Spontaneous Clustering via Minimum γ-divergence

Akifumi Notsu

notsu@ism.ac.jp

Department of Statistical Science, The Graduate University for Advanced Studies, Tachikawa, Tokyo 190-8562, Japan

Osamu Komori

komori@ism.ac.jp

The Institute of Statistical Mathematics, Tachikawa, Tokyo 190-8562, Japan

Shinto Eguchi

eguchi@ism.ac.jp

The Institute of Statistical Mathematics and The Graduate University for Advanced Studies, Tachikawa, Tokyo 190-8562, Japan

Keywords: cluster analysis, clustering, divergence, γ-divergence, power entropy

Abstract

We propose a new method for clustering based on the local minimization of the γ-divergence, which we call the spontaneous clustering. The greatest advantage of the proposed method is that it automatically detects the number of clusters that adequately reflect the data structure. In contrast, exiting methods such as K-means, fuzzy c-means, and model based clustering need to prescribe the number of clusters. We detect all the local minimum points of the γ-divergence, which are defined as the centers of clusters. A necessary and sufficient condition for the γ-divergence to have the local minimum points is also derived in a simple setting. A simulation study and a real data analysis are performed to compare our proposal with existing methods.

1 Introduction

Cluster analysis is a common procedure for grouping similar objects in unsupervised learning (Jain et al., 1999; Xu and Wunsch, 2005; Hastie et al., 2009). The procedure stably produces a classification, and is frequently used as a preprocessing before supervised learning. Cluster analysis has wide applications over many disciplines in exploratory data analysis. See, for example, Jin et al. (2011) and Wu et al. (2011) for recent developments. There are mainly two approaches in cluster analysis. One is the hierarchical approach which describes a tree structure called dendrogram. The other is the approach of data space partition such as K-means algorithm. This paper focuses on

the latter approach from a view point of statistical pattern recognition.

We propose what we call the spontaneous clustering. It starts with finding centers of clusters in a data set. For this purpose, we employ a loss function derived from the power entropy with the power index γ. It is referred to the γ-loss function (Fujisawa and Eguchi, 2008; Eguchi and Kato, 2010). Here is a motivational example for the proposal of the spontaneous clustering. Consider the problem of estimating Gaussian mean parameter μ. The maximum likelihood estimator (MLE) of μ is given by the arithmetic mean of the data set as the unique maximum point of the log likelihood function. It is known that the MLE poorly behaves in various situations where Gaussianity assumption is inappropriate. For example, the log likelihood function suggests rather a misleading summary as seen in panel (a) of Figure 1. Alternatively, the γ-loss function properly reflects the data shape. For the same data set in panel (a) of Figure 1, panel (b) shows that the γ-loss function has two local minimum points corresponding to the two normal distributions. We will propose to determine the centers of clusters by such local minimum points.

Almost all procedures via data space partition need the number of clusters a priori. The selection of the number of clusters is a major challenge in cluster analysis. A lot of methods have been proposed in the literature (Xu and Wunsch, 2005). Our clustering method can find the number of clusters automatically as long as the value of γ is properly fixed. The name of the spontaneous clustering comes from this property. Instead of the number of clusters, the value of power index γ should be determined. We will propose two methods to accomplish this aim. One is a heuristic choice of γ that merely relies on the range of the data, and the other is a more sophisticated method based on

Akaike Information Criterion (AIC).

This paper is organized as follows. Section 2 describes the algorithm of the spontaneous clustering and selection procedure of the value of γ. In section 3 the existence of the local minimum points is discussed. Section 4 investigates the numerical properties of the spontaneous clustering. In section 5 a real data analysis is given. Further a discussion is presented in section 6.

2 Spontaneous Clustering

We begin with a statistical formulation of cluster analysis. Suppose the p-dimensional density function of the population distribution is given by

$$g(x) = \sum_{k=1}^{K} \tau_k f_k(x),\ \sum_{k=1}^{K} \tau_k = 1,\ \tau_k > 0,\ k = 1, \ldots, K, \tag{1}$$

where $f_k(x)$ is a density function. Let $\{x_1, \ldots, x_n\}$ be a data set generated from g. We apply the γ-estimation method to this data set. The γ-loss function for the normal distribution with the identity covariance matrix is given by

$$L_\gamma(\mu) = -\frac{1}{n} \sum_{i=1}^{n} \exp\left(-\frac{\gamma}{2}\|x_i - \mu\|^2\right), \tag{2}$$

apart from a constant, where μ and $\|\cdot\|$ denote the mean vector and the Euclidean norm, respectively. In the remainder of the paper, we omit a constant term that does not affect the optimization. In panel (b) of Figure 1, $L_\gamma(\mu)$ is illustrated. See appendix B for a general introduction to the γ-loss function. It is expected that the γ-loss function $L_\gamma(\mu)$ has K local minimum points corresponding to K mean vectors with respect to $f_1, \ldots, f_K$. Then we expect that the local minimum points can help us to define

the centers of K clusters and to build K clusters in a similar way to the K-means algorithm. The covariance structure of the data set is taken into consideration in a subsequent discussion.

2.1 γ-loss Function for the Normal Distribution

We consider the γ-loss function for the normal distribution with mean vector μ and covariance matrix Σ,

$$L_\gamma(\mu, \Sigma) = -\det \Sigma^{-\frac{\gamma}{2(1+\gamma)}} \sum_{i=1}^{n} \exp\left(-\frac{\gamma}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu)\right).$$

An iteration algorithm to find the local minimum points of $L_\gamma(\mu, \Sigma)$ is proposed in Fujisawa and Eguchi (2008) and Eguchi and Kato (2010). It is obtained by differentiating $L_\gamma(\mu, \Sigma)$ with respect to μ and Σ^{-1} and setting the derivatives to 0. The algorithm is a concave-convex procedure (CCCP) (Yuille and Rangarajan, 2003), so that it is guaranteed to decrease the γ-loss function monotonically as the iteration step t increases. It is described as follows.

Step 1 Set appropriate μ_0 and Σ_0 as initial values.

Step 2 Given μ_t and Σ_t, calculate μ_{t+1} and Σ_{t+1} by the following update formula,

$$\mu_{t+1} = \sum_{i=1}^{n} w_\gamma(x_i, \mu_t, \Sigma_t) x_i, \tag{3}$$

$$\Sigma_{t+1} = (1+\gamma) \sum_{i=1}^{n} w_\gamma(x_i, \mu_t, \Sigma_t)(x_i - \mu_{t+1})(x_i - \mu_{t+1})^\top, \tag{4}$$

where

$$w_\gamma(x, \mu, \Sigma) = \frac{\exp\left(-\frac{\gamma}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{\sum_{j=1}^{n} \exp\left(-\frac{\gamma}{2}(x_j-\mu)^\top \Sigma^{-1}(x_j-\mu)\right)}.$$

Step 3 For a sufficiently small number ε, repeat Step 2 while

$$\|\mu_{t+1} - \mu_t\| + \|\Sigma_{t+1} - \Sigma_t\|_{\mathrm{F}} < \varepsilon,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

If $\gamma = 0$, then the right hand sides of equations (3) and (4) are equal to the sample mean vector and covariance matrix, respectively, which are nothing but the MLEs. If our aim is to obtain the local minimum points of $L_\gamma(\mu)$, then we only have to update μ_t and fix Σ_t to be the identity matrix I. Similarly if our aim is to obtain the local minimum points of $L_\gamma(\mu, \Sigma)$ with fixed μ, then we only have to update Σ_t and fix $\mu_t = \mu$.

2.2 Algorithm of the Spontaneous Clustering

In general, the spontaneous clustering based on a density function $f(x, \theta)$ with parameter θ is defined as follows.

