Semi-supervised model-based clustering with controlled clusters leakage

Marek Śmieja Łukasz Struski Jacek Tabor

[1]Faculty of Mathematics and Computer Science
Jagiellonian University
Łojasiewicza 6, 30-348 Krakow, Poland

Abstract

In this paper, we focus on finding clusters in partially categorized data sets. We propose a semi-supervised version of Gaussian mixture model, called C3L, which retrieves natural subgroups of given categories. In contrast to other semi-supervised models, C3L is parametrized by user-defined leakage level, which controls maximal inconsistency between initial categorization and resulting clustering. Our method can be implemented as a module in practical expert systems to detect clusters, which combine expert knowledge with true distribution of data. Moreover, it can be used for improving the results of less flexible clustering techniques, such as projection pursuit clustering. The paper presents extensive theoretical analysis of the model and fast algorithm for its efficient optimization. Experimental results show that C3L finds high quality clustering model, which can be applied in discovering meaningful groups in partially classified data.

1 Introduction

Model-based clustering aims at finding a mixture of probability models, which optimally estimates true probability distribution on data space. Contrary to other clustering techniques, it does not only recover meaningful groups, but also gives a rule (probability model) for generating elements from clusters. Therefore, it is commonly used in various areas of machine learning and data analysis [28, 34, 40].

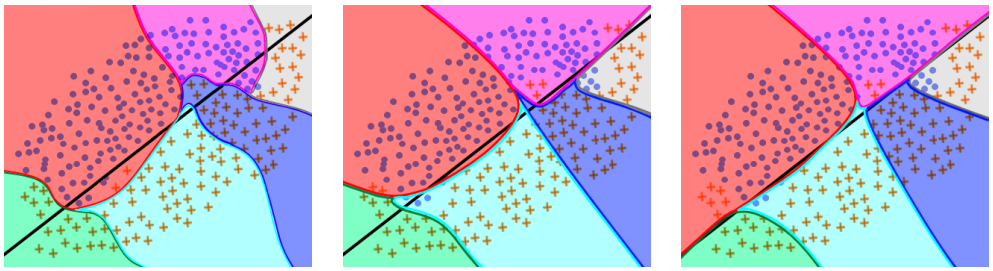
Although clustering is an unsupervised technique, one can introduce additional information to guide the algorithm what is the expected structure of clusters. Semi-supervised learning methods usually use partial labeling [16] or pairwise constraints

[20] to transfer expert knowledge into clustering process, while consensus and alternative clustering gather information from several partitions of data into one general view [9, 22]. In this paper, we assume that we have the knowledge about division of data set into two categories and focus on the following problem: *How to find the best model of clusters that preserves a fixed amount of information about existing categories?* In other words, we focus on finding interesting clusters, which are very likely to belong to one category.

To explain a basic motivation behind our model, let us consider an expert system used for automatic text translation. It is a common practice to construct several translation models, each designed for one cluster retrieved from a data set [1]. Alternatively, since texts are often categorized into specific domains, e.g. sport, politics, etc., then each translator can be fitted to one of these categories. To consider together both options, we could implement a separate module responsible for finding clusters, which (a) are described by compact models (e.g. Gaussians) and (b) are related with predefined topics. Observe that optimization of these two conflicting goals simultaneously is non-trivial. We cannot cluster elements from each category individually, because this strategy does not lead to optimal solution for the entire data set (in terms of likelihood). Moreover, existing categorization might be inaccurate as well as the interesting groups can cross the boundary between predefined domains. Therefore, a better approach is to incorporate the constraint to the clustering process and always work with the entire data set.

Our method can also be applied to strictly unsupervised situations, where no initial categorization is given. Let us recall that one way to analyze clusters in complex data spaces relies on finding projections onto one dimensional subspaces, where groups can be easily identified. Projection pursuit focuses on choosing such a direction, which optimizes selected statistical index such as kurtosis [26] or skewness [17]. Since one dimensional views generate linear decision boundaries in original data space, it is not possible to find flexible cluster structures. However, we can input such a linear boundary to our model in order to improve existing clusters. Our method directly uses the information from initial splitting, but can extend linear decision surfaces to nonlinear ones generated by probabilistic mixture models.

Following the above motivation, we propose a semi-supervised clustering with controlled clusters leakage model (C3L), which integrates a distribution of data with a fixed division of the space into two categories. C3L focuses on finding a type of Gaussian mixture model (GMM) [21], which maximizes the likelihood function and preserves the information contained in the initial splitting with a predefined probability (leakage level). Intuitively, we allow for the flow of clusters densities over decision surface, but with a full control of total probability assigned to the opposite category, which is defined as the leakage level $\alpha \in (0, 1)$ (see Figure 1).



(a) $\alpha = 0.159$ (b) $\alpha = 0.043$ (c) $\alpha = 0.001$

Figure 1: The effects of C3L for different values of the leakage level α.

This general idea is formulated as a constrained optimization problem (Section 3).

The advantages of C3L can be summarized as follows:

1. It has a closed form solution in a special case of cross-entropy clustering (a type of GMM) [35].

2. It can be efficiently implemented and optimized by a modified on-line Hartigan algorithm (Section 4).

3. The user can directly parametrize C3L by a maximal inconsistency level between initial categorization and final clustering model (leakage level).

4. The selection of the leakage level α allows to move from a strictly unsupervised GMM for $\alpha = 0.5$, where decision boundary has no effect on clustering, to the limiting case of $\alpha \to 0$, where every group is fully condensed in one category (Section 5).

