mathematics

MDPI

Article

Clustering Validation Inference

Pau Figuera [1,*], Alfredo Cuzzocrea [2] and Pablo García Bringas [1]

[1] Faculty of Engineering, University of Deusto, 48007 Bilbao, Spain
[2] iDEA Lab, University of Calabria, 87036 Rende, Italy
* Correspondence: pau.figuera@opendeusto.es

Abstract: Clustering validation is applied to evaluate the quality of classifications. This step is crucial for unsupervised machine learning. A plethora of methods exist for this purpose; however, a common drawback is that statistical inference is not possible. In this study, we construct a density function for the cluster number. For this purpose, we use smooth techniques. Then, we apply non-negative matrix factorization using the Kullback–Leibler divergence. Employing a unique linearly independent uncorrelated observational variable hypothesis, we construct a sequence by varying the dimension of the span space of the factorization only using analytical techniques. The expectation of the limit of this sequence follows a gamma probability density function. Then, identifying the dimension of the factorization of the space span with clusters, we transform the estimation of the suitable dimension of the factorization into a probabilistic estimate of the number of clusters. This approach is an internal validation method that is suitable for numerical and categorical multivariate data and independent of the clustering technique. Our main achievement is a predictive clustering validation model with graphical abilities. It provides results in terms of credibility, thus making it possible to compare results such as expert judgment on a quantitative basis.

Keywords: non-negative matrix factorization; trace sequence limit; clustering validation; inferential clustering validation

MSC: 15B48; 62H86; 62H30

Citation: Figuera, P.; Cuzzocrea, A.; García Bringas, P. Clustering Validation Inference. *Mathematics* **2024**, *12*, 2349. https://doi.org/10.3390/math12152349

Academic Editor: Jin-Ting Zhang

Received: 21 March 2024
Revised: 18 July 2024
Accepted: 19 July 2024
Published: 27 July 2024

1. Introduction

Clustering is a technique attributed to MacQueen et al. [1], who posed the problem of partitioning a set of observations into disjoint subsets such that the within-group variance is as small as possible, and they proposed the benchmark *k*-means method [1]. Currently, there is a well-established corpus in the statistics and machine learning (ML) fields that involves the use of extended and well-known methods with the purpose of *a formal study of algorithms and methods for grouping or clustering objects according to measured or perceived intrinsic characteristics or similarities* [2]. The vertiginous growth experienced in such fields in recent decades, particularly in terms of concepts and computational techniques, is reflected in several excellent books published in this field. One of them, providing detailed explanations of the most relevant contributions, is [3], which clearly reflects the difficulty of conducting simple classifications of existing techniques, as there exists an overlapping of methods aggravated with the introduction of eigenanalysis criteria.

An alternative to partitioning methods can be designed through admitting that, in the presence of uncertainty in observations, they may belong to more than one category (to a greater or lesser extent) and by using supporting generative models. These approaches are based on determining suitable mixtures of *components* in terms of number and parameters, and they are denoted by $f(\mathbf{x};\theta) = \sum \alpha_j g_j(x;\theta_j)$, where g is a density with parameter $\theta \in \Theta$, and α denotes the mixing weights assigned to each density. Two interpretations of this formulation exist: the first supposes that each entity belongs to each cluster with a different

probability [4], while the second concept classifies the observations that most likely belong to each distribution [5].

If the data frame of observations can be handled as a matrix **X**, its transformation to the probability space yields a stochastic matrix or *probabilistic image* of the data matrix, thus making non-negative matrix factorization (NMF) techniques especially apt for handling probabilistic structures. Many authors have attributed the introduction of these techniques to [6], while others attribute their introduction to [7]—a study clearly centered on ML. We prefer to attribute their introduction to Chen [8], who was influential in establishing conditions for the existence of stochastic matrices [9]. The main idea is to factorize the stochastic matrix as the product of the other two matrices and of non-negative inputs. This factorization depends on the dimension of the space span (the dimension of the factorization space), which is not determined a priori. Then, the problem of parameter estimation is translated to determining the number of suitable mixtures, or the space span of the factorization, converting the problem to the unsupervised classification of the parameter estimation that assigns probabilities of membership to each item in each cluster. Pioneering applications of clustering using NMF techniques have been demonstrated in studies on the classification of textures [10], bioinformatics [11], instrument classification [12], facial recognition [13], and spectroscopy [14].

Independent of the classification method, an important step is determining the quality of the classification. In this context, the number of clusters is one of the most relevant aspects. This problem, called *clustering validation*, is of active and increasing interest in this research field, and it is critical for unsupervised cases. Additionally, many validation methods exist. A drawback shared by many validation methods is that *the use of one or more criteria may inadvertently satisfy different algorithms. Thus, clustering is a problem in which precise quantification is often not possible because of its unsupervised nature* [3] (pg. 22). The practical consequence is that many existing validation procedures seem to be less stringent with respect to the clustering methods that they validate. From a theoretical viewpoint, the relationships between many of these criteria have not yet been well established. In contrast, qualitative methods exist—usually based on graphical criteria—that provide good estimates but suffer from drawbacks in terms of providing quantitative results [15]. Moreover, for large datasets, the presence of noise can be significant, and making estimations and results that are formally correct can be unrealistic.

A line of work that does not share these problems is based on the idea of stability. Stability is an internal measure that evaluates the results for a choice of the number of clusters. The main idea is that a dataset is a sample of a statistical distribution. The sample is stable if the underlying distribution is the same for different samples obtained from the data [16]. A review work that provides insight into these methods is [17]. A more recent work to determine the partition from the underlying distribution is [18]. This work redefines the concept of clustering as the *partitioning of data into groups so that the partition is stable*, proposing the stadion index. In essence, this index maximizes a function of the stability between clusters and minimizes it within cluster.

In our approach, we derive a probability density function (pdf) to assess the degree of validity for classifications. The first key step is the formulation of the problem in light of the NMF. The second key step is to obtain a sequence of traces by varying the space span of the NMF. Then, the expectation of the limit behavior is given as a pdf.

While the term *inference* is not vague, it has been given different definitions. It is extensively used in logic, referring to the correct result of reasoning with several propositions. In statistics, it refers to *the branch... which deals with the generalizations from samples to the population* parameters, being *the relative frequency interpretation* in the Bayesian context, or *the interpretation of probability* in a subjective sense, which *may estimate the probability of an event on the basis of our experience* or credibility [19]. In the ML context, according to the dictionary of Google (https://developers.google.com/machine-learning/glossary?hl=es-419#i, URL accessed 15 September 2023), it is *the process of making predictions by applying a trained model to unlabeled examples*. This definition implies a simple descriptive statistical framework, re-

stricting the clustering validation problem to the best data structure description in order to handle the data while sacrificing predictions. In this manuscript, the definition of inference is related to inferential statistics.

The remainder of this manuscript is structured as follows: Studies covering similar topics are presented in Section 2. Basic formulas that are well known and extended for use with NMF practitioners are presented in Section 3. In this section, we also propose an error bound for the case of low-rank approximations. The manipulations necessary to obtain the traces and the underlying hypothesis (linear independence of observational variables) are explained in Section 3.4, as well as the main result; that is, the sequence of NMF traces follows a gamma pdf when the dimension of the space span varies, and it is the posterior of a Poisson distribution. We provide some indications for the application of this result for clustering validation in Section 4. In Section 5, we illustrate how the pdf operates in the context of several data configurations. We discuss several gaps and open questions in Section 6 before presenting our conclusions.

A key contribution of this manuscript is the generalization of NMF to real input matrices achieved through transformations to probabilistic space, which can always be justified in the presence of data uncertainty. However, the most important contribution of this work is the provision of a probabilistic validation criterion that allows for credibility to be assigned to clustering results independently of the clustering method. This result enables the construction of credible or acceptance regions, thereby allowing for a comparison of hypotheses or acceptance intervals to be carried out.

The existence of a pdf extends the validation problem to the evaluation on a quantitative basis of criteria other than purely numerical ones, including expert criteria. In the case of serious discrepancies, it allows justifying the re-analysis of the problem with more relevant data. Furthermore, in the case of big data, it may be more appropriate to provide a confidence interval than a single value for classification.

2. Related Work

The problem of clustering validation is as old as the techniques of clustering. Although there are no unified criteria to determine whether it is a geometric problem related to partitioning a set based on similarities or a problem of label assignment, this conceptual nuance does not affect the practical results. Validation requires that several properties (e.g., sensibility, cluster number impact, invariance) be determined in order to establish the capability of a cluster method for various data structures. There exist many studies that have justified these criteria, such as [3] (chap. 23) and [20].

One of the first statistical attempts to use multivariate techniques to determine the number of clusters was detailed in [21]. At this time, the use of statistical concepts to construct validation indices was quasi-mandatory [22]. A later work by the same author pointed out the effects of sample size, data dimension, and cluster spread [23]. These ideas are currently relevant, considering the increasing size of datasets. In [24], the difference between graphical methods and those that postulate the existence of an underlying parametric statistical model was pointed out. This idea is relevant when using the likelihood ratio test, which is the starting point for many fuzzy and probabilistic validation criteria. Furthermore, the current classification of validation methods was introduced in this work. One such criterion is the classical silhouette index, which evaluates the within and between variances of the groups into which the data have been divided for the selection of the cluster number [25]. Further studies focusing on this viewpoint of statistical techniques for clustering validation include Har [26], which introduced an indicator function for observations based on kernelization and used the null hypothesis as a classifier.

In Smyth [27], likelihood cross-validation was used to infer information on the number of model components. At this time, and as a consequence of the expansion of the Internet, other problems appeared regarding what constitutes a good measure of similarity. In Halkidi, a detailed survey analyzing the problem of data similarity was provided. This study raised the problem of how to measure the goodness of fit of the data relative to

the groups that use concepts of compactness and separation. This work serves to create a consensus according to the types of indices: external validation (the availability of additional information, such as a subset of labeled data) and internal validation (the case in which only information on the data is available) [28]. This work also provided many examples. A classical index derived from these ideas is the gap statistic, which compares the within-cluster dispersion to the expectation of a reference distribution [29].

Using a different approach, the similarity between clusters can be evaluated using the χ^2 statistic between probabilistic classifications [30]. In Brun, a model for error prediction based on the correlation error rate vs. several validity measures was provided. The Kendall rank correlation index can be used to quantify the degree of similarity between the validation indices and the classification errors [31]. By introducing an index and assuming that each cluster is generated by a parametric distribution, the minimum can be taken as the validation index [32]. The effect of clusters with different sizes and densities was studied in [33].

More recent works include [34], which, within the scope of astronomical observations and under the hypothesis of normality and the existence of a correlation, presented an algorithm in which the posterior distribution of the correlations follows a gamma pdf. A quantitative discriminant method using the elbow point for the determination of the optimal cluster number was presented in [35]. In [36], a purely algebraic approach was presented, in which elements are clustered according to their co-linearity. A review centered on the impact and importance of clustering validation in the context of the recent growth of bioinformatics was presented in [37], with the evaluation of the number of clusters being a more recent development [38].

3. Probabilistic Space Domain Transformation, Parametrization, and Non-Negative Factorization Sequence Traces

The transformation of data to the probabilistic space allows one to search for differences or similarities in terms of probabilities. This transformation is justified when uncertainty exists in the data estimation. Several techniques that involve transformation based on a kernel are extensively used in the ML context [3] (p. 10). Additionally, well-established methods for smoothing represent a commonly employed solution to obtain good results.

One of the relevant issues in the statistics literature is the determination of the underlying pdf of a sample. In the multivariate case, the data frame containing this sample is the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$, in which m observations are formulated as i ($i = 1, \ldots, m$) rows obtained under the j ($j = 1, \ldots, n$) observational criteria. Thus, an observation under the j^{th} criterion is the column vector $\mathbf{x}_j$, while the row vectors provide a description of an item or observation. We represent these observations as i' as $x_{i'1}, \ldots, x_{i'n}$. Additionally, we assume the linear independence of the columns of $\mathbf{X}$.

3.1. Probabilistic Image of a Real-Valued Data Matrix

3.1.1. Real Field Domain

The transformation to the probabilistic space for the case in which the data frame of observations can be written as the matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ requires the determination of the underlying pdfs $f_1, \ldots, f_n$, such that

$$[\mathbf{X}]_{ij} = \begin{bmatrix} f_1^{-1}(x_1) & \ldots & f_n^{-1}(x_1) \\ \vdots & \ddots & \vdots \\ f_1^{-1}(x_m) & \ldots & f_n^{-1}(x_m) \end{bmatrix} \tag{1}$$

$$= \left[f_1^{-1}(\mathbf{x}) | \ldots | f_n^{-1}(\mathbf{x}) \right]. \tag{2}$$

Each column vector is associated with a density corresponding to the observational criteria used to generate the data, which are unknown in the general case.

