# Statistical Analysis via Local Learning with Gamma-Divergence

Akifumi Notsu

Doctor of Philosophy

Department of Statistical Science

School of Multidisciplinary Sciences

The Graduate University for Advanced Studies

2013

**Abstract**

A divergence measure describes discrepancy between two probability distributions. We present a local learning approach with a specific form of the measures called the gamma-divergence. Learning algorithms are divided into two types, global learning algorithms and local learning algorithms. Global learning algorithms employ all the data simultaneously to estimate the whole data structure, while local learning algorithms employ a part of the data to capture the local structure. Estimation with the gamma-divergence has the local learning capability. The gamma-divergence is a generalization of the Kullback-Leibler divergence with the power index gamma. It employs the power transformation of density functions, instead of the logarithmic transformation employed by the Kullback-Leibler divergence. We consider the gamma-divergence between the underlying distribution and a parametric model, where the underlying distribution means the one which data follow. As a function of the parameter, the gamma-divergence has some local minimum points corresponding to the local structure in the data set. Therefore, we can capture the local structure by the local minimum points. We show that the existence of the local minimum points theoretically in some simple settings. The local learning capability of estimation with the gamma-divergence is applied with respect to cluster analysis and detection of heterogeneous correlation structure.

Cluster analysis is aimed to divide data into some groups called clusters. Finding clusters can be regarded as investigation of the local structure of the data set, so we can apply the local learning capability to cluster analysis. We propose a new method for clustering

with local minimization of the gamma-divergence based on the normal distribution, which we call “spontaneous clustering”. The greatest advantage of the spontaneous clustering is that it automatically detects the number of clusters that adequately reflect the data structure. In contrast, existing methods, such as $K$-means, fuzzy $c$-means, or model-based clustering need to prescribe the number of clusters. Instead of the number of clusters, the value of gamma should be determined for the spontaneous clustering. We propose two methods for this purpose. One is a heuristic choice similar to the bandwidth selection in kernel density estimation. The other is based on Akaike Information Criterion (AIC). We detect all the local minimum points of the gamma-divergence, by which we define the cluster centers.

As for the second application we discuss a parameter estimation problem for a Gaussian copula model. A copula is a multivariate distribution function with uniformly distributed marginals on the unit interval and it determines the correlation structure of a multivariate distribution. We consider the heterogeneous correlation structure, that is, the copula of the underlying distribution might be a mixture of some Gaussian copulas. This heterogeneity can be captured by finding the local minimum points of the gamma-divergence based on the Gaussian copula model. We propose a fixed point algorithm to obtain the local minimum points of the gamma-divergence. It is also shown that the gamma-estimation is robust against outliers in terms of the influence function. A feasible form of the gamma-divergence is given that suites the Gaussian copula model.

In both applications, we consider the situation where the underlying distribution might deviate from the statistical model we fit. The statistical model is a single parametric model, while the underlying distribution is represented by a mixture of some distributions in the

model. This is not the standard situation where the statistical model includes the underlying distribution. In this thesis, we show that even in such a situation the estimation is possible by using the gamma-divergence. One of the advantages of this method is that it works well for mixtures of any number of distributions if they are “distinct” enough.

# Contents

# Chapter 1

# Introduction

Divergence measures serve the concept of distance between two probability distributions. The most well-known divergence is the Kullback-Leibler (KL) divergence proposed in Kullback and Leibler (1951). It is well known that the maximum likelihood estimation can be regarded as the minimization of the empirically estimated KL divergence. A number of different divergence measures have been presented in the literature (Rao, 1982; Eguchi, 1985; Amari and Nagaoka, 2000; Zhang, 2004; Cichocki and Amari, 2010). Other divergence measures lead to different estimators in the same way as the maximum likelihood estimator is defined. Divergence measures are used in not only statistical estimation but also other statistical analyses, such as hypothesis test (Pardo, 2006), multivariate analysis (Mollah et al., 2006, 2010), information criteria (Konishi and Kitagawa, 2008), and boosting (Murata et al., 2004).

In this thesis, we focus on applications of the $\gamma$-divergence, which is one of divergence measures employing the power transformation of density functions. The power transforma-

tion has been employed in statistics, information theory, and physics (Tsallis, 1988). For example, in Box and Cox (1964), it is used for transforming data to meet the standard assumptions, such as the normality of the data. A number of divergence measures with the power transformation have been proposed (Rényi, 1961; Sharma and Mittal, 1977; Liese and Vajda, 1987). Some of the divergence measures have been proved to be especially useful for constructing robust methods. The density power divergence was proposed in Basu et al. (1998) for robust parametric estimation. Minami and Eguchi (2002) presented the same divergence independently for robust blind source separation, which is called the $\beta$-divergence. Jones et al. (2001) and Fujisawa and Eguchi (2008) proposed the $\gamma$-divergence for robust parametric estimation. Jones et al. (2001) investigated the robustness of the estimation with the $\gamma$-divergence from the point of view of the influence function, and they compared the properties of the $\beta$-divergence with those of the $\gamma$-divergence. On the other hand, in Fujisawa and Eguchi (2008), the robustness was explored in terms of information geometry. We, however, employ the $\gamma$-divergence not for robustness but for detection of the local structure in the data set.

Here is a simple example that explains the motivation for employing the $\gamma$-divergence to capture the local structure. Consider the problem of estimating the Gaussian mean parameter $\mu$. The maximum likelihood estimator (MLE) of $\mu$ is given by the arithmetic mean of the data set as the unique maximum point of the log likelihood function. Similarly, the $\gamma$-estimator of $\mu$ is defined by the minimum point of the $\gamma$-loss function, which is the empirically estimated $\gamma$-divergence. It is known that the MLE behaves poorly in various situations where the Gaussianity assumption is inappropriate. For example, if the data are

derived from a mixture of two normal distributions while our model is normal, then the estimation with the log likelihood function fails, as shown in Figure 1 (a). This failure results from the unfaithfulness of the model. The estimation with the $\gamma$-loss function, however, captures all the components of the normal mixture, even though based on the unfaithful model. Figure 1 (b) shows that the $\gamma$-loss function has two local minimum points corresponding to the two mean values of the two normal distributions. That is, the two means can be estimated by the two local minimum points. This thesis applies such a property to detect local mean structure and local correlation structure in the data set, i.e. cluster analysis and parameter estimation of a copula model.

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 technique 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 two main approaches in cluster analysis. One is the hierarchical approach, which describes a tree structure called a "dendrogram". The other is the approach of data space partition, such as the $K$-means clustering. This thesis focuses on the latter approach from the point of view of statistical pattern recognition. We propose what we call the spontaneous clustering. It starts with finding cluster centers in a data set. For this purpose, we employ the $\gamma$-loss function of the Gaussian mean parameter. In the spontaneous clustering, we will propose to determine the cluster centers by the local minimum points of the $\gamma$-loss function. Almost all procedures via data space partition require to pre-determine the num-

ber of clusters; the selection of the number of clusters is a major challenge. A number of methods for this purpose 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 $\gamma$ is properly fixed. The name "spontaneous clustering" comes from this property. Instead of the number of clusters, the value of the power index $\gamma$ should be determined. We will propose two methods to accomplish this aim. One is a heuristic choice of $\gamma$ that merely relies on the range of the data, and the other is a more sophisticated method based on Akaike Information Criterion (AIC).

The estimation of the Gaussian copula parameter is the other application of the local learning capability with the $\gamma$-divergence. Applications of copula models have been increasing in number in recent years. There are a variety of applications on finance, risk management (McNeil et al., 2005), and multivariate time series analysis (Zhang et al., 2011). With copula models, the specification of the marginal distributions is parameterized separately from the dependence structure of the joint distribution. Hence it gives a convenient way of the construction of flexible and more general multivariate distributions. As far as we know, there exist only a few works that tackled with the identification and the statistical estimation of the mixture of copula models and most of them rely on MCMC algorithm. In this thesis we focus on a misspecified Gaussian copula model. In other words, a sample follows a distribution mixed with different sources but a statistical model we fit is just a single Gaussian copula. It is very hard to construct multivariate copulas for three or more random variables (Nelsen, 1999), while the Gaussian is an exception. So we start with the Gaussian copula model, but later in the section 4.3.6 we will show our method is closely related to

$t$-copula. As an example of misspecification, we consider that the underlying distribution is

$$\tau c_{\mathrm{G}}(u; P_1) + (1-\tau) c_{\mathrm{G}}(u; P_2), \tag{1.1}$$

where $\tau$ is a mixing proportion and $c_{\mathrm{G}}(u; P)$ denotes the probability density function of the Gaussian copula with the correlation matrix parameter $P$. We see that the MLE for $P$ almost surely converges to $\tau P_1 + (1-\tau)P_2$ under the assumption (1.1), which means that the MLE fails to detect the structure of the underlying distribution. We make use of the $\gamma$-loss function of the Gaussian copula parameter for this problem. Our research shows that even if a single Gaussian copula model is incorrectly fitted to the data from the mixture distribution (1.1), the $\gamma$-loss function can detect both $P_1$ and $P_2$ separately if $P_1$ and $P_2$ are "distinct" enough and $\tau$ is close to 0.5. We, therefore, propose to use these local minimum points to detect $P_1$ and $P_2$.

This thesis is organized as follows. In Chapter 2, we make a review of divergence measures and estimation with the $\gamma$-divergence. Chapter 3 describes the application of the $\gamma$-divergence to cluster analysis, where some existing clustering methods are also discussed. In Chapter 4, we provide a brief summary of copulas and discuss the $\gamma$-estimation for the Gaussian copula model. Summary and discussion are given in Chapter 5.

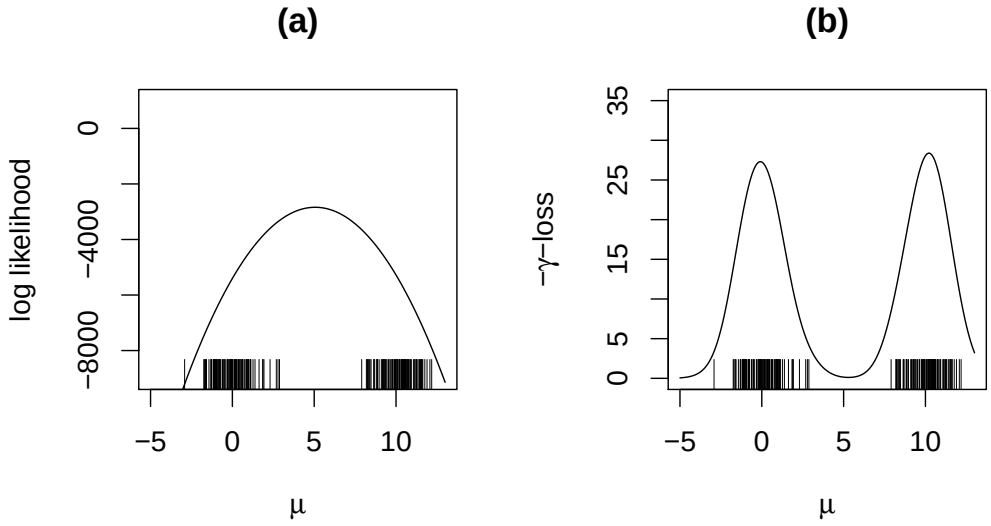



Figure 1.1: (a) Log likelihood function. (b) Minus $\gamma$-loss function $L_\gamma(\mu)$ ($\gamma = 1$). In (a) and (b), the data of size 200 is generated from the mixture of two standard normal distributions centered at 0 and 10, respectively.

# Chapter 2

# Minimum Divergence Estimation

## 2.1 Divergence Measures

In statistics, there are a number of indexes to measure the difference between two objects. Divergence measures reflect the difference between two probability distributions. They are defined by functionals which satisfy the following properties:

$$D(g, f) \geq 0 \text{ with equality if and only if } g \equiv f, \tag{2.1}$$

where $g$ and $f$ are probability density functions. Although a large number of divergence measures have been proposed in the literature, we present an introduction to two wide classes of divergence measures, Bregman divergence and $U$-divergence. The $\gamma$-divergence is derived from the $\beta$-divergence, which is one of the $U$-divergence.

**Bregman divergence**

Bregman (1967) introduced a family of divergence measures in the following way,

$$B_\psi(g, f) = \int \psi(f(x)) - \psi(g(x)) - \psi'(g(x))(f(x) - g(x))dx \tag{2.2}$$

for any differentiable convex function $\psi$ with $\psi(0) = \lim_{t\to 0} \psi(t) \in (-\infty, \infty)$. Note that $B_\psi(g, f)$ satisfies condition (2.1) due to the convexity of $\psi$ (see Figure 2.1).

**$U$-divergence**

The $U$-divergence is defined similar to the Bregman divergence (Murata et al. (2004)). Let $U$ be a differentiable and strictly convex function. Then its derivative $u = U'$ is a monotonic function, which has the inverse function $\xi = (u)^{-1}$. The $U$-divergence with respect to $U$ is defined as

$$\begin{aligned} D_U(g, f) &= \int U(\xi(f(x))) - U(\xi(g(x))) - U'(\xi(g(x)))(\xi(f(x)) - \xi(g(x)))dx \\ &= \int U(\xi(f(x))) - U(\xi(g(x))) - g(x)(\xi(f(x)) - \xi(g(x)))dx. \end{aligned} \tag{2.3}$$

We obtain the $U$-divergence by substituting $\psi = U, g(x) = \xi(g(x))$, and $f(x) = \xi(f(x))$ in (2.2). The advantage of the form (2.3) is allowing us to plug in the empirical distribution directly.

### $\beta$-divergence

When $U(t) = U_\beta(t) = \{1/(1+\beta)\}(1+\beta t)^{(1+\beta)/\beta}$ $(\beta > 0)$, the $U$-divergence becomes

$$D_{U_\beta}(g, f) = \frac{1}{\beta} \int \left\{ g(x)^\beta - f(x)^\beta \right\} g(x) dx - \frac{1}{1+\beta} \int g(x)^{1+\beta} - f(x)^{1+\beta} dx, \quad (2.4)$$

which is called the $\beta$-divergence (Basu et al., 1998; Minami and Eguchi, 2002).

### $\gamma$-divergence

The $\gamma$-divergence is derived from the $\beta$-divergence, which can lead more robust methods than the $\beta$-divergence (Jones et al., 2001; Fujisawa and Eguchi, 2008). The $\gamma$-divergence is defined as

$$D_\gamma(g, f) = -\kappa_\gamma \int g(x) f(x)^\gamma dx + \left( \int g(x)^{1+\gamma} dx \right)^{\frac{1}{1+\gamma}}, \quad (2.5)$$

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

The derivation of the $\gamma$-divergence is as follows. Consider the following optimization problem, $\underset{v>0}{\operatorname{argmin}} D_{U_\beta}(g, vf)$. The first derivative of $D_{U_\beta}(g, vf)$ with respect to $v$ becomes

$$\frac{d}{dv} D_{U_\beta}(g, vf) = -v^{\beta-1} \int f(x)^\beta g(x) dx + v^\beta \int f(x)^{1+\beta} dx.$$

Set the derivative to 0. Then we have

$$\min_{v>0} D_{U_\beta}(g, vf) = \frac{1}{\beta(1+\beta)} \left( \int g(x)^{1+\beta} dx - \frac{\left( \int f(x)^\beta g(x) dx \right)^{1+\beta}}{\left( \int f(x)^{1+\beta} dx \right)^\beta} \right). \quad (2.6)$$

The logarithmic of the ratio of the first and second terms of (2.6) is equal to

$$\begin{aligned} & \frac{1}{\beta(1+\beta)} \log \frac{(\int g(x)^{1+\beta}dx)(\int f(x)^{1+\beta}dx)^{\beta}}{(\int f(x)^{\beta}g(x)dx)^{1+\beta}} \\ = \; & \frac{1}{\beta}\left\{ -\log \int \frac{g(x)f(x)^{\beta}dx}{(\int f(x)^{1+\beta})^{\beta/(1+\beta)}} + \log \left( \int g(x)^{1+\beta}dx \right)^{1/(1+\beta)} \right\}. \end{aligned}$$

Then we consider the value

$$-\frac{1}{(\int f(x)^{1+\beta})^{\beta/(1+\beta)}} \int g(x)f(x)^{\beta}dx + \left( \int g(x)^{1+\beta}dx \right)^{1/(1+\beta)},$$

which corresponds to $D_{\gamma}(g, f)$ if $\beta = \gamma$. Note that $D_{\gamma}(g, f)$ satisfies condition (2.1) from this derivation.

## 2.2 $\gamma$-Estimation

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 $\theta$. The $\gamma$-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 $\gamma$-cross entropy and the Boltzmann-Shannon cross entropy have the following relation since $\kappa_\gamma(\theta)$ converges to 1 if $\gamma$ 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 $\gamma$-cross entropy can be regarded as an extension of the Boltzmann-Shannon cross entropy. The $\gamma$-entropy of $g$ is defined as $H_\gamma(g) = C_\gamma(g, g)$. Then the $\gamma$-divergence between $g$ and $f(\cdot, \theta)$ becomes

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

Recall that the $\gamma$-divergence $D_\gamma(g, f(\cdot, \theta))$ is nonnegative, and $D_\gamma(g, f(\cdot, \theta))$ is equal to $0$ if and only if $\theta$ 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 $\gamma$-divergence, we only have to find the global minimum point of $D_\gamma(g, f(\cdot, \theta))$ with respect to $\theta$, which is equal to that of $C_\gamma(g, f(\cdot, \theta))$.

