Hierarchical clustering and its asymptotic behaviors in high-dimensional settings

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

[a]**Degree Programs in Pure and Applied Sciences, Graduate School of Science and Technology, University of Tsukuba**
[b]**Institute of Mathematics, University of Tsukuba**

1 Introduction

Hierarchical clustering is a methodology to group a set of datas by building dendrogram based on a similarity or a dissimilarity between clusters so that datas in a cluster are similar in the sense of pre-determined linkage function. In hierarchical clustering, one can observe a process how a cluster is combined or devided through dendrogram on graphic. Hierarchical clustering has been approved as useful tool for analysis of gene expression microarray data. In fact, applications of hierarchical clustering on gene expression microarray data are given by Eisen et al. [4], Perou et al. [9], Bhattacharjee et al. [2], among others. A characteristic of datas used in Eisen et al. [4], Perou et al. [9] and Bhattacharjee et al. [2] is that the number of variables is much larger than sample size. This type of data represented by gene expression microarray data is called high-dimension, low-sample-size (HDLSS) data. Substantial work about clustering has been done on HDLSS asymptotics in recent years. Liu et al. [7] proposed a two-way split clustering called "statistical significance of clustering(SigClust)" especially for HDLSS data. Ahn et al. [1] proposed a hierarchical divisive clustering and considered its high dimensional asymptotics. Huang et al. [5] developed the SigClust by Liu et al. [7] with soft thresholding approach. Kimes et al. [6] proposed a methodology to sequentialy test statistical significance of hierarchical clustering controling the family-wise error rate in HDLSS settings. Yata and Aoshima [11] gave consistency properties of sample principal component scores and applied it to clustering under high dimensional settings. Nakayama et al. [8] investigated clustering by kernel principal component analysis for HDLSS data. Borysov et al. [3] studied behaviors of hierarchical clustering under several asymptotic settings from moderate dimension through HDLSS, nevertheless it is considered that theoretical assumptions are strict for HDLSS data due to having discussions on several asymptotic settings at once. Given this background, we focus on HDLSS settings and consider asymptotic properties of hierarchical clustering with several linkage functions.

In this talk, we investigate the hierarchical clustering theoretically in the HDLSS

context as dimension goes to infinity while sample size is fixed.

2 Formulation of Hierarchical Clustering

Hierarchical clustering is generally classified into two types : agglomerative clustering and divisive one. Hierarchical agglomerative clustering is a bottom-up approach. At first, every data point is considered as a cluster of its own. Then, two nearest clusters are combined at each procedure. In the end, all datas belong to one single cluster. Hierarchical divisive clustering is so-called a top-down approach. At first, every data point is considered as one single cluster. Then, a clusters is split up into two clusters at each procedure. In the end, every data belongs to a single cluster of its own. To proceed the hierarchical agglomerative clustering, we need to consider $O(n^3)$ computations in whole process when the data set contains n samples. On the other hand, to proceed the hierarchical divisive clustering, it is necessary to consider the all devisions of the dataset into two nonempty subsets which require $2^{n-1} - 1$ computations when the data set contains n samples. The computation number grows exponentially and easily become prohibitive. Thus, we focus on the hierarchical agglomerative clustering in this talk.

2.1 Hierarchical clustering

The function to measure a similarity or a dissimilarity between clusters is called linkage function, which are introduced in Section 2.2. Generally, a dendrogram built by hierarchical clustering shows an arrangement of clusters and distances between clusters. An intersection indicates that two clusters are combined and a height of horizontal segment at any intersection stands for a distance between clusters. A dendrogram shows a process of hierarchical clustering from the bottom toward the top, which means that lower an intersection is, earlier the cluster are combined.

2.2 Linkage function

In this talk, we only take the Euclidean distance to define linkage functions. Suppose we have two sets $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{in_i}\}$ of size n_i for $i = 1, 2$ and $d(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is a distance between samples $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. In this talk, $d(\boldsymbol{z}_1, \boldsymbol{z}_2) = \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$, where $\|\cdot\|$ denotes the Euclidean norm. Then, the linkage functions, the distance between clusters, can be defined as follows.

1. Single linkage function

$$D_1(\boldsymbol{X}_1, \boldsymbol{X}_2) = \min_{\boldsymbol{x}_1 \in \boldsymbol{X}_1, \boldsymbol{x}_2 \in \boldsymbol{X}_2} d(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

2. Average linkage function

$$D_2(\boldsymbol{X}_1, \boldsymbol{X}_2) = \frac{1}{n_1 n_2} \sum_{\boldsymbol{x}_1 \in \boldsymbol{X}_1} \sum_{\boldsymbol{x}_2 \in \boldsymbol{X}_2} d(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

3. Ward's linkage function

$$D_3(\boldsymbol{X}_1, \boldsymbol{X}_2) = \left(2\left(SS(\boldsymbol{X}_1 \cup \boldsymbol{X}_2) - SS(\boldsymbol{X}_1) - SS(\boldsymbol{X}_2)\right)\right)^{1/2}$$

where $SS(Z) = \sum_{\boldsymbol{z} \in Z} \|\boldsymbol{z} - \sum_{\boldsymbol{z}' \in Z} \boldsymbol{z}'/|Z|\|^2$ and $|Z|$ is the number of data in Z for any set Z. Single linkage function measures distance between clusters as closest distance between datas from each clusters. Average linkage function measures distance between clusters as average distance between datas from each clusters. Example of application of the hierarchical clustering with average linkage function on gene expression microarray data is found in Eisen et al. [4]. Ward's linkage function is proposed by Ward [10]. We emphasize that Ward's linkage function can be expressed as

$$D_3(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\frac{2n_1 n_2}{n_1 + n_2}} \left\| \frac{1}{n_1} \sum_{j=1}^{n_1} \boldsymbol{x}_{1j} - \frac{1}{n_2} \sum_{j=1}^{n_2} \boldsymbol{x}_{2j} \right\|.$$

This indicates that Ward's linkage function magnifies a distance between clusters due to the coefficient $\sqrt{2n_1 n_2/(n_1 + n_2)}$ in general. Although there are a large number of linkage functions besides the above, we focus on the single, average, and Ward's linkage functions here.

