IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 February 2023, accepted 10 March 2023, date of publication 14 March 2023, date of current version 17 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257163

RESEARCH ARTICLE

Finding the Number of Clusters Using a Small Training Sequence

DONG SIK KIM, (Senior Member, IEEE)
Department of Electronics Engineering, Hankuk University of Foreign Studies, Yongin 17579, South Korea
e-mail: dskim@hufs.ac.kr

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MISP) under Grant 2020R1A2C1009895, and in part by the Hankuk University of Foreign Studies Research Fund of 2023.

ABSTRACT In clustering the training sequence (TS), K-means algorithm tries to find empirically optimal representative vectors that achieve the empirical minimum to inductively design optimal representative vectors yielding the true optimum for the underlying distribution. In this paper, the convergence rates on the clustering errors are first observed as functions of $\beta^{-\alpha}$, where β is the training ratio that relates the training sequence size to the number of representative vectors, and α is a non-negative constant. From the convergence rates, we can observe the training performance for a finite TS size. If the TS size is relatively small, errors occur in finding the number of clusters. In order to reduce the errors from small TS sizes, a compensation constant $(1-\beta^{-\alpha})^{-1}$ for the empirical errors is devised based on the rate analyses and a novel algorithm for finding the number of clusters is proposed. The compensation constant can be applied to other clustering applications especially when the TS size is relatively small.

INDEX TERMS Clustering, K-means algorithm, number of clusters, small training sequence, training ratio, β-compensation.

I. INTRODUCTION

The clustering of samples is a method of grouping or segmenting samples into subsets or clusters to efficiently represent the samples [1], [2], [3] and improve deep learning performances. Clustering samples can be a scheme for self-organizing or unsupervised learning [4], [5]. We assume that a sequence of samples, which is called the training sequence (TS), is realized from a random vector with the underlying distribution. The K-means algorithm can efficiently conduct clustering by iteratively decreasing an empirical error from TS [6]. The algorithm divides the TS into a finite number of disjoint clusters, of which centroids are the representative vectors, based on the nearest neighbor search. Here, we call the finite set of these vectors the codebook. The K-means algorithm can asymptotically design optimal codebooks as the TS size increases in an inductive approach [7], [8]. Hence, the K-means algorithm can be an efficient approach to reduce both linear and nonlinear correlations [9], and combined with the conventional linear decorrelation methods, such as the principle component analysis, Karhunen-Loeve transform, and discrete cosine transform for autoregressive signals [10].

Based on the consistency property of a sequence of trained codebooks as the TS size increases [8], [11], the K-means algorithm tries to find empirically optimal codebooks, which achieve the empirical minimum for the given TS, to inductively design an optimal codebook for the underlying distribution [12]. However, the K-means algorithm often yields locally optimal codebooks depending on initial guesses. Locally optimal codebooks do not guarantee a convergence to an optimal codebook even though the TS size increases. Hence, it is necessary to devise an algorithm, which can achieve the global optimum. Vaisey and Gersho [13] utilize the simulated annealing approach to alleviate the local minimum problem. In inductively training codebooks, the training ratio β, which is defined by the ratio of the TS size to the codebook size, is a more important parameter rather than the TS size itself [9]. Furthermore, depending on search structures, β can be different [14], [15], [16].

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

In this paper, in order to formulate the convergence rates of the errors yielded by the trained codebooks as a function of β, previous research regarding the effects of a finite TS size on the clustering performance is first surveyed. For this paper, the main contributions are as follows. (1) It is shown that the convergence rate in clustering TS is a function of $\beta^{-\alpha}$, where $1/2 \leq \alpha \leq 1$. (2) In testing the trained codebook, a similar rate is also observed and a convergence rate in the minimax sense is derived; the trained codebook shows a rate of β^{-1} in a minimax sense. (3) Based on the convergence rate analyses with β, a novel algorithm to find the number of clusters is proposed, where a compensation for the K-means algorithm is considered especially for small training ratios of β. Note that, as the TS size becomes relatively small, more errors occur in finding the correct number of clusters.

This paper is organized in the following way. In Section II, several definitions in training codebooks are shown. The convergence rates of β in training codebooks are shown in Section III and the rates in testing the trained codebooks are shown in Section IV. In Section V, a novel algorithm to find the number of clusters is proposed. The conclusion is given in the last section.

II. PRELIMINARY

In early works on statistical learning, the Vapnik-Chervonenkis (VC) dimension [17] has been introduced for a set of indicator functions, which can be employed in the pattern recognitions [2]. Cohn et al. [18] have framed the clustering as a classification problem and proposed a bound by adopting the VC dimension. In this section, several definitions on trained codebooks are introduced and the convergence rates with β are observed by surveying the previous research.

Let F denote the underlying distribution and $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^k$ as a dissimilarity measure. The codebook design problem for F is to find a set C that minimizes a distortion defined by a mean square error as

$$D(C) := \int \min_{\boldsymbol{y}\in C} \|\boldsymbol{x}-\boldsymbol{y}\|^2 dF(\boldsymbol{x}) \tag{1}$$

over all possible choices of $C \subset \mathbb{R}^k$, in which the size of C is less than or equal to a positive integer n. We call C the codebook and its elements the representative vectors. Let C^* denote an optimal codebook if $D(C^*) = \inf_C D(C)$. We call $D(C^*)$ the optimum and simply denote the optimum as D^*, which is assumed $D^* \neq 0$. From the source coding theorem, D^* converges to the Shannon lower bound as the vector dimension k increases [9], [19]. To find C^*, Max [20] and Lloyd [21] proposed algorithms for a given distribution F.

Assume that $\boldsymbol{X}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_m$ are independent, and identically distributed random vectors taking values in $\mathbb{R}^k$ with distribution F. Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m$ denote a finite TS and a positive integer m denote the TS size. For a codebook C, we define the empirical distortion, which is an empirical mean square error, as

$$D_m(C) := \frac{1}{m}\sum_{\ell=1}^{m} \min_{\boldsymbol{y}\in C} \|\boldsymbol{X}_\ell - \boldsymbol{y}\|^2, \tag{2}$$

where we suppose that $E\{\|\boldsymbol{X}\|^2\} < \infty$. Note that D_m is a random variable defined on the underlying sample space. There exists a codebook that achieves $\inf_C D_m(C)$ for a given TS [12]. Let C_m^* denote such a codebook that minimizes the empirical distortion D_m. We call $D_m(C_m^*)$ ($= \min_C D_m(C)$) the empirical minimum. In order to find C^* for an unknown F, an inductive method that minimizes D_m is usually considered in the traditional codebook design approaches; because, for a sequence of C_m^*, $D(C_m^*)$ and $D_m(C_m^*)$ converge to D^* almost surely (a.s.) under several conditions [8], [11]. To find an empirical optimum C_m^*, Linder et al. [7] proposed an iterative algorithm by using TS [9, p. 366]. This algorithm is equivalent to a clustering algorithm, which is called the isodata or K-means algorithm in the area of pattern recognition [22, p. 482]. K-means algorithms usually search the cluster centers of C_m^* by minimizing $D_m(C)$.

