Applied Mathematics and Computation 256 (2015) 591–601

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Clusterability assessment for Gaussian mixture models

Ewa Nowakowska [a,*], Jacek Koronacki [a], Stan Lipovetsky [b]

[a] Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
[b] GfK Custom Research North America, Marketing & Data Sciences, 8401 Golden Valley Rd., Minneapolis, MN 55427, USA

ARTICLE INFO

Keywords:
Clusterability
Gaussian mixture models
Fisher's discriminant
Principal component analysis

ABSTRACT

There are numerous measures designed to evaluate quality of a given data grouping in terms of its distinctness and between-cluster separation. However, there seems to be no efficient method to assess distinctness of the intrinsic structure within data (clusterability) before actual clustering is determined. Based on recent findings, we propose such measure in terms of covariance matrix decomposition for appropriately transformed data. The data is assumed to come from a Gaussian mixture model. The transformation reshapes the data so that unsupervised technique of principal component analysis is able to uncover information directly indicative of the data clusterability characteristics. In this work we propose the measure and explain the motivation as well as the relation to supervised structure distinctness coefficients. We also show how the measure can be applied for number of clusters and feature selection tasks.

1. Introduction

1.1. State-of-the-art

Literature on clusterability and related distinctness assessment problems suffers from inconsistent and ambiguous terminology. What can be currently found under the term of clusterability is gathered in [1]. The indices analyzed there though, are meant to assess a given data partition, while the term clusterability should rather be confined to measures that are partition-independent. It should refer to assessing the possibility to cluster efficiently rather than to quality evaluation of a previously determined solution. In what follows, we will follow this distinction and refer to measures based on certain partition as structure distinctness coefficients. In other words, for structure distinctness coefficients we will require the model parameters to be known or estimated. The term clusterability we will use exclusively to refer to measures that do not need this kind of information and therefore can operate on raw data, without the knowledge of partition in particular. Based on some of the indices of [1], partition-independent measures can be obtained, as it was done for instance in [2]. Somewhat unfortunately, this heuristic method tends to fail in actual applications when the number of dimensions increases. More formal inspiration for clusterability analysis is provided by considerations on the number of clusters (main references include [3–6]) and component overlap analysis for mixtures of normal distributions (see [7–10]). However, both approaches originally assume the underlying partition to be already determined.

Direct inspiration for this work comes from a series of works on learning mixture parameters in a selected subspace. It started with the challenge of random projections considered first in one-dimensional space – by Kalai et al. [11] for two and

* Corresponding author.
E-mail addresses: ewa.nowakowska@ipipan.waw.pl (E. Nowakowska), jacek.koronacki@ipipan.waw.pl (J. Koronacki), stan.lipovetsky@gfk.com (S. Lipovetsky).

http://dx.doi.org/10.1016/j.amc.2014.12.038
0096-3003/

Moitra and Valiant [12] for arbitrary number of clusters. Then [13] suggested random projections to substantially lower potentially more than one-dimensional subspace based on Johnson–Lindenstrauss (concentration) theorem. In [14] some of the distributional assumptions were relaxed but as the concentration theorem was still used, the assumption of high initial cluster separation had to be maintained. This was only relaxed by Brand and Huang [15], where random projections were replaced with spectral approach. These results were further applied and developed in [16–18]. The key insight for this work comes from [19], where a preliminary data transformation was used to enhance the unknown structure in data in order to improve performance of a parameter learning algorithm. This proved that it is possible to use the structure in data without actually knowing it and inspired [20], which in turn became the basis for the clusterability assessment method proposed in this work.

1.2. Model and notation

We assume that the data $X = (x_1, \ldots, x_n)^T,\ X \in \mathbb{R}^{n\times d}$ – consisting of n observations – comes from a mixture of k d-dimensional normal distributions

$$f(x) = \pi_1 f_1(\mu_1, \Sigma_1)(x) + \cdots + \pi_k f_k(\mu_k, \Sigma_k)(x),$$

where

$$f_l(\mu_l, \Sigma_l)(x) = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det \Sigma_l}} e^{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)}.$$

We refer to each $f_l(\mu_l, \Sigma_l),\ l = 1, \ldots, k$ as a component of the mixture (see [21] or [22] for model details and [23] or [24] for example alternatives). We call $\pi_l,\ l = 1, \ldots, k$ a mixing factor of the corresponding component. We assume equal mixing factors for all the components $\pi_1 = \cdots = \pi_k = \frac{1}{k}$, however we allow different covariance matrices Σ_l. Additionally we assume the space dimension to be large with respect to the number of components $d > k - 1$. We also assume the number of observations is large with respect to d or k, that is $n \gg d$. We take the number of components k as known, which puts no constrains on our considerations they can easily be repeated for all k within the range of interest.

We assume the cluster centers to be independent in terms of point independence. Also, we assume covariance matrix to be of full rank, $\Sigma_X = d$. Let $T_X = n\Sigma_X$ be the total scatter matrix for X. We say that data is in *isotropic position* if $\mu_X = \mathbf{0}$ and $T_X = \mathbf{I}$. We recall the known fact that $T_X = W_X + B_X$, which decomposes the total scatter to its within (W_X) and between (B_X) cluster components. For notation ease we will often assume the data is centered and the origin and indicate that with a zero subscript X_0 or Z_0.

A cluster – or a class – is understood as a subset of observations that are similar (close in the space) to one another but different (far in the space) from other observations. A grouping that divides observations into clusters is called a *cluster solution* or a *cluster structure*. To set up a link between the theoretical model and the data, we assume that each mixture component corresponds to one cluster.