The estimate of each density in (2) using a smoothing technique requires the j density to be approximated by p mixtures of a kernel density function (kdf), providing an estimate $\widehat{f}_j$ that approximates f_j as

$$\widehat{f}_j(x;h) = \frac{1}{p}\sum_{p} K_h\Big(x - x_p;h\Big), \tag{3}$$

where K_h denotes the kdf. This adjustment involves the choice of the shape of the kdf and the selection of the smoothing parameter h. This point is controversial, as a small value provides a good estimation for each point but increases the variance, while large values reduce the variance but increase the bias and obscure the underlying data structure. An extended criterion for its selection involves the estimation of the mean integrated standard error (MISE), defined as

$$\text{MISE}\,(\hat{f}) = \mathbb{E}\Big\{\int_{\mathcal{D}}(\hat{f} - f)dx\Big\}^2, \tag{4}$$

where $\hat{f}$ is the MISE estimate of f.

After juxtaposing the densities $\widehat{f}_j(\mathbf{x};h)$, the result is the following stochastic column matrix:

$$[\widetilde{\mathbf{Y}}]_{ij} = \big[\widehat{f}_1(\mathbf{x};h_1)|\ldots|\widehat{f}_n(\mathbf{x};h_n)\big] \qquad \big(\text{s.t. } \|\widehat{f}_j(\mathbf{x};h_j)\|_1 = 1 \text{ for all } j{=}\,1,\ldots,n\big) \tag{5}$$

where $\|\cdot\|_1$ is the L_1 norm in the Shalten sense.

This case is univariate, while the multivariate case involves considering the x_{ij} entries of (1) as $x_{ij} = f_j^{-1}(x_i)$. Smoothing with the kdf is carried out:

$$\widehat{f}(\mathbf{X};\mathbf{H}_{\mathcal{K}}) = \frac{1}{p}\sum_{p} K_{\mathbf{H}_{\mathcal{K}}}(\mathbf{x}_p - \mathbf{x}), \tag{6}$$

where $\mathbf{H}_{\mathcal{K}}$ is the matrix of smooth parameters, usually written as $\mathbf{H}$, in which we add the sub-index $\mathcal{K}$ referring to the kernel in order to avoid confusion with the matrix $\mathbf{H}$ used in the NMF context. Additionally, $\mathbf{x}_p$ is a row vector in (1), and $\mathbf{x}$ represents the other row vectors weighted by the kdf. $\mathbf{H}_{\mathcal{K}}$ is a positive semi-definite matrix containing the smooth parameters:

$$\mathbf{H}_{\mathcal{K}} = \big[\mathbf{h}_1,\ldots,\mathbf{h}_n\big]. \tag{7}$$

Moreover, the probabilistic image is

$$[\mathbf{Y}]_{ij} = f^{-1}(\mathbf{X};\mathbf{H}_{\mathcal{K}}). \tag{8}$$

The relationship between (5) and (8) is determined by

$$[\widetilde{\mathbf{Y}}]_{ij}\mathbf{D}_C = [\mathbf{Y}]_{ij}, \tag{9}$$

where $\mathbf{D}_C$ satisfies $\|\mathbf{D}_C\|_1 = 1$. Thus, pseudo-inverses are taken:

$$\mathbf{I}\mathbf{D}_C = \Big([\widetilde{\mathbf{Y}}]'_{ij}[\widetilde{\mathbf{Y}}]_{ij}\Big)^{-1}\Big([\widetilde{\mathbf{Y}}]'_{ij}[\mathbf{Y}]_{ij}\Big), \tag{10}$$

where $\mathbf{I}$ is an identity matrix of suitable dimension, and

$$\mathbf{D}_C = \text{diag}\Big[\frac{1}{m}\Big] \qquad (\text{s.t. } \mathbf{D}_C \in \mathbb{R}^{m\times m}). \tag{11}$$

More details on smoothing techniques have been presented in [39], which contains many examples and the R code. The multivariate case has been explained in depth in [40].

Additionally, the probabilistic sense of the matrices in (5) and (8) can be formally justified by considering the set $\mathcal{B} = \{x_i\}$ (or $\{x_{ij}\}$), which takes values in all possible outcomes of the set Ω. Thus, the problem can be stated in terms of the triplet $(\Omega, \mathcal{B}, P)$, with P being a measure or probability and $\mathcal{B}$ being a Borel σ-algebra. The map of the inverse image of $\mathcal{B}$ is $F^{-1}(\mathcal{B}) = \{\omega \text{ s.t. } F(\omega) \in \mathcal{B}\}$ $(\omega \in \Omega)$. The Borel sets define probabilities $P(\mathbf{x} \in \mathcal{B}) = P(F^{-1}(\mathcal{B}))$ with distributions $P(\mathbf{x} \leq x)$ and $P(\mathbf{x}_j) = 1$; this allows for the estimation of the density f_j (or f) associated with the distribution P, which generates the data [41] (Chap. 1).

3.1.2. The Particular Case of the Positive Integer Domain

A classical probabilistic transformation widely used in many contexts consists of estimating the relative frequencies of a data frame $N(d_i, w_j)$, obtained from a corpus of d_i $(i = 1, \ldots, m)$ documents when crossed with a thesaurus of w_j $(j = 1, \ldots, n)$ words [42,43]. This approach has been extended to many other cases in which the data domain takes values in the positive integers [44], many of which have been explained in our recent work [45]. Then, the relative frequencies are the following probabilities:

$$P(d_i, w_j) = \frac{N(d_i, w_j)}{\sum_j \sum_i N(d_i, w_j)}, \tag{12}$$

which are estimated via joint probability $P(d_i, w_j)$.

The Bayes rule provides the following:

$$P(d_i, w_j) = P(d_i | \, w_j) P(w_j) \qquad \left(\text{with } \sum_i P(d_i | \, w_j) = 1 \text{ for all } j\right). \tag{13}$$

The introduction of matrix notation requires the identification of $\mathbf{N} \sim P(d_i, w_j)$ and $\widetilde{\mathbf{N}} \sim P(d_i | \, w_j)$. Each column of $\widetilde{\mathbf{N}}$ has an L_1 norm of 1. In this case, we have the following:

$$[\mathbf{N}]_{ij} = [\widetilde{\mathbf{N}}]_{ij} \mathbf{D}_C, \tag{14}$$

with $\mathbf{D}_C$ as defined in (11).

Furthermore, if $\mathbf{D}_N = \text{diag}[1/\sum_j N(d_i, w_j)]_{j=1}^n$, the transformation (12) can be written as

$$\frac{N(d_i, w_j)}{\sum_j \sum_i N(d_i, w_j)} = N(d_i, w_j) \mathbf{D}_C \mathbf{D}_N. \tag{15}$$

Transformation (12) is based on Laplace's probability definition, and its use is limited to the case of count matrices or contingency tables. This restriction greatly limits the use of NMF techniques.

3.1.3. Equivalence

The probabilistic transformations given by (9) and (14) for the case of positive integer and real-entry matrices are equivalent, and equality between both results can be obtained if the entries of $\widetilde{\mathbf{N}}$ and $\widetilde{\mathbf{Y}}$ are the same. This result can be obtained by introducing a triangular kdf in (3) and a suitable smoothing parameter h.

The triangular kdf is based on the triangular function, as shown in Figure 1. It was investigated in [46,47], in which it was stated that it corresponds to a discrete pdf, while it was found to be continuous in [41] (Chap. 13). This question depends on the conditions of the definition of the variable domain and its support.

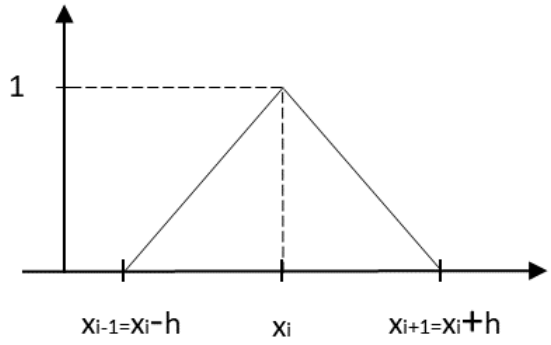


Figure 1. Triangular kernel. The triangular kernel is based on the triangular function, defined as $f(x) = 1 - |x|$ on the set [−1,1]. For each x_i, the smoothing parameter h ensures that $f(x_i) = 1$, whereas it is $f(x_{i-1}) = f(x_{i+1}) = 0$. The figure shows the behavior in the sense of Formula (16).

The triangular kernel is introduced as follows [46,47]:

$$K(x;h) = \begin{cases} \dfrac{h - |x - x_i|}{h^2} & (\text{if } x_i - h \leq x \leq x_i + h)\ (h > 0) \\ 0 & \text{otherwise} \end{cases}. \tag{16}$$

Writing the difference $|x - x_i| = x_i + |h|\, x_i$ for $h = 1$ the *grid* (i.e., the values at which the density is estimated) $x_i = (x_1, \ldots x_m)$, the interval $[x - h, x + h]$ contains a unique point. Then, we have the following:

$$K(x) = x_i, \tag{17}$$

with the density estimate

$$\widehat{f}(x_i;h) = \frac{x_i}{n_{ri}}, \tag{18}$$

where n_{ri} is the number of observations at x_i. To obtain the probabilistic image, this procedure must be carried out for all the columns of (1).

The multivariate case requires the consideration of a multivariate kernel in the form of $K(\mathbf{X}; \mathbf{H}_{\mathcal{K}})$. According to our literature search, the d-dimensional triangular kernel has not been defined; although we think that it will not be difficult to obtain, it remains an unresolved issue. We point out that once (18) is obtained, the analogous multivariate case introducing the diagonal matrix $\mathbf{D}_C$ (as defined in (11)) can be immediately obtained. This matrix corresponds to a uniform distribution, with equal weights assigned to all observational variables. This result demonstrates that the Laplace rule is a particular case of the wider transformation to the probabilistic space.

We call the stochastic matrices $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}$ the *probabilistic image* and *column probabilistic image*, respectively. Moreover, they are the *transformations to the probabilistic space* or simply *probabilistic transformations* of the data.

3.2. Parametrization

To achieve the factorization of (8) while maintaining a probabilistic sense, one must impose the NMF on $\mathbf{Y}$ to obtain the matrices $\mathbf{W}$ and $\mathbf{H}$ as the following product:

$$[\widehat{\mathbf{Y}}]_{ij} = [\mathbf{W}]_{ik}[\mathbf{H}]_{kj} \qquad \left(\text{for } \|[\mathbf{WH}\|_1 = 1 \text{ with } \mathbf{W}, \mathbf{H} \geq 0 \text{ for all } k = 1, 2, \ldots\right) \tag{19}$$

$$\approx [\mathbf{Y}]_{ij}, \tag{20}$$

which minimizes an objective or cost function. (The standard NMF formulation is usually stated as the product $\mathbf{Y} = \mathbf{WH} + \mathbf{E}$, being the objective to minimize $\mathbf{E}$ [48] (p. 8). In this work we have preferred to ignore the matrix $\mathbf{E}$ and formulate it strictly as an approximation problem.) This factorization always exists for non-negative real matrices [8], and it has a probabilistic sense if the normalization conditions of (19) are fulfilled.

The Kullback–Leibler (KL) divergence is chosen as the following objective function [49]:

$$D_{KL}\big([\mathbf{Y}]_{ij}\big\|[\mathbf{WH}]_{ij}\big) = \sum_j \sum_i [\mathbf{Y}]_{ij} \odot \log \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} \tag{21}$$

$$= \sum_j \sum_i \Big([\mathbf{Y}]_{ij} \odot \log[\mathbf{Y}]_{ij} - [\mathbf{Y}]_{ij} \odot \log[\mathbf{W}\,\mathbf{H}]_{ij}\Big), \tag{22}$$

where $\odot$ is the Hadamard (or element-wise) product, and the matrix fraction is the quotient for equal sub-index entries.

Then, (21) is minimized according to the Karush–Kuhn–Tucker conditions [48] (p. 141):

$$[\mathbf{W}]_{ik} \odot \nabla_{\mathbf{W}} D_{KL} = 0 \quad \Big(\text{for } \nabla_{\mathbf{W}} D_{KL}\big([\mathbf{Y}]_{ij}\big\|[\mathbf{WH}]_{ij}\big) \geq 0 \text{ with } \mathbf{W} \geq 0\Big) \tag{23}$$

and

$$[\mathbf{H}]_{kj} \odot \nabla_{\mathbf{H}} D_{KL} = 0 \quad \Big(\text{for } \nabla_{\mathbf{H}} D_{KL}\big([\mathbf{Y}]_{ij}\big\|[\mathbf{WH}]_{ij}\big) \geq 0 \text{ with } \mathbf{H} \geq 0\Big), \tag{24}$$

which provides the following solutions [50]:

$$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} \odot \left(\frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} [\mathbf{H}]'_{kj}\right) \tag{25}$$

$$[\mathbf{H}]_{kj} \leftarrow [\mathbf{H}]_{kj} \odot \left([\mathbf{W}]'_{ik} \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}}\right). \tag{26}$$

This procedure is a particular case of NMF in which divergences are used, which are known as *iterative updates* or *multiplicative iterative algorithms* [48] (Chap. 3). In [51], it was demonstrated that this procedure maintains the normalization conditions in the iterative process.

To adjust the product (19) or approximation (20), it is necessary to select a value for k, initialize the matrices $\mathbf{W}$ and $\mathbf{H}$, and then apply an iterative process (switching between (25) and (26)) until a condition is reached where the approximation degree of (19) is achieved.