Because the empirical minimum $D_m(C_m^*)$ is readily available in training the codebook, $D_m(C_m^*)$ can be employed as a performance measure for the trained codebooks, rather than the mean distortion $D(C_m^*)$ [23], [24]. However, because $D_m(C_m^*)$ is a biased estimate of $D(C_m^*)$, using a relatively small size of a separate test or validating sequence, we can efficiently estimate $D(C)$ to evaluate the performance of C [25].

Let the training ratio β be defined as $\beta := m/n$, which is a normalized TS size by the codebook size [9, p. 364]. The distortion difference between the optimum and that of the trained codebook is dependent on the ratio β [16], [24], [26]. Hence, we may estimate the achievable performance of trained codebooks by observing β. In the special case when $n = 1$ and $\beta \geq 1$, we obtain explicit relations: $D^* - E\{D_m(C_m^*)\} = c\beta^{-1}$ and $E\{D(C_m^*)\} - D^* = c\beta^{-1}$, where $c = D^*$ is equal to the trace of the covariance matrix of $\boldsymbol{X}$ [24].

For a more general case, when $n \geq 2$, any explicit derivation regarding $D_m(C_m^*)$ or $D(C_m^*)$ is not shown in the literature except for several special cases; $D_m(C_m^*) = 0$ if $\beta \leq 1$. Instead of providing such exact values, several bounds in some senses have been derived in order to observe the rate as a function of m or β. By constructing a discrete distribution, Linder [23] derived a lower bound on $D^* - E\{D_m(C_m^*)\}$, as a function of $m^{-1/2}$. Earlier results also provide upper bounds of the same order. From the bounds, Linder [23] showed that $D^* - E\{D_m(C_m^*)\}$ has a rate of $\beta^{-1/2}$ by introducing worst case bounds. Based on a theorem of [27], Chou [28] studied the mean distortion of C_m^*, and showed that $D(C_m^*) - D^*$ has the rate m^{-1} in a sense of distribution. Linder et al. [29] proposed an upper bound on $E\{D(C_m^*)\} - D^*$, and Bartlett et al. [30] suggested the rate $\beta^{-1/2}$ in the minimax sense. In comparison to the worst case bound of Linder [23], Kim and Bell [24] derived a pointwise lower bound on $D^* - E\{D_m(C_m^*)\}$ as

a function of any distribution F. They also derived a lower bound, which has the rate m^{-1}.

In experimental research, we find several observations regarding the rate, especially for image signals. Cosman et al. [31] numerically investigated the performance of the trained codebooks and suggested an algebraic decay of the form $m^{-\alpha}$ for a positive α. Cohn et al. [18] also observed the trained codebook performance based on a discrete source model for images. Collura and Tremain [26] numerically observed the appropriate values of β for designing a full-search codebook based on spectral data and showed that constrained-search structures have a better training performance than the full-search case. Note that β can be different depending on the search structures. Kim [16] observed such training ratios for different search structures, and compared their performances in terms of training. Kim and Bell [24] also showed several numerical results for the uniform, Gaussian, and Laplacian densities of F with fitted curves of the form $\beta^{-\alpha}$. Based on a constrained-search structure, the product quantization [14], [15] can provide good characteristics both in terms of training performance and search efficiency [32].

III. RATES ON THE EMPIRICAL MINIMUM

In this section, the convergence rates of the empirical minimum $D_m(C_m^*)$ of (2) for the empirical optimum C_m^* is observed.

A. EMPIRICAL MINIMUM OF RATE $\beta^{-\alpha}$

Let us consider n points $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$ as a codebook $\{\boldsymbol{y}_i\}$, and the corresponding partition $\{S_i\}$ of $\mathbb{R}^k$. The corresponding element of the region S_i is y_i, and the codebook size is n. Note that the partition is a finite, disjoint class $\{S_i\}$ whose union is $\mathbb{R}^k$ (or includes the support of a density function of F). Let P_i denote the probability that X belongs to S_i, i.e., $P_i := \int_{S_i} dF(\boldsymbol{x})$, and $\mathcal{I}$ be an index set $\mathcal{I} = \{i : P_i \neq 0, i = 1, \cdots, n\}$. Define the ith partial distortion δ_i as $\delta_i := \int_{S_i} \|x - y_i\|^2 dF(\boldsymbol{x})$, for $i \in \mathcal{I}$. The summation of the partial distortions, $\sum_{i\in\mathcal{I}} \delta_i$, is equal to the mean distortion of the codebook $\{\boldsymbol{y}_i\}$ and the partition $\{S_i\}$ for F. Here, we assume that $\sum_{i\in\mathcal{I}} \delta_i \neq 0$. For the underlying codebook $\{\boldsymbol{y}_i\}$ and partition $\{S_i\}$, let us consider a random vector $\boldsymbol{Y}_i^o$ that is defined as

$$\boldsymbol{Y}_i^o := \begin{cases} (0, \cdots, 0), & \text{if } m_i = 0 \\ \dfrac{1}{m_i} \displaystyle\sum_{\ell=1}^{m} I_{S_i}(X_\ell) X_\ell, & \text{otherwise,} \end{cases} \tag{3}$$

where m_i is a random variable defined as $m_i := \sum_{\ell=1}^m I_{S_i}(X_\ell)$, for each S_i. Here, $I_S(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in S$, and $I_S(\boldsymbol{x}) = 0$ otherwise. Let C_m^o denote the set of $\boldsymbol{Y}_i^o$ as $C_m^o := \{\boldsymbol{Y}_i^o\}$. For the set C_m^o, define an empirical distortion Λ_m as $\Lambda_m := m^{-1} \sum_{i\in\mathcal{I}} \sum_{\ell=1}^m I_{S_i}(X_\ell) \|X_\ell - \boldsymbol{Y}_i^o\|^2$. For any distribution F, and the underlying $\{\boldsymbol{y}_i\}$ and $\{S_i\}$,

$$E\{\Lambda_m\} \leq \sum_{i\in\mathcal{I}} \delta_i \left[1 - \frac{1 - (1 - P_i)^m}{mP_i}\right] \tag{4}$$

holds [24], where $k, m, n \geq 1$. If each $\boldsymbol{y}_i$ is the centroid of S_i, i.e., $\boldsymbol{y}_i = \int_{S_i} \boldsymbol{x} dF(\boldsymbol{x}) / \int_{S_i} dF(\boldsymbol{x})$, then the equality in (4) holds and a relationship $D(C) - E\{\Lambda_m\} = c_0 \beta^{-1}$ is obtained, where a positive c_0 is defined as $c_0 := \sum_{i\in\mathcal{I}} \delta_i [1 - (1 - P_i)^m] / nP_i$. Note that $c_0 \uparrow c_\infty := \sum_{i\in\mathcal{I}} \delta_i / nP_i$ as m increases. Because $\lim_{m\to\infty} |c_0 - c_\infty| / \beta = 0$, the sequence $(c_0)_m$ converges to c_∞ at a faster rate than that of $\beta^{-1} \to 0$.