Experimental studies confirm that the proposed approach builds a high quality model under a given constraint in terms of inner clustering measures, such as Bayesian Information Criterion (Section 6.1). It can be successfully used to discover meaningful groups in partially classified data (Section 6.2) as well as to improve existing clusters obtained by applying projection techniques (Section 6.3). We present a real-life case study, in which the use of C3L allows to detect subgroups of chemical space given their division into active and inactive classes (Section 6.4).

2 Related work

Semi-supervised clustering incorporates the knowledge about class labels to partitioning process [5]. This information can be presented as partial labeling, which gives a division of a small portion of data into categories, or as pairwise constraints,

which indicate whether two data points originate from the same (must-links) or distinct classes (cannot-links). Although pairwise constraints provide less amount of information than partial labeling, it is easier to assess whether two instances come from the same group than assign them to particular classes.

Clustering with pairwise constraints was introduced by Wagstaff et al. [37], who created a variant of k-means, which focuses on preserving all constraints. Shental et al. [29] constructed a version of Gaussian mixture model, which gathers data points into equivalence classes (called chunklets) using must-link relation and then applied EM algorithm on such generalized data set of chunklets. This approach was later modified to multi-modal clustering models [32]. The aforementioned methods work well with noiseless side information, but deteriorate the results when some constraints are mislabeled. To overcome this problem, the authors of [6, 19] applied hidden Markov random fields (HMRF) to construct more sophisticated dependencies between linked points. However, the use of HMRF leads to complex solutions, which are difficult to optimize. In recent years, Asafi and Cohen-Or. [4] suggested reducing distances between data points with a must-link constraint and adding a dimension for each cannot-link constraint. After updating all other distances to, e.g., satisfy the triangle inequality, the thus obtained pairwise distance matrix can be used for unsupervised learning. Wang and Davidson [38] proposed a version of spectral clustering, which relies on solving a generalized eigenvalue problem.

Partial labeling is used in clustering to define sample data points from particular classes. Liu and Fu [16] added additional attributes to feature vectors and proposed modified k-means algorithm. There is also a semi-supervised version of fuzzy c-means [24, 25], where the authors supplied the cost function with a regularization term that penalizes fuzzy partitions that are inconsistent with the side information. GMMs can be adapted to make use of class labels by combining the classical unsupervised GMM with a supervised one [2, 41].

Since assigning data points to classes or labeling pairwise constraints requires extensive domain knowledge, then many clustering methods were adapted to use additional information about data, which does not require human intervention. One example is consensus clustering, which considers gathering information coming from different domains [22]. On the other hand, complementary (alternative) clustering aims at finding groups which provide a perspective on the data that expands on what can be inferred from previous partitions [9].

C3L is a version of Gaussian mixture model, which uses side information given by class labels or more generally by a decision boundary between classes. In contrast to classical methods applying partial labeling, it focuses on finding subgroups of original classes. This goal is similar to information bottleneck method [7, 36]. Roughly speaking, this approach tries to construct compact clusters (compressed represen-

tation), which contain high amount of information about existing classes (auxiliary variable). While information-theoretic approaches use mutual information (or conditional entropy) to preserve the consistency with an initial categorization, C3L explicitly defines maximal probability of inconsistency (leakage level). The leakage level can be understood as Bayes error in classification or significance level in hypothesis testing and restricts every cluster model to be assigned to one of two initial classes with a predefined probability. Subgroups could also be detected by using cannot-link constraints, clustering with pairwise constraints does not allow to input maximal level of error. Moreover, computational complexity of applying cannot-link constraints to GMM is usually high, while C3L works in a comparable time to classical unsupervised mixture models.

C3L can be naturally combined with projection pursuit approach, which focuses on selecting low dimensional projections of data for finding clusters (or other meaningful characteristics of data). Such a projection can be determined by optimizing selected statistical coefficients e.g. kurtosis [11, 26] or skewness [17, 18]. Since every one dimensional view induces linear decision boundary in the original space, this technique may not be sufficient to detect complex data patters. C3L allows to take such a rough linear splitting of data and correct simple decision boundaries to nonlinear ones.

3 Theoretical model

In this section, we introduce our model and discuss its possible extensions and applications.

Let a data space $\mathbb{R}^N$ be divided by a codimension one hyperplane[1] H given by:

$$H = \{x \in \mathbb{R}^N : h^T x = a\},$$

for fixed $h \in \mathbb{R}^N$ and $a \in \mathbb{R}$. The hyperplane H induces hard classification rule: the class label of each point $x \in \mathbb{R}^N$ is determined by

$$\text{class}_H(x) = \text{sign}(h^T x - a). \tag{1}$$

This splits a dataset $X \subset \mathbb{R}^N$ into two groups X_+, X_- given by

$$X_\pm = \{x \in X : \text{class}_H(x) = \pm 1\}. \tag{2}$$

Alternatively, we can consider the initial classification of entire space $\mathbb{R}^N$ as

$$H_\pm = \{x \in \mathbb{R}^N : \text{class}_H(x) = \pm 1\},$$

[1] $(N-1)$-dimensional hyperplane

and put $X_{\pm} = X \cap H_{\pm}$. Note that the uncertainty of class label is usually higher for elements localized closer to the barrier than for those with larger distance from H.

In a model-based clustering we focus on estimating a density of data space with a use of mixture of k densities, $g = \sum_{i=1}^{k} p_i g_i$, where every g_i belongs to a given parametric family of densities (usually Gaussian) and p_i are prior probabilities [21]. This goal can be practically realized by maximizing the likelihood function. Let us define the inconsistency between a cluster density g_i and the initial classification:

$$\alpha_i = \min\{\int_{H_-} g_i(x)dx, \int_{H_+} g_i(x)dx\}.$$

The above formula gives the amount of probability that is spread to opposite class and it is related with Bayes error of Gaussian model assuming that H predicts the class membership correctly.