This procedure of approximating $\mathbf{WH}$ to $\mathbf{Y}$ in order to maintain matrix normalization involves an approximation of $\widehat{\mathbf{Y}}$ to $\mathbf{Y}$, which are densities. Thus, it is a sufficient (not minimal) statistic—roughly speaking, the density contains the full sample's information. Therefore, the products $\{\mathbf{W}, \mathbf{H}\}$ are distributional parameters with no distributional assumption, constituting a non-parametric approach (in the sense that no hypothesis is made for the parameters). We also refer to the columns of $\mathbf{W}$ (or $\widetilde{\mathbf{W}}$) and the rows of $\mathbf{H}$ (or $\widetilde{\mathbf{H}}$) as *components*.

3.3. Convergence of Solutions

The approximation of (19) requires the following:

$$\Big\|[\widehat{\mathbf{Y}}]_{ij} - [\mathbf{Y}]_{ij}\Big\|_1 \leq \epsilon \qquad (\epsilon > 0). \tag{27}$$

Thus, convergence in the iterative process of approximating $\widehat{\mathbf{Y}} \to \mathbf{Y}$ must be proven.

Usually, these proofs are based on the same strategy as the Expectation-Maximization (EM) algorithm, which always converges [52]. Basically, they consist of postulating the existence of a function G accomplishing $G(h, h') \geq F(h)$ and $G(h, h) = F(h)$, which leads to the sequence $h^{p+1} = \arg\min G(h, h')$. This procedure is similar to that detailed in [7].

Based on this proof, it is usually stated that factorization (19) with the KL divergence converges to the matrix $\mathbf{Y}$ through a simple observation of the expansion of the KL divergence of (22): While the first term is a constant, the second is the log-likelihood. A more general conclusion has been derived by [53], showing the asymptotic properties of this solution.

An alternative naive proof can be obtained by taking into account the fact that the divergence (22) takes positive values (divergences are non-negative). Thus, $\|\mathbf{Y}\|_1 > \|\mathbf{WH}\|_1$. Then, for matrices satisfying this condition, Formulas (25) and (26) can be rewritten for the p^{th} iteration as follows:

$$\mathbf{W}^{(p+1)} = \mathbf{W}^{(p)} \odot \mathbf{W}^{(a)}, \tag{28}$$

$$\mathbf{H}^{(p+1)} = \mathbf{H}^{(p)} \odot \mathbf{H}^{(a)}, \tag{29}$$

interpreting $\mathbf{W}^{(a)}$ and $\mathbf{H}^{(a)}$ as actualization weight matrices.

As all the entries of the matrices $\mathbf{W}^{(a)}$ and $\mathbf{H}^{(a)}$ take values in $[0, 1]$, it holds that $\mathbf{W}^{(p+1)} < \mathbf{W}^{(p)}$ (also for the matrices in (29)). Then, the product of the matrices $\mathbf{W}$ and $\mathbf{H}$ is monotonically decreasing, and the difference (27) also decreases.

A consequence of convergence is an approximation error bound. This requires the consideration of two cases. First, for $k \geq \min(m, n)$, the convergence $\widehat{\mathbf{Y}} \to \mathbf{Y}$ is well known (a formal proof can be found in [50]).

The case $k < \min(m, n)$ is known as *low-rank* approximation, which has particular interest in practice. It poses several problems, as reported in the literature. In the context of the probabilistic latent semantic analysis (PLSA), it has been pointed out that the convergence limit does not necessarily occur at a global optimum, thus providing sub-optimal results [54].

Our proposal to establish a convergence limit is similar to that used by Schmidt for the approximation theorem [55].

Theorem 1 below was formulated according to the singular value decomposition (SVD) theorem for real-valued matrices [56] (p. 275).

Theorem 1. *Let* $\mathbf{X} \in \mathbb{R}^{m \times n}$. *Then, orthogonal (or unitary) matrices* $\mathbf{U} \in \Re^{m \times m}$ *and* $\mathbf{V} \in \Re^{n \times n}$ *exist such that*

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}' \qquad \Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_r)$ *with diagonal entries*

$$\sigma_1 \geq, \ldots, \geq \sigma_r > 0 \quad r = \text{rank}(\mathbf{A}).$$

The approximation for $r' < r$ is known as the *low-rank SVD approximation*. Additionally, it is necessary to consider that the spectral norm of a matrix is the sum of its eigenvalues (or trace). Thus, by writing the inequality

$$\|\mathbf{Y} - \mathbf{WH}\|_1 \leq \left\|\mathbf{Y}\right\|_1 - \left\|\mathbf{WH}\right\|_1 \tag{30}$$

and taking into account approximation (27) for the eigenvalues of the second term of (30),

$$\epsilon \leq \left(\sigma_1([\mathbf{Y}]_{ij}) + \cdots + \sigma_r([\mathbf{Y}]_{ij})\right) - \left(\sigma_1([\widehat{\mathbf{Y}}]_{ij}) + \cdots + \sigma_{r'}([\widehat{\mathbf{Y}}]_{ij})\right) \tag{31}$$

$$= \left(\sigma_1([\mathbf{Y}]_{ij}) - \sigma_1([\widehat{\mathbf{Y}}]_{ij})\right) + \cdots + \left(\sigma_{r'}([\mathbf{Y}]_{ij}) - \sigma_{r'}([\widehat{\mathbf{Y}}]_{ij})\right) + \sigma_{r'+1}([\mathbf{Y}]_{ij}) + \cdots + \sigma_r([\mathbf{Y}]_{ij}), \tag{32}$$

as the common eigenvalues are the same [56] (p. 277):

$$\epsilon \leq \sum_{r=r'+1}^{r} \sigma_r([\mathbf{Y}]_{ij}), \tag{33}$$

Formula (33) provides the error bound for the approximation case $k < \min(m, n)$. Furthermore, in [43], it was demonstrated that

$$[\mathbf{WH}]_{ij} = \frac{1}{k}[\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj}. \tag{34}$$

3.4. Non-Negative Factorization Sequence Traces

In Formula (19), the trace

$$\mathrm{tr}(\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}) = \sum_k \left(\mathrm{diag}([\mathbf{W}]_{ik}[\mathbf{H}]_{kj})'([\mathbf{W}]_{ik}[\mathbf{H}]_{kj})\right)^{1/2} \tag{35}$$

$$= \left\|\mathrm{diag}\left(([\mathbf{W}]_{ik}[\mathbf{H}]_{kj})'([\mathbf{W}]_{ik}[\mathbf{H}]_{kj})\right)^{1/2}\right\|_1 \tag{36}$$

leads to a sequence with varying k values in the NMF factorization space span.

$$\left\{z_{[k]}\right\}_k = \left\{\mathrm{tr}\left(([\mathbf{W}]_{i1}[\mathbf{H}]_{1j})'([\mathbf{W}]_{i1}[\mathbf{H}]_{1j})\right)^{1/2}, \mathrm{tr}\left(([\mathbf{W}]_{i2}[\mathbf{H}]_{2j})'([\mathbf{W}]_{i2}[\mathbf{H}]_{2j})\right)^{1/2}, \ldots\right\} \tag{37}$$

$$= \left\{z_1, z_2, \ldots\right\}. \tag{38}$$

Here, the sub-index brackets indicate that terms are given in increasing order. We can write z as

$$z = \mathrm{tr}\left(([\mathbf{Y}]'_{ij}[\mathbf{Y}]_{ij})\right)^{1/2}. \tag{39}$$

In Formulas (38) and (39), the quotient

$$\tilde{z} = \left\{\frac{z_{[k]}}{z}\right\}_k \tag{40}$$

leads to a monotonically decreasing sequence. To achieve a reliable sequence, it is necessary to establish the same convergence condition for all of the terms. This requires consideration of the following cases.

If $k \geq \min(m, n)$ occurs, then $\widehat{\mathbf{Y}} \to \mathbf{Y}$ during the iteration process. By imposing the same approximation condition for ϵ on all products of **WH** obtained with different values of k and introducing

$$\|[\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\|_1 = \frac{1}{k}\|[\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj}\|_1, \tag{41}$$

the inequality

$$\frac{1}{k^2}\mathrm{tr}\left(([\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj})'([\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj})\right)^{1/2} > \frac{1}{(k')^2}\mathrm{tr}\left(([\widetilde{\mathbf{W}}]_{ik'}[\widetilde{\mathbf{H}}]_{k'j})'([\widetilde{\mathbf{W}}]_{ik'}[\widetilde{\mathbf{H}}]_{k'j})\right)^{1/2} \tag{42}$$

holds only if $k' > k$.

If $k < \min(m, n)$, then a convergence error bound given by (33) exists. Thus, it is necessary to impose a wider condition for the approximation of (27) (i.e., $\epsilon = O(1/m)$), which is a jump in the empirical univariate distribution of the columns of $\widetilde{\mathbf{Y}}$. As the convergence of the empirical distribution to each component (column) of $\widetilde{\mathbf{Y}}$ is almost sure (a.s.), the difference is bounded. After taking the maximum difference, a decreasing behavior is observed when we carry out division by increasing values of k.

3.4.1. Trace Sequence Limit Behavior

For sequence (40) obtained from a full-rank matrix, the function

$$\varphi(z_{[k]}) = \left(\frac{z_{[k]}}{z}\right)^{-z} \tag{43}$$

represents the similarity between z and z_k in terms of the inverse of the (non-logarithmic) likelihood. This function can be written as

$$\varphi(z_{[k]}) = \left(1 + \frac{z_k - z}{z}\right)^{-z}. \tag{44}$$

By introducing the following transformations

$$\lambda = z_1 - z \tag{45}$$

$$\frac{1}{\nu} = \frac{z_k - z}{z_1 - z} \qquad (k \neq 1), \tag{46}$$

with the Jacobian

$$|J| = \det \begin{bmatrix} \frac{\partial z_1}{\partial \lambda} & \frac{\partial z_1}{\partial \nu} \\ \frac{\partial z_k}{\partial \lambda} & \frac{\partial z_k}{\partial \nu} \end{bmatrix} \tag{47}$$

$$= \det \begin{bmatrix} 1 & 0 \\ 0 & \frac{\lambda}{\nu^2} \end{bmatrix} \tag{48}$$

$$= \frac{\lambda}{\nu^2}, \tag{49}$$

we can express (43) as a function of the new variables as follows:

$$\varphi(\lambda, \nu) = \frac{\lambda}{\nu^2}\left(1 + \frac{\lambda}{z\nu}\right)^{-z} \qquad \left(1 \leq \nu < +\infty\right). \tag{50}$$

Formula (45) is merely a displacement, and Relation (46) transforms the domain of z_k to a set with lower bound 1 but no upper bound. This transformation does not change the dimension of the space, as ν depends on λ.

Hence,

$$\varphi(\lambda, \nu) = \frac{z}{z-1}\frac{\partial}{\partial \nu}\left(1 + \frac{\lambda}{z\nu}\right)^{1-z}. \tag{51}$$

As

$$\left(1+\frac{\lambda}{z\nu}\right) \xrightarrow[\nu\to\infty]{} \exp\left(\frac{\lambda}{z\nu}\right), \tag{52}$$

by substituting (52) into (51) and taking into account the fact that z is a constant, we obtain

$$\varphi(\lambda,\nu) = c\,\frac{\partial}{\partial\nu}\,\exp\left(-\frac{\lambda}{c\nu}\right) \qquad \left(\text{s.t. } c=\frac{z}{z-1}\right). \tag{53}$$

With the sole purpose of facilitating further calculations and avoiding an incomplete inverse gamma, we change the variables $y=\nu-1$, which ensures that the variation domain is $(0,+\infty)$, with no effect on the scale (the Jacobian is 1). Then, we take

$$x=\frac{1}{cy}. \tag{54}$$

Function (53) can be rewritten as follows:

$$\varphi(\lambda,x) = \frac{1}{c}\,\frac{\partial}{\partial\,x}\frac{\partial^2}{\partial\lambda^2}\,e^{-\lambda\,x}. \tag{55}$$

More generally,

$$\frac{\partial^{r-1}}{\partial\,x^{r-1}}\frac{\partial^{p-2}}{\partial\lambda^{p-2}}\varphi(\lambda,x) = \frac{1}{c}\frac{\partial^r}{\partial\,x^r}\frac{\partial^p}{\partial\lambda^p}\,e^{-\lambda\,x} \quad (p\geq 2 \text{ and } r\geq 1). \tag{56}$$

3.4.2. Expectation of Trace Sequence Limit

To solve (56), we take into account (55). As exponential functions are sufficiently regular to interchange the derivative signs, considering Formula (55), we introduce

$$\zeta(\lambda,x) = \frac{\partial^r}{\partial\,x^r}\varphi(\lambda,x). \tag{57}$$

After taking a derivative, (57) becomes

$$\frac{\partial^{p-2}}{\partial\lambda^{p-2}}\zeta(\lambda,x) = \frac{\lambda}{c}(-1)^p x^p e^{-\lambda\,x}. \tag{58}$$

The Laplace transform of (58) is

$$\zeta_{(p-2)}(s) = \frac{\lambda}{c}(-1)^p\int_0^{+\infty} e^{-sx}x^p e^{-\lambda\,x}\,dx \qquad (s>0) \tag{59}$$

$$= \frac{\lambda}{c}(-1)^p\left(\frac{1}{s+\lambda}\right)^{p+1}. \tag{60}$$

From (55), we find

$$\zeta(s) = \frac{\lambda}{c}\left(\frac{1}{s+\lambda}\right). \tag{61}$$

The relationship between (60) and (61) provides the following recursive formula:

$$\zeta(s) = \frac{(-1)^{p+1}}{(p+1)!}\zeta_{(p-2)}^{(p+1)}(s), \tag{62}$$

where we indicate the order of the derivative in parentheses to avoid confusion with the exponents.