The $\gamma$-loss function is defined by an estimator of the $\gamma$-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 $\gamma$-loss function for $f(\cdot, \theta)$ associated with the $\gamma$-divergence is given by

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

We extend the definition of the $\gamma$-cross entropy to any distributions. For any distribution function $G$, the $\gamma$-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 $L_\gamma(\theta)$ is equal to $C_\gamma(\hat{G}, f(\cdot, \theta))$ 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 $\gamma$-estimator of $\theta$ is defined by the global minimum point of $L_\gamma(\theta)$ (Eguchi and Kato, 2010). From the definition of the $\gamma$-estimator, it satisfies Fisher consistency. If the density function $g$ belongs to the statistical model $\{f(\cdot, \theta)\}$, then the $\gamma$-estimator satisfies asymptotic consistency and normality. The $\gamma$-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 $\gamma$-estimator can be seen as an

extension of the MLE.

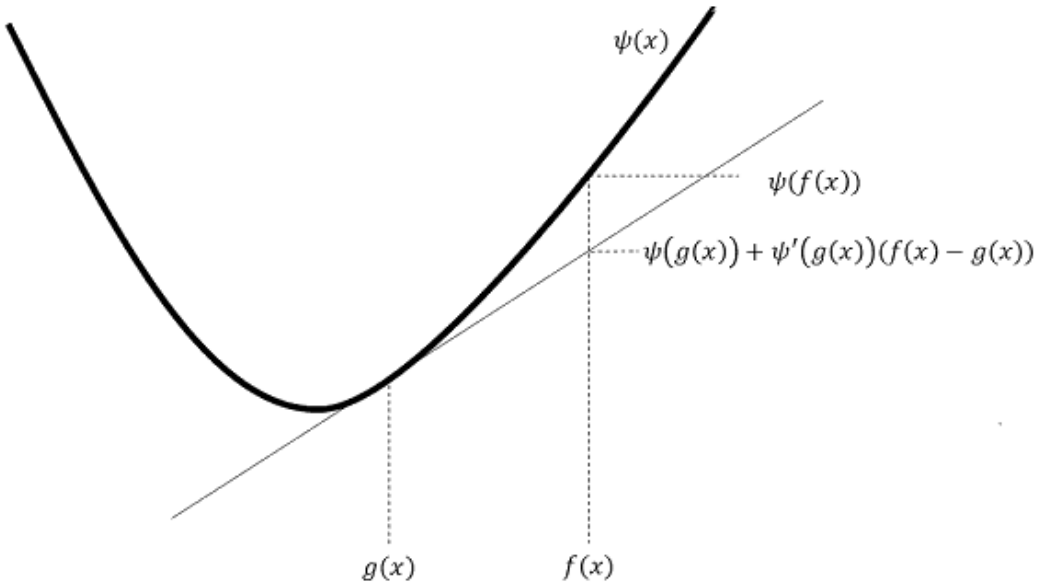



Figure 2.1: Illustration of the Bregman divergence.

# Chapter 3

# Cluster Analysis

## 3.1 Existing Methods

In this section, we make a review of three clustering algorithms, the $K$-means, model-based, and mean shift clustering. The model-based clustering is based on some parametric models while the $K$-means clustering considers the Euclidean distance among the data points. In the mean shift clustering, density estimates are used for finding cluster centers.

### 3.1.1 $K$-Means Clustering

The $K$-means clustering is to minimize the lack of homogeneity of each cluster based on the Euclidean distance. For the $K$-means algorithm, the number of clusters $K$ has been fixed by the investigator.

Let $\{x_1, x_2, \ldots, x_n\}$ be a data set, and $K$ be the number of clusters. The dispersion

matrix based on the data set is defined as

$$T_K = \sum_{k=1}^{K} \sum_{x \in C_k} (x - \bar{x})(x - \bar{x})^\top,$$

where $C_k$ is the $k$th cluster, and $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$. The dispersion matrix represents the total dispersion, and it can be decomposed into two matrices, the with-in cluster dispersion matrix $W_K$ and the between-cluster dispersion matrix $B_K$,

$$W_K = \sum_{k=1}^{K} \sum_{x \in C_k} (x - \bar{x}_k)(x - \bar{x}_k)^\top, \; B_K = \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top,$$

so that $T_K = W_K + B_K$, where $n_k$ is the number of objects in $C_k$, and $\bar{x}_k = (1/n_k) \sum_{x \in C_k} x$.

For cluster analysis, there are a lot of criteria based on $W_K$ and $B_K$, for example, minimization of $\det(W_K)$, and maximization of $\text{tr}(B_K W_K^{-1})$ (see Everitt et al. (2011) for more detailed discussion). The criterion for the $K$-means algorithm is minimization of $\text{tr}(W_K)$. This criterion is equivalent to minimization of the lack of homogeneity of clusters, that is,

$$\text{tr}(W_K) = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{x, x' \in C_k} \|x - x'\|^2.$$

In practice, the investigators will have to estimate the number of clusters in the data set. It is of great importance to select the number of clusters, because the clustering results may change drastically as the number of clusters increases. Two criteria will be shown for selection of the number of clusters for the $K$-means clustering.

**CH**

In Caliński and Harabasz (1974), a criterion $\mathrm{CH}(k)$ is defined by

$$\mathrm{CH}(k) = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)}.$$

This criterion is analogous to the $F$-statistic for analysis of variance in the univariate case. They propose to select the number of clusters $k$, which maximizes $\mathrm{CH}(k)$.

**Gap Statistic**

The within-cluster sum of squares $\mathrm{tr}\,(W_k)$ monotonically decreases as $k$ increases, but there exists $k^*$ such that for $k \geq k^*$, $\mathrm{tr}(W_k)$ decreases smaller than for $k < k^*$. Such a $k^*$ is used as an optimal value for the number of clusters. In Tibshirani et al. (2001), they provide a more sophisticated procedure to formulate this heuristic.

The gap statistic is defined by

$$\mathrm{Gap}_n(k) = E_n^*(\log(\mathrm{tr}(W_k))) - \log(\mathrm{tr}(W_k)),$$

where $E_n^*$ denotes the expectation under a sample size $n$ from a reference distribution. They propose the optimal value of $k$, which maximizes the gap statistic with taking the sampling distribution into account. Then $E_n^*(\log(\mathrm{tr}(W_k)))$ is calculated by a Monte Carlo approximation. In practice, the selected number of clusters is the smallest $k$ such that

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1},$$

where $s_{k+1}$ is a standard error.

### 3.1.2 Model-Based Clustering

A model-based clustering is to postulate a mixture density function for the population distribution, from which the data are sampled. Then the parameters in the mixture density function are estimated, and the posterior probabilities are calculated by plugging-in the estimators for the counterparts in the mixture densities. Each object is assigned to the cluster which maximizes the estimated posterior probability that the object is in the cluster.

Let $f_k(x, \theta_k)$ be a density function parametrized by $\theta_k$, and $g(x, \tau, \theta)$ be a mixture density function,

$$g(x, \tau, \theta) = \sum_{k=1}^{K} \tau_k f_k(x, \theta_k),$$

where $\tau = (\tau_1, \tau_2, \ldots, \tau_K)^\top$ and $\theta = (\theta_1^\top, \theta_2^\top, \ldots, \theta_K^\top)^\top$. Here $\tau_1, \tau_2, \ldots, \tau_K$ are the mixing proportions, that is, they are nonnegative and satisfy that $\sum_{k=1}^{K} \tau_k = 1$. The model-based clustering postulates $g(x, \tau, \theta)$ for the population distribution, and we estimate the parameter $\tau$ and $\theta$.

Although there are a lot of estimation methods, we focus on the maximum likelihood estimation. Since the log likelihood function for $g(x, \tau, \theta)$ is often very complicated, it is hard to calculate the maximum likelihood estimator (MLE) by using the log likelihood for $g(x, \tau, \theta)$. An alternative to obtain the MLE is the EM-algorithm (see Dempster et al. (1977)). From this, we focus on the situation where the component densities of the mixture density are normal. Let $\phi(x, \mu, \Sigma)$ be the density function of the normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$. We postulate $\phi(x, \mu_k, \Sigma_k)$ as the $k$th component

density of the mixture density function. Then the EM-algorithm for normal mixture is given as follows.

**EM-algorithm for normal mixture**

**Step 1** Set appropriate $\tau_1^{(0)}, \tau_2^{(0)}, \ldots, \tau_K^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_K^{(0)}, \Sigma_1^{(0)}, \Sigma_2^{(0)}, \ldots, \Sigma_K^{(0)}$.

**Step 2** Given $\tau_1^{(t)}, \ldots, \Sigma_K^{(t)}$, calculate $\tau_1^{(t+1)}, \ldots, \Sigma_K^{(t+1)}$ by the following update formula.

$$\begin{aligned}
\tau_k^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n}\tau_{ki}^{(t)}, \\
\tau_{ki}^{(t)} &= \frac{\tau_k^{(t)}\phi(x_i, \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K}\tau_j^{(t)}\phi(x_i, \mu_j^{(t)}, \Sigma_j^{(t)})}, \\
\mu_k^{(t+1)} &= \sum_{i=1}^{n}\frac{\tau_{ki}^{(t)}}{\sum_{j=1}^{n}\tau_{kj}^{(t)}}x_i, \\
\Sigma_k^{(t+1)} &= \sum_{i=1}^{n}\frac{\tau_{ki}^{(t)}}{\sum_{j=1}^{n}\tau_{kj}^{(t)}}(x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^\top.
\end{aligned}$$

**Step 3** Repeat Step 2 until all parameters converge.

For selection of $K$, we can use information criteria, such as AIC and BIC,

$$\begin{aligned}
\text{AIC} &= -2\sum_{i=1}^{n}\log f(x_i, \hat{\theta}, \hat{\tau}) + 2(\text{number of parameters}), \\
\text{BIC} &= -2\sum_{i=1}^{n}\log f(x_i, \hat{\theta}, \hat{\tau}) + \log(n)(\text{number of parameters}),
\end{aligned}$$

and select $k$ minimizing those criteria.

### 3.1.3 Mean Shift Clustering

A mean shift clustering (MSC) with the Gaussian kernel is to determine the cluster centers by the modes of the density estimate defined by

$$\hat{f}_h(x) = \frac{1}{nh^p} \sum_{i=1}^{n} \frac{\exp\left(-\frac{1}{2h^2}\|x - x_i\|^2\right)}{(2\pi)^{p/2}}. \tag{3.1}$$

We suppose $x_i^{(m)}$ is the position of $x_i$ at stage $m$ of the procedure, where $x_i^{(0)} = x_i$. Then $x_i^{(m)}$ is updated by

$$x_i^{(m+1)} = \sum_{j=1}^{n} \frac{\exp\left(-\|x_i^{(m)} - x_j\|^2/(2h^2)\right)}{\sum_{\ell=1}^{n} \exp\left(-\|x_i^{(m)} - x_\ell\|^2/(2h^2)\right)} x_j.$$

Note that each $x_i^{(m)}$ will converge to a mode of the density estimate defined by (3.1). The set $\{x_i^{(m)} : m \in \mathbb{N}\}$ is called the trajectory of $x_i$. Let $\{x_1^{(\infty)}, \ldots, x_n^{(\infty)}\}$ be $\{c_1, \ldots, c_{K_h}\}$. Then we define $c_k$ as the cluster center, and each $x_i$ is assigned to the cluster of which the center $c_k$ is equal to $x_i^{(\infty)}$.

For MSC, a bandwidth selection by Einbeck (2011) can be used. Suppose $c_{i,h}$ is the cluster center to which $x_i$ is assigned. The self-coverage for cluster analysis is defined as

$$S(h) = \frac{1}{n} \sum_{i=1}^{n} 1(\|x_i - c_{i,h}\| \leq h),$$

where $1(\cdot)$ is the indicator function. Assume we have evaluated $S(h)$ over a grid of band-

widths $h_1 < \cdots < h_L$. The curvature of $S(h)$ is approximated by

$$\triangle^2 S(h_\ell) = S(h_{\ell+1}) - 2S(h_\ell) + S(h_{\ell-1}).$$

Let $h_{(j)}$ be the bandwidth yielding the $j$th lowest of $\triangle^2 S(h_\ell), h = 1, \ldots, L$, under the constraint

$$S(h_\ell) > \max\{S(h_1), \ldots, S(h_{\ell-1}), s\},$$

where $s \in (0, 1)$ is a pre-determined constant. A value of $s = 1/3$ is recommended (Einbeck, 2011). Then $h_{(1)}$ is used as the optimal value.

## 3.2 Spontaneous Clustering

This section is based on the paper (Notsu et al., 2014). We begin with reconsidering the motivational example in the introduction from the point of view of cluster analysis. First, we consider a trivial situation, where the number of clusters is one. For example, assume that $x_1, \ldots, x_n$ in $\mathbb{R}^p$ follow a normal distribution with the mean vector $\mu$ and the identity covariance matrix. Then the log likelihood function multiplied by $-1/n$ is given by

$$L_0(\mu) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp\left(-\frac{1}{2}\|x_i - \mu\|^2\right)}{(2\pi)^{p/2}}\right),$$

which is equal to $1/(2n)(\sum_{i=1}^n \|x_i - \mu\|^2 + p\log 2\pi)$, where $\|\cdot\|$ denotes the Euclidean norm. The MLE of $\mu$ is just the sample mean, by which the cluster center can be determined. However, if there is more than one cluster, then the MLE does not work. We take another

estimator of $\mu$, the $\gamma$-estimator (Eguchi and Kato, 2010). In general, for a location family $\{f(x-\mu) : \mu \in \mathbb{R}^p\}$, where $f(x)$ is a probability density function, the $\gamma$-loss function is defined as

$$L_{\gamma,f}(\mu) = -\frac{1}{n}\kappa_\gamma \sum_{i=1}^{n} f(x_i - \mu)^\gamma, \tag{3.2}$$

where $\kappa_\gamma = \left(\int f(x)^{1+\gamma} dx\right)^{-\frac{\gamma}{1+\gamma}}$. If $f(x)$ is the normal density function with mean vector $0$ and the identity covariance matrix, then the $\gamma$-loss function becomes

$$L_\gamma(\mu) = -\frac{1}{n}\{(1+\gamma)(2\pi)^\gamma\}^{\frac{\gamma p}{2(1+\gamma)}} \sum_{i=1}^{n} \left( \frac{\exp\left(-\frac{1}{2}\|x_i - \mu\|^2\right)}{(2\pi)^{p/2}} \right)^\gamma, \tag{3.3}$$

where the subscript $f$ is omitted for simplicity. The $\gamma$-estimator of the normal mean $\mu$ is the value which minimizes $L_\gamma(\mu)$.

We consider a standard situation of $K$ clusters, where the density function of the population distribution has $K$ modes, for example,

$$g(x) = \sum_{k=1}^{K} \tau_k f_k(x),\ \sum_{k=1}^{K} \tau_k = 1,\ \tau_k > 0,\ k = 1, \ldots, K, \tag{3.4}$$

where $f_k(x)$ is a unimodal density function. As stated above, the MLE does not work in this situation. It is expected that the $\gamma$-loss function $L_\gamma(\mu)$ has $K$ local minimum points corresponding to $K$ mean vectors with respect to $f_1, \ldots, f_K$. Figure 1 (b) shows that $L_\gamma(\mu)$ has two local minimum points when the data have two clusters. Thus the cluster centers defined by the local minimum points lead to a clustering method similar to the $K$-means

clustering.

The proposed procedure appears to be similar to density-based clustering methods, for example, the mean shift clustering (Cheng, 1995), since the $\gamma$-loss function $L_\gamma(\mu)$ resembles the kernel density estimate with the Gaussian kernel (3.1). If $\mu = x$, and $h^2 = \gamma^{-1}$, then $L_\gamma(\mu)$ and $\hat{f}_h(x)$ are essentially the same, apart from a constant. Since the mean shift clustering defines the cluster centers by modes of the density estimate (3.1), the proposed procedure is the same with the mean shift clustering, that is, finding modes of equation (3.3).

There are some differences between them, however. We employ the $\gamma$-loss function, not density estimates, so that we will naturally estimate covariance structures of clusters by incorporating the $\gamma$-loss function for the covariance matrix of the normal distribution. In addition, we will propose the selection for the power index $\gamma$ based on the theory of the $\gamma$-loss function, which also gives a new insight or different view to the selection of the bandwidth $h$ for the density estimation. $L_\gamma(\mu)$ is a loss function for the normal mean $\mu$; $\hat{f}_h(x)$ is a density estimate obtained by smoothing the histogram in terms of the Gaussian kernel function. In general, kernel density estimates are given by

$$\frac{1}{nh^p}\sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right), \tag{3.5}$$

where $W$ is a kernel function. Two equations (3.2) and (3.5) are quite different forms derived from different ideas.

