

Enhanced Gaussian kernel k-means for high-dimensional, low-sample-size data

Kento Egashira^a, Kazuyoshi Yata^b, Makoto Aoshima^b

^aDepartment of Information Science and Technology, Tokyo University of
Science

^bInstitute of Mathematics, University of Tsukuba

1 Introduction

Clustering high-dimensional data plays an important role in a wide range of applications, including gene expression analysis, where the dimension can be substantially larger than the sample size. In such high-dimensional, low-sample-size (HDLSS) settings, clustering methods are routinely used to identify latent groups of observations. Among them, k-means is one of the most widely used clustering methods because of its simple formulation and computational efficiency. However, high dimensionality can substantially affect the behavior of clustering procedures, and a theoretical understanding of clustering methods under HDLSS asymptotics is therefore important.

Recent research has extensively investigated the asymptotic properties of clustering methods for HDLSS data. The theoretical analysis of k-means clustering in high-dimensional settings has been conducted predominantly under sparsity assumptions. In other words, many existing studies rely on the premise that only a relatively small subset of variables contributes to the clustering structure. Within this framework, Sun et al. [6] proposed a regularized k-means procedure and established its asymptotic consistency. Arias-Castro and Pu [1] developed a sparse k-means method based on a different formulation. Chang et al. [3] introduced a regularization scheme based on an ℓ_∞/ℓ_0 -type penalty. Chakraborty and Das [2] considered a lasso-weighted k-means algorithm for high-dimensional data. Yang and Benjamin [7] proposed a sparse extension of the possibilistic c-means algorithm originally introduced by Krishnapuram and Keller [5], adapting it to high-dimensional settings. Despite these advances, the asymptotic properties of the standard k-means algorithm in HDLSS settings remain less well understood, particularly without imposing sparsity assumptions.

In our previous study, Egashira et al. [4], we theoretically investigated the asymptotic properties of k-means under HDLSS settings without imposing sparsity assumptions. We showed that, even when appropriate initial observations are selected, the clustering criterion can contain a bias induced by differences in the within-cluster covariance structures. This bias can affect the clustering result and prevent the recovery of the underlying population structure. To overcome this problem, in Egashira et al. [4], we proposed a bias-corrected k-means procedure that removes the bias arising from the within-cluster covariance structures. We also considered k-means with multiple initial observations and

showed that using multiple observations for initialization can relax the conditions required for consistent clustering. These results indicate that both the initialization strategy and the bias inherent in the clustering criterion play important roles in the asymptotic behavior of k-means in HDLSS settings.

Kernel methods provide a natural extension of k-means for capturing nonlinear relationships through a feature-space representation. In particular, Gaussian kernel k-means is widely used for nonlinear clustering. However, our previous theoretical analysis showed that, under HDLSS asymptotics, Gaussian kernel k-means does not necessarily provide an asymptotic advantage over conventional k-means. More specifically, the consistency conditions of Gaussian kernel k-means and conventional k-means are asymptotically equivalent under the HDLSS setting. This result indicates that, although kernelization provides a nonlinear feature-space representation, it does not by itself eliminate the fundamental difficulty associated with the high-dimensional bias.

The above observation naturally raises the question of whether Gaussian kernel k-means can be improved by incorporating mechanisms that explicitly address the difficulties identified for conventional k-means. In particular, the bias-correction and multiple-initialization strategies developed in our previous study provide natural candidates for improving Gaussian kernel k-means. However, it is not immediately clear whether these strategies retain their effectiveness after kernelization, since the transformation into a feature space changes the form of the clustering criterion and its asymptotic behavior.

In this talk, we investigate this question by incorporating two ideas into Gaussian kernel k-means: multiple initial observations and bias correction. We develop enhanced Gaussian kernel k-means procedures based on these two ideas and derive their asymptotic properties under HDLSS settings. In particular, we clarify how multiple initial observations and bias correction affect the consistency conditions of Gaussian kernel k-means and whether these modifications can improve its asymptotic performance.

The proposed procedures are theoretically compared with conventional k-means, bias-corrected k-means, and Gaussian kernel k-means. These comparisons clarify the relationship between kernelization, initialization, and bias-correction in high-dimensional clustering. Numerical studies are also conducted to examine the finite-sample performance of the proposed procedures.

2 Asymptotic properties of kernel k-means

In this section, we introduce the kernel k-means algorithm and review its asymptotic performance under the HDLSS setting. We first summarize the results for conventional k-means established in Egashira et al. [4], and then present the corresponding asymptotic results for Gaussian kernel k-means, thereby clarifying the relationship between the two approaches.

