Asymptotic properties of kernel k-means for high dimensional data

Kento Egashira[a], **Kazuyoshi Yata**[b], **Makoto Aoshima**[b]

[a]**Department of Information Sciences, Tokyo University of Science**
[b]**Institute of Mathematics, University of Tsukuba**

1 Introduction

Cluster analysis can be divided into two types: hierarchical and partitional. Hierarchical clustering groups data into dendrograms based on their cluster similarities determined by a preset linkage function. A dendrogram enables the observation of the process of merging or dividing clusters. For discussions on hierarchical cluster analyses, see the works of Everitt et al. [6] and Hastie et al. [8], among others. Partitional clustering, as its name suggests, divides data into a pre-determined number of clusters. K-means can be given on behalf of partitional clustering. Notably, k-means has been approved as a useful tool for analyzing microarray gene expression data. A characteristic of such data is that the number of variables was considerably larger than the sample size, giving high-dimensional, low-sample-size (HDLSS) scenarios. Substantial work on HDLSS asymptotic clustering has been performed in recent years. For example, Liu et al. [10] proposed a two-way split statistical-significance-of-clustering (SigClust) method for HDLSS data. Ahn et al. [1] proposed hierarchical divisive clustering for high-dimensional asymptotics. Huang et al. [7] modified SigClust using a soft thresholding approach. Kimes et al. [9] proposed a method for sequentially testing the statistical significance of hierarchical clustering by controlling the family-wise error rate in HDLSS settings. Yata and Aoshima [14] presented the consistency properties of sample principal component scores and applied them to clustering in high-dimensional settings. Nakayama et al. [12] investigated HDLSS clustering using kernel principal component analysis. Borysov et al. [3] studied the behaviors of hierarchical clustering under several asymptotic settings from a moderate dimension for HDLSS; however, the theoretical assumptions were considered to be strict for HDLSS data owing to several simultaneous asymptotic settings. Egashira et al. [5] explores practical assumptions to indicate the behavior of hierarchical clustering and obtained theoretical results in multiclass settings. Given this background, asymptotic properties of k-means in the HDLSS settings seems to have not been studied sufficiently.

In this talk, we investigate k-means when both the dimension and sample size approach infinity at first. Then, we explores kernel k-means in the HDLSS context theoretically. Especially, we mension kernel k-means with gaussian kernel function and compare performance of it to conventional k-means in the multiclass HDLSS context.

2 Introduction of k-means

In this section, we introduce k-means. The k-means algorithm is a clustering method used to divide a dataset into distinct clusters. It aims to minimize the within-cluster variance, which is a measure of how similar the data points within each cluster are to each other.

The k-means algorithm applied to a given dataset $\boldsymbol{X}$ can be formulated as the following optimization problem, using a pre-defined number of clusters k. The mathematical formulation of k-means is given by

$$\{\widehat{\boldsymbol{C}}_1, ..., \widehat{\boldsymbol{C}}_k\} = \underset{\boldsymbol{C}_1,...,\boldsymbol{C}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x}\in\boldsymbol{C}_i} \|\boldsymbol{x} - \overline{\boldsymbol{C}}_i\|^2$$
$$\text{subject to } \cup_{i=1}^{k} \boldsymbol{C}_i = \boldsymbol{X},\ \boldsymbol{C}_i \cap \boldsymbol{C}_j = \emptyset \ \ (i \neq j).$$

where $\overline{\boldsymbol{C}}_i =$ and $\|\cdot\|$is Euclidean norm. $\{\widehat{\boldsymbol{C}}_1, ..., \widehat{\boldsymbol{C}}_k\}$ is given as the result of clustering by k-means.

The optimization problem above is generally solved by the following k-means algorithm using k initial centroids.

Here is a step-by-step explanation of the k-means algorithm:

Initialize: Set k initial centroids $\boldsymbol{c}_i$ $(\in \boldsymbol{X})$, $i = 1, ..., k$.

Assign: For each given data point $\boldsymbol{x} \in \boldsymbol{X}$, assign it to $\boldsymbol{C}_i$ if $i = \arg\min_{j=1}^{k} \|\boldsymbol{x} - \boldsymbol{c}_j\|^2$. Repeat this process for all data points to construct sets $\boldsymbol{C}_i, i = 1, ..., k$.

Update: Treat the arithmetic mean of the data points within each $\boldsymbol{C}_i$ as the new initial value, and execute Step 2 to update sets $\boldsymbol{C}_i, i = 1, ..., k$. Repeat this step until the sets from the previous step match the updated sets.

Terminate: Define the converged sets from Step 3 as $\widehat{\boldsymbol{C}}_i, i = 1, ..., k$ which is the result of the k-means algorithm.