### 3.2.1 $\gamma$-Loss Function for the Normal Distribution

We consider the $\gamma$-loss function for the normal distribution with the mean vector $\mu$ and the covariance matrix $\Sigma$,

$$L_\gamma(\mu, \Sigma) = -\det \Sigma^{-\frac{\gamma}{2(1+\gamma)}} \sum_{i=1}^{n} \exp\left(-\frac{\gamma}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu)\right), \tag{3.6}$$

apart from a constant. In the remainder of the thesis, we omit a constant term that does not affect the optimization. An iteration algorithm to find the local minimum points of $L_\gamma(\mu, \Sigma)$ has been proposed in Fujisawa and Eguchi (2008) and Eguchi and Kato (2010). It is obtained by differentiating $L_\gamma(\mu, \Sigma)$ with respect to $\mu$ and $\Sigma^{-1}$ 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 $\gamma$-loss function monotonically as the iteration step $t$ increases. It is described below.

**Step 1** Set appropriate $\mu_0$ and $\Sigma_0$ as initial values.

**Step 2** Given $\mu_t$ and $\Sigma_t$, calculate $\mu_{t+1}$ and $\Sigma_{t+1}$ by the following update formulas,

$$\mu_{t+1} = \sum_{i=1}^{n} w_\gamma(x_i, \mu_t, \Sigma_t) x_i, \tag{3.7}$$

$$\Sigma_{t+1} = (1+\gamma) \sum_{i=1}^{n} w_\gamma(x_i, \mu_t, \Sigma_t)(x_i - \mu_{t+1})(x_i - \mu_{t+1})^\top, \tag{3.8}$$

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 $\varepsilon$, 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 (3.7) and (3.8) are equal to the sample mean vector and covariance matrix, respectively, which are just the MLEs. If our aim is to obtain the local minimum points of $L_\gamma(\mu)$, then we only have to update $\mu_t$ and fix $\Sigma_t$ to be the identity matrix $I$. Similarly, if our aim is to obtain the local minimum points of $L_\gamma(\mu, \Sigma)$ with fixed $\mu$, then we only have to update $\Sigma_t$ and fix $\mu_t = \mu$.

### 3.2.2 Spontaneous Clustering Algorithm

In general, the spontaneous clustering based on a density function $f(x, \theta)$ with parameter $\theta$ 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 $\gamma$-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.

As a special case, the spontaneous clustering based on the normal distribution is defined as follows. We set $\Theta_\mu$ and $\Theta_{(\mu,\Sigma)}$ to be the empty sets at the start of the algorithm. The algorithm of section 3.2.1 is employed in the spontaneous clustering below.

**Spontaneous Clustering Based on the Normal Distribution**

**Step 1-1** If $\Theta_\mu$ 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 section 3.2.1 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 $\Theta_\mu$.

**Step 1-3** Repeat Step 1-1 and 1-2 until the number of elements in $\Theta_\mu$ 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 $\Sigma$, denoted by $\hat{\Sigma}$, with the algorithm in section 3.2.1. 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

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

The centers and the covariance matrices of clusters are defined as $(\hat{\mu}_k, \hat{\Sigma}_k)$. In the remainder of this chapter, we focus on the spontaneous clustering based on the normal distribution.

### 3.2.3 Selection Procedure for $\gamma$

The value of the power index $\gamma$ plays a key role in the spontaneous clustering, because $\gamma$ affects the number of clusters obtained by the spontaneous clustering. We propose two methods to select the value of $\gamma$. One is a heuristic choice of $\gamma$ 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 $\mu_1$ and $\mu_2$ with the identity covariance matrix and the equal mixing proportions, respectively. Our simulation result suggests that if $\|(\mu_1 - \mu_2)/2\| = 3\sqrt{2}/2 \doteq 2.12$, then the value of $\gamma$ needs to be greater than or equal to 1 for two local minimum points of $L_\gamma(\mu)$ to exist. Proposition 3.2.1 states 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 $\gamma$ needs to be greater 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 $\gamma$ defined as

$$\hat{\gamma} = \left( \frac{r}{\frac{3\sqrt{2}}{2}} \right)^{-2} = \frac{9}{2r^2}. \tag{3.9}$$

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 shown in Figure 3.1. 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 a complete theoretical background is missing.

We also propose a more sophisticated method based on AIC. The value of $\gamma$ which minimizes AIC is recommended as the optimal selection of $\gamma$. Let $K_\gamma$ 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 the mean vector $\mu$ and the covariance matrix $\Sigma$. Then $\phi(x, \hat{\mu}_{\gamma k}, \hat{\Sigma}_{\gamma k})$ serves as a density estimator of the mixture component $f_k(x)$ in (3.4). 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 (3.4),

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

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

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

We claim that the value of $\gamma$ that minimizes $\text{AIC}_\gamma$ is the optimal one.

### 3.2.4 Behavior of the $\gamma$-Loss Function

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

#### The Existence of Local Minimum Points

In this section, we consider the conditions for the existence of local minimum points of $L_\gamma(\mu)$. As we discussed in section 3.2.2, the cluster centers are defined at the local minimum points of $L_\gamma(\mu)$, so it is important to know when the $\gamma$-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 the 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 ease of calculation, we consider $n = \infty$. As $n$ tends to $\infty$, $L_\gamma(\mu)$ almost surely converges to the $\gamma$-cross entropy defined by

$$C_\gamma(g, \phi(\cdot, \mu, I)) = -\kappa_\gamma \int g(x)\phi(x, \mu, I)^\gamma dx, \tag{3.10}$$

where $\kappa_\gamma = \left(\int \phi(x, 0, I)^{1+\gamma} dx\right)^{-\frac{\gamma}{1+\gamma}}$. Section 2.2 contains a general introduction to the

$\gamma$-cross entropy. Then $C_\gamma(g, \phi(\cdot, \mu, I))$ 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 just minus the 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 3.2 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 $\gamma$. 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 3.2.1** *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:*

$$d > 0, \tag{3.11}$$

$$\exp\left(\frac{2\gamma}{1+\gamma\sigma^2}\|\nu\|\sqrt{d}\right) > \frac{\gamma}{1+\gamma\sigma^2}\left(\|\nu\| + \sqrt{d}\right)^2 \frac{\tau_1}{\tau_2}, \tag{3.12}$$

$$\exp\left(-\frac{2\gamma}{1+\gamma\sigma^2}\|\nu\|\sqrt{d}\right) < \frac{\gamma}{1+\gamma\sigma^2}\left(\|\nu\| - \sqrt{d}\right)^2 \frac{\tau_1}{\tau_2}. \tag{3.13}$$

*In particular, if $\tau_1 = \tau_2$, then (3.12) and (3.13) hold for any $d > 0$. When the two local*

*minimum points exist, they lie on the segment between $\mu_1$ and $\mu_2$. The one closer to $\mu_1$ and the other closer to $\mu_2$ are denoted by $\mu_1^*$ and $\mu_2^*$, respectively. Then $\|\mu_1 - \mu_1^*\|$ and $\|\mu_2 - \mu_2^*\|$ are bounded above by*

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

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

$$\begin{aligned}\frac{\partial C_\gamma(g, \phi(\cdot, \mu, I))}{\partial \mu} \quad \propto \quad & \tau_1 \phi(\mu, \mu_1, (\sigma^2 + 1/\gamma)I)(\mu - \mu_1) \\ & + \tau_2 \phi(\mu, -\mu_1, (\sigma^2 + 1/\gamma)I)(\mu + \mu_1). \end{aligned} \tag{3.14}$$

From (3.14), every local minimum point of $C_\gamma(g, \phi(\cdot, \mu, I))$ should exist on the segment between $-\mu_1$ and $\mu_1$. The Hessian matrix of $C_\gamma(g, \phi(\cdot, \mu, I))$ is given by

$$\begin{aligned}\frac{\partial^2 C_\gamma(g, \phi(\cdot, \mu, I))}{\partial \mu \partial \mu^\top} \quad \propto \quad & -\tau_1 \phi(\mu, \mu_1, (\sigma^2 + 1/\gamma)I) \frac{\gamma}{1 + \sigma^2 \gamma} (\mu - \mu_1)(\mu - \mu_1)^\top \\ & -\tau_2 \phi(\mu, -\mu_1, (\sigma^2 + 1/\gamma)I) \frac{\gamma}{1 + \sigma^2 \gamma} (\mu + \mu_1)(\mu + \mu_1)^\top \\ & +\tau_1 \phi(\mu, \mu_1, (\sigma^2 + 1/\gamma)I) I \\ & +\tau_2 \phi(\mu, -\mu_1, (\sigma^2 + 1/\gamma)I) I. \end{aligned} \tag{3.15}$$

Let $\mu(t) = t\mu_1$. From (3.15), $\mu(t)$ 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

$$\begin{aligned} & \frac{d}{dt}C_\gamma(g, \phi(\cdot, \mu(t), I)) > 0 \\ \iff & \exp\left(-C(t+1)^2 + C(t-1)^2\right) > \frac{(1-t)\tau_1}{(t+1)\tau_2} \\ \iff & -4Ct + \log(t+1) - \log(1-t) - \log\frac{\tau_1}{\tau_2} > 0. \end{aligned} \tag{3.16}$$

Let $h(t)$ be the left hand side of inequality (3.16). 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

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

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

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

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

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

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 $1 - 1/(2C) > 0$

holds. □

If $\mu_1$ and $\mu_2$ are distinct enough, then conditions (3.11), (3.12), and (3.13) hold. Condition (3.11) means that the distance between $\mu_1$ and $\mu_2$ should be large for the existence of two local minimum points; condition (3.12) and (3.13) mean that if $\tau_1 \neq \tau_2$, then the distance between $\mu_1$ and $\mu_2$ should be larger compared to the case when $\tau_1 = \tau_2$.

By proposition 3.2.1, for any $\sigma^2$, if $\mu_1$ and $\mu_2$ are distinct enough, then there exists $\gamma$ 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, although the center of a cluster $\mu_k^*$ does not coincide with the normal mean $\mu_k$ ($k = 1, 2$), it becomes arbitrarily close to $\mu_k$, when $\|\mu_1 - \mu_2\|$ becomes large.

### 3.2.5 Comparison among Spontaneous Clustering and Existing Methods

In this section, we clarify the differences among the three clustering methods, the spontaneous clustering based on the normal distribution, the mean shift clustering with the Gaussian kernel, and the $K$-means clustering. For a given number of clusters $K$, the $K$-means clustering determines the cluster centers $c_1, \ldots, c_K$ by

$$\underset{c_1,\ldots,c_K \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{k \in \{1,\ldots,K\}} \|x_i - c_k\|^2. \tag{3.17}$$

Each $x_i$ is assigned to the cluster of which the center $c_k$ is the nearest to $x_i$ in terms of the Euclidean distance.

To find the cluster centers, the spontaneous clustering and the mean shift clustering use the modes of the same function since $L_\gamma(\mu)$ and $\hat{f}_h$ are essentially the same apart from a constant. On the other hand, the $K$-means clustering uses the minimum point defined by (3.17). After determining the cluster centers, to assign the data to clusters, the $K$-means clustering uses the Euclidean distance, but the spontaneous clustering uses the Mahalanobis distance. The mean shift clustering employs the mean shift trajectories for assignment. Table 3.1 summarizes the comparison among the three clustering methods.

## 3.3 Simulation and Data Analysis

### 3.3.1 Simulation 1: The Case of Spherical Clusters

We demonstrate the performance of the spontaneous clustering (SC) in comparison with the $K$-means clustering and the mean shift clustering (MSC). In this simulation, we suppose that the covariance matrices of clusters are known to be the identity matrix. Hence, in SC, the covariance matrices of clusters were not estimated and fixed to be the identity matrix. The performance of clustering is measured by BHI as defined below.

The value of $\gamma$ for SC was determined by the range of data (R) and AIC described in section 3.2.3 and a heuristic bandwidth selection (HBS) in kernel density estimation. The value selected by HBS is given by

$$\hat{\gamma} = \left(\frac{n(p+2)}{4}\right)^{\frac{2}{p+4}} / \hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is the average of sample variances for each variable (see Silverman (1986) page 87). For the $K$-means clustering, the method by Caliński and Harabasz (1974) and the gap statistic by Tibshirani et al. (2001) were used to fix the number of clusters.

We considered four different simulation settings with the sample size 200. The samples were generated from the mixture of five standard normal distributions with

(a) the mean vectors $(0,0,0,0,0)^\top, (4,4,4,4,4)^\top, (-4,-4,4,4,4)^\top$, $(4,-4,-4,4,4)^\top$, and $(4,4,-4,-4,4)^\top$, and equal mixing proportions;

(b) the same mean vectors as (a) but different mixing proportions 0.025, 0.025, 0.375, 0.375, and 0.2;

(c) the mean vectors $(0,0,0,0,0,0,0,0,0,0)^\top, (4,4,4,4,4,4,4,4,4,4)^\top$, $(-4,-4,-4,4,4,4,4,4,4,4)^\top, (4,-4,-4,-4,4,4,4,4,4,4)^\top$, and $(4,4,-4,-4,-4,4,4,4,4,4)^\top$, and equal mixing proportions;

(d) the same mean vectors as (c) but different mixing proportions 0.025, 0.025, 0.375, 0.375, and 0.2.

Figure 3.3 displays a sample from (a). We simulated 100 runs for each setting and compared clustering results from SC with those from the $K$-means clustering and MSC. Figure 3.4 shows the value of AIC and the number of clusters resulting from SC for the sample in Figure 3.3. The selected value of $\gamma$ based on AIC was $0.15$. To measure the performance of clustering, we used the biological homogeneity index (BHI) (Wu, 2011), which measures the homogeneity between the cluster $\mathcal{C} = \{C_1, \ldots, C_K\}$ and the true category

$\mathcal{B} = \{B_1, \ldots, B_L\}$,

$$\mathrm{BHI}(\mathcal{C}, \mathcal{B}) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k(n_k - 1)} \sum_{i \neq j, i, j \in C_k} 1(B^{(i)} = B^{(j)}), \tag{3.18}$$

where $B^{(i)} \in \mathcal{B}$ is the category for the observation $x_i$, and $n_k$ is the number of the observations in $C_k$. In this simulation, $L$ is equal to the number of components of the normal mixture, and $B^{(i)}$ corresponds to its component label in the data generation. This index is bounded above by 1, which means the perfect homogeneity between the clusters and the true categories. We should check the estimated number of clusters as well as BHI, since, occasionally, the value of BHI becomes 1, even though the estimated number of clusters is larger than the true number of clusters.

Table 3.2 displays the frequency of choosing $K$ clusters, the mean value and the standard deviation (SD) of BHI over 100 runs. When the mixing proportions are not equal, the $K$-means with CH, $K$-means with Gap, or MSC do not detect the correct number of clusters. When the mixing proportions are equal, all the clustering methods work well. SC with R and SC with HBS behave almost similarly in these simulation settings. They detect the correct number of clusters in the low dimension case, but do not in the high dimension case with different mixing proportions. SC with AIC can detect the correct number of clusters in all the settings.

### 3.3.2 Simulation 2: The Case of Ellipsoidal Clusters

We demonstrate the performance of SC in comparison with the model-based clustering (MBC) with normal components and MSC. We suppose that the covariance matrices of clusters are heterogeneous and unknown. Hence, in SC, the covariance matrices of clusters were also estimated. The value of $\gamma$ for SC was determined by AIC, and the number of clusters for MBC was determined based on AIC and the Bayesian information criterion (BIC). For MSC, the bandwidth $h$ was determined by the self-coverage.

We considered two different simulation settings with the sample size 400. The samples were generated from the mixture of five normal distributions

(a) with mean vectors $(0,0,0,0,0)^\top$, $(6,6,6,6,6)^\top$, $(6,-6,-6,6,6)^\top$, $(6,6,-6,-6,6)^\top$, and $(6,6,6,-6,-6)^\top$, covariance matrices $S_1$, $S_2$, $S_1$, $S_2$, and $S_1$, where

$$S_1 = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}, S_2 = \begin{pmatrix} 2 & -0.3 & -0.3 & -0.3 & -0.3 \\ -0.3 & 2 & -0.3 & -0.3 & -0.3 \\ -0.3 & -0.3 & 2 & -0.3 & -0.3 \\ -0.3 & -0.3 & -0.3 & 2 & -0.3 \\ -0.3 & -0.3 & -0.3 & -0.3 & 2 \end{pmatrix},$$

and mixing proportions 0.2, 0.2, 0.2, 0.2, and 0.2;

(b) with the same mean vectors and covariance matrices as (a) but different mixing proportions 0.05, 0.05, 0.35, 0.35, and 0.2.

Figure 3.5 displays a sample from (a), and Figure 3.6 shows the value of AIC and the number of clusters resulting from SC for the sample. Note that we used two values $\gamma_1$ and $\gamma_2$ as the power index $\gamma$, where $\gamma_1$ was used for $L_\gamma(\mu)$ when defining the cluster centers, and $\gamma_2$ for $L_\gamma(\mu, \Sigma)$ when defining the covariance matrices. The selected values of $\gamma_1$ and $\gamma_2$ for the sample in Figure 3.5 were $\gamma_1 = 0.1$ and $\gamma_2 = 0.2$. We simulated 100 runs for each simulation setting and compared the clustering result from SC with those from MBC and MSC.