Our question is: how to find a clustering model that optimizes a likelihood function and provides high consistency with initial classification? If we knew that H gives a perfect classification rule, then we could try to keep every model g_i maximally consistent with H, i.e. perform a separate clustering of every category. However, this is usually not the case and this strategy does not guarantee to obtain optimal solution (in terms of likelihood) for the entire data set. Moreover, some interesting groups can cross the decision boundary. Therefore, we should allow for the flow of corresponding densities over a decision surface, but with the full control of the total probability assigned to the opposite class. In our approach, we formulate a constrained optimization problem, where we aim at finding such a mixture model g that maximizes the quality of density estimation and preserves a fixed inconsistency level α, i.e. every component g_i has to satisfy $\alpha_i \leq \alpha$.

One could probably try to realize the above goal by a classical GMM approach, however, at very high numerical and theoretical cost. It would be impossible to get a closed form solution and complex non-linear optimization would be needed. Therefore, in this paper we have decided to use cross-entropy clustering (CEC) [30, 33,35], which similarly to GMM divides data with respect to Gaussian distributions. Contrary to GMM, in CEC the clusters do not "cooperate" one with another to build the global cost function[2] and consequently it is enough to calculate the cost function for each cluster individually.

CEC is based on Minimum Description Length Principle (MDLP) [27] and focuses on minimizing the generalized cross-entropy function, by selecting optimal Gaussian probability distribution for each cluster[3]. Given a single cluster X and

[2]Instead of optimizing a density $p_i g_1 + \ldots + p_k g_k$, CEC finds optimal subdensity $\max\{p_i g_1, \ldots, p_k g_k\}$, i.e. every point x is linked with exactly one model g_i.

[3]The minimization of cross-entropy is equivalent to the maximization of likelihood function.

corresponding density g, the empirical cross-entropy function equals:

$$h^{\times}(X\|g) = -\frac{1}{|X|}\sum_{x\in X}\ln(g(x)).$$

In the case of Gaussian densities, $g = \mathcal{N}(m,\Sigma)$, the above formula can be reduced to its closed form:

$$\begin{array}{l} h^{\times}(X\|\mathcal{N}(m,\Sigma)) = \\ \frac{N}{2}\ln(2\pi) + \frac{1}{2}\|\hat{m}_X - m\|_{\Sigma} + \frac{1}{2}\mathrm{tr}(\Sigma^{-1}\hat{\Sigma}_X) + \frac{1}{2}\mathrm{lndet}(\Sigma), \end{array} \tag{3}$$

where

$$\begin{array}{l} \hat{m}_X := \frac{1}{|X|}\sum\limits_{x\in X} x, \\ \hat{\Sigma}_X := \frac{1}{|X|}\sum\limits_{x\in X}(x-\hat{m}_X)(x-\hat{m}_X)^T, \end{array} \tag{4}$$

denote the sample mean and covariance of X and

$$\|x\|_{\Sigma} := x^T\Sigma^{-1}x$$

is the Mahalanobis norm. Given k clusters $X_1,\ldots,X_k$ the overall cross-entropy function equals

$$\sum_i p_i(-\ln p_i) + p_i h^{\times}(X_i\|g_i), \tag{5}$$

where g_i is a Gaussian density with parameters given by (4) for X_i and $p_i = \frac{|X_i|}{|X|}$.

The term $p_i(-\ln p_i)$ adds a cost for maintaining a cluster. In consequence, the method tends to keep the model simple and allows for the reduction of redundant clusters. Therefore, CEC cost function (5) combines the model accuracy with its complexity.

With this, we are ready to define our C3L model. First, we give a definition of a linear constraint, which restricts every cluster model to one category with a fixed probability.

Definition 1. *Let H be a codimension one hyperplane on $X \subset \mathbb{R}^N$ and let $\alpha > 0$. We say that a density g satisfies a linear constraint (H,α), if*

$$\textit{either } \int_{H_-} g(x)dx \geq 1-\alpha \textit{ or } \int_{H_+} g(x)dx \geq 1-\alpha. \tag{6}$$

The number α will be referred as the leakage level.

The above definition of linear constraint is analogical to linear separability with power α given in [26, Section 2] in the context of projection pursuit.

If $\alpha \geq \frac{1}{2}$ then an arbitrary density satisfies one of the conditions given by (6). Therefore, *a strict constraint* is given by $\alpha < \frac{1}{2}$. Observe that the above constraint is reminiscent of a typical approach used in hypothesis testing, where we accept a given hypothesis if it lies within a predefined percentage of density. *In our case we consider only those density cluster models, which lie with a probability* $(1-\alpha)$ *on one side of decision boundary.*

We now introduce a linear constraint to CEC framework. First, we define the criterion function for a single cluster:

Definition 2. *(One cluster* C3L *cost function) Let* $X \subset \mathbb{R}^N$ *be divided by a codimension one hyperplane* H*. Given* $\alpha > 0$ *and a family* $\mathcal{G}$ *of Gaussian densities,* C3L *cost function for* X *is defined by*

$$\begin{array}{l} E_H^{\alpha}(X\|\mathcal{G}) := \\ \inf\{h^{\times}(X\|g) : g \in \mathcal{G} \text{ which satisfies constraint } (H,\alpha)\}, \end{array} \tag{7}$$

where $h^{\times}(X\|g)$ *is given by* (3)*.*

We always assume that a covariance matrix of X is nonsingular, i.e., $\det(\hat{\Sigma}_X) \neq 0$. This prevents from the situation when X lies in the subspace of $\mathbb{R}^N$, which might lead to degenerate solutions.