By $PC(k-1)$ we denote the subspace spanned by the first $k-1$ principal components (i.e. $k-1$ eigenvectors of the matrix Σ_X corresponding to its $k-1$ largest eigenvalues). By S^* we denote the *Fisher's discriminant (Fisher's subspace)*, which is a $(k-1)$-dimensional subspace that best discriminates k given classes as

$$S^* = \underset{\substack{S\subset\mathbb{R}^d \\ \dim(S)=k-1}}{\operatorname{argmax}} \frac{\sum_{j=1}^{k-1} v_j^T B_X v_j}{\sum_{j=1}^{k-1} v_j^T T_X v_j}, \tag{1}$$

where $v_1, \ldots, v_{k-1}$ is the orthonormal basis for S. Details are given in [25], while the concept is discussed in [21]. In the course of the work we will use the known result that S^* may be represented as a solution of an eigenproblem defined by $T_X^{-1}B_X$. The proof under the assumed model can be found in [26].

Following [20], we define *structure distinctness coefficient* in terms of Fisher's discriminant as

$$\bar{\lambda}^X = \frac{1}{k-1}\sum_{j=1}^{k-1} \lambda_j^{T_X^{-1}B_X}, \tag{2}$$

which is the average eigenvalue over $k-1$ largest eigenvalues of the $T_X^{-1}B_X$ eigenproblem and the mean variability in the Fisher's subspace at the same time. As [26] shows, (2) well reflects the behavior of the generic integral overlap measure (also referred to as MLE-misclassification rate) while possessing the advantage of being analytically tractable at the same time. It allows it to be included in formal derivations and the analysis of the system behavior.

1.3. Concept and related results

The idea of clusterability assessment proposed here is based on the results obtained and presented in [20]. In this work a data transformation is derived that preserves structure distinctness – as defined by (2) – up to a negligible error. The transformation consists of two steps – isotropic transformation and weighting. The first step de-correlates the data – it can easily be shown (see for instance [20]) that the following transformation brings the data to isotropic position so $\mu_Y = \mathbf{0}$ and $T_Y = \mathbf{I}$

$$Y = X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}}, \tag{3}$$

where $T_{X_0} = A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T$ is the eigenvalue decomposition of the data matrix X_0, centered at the origin. The second step is designed to introduce slight perturbation and bring the directions of largest dispersion close to the directions of best between cluster separation. Namely, it is meant to reduce variability in all the directions but the ones that are determined by the cluster centers. Following [20] the weighting is performed as $Z = \mathrm{diag}(\omega)Y$, where weights are defined as

$$\omega_i = \sqrt{\frac{1}{1 + \frac{1}{\alpha}\|y_i\|^2}} \tag{4}$$

and the parameter α is typically set to $\alpha = 0.5$. It is shown in [20] that the transformation makes the principal components approximate the directions of best discrimination and consequently brings $PC(k-1)$ close to S^*. It is also proved analytically in [20] that

$$|\bar{\lambda}^Z - \bar{\lambda}^X| \leqslant \frac{1}{\sqrt{n}}\left(\frac{d}{\alpha}\left(\bar{\lambda}^X + \sqrt{k}\right)\right). \tag{5}$$

Since the sample size n is assumed to be large with respect to the number of dimensions d and the number of clusters k, the resulting value of the upper bound in (5) is very small, which indicates that the transformation preserves structure distinctness up to negligible error. Altogether, it is expected that the structure distinctness defined by (2) in terms of Fisher's discriminant can be approximated – by means of PCA – with the eigenvalues of the covariance matrix $\mathbf{\Sigma}_Z$ for the transformed data Z. And as such it does not require the knowledge of cluster assignments to perform clusterability assessment. The details of the data transformation as well as the results regarding its properties are described in [20].

Note though, that even if after the transformation the two spaces – S^* and $PC(k-1)$ – coincide and the projection to $PC(k-1)$ discriminates the unknown classes best, the partition is still unknown. As such, clusterability coefficient defined in terms of variability in $PC(k-1)$ cannot constitute an estimate of structure distinctness (2) in a formal sense. However, due to specific properties of the proposed data transformation, we are able show evident relation between these two quantities.

Note also that the approximation would not be possible without the data transformation. The concepts of Fisher subspace S^* and principal component subspace $PC(k-1)$ are independent in their essence and in a standard setup there is no direct correspondence between them. In particular, a standard projection to $PC(k-1)$ does not have to take into account the cluster structure existing in data and as such a simple projection to $PC(k-1)$ is not expected to preserve cluster structure or any characteristics of it. Hence the data transformation that establishes this link and preserves the structure distinctness is inevitable to make it feasible to approximate concepts defined in terms of Fisher discriminant by means of principal component analysis.

It would be most desirable to project the data to S^* as it is by definition the subspace that discriminates the groups best. However, it is infeasible in practice as S^* is defined by cluster structure which is unknown before the learning task is performed. Therefore the objective is to approximate S^* with $PC(k-1)$, as the latter does not require the information on the clustering structure and as such can be easily constructed for a data set with unknown clustering. Let us emphasize here, that we assume clusters (classes) to be known, which is inevitable to examine the measure's properties and its relation to structure distinctness coefficient (2). However, the ultimate algorithm, operates on raw data only and does not require the knowledge of cluster belongings. Note also, that when speaking of motivation we use theoretical concepts at population level, however the actual calculations are made for given data, i.e. at sample level.

The content of this work is organized as follows. Section 2 provides the definition and motivation for the clusterability coefficient, it also analyzes its relation with structure distinctness and studies the impact of data parameters on the coefficient value, it concludes with the clusterability assessment algorithm. In Section 3 examples of applications for feature and number of clusters selection are given. Finally Section 4 summarizes the findings and points to possible directions of future development.

2. Clusterability coefficient

2.1. Definition

The clusterability coefficient as proposed here is based on variability in $PC(k-1)$ subspace for the transformed data referred to as Z. It equals average variance in $PC(k-1)$ subspace, which corresponds to the mean eigenvalue over the first $k-1$ largest eigenvalues of covariance matrix decomposition. As data is initially standardized with isotropic transformation and then only slightly perturbed with weighting, additional standardization is redundant. What is more, for not standardized Z data the effect of weighting is also captured by the eigenvalues of covariance matrix, which is highly desirable. The raw clusterability coefficient is then expressed by the following formula

$$\bar{\gamma}^{Z_0} = \frac{1}{k-1}\sum_{l=1}^{k-1}\gamma_l^{Z_0}, \tag{6}$$

where $\gamma_l^{Z_0} \in \mathbb{R}$ for $l = 1, \ldots, d$ are eigenvalues of covariance matrix $\frac{1}{n}T_{Z_0}$, enumerated in non-increasing order. We refer to (6) as raw coefficient as it captures the core information, however it will need further transformation to allow for value comparisons between different configurations of model and data parameters (such as k, d or n).