In the next section, we show asymptotic properties of hierarchical clustering with single, average and Ward's linkage functions under HDLSS settings.

3 Asymptotic behaviors

Borysov et al. [3] supposed that the number of populations is potentially two and proposed three asymptotic behaviors about hierarchical clustering as follows:

(A): Every data point from one population combined only with any cluster from the same population and every data point from the other population combined only with any cluster from the same population before the last step. Then, one cluster and the other cluster will be combined in the last step.

(B): Every data point from one population combined with any cluster from the same population. Then, any point from the other population will be added sequentially one by one.

(AB): Every data point from one population combined with any cluster from the same population. Then, after that, one subcluster consist of data points only from the other population created at least.

The behavior (A) occurs when the every distance between data points from one population are smaller than from the other population and any distances between data points from one population and the other population and the every distance in data points from the other population are smaller than any distance between data points from one population and the other population. The behavior (B) occurs when the every distance between data points from one population are smaller than from the other population and any distances between data points from one population and the other population and the any distances in data points from the other population are larger than any distance between data points from one population and the other population. The behavior (AB) is an event between (A) and (B). (AB) occurs when the every distance between data points from one population are smaller than data points from the other population and any distances between data points from one population and the other population and some distances between data points from the other population are smaller than a distance between mixed cluster of all points from one population and some points from the other population and any point from the other population.

Borysov et al. [3] studied the difference of the behavior theoretically under several asymptotic settings from moderate dimension through HDLSS. However, Borysov et al. [3] considered strict assumptions especially for HDLSS data due to having discussions on several asymptotic settings at once. We derive asymptotic properties of hierarchical clustering under mild and practical assumptions.

Suppose we have two independent and d-variate populations, Π_i having unknown mean vector $\boldsymbol{\mu}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}_i$ for $i = 1, 2$. We suppose $\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ for simplicity. We consider 2 classes case in this paper which is the fundamental case to generalize a latent number of populations.

Suppose that we have independent and identically distributed (i.i.d.) observations, $\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}$ from Π_1 and $\boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}$ from Π_2. Let $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}\}$ and $K_i = \mathrm{Var}[\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2]$ for $i = 1, 2$. As necessary, we assume for the asymptotic setting that only d grows that

(A-i): $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_M^2 \to 0$ for $i = 1, 2$ as $d \to \infty$ and n_1 and n_2 are fixed;

(A-ii): $K_i/\Delta_M^2 \to 0$ for $i = 1, 2$ as $d \to \infty$ and n_1 and n_2 are fixed,

where $\Delta_M = \max\{\Delta, \Delta_\Sigma\}$, $\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, and $\Delta_\Sigma = |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|$. These assumptions are fairly common under HDLSS settings. Then, we have the following results.

Theorem 3.1. *Assume (A-i), (A-ii) and some regularity conditions.*

(1) *If* $\limsup_{d\to\infty} \dfrac{\Delta_\Sigma}{\Delta} < 1$, *the probability of hierarchical behavior (A) when single or average linkage function is used converges to* 1 *as* $d \to \infty$ *when* n_1 *and* n_2 *are fixed.*

(2) *If* $\liminf_{d\to\infty} \dfrac{\Delta_\Sigma}{\Delta} > 1$, *the probability of hierarchical behavior (B) when single or average linkage function is used converges to* 1 *as* $d \to \infty$ *when* n_1 *and* n_2 *are fixed.*

We obtain the threshold to devide the asymptotic behavior (A) and (B) under mild conditions (A-i) and (A-ii). Behavior (A) happens when the distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are larger than the difference between $\mathrm{tr}(\boldsymbol{\Sigma}_1)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_2)$. Behavior (B) happens when the distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are smaller than the difference between $\mathrm{tr}(\boldsymbol{\Sigma}_1)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_2)$. It follows from Theorem 3.1 that behavior for hierarchical clustering by single and average linkage functions are asymptotically same. But, we will observe that convergence rates differ depending on linkage functions by numerical simulations in talk.

Theorem 3.2. *Assume (A-i), (A-ii) and some regularity conditions.*

(1) *If* $\limsup_{d\to\infty} \dfrac{\Delta_\Sigma}{n_1\Delta} < 1$, *the probability of hierarchical behavior (A) when Ward's linkage function is used converges to* 1 *as* $d \to \infty$ *when* n_1 *and* n_2 *are fixed.*

(2) *If* $\liminf_{d\to\infty} \dfrac{\Delta_\Sigma}{n_1\Delta} > 1$, *the probability of hierarchical behavior (B) when Ward's linkage function is used converges to* 1 *as* $d \to \infty$ *when* n_1 *and* n_2 *are fixed.*

Unlike Theorem 3.1 with single and average linkage functions, the boundary condition depends on the sample size of the cluster with smaller variance than the other. This difference from the case with single and average linkage functions makes the hierarchical clustering with Ward's linkage function prone to fall into asymptotic behavior (A). It is considered as natural consequences because the distance between two clusters measured by Ward's linkage function is generally expanded under the influence of the sample sizes.

Acknowledgments

The research of the first author was partially supported by JST SPRING under Grant Number JPMJSP2124. 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] Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 13790–13795.

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

[4] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14863–14868.

[5] 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.

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

[7] 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.

[8] 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.

[9] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406, 747–752.

[10] Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

[11] 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.