The overall C3L cost is defined as follows:

Definition 3. *(Overall* C3L *cost function for clustering) Let* $X \subset \mathbb{R}^N$ *be divided by a codimension one hyperplane* H*. Given a splitting of* X *into clusters* $X_1, \ldots, X_k$*, a family* $\mathcal{G}$ *of Gaussian densities and* $\alpha > 0$*, the total* C3L *clustering cost equals*

$$\begin{array}{l} E_H^{\alpha}(X_1, \ldots, X_k\|\mathcal{G}) = \\ \sum_{i=1}^{k} p_i(-\ln p_i + E_H^{\alpha}(X_i\|\mathcal{G})), \text{ where } p_i = \frac{|X_i|}{|X|}. \end{array} \tag{8}$$

The optimal clustering is the one, which minimizes the above cost function (the optimization problem will be the subject of the next section). We emphasize that C3L accepts arbitrary Gaussian densities as mixture components, which satisfy the linear constraint. In particular, each Gaussian can have distinct covariance matrix.

Let us observe that introduced model can be applied in the case of any decision boundary (not only linear hyperplane). If f_+ and f_- are two decision functions that quantify the chance of assigning data points to positive and negative classes, then the label of an instance $x \in \mathbb{R}^N$ is chosen as

$$\text{class}(x) = \arg\max_{j=\pm} f_j(x). \tag{9}$$

Clearly, for two class problem we can define one discriminant $f = f_+ - f_-$. In consequence, the formula (9) can be simplified to:

$$\text{class}(x) = \text{sign} f(x). \tag{10}$$

Next, we extend the input space $\mathbb{R}^N$ to $\mathbb{R} \times \mathbb{R}^N$ and embed our data set X into this space by:

$$X \ni x \to (f(x), x) \in \mathbb{R} \times \mathbb{R}^N.$$

Observe that a hyperplane $H = \{0\} \times \mathbb{R}^N$ gives the same classification rule in $\mathbb{R}^{N+1}$ to the formula (10) in $\mathbb{R}^N$.

Let $\mathcal{G}_N$ denotes the set of all Gaussian densities on $\mathbb{R}^N$ and let $\mathcal{G}_{1,N}$ be the set of Gaussian densities, that can be factorized into two independent components, defined by:

$$\mathcal{G}_{1,N} := \{g(x) = g_1(x_1) \cdot g_N(x_{2:N+1}) : g_1 \in \mathcal{G}_1, g_N \in \mathcal{G}_N\}, \tag{11}$$

where $x_{k:l} = (x_k, \ldots, x_l)$, for $x = (x_1, \ldots, x_{N+1})$. If we consider a mixture model $g = \sum_i p_i g_i$, where $g_i = g_1^i \cdot g_N^i \in \mathcal{G}_{1,N}$, then the first component g_1^i will describe a distribution of discrimination function f while g_N^i will describe a density in the original space X. Therefore, the use of $\mathcal{G}_{1,N}$ allows to model a distribution of data and discriminant function individually. We will use the family $\mathcal{G}_{1,N}$ in the next section.

4 Optimization

Without loss of generality, we assume that a decision boundary of $X \subset \mathbb{R}^N$ is given by a hyperplane[4] $H = \{0\} \times \mathbb{R}^{N-1}$. From now on, our attention is restricted to the class of Gaussian densities $\mathcal{G}_{1,N-1}$ defined by (11), i.e. we assume that every component g is of the form $g = g_1 \cdot g_{N-1}$, where g_1 is 1-dimensional and g_{N-1} is $(N-1)$-dimensional Gaussian density. This model suits perfectly to the case of arbitrary decision boundary described at the end of section 3, where we model the data distribution and the values of decision support function separately.

We are going to show that in the case of $\mathcal{G}_{1,N-1}$ we can compute one cluster cost function analytically. Let us first observe that the selection of 1-dimensional density $g_1 \in \mathcal{G}_1$ and $(N-1)$-dimensional density $g_{N-1} \in \mathcal{G}_{N-1}$ can be done separately in C3L clustering. First, we verify that the linear constraint is independent of g_{N-1}, i.e.,

$$\int_{[0,+\infty)\times\mathbb{R}^{N-1}} g(x)dx = \int_0^{+\infty} g_1(x_1)dx_1,$$
$$\int_{(-\infty,0]\times\mathbb{R}^{N-1}} g(x)dx = \int_{-\infty}^0 g_1(x_1)dx_1.$$

[4]Observe that given an arbitrary hyperplane one can always shift the original data and change the basis of $\mathbb{R}^N$ in an orthonormal way to obtain this situation.

Then, observe that the cross-entropy between X and g can be calculated as:

$$h^{\times}(X\|g) = h^{\times}(X[1]\|g_1) + h^{\times}(X[2:N]\|g_{N-1}),$$

where $X[k:l]$ denotes a data set X restricted to the attributes from k to l. This follows from

$$\begin{aligned}
&h^{\times}(X\|g)\\
&= \sum_{x\in X} -\ln g(x) = \sum_{x\in X} -\ln(g_1(x_1)\cdot g_{N-1}(x_{2:N}))\\
&= \sum_{x\in X} \big(-\ln(g_1(x_1)) - \ln(g_{N-1}(x_{2:N}))\big)\\
&= \sum_{x_1\in X[1]} -\ln(g_1(x_1)) + \sum_{x_{N-1}\in X[2:N]} -\ln(g_{N-1}(x_{N-1}))\\
&= h^{\times}(X[1]\|g_1) + h^{\times}(X[2:N]\|g_{N-1}).
\end{aligned}$$

Since the product densities can be selected individually, the optimization subject to the constraint is performed in one dimension.