Spontaneous Clustering

Step 1 Find the local minimum points of $L_\gamma(\theta)$, denoted by $\hat{\theta}_1, \ldots, \hat{\theta}_K$, where $L_\gamma(\theta)$ is the γ-loss function for $f(x, \theta)$.

Step 2 Consider K clusters according to $\hat{\theta}_1, \ldots, \hat{\theta}_K$, and assign the data to the clusters.

In a special case, the spontaneous clustering based on the normal distribution is defined as follows. We set Θ_μ and $\Theta_{(\mu,\Sigma)}$ are the empty sets at the start of the algorithm. The algorithm of subsection 2.1 is employed in the spontaneous clustering below.

Spontaneous Clustering Based on the Normal Distribution

Step 1-1 If Θ_μ is the empty set, choose M initial values $x_{(1)}, \ldots, x_{(M)}$ in the data set $\{x_1, \ldots, x_n\}$ at random. Otherwise, choose initial values in $\{x_1, \ldots, x_n\}$ as follows: $x_{(1)}, \ldots, x_{(M)}$ are M maximum points of $d(\cdot, \Theta_\mu)$, where

$$d(x, \Theta_\mu) = \min_{\hat{\mu} \in \Theta_\mu} \|x - \hat{\mu}\|.$$

Step 1-2 Apply the algorithm in subsection 2.1 to the data set M times with each initial value $x_{(i)}, i = 1, \ldots, M$ to find the local minimum points of $L_\gamma(\mu)$. Then add the obtained local minimum points to Θ_μ.

Step 1-3 Repeat Step 1-1 and 1-2 until the number of elements in Θ_μ does not increase.

Step 1-4 For each local minimum point $\hat{\mu} \in \Theta_\mu$, obtain a minimum point of $L_\gamma(\hat{\mu}, \Sigma)$ with respect to Σ, denoted by $\hat{\Sigma}$, with the algorithm in subsection 2.1. Then add $(\hat{\mu}, \hat{\Sigma})$ to $\Theta_{(\mu,\Sigma)}$.

Step 2 Write $\Theta_{(\mu,\Sigma)}$ by $\{(\hat{\mu}_k, \hat{\Sigma}_k)\}_{k=1}^K$ and assign each observation x_i to the $\hat{k}$-th cluster with

$$\hat{k} = \underset{k=1,\ldots,K}{\operatorname{argmin}} (x_i - \hat{\mu}_k)^\top \hat{\Sigma}_k^{-1} (x_i - \hat{\mu}_k).$$

In the algorithm of the spontaneous clustering, we define $(\hat{\mu}_k, \hat{\Sigma}_k), k = 1, \ldots, K$ as the centers and the covariance matrices of clusters. In the remainder of this paper, we focus on the spontaneous clustering based on the normal distribution.

2.3 Selection Procedure for γ

The value of power index γ plays a key role in the spontaneous clustering, because γ affects the number of clusters obtained by the spontaneous clustering. We propose two

methods to select the value of γ. One is a heuristic choice of γ that depends on the range of the data. Our proposal is $\hat{\gamma} = 72/R^2$, where R is defined by the maximum range:

$$R = \max_{j=1,\ldots,p} \left\{ \left(\max_{i=1,\ldots,n} x_{ij} \right) - \left(\min_{i=1,\ldots,n} x_{ij} \right) \right\},$$

where $x_i = (x_{i1}, \ldots, x_{ip})^\top$. The outline of the derivation of $\hat{\gamma}$ is as follows. Suppose the data set is generated from the mixture of two normal distributions centered at μ_1 and μ_2 with the identity covariance matrix and the same mixing proportion, respectively. Our simulation result suggests that if $\|(\mu_1 - \mu_2)/2\| = 3\sqrt{2}/2 \doteq 2.12$, then the value of γ needs to be more than or equal to 1 for two local minimum points of $L_\gamma(\mu)$ to exist. Proposition 3.1 tells that if all the data are multiplied by a scalar a and the spontaneous clustering is applied to the transformed data, then the value of γ needs to be more than or equal to a^{-2} to guarantee the existence of two local minimum points of $L_\gamma(\mu)$. If $\|(\mu_1 - \mu_2)/2\| = r$, then $a = r/(3\sqrt{2}/2)$. Hence we propose to use the value of γ defined as

$$\hat{\gamma} = \left(\frac{r}{\frac{3\sqrt{2}}{2}} \right)^{-2} = \frac{9}{2r^2}. \tag{5}$$

The value of r can be estimated by the range of the data. Let R_j be the range of the j-th variable. If there are K disjoint clusters lying side by side on a line parallel to the axis of the j-th variable, then we can estimate r by $R_j/(2K)$ as is just illustrated in Figure 2. There are p variables, so p directions have to be considered simultaneously. We use the maximum range R, and estimate r by $R/(2K)$. The value of K can be determined from our prior knowledge about the possible number of clusters. If $K = 2$, we have $\hat{\gamma} = 72/R^2$. We observe that this rule works well in several empirical studies although the discussion does not completely have the theoretical background.

We also propose a more sophisticated method based on AIC. The value of γ which minimizes AIC is recommended as the optimal selection of γ. Let K_γ be the number of clusters and $(\hat{\mu}_{\gamma k}, \hat{\Sigma}_{\gamma k}), k = 1, \ldots, K_\gamma$ be the centers and the covariance matrices of clusters resulting from the spontaneous clustering. Let $\phi(x, \mu, \Sigma)$ be the density function of the normal distribution with mean vector μ and covariance matrix Σ. Then $\phi(x, \hat{\mu}_{\gamma k}, \hat{\Sigma}_{\gamma k})$ is used as a density estimator of mixture component $f_k(x)$ in (1). The result of the spontaneous clustering implies the mixture of normal distributions as an estimator of the density function of the population distribution g in (1),

$$\hat{g}_\gamma(x) = \sum_{k=1}^{K_\gamma} \hat{\tau}_{\gamma k} \phi(x, \hat{\mu}_{\gamma k}, \hat{\Sigma}_{\gamma k}),$$

where $\hat{\tau}_{\gamma k}$ is an estimator of mixing proportion τ_k defined as the proportion of the observations assigned to the k-th cluster. The AIC based on $\hat{g}_\gamma$ is defined as follows.

$$\text{AIC}_\gamma = -2 \sum_{i=1}^{n} \log \hat{g}_\gamma(x_i) + 2 \left\{ K_\gamma \frac{p(p+3)}{2} + K_\gamma - 1 \right\}.$$

The value of γ minimizing AIC_γ is proposed as the optimal selection of γ.

3 Behavior of the γ-loss Function

We provide a justification for the spontaneous clustering by exploring its theoretical aspects. The key fact is that the γ-loss function $L_\gamma(\mu)$ has K local minimum points if the data set consists of K cluster groups.