Assume that $\{\boldsymbol{y}_i\}$ and $\{S_i\}$ in distortion Λ_m are equal to an optimal codebook $C^* := \{\boldsymbol{y}_i^*\}$ and the corresponding Voronoi partition $\{S_i^*\}$, respectively. Then, from $D_m(C_m^*) \leq D_m(C_m^o) \leq \Lambda_m$ and (4),

$$D^* - E\{D_m(C_m^*)\} \geq c_0^* \beta^{-1} \tag{5}$$

holds, where $k, m, n \geq 1$. In (5), c_0^* is obtained from c_0 for $\{S_i^*\}$. For a fixed $n \geq 1$ and F, suppose that

$$D^* - E\{D_m(C_m^*)\} = c_1 \beta^{-\alpha_1}, \tag{6}$$

where $c_1, \alpha_1 > 0$, and the sequence $(c_1)_m$ is bounded. From (5), it is clear that $\alpha_1 \leq 1$. In (6), c_1 can be a function of C^*, k, m, n, and F, and can contribute to making the difference of (6) zero as m increases. For the $n = 1$ case, $c_1 = D^*$ and $\alpha_1 = 1$ hold. However, for the case of $n \geq 2$, we can only guess the constant c_1 from c_0^* in (5). From an upper bound on (6), which is derived by Linder [23], we notice that the minimum of α_1 is $1/2$ for the distributions supported by a given bounded region. Here, we add a condition that $\liminf_m c_1 > 0$ to the constant c_1. Hence, for the empirical minimum case of (6), we can obtain a range of α_1 as

$$\text{Empirical Minimum: } \frac{1}{2} \leq \alpha_1 \leq 1, \tag{7}$$

for the distributions supported by a given bounded region.

The fastest rate β^{-1} of (7) is usually achieved when $n = 1$, independently of the distribution type and the vector dimension. Note that fast rates are attractive because we can obtain a good codebook using a relatively small TS [26]. However, for the case of $n \geq 2$, the rate can be different depending on the distribution as illustrated in the numerical results of [24]. In fact, the codebook size n also affects the rate.

Linder [23] constructed a discrete distribution as a function of the codebook size, and showed the minimum rate $\beta^{-1/2}$. In the worst case of distributions, Linder [23] also derived both upper and lower bounds, and showed that the difference is proportional to $\beta^{-1/2}$. From (5), we can also derive a worst case bound, but as a function of β^{-1},

$$\sup_F \left[D^* - E\{D_m(C_m^*)\}\right] \geq c_0^* \beta^{-1}, \tag{8}$$

where the supremum is taken for all distributions over a ball in $\mathbb{R}^k$. As Kim and Bell [24] showed, even though the rate is faster than $\beta^{-1/2}$, a worst case bound obtained from (8) can be better than that of Linder for practically small training ratios.

IV. RATES ON THE MEAN DISTORTION

In this section, the convergence rates of the mean distortion $D(C_m^*)$ of (1) for the empirical optimum C_m^* is first observed. A lower bound in a minimax sense is next derived.

A. MEAN DISTORTION OF RATE $\beta^{-\alpha}$

In a manner similar to (6), we assume an relationship as

$$E\{D(C_m^*)\} - D^* = c_2\beta^{-\alpha_2}, \tag{9}$$

where $c_2, \alpha_2 > 0$, and the sequence $(c_2)_m$ is bounded and $\liminf_m c_2 > 0$. Linder et al. [29] proposed an upper bound on $E\{D(C_m^*)\} - D^*$ as a function of $(m/\log m)^{-1/2}$, for any distribution F. Bartlett et al. [30] sharpened the upper bound as a function of $m^{-1/2}$. Hence, the minimum of α_2 is given by $1/2$ for a fixed codebook size n.

Bartlett et al. [30] constructed a distribution, and derived a lower bound of the constructed distribution as a function of $\beta^{-1/2}$. However, we can obtain a lower bound based on the result of Chou [28] for a more general class of distributions [27]. If C^* is unique for the distribution, then from [28], $m[D(C_m^*) - D^*] \to w$ in distribution, where w is the sum of squares of normal random variables with zero-mean and a covariance matrix. Hence, we have $\liminf_{m\to\infty} m[E\{D(C_m^*)\} - D^*] \geq E\{w\}$ [33, Theorem 25.11]. Because $E\{D(C_m^*)\} - D^* > 0$ if $D^* \neq 0$, and $E\{w\} > 0$, there is a positive constant c_2' such that $\beta[E\{D(C_m^*)\} - D^*] \geq c_2'$, for a fixed n. Therefore, we can obtain a lower bound as

$$E\{D(C_m^*)\} - D^* \geq c_2'\beta^{-1}. \tag{10}$$

If we confine the input distributions within those of [27], then, from (10), the maximum rate is given by β^{-1}. In other words, for a class of distributions, we can obtain a range of α_2 for (9) as

$$\text{Mean Distortion: } \frac{1}{2} \leq \alpha_2 \leq 1. \tag{11}$$

It is clear that the distribution constructed by Bartlett et al. [30] has the rate $\beta^{-1/2}$. The rate β^{-1} is always achieved when $n = 1$. Therefore from (7) and (11), we notice that both distortion differences have a range of rates from $\beta^{-1/2}$ to β^{-1}. In the following section, it will be shown that β^{-1} can be a rate for the mean distortion case in the minimax sense.

B. RATE β^{-1} IN THE MINIMAX SENSE

In order to investigate the rate of the distortion difference $E\{D(C_m^*)\} - D^*$, Bartlett et al. [30] introduced a notion of the minimax expected distortion redundancy, which expresses the minimal worst case excess distortion that an empirical codebook can have. The main result of [30] is that the difference $E\{D(C_m^*)\} - D^*$ is not a function of β^{-1} in the minimax sense, contrary to the conjecture of Chou's rate β^{-1} [28]. They also proposed the rate $\beta^{-1/2}$ for the difference in the minimax sense, instead of β^{-1}. In this section, however, it is shown that the difference $E\{D(C_m^*)\} - D^*$ can have the rate β^{-1} even in the minimax sense.