Table 3.3 displays the frequency of choosing $K$ clusters, the mean value and SD of BHI over 100 runs. When the mixing proportions are equal, all clustering methods without MBC with AIC can detect five clusters well. When the mixing proportions are not equal, SC with AIC can detect five clusters as in the case of spherical clusters. On the other hand, other clustering methods do not detect the correct number of clusters well. We observed that SC can capture the ellipsoidal cluster structures.

### 3.3.3 Data Analysis

To evaluate the practical performance of SC, we applied it with the fixed identity covariance matrix to real data as well as the $K$-means clustering and MSC. 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 3.7 shows the scatterplot matrix of the data. 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 $\gamma$ for SC was determined by the three methods based on R, AIC, and HBS. The number of clusters for the $K$-means clustering was determined by CH and Gap. The bandwidth for MSC was selected by the self-coverage.

Table 3.4 shows the clustering results. The value of AIC and the number of clusters by SC with AIC are shown in Figure 3.8 (a). SC with R and SC with AIC detect properly three clusters, while SC with HBS does not. In particular, the clustering result from SC with R is the most accurate. 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 Figure 3.8 (b) and (c). The value of CH does not decrease after some number of clusters, so CH does not work well for these data. The $K$-means with Gap detects more than three clusters, while MSC detects properly three clusters and assigns the data perfectly. As a result, we observed SC based on R and AIC and MSC can detect three clusters properly and partition the 45 specimens into clusters corresponding to the three regions.

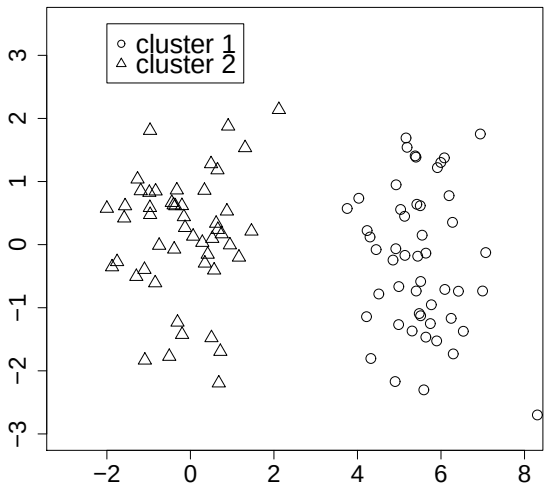



Figure 3.1: Sample generated from the mixture of two normal distributions centered at $(0,0)^\top$ and $(5,0)^\top$ with the identity covariance matrix, respectively.

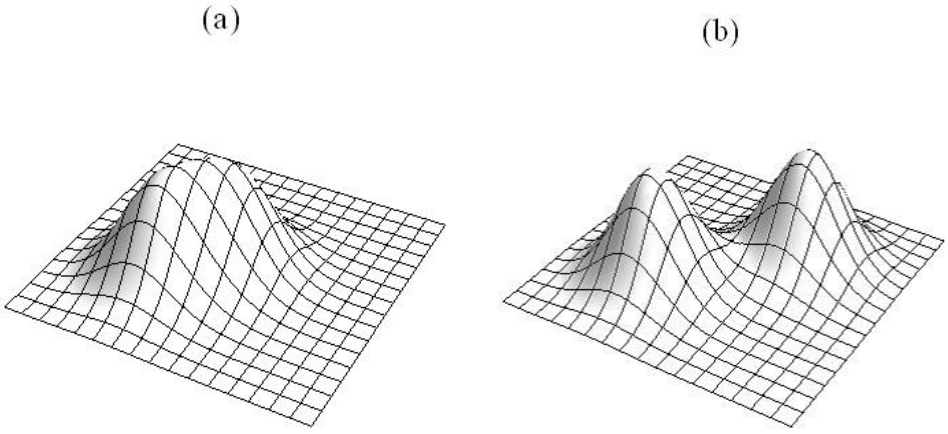



Figure 3.2: Illustration of $-C_\gamma(g, \phi(\cdot, \mu, I))$. In (a), $\mu_1 = (0,0)^\top, \mu_2 = (2,2)^\top, \tau_1 = \tau_2 = 0.5, \gamma = 1, \sigma^2 = 1$. In (b), $\mu_1 = (0,0)^\top, \mu_2 = (4,4)^\top, \tau_1 = \tau_2 = 0.5, \gamma = 1, \sigma^2 = 1$.

Table 3.1: Comparison among Spontaneous Clustering (SC), Mean Shift Clustering (MSC), and $K$-Means Clustering.

| | SC | MSC | $K$-Means |
|---|---|---|---|
| Cluster Center | modes of $-L_\gamma(\mu)$ | modes of $\hat{f}_h$ | equation (3.17) |
| Assignment | Mahalanobis distance | trajectories | Euclidean distance |

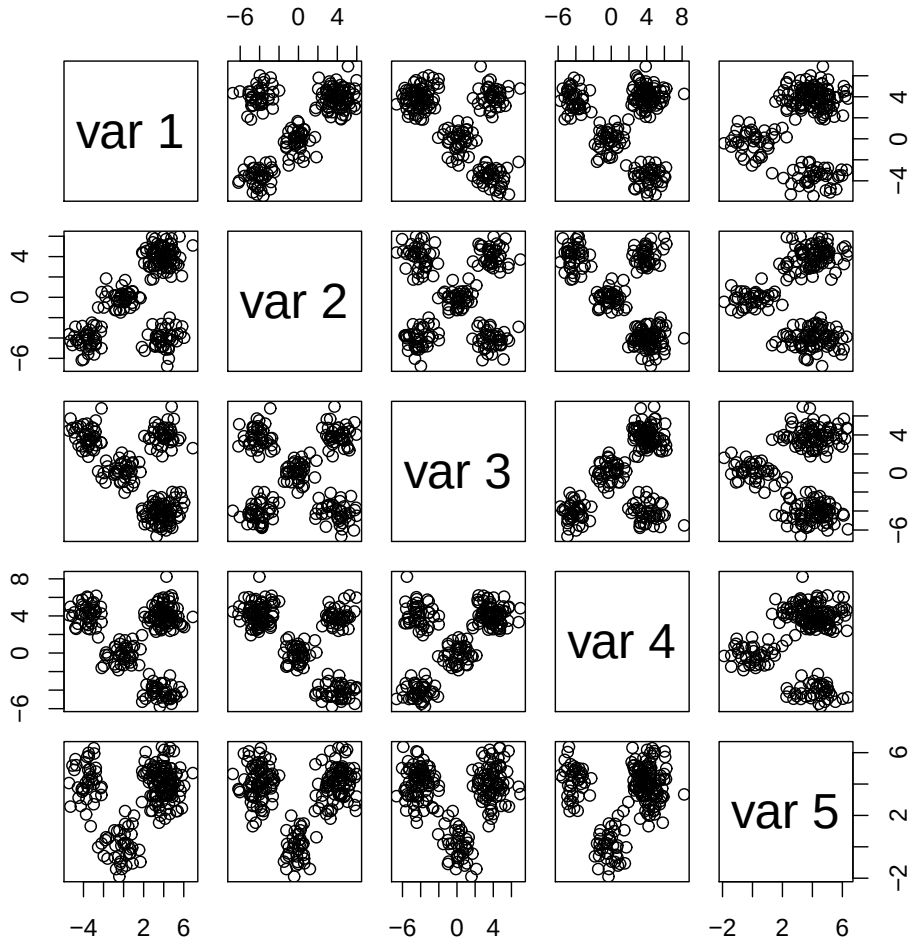



Figure 3.3: Scatterplot matrix of a sample.

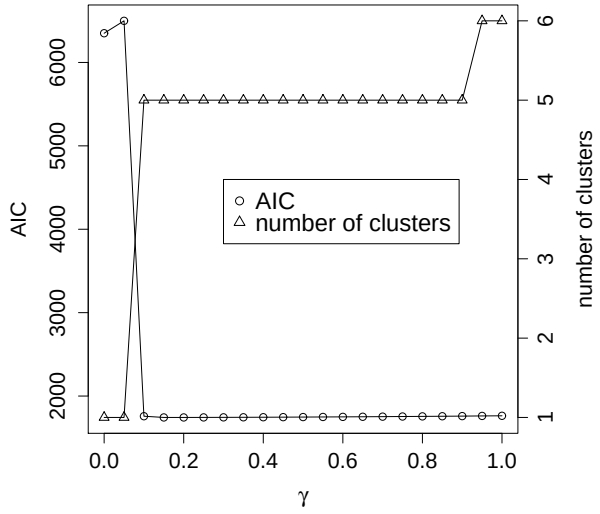



Figure 3.4: The value of AIC and the number of clusters.

Table 3.2: Results of the Clustering.

| Scenario | Method | $K$ | | | | | | BHI (SD) |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | more | |
| (a) | SC with R | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | SC with HBS | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | SC with AIC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | $K$-means with CH | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | $K$-means with Gap | 9 | 0 | 0 | 0 | 91 | 0 | 0.93 (0.23) |
| | MSC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| (b) | SC with R | 0 | 0 | 0 | 0 | 93 | 7 | 1.00 (0.00) |
| | SC with HBS | 0 | 0 | 0 | 0 | 98 | 2 | 1.00 (0.00) |
| | SC with AIC | 0 | 0 | 0 | 0 | 98 | 2 | 1.00 (0.00) |
| | $K$-means with CH | 0 | 0 | 0 | 100 | 0 | 0 | 0.86 (0.00) |
| | $K$-means with Gap | 0 | 0 | 0 | 100 | 0 | 0 | 0.86 (0.00) |
| | MSC | 0 | 0 | 25 | 30 | 44 | 1 | 0.97 (0.03) |
| (c) | SC with R | 0 | 0 | 0 | 0 | 91 | 9 | 1.00 (0.00) |
| | SC with HBS | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | SC with AIC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | $K$-means with CH | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | $K$-means with Gap | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | MSC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| (d) | SC with R | 0 | 0 | 0 | 0 | 7 | 93 | 1.00 (0.00) |
| | SC with HBS | 0 | 0 | 0 | 0 | 11 | 89 | 1.00 (0.00) |
| | SC with AIC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | $K$-means with CH | 0 | 0 | 0 | 100 | 0 | 0 | 0.86 (0.00) |
| | $K$-means with Gap | 9 | 0 | 0 | 91 | 9 | 0 | 0.87 (0.04) |
| | MSC | 0 | 1 | 13 | 80 | 6 | 0 | 0.96 (0.04) |

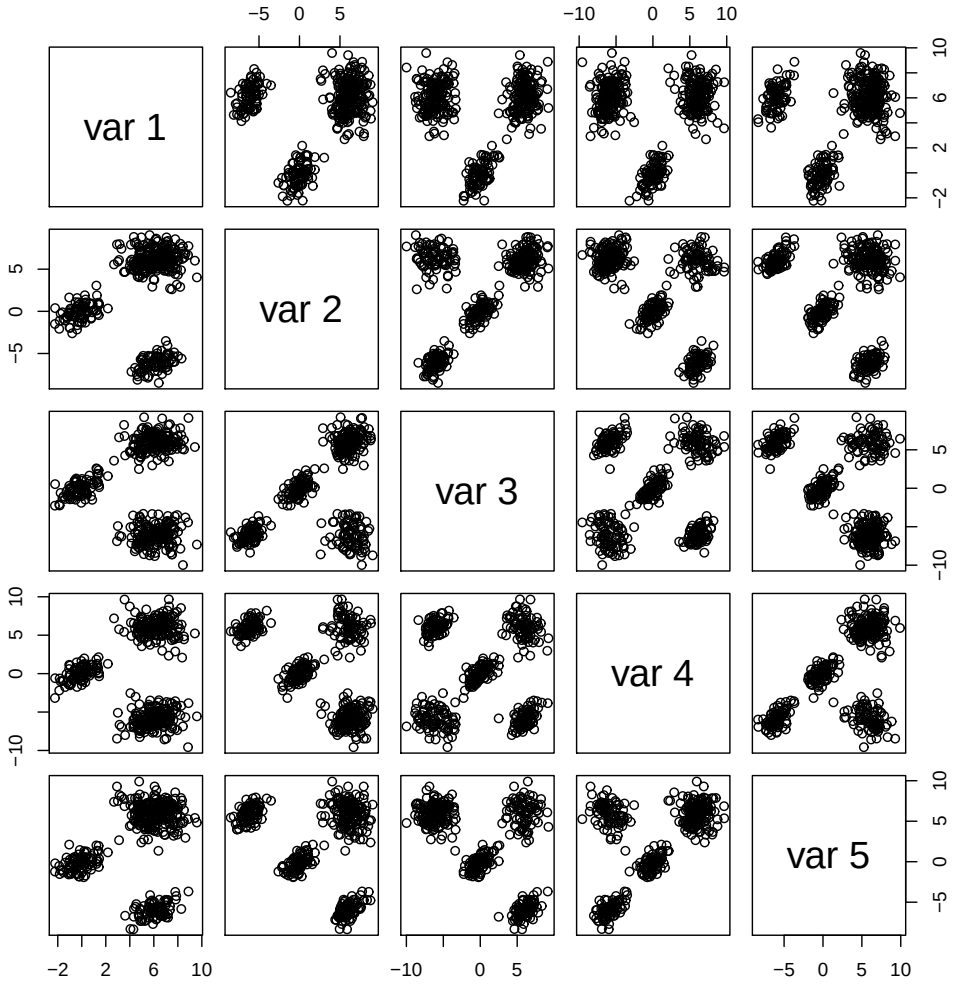
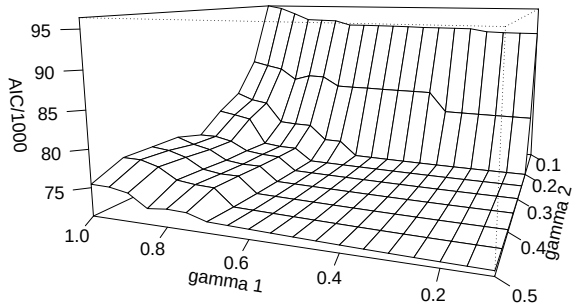



Figure 3.5: Scatterplot matrix of a sample.

**(a)**




**(b)**

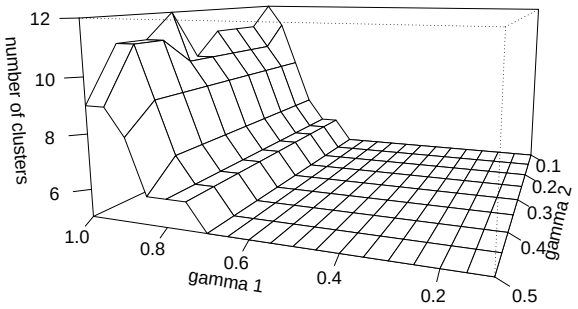



Figure 3.6: (a) The value of AIC. (b) The number of clusters.

Table 3.3: Results of the Clustering.

| Scenario | Method | $K$ | | | | | | BHI (SD) |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | more | |
| (a) | SC with AIC | 0 | 0 | 0 | 0 | 99 | 1 | 1.00 (0.00) |
| | MBC with AIC | 0 | 0 | 0 | 0 | 64 | 36 | 1.00 (0.00) |
| | MBC with BIC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| | MSC | 0 | 0 | 0 | 0 | 100 | 0 | 1.00 (0.00) |
| (b) | SC with AIC | 0 | 0 | 0 | 0 | 98 | 2 | 1.00 (0.01) |
| | MBC with AIC | 0 | 0 | 0 | 0 | 55 | 45 | 1.00 (0.00) |
| | MBC with BIC | 0 | 0 | 0 | 10 | 90 | 0 | 0.99 (0.04) |
| | MSC | 0 | 0 | 8 | 4 | 88 | 0 | 0.99 (0.04) |

Table 3.4: Results of the Clustering.

| Method | Number of clusters | BHI |
|---|---|---|
| SC with R | 3 | 1.00 |
| SC with HBS | 4 | 0.89 |
| SC with AIC | 3 | 0.96 |
| $K$-means with CH | - | - |
| $K$-means with Gap | 4 | 0.88 |
| MSC | 3 | 1.00 |

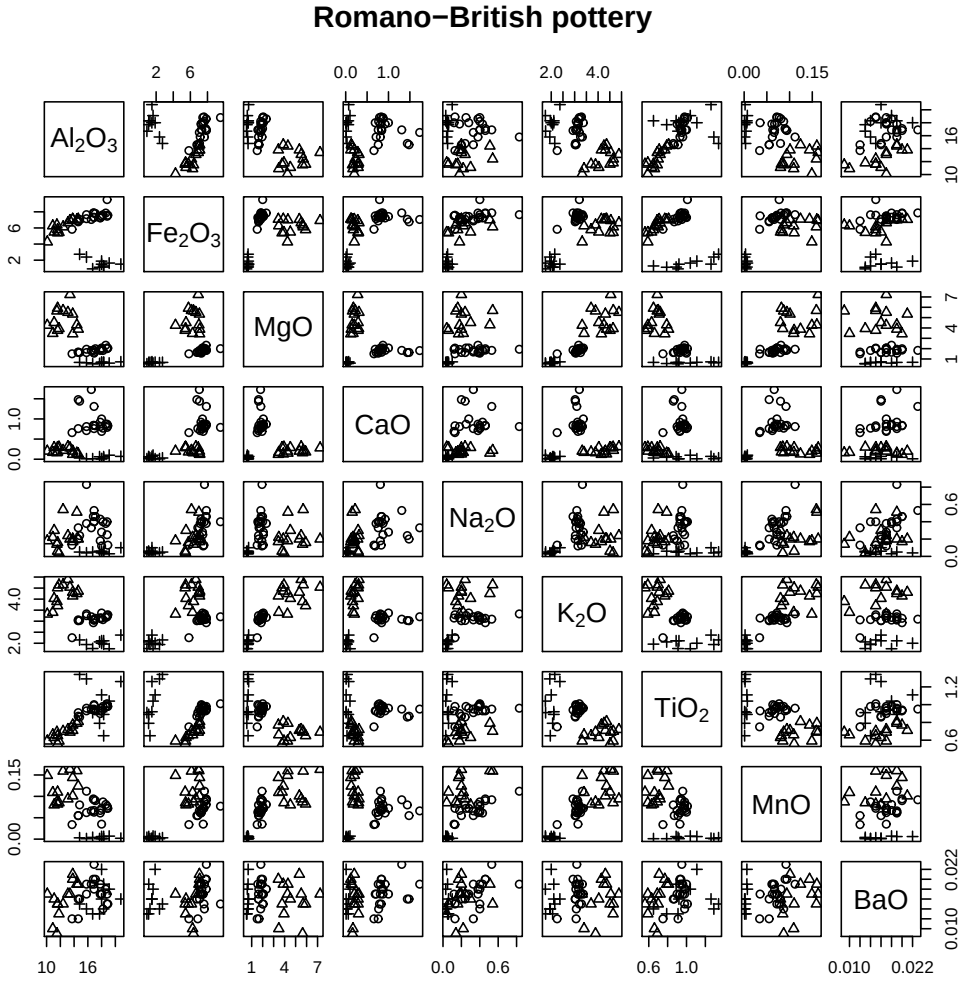



Figure 3.7: Scatterplot matrix of data on Romano-British pottery. ○, △, and + correspond to the three regions.