The intuition behind such measure of clusterability is very simple. For data in isotropic position, variability is constant in all the directions. Then, the weighting – which is implicitly structure-dependent – reduces it unevenly to some extent. If the structure is clear, the principal directions – that coincide with directions of cluster centers in this case – remain basically untouched. The reduction takes place in the remaining directions only, thereby leaving the average variability in $PC(k-1)$ fairly high. On the contrary, when the structure is fuzzy, the reduction takes place relatively evenly over all the directions, reducing variance in $PC(k-1)$ to a considerable extent. However, this is only the intuition that explains the choice of clusterability estimate. A more formal study of the measure's performance, using simulated data, is given in the course of this section.

Note, that variability in Fisher's space – the way we define it – is determined by the ratio of B_X and T_X (B_{Z_0} and T_{Z_0}) matrices, while in principal component subspace it is directly proportional to T_X (T_{Z_0}) matrix. As the between-cluster scatter B_X (B_{Z_0}) is unknown at the stage of preliminary analysis, it cannot be taken into account. Hence, there is an objective gap between structure distinctness coefficient and clusterability estimate. The gap sets an upper bound on the possible accuracy of estimation and deterministic relation between one and the other cannot be expected. However, the data transformation is expected to set up an empirical link and allow for approximations. First, the data transformation followed by projection on $PC(k-1)$ reduces the space to its minimal subspace that preserves (almost) full information on the initial unknown structure. As such, it eliminates vast majority of noise and redundancy so that only the meaningful information is left. Second, the isotropic transformation puts every dataset in the same position from PCA perspective. Hence, the variability which is observed for Z_0 data is the consequence of weighting only, which in turn implicitly depends on the underlying structure. As such, although not directly, the estimate is expected to reflect variability in Fisher's subspace, which in turn captures the distinctness of the unknown structure.

2.2. Relation with structure distinctness

In this subsection a simulation study is used to analyze performance of clusterability estimates in terms of its relation with the structure distinctness coefficients. For each data dimension d from 3 to 15, full range of possible cluster numbers was considered unless it exceeded 10. As such, k ranged from 3 to minimum of d and 10. Sample sizes took values of $n = 100, 300$ and 500 per cluster. The data generating algorithm had a nested form, which means wherever possible a smaller data set was a subset of a larger data set to allow for comparisons between the correlation values. Namely, for each data dimension d largest number of clusters $k = d$ and largest number of observations $n = 500$ was generated and then for varying k and n subsets of data were selected. The entire data generation and computation procedure was repeated 100 times to introduce variability in mixture parameter configuration for each choice of d, k and n and allow for analyzing the relation between the structure distinctness and the clusterability coefficients.

We do not expect the values of clusterability estimates be directly comparable to the coefficients of structure distinctness but we do expect the dynamics of their behavior to be similar. To analyze this relation Pearson's correlation coefficient was used. Scatter plots of one against the other confirm adequacy of the assumed linear relation. The correlations seem to be high in spaces of small dimension and drop with increase in d. However, the correlations are still high if small number of clusters k is considered, even if space dimension d is large. The pattern can be seen in Fig. 1(a) and (b).

Another thing that was observed during the simulations is that the relation between structure distinctness and clusterability estimate strengthens with the increase of the former. In other words, the more distinct the actual structure, the more accurately it may be assessed with clusterability estimate. It is an expected property of a rough estimate to perform better when the structure is clear.

2.3. Impact of data parameters.

Note that the correlations discussed in subsection 2.2, are analyzed for each given set of data parameters – the dimension d, the number of clusters k and the sample size (per cluster) n. As such, their values are not directly indicative of the impact the data parameters may have on the clusterability coefficient. In this subsection we show the nature of this impact and propose a transformation that can eliminate it, making it possible to directly compare values of coefficients for data of different parameter configurations. Note also that the transformations that depend only on data parameters, do not affect correlations as analyzed in subsection 2.2, as for each set of parameters their impact reduces to multiplying by a constant.

In Fig. 2 we can see how clusterability values are affected by data parameters. The variability due to data parameters strongly dominates over the variability that is due to distinctness of each separate structure. This way it may possibly hamper the inference based on the comparison of clusterability values for data sets of different data parameters and must be eliminated. Notice, that we use now for comparisons the total sample size N (as opposed to the sample size per cluster n) to separate the impact of the sample size from the impact of k. Except for the strong impact of N, a slight impact of dimensionality d and number of clusters k can also be spotted when comparing the values between the plots, however this will be analyzed later in subsections 3.2 and 3.1 respectively.

To understand and eliminate the impact of the data parameters, we get back to the derivation of the data transformation and the clusterability coefficient. We recall that due to $T_Y = I$, variance in all the directions equals $1/N$, where N denotes total sample size. The weighting transformation can only shrink variance as well as $PC(k-1)$ eigenvalues. Hence, the clusterability coefficient can only take values from $[1, 1/N]$. Therefore, to eliminate the first-stage impact of the sample size, we must multiply the data by $\sqrt{N}$ or, alternatively, the clusterability coefficient by N. Second, let us look at the weighting process. It can be shown that $\overline{\|y_i\|^2} = \frac{d}{N}$, which implies that

$$\overline{\omega} = \sqrt{\frac{1}{1+\frac{d}{\alpha N}}}, \tag{7}$$

the average value of weight, and consequently the shrinkage degree, is determined by the weighting parameter α, data dimension d and the sample size N. Hence, depending on these parameters the eigenvalues of $PC(k-1)$ and the values of clusterability coefficients differ. Again, to eliminate that we multiply the transformed data by the inverse of the reference value or the clusterability coefficient by its square. The resulting transformation of the clusterability coefficient that eliminates the impact of data and weighting parameters takes the form

$$\overline{\gamma}_{adj} = \overline{\gamma} \cdot N \cdot \left(1 + \frac{d}{\alpha N}\right), \tag{8}$$

where γ is defined as in (6). The behavior of (8) is shown in Fig. 3. The individual differences are now clearly detectable, there is also no visible trend with respect to data parameters. The patterns that can be observed are due to the nested form of the data generating algorithm and to natural fluctuations.