Deriving the rate β^{-1} in the minimax sense is performed by obtaining an upper bound of the mean distortion of C_m^o, which is introduced in (3). In a manner similar to Λ_m, define a mean distortion Λ as

$$\Lambda := \sum_{i\in\mathcal{I}} \int_{S_i} \|\boldsymbol{x} - \boldsymbol{Y}_i^o\|^2 dF(\boldsymbol{x}), \tag{12}$$

for the underlying codebook $\{\boldsymbol{y}_i\}$ and partition $\{S_i\}$. Note that we have a relationship: $D^* \leq D(C_m^o) \leq \Lambda$. However, any relationship between the distortions $D(C_m^*)$ and $D(C_m^o)$ is not known. Hence, it is difficult to derive a pointwise bound on the mean distortion $D(C_m^*)$ as a function of C_m^o.

Assume that each $\boldsymbol{y}_i$ is the centroid of S_i for any given F and $\{S_i\}$. Then

$$E\{\Lambda\} - D(C) = c_3\beta^{-1} \tag{13}$$

holds, where $k, m, n \geq 1$. Here the positive c_3 is given by

$$c_3 := \beta\left[\sum_{i\in\mathcal{I}} \delta_i E\{\xi_i\} + \sum_{i\in\mathcal{I}} P_i\|\boldsymbol{y}_i\|^2(1-P_i)^m\right], \tag{14}$$

where the random variable ξ_i is defined as $\xi_i := 1/m_i$ if $m_i \neq 0$, and $\xi_i = 0$ otherwise, and c_3 converges to c_∞ as $m \to \infty$ at a rate, which is equal to or can be faster than that of $\beta^{-1} \to 0$. The proof of (13) is shown in Appendix.

Assume that $\{\boldsymbol{y}_i\}$ and $\{S_i\}$ in error Λ are equal to an optimal codebook $C^* := \{\boldsymbol{y}_i^*\}$ and the corresponding Voronoi partition $\{S_i^*\}$, respectively. We then obtain the following relationship:

$$E\{\Lambda\} - D^* = c_3^*\beta^{-1}, \tag{15}$$

where $k, m, n \geq 1$. Here, the positive constant c_3^* can be obtained from (14). From (15) and $E\{D(C_m^o)\} \leq E\{\Lambda\}$, $E\{D(C_m^o)\} - D^* \leq c_3^*\beta^{-1}$ holds for any F. Hence, we have an upper bound in the minimax sense as

$$\inf_{C_m}\sup_{F}\left[E\{D(C_m)\} - D^*\right] \leq c_3^*\beta^{-1}, \tag{16}$$

where the infimum is taken over all k-dimensional n-point codebooks trained on m samples, and the supremum is taken for all distributions over a ball in $\mathbb{R}^k$. Bartlett et al. [30] derived a lower bound of rate β^{-1}. Therefore, from (16), in the minimax sense, the difference $E\{D(C_m^*)\} - D^*$ has a rate of β^{-1}.

V. NUMBER OF CLUSTERS

In this section, a distortion compensation in clustering relatively small TS is first introduced based on the convergence rate analysis with the training ratio β. As an example of the compensation, a novel algorithm to find the number of clusters especially for relatively small TS sizes is next proposed.

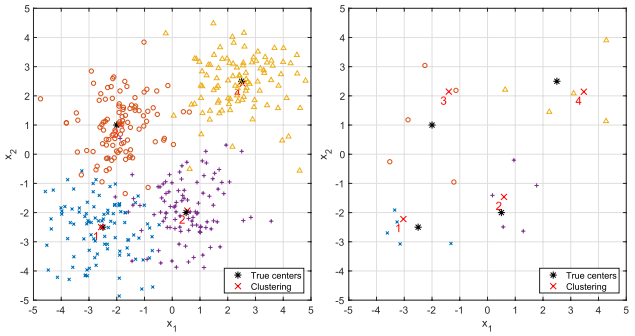

FIGURE 1. Gaussian mixture distribution with 4 clusters and variance of 1 for clustering examples. In the K-means algorithm, a split method for the initial guess was employed ($k = 2$ and $n = 4$) [9]. (a) TS of 100 samples per center ($\beta = 100$). (b) TS of 5 samples per center ($\beta = 5$).

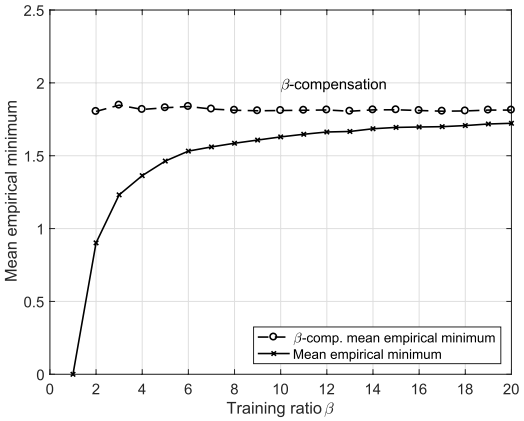

FIGURE 2. Mean of the empirical minima $E\{D_m(C_m^*)\}$ with respect to the training ratio β and the β-compensation for the TS of Fig 1 ($n = 4$ and $\alpha = 1$). The mean distances are reduced to zero as the training ratio β approaches 1; the clustering region is shrinking.

A. β-COMPENSATIONS FOR SMALL TRAINING RATIOS

In order to find better clusters for relatively small TS, a data-driven dissimilarity measure can be used [34]. Instead of such a special measure, a simple compensation of the Euclidean norm is introduced to deal with the problem caused by small training ratios based on the convergence rate analyses of the previous sections.

We consider an empirical optimum assumption that the empirically optimal codebook C_m^* is obtained from the K-means clustering for a TS. If the TS size is relatively small, then the empirical minimum $D_m(C_m^*)$ can have considerable biases due to small β as mentioned in (6). It seems that the distances between the centers and samples are reduced as the training ratio β decreases. If $\beta \leq 1$, then all distances eventually become zero as $D_m(C_m^*) = 0$. In other words, the clustering region is shrinking as β decreases. Hence, it is required to compensate the empirical distortions for the finite TS size. We call this compensation with the training ratio β the β-compensation.