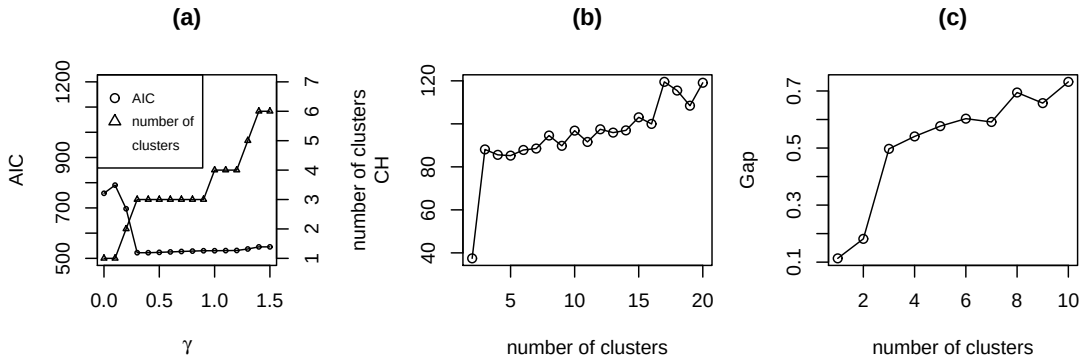



Figure 3.8: (a) AIC and number of clusters. (b) CH. (c) Gap.

# Chapter 4

# Detection of Heterogeneous Correlation Structure

## 4.1 Copulas

A copula is a multivariate distribution function, which determines the correlation or dependence structure of a distribution. Recently, applications of copulas have been increasing due to the simple structure given by the Sklar's theorem. In this section, we will show the backgrounds and basic properties of copulas.

### 4.1.1 Backgrounds

The history of copulas has started about 60 years ago in the study of multivariate distributions with fixed univariate marginals. The term "copula" was employed by Sklar (1959) for the first time. In a theorem named by Sklar's theorem, copulas combine a joint distri-

bution and its marginal distributions. Copulas had been researched in terms of probability theory rather than statistics at first. However since the middle of seventies, copulas have become popular little by little for statisticians. In 1990, the first conference devoted to copulas was held (Dall'Aglio et al., 1991), and a book about copulas was published (Nelsen, 1999), which were to become a standard reference in copula theory. Nowadays, copulas are widely applied to a lot of fields. For example, Song et al. (2009) make use of the Gaussian copula to combine some generalized linear models, one for each response variable, and they apply the proposed method to medical data. In Baŕdossy (2006), groundwater quality is analyzed, where the joint distribution of two observations obtained from different points is represented by a copula.

Finance and risk management are the most active disciplines to apply copulas. In these fields, we often meet problems where there are a lot of products and the modeling of the joint distribution of their values is crucial. Copulas enable us to model their joint distribution flexibly, so copulas have been popular in these fields. For example, McNeil et al. (2005), which is a standard reference for quantitative risk management, devotes one chapter for copulas. In Li (2001), the Gaussian copula was employed to price a new financial instrument "Credit Default Swaps". The seller of the swap agrees to pay off a third party debt if this party defaults. The purchaser of the swap makes payment for this insurance. The joint distribution of default time of the seller and the party was represented by the Gaussian copula.

### 4.1.2 Definitions and Basic Properties

A copula is a multivariate distribution function with standard uniform univariate marginal distributions. For example, if a random vector $x = (x_1, \ldots, x_p)$ has continuous univariate marginal distributions $F_1, \ldots, F_p$, then the distribution function of $(F_1(x_1), \ldots, F_p(x_p))$ is a copula. Copulas combine a joint distribution with its univariate marginal distributions as shown in Sklar's theorem.

**Theorem 4.1.1 (Sklar's theorem)** *Let $F$ be a $p$-dimensional joint distribution function, and $F_1, \ldots, F_p$ be its univariate marginal distribution functions. Then there exists a copula $C$ such that*

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p)), \tag{4.1}$$

*for all $x \in \mathbb{R}^p$. If $F_1, \ldots, F_p$ are continuous, then $C$ is unique. Conversely, if $C$ is a copula and $F_1, \ldots, F_p$ are univariate distribution functions, then the function $F$ defined in (4.1) is a joint distribution function with univariate marginal distribution functions $F_1, \ldots, F_p$.*

By the Sklar's theorem, we can specify a copula and univariate marginals separately in order to construct a multivariate distribution. This is one of the advantages obtained by using copulas.

Although the role of the copula in equation (4.1) is not clear, it determines the correlation or dependence structure of the multivariate distribution as shown below. Let $x = (x_1, \ldots, x_p)^\top$ be a random vector with a joint distribution $F$ and continuous marginals $F_1, \ldots, F_p$. Then, we have the unique copula $C$ satisfying equation (4.1), which is called

the copula of $x$ or the copula of $F$. For the sake of simplicity, we consider the case with $p = 2$. For any $\alpha, \beta \in (0, 1)$, we have

$$P\left(x_1 \leq F_1^{-1}(\alpha), x_2 \leq F_2^{-1}(\beta)\right) = C(\alpha, \beta),$$

where the joint probability that random variables are less than or equal to quantiles depends on the copula only. It means that this kind of dependence is determined by the copula. Kendall's $\tau$ or Spearman's $\rho$ is a representative rank correlation and they merely rely on the copula. That is,

$$\tau = 4\int_0^1 \int_0^1 C(u_1, u_2)dC(u_1, u_2) - 1, \;\; \rho = 12\int_0^1 \int_0^1 (C(u_1, u_2) - u_1u_2)du_1du_2.$$

Note that Pearson's linear correlation is not determined by the copula only.

The copula of $x$ is invariant with respect to monotone increasing transformations of the components of $x$. That is, for any monotone increasing function $T_1, \ldots, T_p$, the copula of $(T_1(x_1), \ldots, T_p(x_p))^\top$ is the same as the one of $x$. Therefore, if we consider copulas of normal distributions, we only have to consider normal distributions with mean $0$ and correlation matrix $P$.

### 4.1.3 Examples of Copulas

We present some well-known families of copulas.

**Gaussian copula**

Gaussian copulas are the copulas of normal distributions. Let $c_{\mathrm{G}}(u, P)$ be the density function of the Gaussian copula, that is, the copula of the normal distribution with mean 0 and correlation matrix $P$,

$$c_{\mathrm{G}}(u, P) = \det P^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x_{\mathrm{G}}(u)^{\top}(P^{-1} - I_p) x_{\mathrm{G}}(u)\right),\ u \in [0,1]^p,$$

where $x_{\mathrm{G}}(u) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p))^{\top}$, $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, and $I_p$ is the identity matrix of size $p$. Let $v(P)$ be the $p(p-1)/2$-dimensional vector which consists of the column-wise stacked lower diagonal elements of $P$. For example, $v(P) = (p_{21}, p_{31}, p_{32})^{\top}$ if $p = 3$. The set $\{v(P) : P \text{ is a correlation matrix of size } m\}$ is a parameter space of the Gaussian copula models.

**$t$ copula**

Let $f_t(x, \nu, P)$ denote the probability density function of $t$-distribution with degrees of freedom $\nu$ and correlation matrix $P$,

$$f_t(x, \nu, P) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \det(\nu\pi P)^{-\frac{1}{2}} \left(1 + \frac{x^{\top} P^{-1} x}{\nu}\right)^{-\frac{\nu+p}{2}}.$$

Let $t_\nu$ be the cumulative distribution function of the $t$-distribution with degrees of freedom $\nu$, $f_t(x, \nu)$ be its density function, and $x_{t,\nu}(u) = (t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_p))^{\top}$. Then the

probability density function of $t$-copula is given by

$$c_t(u, \nu, P) \quad = \quad f_t(x_{t,\nu}(u), \nu, P) / \prod_{i=1}^{p} f_t(t_\nu^{-1}(u_i), \nu).$$

**Archimedean copula**

A continuous, strictly decreasing convex function $\phi : [0, 1] \to [0, \infty]$ satisfying $\phi(1) = 0$ is known as an Archimedean copula generator. It is known as a strict generator if $\phi(0) = \infty$. If $\phi$ is a strict Archimedean copula generator, then

$$C(u_1, \ldots, u_p) = \phi^{-1}(\phi(u_1) + \cdots + \phi(u_p))$$

gives a copula in any dimension $p$ if and only if $\phi^{-1}$ is completely monotonic:

$$(-1)\frac{d^k}{dt^k}\phi^{-1}(t) \geq 0$$

for any $k \in \mathbb{N}$ and $t$. This copula is called Archimedean copula. For example, if $\phi(t) = (-\log t)^\theta$, we have the Gumbel copula, and if $\phi(t) = (t^{-\theta} - 1)/\theta$, we have the Clayton copula. See McNeil et al. (2005) for more details.

## 4.2 Estimation

In this section, we discuss estimation problems for copula models. Suppose $F$ is a multivariate distribution function with continuous density $f$, and $C$ with density $c$ is the copula

of $F$. Let $F_i$ $(i = 1, \ldots, p)$ be the univariate marginal of $F$ with density $f_i$. By the Sklar's theorem, we have

$$f(x_1, \ldots, x_p) = c(F_1(x_1), \ldots, F_p(x_p)) f_1(x_1) \cdots f_p(x_p). \tag{4.2}$$

We make use of this representation (4.2) to construct statistical models.

### 4.2.1 Parametric Models

We assume parametric models for a copula and marginals. In equation (4.2), let $c(u) = c(u, \theta_c)$, $f_i(x_i) = f_i(x_i, \theta_i)$, and $f(x) = f(x, \theta)$, where $\theta = (\theta_c^\top, \theta_1^\top, \ldots, \theta_p^\top)^\top$. Suppose $x_1, \ldots, x_n$ are independently and identically distributed with $f(x, \theta)$, where $x_i = (x_{i1}, \ldots, x_{ip})^\top$. The log likelihood function becomes

$$\begin{aligned} L_0(\theta) &= \sum_{i=1}^{n} \log f(x_i, \theta) \\ &= \sum_{i=1}^{n} \log c(F_1(x_{i1}, \theta_1), \ldots, F_p(x_{ip}, \theta_p), \theta_c) + \sum_{j=1}^{p} \sum_{i=1}^{n} \log f_j(x_{ij}, \theta_j) \\ &= L_0^c(\theta) + \sum_{j=1}^{p} L_0^j(\theta_j), \end{aligned}$$

where

$$L_0^c(\theta) = \sum_{i=1}^{n} \log c(F_1(x_{i1}, \theta_1), \ldots, F_p(x_{ip}, \theta_p), \theta_c), \; L_0^j(\theta_j) = \sum_{i=1}^{n} \log f_j(x_{ij}, \theta_j).$$

The maximum likelihood estimators of the parameters are obtained by solving the following equations,

$$\frac{\partial L_0(\theta)}{\partial \theta_1} = 0, \ldots, \frac{\partial L_0(\theta)}{\partial \theta_p} = 0, \frac{\partial L_0(\theta)}{\partial \theta_c} = 0.$$

These estimating equations are often too complicated to solve. We can consider other estimators based on the decomposition (4.2), which are computationally attractive alternatives to the maximum likelihood estimators. They are obtained by maximizing $L_0^j(\theta_j), (j = 1, \ldots, p)$, substituting the maximizers $\hat{\theta}_1, \ldots, \hat{\theta}_p$ into the counterparts in $L_0^c(\theta)$, and maximizing $L_0^c(\theta_c, \hat{\theta}_1, \ldots, \hat{\theta}_p)$ with respect to $\theta_c$. That is, we solve the following equations,

$$\frac{\partial L_0^1(\theta_1)}{\partial \theta_1} = 0, \ldots, \frac{\partial L_0^p(\theta_p)}{\partial \theta_p} = 0, \frac{\partial L_0^c(\theta_c, \hat{\theta}_1, \ldots, \hat{\theta}_p)}{\partial \theta_c} = 0.$$

These estimators are called inference functions for margins estimators (IFM-estimators) (Joe, 2001). The IFM estimator is consistent and asymptotic normal.

### 4.2.2 Semiparametric Models

We assume a parametric model for a copula but do not make any assumptions for marginals. Let $c(u) = c(u, \theta_c)$ in (4.2). The log likelihood function becomes

$$L_0(\theta_c) \;=\; \sum_{i=1}^{n} \log c(F_1(x_{i1}), \ldots, F_p(x_{ip}), \theta_c),$$

up to constant. An estimator of $\theta_c$ is defined by maximizing $L_0(\theta_c)$ in which $F_1, \ldots, F_p$ are replaced by some nonparametric estimates $\hat{F}_1, \ldots, \hat{F}_p$, respectively. This estimator is

consistent and asymptotic normal (Genest et al., 1995).

## 4.3 $\gamma$-Estimation of the Gaussian Copula Parameter

In this section, we consider the $\gamma$-estimation for the Gaussian copula models. One of the important purposes of using copulas is to represent a variety of dependence structure, such as heavy tail and asymmetric dependence (Demarta and McNeil, 2005). The Gaussian copula is a fundamental copula, but it is neither heavy tail nor asymmetric. Hence, other copulas, such as $t$-copulas and Archimedean copulas, are employed, in which the maximum likelihood estimation is often used. In Yoshiba (2013), the Gaussian copula mixture is used to model heterogeneous correlation structure. These examples can be considered to change the Gaussian copula to other copulas. On the other hand, our idea is to change the maximum likelihood estimation to the $\gamma$-estimation, but keep the Gaussian copula as a statistical model. Due to the change, heterogeneous correlation structure can be captured as shown in this section. We describe this dual relation of changing models or estimation methods precisely in section 4.3.6

For the sake of simplicity, we assume the data $u_1, \ldots, u_n$ are independently and identically distributed with a copula density $c(u)$. If we have the data $x_1, \ldots, x_n$ drawn from a distribution $F$ with marginals $F_1, \ldots, F_p$ and copula density $c$, then we have

$$u_i = (F_1(x_{i1}), \ldots, F_p(x_{ip})).$$

Hence, we can get $u$'s from $x$'s approximately by computing $(\hat{F}_1(x_{i1}), \ldots, \hat{F}_p(x_{ip}))$ with

estimators $\hat{F}_1, \ldots, \hat{F}_p$. This section is based on the paper (Notsu et al., 2013).

### 4.3.1 Maximum likelihood Estimation of the Gaussian Copula Parameter

We consider the MLE for the Gaussian copula model. The log likelihood function multiplied by $-1/n$ is given by

$$L_0(P) \;=\; -\frac{1}{2}\log\det(P)^{-1} + \frac{1}{2n}\sum_{i=1}^{n} {x_i}^{\top}(P^{-1} - I_p)x_i,$$

where $x_i = x_{\mathrm{G}}(u_i)$ for $i = 1, \ldots, n$. It is well known that the MLE does not work well under model misspecification. For example, in the case of (3.4) the MLE for the Gaussian copula model almost surely converges to $\tau P_1 + (1-\tau)P_2$, so we cannot detect neither $P_1$ nor $P_2$. If $\tau = 0.5$ and

$$P_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix},$$

then $\tau P_1 + (1-\tau)P_2$ is equal to the identity matrix, which has no meaning in this situation. We cannot use the MLE in the case of misspecification.