Note that both discussed transformations affect the clusterability coefficient only. Being linear, they do not have their impact on the structure distinctness coefficient, which is defined via scale invariant Fisher's task.

To summarize the findings of this section, the following algorithm for clusterability assessment for a given data set X can be proposed.

Algorithm 1. ClusterabilityAssessment (X)

Step 1: Isotropic transformation
$X_0 \leftarrow FX$ **comment:**Data centering
$T_{X_0} \leftarrow A_{T_{X_0}} L_{T_{X_0}} A_{T_{X_0}}^T$ **comment:**Eigenvalue decomposition
$Y \leftarrow X_0 A_{T_{X_0}} L_{T_{X_0}}^{-\frac{1}{2}}$ **comment:**Data decorrelation
Step 2: Weighting
$\alpha \leftarrow 0.5$ **comment:**Weighting parameter selection
$\omega_i = \sqrt{\frac{1}{1+\frac{1}{\alpha}\|y_i\|^2}}$ **comment:**Weights' calculation
$Z \leftarrow \mathrm{diag}(\omega) Y$ **comment:**Data weighting
Step 3: Clusterability asessment
$Z_0 \leftarrow FZ$ **comment:**Data centering
$\frac{1}{n} T_{Z_0} \leftarrow A_{T_{Z_0}} G_{T_{Z_0}} A_{T_{Z_0}}^T$ **comment:**Covariance matrix eigen decomposition
$\bar{\gamma}^{Z_0} \leftarrow \frac{1}{k-1}\sum_{l=1}^{k-1} \gamma_l^{Z_0}$ for $\gamma^{Z_0} = \mathrm{diag}(G_{T_{Z_0}})$ **comment:**Estimate, see (6)
$\bar{\gamma}^{Z_0}_{\mathrm{adj}} \leftarrow \bar{\gamma}^{Z_0} \cdot N \cdot (1 + \frac{d}{\alpha N})$ **comment:**Estimate, see (8)

3. Applications for model selection tasks

3.1. Number of clusters

One of the potential applications of the clusterability coefficient lies in providing insight in the number of clusters. We recall here [6], which provides a statistical procedure that formalizes heuristic elbow criterion for selecting number of clusters. We observe certain similarities with respect to the behavior of clusterability estimates and we use the work as inspiration for drawing conclusions in our setup.

As the clusterability coefficient approximates the distinctness of the structure in data, we could expect its value for the true number of clusters to stand out. This might be true if the structure is very distinct, however in real world applications this is not necessarily so. As it shows in Fig. 4, for a given data set the values of clusterability decrease with the increasing declared number of clusters. The natural decrease is due to the coefficient formula, which is based on variability in $PC(k-1)$ subspace and for k clusters defined as average value over $k-1$ largest eigenvalues of the covariance matrix decomposition. Hence with the increase in k smaller values enter the mean driving the overall decrease. So, to determine the true number of

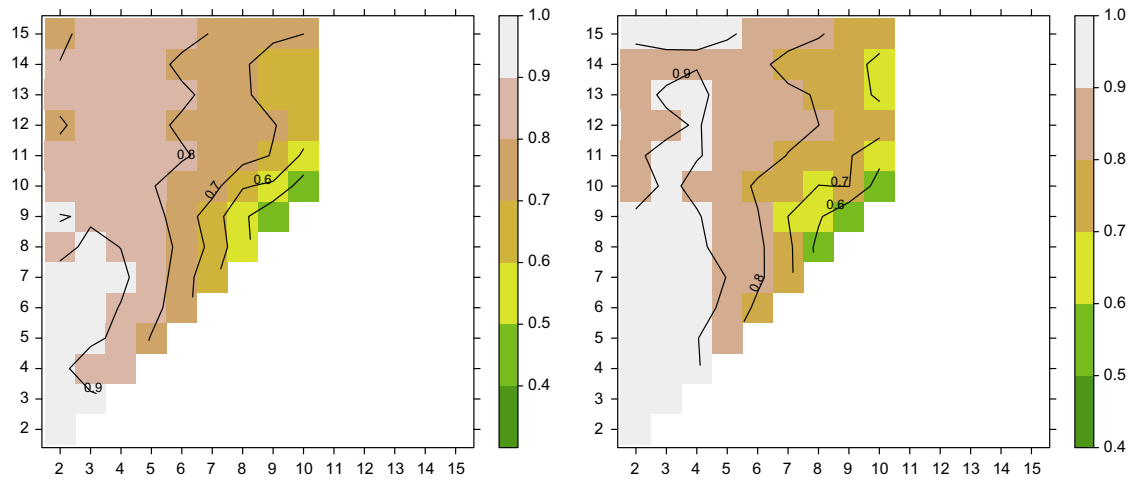

Fig. 1. Topography of correlation between structure distinctness coefficient and its clusterability estimate for given d and k. On x-axis $k = 3, \ldots, \min(d, 10)$, while on y-axis $d = 3, \ldots, 15$.

clusters one has to compare the changes of the values (the pace) rather than the values themselves. Looking at Fig. 4 we can see the decrease is exponential-like until it reaches the true number of clusters, then it is linear, which is an obvious consequence of its origin. Certainly, the difference in the decrease rate is more visible when the clusters are better separated (the pink line). For the green line which reflects the situation of hardly any separation at all the difference in pace is hardly noticeable.