Based on the relationship of (6), assume that the empirical minimum satisfies the following approximation: $D_m(C_m^*)\psi(\beta) \approx D^*$, where ψ denotes the β-compensation constant defined as

$$\beta\text{-Compensation Constant: } \psi(\beta) := (1 - \beta^{-\alpha})^{-1}, \quad (17)$$

for a finite training ratio β, where $1/2 \leq \alpha \leq 1$ from (7). As mentioned in (7), the coefficient for $\beta^{-\alpha}$ is dependent on k and F. However, to simplify the compensation, we set the coefficient $c_1 \approx D^*$ and thus can obtain ψ of (17). For the $n = 1$ case, we use $\alpha = 1$ for the β-compensation. If the clusters are well separated as the fixed bins in Λ_m [35], then $\alpha \approx 1$ from $E\{\Lambda_m\}$ in (4). For a strong compensation, we can set $\alpha = 1/2$.

A β-compensation to obtain D^* can be conducted from $D_m(C_m^*)\psi(\beta)$. We can also consider a β-compensated $\|\cdot\|^2\psi(\beta)$. This compensation is important especially when the training ratio β is too small to accurately estimate D^*. For simulation examples, two cases of TS generated from under a Gaussian mixture distribution with 4 clusters are illustrated in Fig. 1. The TS of Fig. 2(a) has an enough size to be clustered as $\beta = 100$. However, the TS of Fig. 1(b) has a relatively small TS size of $\beta = 5$. A β-compensation example for the Gaussian mixture distribution of Fig. 1 is shown in Fig. 2. We can observe that the distortions obtained from the β-compensation is nearly constant even for small training ratios.

B. FINDING THE NUMBER OF CLUSTERS

We now consider an algorithm to find the number of clusters based on the K-means algorithm for a given finite TS. Let us consider a distribution F having several separate clusters as Fig. 1. In order to find the separate clusters, we introduce the underlying assumption: if the number of clusters is equal to the codebook size n, then a normalized optimum $n^{2/k}D^*$ is lower than the other codebook size case. Here, the optimum should be normalized by multiplying $n^{2/k}$ to alleviate any influence from the codebook size [36]. Hence, for a set of candidates of the number of clusters, we can test $n^{2/k}D^*$ or $n^{2/k}D_m(C_m^*)$ if β is large enough. However, for small training ratios, the β-compensation as $n^{2/k}D_m(C_m^*)\psi(\beta)$ is required.

In order to find the number of clusters, a novel algorithm, which employs the β-compensation constant ψ, is proposed and summarized as follows.

1) NUMBER OF CLUSTERS WITH THE β-COMPENSATION

0) Consider a set $\mathcal{N}$ for the possible number of clusters.
1) Calculate the β-compensation constant $\psi(\beta)$ from (17) for $n \in \mathcal{N}$.
2) Conduct K-means clustering to obtain empirical minima $D_m(C_m^*)$, for $n \in \mathcal{N}$.
3) Select the codebook size such that

$$\min_{n\in\mathcal{N}} n^{2/k} D_m(C_m^*)\psi(\beta) \quad (18)$$

as the number of clusters in F.

Examples of the proposed algorithm are demonstrated in Fig. 3. If the training ratio is large enough as $\beta = 100$ ($n = 4$) of Fig. 1(a), then the compensated empirical minima

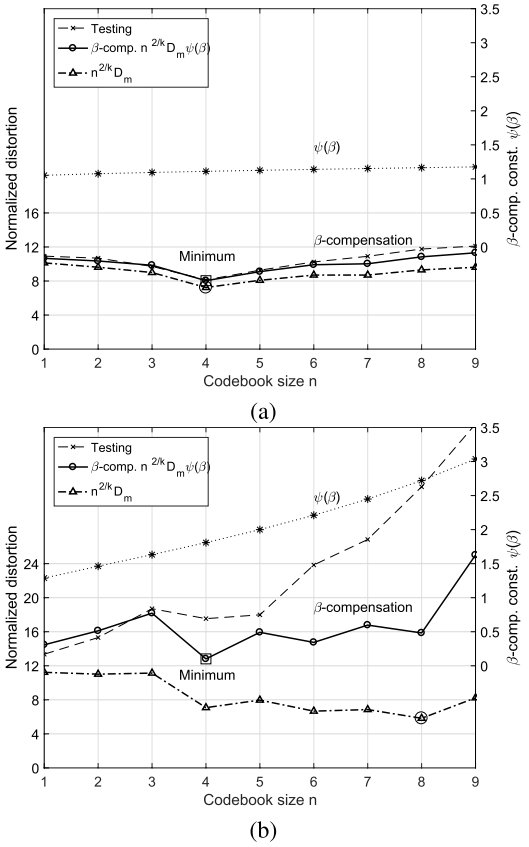

FIGURE 3. Example of the proposed algorithm to find the number of clusters for Fig 1 based on the normalized and β-compensated empirical minima. The β-compensation constants ψ of (17) are also depicted with $\alpha = 1/2$. The sizes that minimize the normalized and β-compensated empirical minima in the proposed algorithm of (18) were 4. (a) 100 samples per center of Fig 1(a) ($\beta = 100$ for $n = 4$). (b) 5 samples per center of Fig 1(b) ($\beta = 5$ for $n = 4$).

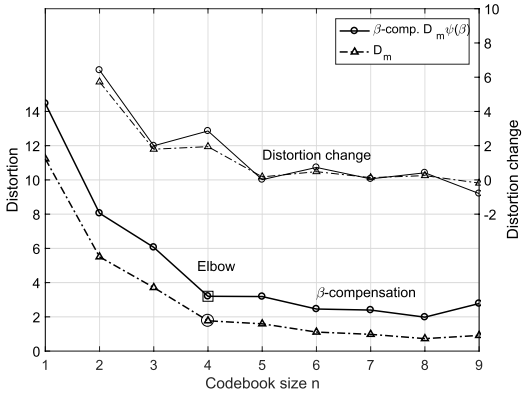

FIGURE 4. β-compensation of the elbow curve [37] for the TS of Fig 1(b) ($\beta = 5$ for $n = 4$).

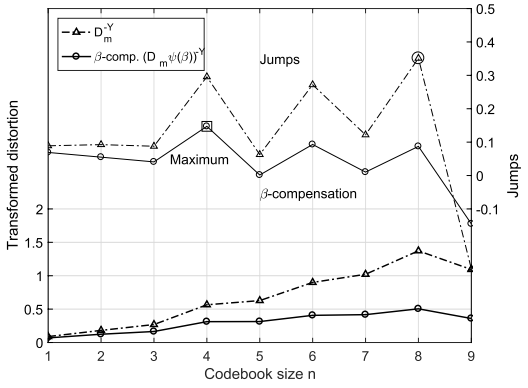

FIGURE 5. β-compensation of the information-theoretic method [40] for the TS of Fig 1(b) ($\beta = 5$ for $n = 4$), where $Y = k/2$.