### 4.3.2 $\gamma$-Estimator of the Gaussian Copula Parameter

Let $u_1, \ldots, u_n$ be a random sample from a copula with the probability density function $c(u)$ while $c_{\mathrm{G}}(u, P)$ is our statistical model. The $\gamma$-loss function for the Gaussian copula is given

by

$$L_\gamma(P) \quad = \quad -\det(P)^{-\frac{\gamma}{2(1+\gamma)}} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-\frac{\gamma}{2} {x_i}^\top P^{-1} x_i\right), \tag{4.3}$$

up to constant. The $\gamma$-estimator is proposed as the set of local minimum points of $L_\gamma(P)$ and interpreted as follows. If $L_\gamma(P)$ has a local minimum, the underlying distribution is estimated by $c_{\mathrm{G}}(u, \hat{P}_\gamma)$ using the minimum point $\hat{P}_\gamma$. If $L_\gamma(P)$ has $\ell$ local minima ($\ell \geq 2$), the underlying distribution is estimated by a mixture of $\ell$ Gaussian copulas. Each Gaussian copula's parameter is estimated by the corresponding local minimum point.

### 4.3.3 An Algorithm to Obtain the $\gamma$-Estimator

We give a fixed point algorithm to obtain the $\gamma$-estimator for the Gaussian copula model using the Lagrange-multiplier method. We can still make use of this algorithm to obtain the MLE just by setting $\gamma = 0$.

**Algorithm**

1. Set an appropriate correlation matrix $P_0$.

2. Given $P_t$, calculate $P_{t+1}$ by the following update formula,

$$P_{t+1} = \Sigma_t + P_t \mathrm{diag}\left(\left(P_t \odot P_t\right)^{-1} \mathrm{Diag}\left(I_p - \Sigma_t\right)\right) P_t, \tag{4.4}$$

where $\odot$ denotes the Hadamard product. $\Sigma_t$ is defined by

$$\Sigma_t \quad = \quad (1+\gamma)\sum_{i=1}^{n} w_\gamma(x_i, P_t) x_i {x_i}^\top,$$

where

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

Here Diag $(M)$ for a square matrix $M$ denotes the column vector which consists of the diagonal elements of $M$ and diag$(a)$ for a vector $a$ denotes the diagonal matrix whose diagonal elements are the components of $a$.

3. For sufficient small given number $\varepsilon$, repeat Procedure 2 while

$$||P_{t+1} - P_t||_{\mathrm{F}} > \varepsilon.$$

4. For all local minimum points, repeat Procedure 1-3 for different initial values $P_0$.

We derive the estimation equation for $P$, which leads to the update formula (4.4). Since $P$ is symmetric and positive definite, there exists a matrix $R$ of size $p$ which satisfies $P = RR^\top$. The $i$th diagonal element of $P$ is expressed by $e_i^\top RR^\top e_i$, where $e_i$ is the $p$-dimensional column vector whose $i$th element is 1 and the other elements are 0. Since the diagonal

elements of $P$ are equal to 1, Lagrange function becomes

$$\begin{aligned}\Lambda(R,\lambda) &= \left(\det R^{-1}\right)^{\frac{\gamma}{(1+\gamma)}} \sum_{i=1}^{n} \exp\left(-\frac{\gamma}{2} {x_i}^\top {R^{-1}}^\top R^{-1} x_i\right) \\ &\quad + \sum_{i=1}^{p} \lambda_i (e_i^\top R R^\top e_i - 1), \end{aligned} \tag{4.5}$$

where $\lambda = (\lambda_1, \ldots, \lambda_p)^\top$ is Lagrange multiplier. We differentiate (4.5) with respect to $R^{-1}$ with the technique in Magnus and Neudecker (1999). The differential of $\sum_{i=1}^{p} \lambda_i e_i^\top R R^\top e_i$, which is defined in Magnus and Neudecker (1999, Section 5.3 and 5.16), is

$$\begin{aligned} d\left(\sum_{i=1}^{p} \lambda_i e_i^\top R R^\top e_i\right) &= d\left(\text{tr}\left(R^\top \text{diag}(\lambda) R\right)\right) \\ &= \text{tr}\left(2R^\top \text{diag}(\lambda)(dR)\right) \\ &= \text{tr}\left(-2RR^\top \text{diag}(\lambda) R (dR^{-1})\right), \end{aligned}$$

where diag($\lambda$) is the diagonal matrix whose diagonal elements are $\lambda_1, \ldots, \lambda_p$. From Table 2 in Magnus and Neudecker (1999, Chapter 9) we have

$$\frac{\partial}{\partial R^{-1}} \sum_{i=1}^{p} \lambda_i e_i^\top R R^\top e_i = -2R^\top \text{diag}(\lambda) R R^\top.$$

Set the derivative of (4.5) to $O$, then we have

$$\begin{aligned}\frac{\partial \Lambda(R,\lambda)}{\partial R^{-1}} &= \frac{\gamma}{(1+\gamma)}\left(\det R^{-1}\right)^{\frac{\gamma}{(1+\gamma)}} R^{\top}\sum_{i=1}^{n}\exp\left(-\frac{\gamma}{2}{x_i}^{\top}{R^{-1}}^{\top}R^{-1}x_i\right)\\ &\quad -\gamma\left(\det R^{-1}\right)^{\frac{\gamma}{(1+\gamma)}} R^{-1}\sum_{i=1}^{n}\exp\left(-\frac{\gamma}{2}{x_i}^{\top}{R^{-1}}^{\top}R^{-1}x_i\right)\\ &\quad x_i{x_i}^{\top} - 2R^{\top}\text{diag}(\lambda)RR^{\top}\\ &= O. \end{aligned} \tag{4.6}$$

Multiply $R$ from the left side of equation (4.6), then (4.6) becomes

$$P = A + aP\text{diag}(\lambda)P,$$

where

$$\begin{aligned} A &= (1+\gamma)\sum_{i=1}^{n} w_{\gamma}(x_i,P)x_i{x_i}^{\top},\\ a &= 2\left(\frac{\gamma}{1+\gamma}\left(\det R^{-1}\right)^{\frac{\gamma}{1+\gamma}}\sum_{i=1}^{n}\exp\left(-\frac{\gamma}{2}{x_i}^{\top}{R^{-1}}^{\top}R^{-1}x_i\right)\right)^{-1}. \end{aligned}$$

From the constraint about the diagonal elements of $P$ we have

$$\text{Diag}(I_p - A) = \text{Diag}(aP\text{diag}(\lambda)P). \tag{4.7}$$

In general, for any square matrices $X$ and $Y$ of size $p$ and $p$-dimensional column vector $x$, we have

$$\text{Diag}(X\text{diag}(x)Y) = (X \odot Y^{\top})x.$$

So (4.7) becomes

$$\lambda = \frac{1}{a}(P \odot P)^{-1}\text{Diag}(I_m - A).$$

Then we have

$$P = A + P\text{diag}\left((P \odot P)^{-1}\ \text{Diag}\ (I_m - A)\right) P,$$

and use this estimation equation as an update formula.

If we consider the estimation problem on Gaussian distributions with mean $0$, the update formula for an iteration algorithm to obtain the $\gamma$-estimator of the covariance matrix $\Sigma$ is given by

$$\Sigma_{t+1} = (1+\gamma)\sum_{i=1}^{n} w_\gamma(x_i, \Sigma_t) x_i {x_i}^\top. \tag{4.8}$$

See Fujisawa and Eguchi (2008) for details. If we consider the optimization problem with the objective function $L_\gamma(P)$ without the constraint that the diagonal elements of $P$ are 1, the same iteration algorithm (4.8) can be deduced. So the second term of the right hand side of the equation (4.4) appears because of the existence of the constraint.

We make a remark on the algorithm to obtain the MLE, or $\gamma$-estimator with $\gamma = 0$. On the main update formula (4.4) in Step 2 $\Sigma_t$ is always the sample covariance matrix $S$ for any $t \geq 1$. Nevertheless we find rather complicated solution of the MLE if we consider a simpler case of $p = 2$. McNeil et al. (2005) show an approximate MLE for the Gaussian copula model because it takes quite a while to solve the constrained optimization problem

in order to obtain the MLE in high dimensions. The approximate MLE is given by

$$\text{diag}(S)^{-\frac{1}{2}} S \text{diag}(S)^{-\frac{1}{2}}, \tag{4.9}$$

where diag$(S)$ is the diagonal matrix whose diagonal elements are equal to those of $S$. We can easily consider an iteration algorithm to obtain an approximate $\gamma$-estimator to combine (4.8) and (4.9). The update formula of the algorithm is given by

$$P^*_{t+1} \quad = \quad \text{diag}(\Sigma^*_{t+1})^{-\frac{1}{2}} \Sigma^*_{t+1} \text{diag}(\Sigma^*_{t+1})^{-\frac{1}{2}},$$

where

$$\Sigma^*_{t+1} \quad = \quad (1+\gamma) \sum_{i=1}^{n} w_\gamma(x_i, P^*_t) x_i {x_i}^\top.$$

If $n$ is infinity, $P_t$ and $P^*_t$ converge to the same correlation matrix when $t$ tends to $\infty$. However $P_t$ and $P^*_t$ are different in general. $P_t$ is preferred to $P^*_t$ in terms of accuracy.

### 4.3.4 Choice of the Carrier Measure

Although the $\gamma$-cross entropy has been defined on the Lebesgue measure in section 2.2, it can be defined on any carrier measure. Here we propose, for Gaussian copula models, the use of a measure, denoted by $Q_{\text{G}}$, of which Radon-Nikodym derivative is given by $J(x_{\text{G}})^{-\gamma}$, where $J(x_{\text{G}})$ is the Jacobian of the transformation $x_{\text{G}}(u)$. From now on we refer this choice to $Q_{\text{G}}$, and explain its rationale by virtue of invariance. The $\gamma$-loss function (4.3) is obtained

by using $Q_{\mathrm{G}}$ as a carrier measure.

Let the $\gamma$-cross entropy with measure $Q$ be denoted by

$$C_\gamma(g, f|Q) = \frac{\int g(x) f(x)^\gamma Q(dx)}{\left\{\int f(x) Q(dx)\right\}^{\frac{\gamma}{1+\gamma}}}.$$

We assume that $x = (x_1, \ldots, x_p)^\top \sim \phi(x, P)$, where $\phi(x, P)$ denotes the probability density function of the $p$-dimensional Gaussian distribution with mean $0$ and correlation matrix $P$. Let $u = (\Phi(x_1), \ldots, \Phi(x_p))^\top$, then $u \sim c_{\mathrm{G}}(u, P)$. If the underlying distribution of $x$ is $g(x)$, then $u \sim c(u)$, where $c(u)$ is given by $g(x_{\mathrm{G}}(u))J(x_{\mathrm{G}})$. It is noteworthy that the $\gamma$-cross entropy between $g(x)$ and $\phi(x, P)$ based on $x$ is not always equal to the $\gamma$-cross entropy between $c(u)$ and $c_{\mathrm{G}}(u, P)$ based on $u$. So the $\gamma$-estimator based on $x$ does not coincide with the $\gamma$-estimator based on $u$.

It is natural for us to require the equivalence of the two $\gamma$-estimators, and therefore we employ the measure $Q_{\mathrm{G}}(u)$. It is striking that the $\gamma$-cross entropy between $c(u)$ and $c_{\mathrm{G}}(u, P)$ calculated under the measure $Q_{\mathrm{G}}$ is equal to the one between $g(x)$ and $\phi(x, P)$ calculated under the Lebesgue measure $Q_{\mathrm{L}}$, that is,

$$C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}}) \;=\; C_\gamma(g, \phi(\cdot, P)|Q_{\mathrm{L}}),$$

which is proportional to

$$-\det(P)^{-\frac{\gamma}{2(1+\gamma)}} \int g(x) \exp\left(-\frac{\gamma}{2} x^\top P^{-1} x\right) dx. \tag{4.10}$$

Obviously there is equalization of the two $\gamma$-estimators. Note that the $\gamma$-loss function associated with (4.10) becomes equation (4.3).

The argument above extends to a general statement. For given one to one transformation $y(x) : x \mapsto y$, $x(y)$ denotes the inverse function of $y(x)$, and $J(y \mapsto x)$ denotes the Jacobian of the transformation $x(y)$. Any nonnegative functions $g(x), f(x)$ satisfy

$$C_\gamma \left( g(x(\cdot))J(y \mapsto x), f(x(\cdot))J(y \mapsto x) | Q \right) = C_\gamma(g, f | Q_{\mathrm{L}}),$$

if and only if the Radon-Nikodym derivative of $Q$ is equal to $J(y \mapsto x)^{-\gamma}$. When $g(x)$ and $f(x)$ are the probability density functions, to consider the $\gamma$-cross entropy on $x$ under the Lebesgue measure is equal to consider the one based on $y$ under the measure having $J(y \mapsto x)^{-\gamma}$ as its Radon-Nikodym derivative.

### 4.3.5 Properties of the $\gamma$-Estimator

The $\gamma$-estimator for the Gaussian copula model under infinite sample size is equal to the set of the local minimum points of $C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$. In this section we leave aside the $\gamma$-loss function $L_\gamma(P)$ for the moment and investigate the property of the $\gamma$-estimator (at $n$ infinity) through $C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$. First we consider the case where there is no misspecification.

**Theorem 4.3.1** *If $c(u) = c_{\mathrm{G}}(u, P_0)$, then $C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_G)$ has the local minimum point $P_0$.*

**Proof.** We see that

$$C_\gamma(c_{\mathrm{G}}(\cdot, P_0), c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}}) \propto -\det P^{\frac{1}{2(1+\gamma)}} \det(P+\gamma P_0)^{-\frac{1}{2}}. \tag{4.11}$$

Consider a monotone transformation of the right hand side of equation (4.11) to obtain

$$\begin{aligned} & -\log\left[\det P^{\frac{1}{2(1+\gamma)}} \det(P+\gamma P_0)^{-\frac{1}{2}}\right] \\ = \; & -\frac{1}{2(1+\gamma)} \log\det P + \frac{1}{2}\log\det(P+\gamma P_0). \end{aligned}$$

For any $P \neq P_0$, let $P_t = (1-t)P_0 + tP, (t>0)$ and define $f(t)$ by

$$f(t) = -\frac{1}{2(1+\gamma)} \log\det P_t + \frac{1}{2}\log\det(P_t+\gamma P_0).$$

We see

$$\begin{aligned} f'(t) &= -\frac{1}{2(1+\gamma)}\mathrm{tr}\left[P_t^{-1}(P-P_0)\right] + \frac{1}{2}\left[(P_t+\gamma P_0)^{-1}(P-P_0)\right] \\ &= \frac{1}{2(1+\gamma)}\mathrm{tr}\left[\left\{\left(\frac{1}{1+\gamma}P_t + \frac{\gamma}{1+\gamma}P_0\right)^{-1} - P_t^{-1}\right\}(P-P_0)\right] \\ &= \frac{1}{2\gamma t}\mathrm{tr}\left[\left\{\left(\frac{1}{1+\gamma}P_t + \frac{\gamma}{1+\gamma}P_0\right)^{-1} - P_t^{-1}\right\}\right. \\ &\quad \left.\left\{P_t - \left(\frac{1}{1+\gamma}P_t + \frac{\gamma}{1+\gamma}P_0\right)\right\}\right] \end{aligned}$$

Let $D_0(c_{\mathrm{G}}(\cdot, P_1), c_{\mathrm{G}}(\cdot, P_2))$ be the KL divergence between $c_{\mathrm{G}}(u, P_1)$ and $c_{\mathrm{G}}(u, P_2)$. It is well known $D_0(c_{\mathrm{G}}(\cdot, P_1), c_{\mathrm{G}}(\cdot, P_2)) \geq 0$ and equal to 0 if and only if $P_1 = P_2$. So for

$P_1 \neq P_2$, we have

$$\begin{aligned} & D_0(c_{\mathrm{G}}(\cdot;P_1), c_{\mathrm{G}}(\cdot;P_2)) + D_0(c_{\mathrm{G}}(\cdot;P_2), c_{\mathrm{G}}(\cdot;P_1)) \\ = \; & \frac{1}{2}\mathrm{tr}\left((P_1 - P_2)(P_2 - P_1)\right) \\ > \; & 0. \end{aligned}$$

If we read as $P_1 = P_t$ and $P_2 = \frac{1}{1+\gamma}P_t + \frac{\gamma}{1+\gamma}P_0$, then we see that $f'(t) > 0$. The proof is complete. □

In this case we note that the $\gamma$-estimator is equal to $\{P_0\}$, which implies Fisher consistency. For asymptotic properties the $\gamma$-estimator has asymptotic consistency and normality.