Kernel k-means is a natural extension of conventional k-means that incorporates a kernel function to measure the similarity between observations. Here, we introduce the kernel k-means algorithm using a kernel function $k(\cdot, \cdot)$ for a given dataset $\mathbf{X}$ and its

asymptotic properties under high-dimensional settings. The kernel k-means algorithm is described as follows.

Algorithm 1 (kernel k-means algorithm).

Step 1: Set K initial centroids $\mathbf{c}_{0i} (\in \mathbf{X})$, $i \in \{1, \dots, K\}$.

Step 2: For each given data point $\mathbf{x} \in \mathbf{X}$, assign $\mathbf{x}$ to a set $\mathbf{C}_i$ if

$$i = \underset{j=1}{\operatorname{argmin}}^K (k(\mathbf{x}, \mathbf{x}) + k(\mathbf{c}_{0j}, \mathbf{c}_{0j}) - 2k(\mathbf{x}, \mathbf{c}_{0j})).$$

Repeat this process for all data points to construct sets $\mathbf{C}_i, i \in \{1, \dots, K\}$.

Step 3: For each given data point $\mathbf{x} \in \mathbf{X}$, assign $\mathbf{x}$ to $\mathbf{C}_i$ if

$$i = \underset{j=1}{\operatorname{argmin}}^K \left(k(\mathbf{x}, \mathbf{x}) + \frac{1}{|\mathbf{C}_j|^2} \sum_{\mathbf{c}_1, \mathbf{c}_2 \in \mathbf{C}_j} k(\mathbf{c}_1, \mathbf{c}_2) - \frac{2}{|\mathbf{C}_j|} \sum_{\mathbf{c}_1 \in \mathbf{C}_j} k(\mathbf{x}, \mathbf{c}_1) \right)$$

to update sets $\mathbf{C}_i, i \in \{1, \dots, K\}$. Repeat this step until the sets from the previous step match the updated sets.

Step 4: Define the converged sets from Step 3 as $\hat{\mathbf{C}}_i, i \in \{1, \dots, K\}$ which is the result of kernel k-means algorithm.

Suppose that there are 2 independent d -variate populations, π_1 and π_2 , where the i th population has an unknown mean vector $\boldsymbol{\mu}_i$ and an unknown covariance matrix $\boldsymbol{\Sigma}_i$ for $i \in \{1, 2\}$. We suppose that

$$\operatorname{tr}(\boldsymbol{\Sigma}_1) \leq \operatorname{tr}(\boldsymbol{\Sigma}_2)$$

and we assume that

$$\limsup_{d \rightarrow \infty} (\|\boldsymbol{\mu}_i\|^2/d) < \infty, \quad \liminf_{d \rightarrow \infty} (\operatorname{tr}(\boldsymbol{\Sigma}_i)/d) > 0, \quad \limsup_{d \rightarrow \infty} (\operatorname{tr}(\boldsymbol{\Sigma}_i)/d) < \infty \quad (1)$$

for $i \in \{1, 2\}$. We have independent and identically distributed (i.i.d.) n_i observations, $\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}$, from π_i for $i \in \{1, 2\}$. Let $\mathbf{X}_i = \{\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i}\}$ for $i \in \{1, 2\}$ and $\mathbf{X} = \cup_{i=1}^2 \mathbf{X}_i$. We assume $n_i \geq 2$ for $i \in \{1, 2\}$. Let $N = n_1 + n_2$. To state the theoretical results, we introduce the following notations. Let $\Delta_{12} = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$, $\Delta_{*12} = \Delta_{12} + \operatorname{tr}(\boldsymbol{\Sigma}_1) + \operatorname{tr}(\boldsymbol{\Sigma}_2)$, and $\Delta_{\Sigma, 12} = |\operatorname{tr}(\boldsymbol{\Sigma}_1) - \operatorname{tr}(\boldsymbol{\Sigma}_2)|$. Additionally, let $L_i = \operatorname{Var}[\|\mathbf{x}_{ij} - \boldsymbol{\mu}_i\|^2]$ for $i \in \{1, 2\}$. We assume $K = 2$ to explore the asymptotic properties of k-means clustering.

We first consider kernel k-means with the linear kernel $k_L(\mathbf{y}_1, \mathbf{y}_2) = \mathbf{y}_1^\top \mathbf{y}_2$ which is equivalent to conventional k-means. The following assumptions were considered.

(A-i): $\operatorname{tr}(\boldsymbol{\Sigma}_i^2)/\Delta_{12}^2 \rightarrow 0$, $i \in \{1, 2\}$, as $d \rightarrow \infty$,

(A-ii): $L_i/\Delta_{12}^2 \rightarrow 0$, $i \in \{1, 2\}$, as $d \rightarrow \infty$.