A similar problem – the so called elbow criterion – is described in [6]. The within sum of squares used as the basis for number of clusters estimation is a decreasing function of k and the objective is to determine the point where the decrease rate changes. To that end, the authors utilize a reference distribution, which is a uniform distribution over the same subspace as for the data of interest. In our setup the exact support is of marginal importance as the data is standardized in the process anyway, so for simplicity we can use the theoretical rather than the empirical support. Also, we do not observe a natural log decrease but a linear decrease instead so we cannot repeat the exact transformations, however we can still use the reference distribution to make comparisons with the obtained values.

The behavior of the reference distribution is presented in Fig. 5. Except for the uniform distribution we also show a single component standard Gaussian distribution as a potential alternative. Not surprisingly the values for both reference distributions decrease as linear functions of the number of clusters. Also not surprisingly the values for Gaussian distribution are smaller than for the uniform, which is expected, given the concentration around the mean (as opposed to even spread) does not facilitate artificial partitioning. Note though that although the model assumes Gaussian mixtures for modeling existing structure, when considering data sets without clear structure we can be comparing the data to observations evenly spread over the space (uniform reference distribution) or to concentrated in the center (Gaussian reference distribution).

The data for the experiments was generated according to the same algorithm as in subsection 2.3. Additionally, the separation between the clusters was one of the controlled parameters and varied from 1 (low) to 4 (high) standard deviations on each of the corresponding dimensions. For each configuration of data parameters the data sets were generated 100 times to allow for averaging the observed tendencies and ensure stability of the results.

The experimental design again includes varying dimensions and assumed number of clusters. Although they do not drive additional clusters, they are still included for comparability reasons, as in the simulation plan considered, the theoretical increase in k increases the sample size. We observe that the function for Gaussian reference distribution tends to decrease faster than for the uniform distribution for small sample sizes and smaller numbers of dimensions. In larger dimensions as well as for larger sample sizes it decreases in parallel to the uniform distribution, which is expected given that both factors tend to increase the relative empirical spread, which makes both reference distributions more similar to each other.

In Figs. 6–9 we compare the data clusterability functions with the reference functions. For well separated clusters, we observe the reference function crosses the clusterability function somewhat below the point of optimal number of clusters, which constitutes the cut point and provides a hint on the number of clusters. Note that the reference distribution comes from a different model which is why the same meaning should not be attached to both values and the reference should rather be treated as a threshold. Not surprisingly, for well separated clusters the values for reference Gaussian distribution are overall substantially lower. If the separation in data is low the uniform reference curve starts to lie above the clusterability function as in Fig. 8, while the Gaussian reference curve tends to cut the clusterability curve below the largest recognizable number of clusters in data (Fig. 9). What we observe here is an example confirmed by wider simulations and leads to the

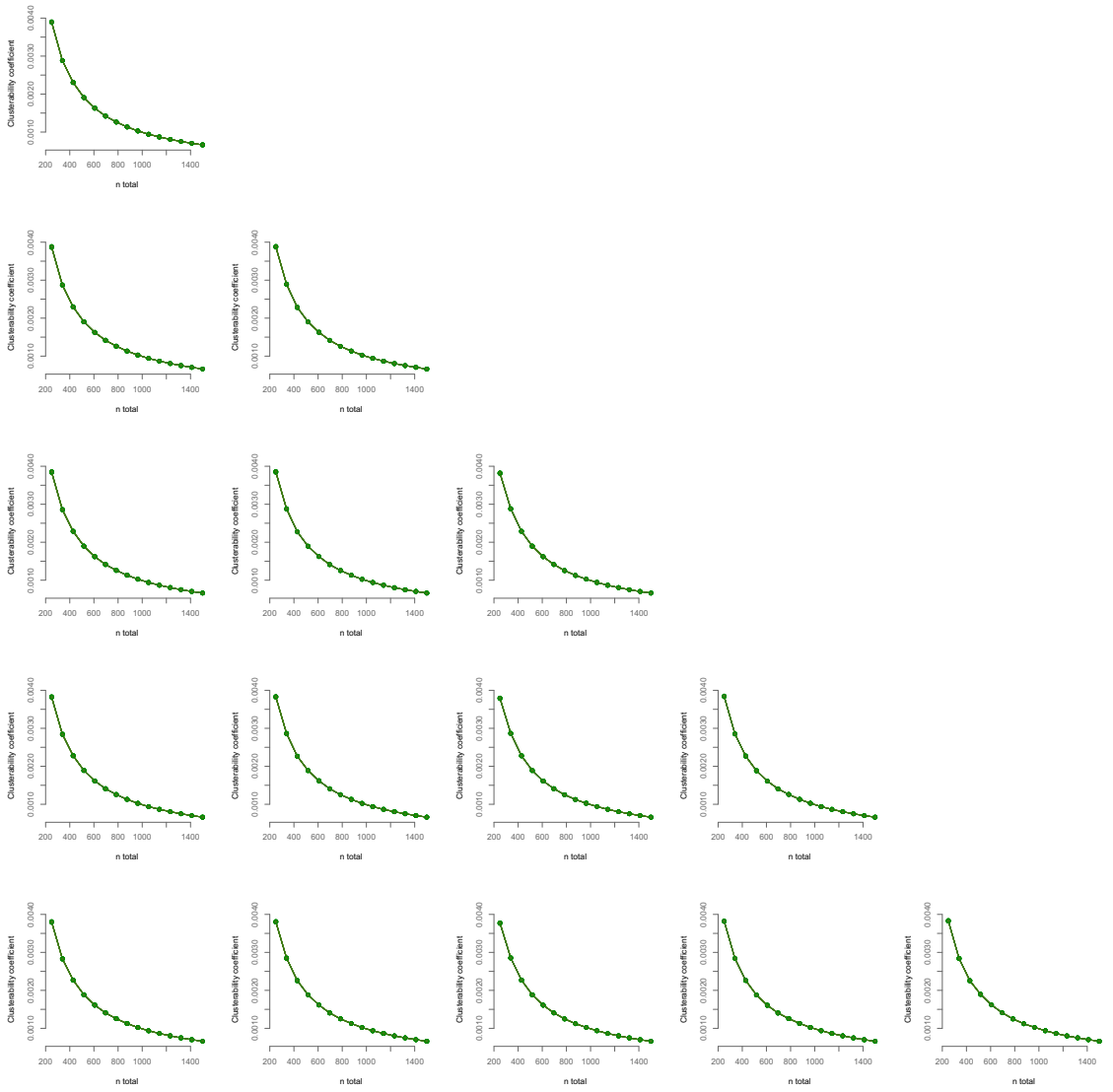

Fig. 2. Clusterability values – increasing **total** sample size on the x-axis. Each row corresponds to fixed dimension $d = 3, \ldots, 7$, while each column to fixed number of clusters $k = 3, \ldots, d$.