3.1 Nonconvexity

We consider the reason why the γ-loss function has local minimum points as illustrated in panel (b) of Figure 1. The optimization problem for a nonconvex function which is

expressed as difference of two convex functions has been considered in Yuille and Rangarajan (2003) and An and Tao (2005). Effective algorithms such as CCCP and DCA have been developed. Actually, a monotonic transformation of the γ-loss function can be expressed as difference of two convex functions, and this expression gives the reason why the γ-loss function has local minimum points. Rewrite $L_\gamma(\mu)$ as

$$L_\gamma(\mu) = -\frac{1}{n}\exp\left[\log\left\{\sum_{i=1}^{n}\exp\left(\gamma x_i^\top \mu - \frac{\gamma}{2}x_i^\top x_i\right)\right\} - \frac{\gamma}{2}\mu^\top\mu\right].$$

The local minimum points of $L_\gamma(\mu)$ are equal to local maximum points of $\Gamma_\gamma(\mu) = \Gamma_\gamma^{(1)}(\mu) - \Gamma_\gamma^{(2)}(\mu)$, where

$$\Gamma_\gamma^{(1)}(\mu) = \log\left\{\sum_{i=1}^{n}\exp\left(\gamma x_i^\top \mu - \frac{\gamma}{2}x_i^\top x_i\right)\right\},\ \Gamma_\gamma^{(2)}(\mu) = \frac{\gamma}{2}\mu^\top\mu.$$

Then $\Gamma_\gamma^{(2)}(\mu)$ is obviously a convex function and has a constant Hessian matrix with positive diagonal elements, which means the surface of $\Gamma_\gamma^{(2)}(\mu)$ is curved. $\Gamma_\gamma^{(1)}(\mu)$ is also a convex function because its Hessian matrix is given by

$$\frac{\partial^2\Gamma_\gamma^{(1)}(\mu)}{\partial\mu\partial\mu^\top} = \gamma^2\sum_{i=1}^{n} w(x_i,\mu,I)(x_i - \overline{x}_{\gamma\mu})(x_i - \overline{x}_{\gamma\mu})^\top, \tag{6}$$

where $\overline{x}_{\gamma\mu} = \sum_{i=1}^{n} w_\gamma(x_i,\mu,I)x_i$, and the Hessian matrix is obviously positive definite. However, the Hessian matrix of $\Gamma_\gamma^{(1)}(\mu)$ varies depending on the data and μ, and becomes close to the zero matrix in a neighborhood where observations are concentrated. This fact is clear from the form of the Hessian matrix (6) and means the surface of $\Gamma_\gamma^{(1)}(\mu)$ is almost flat in such a neighborhood. Difference between the flat surface and the curved surface causes local maximum points of $\Gamma_\gamma(\mu)$. Figure 3 illustrates such a phenomenon, where the red, green, and blue lines show $\Gamma_\gamma^{(1)}(\mu)$, $\Gamma_\gamma^{(2)}(\mu)$, and $\Gamma_\gamma(\mu)$, respectively, with dimension $p = 1$ and $\gamma = 3$. The graphs of $\Gamma_\gamma^{(1)}(\mu)$ and $\Gamma_\gamma(\mu)$ are shifted to take 0 at $\mu = 0$.

3.2 Existence of Local Minimum Points

We consider a condition for the existence of local minimum points of $L_\gamma(\mu)$. As we discussed in subsection 2.2, the local minimum points of $L_\gamma(\mu)$ are defined as the centers of clusters, so it is important to know when the γ-loss function has local minimum points.

To simplify the argument, we assume that the data set is generated from the mixture of two normal distributions with covariance matrix $\sigma^2 I$,

$$g(x) = \tau_1 \phi(x, \mu_1, \sigma^2 I) + \tau_2 \phi(x, \mu_2, \sigma^2 I),\ \tau_1 + \tau_2 = 1,\ \tau_k > 0,\ k = 1, 2.$$

For easy calculation, we consider $n = \infty$. As n tends to ∞, $L_\gamma(\mu)$ almost surely converges to the γ-cross entropy defined by

$$C_\gamma(g, \phi(\cdot, \mu, I)) = -\int g(x)\phi(x, \mu, I)^\gamma dx. \tag{7}$$

See appendix B for the detailed discussion about the γ-cross entropy. $C_\gamma(g, \phi(\cdot, \mu, I))$ becomes

$$\begin{aligned} C_\gamma(g, \phi(\cdot, \mu, I)) &= \sum_{k=1,2} \tau_k C_\gamma(\phi(\cdot, \mu_k, \sigma^2 I), \phi(\cdot, \mu, I)) \\ &\propto -\sum_{k=1,2} \tau_k \phi\left(\mu, \mu_k, \left(\sigma^2 + \frac{1}{\gamma}\right) I\right), \end{aligned}$$

which is nothing but the minus density function of the mixture of two normal distributions with the same covariance matrix $(\sigma^2 + 1/\gamma)I$. Hence the local minimum points of $C_\gamma(g, \phi(\cdot, \mu, I))$ are equal to the modes of the density function of the normal mixture. Figure 4 shows $-C_\gamma(g, \phi(\cdot, \mu, I))$ with dimension $p = 2$, where $-C_\gamma(g, \phi(\cdot, \mu, I))$ has one or two modes depending on the values of $\mu_1, \mu_2, \tau_1, \tau_2$, and γ. For the univariate case, a necessary and sufficient condition that the density function of the mixture of two

normal distributions should be bimodal is given in de Helguero (1904). We use a similar technique as in de Helguero (1904) to obtain a necessary and sufficient condition for $C_\gamma(g, \phi(\cdot, \mu, I))$ to have two local minimum points.

Proposition 3.1 *Let $\nu = (\mu_1 - \mu_2)/2$ and $d = \|\nu\|^2 - (\sigma^2 + 1/\gamma)$. Then $C_\gamma(g, \phi(\cdot, \mu, I))$ has two local minimum points if and only if the following three conditions hold:*

$$d > 0, \tag{8}$$

$$\exp\left(\frac{2\gamma}{1+\gamma\sigma^2}\|\nu\|\sqrt{d}\right) > \frac{\gamma}{1+\gamma\sigma^2}\left(\|\nu\| + \sqrt{d}\right)^2 \frac{\tau_1}{\tau_2}, \tag{9}$$

$$\exp\left(-\frac{2\gamma}{1+\gamma\sigma^2}\|\nu\|\sqrt{d}\right) < \frac{\gamma}{1+\gamma\sigma^2}\left(\|\nu\| - \sqrt{d}\right)^2 \frac{\tau_1}{\tau_2}. \tag{10}$$

Especially, if $\tau_1 = \tau_2$, then (9) and (10) hold for any $d > 0$. When the two local minimum points exist, they lie on the segment between μ_1 and μ_2. One closer to μ_1 and the other to μ_2 are denoted by μ_1^* and μ_2^*, respectively. Then $\|\mu_1 - \mu_1^*\|$ and $\|\mu_2 - \mu_2^*\|$ are bounded above by

$$\|\nu\| - \sqrt{\|\nu\|^2 - \left(\sigma^2 + \frac{1}{\gamma}\right)}.$$

By proposition 3.1, for any σ^2, if μ_1 and μ_2 are distinct enough, then there exists γ that guarantees the existence of two local minimum points of $C_\gamma(g, \phi(\cdot, \mu, I))$, and two clusters are defined at the same instant. In addition, the center of a cluster μ_k^* becomes arbitrarily close to μ_k ($k = 1, 2$), when $\|\mu_1 - \mu_2\|$ becomes large.

4 Simulation

The performance of the spontaneous clustering was investigated through Monte Carlo experiments. A comparison of the spontaneous clustering with the K-means algorithm

and the model based clustering (MBC) was also implemented.

4.1 Case of Spherical Clusters

We demonstrate the performance of the spontaneous clustering in comparison with the K-means algorithm. In this simulation, it is supposed that the covariance matrices of clusters are known to be the identity matrix. The value of γ for the spontaneous clustering is determined by the two methods described in subsection 2.3. The number of clusters for the K-means algorithm is determined by two methods described below. The performance of clustering is measured by BHI defined later.

For the K-means algorithm, the method by Caliński and Harabasz (1974) and the gap statistic by Tibshirani et al. (2001) were used to fix the number of clusters. Let $B(k)$ and $W(k)$ be the between- and within-cluster sums of squares with k clusters. Caliński and Harabasz (1974) propose to select the number of clusters k which maximizes CH(k), where CH(k) is defined as

$$\text{CH}(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}.$$

On the other hand, Tibshirani et al. (2001) propose to choose the value of k which maximizes $\text{Gap}_n(k) = E_n^*(\log(W_k)) - \log(W_k)$, where E_n^* denotes expectation under a sample of size n from the reference distribution.