The following results are obtained for conventional k-means:

Theorem 2.1. *Suppose that the initial observation is $\mathbf{c}_{0i} \in \mathbf{X}_i$ for $i \in \{1, 2\}$. Assume (A-i) and (A-ii) when N is fixed as $d \rightarrow \infty$. When*

$$\limsup_{d \rightarrow \infty} \frac{\Delta_{\Sigma,12}}{\Delta_{12}} < 1 \quad (2)$$

is satisfied, for Algorithm 1 with linear kernel and $K = 2$,

$$P(\{\widehat{\mathbf{C}}_1, \widehat{\mathbf{C}}_2\} = \{\mathbf{X}_1, \mathbf{X}_2\}) \rightarrow 1 \quad (3)$$

holds true as $d \rightarrow \infty$ when N is fixed.

For general kernel k-means, we introduce the following assumption.

(K-i): There exists $\kappa_{ii}, \kappa_{ii}^*$, and $\kappa_{ii'}$ for $i, i' \in \{1, 2\} (i \neq i')$ such that

$$\begin{aligned} k(\mathbf{x}_{ij}, \mathbf{x}_{ij'}) &= \kappa_{ii} + o_p(\Delta_{\kappa,12}) \text{ for } i \in \{1, 2\}, j, j' \in \{1, \dots, n_i\}, j \neq j', \\ k(\mathbf{x}_{ij}, \mathbf{x}_{ij}) &= \kappa_{ii}^* + o_p(\Delta_{\kappa,12}) \text{ for } i \in \{1, 2\}, j \in \{1, \dots, n_i\}, \\ k(\mathbf{x}_{1j}, \mathbf{x}_{2j'}) &= \kappa_{12} + o_p(\Delta_{\kappa,12}) \text{ for } j \in \{1, \dots, n_1\}, j' \in \{1, \dots, n_2\}, \end{aligned}$$

as $d \rightarrow \infty$ when N is fixed where $\kappa_{ii}^* > \kappa_{ii}$ for $i \in \{1, 2\}$ and $\Delta_{\kappa,12} = \kappa_{11} + \kappa_{22} - 2\kappa_{12}$.

Let $\eta_{\kappa,i} = \kappa_{ii}^* - \kappa_{ii}$ for $i, j \in \{1, 2\}$ and $\Delta_{\kappa,\eta,12} = |\eta_{\kappa,1} - \eta_{\kappa,2}|$. The following theorem gives a general consistency condition for kernel k-means.

Theorem 2.2. *Suppose initial observation, $\mathbf{c}_i \in \mathbf{X}_i$ for $i \in \{1, 2\}$. Assume (K-i). When*

$$\limsup_{d \rightarrow \infty} \frac{\Delta_{\kappa,\eta,12}}{\Delta_{\kappa,12}} < 1 \quad (4)$$

is satisfied, (3) holds true as $d \rightarrow \infty$ when N is fixed.

Let $\Omega = \gamma \Delta_{G,12} / \psi$, $\Delta_{G,12} = \kappa_{G11} + \kappa_{G22} - 2\kappa_{G12}$, $\psi = \kappa_{G11}$, $\kappa_{G12} = \exp(-\Delta_{*12}/\gamma)$, $\kappa_{Gii} = \exp(-2\text{tr}(\Sigma_i)/\gamma)$, $\kappa_{Gii}^* = 1$, for $i \in \{1, 2\}$. We introduce the following assumptions:

(B-i): $\limsup_{d \rightarrow \infty} \Delta_{12}/\gamma < \infty$, $\limsup_{d \rightarrow \infty} \Delta_{\Sigma,12}/\gamma < \infty$ as $d \rightarrow \infty$;

(B-ii): $\text{tr}(\Sigma_i^2)/\Omega^2 \rightarrow 0$, $i \in \{1, 2\}$, as $d \rightarrow \infty$;

(B-iii): $L_i/\Omega^2 \rightarrow 0$, $i \in \{1, 2\}$, as $d \rightarrow \infty$.

Under these assumptions, the following proposition establishes (K-i) for the Gaussian kernel, $k_G(\mathbf{y}_1, \mathbf{y}_2) = \exp(-\|\mathbf{y}_1 - \mathbf{y}_2\|^2/\gamma)$, where $\gamma > 0$ is hyperparameter.