$n^{2/k}D_m(C_m^*)\psi(\beta)$ (“β-comp. $n^{2/k}D_m\psi(\beta)$”) with respect to several n are very close to the mean distortions $n^{2/k}D(C_m^*)$ (“Testing”) as shown in Fig. 3(a). Because the minimum of the compensated distortions are achieved at $n = 4$ in Fig. 3(a), we can conclude that the number of clusters is 4 and the clustering result is also depicted in Fig. 1(a). It is clear that, if the TS size is large enough, we can find the correct number of clusters. On the other hand, if the TS size is relatively small as shown in Fig. 1(b), then simply checking the empirical minimum $n^{2/k}D_m(C_m^*)$ can lead to a wrong result as “$n^{2/k}D_m$” of Fig. 3(b). Note that the training ratio β of Fig. 3(b) varies from 20 to $20/9 \approx 2.22$, for $n = 1, \ldots, 9$. Hence, we can observe that the empirical minimum decreases and the test distortion increases as n increases in Fig. 3(b). For this relatively small TS size case, compensating the empirical minimum is important in finding the number of clusters as demonstrated in “β-comp. $n^{2/k}D_m\psi(\beta)$” of Fig. 3(b). Hence, the proposed algorithm can correctly find the number of clusters from the β-compensation especially for small training ratios. In the example of Fig. 3, $\alpha = 1/2$ was set for the β-compensation. However, for the $n = 4$ case, the convergence rate is more close to β^{-1}, i.e., $\alpha \approx 1$. Hence, setting $\alpha = 1$ for further possible sizes of n can be recommended.

In order to find the number of clusters, various methods have been developed [2], [22], [37], [38], [39], [40], [41], [42]. Among the methods, we can apply the β-compensation if the methods utilize the empirical distortion D_m or the Euclidean norm. Observing the elbow curve of D_m can provide an appropriate number of clusters as shown in Fig. 4 for the TS of Fig. 1(b) [43], [44]. From $n = 5$, the changes of D_m are not significant. Hence, we can guess the number of clusters is 4. Applying the β-compensation to D_m can help us find the elbow point more clearly especially when the TS size is relatively small. Sugar [40] proposed an information-theoretic method to find the clusters. The β-compensation can also be applied to the empirical minima and an example is shown in Fig. 5 for the TS of Fig. 1(b). Without the compensation, the distortion change curve yielded a wrong result of 8. However, applying the β-compensation provided the correct value in a similar manner to the proposed algorithm case.

In Fig. 6, the silhouette [38] and gap statistic [2], [39] methods are now discussed regarding the β-compensation. For the silhouette values, if the squared Euclidean norm

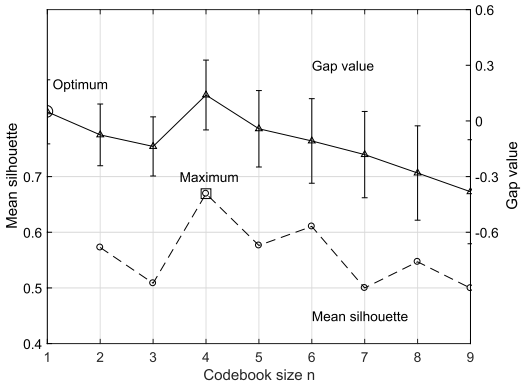

FIGURE 6. Silhouette means [38] and the gap statistic values [39] to find the number of clusters for the TS of Fig 1(b) ($\beta = 5$ for $n = 4$).

is used to calculate distances, the β-compensation constant can also be considered. However, the silhouette values are calculated based on a distance normalization and thus the β-compensation is not required intrinsically compared to the methods in [22]. As we can observe in Fig. 6, the silhouette method can find the correct number of clusters even for the relatively small TS of Fig. 1(b). However, the gap statistic method shows an incorrect optimal value of 1, where the β-correction cannot be applied. Note that, as β increases, the gap statistic method can yield the correct number of clusters.

VI. CONCLUSION

In this paper, the clustering performance of the K-means algorithm was analyzed in both theoretical and numerical observations under the empirical optimum assumption. If this assumption holds, then the analysis can be applied to other clustering algorithms. The clustering performance is

- dominantly dependent on the training ratio $\beta := m/n$,
- less sensitive to the vector dimension k, and
- different depending on the source distribution F.

The convergence rate has a form of $c\beta^{-\alpha}$, where α is dependent on F and $1/2 \le \alpha \le 1$. Positive constant c is a function of k and F, and $\inf_{\beta,k} c > 0$. When clustering TS while minimizing the empirical distortion, using the β-compensation constant $\psi(\beta) = (1-\beta^{-\alpha})^{-1}$ is proposed especially for the small TS or β case. By applying this β-compensation to a clustering algorithm, the number of clusters can be correctly found even for small values of β.

APPENDIX: PROOF OF (13)

In order to prove (13), the expectation of Λ is first derived. In $\Lambda = \sum_{i\in\mathcal{I}} \int_{S_i} \|\boldsymbol{x} - \boldsymbol{Y}_i^o\|^2 dF(\boldsymbol{x})$, replacing $(\boldsymbol{x} - \boldsymbol{Y}_i^o)$ by $(\boldsymbol{x} - \boldsymbol{y}_i) + (\boldsymbol{y}_i - \boldsymbol{Y}_i^o)$ yields

$$\Lambda = \sum_{i\in\mathcal{I}} \int_{S_i} \Big(\|\boldsymbol{x} - \boldsymbol{y}_i\|^2 + \|\boldsymbol{y}_i - \boldsymbol{Y}_i^o\|^2 + \boldsymbol{x}^T\boldsymbol{y}_i - \boldsymbol{x}^T\boldsymbol{Y}_i^o - \boldsymbol{y}_i^T\boldsymbol{y}_i + \boldsymbol{y}_i^T\boldsymbol{Y}_i^o \Big) dF(\boldsymbol{x}).$$

Because $\boldsymbol{y}_i = \int_{S_i} \boldsymbol{x} dF(\boldsymbol{x}) / \int_{S_i} dF(\boldsymbol{x})$ for all i from the assumption, Λ can be rewritten by

$$\Lambda = D(C) + \sum_{i\in\mathcal{I}} P_i \|\boldsymbol{Y}_i^o - \boldsymbol{y}_i\|^2. \tag{A1}$$