The sample of size 200 is generated from the mixture of five standard normal distributions centered at $(0,0)^\top$, $(3,3)^\top$, $(-3,3)^\top, (-3,-3)^\top, (3,-3)^\top$ with equal mixing proportion. Figure 5 displays an example sample. We simulated 100 runs, and compared clustering results from the spontaneous clustering with those from the K-means algorithm. Figure 6 shows the value of AIC and the number of clusters resulting from

the spontaneous clustering for the sample in Figure 5. The selected value of γ based on AIC is 0.7.

Table 1 displays the frequency of choosing K clusters for each of the methods for different values of K. All methods except the K-means algorithm with Gap chose the true number of clusters in almost every simulation run. To measure the performance of the clustering, we used Biological Homogeneity Index (BHI) (Wu, 2011), which measures the homogeneity between the cluster $\mathcal{C} = \{C_1, \ldots, C_K\}$ and the biological category or subtype $\mathcal{B} = \{B_1, \ldots, B_L\}$,

$$\text{BHI}(\mathcal{C}, \mathcal{B}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k(n_k - 1)} \sum_{i \neq j, i, j \in C_k} 1(B^{(i)} = B^{(j)}), \tag{11}$$

where $B^{(i)} \in \mathcal{B}$ is the subtype for the observation x_i and n_k is the number of the observations in C_k. This index is bounded above by 1 meaning the perfect homogeneity between the clusters and the biological categories. The mean value of BHI over 100 simulation runs for each method is shown in Table 2. All methods except the K-means algorithm with Gap have good clustering results. In every simulation run, if each method detected five clusters for a sample, we calculated the Euclidean distance between the center of a cluster and the mean vector of the corresponding normal component of the normal mixture. The mean value of the distance is also shown in Table 2, where DM1, ..., DM5 represent the mean value for cluster $1, \ldots, 5$, respectively. In this simulation setting, the centers obtained by the spontaneous clustering vary more than those obtained by the K-means algorithm.

To summarize, this simulation example shows that the spontaneous clustering with the range and AIC has almost the same performance as the K-means algorithm with CH, and better performance than the K-means algorithm with Gap.

4.2 Case of Ellipsoidal Clusters

We demonstrate the performance of the spontaneous clustering in comparison with the MBC, in which the component density is normal. It is supposed that the covariance matrices of clusters are heterogeneous and unknown. The value of γ for the spontaneous clustering and the number of clusters for the MBC are determined based on AIC.

The sample of size 100 is generated from the mixture of two bivariate normal distributions with mean vectors $(0,0)^\top$, $(3,3)^\top$, and covariance matrices

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \begin{pmatrix} 2 & -0.5 \\ -0.5 & 2 \end{pmatrix}.$$

Figure 7 displays an example sample, and Figure 8 shows the value of AIC and the number of clusters resulting from the spontaneous clustering for the sample. Note that we use two values γ_1 and γ_2 as power index γ. γ_1 is used for $L_\gamma(\mu)$ when defining the centers of clusters, and γ_2 for $L_\gamma(\mu, \Sigma)$ when defining the covariance matrices. The selected values of γ_1 and γ_2 for the sample in Figure 7 are $\gamma_1 = 0.25$ and $\gamma_2 = 0.7$. We simulated 100 runs, and compared the clustering result from the spontaneous clustering with that from MBC.

Table 3 displays the frequency of choosing K clusters for each of the clustering algorithms for different values of K. The spontaneous clustering chose the true number of clusters, while the MBC selected large number of clusters 3-10, 39 frequencies. The mean value of BHI is shown in Table 4. Both clustering algorithms show good performance. In every simulation run, if each clustering method detected two clusters for a sample, two measures were calculated. One is the Euclidean distance between the center of a cluster and the mean vector of the corresponding normal component

of the normal mixture. The other is the Frobenius norm of the covariance matrix of a cluster minus that of the corresponding normal component. The mean values of the Euclidean distance and the Frobenius norm are shown in Table 4, where DV1 and DV2 represent the mean value of the Frobenius norm for cluster 1 and 2, respectively. In this simulation setting, similar to the simulation result in subsection 4.1, the centers and the covariance matrices obtained by the spontaneous clustering vary more than those obtained by MBC.

To summarize, this simulation example reveals that the spontaneous clustering with AIC has almost the same performance as MBC with AIC.

5 Data Analysis

To evaluate the practical performance of the spontaneous clustering, we applied it with the fixed identity covariance matrix to real data as well as the K-means algorithm. The data set consists of the chemical composition of 45 specimens of Romano-British pottery, determined by atomic absorption spectrophotometry, for nine oxides (Tubb et al., 1980). Figure 9 shows the scatterplot matrix of data on Romano-British pottery. In addition to the chemical composition of the specimens, the kiln site at which the specimen was found is known. There exist five kiln sites, and they are from three different regions, so that we use the three regions as class labels. Our aim is to partition the 45 specimens into clusters corresponding to the three classes by using only information about the chemical composition without knowledge about the class labels. The value of γ for the spontaneous clustering is determined by the two methods based on the range

of the data and AIC, respectively. The number of clusters for the K-means algorithm is determined by CH and Gap.

Table 5 shows the result of the spontaneous clustering. The value of AIC and the number of clusters are shown in panel (a) of Figure 10. With optimal values of γ based on the range and AIC, the spontaneous clustering detects three clusters corresponding to the three regions. In particular, the clustering result by the heuristic choice of γ is the most correct. The scatterplot of Al_2O_3 variable suggests that the number of clusters is two, and the maximum range is obtained from the variable. This is associated with the scenario discussed in the derivation of the heuristic method, in which we assume the number of clusters is two. The values of CH and Gap are shown in panels (b) and (c) of Figure 10. They increase almost monotonically as the number of clusters increases, so CH and Gap do not work well for this data. As a result, we observe the spontaneous clustering based on the range and AIC can detect three clusters properly and partition the 45 specimens into clusters corresponding to the three regions.

6 Discussion

We proposed a new clustering algorithm based on the local minimization of the γ-ross function, which we named the spontaneous clustering. In the spontaneous clustering, the local minimum points of the γ-loss function are defined as the centers and covariance matrices of clusters. A large majority of statistical methods use the global minimum or maximum point of objective functions and try to avoid local minimum or maximum points. The convexity of the objective functions plays an important role

in statistics. For example, support vector machine has a convex loss function, and an efficient algorithm to obtain the global minimum point is considered based on the convexity (Bishop, 2006). Although nonconvexity is generally intractable, the spontaneous clustering benefits from the nonconvexity, which makes our method unique and interesting. The idea to use local minimum points of the γ-loss function can be applied to other statistical methods. For example, the idea is applied to principal component analysis (Mollah et al., 2010) and to estimation of Gaussian copula parameter (Notsu et al., 2012).

The spontaneous clustering does not require the information about the number of clusters a priori and can find it automatically if the value of power index γ is properly fixed. In contrast, existing methods such as K-means and model based clustering demand the number of clusters. Instead of the number of clusters, the value of γ has to be determined in the spontaneous clustering. Two methods to determine the value of γ are proposed in this paper. One is a heuristic method which depends on the range of the data. Our simulation research shows that it has good performance in many situations, so we can usually use this heuristic method. A more sophisticated choice based on AIC is also proposed although it requires much computational effort. In the beginning of the research about selection of γ, we considered a cross validation technique, that is one of the common procedures to select the optimal value of a tuning parameter (Hastie et al., 2009). In Mollah et al. (2010) the method using the cross validation is proposed for selection of γ. However, the method does not work well for the spontaneous clustering. Hence we employ AIC for selection of γ. It is demonstrated that our proposal works well by the simulation study and the real data analysis.