Next we consider the misspecification case where the true data generating process is given by equation (1.1). We see that

$$C_\gamma(c, c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}}) \;=\; \tau C_\gamma(c_{\mathrm{G}}(\cdot,P_1), c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}}) + (1-\tau)C_\gamma(c_{\mathrm{G}}(\cdot,P_2), c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}}),$$

which is proportional to

$$\begin{aligned} & -\det P^{-\frac{\gamma}{2(1+\gamma)}} \left[\tau \det P_1^{-\frac{1}{2}} \det\left(P_1^{-1} + \gamma P^{-1}\right)^{-\frac{1}{2}}\right. \\ & \left. +(1-\tau)\det P_2^{-\frac{1}{2}} \det\left(P_2^{-1} + \gamma P^{-1}\right)^{-\frac{1}{2}}\right]. \end{aligned}$$

Then $C_\gamma(c, c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}})$ is a weighted mean of $C_\gamma(c_{\mathrm{G}}(\cdot,P_1), c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}})$ and $C_\gamma(c_{\mathrm{G}}(\cdot,P_2)$ ,$c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}})$. Each component is a unimodal function, bounded above by 0, and has one local minimum point $P_1$ and $P_2$, respectively. We expect $C_\gamma(c, c_{\mathrm{G}}(\cdot,P)|Q_{\mathrm{G}})$ has two local

minimum points, and these local minimum points are near $P_1$ and $P_2$ respectively if $P_1$ and $P_2$ are sufficiently "distinct". However it is hard to formulate such a phenomenon mathematically so we show through easy examples and a graph that such a phenomenon occurs. To obtain numerical solutions, we use the expected (or population) version of the algorithm in section 4.3.3.

**Example 1:** In the case with dimension 2, $C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$ is a univariate function of $\rho$, which is the non-diagonal element of $P$. Let $P_1$ and $P_2$ be

$$P_1 = \begin{pmatrix} 1 & \rho_* \\ \rho_* & 1 \end{pmatrix}, \; P_2 = \begin{pmatrix} 1 & -\rho_* \\ -\rho_* & 1 \end{pmatrix},$$

$\gamma = 1$, and $\tau = 0.5$. If $\rho_* > \sqrt{6 - \sqrt{28}} \doteq 0.842$, then $C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$ has two local minimum points in the interval $(-1, 0)$ and $(0, 1)$, respectively.

**Example 2:** Suppose the true correlation matrices $P_1$ and $P_2$ are given as follows, and $P$ stands for the parameterization of the statistical model we fit,

$$P_1 = \begin{pmatrix} 1 & 0.9 & 0.9^2 \\ 0.9 & 1 & 0.9 \\ 0.9^2 & 0.9 & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 1 & -0.9 & 0.9^2 \\ -0.9 & 1 & -0.9 \\ 0.9^2 & -0.9 & 1 \end{pmatrix},$$

$$P = \begin{pmatrix} 1 & \rho_1 & \rho_1\rho_2 \\ \rho_1 & 1 & \rho_2 \\ \rho_1\rho_2 & \rho_2 & 1 \end{pmatrix}.$$

We also set $\tau = 0.5$ and $\gamma = 1$. Note that $-C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$ is a function of $\rho_1$ and $\rho_2$. Figure 4.1 shows $-C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$. We can see there exist two local maxima at $(0.86, 0.86)$ and $(-0.86, -0.86)$.

**Example 3:** Suppose $\tau = 0.4$ and $\gamma = 1$. If $P_1, P_2$, and $P$ are given by

$$P_1 = \begin{pmatrix} 1 & 0.9 & 0.7 & 0.7 \\ 0.9 & 1 & 0.9 & 0.7 \\ 0.7 & 0.9 & 1 & 0.7 \\ 0.7 & 0.7 & 0.7 & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 1 & -0.9 & 0.7 & 0.7 \\ -0.9 & 1 & -0.9 & -0.7 \\ 0.7 & -0.9 & 1 & 0.7 \\ 0.7 & -0.7 & 0.7 & 1 \end{pmatrix},$$

$$P = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_2 & \rho_4 & 1 & \rho_6 \\ \rho_3 & \rho_5 & \rho_6 & 1 \end{pmatrix},$$

then $C_\gamma(c, c_{\mathrm{G}}(\cdot; P)|Q_{\mathrm{G}})$ has two local minima at

$$P = \begin{pmatrix} 1 & 0.871 & 0.686 & 0.699 \\ 0.871 & 1 & 0.871 & 0.683 \\ 0.686 & 0.871 & 1 & 0.699 \\ 0.699 & 0.683 & 0.699 & 1 \end{pmatrix},$$

and

$$P = \begin{pmatrix} 1 & -0.892 & 0.695 & 0.700 \\ -0.892 & 1 & -0.892 & -0.696 \\ 0.695 & -0.892 & 1 & 0.700 \\ 0.700 & -0.696 & 0.700 & 1 \end{pmatrix}$$

Like these examples, $C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$ has some local minimum points depending on the underlying distribution. Owing to this property we can detect the heterogeneous structure of the underlying distribution under misspecification.

### 4.3.6 Maximum Entropy Copula

So far we have considered the $\gamma$-estimation of the Gaussian copula model. In this section we uncover that the choice of copula model can be characterized in terms of the maximum entropy distribution. In this regard, Zhang et al. (2011) is the most closely related work in which the MLE on meta $t$-distribution is addressed. A $t$-copula is deduced from a multivariate $t$-distribution while the meta $t$-distribution is constructed by linking a $t$-copula to univariate $t$-distributions as its marginal distributions. In our framework, Zhang et al. (2011)'s work can be interpreted as the maximum likelihood estimation of $t$-copulas with the marginals estimated simultaneously. Actually the $\gamma$-estimation of Gaussian copulas and the maximum likelihood estimation of $t$-copulas look very similar and share a common idea.

Eguchi et al. (2011) analyze what the maximum $\gamma$-entropy distributions would be under the given (population) mean vector and covariance matrix. The answer depends on the power index $\gamma$. When $\gamma = 0$, the Gaussian distribution emerges as the maximum $\gamma$-entropy distribution. If $\gamma < 0$, the $t$-distribution comes up. We show that a similar result holds for copulas. Suppose that $\gamma = -2/(\nu + p)$ and $Q_{t,\nu}(du) = J(x_{t,\nu})^{-\gamma} du$, where $\nu$ is the degrees of freedom of $t$-copula. Let $\mathscr{C}_\gamma(P)$ be the set of probability density functions $c(u)$ on $[0,1]^p$ which satisfy the following equation.

$$\int_{[0,1]^p} c(u) x_{t,\nu}(u) x_{t,\nu}(u)^\top du = \frac{\nu}{\nu - 2} P.$$

Then we see that

$$\underset{c \in \mathscr{C}_\gamma(P)}{\operatorname{argmax}} H_\gamma(c|Q_{t,\nu}) = c_t(u, \nu, P).$$

**Proof.** We show that $c_t(\cdot, \nu, P) \in \mathscr{C}_\gamma(P)$. Note that $c_t(u, \nu, P) = f_t(x_{t,\nu}(u), \nu, P) J(x_{t,\nu})$. Then,

$$\begin{aligned}
\int c_t(u, \nu, P) x_{t,\nu}(u) x_{t,\nu}(u)^\top du &= \int f_t(x, \nu, P) x x^\top J(x_{t,\nu}) J(x_{t,\nu}^{-1}) dx \\
&= \int f_t(x, \nu, P) dx \\
&= \frac{v}{v - 2} P.
\end{aligned}$$

Hence $c_t(\cdot, \nu, P) \in \mathscr{C}_\gamma(P)$. We see that

$$\begin{aligned} H_\gamma(c_t(\cdot,\nu,P)|Q_{t,\nu}) &= -\left[\int_{[0,1]^p} c_t(u,\nu,P)^{1+\gamma} Q_{t,\nu}(du)\right]^{\frac{1}{1+\gamma}} \\ &= -\frac{\int_{[0,1]^p} c_t(u,\nu,P)^{1+\gamma} Q_{t,\nu}(du)}{\left[\int_{[0,1]^p} c_t(u,\nu,P)^{1+\gamma} Q_{t,\nu}(du)\right]^{\frac{\gamma}{1+\gamma}}}. \end{aligned} \tag{4.12}$$

The numerator of (4.12) becomes

$$\begin{aligned} \int_{[0,1]^p} c_t(u,\nu,P)^{1+\gamma} Q_{t,\nu}(du) &= \int_{[0,1]^p} c_t(u,\nu,P) c_t(u,\nu,P)^{\gamma} Q_{t,\nu}(du) \\ &= \int_{[0,1]^p} c_t(u,\nu,P) f_t(x_{t,\nu}(u),\nu,P)^{\gamma} du \\ &= \int_{[0,1]^p} c(u) f_t(x_{t,\nu}(u),\nu,P)^{\gamma} du, \end{aligned}$$

for $c \in \mathscr{C}_\gamma(P)$. Hence

$$\begin{aligned} H_\gamma(c_t(\cdot,\nu,P)|Q_{t,\nu}) &= -\frac{\int_{[0,1]^p} c(u) c_t(u,\nu,P)^{\gamma} Q_{t,\nu}(du)}{\left[\int_{[0,1]^p} c_t(u,\nu,P)^{1+\gamma} Q_{t,\nu}(du)\right]^{\frac{\gamma}{1+\gamma}}} \\ &= C_\gamma(c, c_t(\cdot,\nu,P)|Q_{t,\nu}) \\ &\geq H_\gamma(c|Q_{t,\nu}). \end{aligned}$$

□

Note that there exists an element in $\mathscr{C}_\gamma(P)$ except $c_t(\cdot, \nu, P)$. For a given correlation matrix $P$, there exists $\varepsilon' < 0$ such that $(1-\varepsilon')P + \varepsilon' I$ is a positive definite correlation matrix, since $P$ is positive definite. Let $P_1 = (1-\varepsilon')P + \varepsilon' I$ and $\varepsilon = -\varepsilon'/(1-\varepsilon')$. Then we have $(1-\varepsilon)P_1 + \varepsilon I = P$ and $0 < \varepsilon < 1$. Let $c(u) = (1-\varepsilon)c_t(u,\nu,P_1) + \varepsilon c_t(u,\nu,I)$. Then

$c(u)$ satisfies that $c(u) \in \mathscr{C}_\gamma(P)$ and $c(u) \neq c_t(u, \nu, P)$.

If $\gamma \to 0$, then $t_\nu \to \Phi$ and $Q_{t,\nu} \to Q_{\mathrm{L}}$. Hence

$$\underset{c \in \mathscr{C}_0(P)}{\operatorname{argmax}} H_0(c|Q_{\mathrm{L}}) = c_{\mathrm{G}}(u, P).$$

That is, $t$-copula can be characterized as the maximum $\gamma$-entropy distribution on $[0, 1]^p$. Moreover it has limiting equivalence (by letting $\gamma \to 0$) with the Gaussian copula which is tagged with the maximum Boltzmann-Shannon entropy copula. We call these maximum $\gamma$-entropy copulas the $\gamma$-copulas. Let us consider the relationship between the $\gamma$-copula and the $\gamma$-estimation. Our method is discussed on the pair of the Gaussian copula (0-copula) and $\gamma$-estimator. On the other hand Zhang et al. (2011) discussed on the pair of $\gamma$-copula model ($\gamma < 0$) and the MLE (0-estimator). We see a sort of duality relationship between two choices of the pair.

### 4.3.7 Robustness of the $\gamma$-Estimator

We examine robustness of the $\gamma$-estimator for the Gaussian copula model through its influence function. The influence function measures the asymptotic bias caused by contamination at the $x$. The boundedness of the influence function means boundedness of the influence from the outlier, hence its robustness. The influence function of the $\gamma$-estimator is given. We show that it is bounded when $\gamma > 0$. A brief simulation is also performed.

The $\gamma$-estimator for the Gaussian copula model can be regarded as a functional $T(g)$ of

a distribution $g$ defined by

$$\underset{P}{\text{argsolve}} \int \exp\left(-\frac{\gamma}{2} x^\top P^{-1} x\right) v(P^{-1} - (1+\gamma) P^{-1} x x^\top P^{-1}) g(x) dx = 0. \quad (4.13)$$

Let $\psi_\gamma(x, P)$ be

$$\psi_\gamma(x, P) = \exp\left(-\frac{\gamma}{2} x^\top P^{-1} x\right) v(P^{-1} - (1+\gamma) P^{-1} x x^\top P^{-1}).$$

Then the influence function $\text{IF}(x, T, g)$ of the $\gamma$-estimator is given by

$$\text{IF}(x, T, g) = -\left[\int \dot{\psi}_\gamma(x, T(g)) g(x) dx\right]^{-1} \psi_\gamma(x; T(g)),$$

where $\dot{\psi}_\gamma(x, P) = \frac{\partial}{\partial v(P)} \psi_\gamma(x, P)$. See Huber (1981) for details. The boundedness of the influence function is equivalent to the boundedness of $\psi_\gamma(x, P)$. The following theorem gives a bound of $\psi_\gamma(x, P)$.

**Theorem 4.3.2** *When $\gamma = 0$, that is, for the MLE, the influence function is not bounded. When $\gamma < 0$, the influence function is not bounded. When $\gamma > 0$, the influence function is bounded and a bound is given by*

$$\|\psi_\gamma(x, P)\| \leq \|v(P^{-1})\| + \frac{2(1+\gamma)}{e\gamma} \|P^{-\frac{1}{2}} \otimes P^{-\frac{1}{2}}\|,$$

*where $\otimes$ denotes the Kronecker product and $\|h\|$ for an $m$-dimensional vector $h$ denotes the Euclidean norm defined by $\sqrt{h^\top h}$*

**Proof.** If $\gamma = 0$ we see

$$\psi_0(x, P) = v(P^{-1} - P^{-1}xx^\top P^{-1}).$$

It is obvious that $\|\psi_0(x, P)\|$ is not bounded with respect to $x$. Next if $\gamma \neq 0$, then let $P^{-1} = (P^{-\frac{1}{2}})^2$, where $P^{-\frac{1}{2}}$ is a symmetric matrix. Set $y = P^{-\frac{1}{2}}x$, then

$$\psi_\gamma(x, P) = \exp\left(-\frac{\gamma}{2}y^\top y\right) v(P^{-1} - (1+\gamma)P^{-\frac{1}{2}}yy^\top P^{-\frac{1}{2}}).$$

Express $y$ in polar coordinate, then

$$y = rp(\theta) = r(\cos\theta_1, \sin\theta_1\cos\theta_2, \ldots, \sin\theta_1\cdots\sin\theta_{m-1})^\top,$$

where $0 \leq r,\ 0 \leq \theta_1, \ldots, \theta_{m-2} \leq \pi,\ 0 \leq \theta_{m-1} \leq 2\pi$. Hence

$$\psi_\gamma(x, P) = \exp\left(-\frac{\gamma}{2}r^2\right) v(P^{-1} - r^2(1+\gamma)P^{-\frac{1}{2}}p(\theta)p(\theta)^\top P^{-\frac{1}{2}}).$$

If $\gamma < 0$ and $r \to \infty$, then we see $\|\psi_\gamma(x, P)\|$ is not bounded. Next if $\gamma > 0$, we see

$$\begin{aligned}\|\psi_\gamma(x, P)\| \quad &\leq \quad \exp\left(-\frac{\gamma}{2}r^2\right)\|v(P^{-1})\| \\ &\qquad + \exp\left(-\frac{\gamma}{2}r^2\right) r^2(1+\gamma)\|v(P^{-\frac{1}{2}}p(\theta)p(\theta)^\top P^{-\frac{1}{2}})\|.\end{aligned}$$

Since $\exp\left(-\frac{\gamma}{2}r^2\right) \leq 1,\ \exp\left(-\frac{\gamma}{2}r^2\right) r^2 \leq 2/(e\gamma)$,

$$\|\psi_\gamma(x,P)\| \leq \|v(P^{-1})\| + \frac{2(1+\gamma)}{e\gamma}\|\mathrm{vec}(P^{-\frac{1}{2}}p(\theta)p(\theta)^\top P^{-\frac{1}{2}})\|,$$

where vec denotes the vec operator. In addition we observe

$$\begin{aligned}\|\mathrm{vec}(P^{-\frac{1}{2}}p(\theta)p(\theta)^\top P^{-\frac{1}{2}})\| &= \|(P^{-\frac{1}{2}}\otimes P^{-\frac{1}{2}})\mathrm{vec}(p(\theta)p(\theta)^\top)\| \\ &\leq \|(P^{-\frac{1}{2}}\otimes P^{-\frac{1}{2}})\|\|\mathrm{vec}(p(\theta)p(\theta)^\top)\|.\end{aligned}$$

Since

$$\|\mathrm{vec}(p(\theta)p(\theta)^\top)\| = \mathrm{tr}(p(\theta)p(\theta)^\top p(\theta)p(\theta)^\top)^{\frac{1}{2}} = 1,$$

we see

$$\|\psi_\gamma(x,P)\| \leq \|v(P^{-1})\| + \frac{2(1+\gamma)}{e\gamma}\|P^{-\frac{1}{2}}\otimes P^{-\frac{1}{2}}\|.$$

□

For example, if $P$ is equal to $I_p$, then $\|\psi_\gamma(x,I_p)\| \leq \frac{2(1+\gamma)}{e\gamma}p$.