It follows that

$$\int_0^{+\infty} e^{-sx}\zeta(\lambda,\, x)\, dx = \frac{(-1)^{p+1}}{(p+1)!}\int_0^{+\infty} e^{-sx}x^p e^{-\lambda\, x}\, dx. \tag{63}$$

Hence,

$$\zeta(x; p, \lambda) = \frac{(-1)^{p+1}}{(p+1)!}x^p e^{-\lambda\, x}. \tag{64}$$

By reversing the change given by (57) for $r = p + 1$, we obtain

$$\varphi(x; p, \lambda) = \frac{\partial^p}{\partial\, x^p}\zeta(x; p, \lambda). \tag{65}$$

The negative signs cancel each other out, and we find

$$\varphi(x; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda\, x} \qquad (\alpha = p + 1). \tag{66}$$

Formula (66) is a solution of Equation (56), and this is the main result of this manuscript. As the Laplace transform can be viewed as an expectation ($E(\exp -sx) = \zeta(s)$), this relation can be interpreted as the expectation of the limit function, obtained from the sequences of a non-negative matrix trace that follows a gamma pdf.

3.4.3. Gamma Parameter Selection

The adjustment of an unbiased (gamma) density for sequence (43) implicitly assumes

$$\varphi(x_m; \alpha, \lambda) = \max\ c\left(\frac{z_{[k]}}{z}\right)^{-z} \quad \left(\text{s.t. } x_m = \arg\max_x \varphi(x; \alpha, \lambda)\right), \tag{67}$$

for c given by Formula (54) and imposes values for parameters α and λ. Additionally, x_m can be obtained in closed form as

$$x_m = \frac{\alpha - 1}{\lambda}. \tag{68}$$

The classical transformation in (66) is introduced as follows:

$$t = \lambda\, x \qquad \left(\text{with } \left|\frac{\partial\, x}{\partial\, t}\right| = \frac{1}{\lambda}\right), \tag{69}$$

which leads to

$$\varphi(t; \alpha) = \frac{1}{\Gamma(\alpha)}t^{\alpha-1}e^{-t}. \tag{70}$$

This is a standard gamma density with the following maximum:

$$\arg\max_t \varphi(t; \alpha) = \alpha - 1 \tag{71}$$

and expectation α.

Figure 2 shows the effects of the parameters on the gamma pdf.

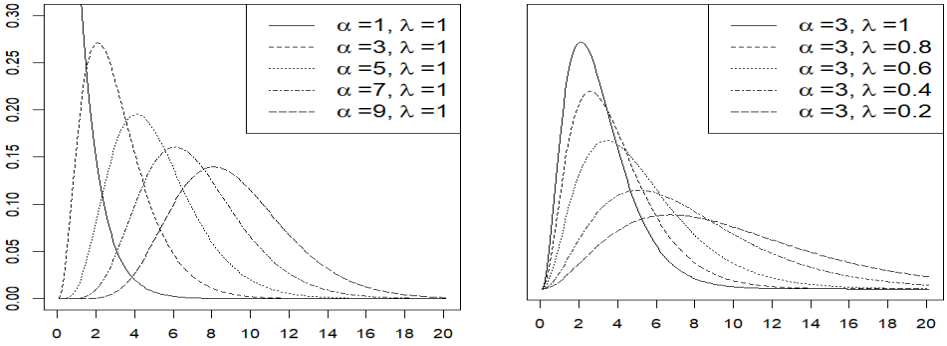


Figure 2. Gamma pdf plots. For both plots, the x-axis shows the gamma density argument; the y-axes are the corresponding probabilities. Left panel shows the effect of parameter α with fixed lambda. Right panel fixes α and varies λ. Both figures share the same axis scale.

If (66) and (70) must reproduce the same shape with $\lambda > 0$, it is necessary to adjust the expectation of (68) to the maximum of (72). Then,

$$\alpha = \frac{\alpha - 1}{\lambda} \tag{72}$$

and

$$\lambda = 1 - \frac{1}{\alpha}. \tag{73}$$

3.5. Relation between Solutions

Another solution for Equation (56) can be obtained by writing

$$\frac{\partial^r}{\partial x^r}\frac{\partial^2}{\partial \lambda^2}\varphi(\lambda, x) = \frac{(-1)^r}{c}\lambda^r \frac{\partial^2}{\partial \lambda^2}\varphi(\lambda, x) \tag{74}$$

and denoting $\psi(\lambda, x) = \partial^2\varphi(\lambda, x)/\partial\lambda^2$. The relation between derivatives with respect to x is

$$\frac{\partial^r}{\partial x^r}\psi(\lambda, x) = \frac{(-1)^r}{c}\lambda^r \psi(\zeta, x). \tag{75}$$

Following the same reasoning described in Section 3.4.2 and considering that $\partial^2\varphi(\lambda, x)/\partial\lambda^2 = x^2\exp(-\lambda x)$, the Laplace transforms of $\psi^{(p)}$ and ψ are

$$\psi_{(r)}(s) = (-1)^r\frac{\lambda^r}{c}\int_0^{+\infty} e^{-sx}\psi^{(r)}(\lambda, x)dz \tag{76}$$

$$= (-1)^r\frac{\lambda^r}{c}\left(\frac{1}{s+\lambda}\right)^{r+1} \tag{77}$$

$$\psi(s) = \frac{1}{c}\int_0^{+\infty} e^{-sx}\psi(\lambda, x)dx \tag{78}$$

$$= \frac{1}{c}\left(\frac{1}{s+\lambda}\right), \tag{79}$$

respectively.

Comparing (77) and (79), we find

$$\psi(s) = \frac{(-1)^r}{r!}\lambda^r\psi^{(r)}(s), \tag{80}$$

which leads to

$$\psi(x;\lambda) = \frac{1}{r!}\lambda^r e^{-\lambda x}. \tag{81}$$

Equation (81) is a Poisson equation with parameter λ ($\lambda > 0$), and it also solves the differential Equation (56).

The pdfs that solve Equation (56) state the problem in a Bayesian context. The gamma and Poisson densities are conjugate. The interpretation in this context is immediate when comparing (66) and (81). For a value of $r = p + 1$, we achieve equality by introducing a factor $\prod_p x$, and

$$\frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x} = \left(\prod_p x\right)\frac{1}{(p+1)!}\lambda^{p+1}e^{-\lambda x}. \tag{82}$$

The Bayesian theorem can be written as follows:

$$P(\theta|\, x) = L(\theta)\, P(x|\theta), \tag{83}$$

where $P(\theta|\, x)$ is the posterior factor, $L(\theta)$ is the likelihood factor, and $P(x|\theta)$ is the prior factor. We identify these factors as follows:

$$P(\theta|\, x) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x} \tag{84}$$

$$P(x|\theta) = \frac{1}{(p+1)!}\lambda^{p+1}e^{-\lambda x} \tag{85}$$

$$L(\theta) = \prod_p x \tag{86}$$

$$= x^{\alpha-1}, \tag{87}$$

where $\theta = \{\alpha, \lambda\}$ is the minimal sufficient statistic.

4. Moving to Clustering Validation

Transferring the previous results to the clustering validation problem requires some considerations. Clustering denotes the task of assigning each item of $\mathbf{X}$ to a subset $l_k \in \mathcal{Y} = \{l_1, \ldots, l_K\}$ or *cluster* based on the geometric or probabilistic properties of vectors $\mathbf{x}_i$. Formally, it is a map [4]:

$$\begin{aligned} h : \mathbf{x_i} &\longrightarrow \mathcal{Y} \qquad (\mathbf{x}_i \in \mathbf{X}) \\ h(\mathbf{x}_i) &\mapsto l_k, \end{aligned} \tag{88}$$

This map provides the assignment $\{(\mathbf{x}_1, l_k), \ldots, (\mathbf{x}_m, l_{k'})\}$. The evaluation of the quality of this assignment in a probabilistic clustering model is stated in terms of the minimization of a loss function, usually the log-likelihood; in particular, the parameters that maximize the log-likelihood are those that provide the best clustering result. It can be immediately observed that (43) is the inverse of the non-logarithmic likelihood. From this point, we obtain the expectation of the limit, and the result is a pdf.

The existence of a pdf allows for the construction of acceptance regions, and its utilization—with or without disjoint subsets—allows for the division of the original dataset. This fact allows for the definition of acceptance regions for the distributional parameters. This problem, which is classic in the field of statistics, is usually known as hypothesis testing. It is stated as a set of parameter values $\theta \in \Theta_0$, for which certain values Θ_0 constitute

the null hypothesis or acceptance region. Its complementary in Θ is Θ_1, constituting the rejection of the null hypothesis. In particular, the acceptance region $\mathcal{R}$ is

$$\mathcal{R} = P(\theta \in \Theta_0) \tag{89}$$

$$= \left\{\theta \text{ s.t. } \theta \in f_{\theta_0 \in \Theta}(x; \theta_0)\right\} \tag{90}$$

and, for a sample, dependence on the data $\theta = \theta(x)$ leads to

$$\mathcal{C} = \left\{x \text{ s.t. } f_{\theta_0 \in \Theta}(x; \theta_0) \geq \alpha\right\}, \tag{91}$$

providing the probability α level confidence interval $(\theta_1^{-1}(x), \theta_2^{-1}(x))$, with those values being the boundaries of $\mathcal{C}$. Then, the confidence level is

$$1 - \beta = \int_{\theta_1^{-1}(x)}^{\theta_2^{-1}(x)} f(x; \theta) d\, x. \tag{92}$$

In the literature, β is usually referred to as α. We use this notation to avoid confusion with the parameters of the gamma pdf. A detailed pedagogical exposition of these concepts, containing many examples, has been presented in [57]. This text deals with point estimation, the construction of hypothesis tests, and the construction of confidence intervals, emphasizing the equivalence between them in a standard statistical exposition.

When taking into account that only positive integers make sense in the clustering problem, Formula (66) should be rewritten as $\varphi(x; \alpha, \lambda) = \mathbf{J}/\Gamma(\alpha)\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}$, with $\mathbf{J}$ being a suitable column matrix. We follow the statistical convention to omit the one-dimensional basis.

Moreover, it is important that the normalization of the data matrix must be carried out in the same space. As a consequence of smoothing via columns, the norm $\| \widehat{\mathbf{Y}}' \widehat{\mathbf{Y}} \| = n_r$, where n_r is the rank of the image of matrix $\mathbf{X}$ of (1). To achieve accurate results, it is necessary to transform to the same representation space and the following:

$$\|\mathbf{Y}'\mathbf{Y}\|_1 = \frac{1}{n_r} \left\| \widehat{\mathbf{Y}}' \widehat{\mathbf{Y}} \right\|_1. \tag{93}$$

This results in a re-estimation of sequence (38):

$$z^{\star}_{[k]} = \sqrt{n_r}\, z_{[k]}. \tag{94}$$

This nuance has no theoretical effects since the limit of the sequence (43) is also a gamma, but it causes bias in the estimation of parameters.

On the other hand, the clustering validation problem makes sense for positive integers. The solution needs to be re-scaled for transformation to this space, for which we take the following:

$$c_r = \frac{z}{1 - z} \sqrt{m n_r}. \tag{95}$$

These results allow the evaluation of the number of clusters for any data frame that can be written in matrix form, with the only restriction being that the observational variables are linearly independent. This makes it possible to evaluate data structures that simultaneously contain continuous and categorical variables and do not require distributional assumptions or their independence on the parameters.

4.1. Computational Remarks

Several parameters affecting the reliability of the gamma density are required. To better clarify their effect, we divide the process into three phases. First, we (i) obtain the

sequence of traces, (ii) evaluate the limit sequence, and (iii) adjust the gamma density, for which it is necessary to reproduce the shape—especially the maximum and a sufficient number of values for the queue, which is the value of the ancillary statistic (k).

4.1.1. NMF Parameters

The first step is to obtain an NMF with random initialization (otherwise, it would not make sense), varying k from 1 to the ancillary statistic in increasing order. For each factorization, the sequence (43) is computed. This process is repeated, and the results are kept in a matrix. In this phase, the parameters of matrix $\mathbf{X}$ are also obtained, including the matrix dimension, the matrix rank, and z. Algorithm 1 provides details of this process.

As the computational cost required to obtain a reliable approximation of (19) is high, an alternative approach for the trace sequence is to relax the approximation degree and re-compute the terms with q random re-initializations of the matrices in formulas (25) and (26) (otherwise, the re-estimation would not make sense). Then, by selecting the statistic $z^\star$ such that $E\, z^\star = z$, we obtain an estimate sequence.