A Proof of Proposition 3.1

No generality is lost by assuming $\mu_2 = -\mu_1$. The gradient of $C_\gamma(g, \phi(\cdot, \mu, I))$ is given by

$$\begin{aligned}\frac{\partial C_\gamma(g, \phi(\cdot, \mu, I))}{\partial \mu} \quad &\propto \quad \tau_1 \phi(\mu, \mu_1, (\sigma^2 + 1/\gamma)I)(\mu - \mu_1) \\ &\quad + \tau_2 \phi(\mu, -\mu_1, (\sigma^2 + 1/\gamma)I)(\mu + \mu_1). \end{aligned} \tag{12}$$

From (12), every local minimum point of $C_\gamma(g, \phi(\cdot, \mu, I))$ should exist on the segment between $-\mu_1$ and μ_1. The Hessian matrix of $C_\gamma(g, \phi(\cdot, \mu, I))$ is given by

$$\begin{aligned}\frac{\partial^2 C_\gamma(g, \phi(\cdot, \mu, I))}{\partial \mu \partial \mu^\top} \quad \propto \quad & -\tau_1 \phi(\mu, \mu_1, (\sigma^2 + 1/\gamma)I) \frac{\gamma}{1 + \sigma^2 \gamma} (\mu - \mu_1)(\mu - \mu_1)^\top \\ & -\tau_2 \phi(\mu, -\mu_1, (\sigma^2 + 1/\gamma)I) \frac{\gamma}{1 + \sigma^2 \gamma} (\mu + \mu_1)(\mu + \mu_1)^\top \\ & +\tau_1 \phi(\mu, \mu_1, (\sigma^2 + 1/\gamma)I) I \\ & +\tau_2 \phi(\mu, -\mu_1, (\sigma^2 + 1/\gamma)I) I. \end{aligned} \tag{13}$$

Let $\mu(t) = t\mu_1$. From (13), $\mu(t)$ is a local minimum point of $C_\gamma(g, \phi(\cdot, \mu, I))$ if and only if t is a local minimum point of $C_\gamma(g, \phi(\cdot, \mu(t), I))$ with respect to t. $C_\gamma(g, \phi(\cdot, \mu(t), I))$ becomes

$$C_\gamma(g, \phi(\cdot, \mu(t), I)) \propto -\tau_1 \exp(-C(t-1)^2) - \tau_2 \exp(-C(t+1)^2),$$

where C is equal to $\|\mu_1\|^2 \gamma / (2(1 + \sigma^2 \gamma))$. The derivative of $C_\gamma(g, \phi(\cdot, \mu(t), I))$ is given by

$$\frac{d}{dt} C_\gamma(g, \phi(\cdot, \mu(t), I)) \quad \propto \quad \tau_1 \exp(-C(t-1)^2)(t-1) + \tau_2 \exp(-C(t+1)^2)(t+1).$$

It is possible to restrict $-1 < t < 1$. Then

$$\begin{aligned} & \frac{d}{dt} C_\gamma(g, \phi(\cdot, \mu(t), I)) > 0 \\ \iff & \exp\left(-C(t+1)^2 + C(t-1)^2\right) > \frac{(1-t)\tau_1}{(t+1)\tau_2} \\ \iff & -4Ct + \log(t+1) - \log(1-t) - \log\frac{\tau_1}{\tau_2} > 0. \end{aligned} \tag{14}$$

Let $h(t)$ be the left hand side of inequality (14). The derivative of $h(t)$ is given by

$$h'(t) = -4C + \frac{1}{t+1} + \frac{1}{1-t},$$

and

$$\begin{aligned} h'(t) > 0 \iff & -4C(1-t^2) + (1-t) + (1+t) > 0 \\ \iff & t^2 - \left(1 - \frac{1}{2C}\right) > 0. \end{aligned}$$

If $1 - 1/(2C) \leq 0$, then $h'(t) \geq 0$, and $C_\gamma(g, \phi(\cdot, \mu(t), I))$ has one local minimum point. Hence $C_\gamma(g, \phi(\cdot, \mu(t), I))$ has two local minimum points if and only if

$$1 - \frac{1}{2C} > 0,\ h(-D) > 0,\ h(D) < 0,$$

where D is the positive solution of equation $h'(t) = 0$, that is $D = \sqrt{1 - 1/(2C)}$. Condition $1 - 1/(2C) > 0$ is equivalent to $\|\mu_1\|^2 - (\sigma^2 + 1/\gamma) > 0$. Condition $h(-D) > 0$ is equivalent to

$$\begin{aligned} & \exp\left(\frac{2\gamma}{1+\sigma^2\gamma}\|\mu_1\|\sqrt{\|\mu_1\|^2 - \left(\sigma^2 + \frac{1}{\gamma}\right)}\right) \\ & \quad > \frac{\gamma}{1+\sigma^2\gamma}\left(\|\mu_1\| + \sqrt{\|\mu_1\|^2 - \left(\sigma^2 + \frac{1}{\gamma}\right)}\right)^2 \frac{\tau_1}{\tau_2}, \end{aligned}$$

and condition $h(D) < 0$ is equivalent to

$$\begin{aligned} & \exp\left(-\frac{2\gamma}{1+\sigma^2\gamma}\|\mu_1\|\sqrt{\|\mu_1\|^2 - \left(\sigma^2 + \frac{1}{\gamma}\right)}\right) \\ & \quad < \frac{\gamma}{1+\sigma^2\gamma}\left(\|\mu_1\| - \sqrt{\|\mu_1\|^2 - \left(\sigma^2 + \frac{1}{\gamma}\right)}\right)^2 \frac{\tau_1}{\tau_2}. \end{aligned}$$

Note that μ_1^* is on the line between $D\mu_1$ and μ_1. Similarly $(-\mu_1)^*$ is on the line between $-\mu_1$ and $-D\mu_1$. Then

$$\|\mu_1^* - \mu_1\| \leq (1-D)\|\mu_1\| = \|\mu_1\| - \sqrt{\|\mu_1\|^2 - \left(\sigma^2 + \frac{1}{\gamma}\right)}.$$

If $\tau_1 = \tau_2$, then $h(\pm 1) = \pm\infty, h(0) = 0$. Condition $1 - 1/(2C) > 0$ is equivalent to $h'(0) < 0$. Hence two conditions $h(-D) > 0, h(D) < 0$ hold whenever condition $1 - 1/(2C) > 0$ holds. □

B γ-divergence and γ-loss Function

The aim of this section is to give a general introduction to the γ-divergence and the γ-loss function. A more detailed discussion can be found in Eguchi and Kato (2010).

B.1 γ-divergence

Suppose a random sample is generated from a population distribution with density function g. Let $\{f(\cdot, \theta)\}$ be a family of density functions indexed by parameter θ. The γ-cross entropy between g and $f(\cdot, \theta)$ is defined as

$$C_\gamma(g, f(\cdot, \theta)) = -\kappa_\gamma(\theta) \int g(x) f(x, \theta)^\gamma dx,$$

with power index $\gamma > 0$, where $\kappa_\gamma(\theta)$ is the normalizing constant defined as

$$\kappa_\gamma(\theta) = \left(\int f(x, \theta)^{1+\gamma} dx \right)^{-\frac{\gamma}{1+\gamma}}.$$

The Boltzmann-Shannon cross entropy between g and $f(\cdot, \theta)$ is defined by

$$-\int g(x) \log f(x, \theta) dx.$$

The γ-cross entropy and the Boltzmann-Shannon cross entropy have the following relation since $\kappa_\gamma(\theta)$ converges to 1 if γ tends to 0.

$$\begin{aligned}\lim_{\gamma\to 0}\frac{C_\gamma(g, f(\cdot,\theta))+1}{\gamma} &= -\int g(x)\lim_{\gamma\to 0}\left(\frac{f(x,\theta)^\gamma-1}{\gamma}\right)dx \\ &= -\int g(x)\log f(x,\theta)dx.\end{aligned}$$

Hence the Boltzmann-Shannon cross entropy can be seen as the 0-cross entropy, and the γ-cross entropy can be regarded as an extension of the Boltzmann-Shannon cross entropy. The γ-entropy of g is defined as $H_\gamma(g) = C_\gamma(g, g)$; the γ-divergence between g and $f(\cdot,\theta)$ is defined as

$$D_\gamma(g, f(\cdot,\theta)) = C_\gamma(g, f(\cdot,\theta)) - H_\gamma(g).$$

Note that the γ-divergence $D_\gamma(g, f(\cdot,\theta))$ is nonnegative, and $D_\gamma(g, f(\cdot,\theta))$ is equal to 0 if and only if θ satisfies that $g(x) = f(x,\theta)$ almost everywhere x. From these properties, $D_\gamma(g, f(\cdot,\theta))$ can be seen as a kind of distance between g and $f(\cdot,\theta)$ although it does not satisfy the symmetry. When our aim is to find the closest distribution to g in model $\{f(\cdot,\theta)\}$ with respect to the γ-divergence, we only have to find the global minimum point of $D_\gamma(g, f(\cdot,\theta))$ with respect to θ, which is equal to that of $C_\gamma(g, f(\cdot,\theta))$.