following conclusion. The uniform reference distribution based on the same number of observations in the same dimensions as the original data provides a cut point for well separated clusters. Whatever is above this reference curve is separated better than a uniform distribution so the structure is visible in data. Unless too large, one is typically interested in the largest number of visible clusters, so usually the last cluster point above the reference curve is of primary interest. On the other hand, the Gaussian reference curve provides the lower bound for any possible structure. Whatever is below the curve is most likely impossible to cluster in a meaningful way as its distinctness is lower than that of a standard single component mixture. Note that this approach not only allows for a relative comparison of the decreasing values on the clusterability curve but also provides absolute benchmarks that facilitate general assessment of the considered data set. The approach is less formalized than that presented in [6] but it is based on the same assumptions and follows the same reasoning. At the same time it leaves certain degree of flexibility to the researcher allowing for some trade-offs based on the specific objectives and context of the expected segmentation.

3.2. Feature selection

Another practical application of the clusterability coefficient is the task of feature selection. The coefficient can be calculated for multiple subsets of features and based on the values of the clusterability coefficient certain sets might be expected

to better support the clustering task than others and we might want to be able to determine the most distinctive subspace. In this case there is no theoretical reason for the clusterability values to change with the increase in the number of dimensions d as the linear transformation (8) eliminates the impact of d on the ultimate values. However, technically the separability of

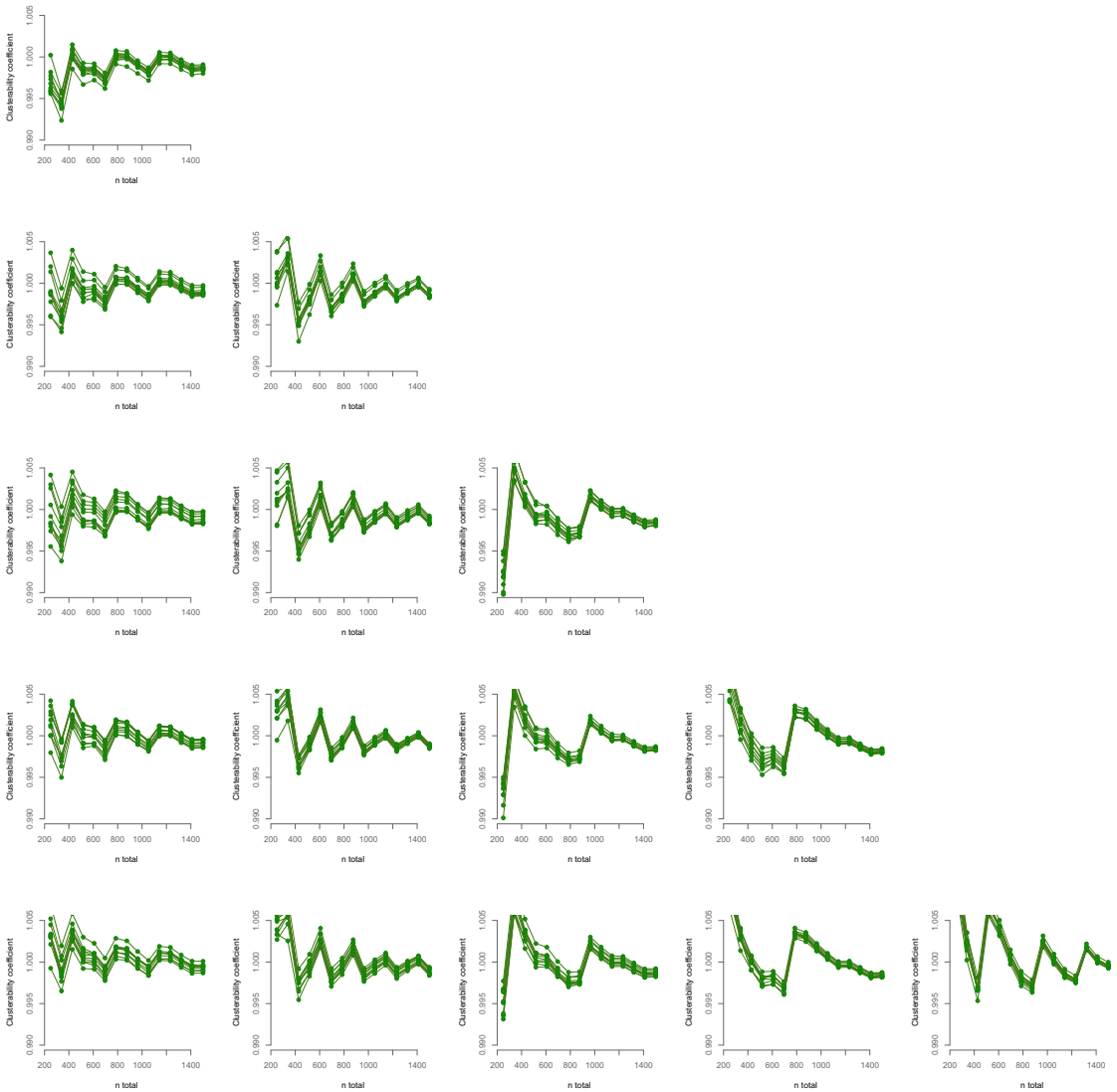

Fig. 3. On y-axis $\overline{\gamma} \cdot N \cdot \left(1 + \frac{d}{2N}\right)$, on x-axis – increasing **total** sample size. Each row corresponds to fixed dimensionality $d = 3, \ldots, 7$, while each column to fixed number of clusters $k = 3, \ldots, d$.