Proposition 2.1. *Under (B-i), (B-ii), and (B-iii), (K-i) holds as follows:*

$$k_G(\mathbf{x}_{ij}, \mathbf{x}_{ij'}) = \kappa_{Gii} + o_p(\Delta_{G,12}), \quad i \in \{1, 2\}, j, j' \in \{1, \dots, n_i\}, j \neq j',$$

$$k_G(\mathbf{x}_{ij}, \mathbf{x}_{ij}) = \kappa_{Gii}^*, i \in \{1, 2\}, j \in \{1, \dots, n_i\},$$

$$k_G(\mathbf{x}_{1j}, \mathbf{x}_{2j'}) = \kappa_{G12} + o_p(\Delta_{G,12}), j \in \{1, \dots, n_1\}, j' \in \{1, \dots, n_2\}.$$

Let $\eta_{G,i} = \kappa_{Gii}^* - \kappa_{Gii}$ $i \in \{1, 2\}$ and $\Delta_{G,\eta,12} = |\eta_{G,1} - \eta_{G,2}|$ Note that

$$\frac{\Delta_{G,12}}{\psi} = (1 - \delta_{\Sigma,12})^2 + 2(1 - \delta_{12})\delta_{\Sigma,12} \quad \text{and} \quad \frac{\Delta_{G,\eta,12}}{\psi} = 1 - \delta_{\Sigma,12}^2$$

where $\delta_{12} = \exp(-\Delta_{12}/\gamma)$ and $\delta_{\Sigma,12} = \exp(-\Delta_{\Sigma,12}/\gamma)$. Consequently, Theorem 2.2 yields the following corollary.

Corollary 2.1. *Suppose initial observation, $\mathbf{c}_i \in \mathbf{X}_i$ for $i \in \{1, 2\}$. Assume (B-i), (B-ii), and (B-iii). When*

$$\limsup_{d \rightarrow \infty} \frac{1 - \delta_{\Sigma,12}^2}{(1 - \delta_{\Sigma,12})^2 + 2(1 - \delta_{12})\delta_{\Sigma,12}} < 1 \quad (5)$$

is satisfied, (3) holds true as $d \rightarrow \infty$ when N is fixed.

We now compare the asymptotic consistency condition of Gaussian kernel k-means with that of conventional k-means.

Remark 1. Condition (5) holds for any $\gamma > 0$ if and only if the corresponding consistency condition for conventional k-means, (2), holds.

Thus, although Gaussian kernel k-means introduces a nonlinear transformation of the observations, its asymptotic consistency condition is equivalent to that of conventional k-means under the considered HDLSS setting. In particular, an appropriate choice of the Gaussian kernel parameter γ does not, by itself, eliminate the fundamental high-dimensional difficulty associated with the heterogeneity of the covariance structures.

The above result provides an important limitation of kernelization in HDLSS clustering. Although kernel methods are useful for incorporating nonlinear relationships, the use of a Gaussian kernel alone does not overcome the asymptotic bias arising in high-dimensional clustering. This observation naturally raises the following question: Can Gaussian kernel k-means be improved by incorporating procedures that directly address the high-dimensional bias? In this talk, we investigate two approaches to this problem. First, we consider multiple initial observations in Gaussian kernel k-means to reduce the effect of the initialization. Second, we develop a bias-corrected version of Gaussian kernel k-means by explicitly removing the asymptotic bias in the kernel-based clustering criterion.

Acknowledgments

This research of the first author was partially supported by Grant-in-Aid for Early-Career Scientists, JSPS, under Contract Number 24K20748. This research of the third author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Research (Exploratory), JSPS, under Contract Numbers 25H01107 and 26K23798.

References

- [1] Arias-Castro, E., Pu, X. (2017) A simple approach to sparse clustering. *Computational Statistics & Data Analysis*, 105, 217–228.
- [2] Chakraborty, S., Das, S. (2022). Detecting meaningful clusters from high-dimensional data: a strongly consistent sparse center-based clustering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 2894–2908.
- [3] Chang, X., Wang, Y., Li, R., Xu, Z. (2018). Sparse k -means with l_∞/l_0 penalty for high-dimensional data clustering. *Statistica Sinica*, 28, 1265–1284.
- [4] Egashira, K., Yata, K., Aoshima, M. (2026). Asymptotic properties of k -means and its bias correction for high-dimension, low-sample-size data. *Journal of Multivariate Analysis*, 216, 105675.
- [5] Krishnapuram, R., Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1, 98–110.
- [6] Sun, W., Wang, J., Fang, Y. (2012). Regularized k -means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6, 148–167.
- [7] Yang, M.-S., Benjamin, J.B. (2023). Sparse possibilistic c -means clustering with lasso. *Pattern Recognition*, 138, 109348.