\frac{\lambda_1+\cdots+\lambda_d}{p}\right)^{-\frac{d\beta}{2}}\left\{1+\frac{\beta}{2-d\beta}(\lambda_1+\cdots+\lambda_d)\right\}, \ (\because \beta>0) \tag{A33}$$

$$= \quad \Lambda^{-\frac{d\beta}{2}}p^{\frac{d\beta}{2}}\left(1+\frac{\beta}{2-d\beta}\Lambda\right), \tag{A34}$$

where $\Lambda = \lambda_1 + \cdots + \lambda_d \geq 0$. The last term attains the minimum at $\Lambda = d$, leading to

$$\frac{2}{2-d\beta} \leq |\Sigma_k^{-1}\Sigma_k^{(t+1)}|^{-\frac{\beta}{2}}\left\{1+\frac{\beta}{2-d\beta}\text{trace}(\Sigma_k^{-1}\Sigma_k^{(t+1)})\right\}. \tag{A35}$$

As for π_k we define

$$R(\theta|\theta^{(t)}) = Q(\theta|\theta^{(t)}) + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

and we have

$$\frac{\partial R(\theta|\theta^{(t)})}{\partial \pi_k}\Big|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)},\pi_k=\pi_k^{(t+1)}} = 0 \tag{A36}$$

$$\frac{\partial^2 R(\theta|\theta^{(t)})}{\partial \pi_k^2}\Big|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)}} = \frac{1+\beta}{\tau}\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}\pi_k^{-\beta-2}|\Sigma_k^{(t+1)}|^{\frac{\beta}{2}} \times \{1+\tau\beta(x_i-\mu_k^{(t+1)})^\top {\Sigma_k^{(t+1)}}^{-1}(x_i-\mu_k^{(t+1)})\} \tag{A37}$$

$$\geq 0, \tag{A38}$$

where

$$\pi_k^{(t+1)} = \frac{\{\sum_{i=1}^{n}\{q_k^{(t)}(x_i)\}^{1+\beta}w(x_i,\mu_k^{(t+1)},\Sigma_k^{(t+1)})^{-\beta}\}^{\frac{1}{1+\beta}}}{\sum_{\ell=1}^{K}\{\sum_{i=1}^{n}\{q_\ell^{(t)}(x_i)\}^{1+\beta}w(x_i,\mu_\ell,\Sigma_\ell^{(t+1)})^{-\beta}\}^{\frac{1}{1+\beta}}}. \tag{A39}$$

Hence, we have

$$Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)}} - Q(\theta|\theta^{(t)})|_{\mu_k=\mu_k^{(t+1)},\Sigma_k=\Sigma_k^{(t+1)},\pi_k=\pi_k^{(t+1)}} \geq 0. \tag{A40}$$

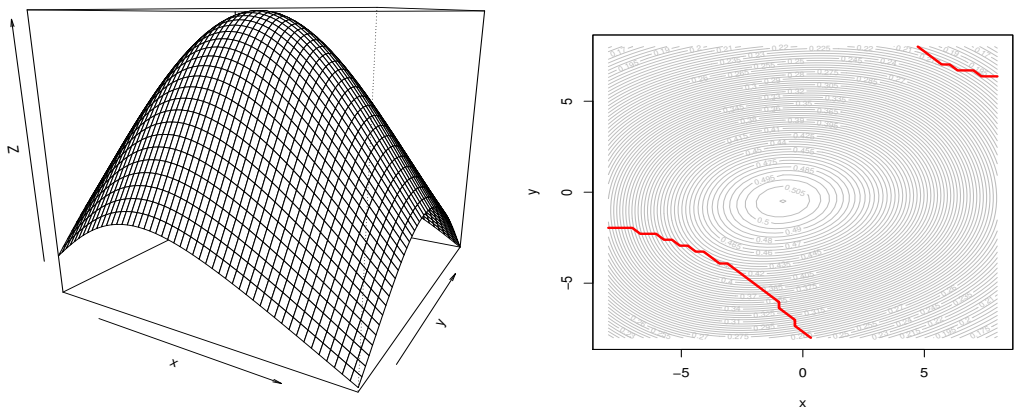
Inequations (A18), (A35) and (A40) hold for $k = 1, \ldots, K$. As a result, we have

$$Q(\theta|\theta^{(t)}) - Q(\theta^{(t+1)}|\theta^{(t)}) \geq 0, \ \forall\theta \tag{A41}$$