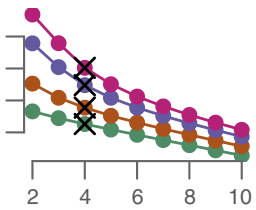

Fig. 4. Clusterability coefficient wrt growing assumed number of clusters from low (green) to high (pink) cluster separation, X marks the true k. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

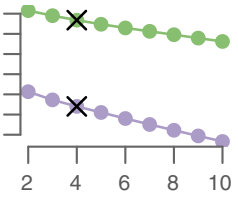

Fig. 5. Behavior of clusterability coefficient wrt growing assumed number of clusters for uniform (green) and Gaussian (purple) reference distributions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

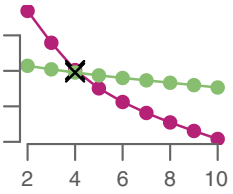

Fig. 6. High separation, uniform reference.

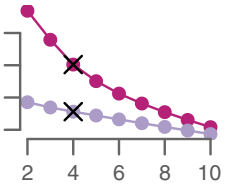

Fig. 7. High separation, Gaussian reference.

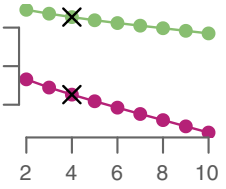

Fig. 8. Low separation, uniform reference.

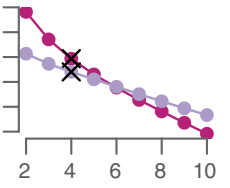

Fig. 9. Low separation, Gaussian reference.

clusters always increases with increase in d even if the new information has only negligible contribution. Hence we do observe slight increase in clusterability as d increases but it is desirable to determine the point beyond which adding further features is not worth the increased model complexity.

There is also a technical problem with typically large number of possible combinations of features to examine. In this very case the experimental design allowed for ordering features based on importance and without the loss of generality the task was reduced to finding the cut off points. In practical applications the features are not sorted and there are multiple subsets to consider, hence the expert knowledge is required at the preprocessing stage to set limits on the number of configurations to analyze.

The design of the simulations allowed for control over differentiating dimensions. Each cluster was differentiated from others based on a subset of attributes. Space dimension permitting, there was a set of features that did not differentiate the clusters more than at a noise level. Without the loss of generality the dimensions were ordered (the differentiating dimension first, then the non-differentiating dimensions) to allow for easy presentation of the summary results. As such, the general task was reduced to detecting the cut off point. Note that one of the assumptions of the underlying model, is the point independence of the cluster centers. This sets a limit on the minimal number of features required to assess clusterability for k assumed clusters as equal to $k-1$. Hence the simulations do not show any results beyond this point. Also, for k clusters independent in terms of point independence, $k-1$ is the minimal number of differentiating dimensions. The simulations were repeated for the previously used range of controlled between cluster separation. Up to those remarks, data generation followed the algorithm of subsection 2.2. For each space dimension d and each number of clusters k, the number of features considered was between $k-1$ and d.

Fig. 10 shows an example of the results for space dimension $d=14$ and $k=4$ clusters, k first features differentiated the clusters in this case. Analogously to the number of cluster selection task, we can see a monotonous increase in values however the pace changes substantially. What is a dramatic increase in clusterability values for the meaningful features turns into a steady almost negligible growth for irrelevant attributes. The change is even more noticeable for well separated clusters, however it is evident even for mixtures of low between cluster separation. Also, the increase induced by irrelevant

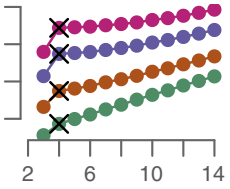

Fig. 10. Clusterability coefficient wrt increasing number of dimensions (X marks the true number of differentiating features), from low (green) to high (pink) cluster separation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

dimensions decreases with the increase in the sample size. This is as well to be expected given the contribution of a single dimension decreases with the increase in the number of dimensions.

4. Conclusions

In this work we motivate and present a clusterability measure, which uses transformed data in order to assess expected distinctness of its intrinsic cluster structure. As opposed to other methods available in the literature, it does not require the knowledge of cluster belongings and operates on raw data only. We show that the (unsupervised) clusterability measure is strongly connected to (supervised) structure distinctness coefficient, which makes the former a meaningful indicator of how distinct the clusters in data might be. Note also that as the data transformation changes the structure distinctness only to a negligible extent it also has a marginal impact on actual clusterability values making it possible to compare original X data based on the assessment results for the transformed data Z. We also show that using reference distributions and analyzing the increase/decrease rate, we can apply the coefficient for number of cluster or feature selection tasks.

Although the method already presents a closed tool, there are still open questions that might make it even more useful or interpretable. One of such questions being a possible rescaling that would allow for more spread in the clusterability values, which are typically concentrated in the upper part of the $[0, 1]$ interval. The analysis of distribution of the clusterability values will probably be one of the future research directions.

Acknowledgments

This work was supported by National Science Center of Poland, Grant No. DEC-2011/01/N/ST6/04174.