We now derive the mean of Λ, $E\{\Lambda\}$. In order to simplify the derivation, let $|\mathcal{I}| = n$. The proof in the case of $|\mathcal{I}| < n$ is similar. Let $B_\nu \subset \mathbb{R}^{km}$ be the m-fold Cartesian product of S_i's, i.e., $B_\nu = S_{i_{\nu,1}} \times \cdots \times S_{i_{\nu,m}}$, where $\nu = (i_{\nu,1} - 1) + (i_{\nu,2} - 1)n + \cdots + (i_{\nu,m} - 1)n^{m-1}$ and $i_{\nu,j} \in \{1, 2, \cdots, n\}$. Let us consider a product measure of order m as $F \times \cdots \times F$. Then the second term of the right-hand side of (A1) satisfies

$$\begin{aligned} &E\left\{ \sum_{i=1}^{n} P_i \|\boldsymbol{Y}_i^o - \boldsymbol{y}_i\|^2 \right\} \\ &= \sum_{\nu=0}^{n^m-1} \int_{B_\nu} \sum_{i=1}^{n} P_i \|\boldsymbol{y}_i^o - \boldsymbol{y}_i\|^2 dF(\boldsymbol{x}_1)\cdots dF(\boldsymbol{x}_m), \\ &=: \sum_{\nu=0}^{n^m-1} \sum_{i=1}^{n} P_i \psi_{i,\nu}, \end{aligned}$$

where $\psi_{i,\nu} := \int_{B_\nu} \|\boldsymbol{y}_i^o - \boldsymbol{y}_i\|^2 dF(\boldsymbol{x}_1)\cdots dF(\boldsymbol{x}_m)$, and $\boldsymbol{y}_i^o$ is the function of $\boldsymbol{x}_\ell$'s as in (3). Let $m_{i,\nu}$ denote the number of i in $\{i_{\nu,j}\}_{j=1}^m$. In the case when $m_{i,\nu} \neq 0$, by rearranging the parameters x_ℓ, $\psi_{i,\nu}$ can be expanded as

$$\psi_{i,\nu} = \prod_{\substack{j=1 \\ j\neq i}}^{n} P_j^{m_{j,\nu}} \int_{(S_i)^{m_{i,\nu}}} \|\boldsymbol{y}_i^o - \boldsymbol{y}_i\|^2 dF(\boldsymbol{x}_1)\cdots dF(\boldsymbol{x}_{m_{i,\nu}}), \tag{A2}$$

and $\psi_{i,\nu} = \prod_{j=1}^{n} P_j^{m_{j,\nu}} \|\boldsymbol{y}_i\|^2$ otherwise. Here, $\boldsymbol{y}_i^o := (1/m_{i,\nu}) \sum_{\ell=1}^{m_{i,\nu}} x_\ell$. Let $\xi_{i,\nu}$ be defined as $\xi_{i,\nu} := 1/m_{i,\nu}$ if $m_{i,\nu} \neq 0$, and 0 otherwise, and let $\gamma_{i,\nu}$ be defined as $\gamma_{i,\nu} :=$ 1 if $m_{i,\nu} \neq 0$, and 0 otherwise. $\psi_{i,\nu}$ can then be rewritten by

$$\begin{aligned} \psi_{i,\nu} = \frac{\xi_{i,\nu}}{P_i} \prod_{j=1}^{n} P_j^{m_{j,\nu}} \int_{S_i} \|\boldsymbol{x} - \boldsymbol{y}_i\|^2 dF(\boldsymbol{x}) \\ + (1 - \gamma_{i,\nu}) \prod_{j=1}^{n} P_j^{m_{j,\nu}} \|\boldsymbol{y}_i\|^2. \end{aligned} \tag{A3}$$

Because $\sum_\nu \prod_j P_j^{m_{j,\nu}} \gamma_{i,\nu} = 1 - (1 - P_i)^m$ [24], we obtain

$$E\{\Lambda\} = \sum_{i\in\mathcal{I}} \delta_i (1 + E\{\xi_i\}) + \sum_{i\in\mathcal{I}} P_i \|\boldsymbol{y}_i\|^2 m (1 - P_i)^m, \tag{A4}$$

where the random variable ξ_i is defined as $\xi_i := 1/m_i$ if $m_i \neq 0$, and 0 otherwise. From (A4), we can obtain a relationship: $E\{\Lambda\} - D(C) = c_3 \beta^{-1}$, where c_3 is given by

$$c_3 := \frac{1}{n} \sum_{i\in\mathcal{I}} \delta_i m E\{\xi_i\} + \frac{1}{n} \sum_{i\in\mathcal{I}} P_i \|\boldsymbol{y}_i\|^2 m (1 - P_i)^m. \tag{A5}$$

It is clear that the second term in the right-hand side of (A5) goes to 0 as $m \to \infty$ because $m(1 - P_i)^m \to 0$.

We now consider the term $mE\{\xi_i\}$ in (A5). A lower bound of ξ_i is given by $\gamma_i/(m_i+1) \le \xi_i$, where $\gamma_i = 1$ if $m_i \neq 0$, and

0 otherwise. Hence, $E\{\gamma_i/(m_i+1)\} \le E\{\xi_i\}$ holds. Because $\gamma_i/(m_i+1)$ has a multinomial distribution,

$$E\left\{\frac{\gamma_i}{m_i+1}\right\} = \sum_{\ell=1}^{m} \frac{1}{\ell+1}\frac{m!}{\ell!(m-\ell)!}(P_i)^{\ell}(1-P_i)^{m-\ell} = \frac{1-(1-P_i)^{m+1}-(m+1)P_i(1-P_i)^m}{(m+1)P_i}$$

holds. Hence, we have

$$\liminf_{m\to\infty} mE\{\xi_i\} \ge \frac{1}{P_i}. \tag{A6}$$

An upper bound on ξ_i is given by $\xi_i \le \gamma_i/(m_i+1) + 3\gamma_i/(m_i+1)(m_i+2)$ [45]. Hence,

$$E\{\xi_i\} \le E\left\{\frac{\gamma_i}{m_i+1}\right\} + E\left\{\frac{3\gamma_i}{(m_i+1)(m_i+2)}\right\}. \tag{A7}$$

The second term of the right-hand side in (A7) can be expanded as

$$E\left\{\frac{3\gamma_i}{(m_i+1)(m_i+2)}\right\} = \sum_{\ell=1}^{m} \frac{1}{(\ell+1)(\ell+2)}\frac{m!}{\ell!(m-\ell)!}(P_i)^{\ell}(1-P_i)^{m-\ell} = \frac{1}{(m+1)(m+2)(P_i)^2}\Big[1-(1-P_i)^{m+2} -(m+2)P_i(1-P_i)^{m+1} -(m+1)(m+2)(P_i)^2(1-P_i)^m/2\Big].$$

Hence, $\lim_{m\to\infty} mE\{3\gamma_i/(m_i+1)(m_i+2)\} = 0$, which yields

$$\limsup_{m\to\infty} mE\{\xi_i\} \le \frac{1}{P_i}. \tag{A8}$$

Consequently, from (A6) and (A8), $mE\{\xi_i\} \to 1/P_i$ for $i \in \mathcal{I}$. Therefore, $c_3 \to \sum_{i\in\mathcal{I}} \delta_i/nP_i$, which is equal to $c_\infty = \lim_{m\to\infty} c_0$. □

REFERENCES

[1] J. M. Pena, J. A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm," *Pattern Recognit. Lett.*, vol. 20, no. 10, pp. 1027–1040, 1999.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2001.