Hence we have

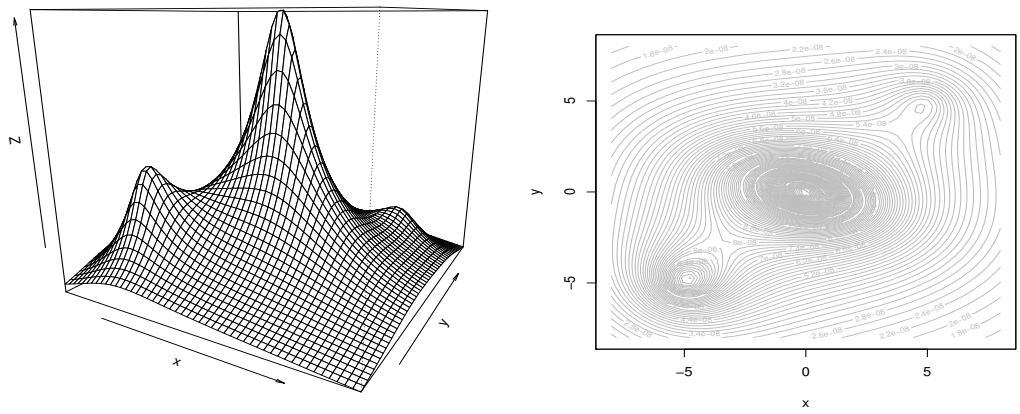
$$L_{\tau,\beta}(\theta^{(t+1)}) \leq Q(\theta^{(t+1)}|\theta^{(t)}) \leq Q(\theta^{(t)}|\theta^{(t)}) = L_{\tau,\beta}(\theta^{(t)}). \tag{A42}$$

Appendix C. Perspective Plots and Contour Plots for $p_{\tau,\beta}(\theta^*)$

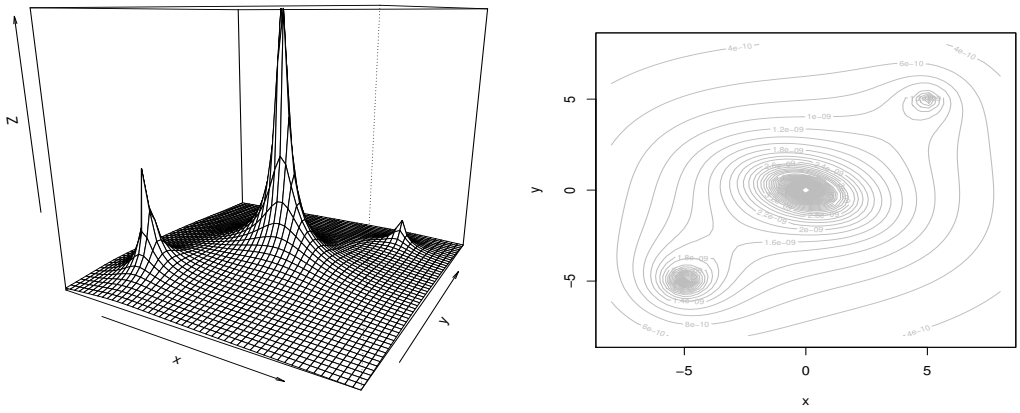


(**a**) $\tau = 0.01, \beta = 1$

Figure A1. *Cont.*



(**b**) $\tau = 0.01$, $\beta = 100$



(**c**) $\tau = 0.5$, $\beta = 100$

Figure A1. Perspective plots (left panels) and contour plots (right panels) for $p_{\tau,\beta}(\theta^*)$.

References

1. Rokach, L.; Maimon, O. Clustering Methods. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2005.
2. Tukey, J.W. We need both exploratory and confirmatory. *Am. Stat.* **1980**, *314*, 23–25.
3. Dubes, R.; Jain, A.K. Clustering methodologies in exploratory data analysis. *Adv. Comput.* **1980**, *19*, 113–228.
4. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
5. Ghosh, S.; Dubey, S. Comparative analysis of k-means and fuzzy c-means algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2013**, *4*, 35–39. [CrossRef]
6. Komori, O.; Eguchi, S.; Ikeda, S.; Okamura, H.; Ichinokawa, M.; Nakayama, S. An asymmetric logistic regression model for ecological data. *Methods Ecol. Evol.* **2016**, *7*, 249–260. [CrossRef]
7. Komori, O.; Eguchi, S.; Saigusa, Y.; Okamura, H.; Ichinokawa, M. Robust bias correction model for estimation of global trend in marine populations. *Ecosphere* **2017**, *8*, 1–9. [CrossRef]
8. Omae, K.; Komori, O.; Eguchi, S. Quasi-linear score for capturing heterogeneous structure in biomarkers. *BMC Bioinform.* **2017**, *18*, 308. [CrossRef] [PubMed]
9. Naudts, J. *Generalised Thermostatistics*; Springer: London, UK, 2011.
10. Rose, K.; Gurewitz, E.; Fox, G.C. Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* **1990**, *65*, 945–948. [CrossRef]
11. Beirlant, J.; Goegebeur, Y.; Segers, J.; Teugels, J.L.; Waal, D.D.; Ferro, C. *Statistics of Extremes: Theory and Applications*; Wiley: Hoboken, NJ, USA, 2004.
12. Cox, D.R. Note on grouping. *J. Am. Stat. Assoc.* **1957**, *52*, 543–547. [CrossRef]