Corollary 1. *Let (H,α) be a linear constraint defined on dataset $X\subset\mathbb{R}^N$, where $H=\{0\}\times\mathbb{R}^{N-1}$. Then the one cluster* C3L *cost function of X is given by:*

$$E^{\alpha}_{H}(X\|\mathcal{G}_{1,N-1}) = E^{\alpha}_{\{0\}}(X[1]\|\mathcal{G}_1) + h^{\times}(X[2:N]\|\mathcal{G}_{N-1}),$$

where $E^{\alpha}_{\{0\}}((X[1]\|\mathcal{G}_1)$ is one cluster C3L *cost function* (7) *calculated in one dimensional situation.*

To complete the formula given in Corollary 1, the C3L cost function in one dimensional case has to be calculated. To facilitate the calculation we first give an equivalent form of linear constraint.

Given $\alpha>0$, let us denote by p^{α} the corresponding quantile:

$$p^{\alpha} := \Phi^{-1}_{\mathcal{N}(0,1)}(1-\alpha), \tag{12}$$

where $\Phi_{\mathcal{N}(m,\sigma)}(\cdot)$ denotes a cumulative distribution function of $\mathcal{N}(m,\sigma)$. Making use of elementary calculations we get that:

$$\alpha = \Phi_{\mathcal{N}(m,\sigma)}(m - p^{\alpha}\sigma),$$

for any $m\in\mathbb{R}$ and $\sigma>0$. Then, one dimensional density $\mathcal{N}(m,\sigma)$ satisfies the constraint $(\{0\},\alpha)$, iff

$$|m| \geq p^{\alpha}\sigma, \tag{13}$$

In other words, the distance between the mean m and the barrier is at least $p^{\alpha}\sigma$.

To calculate the optimal one dimensional cost function, we must observe that the cross-entropy (3) between a distribution of dataset $X \subset \mathbb{R}$ with mean $\hat{m}_X$ and standard deviation $\hat{\sigma}_X$ and a Gaussian density $\mathcal{N}(m, \sigma)$ equals:

$$h^{\times}(X \| \mathcal{N}(m,\sigma)) = \tfrac{1}{2}\left(\tfrac{\hat{\sigma}_X^2 + (m-\hat{m}_X)^2}{\sigma^2} + \ln(\sigma^2) + \ln(2\pi)\right). \tag{14}$$

The optimal parameters of $\mathcal{N}(m, \sigma)$ are obtained by the minimization of the above function under the restriction (13):

Theorem 1. *Let $X \subset \mathbb{R}$ be a dataset with the mean $\hat{m}_X \neq 0$ and the standard deviation $\hat{\sigma}_X > 0$. We assume that $(\{0\}, \alpha)$ denotes the linear constraint on X, for $\alpha > 0$, and $p^\alpha = \Phi^{-1}_{\mathcal{N}(0,1)}(1-\alpha)$.*

If $|\hat{m}_X| \geq p^\alpha \hat{\sigma}_X$, then put $m_X^\alpha := \hat{m}_X$, $\sigma_X^\alpha := \hat{\sigma}_X$, otherwise

$$\begin{array}{rl} m_X^\alpha := & \frac{-(p^\alpha)^2 \hat{m}_X + \mathrm{sign}(\hat{m}_X) p^\alpha \sqrt{((p^\alpha)^2+4)\hat{m}_X^2 + 4\hat{\sigma}_X^2}}{2}, \\ \sigma_X^\alpha := & \frac{|\hat{m}_X^\alpha|}{p^\alpha}. \end{array} \tag{15}$$

Then, the normal density $\mathcal{N}(m_X^\alpha, \sigma_X^\alpha)$ minimizes the value of $h^{\times}(X \| \mathcal{N}(m,\sigma))$, given by (14)*, under the restriction $|m| \geq p^\alpha \sigma_X$ (*C3L *cost function $E^\alpha_{\{0\}}(X \| \mathcal{G})$).*

Proof. Our aim is to find the minimum of the function

$$h(m,\sigma) = h^{\times}(X \| \mathcal{N}(m,\sigma)) \text{ under the condition } |m| \geq p^\alpha \sigma. \tag{16}$$

It is obvious that the above function has the derivative zero only at its global minimum which is given by a pair (m_X, σ_X). Consequently, if $m = \hat{m}_X, \sigma = \hat{\sigma}_X$ satisfies the constraint $|m| \geq p^\alpha \sigma$, then we have found the minimum.

In the opposite case, we only need to verify what happens on the boundary of the constraints (since we do not have any local minimum inside $|m| > p^\alpha \sigma$), that is when $p^\alpha \sigma = |m|$. Then, by (14), the function (16) simplifies to

$$h(m) = \frac{1}{2}\big(\frac{\hat{\sigma}_X^2 + (m-\hat{m}_X)^2}{m^2}(p^\alpha)^2 + \ln\frac{m^2}{(p^\alpha)^2} + \ln(2\pi)\big).$$

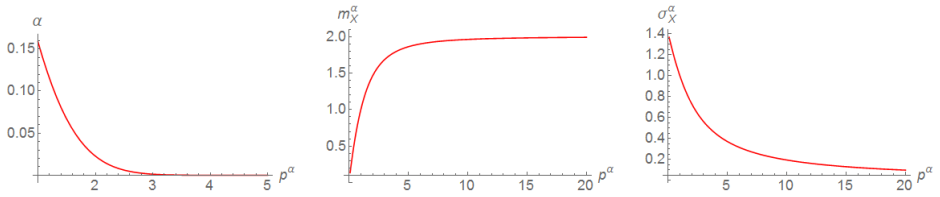
Then

$$h'(m) = -\frac{\hat{\sigma}_X^2 + \hat{m}_X^2}{m^3}(p^\alpha)^2 + \frac{\hat{m}_X}{m^2}(p^\alpha)^2 + \frac{1}{m}.$$