B.2 γ-loss Function

The γ-loss function is defined by an estimator of the γ-cross entropy. Let $\{x_1, x_2, \ldots, x_n\}$ be a random sample generated from a population distribution with density function g and $\{f(\cdot,\theta)\}$ be our statistical model. The γ-loss function for $f(\cdot,\theta)$ associated with the γ-divergence is given by

$$L_\gamma(\theta) = -\kappa_\gamma(\theta)\frac{1}{n}\sum_{i=1}^{n} f(x_i,\theta)^\gamma.$$

We extend the definition of the γ-cross entropy to any distributions. For any distribution function G, the γ-cross entropy between G and $f(\cdot, \theta)$ is defined as

$$C_\gamma(G, f(, \theta)) = -\kappa_\gamma(\theta) \int f(x, \theta)^\gamma dG(x).$$

Note that $L_\gamma(\theta)$ equals $C_\gamma(\hat{G}, f(\cdot, \theta))$ with empirical distribution function $\hat{G}$, so that $E(L_\gamma(\theta)) = C_\gamma(g, f(\cdot, \theta))$, and $L_\gamma(\theta)$ almost surely converges to $C_\gamma(g, f(\cdot, \theta))$. The γ-estimator of θ is defined by the global minimum point of $L_\gamma(\theta)$ (Eguchi and Kato, 2010). From the definition of the γ-estimator, it satisfies Fisher consistency. If the density function g belongs to the statistical model $\{f(\cdot, \theta)\}$, then the γ-estimator satisfies asymptotic consistency and normality. The γ-loss function and the log likelihood function satisfy the following relation

$$\lim_{\gamma \to 0} \frac{L_\gamma(\theta) + 1}{\gamma} = -\frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \theta).$$

Hence the MLE can be regarded as the 0-estimator and the γ-estimator can be seen as an extension of the MLE.

References

An, L. T. H. & Tao, P. D. (2005). The DC (Difference of Convex Functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Caliński, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27.

de Helguero, F. (1904). Sui massimi delle curve dimorfiche. *Biometrika*, 3(1):84–98.

Eguchi, S. & Kato, S. (2010). Entropy and divergence associated with power function and the statistical application. *Entropy*, 12:262–274.

Fujisawa, H. & Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer, second edition.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Jin, D., Peng, J., & Li, B. (2011). A new clustering approach on the basis of dynamical neural field. *Neural Computation*, 23:2032–2057.

Mollah, M. N. H., Sultana, N., Minami, M., & Eguchi, S. (2010). Robust extraction of local structures by the minimum β-divergence method. *Neural Networks*, 23(2):226–238.

Notsu, A., Kawasaki, Y., & Eguchi, S. (2012). Detection of heterogeneous structures on the Gaussian copula model using projective power entropy. *submitted*.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Jounal of the Royal Statistical society: Series B*, 63(2):411–423.

Tubb, A., Parker, A. J., & Nickless, G. (1980). The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry*, 22:153–171.

Wu, H. (2011). On biological validity indices for soft clustering algorithms for gene expression data. *Computational Statistics and Data Analysis*, 55(5):1969–1979.

Wu, J., Zivari-Piran, H., Hunter, J. D., & Milton, J. G. (2011). Projective clustering using neural networks with adaptive delay and signal transmission loss. *Neural computation*, 23:1568–1604.

Xu, R. & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

Yuille, A. L. & Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, 15:915–936.

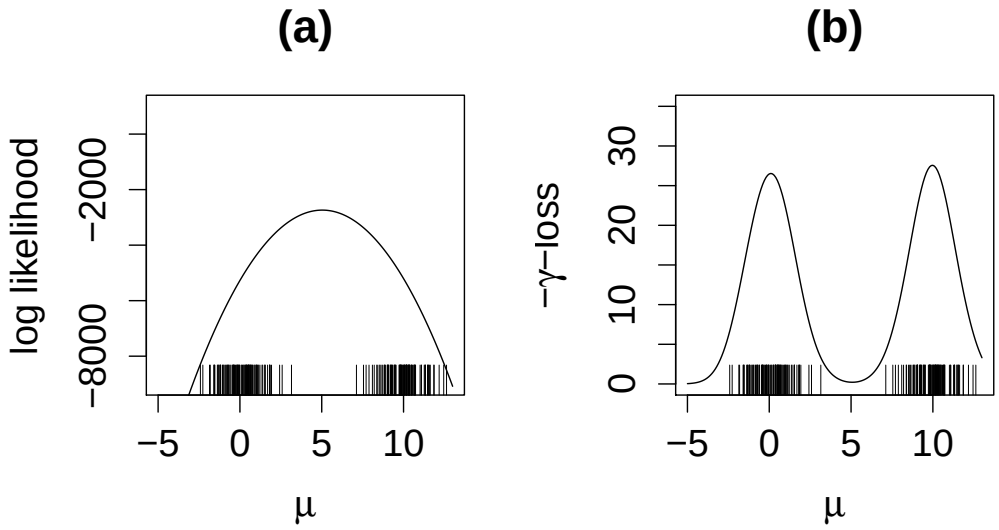

Figure 1: (a) Log likelihood function. (b) Minus γ-loss function ($\gamma = 1$). In panels (a) and (b) the data of size 200 is generated from the mixture of two standard normal distributions centered at 0 and 10, respectively.

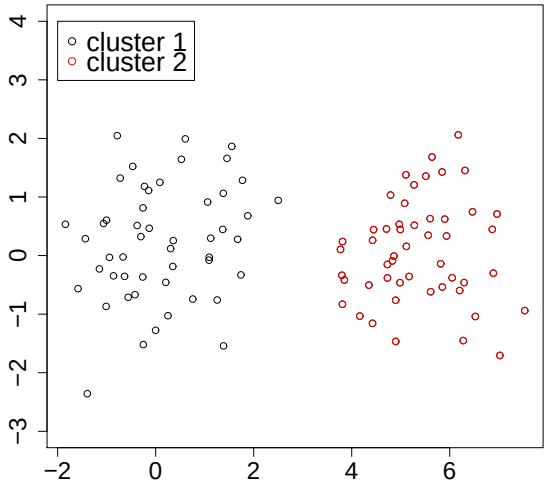

Figure 2: Example data generated from the mixture of two normal distributions centered at $(0, 0)^\top$ and $(5, 0)^\top$ with the identity covariance matrix, respectively.