References

[1] M. Ackerman, S. Ben-David, Clusterability: a theoretical study, in: N. Lawrence (Ed.), International Conference on Artificial Intelligence and Statistics, JMLR: Workshop and Conference Proceedings, vol. 5, 2009, pp. 1–8. <http://jmlr.org/proceedings/papers/v5/ackerman09a/ackerman09a.pdf>.
[2] S. Epter, M. Krishnamoorthy, M. Zaki, Clusterability detection and initial seed selection in large datasets, Tech. Rep. 99-6, Rensselaer Polytechnic Institute, 1999.
[3] A. Cuevas, M. Febrero, R. Fraiman, Estimating the number of clusters, Can. J. Stat./La Rev. Can. Stat. 28 (2) (2000) 367–382. <http://www.jstor.org/stable/3315985>.
[4] P. McCullagh, J. Yang, How many clusters?, Bayesian Anal 3 (1) (2008) 101–120.
[5] C. Fraley, A.E. Raftery, How many clusters? Which clustering method? Answers via model-based cluster analysis, Comput. J. 41 (8) (1998) 578–588, http://dx.doi.org/10.1093/comjnl/41.8.578. <http://comjnl.oxfordjournals.org/content/41/8/578.abstract>.
[6] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 63 (2) (2001) 411–423, http://dx.doi.org/10.1111/1467-9868.00293. http://dx.doi.org/10.1111/1467-9868.00293.
[7] T.W. Anderson, R.R. Bahadur, Classification into two multivariate normal distributions with different covariance matrices, Ann. Math. Stat. 33 (2) (1962) 420–431, http://dx.doi.org/10.1214/aoms/1177704568. <http://dx.doi.org/10.1214/aoms/1177704568>.
[8] S. Ray, B.G. Lindsay, The topography of multivariate normal mixtures, Ann. Stat. 33 (5) (2005) 2042–2065, http://dx.doi.org/10.1214/009053605000000417. http://dx.doi.org/10.1214/009053605000000417.
[9] H.-J. Sun, M. Sun, S.-R. Wang, A measurement of overlap rate between Gaussian components, in: International Conference on Machine Learning and Cybernetics, vol. 4, 2007, pp. 2373–2378. http://dx.doi.org/10.1109/ICMLC.2007.4370542.
[10] H. Sun, S. Wang, Measuring the component overlapping in the Gaussian mixture model, Data Min. Knowl. Discov. 23 (3) (2011) 479–502, http://dx.doi.org/10.1007/s10618-011-0212-3. http://dx.doi.org/10.1007/s10618-011-0212-3.
[11] A.T. Kalai, A. Moitra, G. Valiant, Efficiently learning mixtures of two Gaussians, in: L.J. Schulman (Ed.), STOC, 2010, pp. 553–562. http://dx.doi.org/10.1145/1806689.1806765, <http://doi.acm.org/10.1145/1806689.1806765>.
[12] A. Moitra, G. Valiant, Settling the polynomial learnability of mixtures of Gaussians, in: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, 2010, pp. 93–102. http://dx.doi.org/10.1109/FOCS.2010.15.
[13] S. Dasgupta, Learning mixtures of Gaussians, in: 40th Annual Symposium on Foundations of Computer Science, 1999, pp. 634–644. http://dx.doi.org/10.1109/SFFCS.1999.814639.
[14] S. Arora, R. Kannan, Learning mixtures of separated nonspherical gaussians, Ann. Appl. Probab. 15 (1A) (2005) 69–92, http://dx.doi.org/10.1214/105051604000000512. <http://dx.doi.org/10.1214/105051604000000512>.
[15] M. Brand, K. Huang, A unifying theorem for spectral embedding and clustering, in: C.M. Bishop, B.J. Frey (Eds.), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, 2003. <http://research.microsoft.com/en-us/um/cambridge/events/aistats2003/proceedings/189.pdf>.
[16] S. Vempala, G. Wang, A spectral algorithm for learning mixtures of distributions, in: Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on, 2002, pp. 113–122. <http://dx.doi.org/10.1109/SFCS.2002.1181888>.

[17] D. Achlioptas, F. McSherry, On spectral learning of mixtures of distributions, in: Learning Theory, Lecture Notes in Computer Science, vol. 3559, Springer, Berlin, 2005, pp. 458–469, http://dx.doi.org/10.1007/1150341531. <http://dx.doi.org/10.1007/1150341531>.
[18] R. Kannan, H. Salmasian, S. Vempala, The spectral method for general mixture models, in: P. Auer, R. Meir (Eds.), Learning Theory, Lecture Notes in Computer Science, vol. 3559, Springer, Berlin, Heidelberg, 2005, pp. 444–457, http://dx.doi.org/10.1007/1150341530. <http://dx.doi.org/10.1007/1150341530>.
[19] S. Brubaker, S. Vempala, Isotropic pca and affine-invariant clustering, in: M. Grötschel, G. Katona, G. Sági (Eds.), Building Bridges, Bolyai Society Mathematical Studies, vol. 19, Springer, Berlin, Heidelberg, 2008, pp. 241–281, http://dx.doi.org/10.1007/978-3-540-85221-68. <http://dx.doi.org/10.1007/978-3-540-85221-68>.
[20] E. Nowakowska, J. Koronacki, S. Lipovetsky, Dimension reduction for data of unknown cluster structure. Available from: arXiv:1407.7811.
[21] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Probability and Mathematical Statistics: A Series of Monographs and Textbooks, Academic Press [Harcourt Brace Jovanovich, Publishers], 1979.
[22] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Springer Series in Statistics, Springer, New York, 2009, http://dx.doi.org/10.1007/978-0-387-84858-7. <http://dx.doi.org/10.1007/978-0-387-84858-7>.
[23] S. Lipovetsky, Additive and multiplicative mixed normal distributions and finding cluster centers, Int. J. Mach. Learn. Cybern. 4 (1) (2013) 1–11, http://dx.doi.org/10.1007/s13042-012-0070-3. <http://dx.doi.org/10.1007/s13042-012-0070-3>.
[24] S. Lipovetsky, Finding cluster centers and sizes via multinomial parameterization, Appl. Math. Comput. 221 (2013) 571–580, http://dx.doi.org/10.1016/j.amc.2013.06.098.
[25] K. Fukunaga, Introduction to Statistical Pattern Recognition. Computer Science and Scientific Computing, second ed., Academic Press Inc., Boston, MA, 1990.
[26] E. Nowakowska, J. Koronacki, S. Lipovetsky, Tractable measure of component overlap for gaussian mixture models. <arXiv:1407.7172v1>.