Finally, the solution of $h'(m) = 0$ which minimizes the value of $h(m)$, is given by

$$m = \frac{-(p^\alpha)^2 \hat{m}_X + \mathrm{sign}(\hat{m}_X) p^\alpha \sqrt{((p^\alpha)^2+4)\hat{m}_X^2 + 4\hat{\sigma}_X^2}}{2},$$

which completes the proof. □



(a) (b) (c)

Figure 2: The influence of the leakage level α on the parameters p^α, m_X^α and σ_X^α.

The above analysis shows how to calculate the best model of clusters for a given partition. However, finding an optimal partition is NP-hard problem, where heuristic iterative algorithms are commonly used. One can apply a slight modification of Hartigan approach to optimize C3L cost function (see A for details). Similar algorithm is used in optimization of CEC and k-means methods.

5 Theoretical analysis

In this section we present a theoretical analysis of C3L model in its simplified form. We start with investigating the convergence of cluster parameters with respect to the leakage level α. Then, we show that under certain assumptions a decision boundary determined by C3L model with two clusters converges to the initial barrier, when α approaches to 0.

In order to accommodate the constraint one-dimensional density cluster model modifies its mean and standard deviation according to Theorem 1. The relation between m_X^α and σ_X^α (given by (15)) is inversely proportional, i.e., the increase of m_X^α results in the decrease of σ_X^α and vice versa (see Figure 2). However, the most important fact is that $|m_X^\alpha|$ does not grow infinitely, but converges to a finite number dependent on a data set. To prove it formally, let us first consider a one dimensional case.

Lemma 1. *We assume that $X \subset \mathbb{R}$ is a data set with a mean $\hat{m}_X \neq 0$ and standard deviation $\hat{\sigma}_X > 0$. Let $g^\alpha = \mathcal{N}(m_X^\alpha, \sigma_X^\alpha)$ denote a density minimizing one cluster* C3L *cost function under the linear constraint $(\{0\}, \alpha)$, i.e, m_X^α and σ_X^α are given by Theorem 1.*

Then:

$$\begin{array}{ll} m_X^\alpha & \to \hat{m}_X + \frac{\hat{\sigma}_X^2}{\hat{m}_X} \\ \sigma_X^\alpha & \to 0 \end{array} \quad , \text{ as } \alpha \to 0.$$

Proof. The proof is included in B. □

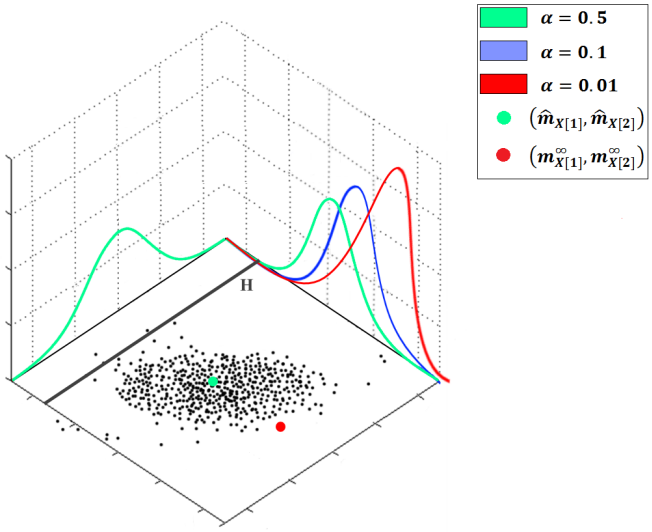


Figure 3: The change of α affects only the form of a density model which is orthogonal to the barrier. Its mean converges to the limiting value (marked with red dot) given by Corollary 2 while its standard deviation grows to infinity.

If we combine the above result with the fact that the mean and the covariance of $(N-1)$ dimensional density g_{N-1}, for a model $g = g_1 \cdot g_{N-1}$, do not depend on the linear constraint, but are the maximum likelihood estimators of data, we get the following corollary:

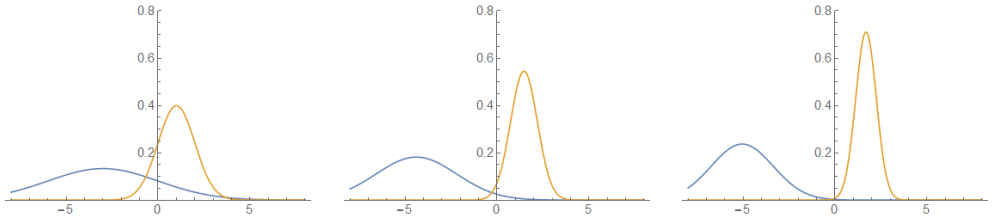
Corollary 2. *We assume that $X \subset \mathbb{R}^N$ is a data set, with a mean $\hat{m}_X$ and a covariance matrix $\hat{\Sigma}_X$, where $\hat{m}_{X[1]} \neq 0$ and $\hat{\sigma}^2_{X[1]} = \hat{\Sigma}_{X[1]}$. Let $g^\alpha = g_1^\alpha \cdot g_{N-1} \in \mathcal{G}_{1,N-1}$ denote a density minimizing one cluster* C3L *cost function under the linear constraint $(\{0\} \times \mathbb{R}^{N-1}, \alpha)$, i.e., $g_1^\alpha = \mathcal{N}(m^\alpha_{X[1]}, \sigma^\alpha_{X[1]})$ is given by Theorem 1 and $g_{N-1} = \mathcal{N}(m^\alpha_{X[2:N]}, \Sigma^\alpha_{X[2:N]})$.*