13. MacQueen, J. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; Cam, L.M.L., Neyman, J., Eds.; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
14. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **1984**, *10*, 191–2003. [CrossRef]
15. Hathaway, R.J.; Bezdek, J.C. Optimization of clustering criteria by reformulation. *IEEE Trans. Fuzzy Syst.* **1995**, *3*, 241–245. [CrossRef]
16. Yu, J. General C-means clustering model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1197–1211. [PubMed]
17. Hunter, D.R.; Lange, K. A tutorial on MM algorithms. *Am. Stat.* **2004**, *58*, 30–37. [CrossRef]
18. Eguchi, S.; Komori, O. Path Connectedness on a Space of Probability Density Functions. In *Geometric Science of Information: Second International Conference, GSI 2015*; Nielsen, F., Barbaresco, F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; p. 615.
19. Komori, O.; Eguchi, S.; Saigusa, Y.; Kusumoto, B.; Kubota, Y. Sampling bias correction in species distribution models by quasi-linear Poisson point process. *Ecol. Inform.* **2020**, *55*, 1–11. [CrossRef]
20. Nelsen, R.B. *An Introduction to Copulas*; Springer: New York, NY, USA, 2006.
21. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser.* **1977**, *39*, 1–38.
22. Scrucca, L.; Fop, M.; Murphy, T.B.; Raftery, A.E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **2016**, *8*, 289–317. [CrossRef]
23. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
24. Hartigan, J.A.; Wong, M.A. A k-means clustering algorithm. *J. R. Stat. Soc. Ser.* **1979**, *28*, 100–108.
25. Reynolds, A.P.; Richards, G.; de la Iglesia, B.; Rayward-Smith, V.J. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *J. Math. Model. Algorithms* **2006**, *5*, 475–504. [CrossRef]
26. Fränti, P.; Rezaei, M.; Zhao, Q. Centroid index: Cluster level similarity measure. *Pattern Recognit.* **2014**, *47*, 3034–3045. [CrossRef]
27. Sofaer, H.R.; Hoeting, J.A.; Jarnevich, C.S. The area under the precision-recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* **2019**, *10*, 565–577. [CrossRef]
28. Amigó, E.; Gonzalo, J.; Artiles, J.; Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.* **2009**, *12*, 461–486. [CrossRef]
29. Van Rijsbergen, C. Foundation of evaluation. *J. Doc.* **1974**, *30*, 365–373. [CrossRef]
30. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109. [CrossRef]
31. Chib, S.; Greenberg, E. Understanding the Metropolis-Hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
32. Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets. *Appl. Intell.* **2018**, *48*, 4743–4759. [CrossRef]
33. Yang, B.; Fu, X.; Sidiropoulos, N.D.; Hong, M. Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 7–9 August 2017; Precup, D., Teh, Y.W., Eds.; 2017; Volume 70, pp. 3861–3870.
34. Mohsen, H.; El-Dahshan, E.S.A.; El-Horbaty, E.S.M.; Salem, A.B.M. Classification using deep learning neural networks for brain tumors. *Future Comput. Inform. J.* **2018**, *3*, 68–71. [CrossRef]
35. Gorsevski, P.V.; Gessler, P.E.; Jankowski, P. Integrating a fuzzy k-means classification and a Bayesian approach for spatial prediction of landslide hazard. *J. Geogr. Syst.* **2003**, *5*, 223–251. [CrossRef]
36. Kwok, T.; Smith, K.; Lozano, S.; Taniar, D. Parallel Fuzzy c- Means Clustering for Large Data Sets. In *Euro-Par 2002 Parallel Processing*; Monien, B., Feldmann, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2002; pp. 365–374.
37. Mollah, M.N.H.; Eguchi, S.; Minami, M. Robust Prewhitening for ICA by Minimizing β-Divergence and Its Application to FastICA. *Neural Process. Lett.* **2007**, *25*, 91–110. [CrossRef]
38. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman Divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
39. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [CrossRef]
40. Notsu, A.; Eguchi, S. Robust clustering method in the presence of scattered observations. *Neural Comput.* **2016**, *28*, 1141–1162. [CrossRef] [PubMed]
41. Pernkopf, F.; Bouchaffra, D. Genetic-based EM algorithm for learning Gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1344–1348. [CrossRef] [PubMed]
42. Likas, A.; Vlassis, N.; Verbeek, J.J. The global k-means clustering algorithm. *Pattern Recognit.* **2003**, *36*, 451–461. [CrossRef]
43. Krishna, K.; Murty, M.N. Genetic K-means algorithm. *IEEE Trans. Syst. Man Cybern. Part (Cybern.)* **1999**, *29*, 433–439. [CrossRef] [PubMed]