The convergence condition of (27) is somewhat complex in practice. A tight condition provides good values when $k > \min(m, n)$, but it causes problems if $k < \min(m, n)$, and the iterative process may not stop. Conversely, an overly wide condition provides no representative values. To achieve reasonable results, it is sufficient to take a value of $\epsilon \sim O(1/m)$. Furthermore, deciding the rank of the matrix is not a trivial problem. This is a classical situation illustrating a well-solved problem in theory that faces problems in practice. This issue is intensified by transformations to the probabilistic space. We use SVD, selecting the significant eigenvalues for this situation.

Algorithm 1: Data Parameters and Trace Estimation.

Input: Data Matrix $\mathbf{X}$
Approximation condition ϵ
Number of model components k
Number of re-estimations q

$(\mathbf{A})$ Obtain Data Matrix Parameters:
1: Dimension of $\mathbf{X}$: m and n;
2: Transform $\mathbf{X}$ to $\mathbf{Y}$ (probabilistic image);
3: Compute trace of $\mathbf{Y}$;
4: Obtain a full range matrix $\mathbf{Y}_r$ from $\mathbf{Y}$;
5: Obtain n_r (number of linear independent columns of matrix $\mathbf{Y}_r$);
6: Define Variables z_k (vector of estimations $z_1, z_2, \ldots, z_k$);
7: Compute correlation $\det(\mathbf{C})$ (matrix correlations determinant);;
8: Estimate Overlap ϑ

for *each* $q = 1, \ldots, q$ (re-estimations) **do**
  for *each* $k = 1, 2, \ldots$ (construction of sequence) **do**
    1: Random initialize $\mathbf{W}$ of dimension $m \times k$ and $\mathbf{H}$ of dimension $k \times n$;
    2: Factorize $\mathbf{Y}$ obtaining $\mathbf{W}$ and $\mathbf{H}$
    3: Compute the trace $z_k = \mathrm{tr}(\mathbf{WH})$;
  if $\widehat{\mathbf{Y}}$ *is column normalized* **then**
    $z_k = \sqrt{n_r}\, z_k$
  in each row of $\mathbf{Z}$ put z_k

Output: $\mathbf{Z}$, n_r, $\vartheta, \det(\mathbf{C})$

4.1.2. Sequence of Traces

This step involves selecting an estimator for the matrix containing the trace sequences, which is calculated for each value of k. When the estimator is computed, we handle the results using local likelihood regression with the help of the *sm R* package [58]. The selected smooth parameter is $h = 1$, which has no effect on the data. It is also necessary to define a support (*grid*) that contains the positive integers valued at those same points, as detailed in Algorithm 2.

Algorithm 2: Trace Sequence.

Input: $\mathbf{Z}$
n_r
1: support $= \left(z/(1-z)\right) \times \sqrt{mn_r} \times$ support
for *each column of* $\mathbf{Z}$ **do**
  $\mathbf{s} = \left(1/q\right) \sum_k \mathbf{Z}$
2: Assign pairs $S = (\text{support}, \mathbf{s})$
3: $S = \text{normalize}(S)$
Output: S

4.1.3. Gamma Density Adjustment

In this step, we estimate the peak of (67) according to the criteria discussed in Section 3.4.3 with no further problems. Algorithm 3 details this simple procedure.

Algorithm 3: Gamma Parameters.

Input: S
1: $\alpha = \max(S) + 1$
2: $\lambda = 1 - 1/\alpha$
Output: α,λ (parameters of gamma pdf)

4.1.4. Overlapping

The application of the previous theoretical results for the clustering validation is based solely on the hypothesis of the linear independence of the observational variables. However, this situation may be very different from that which occurs in practical situations.

Assuming that the probabilistic image has been obtained from a matrix for which its columns are linearly independent (the information contained in the sample is preserved by omitting the linearly dependent observational variables), the existence of overlap plays an important role in obscuring the estimation of the parameters of (66), even in the case where the hypotheses required for its derivation are fulfilled. This is because the overlap between variables affects their number and the contained information. In the case of uncorrelated variables, this effect is dramatic. Let us consider the case of the mixture $f(x) = \lambda\, g_1(x) + (1-\lambda) g_2(x)$ with strong overlap. In this case, $g_2(x) \approx g_1(x)$; thus, $f(x) \approx g_1(x)$. This drawback can be corrected by introducing the overlapping effect as a factor.

Furthermore, in multivariate structures, the correction must consider the structure of the distribution of the information: it can be within variables, between variables, or both. This means that considering $\text{sign}(\alpha - n_r)$, which provides information on the distributional behavior of the information (in the case where the number of non-overlapping informative variables is less than α, the information is mainly contained within each variable; otherwise, the existence of more variables provides the main source of information between them). This suggests the introduction of the following factor:

$$\delta = \begin{cases} n_r \vartheta^{\,\text{sign}(\alpha - n_r)} & \text{if } \det(\mathbf{C}) = 1 \\ \vartheta^{\,\text{sign}(\alpha - n_r)} & \text{otherwise} \end{cases} \tag{96}$$

where $\text{sign}(\alpha - n_r) = 1$ if $\alpha - n_r \geq 0$, $\text{sign}(\alpha - n_r) = -1$ if $\alpha - n_r < 0$, ϑ is the maximum overlapping, and $\det(\mathbf{C})$ is the determinant of the correlation matrix. It is noted that $\det(\mathbf{C}) = \mathbf{1}$ indicates non-correlation, while smaller values correspond to correlation. This correction is also described in Algorithm 4.

Algorithm 4: Overlap and Correlation.

Input: α
$\det \mathbf{C}$
ϑ
n_r
if $\vartheta > \vartheta_0$ **then**
  if $\det \mathbf{C} = \mathbf{1}$ **then**
    $\delta = nr\vartheta$
  else
    $\det \mathbf{C} < \mathbf{1}$
  $\delta = \sqrt{n_r - 1}\,\vartheta$
$\alpha^\star = \delta\alpha$
$\lambda^\star = 1 - 1/\alpha^\star$
Output: $\alpha^\star$,$\lambda^\star$ (parameters of gamma pdf)

On the other hand, we indicate that this correction requires the choice of a value ϑ_0 for the overlap parameter, and the choice of the distributional criteria is an open question.

5. Examples

The examples have the purpose of evaluating the pros and cons of the proposed method. Before performing experiments, we provide a classical toy example related to the classification of documents.

Example 1. *For a co-occurrence data frame* $\mathbf{X}$ *consisting of a corpus of seven (*$d1$*–*$d7$*) documents containing letters* $\{a, b, c, d, e, f\}$*, which we assimilate into words in a thesaurus, the probability image* $\mathbf{Y}$ *is as follows:*

$$\mathbf{X} = \begin{array}{c} \\ d1 \\ d2 \\ d3 \\ d4 \\ d5 \\ d6 \\ d7 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{pmatrix} 4 & 3 & 4 & 0 & 0 & 0 \\ 5 & 3 & 3 & 0 & 0 & 0 \\ 4 & 3 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 8 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 & 4 \\ 0 & 0 & 0 & 0 & 4 & 3 \end{pmatrix} \end{array} \qquad \mathbf{Y} = \begin{bmatrix} 0.06 & 0.05 & 0.06 & 0.00 & 0.00 & 0.00 \\ 0.08 & 0.05 & 0.05 & 0.00 & 0.00 & 0.00 \\ 0.06 & 0.05 & 0.05 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.14 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.12 & 0.03 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.05 & 0.06 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.06 & 0.05 \end{bmatrix}$$

The objective is to find the suitable number of subjects that classifies the documents, with the number of subjects being unknown.

Simple visual inspection of the data frame $\mathbf{X}$ *indicates three or four clusters. We first perform NMF clustering for the choices of* $k = 3$*.*

The results for $k = 3$ *are as follows:*

$$\mathbf{W} = \begin{bmatrix} 0.34 & 0.00 & 0.00 \\ 0.34 & 0.00 & 0.00 \\ 0.31 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.48 \\ 0.00 & 0.02 & 0.52 \\ 0.00 & 0.49 & 0.00 \\ 0.00 & 0.49 & 0.00 \end{bmatrix} \quad \mathbf{W}_{docs} = \begin{bmatrix} d2 & - & - \\ d1 & - & - \\ d3 & - & - \\ - & - & d4 \\ - & d5 & d5 \\ - & d6 & - \\ - & d7 & - \end{bmatrix} \quad \mathbf{W}_{prob} = \begin{bmatrix} d2 & d6 & d4 \\ d1 & d7 & d5 \\ d3 & d5 & - \\ - & - & - \\ - & - & - \\ - & - & - \\ - & - & - \end{bmatrix}$$

The matrix $\mathbf{W}$ *represents the probabilities of documents vs. model components or clusters, and they are assimilated to the matrix's columns. Once obtained, the qualitative matrix* $\mathbf{W}_{docs}$ *is a probabilistic classification per cluster. In each column, lines are the probabilities close to zero. This*

matrix is usually written as lists in decreasing order, and it is a probabilistic classification shown in matrix $\mathbf{W}_{prob}$.

Figure 3 shows the gamma pdf that provides credibility for the model components (or the number of clusters) used to achieve the NMF.

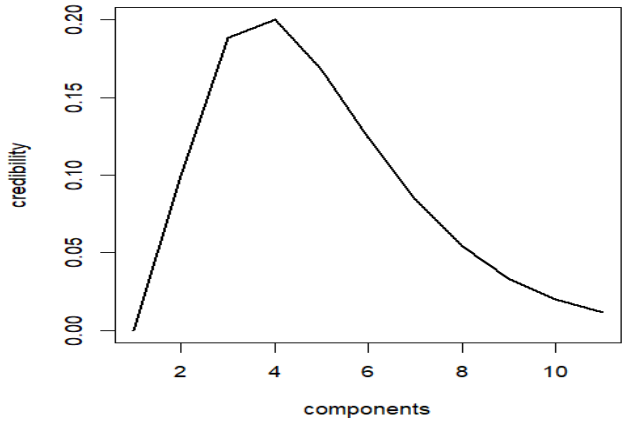


Figure 3. Gamma pdf credibility.

The parameters for the gamma pdf in Figure 3 are $\alpha = 3$ *and* $\lambda = 0.67$, *providing an acceptance hypothesis test of* $\beta = 0.90$ *for the confidence* β*-level acceptance interval of* $[3, 4]$.

From the matrix $\mathbf{W}$, *it is possible to obtain a disjoint classification through the introduction of a Bayes classifier [43]. For the Bayes classifier* $\mathbf{w}_i = max_{w_i \in W}(\mathbf{W})$, *the qualitative matrix containing the documents classified according its weight in the span space are*

$$\mathbf{W}_{Bayes} = \begin{bmatrix} d1 & - & - \\ d2 & - & - \\ d3 & - & - \\ - & - & d4 \\ - & - & d5 \\ - & d6 & - \\ - & d7 & - \end{bmatrix}.$$

This procedure illustrates the equivalence of NMF and k-means clustering. Despite NMF clustering being related to probabilistic and k-means clustering, the construction of the gamma pdf has no hypothesis that goes beyond the construction of a statistic, and it does not depend on the clustering method.

To examine the examples, we also compared the proposed criterion with indices that provide good results: the Dunn and silhouette indices and the gap statistic. These indices were chosen due to their widespread use in the related literature and the good results they provide: notably, these indices provide a single value. We took the maximum of (66) as $\kappa = \arg\max f(x; \alpha, \lambda)$ to make comparisons possible. We also considered the number of labels or distributions for each dataset or the distributions that generate them.

The Dunn index is an ancient internal evaluation index for identifying compact sets of clusters, with a small variance within them [59]. The silhouette index evaluates the compactness and separation of clusters [25]. The underlying idea of the gap statistic is to obtain the expected differences in the k clusters, with each of them having r elements; then, they are compared with a reference distribution [29]. These indices assume the optimization of a loss function related to the clustering method. We assume a k-means underlying partition schema.

We aid with the *clValid*-package of the R computing environment to obtain the Dunn and silhouette indices [60]. For the gap statistic, we use the *factoextra* package [61] with a brute force recursive procedure consisting of a loop varying the number of clusters.

We conducted experiments in two contexts. The first involved illustrative examples, and we selected four examples from well-known datasets in the UCI repository (https://archive.ics.uci.edu/datasets, accessed on 2 September 2023), which is currently of a reduced size. Using these examples, we illustrate the validation ability of the proposed criterion. Meanwhile, the other examples possessed more complex data structures (first used in [20], and it is available at http://cs.joensuu.fi/sipu/datasets, last accession 2 September 2023). We examined the validation criterion against overlapping, dimension variation, and different numbers of observations per cluster.

5.1. Simple Examples

The four datasets from the UCI repository are briefly described in Table 1. The selection criterion was the non-existence of missing values in order to avoid the need for a pre-processing step.

Table 1. Overview of the main features of selected datasets. The sub-categories of the dataset glass correspond to the presence of additional features. The size refers to the number of rows (m) and columns (n) of the data matrix. n_r is the rank of the probabilistic image, $\det(\mathbf{C})$ is the determinant of the correlation matrix, and ϑ is the maximum overlapping.