[3] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2008.

[4] R. Hecht-Nielsen, *Neurocomputing*. New York, NY, USA: Addison-Wesley, 1990.

[5] T. Kohonen, *Self-Organizing Maps*, 3rd ed. New York, NY, USA: Springer, 2001.

[6] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. New York, NY, USA: Addison-Wesley, 1974.

[7] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

[8] D. Pollard, "Quantization and the method of k-means," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 199–205, Mar. 1982.

[9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA, USA: Kluwer, 1992.

[10] N. S. Jaynat and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Upper Saddle River, NJ, USA: Prentice-Hall, 1984.

[11] E. F. Abaya and G. L. Wise, "Convergence of vector quantizers with applications to optimal quantization," *SIAM J. Appl. Math.*, vol. 44, no. 1, pp. 183–189, Feb. 1984.

[12] X. Wu and K. Zhang, "Quantizer monotonicities and globally optimal scalar quantizer design," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 1049–1053, May 1993.

[13] J. Vaisey and A. Gersho, "Simulated annealing and codebook design," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Apr. 1988, pp. 1176–1179.

[14] D. S. Kim and N. B. Shroff, "Quantization based on a novel sample-adaptive product quantizer (SAPQ)," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2306–2320, Nov. 1999.

[15] D. Sik Kim and N. B. Shroff, "Sample-adaptive product quantization: Asymptotic analysis and examples," *IEEE Trans. Signal Process.*, vol. 48, no. 10, pp. 2937–2947, Oct. 2000.

[16] D. S. Kim, "Training ratio and comparison of trained vector quantizers," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1632–1641, Jun. 2003.

[17] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.

[18] D. Cohn, E. A. Riskin, and R. Ladner, "Theory and practice of vector quantizers trained on small training sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 1, pp. 54–65, Jan. 1994.

[19] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1971.

[20] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.

[21] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[22] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. New York, NY, USA: Academic, 1999.

[23] T. Linder, "On the training distortion of vector quantizers," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1617–1623, Jul. 2000.

[24] D. S. Kim and M. R. Bell, "Upper bounds on empirically optimal quantizers," *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 1037–1046, Apr. 2003.

[25] D. S. Kim, T. Kim, and S. U. Lee, "On testing trained vector quantizer codebooks," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 398–406, Mar. 1997.

[26] J. Collura and T. Tremain, "How good is your β?-Observations on VQ training ratios," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 1995, pp. 744–747.

[27] D. Pollard, "A central limit theorem for k-means clustering," *Ann. Probab.*, vol. 10, no. 4, pp. 919–926, 1982.

[28] P. Chou, "The distortion of vector quantizers trained on n vectors decreases to the optimum as O_p(1/n)," in *Proc. IEEE Int. Symp. Inform. Theory*, Jun./Jul. 1994, p. 457.

[29] T. Linder, G. Lugosi, and K. Zeger, "Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1728–1740, 1994.

[30] P. L. Bartlett, T. Linder, and G. Lugosi, "The minimax distortion redundancy in empirical quantizer design," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1802–1813, Sep. 1998.

[31] P. Cosman, K. Perlmutter, S. Perlmutter, R. Olshen, and R. Gray, "Training sequence size and vector quantizer performance," in *Proc. Conf. Rec. 25th Asilomar Conf. Signals, Syst. Comput.*, Nov. 1991, pp. 434–438 vol. 1.

[32] F. Groh, L. Ruppert, P. Wieschollek, and H. P. A. Lensch, "GGNN: Graph-based GPU nearest neighbor search," *IEEE Trans. Big Data*, vol. 9, no. 1, pp. 267–279, Feb. 2023.

[33] P. Billingsley, *Probability and Measure*. New York, NY, USA: Wiley, 1995.

[34] S. Sarkar and A. K. Ghosh, "On perfect clustering of high dimension, low sample size data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2257–2272, Sep. 2020.

[35] D. S. Kim, "Convergence rate on a nonparametric estimator for the conditional mean," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2009, pp. 453–457.

[36] P. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 139–149, Mar. 1982.

[37] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, Dec. 1953.
[38] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[39] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc., B (Stat. Methodol.)*, vol. 63, no. 2, pp. 411–423, 2001.
[40] C. A. Sugar and G. M. James, "Finding the number of clusters in a dataset," *J. Amer. Stat. Assoc.*, vol. 98, no. 463, pp. 750–763, Sep. 2003.
[41] L. Wang, C. Leckie, K. Ramamohanarao, and J. Bezdek, "Automatically determining the number of clusters in unlabeled data sets," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 3, pp. 335–350, Mar. 2009.
[42] Z. He, A. Cichocki, S. Xie, and K. Choi, "Detecting the number of clusters in n-way probabilistic clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2006–2021, Nov. 2010.
[43] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J. Wireless Commun. Netw.*, vol. 2021, no. 1, Dec. 2021.
[44] F. Liu and Y. Deng, "Determine the number of unknown targets in open world based on elbow method," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, May 2021.
[45] E. L. Grab and I. R. Savage, "Tables of the expected value of $1/X$ for positive Bernoulli and Poisson variables," *J. Amer. Stat. Assoc.*, vol. 49, no. 265, pp. 169–177, 1954.

DONG SIK KIM (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1986, 1988, and 1994, respectively. Since 1986, he has been the Research Director with Automan Company Ltd., South Korea, where he has conducted RF circuit design projects. From 1998 to 1999, he was a Visiting Assistant Professor with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. He is currently a Professor with Hankuk University of Foreign Studies, South Korea. His research interests include the theory of quantization, biomedical image processing, medical physics, sensor networks, and smart grid based on statistical signal processing. He was a co-recipient of the 2003 International Workshop on Digital Watermarking Best Paper Prize.