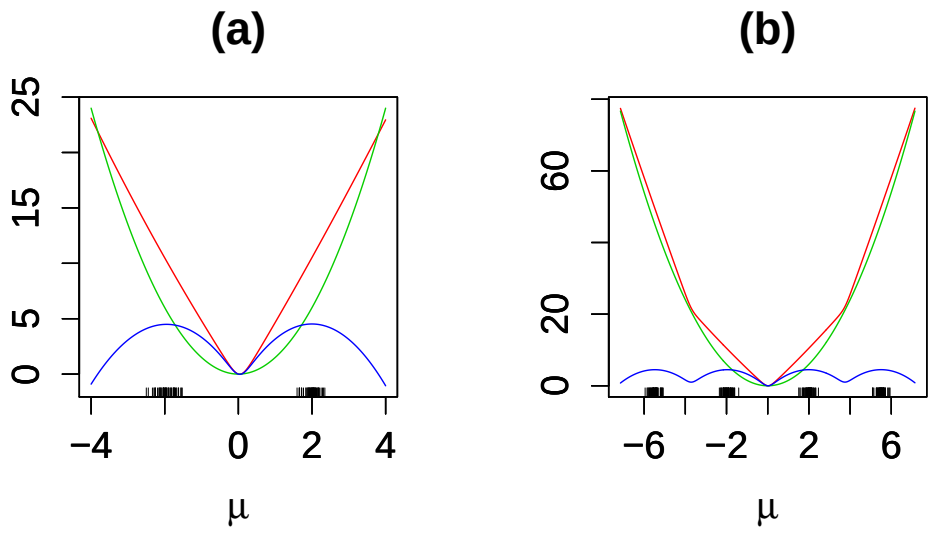

Figure 3: Visualization of $\Gamma_\gamma^{(1)}(\mu)$, $\Gamma_\gamma^{(2)}(\mu)$, and $\Gamma_\gamma(\mu)$. In panel (a) the sample of size 100 is generated from normal mixture $0.5\phi(x,-2,0.04)+0.5\phi(x,2,0.04)$. In panel (b) the sample of size 200 is generated from normal mixture $0.25\phi(x,-5.5,0.04)+0.25\phi(x,-2,0.04)+0.25\phi(x,2,0.04)+0.25\phi(x,5.5,0.04)$.

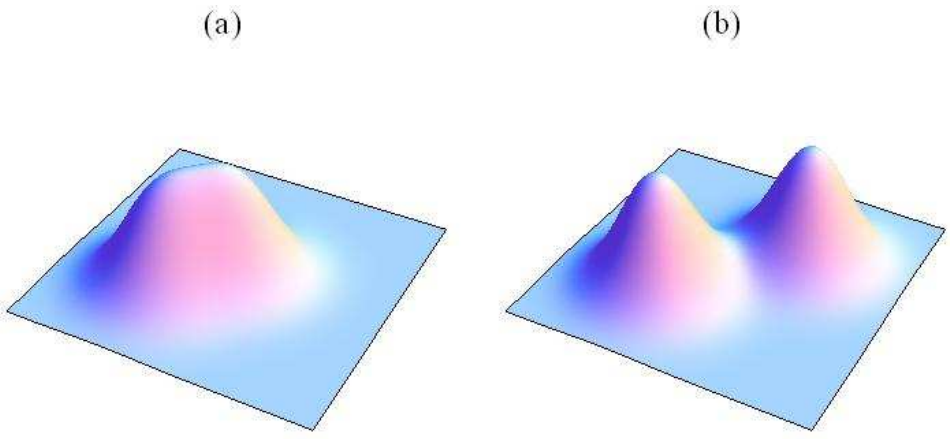

Figure 4: Illustration of $-C_\gamma(g,\phi(\cdot,\mu,I))$. In panel (a) $\mu_1=(0,0)^\top,\mu_2=(2,2)^\top,\tau_1=\tau_2=0.5,\gamma=1,\sigma^2=1$. In panel (b) $\mu_1=(0,0)^\top,\mu_2=(4,4)^\top,\tau_1=\tau_2=0.5,\gamma=1,\sigma^2=1$.

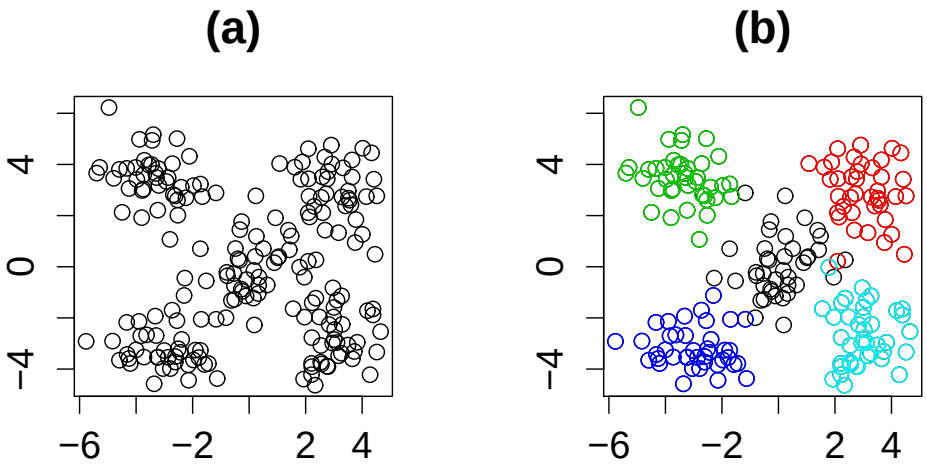

Figure 5: (a) Five clusters. (b) Same as (a) but colored according to cluster.

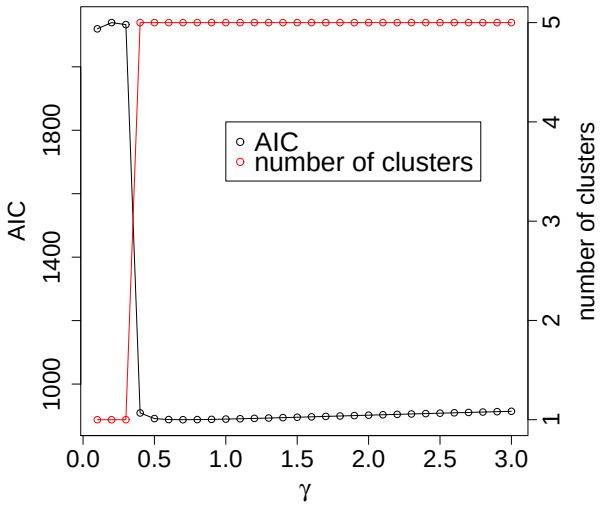

Figure 6: Value of AIC and number of clusters.

Table 1: Frequencies of Choosing K Clusters.

K	1	2	3	4	5
Spontaneous clustering with the range	0	0	0	9	91
Spontaneous clustering with AIC	0	0	0	1	99
K-means with CH	0	0	0	0	100
K-means with Gap	91	7	0	0	2

Table 2: Mean Value of BHI and DM1-DM5.

	BHI	DM1	DM2	DM3	DM4	DM5
Spontaneous clustering with the range	0.93	0.38	0.38	0.37	0.33	0.34
Spontaneous clustering with AIC	0.94	0.34	0.32	0.28	0.27	0.26
K-means with CH	0.95	0.25	0.23	0.21	0.21	0.21
K-means with Gap	0.22	0.16	0.49	0.23	0.41	0.21

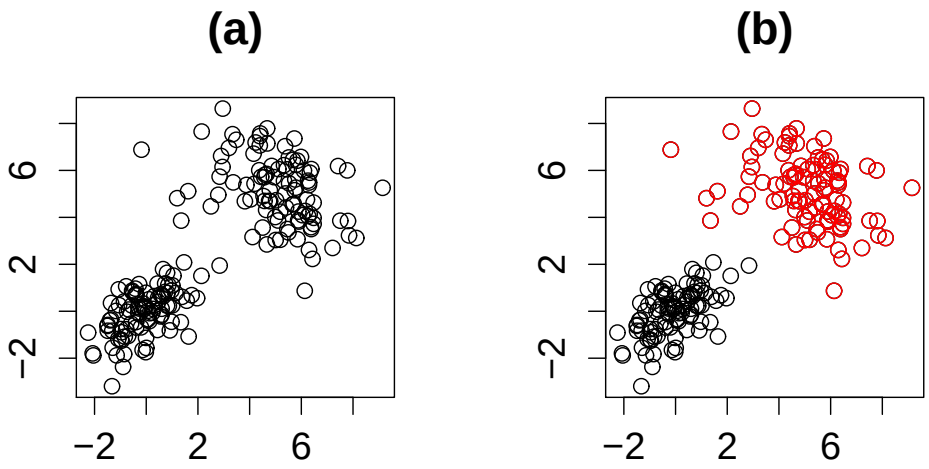

Figure 7: (a) Two clusters. (b) Same as (a) but colored according to cluster.