Dataset	Size ($m \times n$)	Labels	Sub-Categories	n_r	$\det(\mathbf{C})$	ϑ
iris	150×5	3	no	2	0.42	0.66
seeds	210×8	3	no	2	0.67	0.11
ecoli	336×9	6	no	4	0.53	0.71
glass	214×19	7	yes	3	0.00	0.60

The *iris* dataset is one of the most popular and best-studied datasets, and it has been used in many statistical studies. It contains three species of labeled flowers. Its original use can be attributed to Fisher. A discussion of how it was obtained and the conditions under which various authors attribute different results (2–3 categories, while some provide 4) can be found in [62]. The *ecoli* dataset was proposed to illustrate the unsupervised classification of biological functions [63]. The dataset *seeds* was first used to analyze several clustering algorithms on a real dataset of seed images [64]. The *glass* dataset allows for the study of the identification of seven types of glass and the effects of certain additives. Figure 4 provides a graphical description of the datasets.

Table 2 shows the results obtained for the considered validation indices. It corresponds to $k = 1, \ldots, 10$ clusters. The value κ refers to the $\arg\max\ f(x; \alpha, \lambda)$ of Formula (66).

The results for the glass (a) dataset assume the selection of the first *seven* columns of the data frame. It was found that *glass* contains *seven* classes of glasses, identifying them with the labels $1, \ldots, 7$. In addition, there are additional treatments consisting of additives. Despite the quantities not being the same, it is easy to identify them (columns 8 and 9 of the data frame) as binary variables. Table 3 presents both results for this dataset, and in Figure 5, we present the case alone. All results obtained with our validation criteria were similar to those of the gap statistic; however, the Dunn and silhouette index and the gap statistic did not capture the presence of the binary variable. This illustrates, from a practical point of view, an advantage of our less restrictive assumption.

A comparison is presented in Table 3. Figure 5 shows the behavior of the pdf compared to the non-parametric curve obtained with Equation (43), in which the relation with both maximums can be seen.

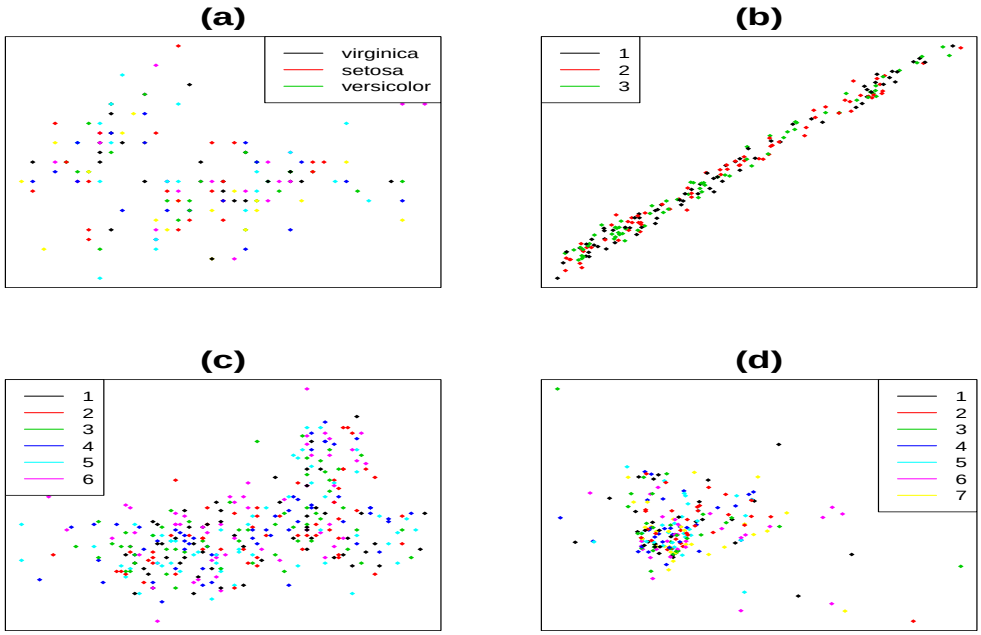


Figure 4. Biplots (first two columns) of each dataset: (**a**) iris; (**b**) seeds; (**c**) ecoli; (**d**) glass.

Table 2. Results for Dunn (minimize), silhouette (maximize), and gap statistic (maximize), values of the non-parametric sequence of (69) (maximize), and gamma pdf versus the number of clusters (maximize). The dataset glass (a) omits thermal treatments, but they are included in (b). The maximum for the gap statistic in cases (a) and (b) was 14, which does not appear in the table. The gamma pdf and sequence of (43) for case (b) were 14, which also do not appear in the table.

Dataset	Validation Index	Clusters								
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
iris	Dunn	0.077	0.099	0.137	0.082	0.085	0.087	0.087	0.062	0.068
	silhouette	0.582	0.460	0.419	0.354	0.345	0.315	0.346	0.318	0.318
	gap	0.473	0.495	0.442	0.405	0.455	0.450	0.466	0.479	0.473
	sequence	0.131	0.225	0.191	0.129	0.094	0.070	0.055	0.043	0.340
	gamma	0.228	0.234	0.180	0.124	0.079	0.049	0.029	0.017	0.010
seeds	Dunn	0.055	0.086	0.049	0.056	0.044	0.078	0.084	0.089	0.089
	silhouette	0.466	0.404	0.321	0.321	0.275	0.253	0.284	0.244	0.250
	gap	0.312	0.415	0.395	0.367	0.374	0.373	0.361	0.361	0.356
	sequence	0.048	0.219	0.207	0.138	0.092	0.066	0.050	0.039	0.031
	gamma	0.152	0.184	0.167	0.135	0.103	0.075	0.053	0.037	0.025
ecoli	Dunn	0.080	0.087	0.041	0.065	0.072	0.071	0.053	0.086	0.085
	silhouette	0.347	0.396	0.338	0.341	0.249	0.244	0.223	0.242	0.221
	gap	0.553	0.639	0.621	0.633	0.635	0.637	0.621	0.637	0.628
	sequence	0.030	0.060	0.090	0.120	0.105	0.089	0.073	0.057	0.053
	gamma	0.006	0.044	0.100	0.142	0.156	0.146	0.121	0.093	0.067
glass (a)	Dunn	0.190	0.089	0.167	0.154	0.137	0.154	0.055	0.055	0.061
	silhouette	0.388	0.421	0.447	0.432	0.328	0.300	0.246	0.373	0.246
	gap	0.929	0.945	0.990	0.997	0.069	0.111	0.118	0.152	0.154
	sequence	0.025	0.051	0.076	0.102	0.127	0.152	0.138	0.124	0.110
	gamma	0.005	0.026	0.063	0.102	0.129	0.138	0.130	0.112	0.089
glass (b)	Dunn	0.142	0.085	0.133	0.132	0.131	0.054	0.054	0.054	0.039
	silhouette	0.448	0.367	0.391	0.298	0.329	0.218	0.243	0.342	0.227
	gap	0.957	1.009	1.000	1.053	1.069	1.085	1.090	1.094	1.300
	sequence	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.105	0.118
	gamma	0.000	0.000	0.000	0.000	0.001	0.003	0.008	0.019	0.034

Table 3. Comparison of results. Overview of the results obtained with the used indices. Dataset glass (a) corresponds to the selection of the independent identically distributed (i.i.d.) variables, while dataset glass (b) introduces binary variables. These binary variables are captured by the gamma pdf and the gap statistic. $\kappa = \arg\max f(x; \alpha, \lambda)$.

Dataset	Labels	Dunn	Silhouette	Gap	κ
iris	3	4	2	3	3
seeds	3	9	2	3	3
ecoli	6	3	3	2	6
glass (a)	7	2	4	5	7
glass (b)	7	2	2	14	14

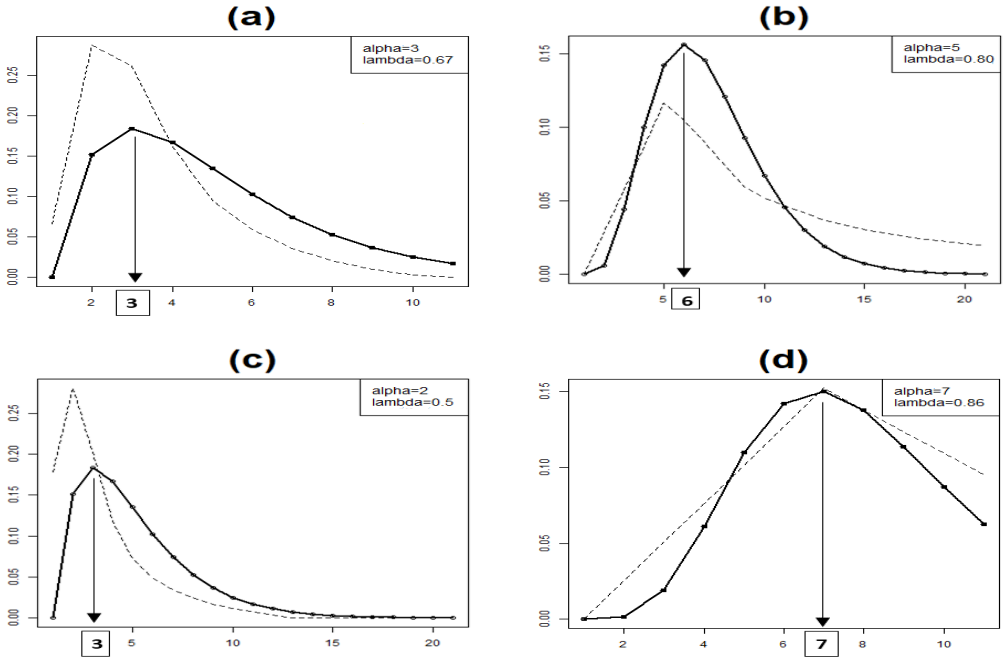


Figure 5. Graphical representation of the credibility of the number of clusters for the UCI data sets iris (**a**), seeds (**b**), ecoli (**c**), and glass (**d**). Dotted line corresponds to sequence (43) multiplied by the factor c defined in the Formula (53). The continuous line is the gamma pdf of (66). Multiplication by factor c leads to represent the sequence (43) in the same support as the gamma pdf. The maximum credibility is the maxima of each curve and can be used to compare the results obtained with other indices. Values of α and λ are according to the results mentioned in Sections 3.4.3 and 4.

Figure 6 shows the NMF clustering with the iris dataset. This dataset is useful to illustrate the independence of our validation criterion and the clustering method. To obtain the density of Equation (46), we set $p = 3$ iterations in the switching process to adjust (34) and $q = 20$ re-estimates, while to obtain the classification of Figure 6, we needed $O(10^5)$ iterations. This shows that the underlying matrices of (40) are not necessary a plausible clustering and cannot be used to obtain the corresponding classification (except for the case where a good approximation of (30) is reached). This is due to the different speeds of convergence of the adjustment process to obtain the factorization and the sequence $z_{[k]}$.

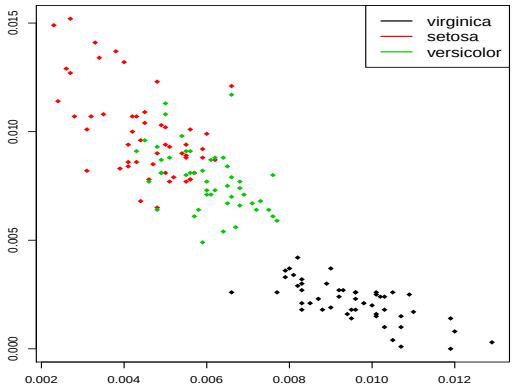


Figure 6. Iris dataset clusterization. Axes are probabilities and have no geometrical significance. Also, they are not orthogonal, and the angles are the dot product of the first two column vectors of the matrix **H**.

5.2. More Complex Data Structures

To show the effects of overlapping, dimension variation, and different numbers of observations per cluster, we selected several synthetic datasets created in [20] to study validation criteria behavior (available at http://cs.joensuu.fi/sipu/datasets, last accession on 2 September 2023). These datasets were generated using a certain number of distributions, in which the overlaps and number of generating distributions vary.

The datasets include four types of artificial data configurations in which the overlapping effect increases (datasets $s1$, $s2$, $s3$, and $s4$), the number of clusters varies (datasets $a1$, $a2$, and $a3$), and the dimension varies (datasets $dim064$, $dim128$, $dim256$, $dim512$, and $dim1024$). These datasets are balanced (i.e., equal number of entities in each cluster). An *unbalanced* dataset allows us to examine the effect of different numbers of observations per cluster.