## 4.4 Simulation

### 4.4.1 Simulation 1: Robustness of the $\gamma$-estimator

This section describes the results of Monte Carlo simulations carried out in order to examine the robustness of the $\gamma$-estimator for the Gaussian copula model. We generate 500 pseudo-random samples of size 500 from distribution

$$0.9c_{\mathrm{G}}(u, P) + 0.1c_{\mathrm{G}}(u, I_{10}),$$

where $c_{\mathrm{G}}(u, I_{10})$ is equal to the independent copula and $P$ is given by

$$P = \begin{pmatrix} 1 & 0.846 & \cdots & \cdots & 0.846 \\ & 1 & 0.846 & \cdots & 0.846 \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \vdots \\ & & & & 1 \end{pmatrix}.$$

For each sample, we calculate the $\gamma$-estimator $\hat{P}_{\mathrm{GE}}$ for the Gaussian copula model with $\gamma = 0.5$ and the MLE $\hat{P}_{\mathrm{MLE}}$ for the Gaussian copula model. We use the norm $\|\hat{P} - P\|$ as the accuracy measure. Table 4.1 shows the root mean squared error (RMSE) of the norm for the $\gamma$-estimator and MLE. We can see that the norm for the $\gamma$-estimator is less than that for the MLE, so we see that the $\gamma$-estimator is more robust than the MLE.

### 4.4.2 Simulation 2: Detection of Heterogeneous Correlation Structure

The property of the $\gamma$-estimator to detect heterogeneous structure is investigated by a bunch of simulations. A comparison of the $\gamma$-estimator with the MLE for a mixture Gaussian copula (1.1) is also discussed. We conducted two kinds of simulation.

**Simulation 2.1:** The underlying distribution was constructed based on the one factor Gaussian copula model (Hull and White, 2004). Suppose

$$x_i = a_i W + \sqrt{1 - a_i^2}\varepsilon_i,\ i = 1, \ldots, p$$

where $W, \varepsilon_1, \ldots, \varepsilon_p$ have independently the standard normal distribution. Then we see $x \sim \Phi(x, P), u = (\Phi^{-1}(x_1), \ldots, \Phi^{-1}(x_p)) \sim c_{\mathrm{G}}(u, P)$, where $P = (p_{ij})_{ij}$ satisfies

$$p_{ij} = \begin{cases} 1 & i = j \\ & \\ a_i a_j & i \neq j \end{cases}.$$

Let the underlying distribution be equation (1.1), where $c_{\mathrm{G}}(u, P_1)$ and $c_{\mathrm{G}}\ (u, P_2)$ are made from the one factor Gaussian copula model. This model means the dependence structure is expressed by the mixture of Gaussian copulas. Assume $\tau = 0.5$, $P_1$ is made with

$$a = (0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.92, 0.92)^\top,$$

and $P_2$ with

$$a = (-0.92, 0.92, -0.92, 0.92, -0.92, 0.92, -0.92, 0.92, -0.92, 0.92)^\top.$$

Then we have

$$P_1 = \begin{pmatrix} 1 & 0.846 & \cdots & 0.846 \\ & 1 & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 1 & -0.846 & 0.846 & \cdots & -0.846 \\ & 1 & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & & 1 \end{pmatrix}.$$

The $\gamma$-estimator for the Gaussian copula model with $\gamma = 0.7$ is investigated. Initial values of $P$ which are used in calculating the $\gamma$-estimator are

$$AR(\pm 0.1), AR(\pm 0.3), \cdots, AR(\pm 0.9),$$

where $AR(\rho)$ is the correlation matrix whose $(i, j)$ component ($i < j$) is equal to $\rho^{j-i}$. If the $\gamma$-estimator has two components $G_1$ and $G_2$ such that

$$||G_1 - P_1|| < ||G_2 - P_1||,$$

then $G_1$ is thought of as an estimator of $P_1$ and denoted by $\hat{P}_{1,\mathrm{GE}}$. Similarly $G_2$ for $P_2$ and denoted by $\hat{P}_{2,\mathrm{GE}}$.

We adopt the MLE for a mixture Gaussian copula model (1.1). Although $P_1$ and $P_2$ are

the correlation matrices, we tentatively view them to be the covariance matrices and use EM-Algorithm to obtain an approximate MLE. The obtained estimators $\hat{P}_1$ and $\hat{P}_2$ are not necessarily the correlation matrices, so they are transformed into the correlation matrices by

$$\text{diag}(\hat{P}_i)^{-\frac{1}{2}}\hat{P}_i\text{diag}(\hat{P}_i)^{-\frac{1}{2}},$$

which is denoted by $\hat{P}_{i,\text{MLE}}$ for $i = 1, 2$. The initial value of $(\tau, P_1, P_2)$ which is used in calculating the MLE is set to $(0.5, AR(0.5), AR(-0.5))$.

A set of data of size $n$ ($n = 200$ or $500$) was generated from (1.1), and the norm of $\hat{P}_{1,\text{GE}} - P_1$, $\hat{P}_{2,\text{GE}} - P_2$, $\hat{P}_{1,\text{MLE}} - P_1$, and $\hat{P}_{2,\text{MLE}} - P_2$ were calculated. 500 simulations were carried out, and then, we calculated the RMSE of the norm based on 500 norm values obtained by simulation. The results are shown in Table 4.3.

Table 4.2 shows the ratio for the $\gamma$-estimator to detect two correlation matrices. For $n = 500$ nearly 80 percent was successful, and for $n = 1000$ it worked out almost perfectly. From Table 4.3, the MLE had better performance than the $\gamma$-estimator. However this is natural because the MLE is used under no misspecification.

**Simulation 2.2:** Suppose that the underlying distribution is

$$c(u) = \tau_1 c_{\text{G}}(u, P_1) + \tau_2 c_{\text{G}}(u, P_2) + (1 - \tau_1 - \tau_2)c_{\text{G}}(u, I_{10}), \tag{4.14}$$

where $\tau_1 = \tau_2 = 0.45$ and $P_1, P_2$ are the same in Simulation 2.1. The other settings are the same as in Simulation 2.1. The results are shown in Table 4.5.

Table 4.4 shows the ratio for the $\gamma$-estimator to detect two correlation matrices. Com-

pared to the result of Simulation 2.1 the detection rate at $n = 500$ gets worse while at $n = 1000$ the result is almost alike in Table 4.2. From Table 4.5, we find the MLE is considerably underperforming and the $\gamma$-estimator is much better.

Figure 4.1: Illustration of $-C_\gamma(c, c_{\mathrm{G}}(\cdot, P)|Q_{\mathrm{G}})$

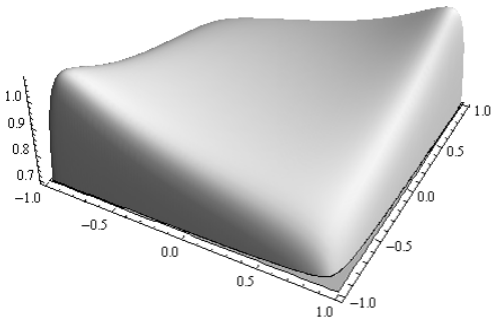



Table 4.1: RMSE of the norm for the $\gamma$-estimator and MLE.

| | $\hat{P}_{\mathrm{GE}}$ | $\hat{P}_{\mathrm{MLE}}$ |
|---|---|---|
| RMSE | 0.155 | 0.808 |

Table 4.2: Ratio of the number of success for the $\gamma$-estimator to detect two correlation matrices.

| $n$ | 500 | 1000 |
|---|---|---|
| ratio | 0.768 | 0.968 |

Table 4.3: RMSE of the norm of $\hat{P}_{1,\mathrm{GE}} - P_1$, $\hat{P}_{1,\mathrm{MLE}} - P_1$, $\hat{P}_{2,\mathrm{GE}} - P_2$, and $\hat{P}_{2,\mathrm{MLE}} - P_2$.

| | $n$ | $\hat{P}_{1,\mathrm{GE}}$ | $\hat{P}_{1,\mathrm{MLE}}$ | $\hat{P}_{2,\mathrm{GE}}$ | $\hat{P}_{2,\mathrm{MLE}}$ |
|---|---|---|---|---|---|
| RMSE | 500 | 0.600 | 0.184 | 0.476 | 0.186 |
| | 1000 | 0.479 | 0.127 | 0.431 | 0.129 |

Table 4.4: Ratio of the number of success for the $\gamma$-estimator to detect two correlation matrices.

| $n$ | 500 | 1000 |
|---|---|---|
| ratio | 0.61 | 0.966 |

Table 4.5: RMSE of the norm of $\hat{P}_{1,\text{GE}} - P_1$, $\hat{P}_{1,\text{MLE}} - P_1$, $\hat{P}_{2,\text{GE}} - P_2$, and $\hat{P}_{2,\text{MLE}} - P_2$.

| | $n$ | $\hat{P}_{1,\text{GE}}$ | $\hat{P}_{1,\text{MLE}}$ | $\hat{P}_{2,\text{GE}}$ | $\hat{P}_{2,\text{MLE}}$ |
|---|---|---|---|---|---|
| RMSE | 500 | 0.494 | 0.946 | 0.563 | 0.952 |
| | 1000 | 0.468 | 1.010 | 0.438 | 1.032 |

# Chapter 5

# Summary and Discussion

We have considered the local learning based on local minimization of the $\gamma$-divergence. Applying this local minimization method to cluster analysis, the spontaneous clustering is proposed. In the spontaneous clustering, the centers of clusters are defined as the local minimum points of the $\gamma$-loss function. On the other hand, we apply the local minimization method to the $\gamma$-estimation of the Gaussian copula parameter to detect heterogeneous correlation structure. In this case, the local minimum points of the $\gamma$-loss function are employed to estimate each correlation matrix. 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, the 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 non-convexity is generally intractable, the proposed methods benefit from the non-convexity, which makes our method unique and interesting. The idea to use local

minimum points of the $\gamma$-loss function can be applied to other statistical methods. For example, the idea is applied to principal component analysis (Mollah et al., 2010) and could be applied to regression analysis.

The spontaneous clustering does not require the information about the number of clusters *a priori* and can find it automatically if the value of the power index $\gamma$ is properly fixed. In contrast, existing methods such as $K$-means and model-based clustering need the number of clusters. Instead of the number of clusters, the value of $\gamma$ has to be determined in the spontaneous clustering. Two methods to determine the value of $\gamma$ have been proposed in this thesis. One is a heuristic method, which depends on the range of the data. Our simulations show that this method has satisfactory performance and can thus be used in most situations. A more sophisticated choice based on AIC is also proposed, although it requires more computational effort. When selecting $\gamma$, we first considered a cross validation technique, one of the common procedures to select the optimal value of a tuning parameter (Hastie et al., 2009). Mollah et al. (2010) proposed using cross validation for $\gamma$ selection. However, the method does not work well for the spontaneous clustering. Hence we employ AIC instead. We have demonstrated that the proposed clustering works well by the simulations and the application to the data. Though we did not consider how to determine the value of $\gamma$ for the $\gamma$-estimation of the Gaussian copula parameter, the method based on AIC could be possible for this problem, but it is currently a future problem.

In the spontaneous clustering, the proposed method employs the local minimum points of equation (3.2) or (3.6). Then, it assigns the data into clusters with the Mahalanobis distance. We have proposed an iteration algorithm to find the local minimum points. There are,

however, a bunch of possibility for optimization and cluster assignment. For optimization, we could, for example, replace Step 1-1 to 1-3 in section 3.2.2 with $n$ initial values using all the data. This is nothing but MSC with estimation of covariance matrices. For cluster assignment, we could use MSC's trajectory-based assignment (Chen et al., 2013).

In the $\gamma$-estimation of the Gaussian copula model, we choose the measure in terms of invariance. However the $\gamma$-estimator obtained is equal to the estimator with normal distribution as a statistical model, so it seems natural. If we use Lebesgue measure in calculating the $\gamma$-estimator for Gaussian copula model, we cannot calculate the projective power entropy for all the value of $\gamma$ and $P$.

Another issue in the $\gamma$-estimation of the Gaussian copula model is to what extent the methodology here works for time series data. Because the basic premise of this problem is that we have data as quantiles, our method would fit, for example, the modeling of unconditional loss distribution (McNeil et al. (2005), p.28). Such a case is of particular interest when the time horizon over which we measure our losses is relatively large. When we are working on the conditional modeling, our method should be regarded as a tool for the post analysis. As a typical case, we may want to apply our mixture copula approach to multivariate log-return series which are appropriately standardized and declustered by the multivariate GARCH model fitted to them. See Zhang et al. (2011) for more details.

# Acknowledgements

I would like to express my sincere gratitude to Prof. S. Eguchi for giving me the great opportunity to do research with him at The Graduate University for Advanced Studies. I am very grateful for his valuable guidance, useful advice, and financial support. I deeply appreciate Assoc. Prof. Y. Kawasaki , Prof. S. Kuriki, and Assoc. Prof. Y. Nishiyama, who were my secondary supervisors and gave me lectures about time series analysis, basics of mathematical statistics, and martingale theory. I would like to provide special thanks to Assoc. Prof. Y. Kawasaki, who proofread my two academic papers.

Thanks are due to Prof. H. Fujisawa for valuable comments in almost all my presentations, Dr. O. Komori for useful discussion about replies to reviewer's comments, and my colleague Md. Ashad Alam for his help in writing English. My colleagues always encouraged and relaxed me, especially F. Kobayashi. I. Kawaji, Y. Hasebe, N. Watanabe, and other staff of The Institute of Statistical Mathematics have supported me a lot. I would like to thank them.

I would like to take this opportunity to express my sincere gratitude to S. Nagumo and F. Nagumo, who are the owners of the restaurant near my apartment. They treated me as a family member. Owing to their encouragement, my thesis was finally completed.

Last but not least, I would like to express my best thank to my parents, Yoshifumi and Yukiko Notsu, and my brothers and sister, for financial support and warm encouragement. Without their help, I could not have completed this thesis.

# Bibliography

Amari, S. & Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society.

Bańdossy, A. (2006). Copula-based geostatistical models for groundwater quality parameters. *WATER RESOURCES RESEARCH*, 42(11).

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559.

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

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217.

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

Chen, T., Hsieh, D., Hung, H., Tu, I., Wu, P., Wu, Y., Chang, W., & Huang, S. (2013). gamma-SUP: a clustering algorithm for cryo-electron microscopy images of asymmetric particles. *To appear in The Annals of Applied Statistics*.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 17(8):790–799.

Cichocki, A. & Amari, S. (2010). Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568.

Dall'Aglio, G., Kotz, S., & Salinetti, G. (1991). *Advances in Probability Distributions with Given Marginals*, volume 67 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht.

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

Demarta, S. & McNeil, A. J. (2005). The $t$ copula and related copulas. *International Statistical Review*, 73(1):111–129.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Eguchi, S. (1985). A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima Mathematical Journal*, 15(2):341–391.

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

Eguchi, S., Komori, O., & Kato, S. (2011). Projective power entropy and maximum Tsallis entropy distributions. *Entropy*, 13(10):1746–1764.

Einbeck, J. (2011). Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of pattern recognition research*, 6(2):175–192.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 5 th edition.

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

Genest, C., Ghoudi, K., & Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

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

Huber, P. J. (1981). *Robust Statistics*. John Wiley.

Hull, J. & White, A. (2004). Valuation of a CDO and an $n$-th to Default CDS Without Monte Carlo Simulation. *The Journal of Derivatives*, 12(2):8–23.

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.

Joe, H. (2001). *Multivariate Models and Dependence Concepts*. Chapman and Hall.

Jones, M. C., Hjort, N. L., Harris, I. R., & Basu, A. (2001). A comparison of related density ‐ based minimum divergence estimators. *Biometrika*, 88(3):865–873.

Konishi, S. & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Li, D. (2001). On default correlation: a copula function approach. *Journal of Fixed Income*, 9:43–54.

Liese, F. & Vajda, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

Magnus, J. R. & Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, revised edition.

McNeil, A. J., Frey, R., & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

Minami, M. & Eguchi, S. (2002). Robust blind source separation by beta divergence. *Neural Computation*, 14:1859–1886.

Mollah, M. N. H., Minami, M., & Eguchi, S. (2006). Exploring latent structure of mixture ICA models by the minimum $\beta$-divergence method. *Neural Computation*, 18:166–190.

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

Murata, N., Takenouchi, T., Kanamori, T., & Eguchi, S. (2004). Information geometry of $U$-boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer.

Notsu, A., Kawasaki, Y., & Eguchi, S. (2013). Detection of heterogeneous structures on the Gaussian copula model using projective power entropy. *ISRN Probability and Statistics*.

Notsu, A., Komori, O., & Eguchi, S. (2014). Spontaneous clustering via minimum gamma-divergence. *Neural Computation*, 26(2).

Pardo, L. (2006). *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43.

Rényi, A. (1961). On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561.

Sharma, B. D. & Mittal, D. P. (1977). New non-additive measures of relative information. *Journal of Combinatory Information and Systems Science*, 2:122–133.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. CHAPMAN and HALL.

Sklar, A. (1959). Fonctions de répartition à $n$ dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.

Song, P. X.-K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, 65(1):60–68.

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.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487.

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.

Yoshiba, T. (2013). Risk aggregation by a copula with a stressed condition. *Bank of Japan Working Paper Series*.

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

Zhang, J. (2004). Divergence function, duality, and convex analysis. *Neural Computation*, 16:159–195.

Zhang, R., Czado, C., & Min, A. (2011). Efficient maximum likelihood estimation of copula based meta $t$-distributions. *Computational Statistics and Data Analysis*, 55(3):1196–1214.