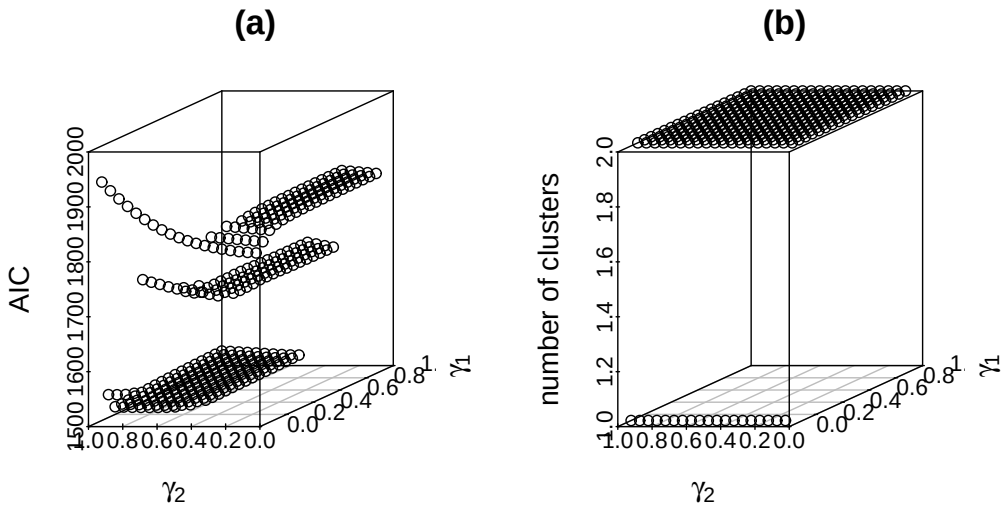

Figure 8: (a) Value of AIC. (b) Number of clusters.

Table 3: Frequencies of Choosing K Clusters.

K	1	2	3	4	5	6	7	8	9	10
Spontaneous clustering	0	100	0	0	0	0	0	0	0	0
MBC	0	61	13	3	4	4	3	4	5	3

Table 4: Mean Value of BHI and DM1, DM2, DV1, and DV2.

	BHI	DM1	DM2	DV1	DV2
Spontaneous clustering	1.00	0.12	0.20	0.33	0.58
MBC	0.99	0.10	0.16	0.22	0.48

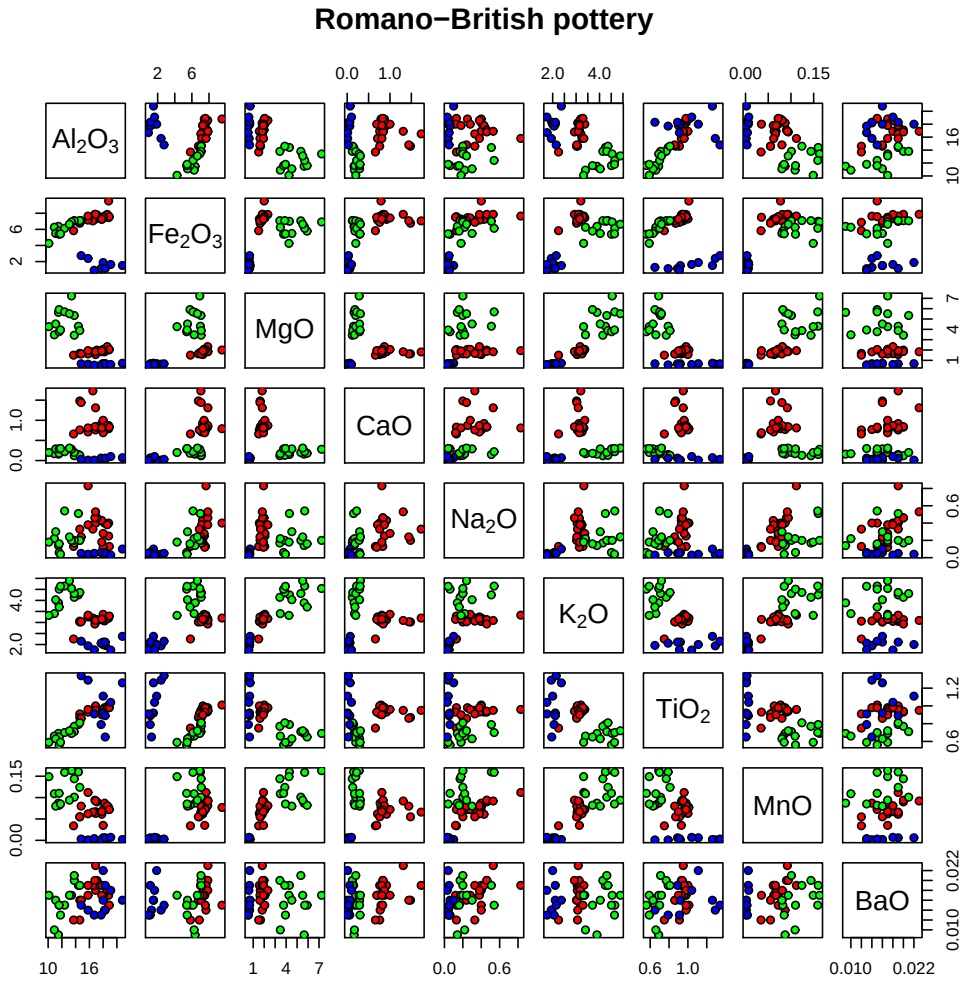

Figure 9: Scatterplot matrix of data on Romano-British pottery. The red, blue, and greed circles correspond to the three regions.

Table 5: Result of the Spontaneous Clustering.

Method	γ	Number of clusters	BHI
Range	0.63	3	1
AIC	0.35	3	0.96

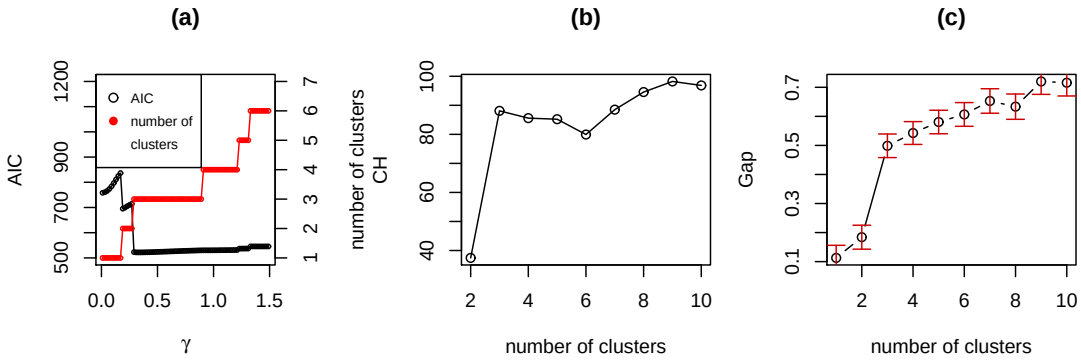

Figure 10: (a) AIC and number of clusters. (b) CH. (c) Gap.