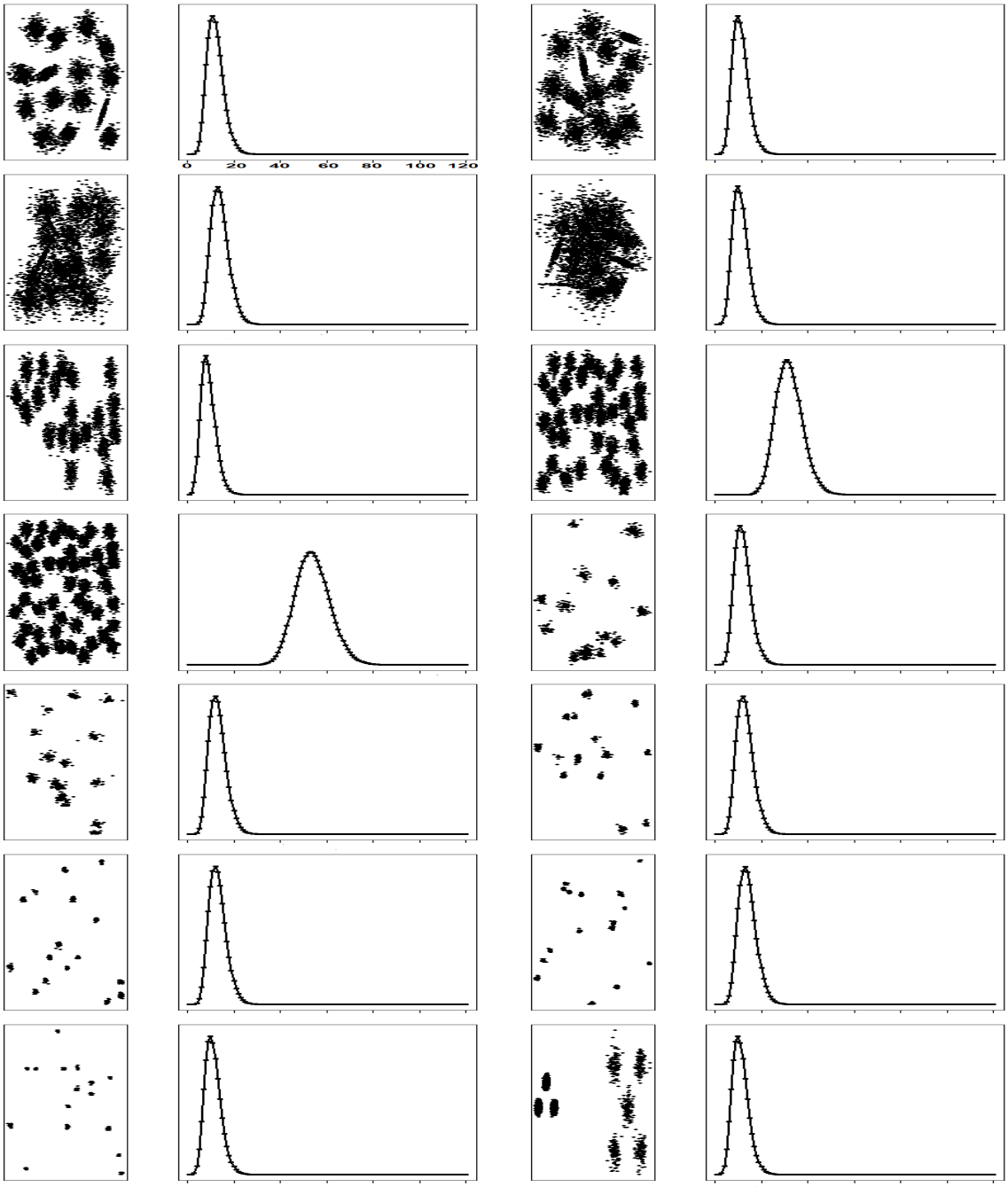
To proceed with the examples, we selected the same parameters for each case: the number of component estimations $k = 1, \ldots, 120$, $q = 200$ re-estimations, no condition on the degree of approximation, and $p = 100$ iterations in the process of switching Equations (25) and (26). To determine the rank of the matrix, we considered the number of relevant eigenvalues of matrix **Y** with the help of the *condition number* (i.e., the quotient of the largest involved eigenvalue).

Running Algorithms 1–3, we obtain the parameter α. These values are acceptable if the overlapping between observational variables does not exists. In the case of the existence of overlapping, it is necessary to adjust them according Formula (96), providing new values for $\alpha^\star$ in Table 4.

Table 4. Results of the execution of Algorithms 1–4. The values of α correspond to the output provided by Algorithm 3. The overlapping and correlation parameters (ϑ and $det(\mathbf{C})$, respectively) are obtained from the probabilistic image **Y** in Algorithm 1. The elevation factor δ according Formula (94) provides the new parameter $\alpha^\star$. For overlapping, we consider the values $\vartheta \geq 0.7$.

	a1	**a2**	**a3**	**a4**	**s1**	**s2**	**s3**	**d32**	**d64**	**d128**	**d256**	**d512**	**d1024**	**unb.**
α	20	22	15	9	8	29	30	11	12	12	13	13	8	8
n_r	2	2	2	2	2	2	2	9	10	10	12	12	13	2
$\det(\mathbf{C})$	0.998	0.982	0.886	0.995	0.853	1	1	0.356	0.427	0.686	0.558	0.453	0.011	0.982
ϑ	0.74	0.70	0.78	0.89	0.34	0.40	0.81	0.69	0.74	0.74	0.84	0.81	0.84	0.17
$\alpha^\star$	15	15	12	8	8	29	48	16	16	16	16	16	16	8

Figure 7 shows the biplots of the data and the corresponding distributions for estimating the credibility of the number of clusters according to the proposed criteria.



	a1	a2	a3	a4	s1	s2	s3	d32	d64	d128	d256	d512	d 1024	unb.
α	14	14	11	7	7	28	47	15	15	15	15	15	15	7
λ	0.934	0.934	0.914	0.864	0.864	0.964	0.984	0.934	0.934	0.934	0.934	0.934	0.934	0.86

Figure 7. Graphical results. The left panel is the biplot (representation of first two dimensions) in the XY plane for each dataset. Right panels show the clustering credibility: Solid lines are densities according to Formula (66), points are the positive-integer values, and dashed lines represent the sequence given by Formula (43). Corresponding gamma parameters are provided at the bottom. All figures shares the xame x-axis.

5.3. Comparative

Comparisons illustrate the pros and cons of the proposed method versus the Dunn, silhouette, and gap statistic indices. In all cases, the most suitable value is chosen to minimize the corresponding loss function. The number of re-sampling steps in the bootstrap procedure used to estimate the reference density for determining the gap statistic was

$B = 100$. Those cases corresponded to the default parameters of the used R packages. Additionally, as in the previous examples, we took the value of the maximum of the gamma pdf for comparison with the other indices, with the results summarized in Table 5.

Table 5. Comparative. The values m and n refer to the size of the data frame (rows and columns, respectively). D is the number of clusters used to generate the datasets (provided by Franti in [20]). Values of the silhouette index and gap statistic are given in the respective columns. κ denotes the results for the maximum of the gamma pdf according to (66).

Data Set (Size)	D	Dunn	Silhouette	Gap	κ
a1 (5000 × 2)	15	15	15	19	15
a2 (5000 × 2)	15	15	15	18	15
a3 (5000 × 2)	15	52	17	17	12
a4 (5000 × 2)	15	68	16	15	8
s1 (3000 × 29	25	21	20	53	8
s2 (5250 × 2)	35	44	35	37	29
s3 (7500 × 2)	50	65	50	57	48
dim32 (1024 × 32)	16	16	16	16	16
dim64 (1024 × 64)	16	16	16	16	15
dim128 (1024 × 128)	16	16	16	16	16
dim256 (1024 × 256)	16	16	16	16	16
dim512 (1024 × 512)	16	16	16	16	18
dim1024 (1024 × 1024)	16	16	16	16	16
unbalance (6500 × 2)	8	2	2	15	8

A quantitative comparison of the results can be carried out using the Adjusted Rand Index (ARI), defined as [65]

$$\mathrm{ARI} = \frac{k - E(k)}{\max k - E(k)}, \tag{97}$$

with k being the number of clusters and $E(k)$ denoting its expectation, which is a measure of the similarity of clustering results. To avoid negative values, and to be consistent with the L_1 metric in the context of non-negativity, we define the *relative error* in the L_1 space as:

$$\mathrm{Rel}_{L_1} = \frac{|\, k - E(k)|}{\max k - \min k} \tag{98}$$

with $\max(k)$ and $\min(k)$ being the maximal and minimal optimal number of clusters provided by a method, respectively. This index provides values over the interval $[0, 1]$, and values close to zero are preferable. The results are shown in Table 6.

On the other hand, we provide values of the $1 - \beta$ levels of κ, assuming that the reported number of densities (D) is the correct value. These examples show that the proposed validation criteria do not assume any partition or fuzzy scheme. It evaluates the number of clusters and provides no classification or clusterization.

It can be obeserved that the proposed procedure captured overlapping, providing underestimation in these cases and exhibiting stability when the dimension increased (compared with D). Therefore, the procedure works well in the unbalanced case. A discussion of the computational cost to obtain the NMF solutions is shown in [50], and some tricks are explained in [45].

Also, while the examples show the behavior of the proposed validation method and open the door to statistical inference, they do not constitute a numerical confirmation. Performing confirmatory numerical experiments would require a study that should include real-world examples including other recent validation indices, such as the Wemmert–Gancarski index [66], SpecialK [67], and Stadion [18], which should be used to establish more rich conclusions.

Table 6. Comparative of clustering validation methods. The Rel_{L1} is computed according Formula (97) for Dunn, silhouette, gap, and κ. The $1-\beta$ value assumes that the expectation of the number of clusters is D, and it is computed for the cases such that $\kappa \neq D$. The acronym *ns* means that the value of $1-\beta$ is not significant (exactly zero).

Dataset	$E(x)$	Dunn	Silhouette	Gap	κ	$1-\beta$
iris	3	0.500	0.500	0.000	0.000	*ns*
seeds	3	0.857	0.143	0.000	0.000	*ns*
ecoli	6	0.750	0.750	1.000	0.000	*ns*
glass (a)	7	1.000	0.600	0.400	0.000	*ns*
glass (b)	14	1.000	1.000	0.000	0.000	*ns*
a1	15	0.000	0.000	1.000	0.000	*ns*
a2	15	0.000	0.000	1.000	0.000	*ns*
a3	15	0.925	0.050	0.050	0.075	0.10
a4	15	0.883	0.017	0.000	0.117	0.30
s1	25	0.089	0.111	0.622	0.378	$O\,(10^{-7})$
s2	35	0.474	0.526	0.105	0.316	0.09
s3	50	0.082	0.000	0.412	0.118	0.30
dim32	16	0.000	0.000	0.000	0.000	*ns*
dim64	16	0.000	0.000	0.000	0.000	*ns*
dim128	16	0.000	0.000	0.000	0.000	*ns*
dim256	16	0.000	0.000	0.000	0.000	*ns*
dim512	16	0.000	0.000	0.000	0.000	*ns*
dim1024	16	0.000	0.000	0.000	0.000	*ns*
unbalance	8	0.462	0.462	0.538	0.000	*ns*

6. Discussion

Usually, the gamma density is obtained as a sum of exponentials [41] (p. 179). It is possible to prove that the independence of sums and sums of squares implies that the coefficient of variation follows this distribution [68,69], which is also true for the harmonic mean of the posterior distribution [70]. Recently, [71] focused on reducing the variance in such cases. Our approach aims to evaluate the expectation of a trace sequence in the probabilistic space. In fact, the Laplace transform confirms this interpretation. The same result can be reached by taking derivatives and carrying out normalization (56); however, we minimized the statistical assumptions and procedures needed to achieve this result. Additionally, we present an approach for applications in the context of clustering validations.

Our approach does not result in any hypotheses in terms of the space of parameters—that is, in neither the distribution nor the dimensionality of the parameters—and, therefore, constitutes a non-parametric approach. One of the advantages of such an approach is that it supports any type of variable. The unique hypothesis, in this sense, is the independence of observational variables. Conversely, the transformation to the probabilistic space and use of the KL divergence provides a maximum-likelihood estimate. However, some issues need to be highlighted.

First, it seems that the main drawback of this method results from the NMF approach being a slow iterative process with high computational costs. We alleviated this through the introduction of several random re-estimations. Another problem is that the gamma pdf is continuous on the compact set $[0, +\infty)$, while the clustering problem is appropriate only for non-negative integers. Additionally, the type of smoothing used in this method implicitly assumes Euclidean distances, while the approximation problem requires the L_1 norm. Although some works have focused on this issue [72,73], according to our literature search, such works have received little attention in recent years. We recall that this choice is critical to reproduce the data structure and is related to the variance.

Although we think that, from a practical viewpoint, the provided examples do not validate our clustering validation criteria or our attempts, they do provide a practical

framework that illustrates several issues. Our proposal worked well for the UCI Machine Learning Repository datasets, but we also selected some of the synthetic datasets developed by Franti to illustrate certain problems. We made this choice as the effects of superposition, stability, and size could be controlled. In this sense, we believe that the greatest difficulties arise when attempting to establish conditions to adjust the model to cases in which compactness is high (datasets $a4$, $s1$, and $s1$).

We did not apply our proposed procedure to other data structures, such as those with well-defined geometric shapes, for which available methods such as CLARANS [74] offer good results and have been extended in the context of Big Data, although we think that this would allow for good exploration. However, according to the experiments of Section 5, our results are comparable to those provided by the selected indices.

Moreover, the provision of a pdf provides several advantages. In unsupervised environments, a graph can be easily interpreted by human operators without analytical skills, allowing them to incorporate other results in the analysis as part of their expert judgment. Expert judgment does not always coincide with the statistical results, leading to controversial situations. Additionally, expert judgment comes from many years of studying a discipline or practice, and it should not be misconsidered. We believe that controversial situations should be placed in the area of the selection of relevant observational variables [75].