Then:

$$\begin{aligned} m^\alpha_{X[1]} &\to \hat{m}_{X[1]} + \frac{\hat{\sigma}^2_{X[1]}}{\hat{m}_{X[1]}}, \\ \sigma^\alpha_{X[1]} &\to 0, \\ m^\alpha_{X[2:N]} &= \hat{m}_{X[2:N]}, \\ \Sigma^\alpha_{X[2:N]} &= \hat{\Sigma}_{X[2:N]}, \end{aligned} \qquad \text{as } \alpha \to 0.$$

The Figure 3 shows the influence of the change of the parameter α on the form of resulting density function.

We now discuss the relations between an initial decision boundary defined by a hyperplane H and a splitting determined by C3L model. For a simplicity, we consider the case of only two clusters. We assume that the hyperplane $H = \{0\} \times$



(a) $\alpha = 0.159$ (b) $\alpha = 0.043$ (c) $\alpha = 0.001$

Figure 4: Convergence of C3L model for $\alpha \to 0$.

$\mathbb{R}^{N-1}$ defines the classification rule (1):

$$\text{class}_H(x) := \text{sign}(x_1), \text{ for } x \in \mathbb{R}^N,$$

which splits a data set $X \subset \mathbb{R}^N$ into two subsets $X_\pm$ (2).

Let the leakage level α be fixed. In a simplified form of C3L model with only two clusters, we assume that density models $p_\pm g_\pm^\alpha$ are chosen so that to maximize one cluster cost functions of $X_\pm$, respectively (not the overall cost of clustering). In the other words, a class density is selected ignoring the influence of objects which belong to the opposite class. Then any incoming object $x \in X$ can be classified to one of two clusters by calculating

$$\text{class}_\alpha(x) := \text{sign}(p_+ g_+^\alpha(x) - p_- g_-^\alpha(x)) = \pm 1,$$

We will show that a decision boundary determined by such model converges to H, as the leakage level α approaches to 0, i.e.,

$$\text{class}_\alpha(x) \xrightarrow{\alpha \to 0} \text{class}_H(x),$$

which is illustrated in Figure 4:

Theorem 2. *We assume that $X \subset \mathbb{R}^N$ is a data set and $H = \{0\} \times \mathbb{R}^{N-1}$ defines a hyperplane in $\mathbb{R}^N$ dividing X into two classes X_-, X_+, where $X_\pm \neq \emptyset$. Let $g_\pm^\alpha \in \mathcal{G}_{1,N-1}$ denote two densities minimizing one cluster* C3L *functions of $X_\pm$ under the constraint (H, α), respectively.*

Then, there exists a constant $C > 0$ such that

$$\text{class}_\alpha(x) \to \text{class}_H(x), \text{ as } \alpha \to 0,$$

for every $x \in \mathbb{R}^N$ satisfying $dist(x, H) \leq C$.

The above convergence holds only for instances, which are localized within the margin of size C around the barrier H. This is a natural situation occurring in every Gaussian discrimination. To prove the above theorem we first consider one dimensional situation, where an exact value of constant C will be given.

Lemma 2. *We assume that $X_\pm \subset \mathbb{R}_\pm$ are two non empty sets with means $\hat{m}_\pm$ and standard deviations $\hat{\sigma}_\pm$. Let $g^\alpha_\pm \in \mathcal{G}_1$ denote two densities minimizing one cluster* C3L *cost functions of $X_\pm$ under the linear constraint $(\{0\}, \alpha)$, respectively.*

Then,

$$\ln g^\alpha_+(x) - \ln g^\alpha_-(x) \to +\infty, \text{ as } \alpha \to \infty,$$

for every $0 < x \leq 2\left(\hat{m}_+ + \frac{\hat{\sigma}^2_+}{\hat{m}_+}\right)$.

Proof. The proof is included in C. □

In one dimensional case the constant C from Theorem 2 equals $C = 2\left(\hat{m}_+ + \frac{\hat{\sigma}^2_+}{\hat{m}_+}\right)$. Below we complete the proof of our main result:

Proof. (of Theorem 2) Let $x \in \mathbb{R}^N$ be such that $0 < x_1 < 2\left(\hat{m}_+ + \frac{\hat{\sigma}^2_+}{\hat{m}_+}\right)$, where $\hat{m}_+, \hat{\sigma}_+$ are the mean and standard deviation of $X_+[1]$. In other words, we assume that x lies at the right side of a decision boundary. We assume that the optimal C3L densities $g^\alpha_\pm \in \mathcal{G}_{1,N-1}$ for $X_\pm$ equal

$$g^\alpha_\pm = (g^\alpha_\pm)_1 \cdot (g^\alpha_\pm)_{N-1}$$

with the priors $p_\pm$. We will show that

$$\ln p_+ g^\alpha_+(x) > \ln p_- g^\alpha_-(x), \tag{17}$$

for sufficiently small $\alpha > 0$.