The final result of the k-means algorithm is a set of k clusters, each represented by its centroid. The algorithm strives to minimize the sum of squared distances between the data points and their assigned centroids. It's important to note that the k-means algorithm can be sensitive to the initial placement of the centroids and may converge to a suboptimal solution. To mitigate this, it is common to run the algorithm multiple times with different initializations and choose the clustering result with the lowest overall within-cluster variance. We acknowledge the importance of considering computational complexity and convergence speed while primarily focusing on investigating the theoretical properties of k-means and kernel k-means in high dimensional settings. In the next section, we show asymptotic properties of k-means under HDLSS settings.

3 Asymptotic Behaviors of k-means for binary class

Suppose we have q independent and d-variate populations, Π_i, with an unknown mean vector $\boldsymbol{\mu}_i$, and an unknown covariance matrix, $\boldsymbol{\Sigma}_i$ for $i = 1, ..., q$. We suppose that

$$\mathrm{tr}(\boldsymbol{\Sigma}_i) \leq \mathrm{tr}(\boldsymbol{\Sigma}_j)$$

for $i < j$ without loss of generality. We have independent and identically distributed observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$ from Π_i for $i = 1, ..., q$. Let $N_q = \sum_{i=1}^{q} n_i$, $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}\}$, $K_i = \mathrm{Var}[\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2]$ for $i = 1, ..., q$, $\Delta_{\Sigma,ij} = |\mathrm{tr}(\boldsymbol{\Sigma}_i) - \mathrm{tr}(\boldsymbol{\Sigma}_j)|$, $\Delta_{ij} = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$ for $i, j = 1, ..., q$.

In this section, we consider asymptotic properties in binary setting, $q = 2$ as $d \to \infty$ when N_2 is fixed. Suppose two class situation, $q = 2$. We introduce the following assumptions:

(A-i): $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_{12}^2 \to 0$, $i = 1, 2$, as $d \to \infty$;

(A-ii): $K_i/\Delta_{12}^2 \to 0$, $i = 1, 2$, as $d \to \infty$.

Note that $K_i = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ when Π_i is Gaussian; thus, (A-i) and (A-ii) are equivalent when $\Pi_i, i = 1, 2$ are Gaussian. These assumptions are fairly common in HDLSS settings. See the works of Aoshima and Yata [2], Nakayama et al. [11], and Egashira et al. [4, 5].

Theorem 3.1. *Suppose $q = 2$ and initial observation, $\boldsymbol{c}_i \in \boldsymbol{X}_i$ for $i = 1, 2$. Assume (A-i), (A-ii) and some regularity conditions. When*

$$\limsup_{d\to\infty} \frac{\Delta_{\Sigma,12}}{\Delta_{12}} < 1$$

holds, the probability, $P(\{\widehat{\boldsymbol{C}}_1, \widehat{\boldsymbol{C}}_2\} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}) \to 1$ as $d \to \infty$ when N_2 is fixed.

Acknowledgments

The research of the second author was partially supported by Grant-in-Aid for Scientific Research (C), JSPS, under Contract Number 22K03412. The research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 20H00576 and 22K19769.

References

[1] Ahn, J., Lee, M.H., Yoon, Y.J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, 22, 443–464.

[2] Aoshima, M., Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66, 983–1010.

[3] Borysov, P., Hannig, J., Marron, J.S. (2014). Asymptotics of hierarchical clustering for growing dimension. *Journal of Multivariate Analysis*, 124, 465–479.

[4] Egashira, K., Yata, K., Aoshima, M. (2021). Asymptotic properties of distance weighted discrimination and its bias correction for high-dimension, low-sample-size data. *Japanese Journal of Statistics and Data Science*, 4, 821–840.

[5] Egashira, K., Yata, K., Aoshima, M. (2023). Asymptotic properties of hierarchical clustering in high-dimensional settings. *Journal of Multivariate Analysis*, in press.

[6] Everitt, B.S., Landau, S., Leese, M. (2001). Cluster Analysis. Arnold, New York.

[7] Huang, H., Liu, Y., Yuan, M., Marron, J.S. (2015). Statistical Significance of Clustering using Soft Thresholding. *Journal of Computational and Graphical Statistics*, 24, 975–993.

[8] Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer, New York.

[9] Kimes, P.K., Liu, Y., Neil, H.D., Marron, J.S. (2017) Statistical significance for hierarchical clustering. *Biometrics*, 73, 811–821.

[10] Liu, Y., Hayes, D.N., Nobel, A., Marron, J.S.(2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association*, 103, 1281–1293.

[11] Nakayama, Y., Yata, K., Aoshima, M. (2017). Support vector machine and its bias correction in high-dimension, low-sample-size settings. *Journal of Statistical Planning and Inference*, 191, 88–100.

[12] Nakayama, Y., Yata, K., Aoshima, M. (2021). Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings. *Journal of Multivariate Analysis*, 185, 104779.

[13] Yata, K., Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, **101**, 2060–2077.

[14] Yata, K., Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics*, 47, 899–921.