Without entering into this discussion—which is philosophical and profound—we merely indicate that graphic visualization can provide relevant considerations for practical situations, and a contribution of the proposed method is the visualization of densities, thereby providing relevant graphical information.

Future research should include our proposal in a more exhaustive comparison, including artificial datasets that reproduce shapes. The effect of superposition is another window to explore.

7. Conclusions

Although inference on clustering is controversial, a pdf was built from the sequence of traces obtained with NMF techniques, the construction of which requires no assumptions beyond linearly independent uncorrelated observational variables. Thus, by transforming observations to the probabilistic space, with the expectation provided by the limit of the distribution on the sequence of traces, varying the dimensions of the space span provides a gamma density. This is the main result of this manuscript.

This result allows us to assign credence to clustering results regardless of the method used. To carry this out, we have established an error bound for the approximation error between the matrix of observations in the probabilistic space and the approximate factorization in the case of low-rank approximation.

Our proposal allows non-skilled humans to visualize the results in a fully unsupervised validation environment, achieved with a single plot of the adjusted gamma density. Additionally, in the context of Big Data and Computer Engineering, an interval of plausible estimations seems more advantageous. This result allows discussions in quantitative terms between different validation results. In practical situations, it allows the verification of whether the selection of observational or experimental criteria is correct.

Author Contributions: Conceptualization, methodology, validation, formal analysis, investigation, writing—original draft preparation, writing—review and editing, P.F.; supervision and project administration, A.C. and P.G.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: https://archive.ics.uci.edu/datasets (accessed on 21 March 2024) (Section 5.1 examples); http://cs.joensuu.fi/sipu/datasets (accessed on 21 March 2024) (Section 5.2 examples).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 27 December 1965–7 January 1966; pp. 281–297.
2. Jain Anil, K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **2010**, *31*, 651–666. [CrossRef]
3. Aggarwal, C.C. *Clustering: Algorithms and Applications*; CRC Press Taylor and Francis Group: Boca Raton, FL, USA, 2014.
4. Dougherty, E.R.; Brun, M. A probabilistic theory of clustering. *Pattern Recognit.* **2004**, *37*, 917–925. [CrossRef]
5. Deng, H.; Han, J. Probabilistic models for clustering. In *Data Clustering*; CRC: Boca Raton, FL, USA, 2018; pp. 61–86.
6. Paatero, P.; Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **1994**, *5*, 111–126. [CrossRef]
7. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791. [CrossRef] [PubMed]
8. Chen, J.C. The nonnegative rank factorizations of nonnegative matrices. *Linear Algebra Its Appl.* **1984**, *62*, 207–217. [CrossRef]
9. Brualdi, R.A.; Parter, S.V.; Schneider, H. The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *J. Math. Anal. Appl.* **1966**, *16*, 31–50. [CrossRef]
10. Guillamet, D.; Bressan, M.; Vitria, J. A weighted non-negative matrix factorization for local representations. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 15 April 2001; Volume 1, pp. I–I.
11. Brunet, J.P.; Tamayo, P.; Golub, T.R.; Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4164–4169. [CrossRef]
12. Benetos, E.; Kotti, M.; Kotropoulos, C. Applying supervised classifiers based on non-negative matrix factorization to musical instrument classification. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 12 July 2006; pp. 2105–2108.
13. Wang, Y.; Jia, Y.; Hu, C.; Turk, M. Non-negative matrix factorization framework for face recognition. *Int. J. Pattern Recognit. Artif. Intell.* **2005**, *19*, 495–511. [CrossRef]
14. Li, H.; Adal, T.; Wang, W.; Emge, D.; Cichocki, A.; Cichocki, A. Non-negative matrix factorization with orthogonality constraints and its application to Raman spectroscopy. *J. Vlsi Signal Process. Syst. Signal, Image, Video Technol.* **2007**, *48*, 83–97. [CrossRef]
15. Bholowalia, P.; Kumar, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. *Int. J. Comput. Appl.* **2014**, *105*, 17–24.
16. Ben-Hur, A.; Elisseeff, A.; Guyon, I. A stability based method for discovering structure in clustered data. In *Biocomputing 2002*; World Scientific: Singapore, 2001; pp. 6–17.
17. Von Luxburg, U. Clustering stability: An overview. *Found. Trends Mach. Learn.* **2010**, *2*, 235–274.
18. Mourer, A.; Forest, F.; Lebbah, M.; Azzag, H.; Lacaille, J. Selecting the number of clusters k with a stability trade-off: An internal validation criterion. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Osaka, Japan, 28 May 2023; pp. 210–222.
19. Cramer, D.; Howitt, D.L. *The Sage Dictionary of Statistics: A Practical Resource for Students in the Social Sciences*; Sage: Thousand Oaks, CA, USA, 2004.
20. Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets. *Appl. Intell.* **2018**, *48*, 4743–4759. [CrossRef]
21. Everitt, B.S. Unresolved problems in cluster analysis. *Biometrics* **1979**, *36*, 169–181. [CrossRef]
22. Dubes, R.; Jain, A.K. Validity studies in clustering methodologies. *Pattern Recognit.* **1979**, *11*, 235–254. [CrossRef]
23. Dubes, R.C. How many clusters are best?-an experiment. *Pattern Recognit.* **1987**, *20*, 645–663. [CrossRef]
24. Hardy, A. On the number of clusters. *Comput. Stat. Data Anal.* **1996**, *23*, 83–96. [CrossRef]
25. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
26. Har-Even, M.; Brailovsky, V.L. Probabilistic validation approach for clustering. *Pattern Recognit. Lett.* **1995**, *16*, 1189–1196. [CrossRef]
27. Smyth, P. Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.* **2000**, *10*, 63–72. [CrossRef]
28. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On clustering validation techniques. *J. Intell. Inf. Syst.* **2001**, *17*, 107–145. [CrossRef]
29. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2001**, *63*, 411–423. [CrossRef]
30. Pallis, G.; Angelis, L.; Vakali, A.; Pokorny, J. A probabilistic validation algorithm for web users' clusters. In Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), Hague, The Netherlands, 10–13 October 2004; Volume 5, pp. 4129–4134.
31. Brun, M.; Sima, C.; Hua, J.; Lowey, J.; Carroll, B.; Suh, E.; Dougherty, E.R. Model-based evaluation of clustering validation measures. *Pattern Recognit.* **2007**, *40*, 807–824. [CrossRef]
32. Fred, A.L.; Jain, A.K. Cluster validation using a probabilistic attributed graph. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
33. Žalik, K.R.; Žalik, B. Validity index for clusters of different sizes and densities. *Pattern Recognit. Lett.* **2011**, *32*, 221–234. [CrossRef]
34. Olivares, J.; Sarro, L.; Bouy, H.; Miret-Roig, N.; Casamiquela, L.; Galli, P.; Berihuete, A.; Tarricq, Y. Kalkayotl: A cluster distance inference code. *Astron. Astrophys.* **2020**, *644*, A7. [CrossRef]

35. Shi, C.; Wei, B.; Wei, S.; Wang, W.; Liu, H.; Liu, J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip J. Wirel. Commun. Netw.* **2021**, *2021*, 1–16. [CrossRef]
36. Usefi, H. Clustering, multicollinearity, and singular vectors. *Comput. Stat. Data Anal.* **2022**, *173*, 107523. [CrossRef]
37. Ullmann, T.; Hennig, C.; Boulesteix, A.L. Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1444. [CrossRef]
38. Modak, D.S. Evaluation of the number of clusters in a data set using p-values from multiple tests of hypotheses. *Commun.-Stat.-Theory Methods* **2024**, 1. [CrossRef]
39. Bowman, A.W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*; OUP Oxford: Oxford, UK, 1997; Volume 18.
40. Chacón, J.E.; Duong, T. *Multivariate Kernel Smoothing and Its Applications*; CRC Press: Boca Raton, FL, USA, 2018.
41. Balakrishnan, N.; Nevzorov, V.B. *A Primer on Statistical Distributions*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
42. Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* **2001**, *42*, 177–196. [CrossRef]
43. Ding, C.; Li, T.; Peng, W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.* **2008**, *52*, 3913–3927. [CrossRef]
44. Mnih, A.; Salakhutdinov, R.R. Probabilistic matrix factorization. *Adv. Neural Inf. Process. Syst.* **2007**, *20*.
45. Figuera, P.; García Bringas, P. Revisiting Probabilistic Latent Semantic Analysis: Extensions, Challenges and Insights. *Technologies* **2024**, *12*, 5. [CrossRef]
46. Senga Kiesse, T.; Cuny, H. Discrete triangular associated kernel and bandwidth choices in semiparametric estimation for count data. *J. Stat. Comput. Simul.* **2014**, *84*, 1813–1829. [CrossRef]
47. Kokonendji, C.; Senga Kiesse, T.; Zocchi, S.S. Discrete triangular distributions and non-parametric estimation for probability mass function. *J. Nonparametr. Stat.* **2007**, *19*, 241–254. [CrossRef]
48. Cichocki, A.; Zdunek, R.; Phan, A.H.; Amari, S.i. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*; Wiley: Hoboken, NJ, USA, 2009.
49. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
50. Figuera, P.; García Bringas, P. On the Probabilistic Latent Semantic Analysis Generalization as the Singular Value Decomposition Probabilistic Image. *J. Stat. Theory Appl.* **2020**, *19*, 286–296. [CrossRef]
51. Ho, N.-D.; Van Dooren, P. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra Appl.* **2008**, *429*, 1020–1025. [CrossRef]
52. Dempster, A.; Laird, N.; Rubin, D. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Stat. Soc.* **1977**, *39*, 1–38. [CrossRef]
53. Amari, S.I. Information geometry of the EM and em algorithms for neural networks. *Neural Netw.* **1995**, *8*, 1379–1408. [CrossRef]
54. Gupta, M.D. Additive non-negative matrix factorization for missing data. *arXiv* **2010**, arXiv:1007.0380v1.
55. Schmidt, E. Zur Theorie der linearen und nichtlinearen Integralgleichungen. In *Integralgleichungen und Gleichungen Mit Unendlich Vielen Unbekannten*; Springer: Berlin/Heidelberg, Germany, 1989; pp. 190–233.
56. Zhang, X.D. *Matrix Analysis and Applications*; Cambridge University Press: Cambridge, MA, USA, 2017.
57. Casella, G.; Berger, R.L. *Statistical Inference*; Cengage Learning: Boston, MA, USA, 2021.
58. Bowman, A.W.; Azzalini, A. *R Package sm: Nonparametric Smoothing Methods (Version 2.2-6.0)*; University of Glasgow: Glasgow, UK; Università di Padova: Padova, Italy, 2024.
59. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
60. Brock, G.; Pihur, V.; Datta, S.; Datta, S. clValid: An R Package for Cluster Validation. *J. Stat. Softw.* **2008**, *25*, 1–22. [CrossRef]
61. Kassambara, A.; Mundt, F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses, 2019. R package version 1.0.6. Available online: https://CRAN.R-project.org/package=factoextra (accessed on 21 March 2024).
62. Unwin, A.; Kleinman, K. The iris data set: In search of the source of virginica. *Significance* **2021**, *18*, 26–29. [CrossRef]
63. Nakai, K.; Kanehisa, M. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **1992**, *14*, 897–911. [CrossRef] [PubMed]
64. Charytanowicz, M.; Niewczas, J.; Kulczycki, P.; Kowalski, P.A.; Łukasik, S.; Żak, S. Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images. In *Information Technologies in Biomedicine*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 2, pp. 15–24.
65. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
66. Desgraupes, B. *Clustering Indices*; University of Paris Ouest-Lab Modal'X: Paris, France, 2013; Volume 1, p. 34.
67. Hess, S.; Duivesteijn, W. k is the magic number—inferring the number of clusters through nonparametric concentration inequalities. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, 16–20 September 2019; pp. 257–273.
68. Hwang, T.Y.; Hu, C.Y. On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation. *Ann. Inst. Stat. Math.* **1999**, *51*, 749–753. [CrossRef]
69. Hwang, T.Y.; Huang, P.H. On new moment estimation of parameters of the gamma distribution using its characterization. *Ann. Inst. Stat. Math.* **2002**, *54*, 840–847. [CrossRef]

70. Raftery, A.E.; Newton, M.A.; Satagopan, J.M.; Krivitsky, P.N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. 2006. Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series. Working Paper 6. Available online: https://biostats.bepress.com/mskccbiostat/paper6 (accessed on 21 March 2024).
71. McEwen, J.D.; Wallis, C.G.; Price, M.A.; Docherty, M.M. Machine learning assisted Bayesian model comparison: Learnt harmonic mean estimator. *arXiv* **2021**, arXiv:2111.12720.
72. Stone, C.J. Consistent nonparametric regression. *Ann. Stat.* **1977**, *5*, 595–620. [CrossRef]
73. Hall, P.; Wand, M. Minimizing L1 distance in nonparametric density estimation. *J. Multivar. Anal.* **1988**, *26*, 59–88. [CrossRef]
74. Ng, R.T.; Han, J. CLARANS: A method for clustering objects for spatial data mining. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 1003–1016. [CrossRef]
75. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]