Since

$$\ln g^\alpha_\pm(x) \quad = \ln(g^\alpha_\pm)_1(x_1) + \ln(g^\alpha_\pm)_{N-1}(x_{2:N}) + \ln p_\pm,$$

the formula (17) can be rewritten as

$$\begin{array}{l} \ln(g^\alpha_+)_1(x_1) - \ln(g^\alpha_-)_1(x_1) + \ln(g^\alpha_+)_{N-1}(x_{2:N}) \\ -\ln(g^\alpha_-)_{N-1}(x_{2:N}) + \ln p_+ - \ln p_- > 0. \end{array} \tag{18}$$

Making use of Lemma 2, we have

$$\ln(g^\alpha_+)_1(x_1) - \ln(g^\alpha_-)_1(x_1) \to +\infty, \text{ as } \alpha \to 0.$$

Because

$$|\ln(g^\alpha_+)_{N-1}(x_{2:N}) - \ln(g^\alpha_-)_{N-1}(x_{2:N}) + \ln p_+ - \ln p_-| < \infty,$$

the LHS of (18) can be arbitrary large, when $\alpha \to 0$, which completes the proof. □

Table 1: Regression UCI data sets and median number of clusters returned by C3L.

	Airfoil	Forest	Music$^+$	Stock
# Instances	1502	517	1059	536
# Features	5	12	5	7
# Clusters	5	3	5	7

$^+$ PCA was used to reduce a dimension of data

6 Experiments

We tested our method on sample examples retrieved from UCI repository [15] and one real data set of chemical compounds [39]. We verified the quality of the clustering model and demonstrated that C3L can be useful in discovering natural subgroups given a partial knowledge about two class division. We also used C3L to extend linear boundary between clusters obtained by projection pursuit technique to non-linear one. We also show its application on real data set of chemical compounds. We compared its performance with related model-based clustering techniques.

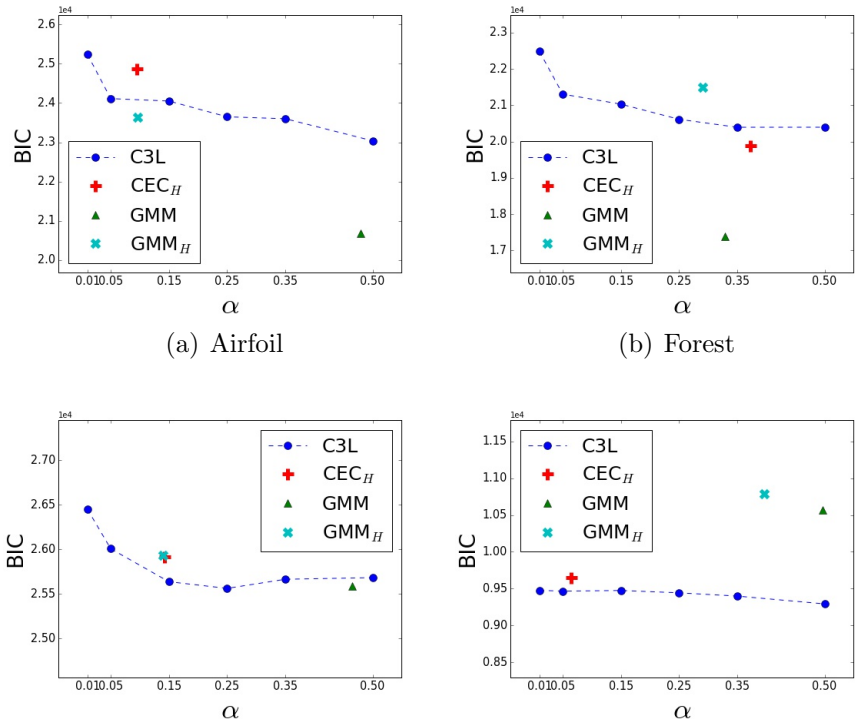
6.1 Quality of the model

In this experiment we consider a scenario, where every instance is assigned to one of two classes based on the value of a fixed decision support function. C3L builds a clustering model which preserves the information of class membership in a sense that every cluster density belongs to one of two classes with a probability greater than $(1 - \alpha)$ (6). We want to verify the quality of such model and compare it with the results produced by related model-based clustering techniques, which however do not allow for a direct specification of the leakage level.

To compare the quality of clustering models we applied Bayesian Information Criterion (BIC), which is a standard criterion for model selection [8]. The lower the BIC is the better the model is.

We used four regression UCI examples, which are summarized in Table 1. A dependent (output) variable was treated as a decision support function, which determines a decision boundary H. More precisely, if $X \times Y$ is a data set, where $X \subset \mathbb{R}^N$ contains explanatory variables and $Y \subset \mathbb{R}$ includes dependent variable, then a decision boundary H is defined by $H = X \times \{\text{median}(Y)\}$, where $\text{median}(Y)$ is the median of attribute Y.

The effects of C3L were compared with those obtained by classical GMM method, which ignores the presence of existing decision boundary. To introduce a decision boundary to the model, we also considered the second variant of GMM (which is



(a) Airfoil (b) Forest

(c) Music (d) Stock

Figure 5: Quality of the clustering models measured by Bayesian Information Criterion (BIC).

referred to as GMM$_H$): given a linear hyperplane H, which divides a data set X into two regions X_-, X_+ (see (2)), GMM$_H$ is defined as follows:

- GMM is applied to X_- and X_+ separately, which give two models $g_\pm = p_1^\pm g_1^\pm + \ldots + p_k^\pm g_k^\pm$.
- These models are combined into a single one by $g = \frac{|X_-|}{|X|} g_- + \frac{|X_+|}{|X|} g_+$.
- Finally, every point $x \in X$ is assigned to the most probable cluster by calculating $\frac{|X_\pm|}{|X|} p_i^\pm g_i^\pm(x)$.

Analogical strategies were also applied to CEC. Both variants of GMM use general Gaussian densities to model clusters distributions, while CEC-based methods use the same densities as C3L method, i.e. densities from the family $\mathcal{G}_{1,N-1}$.

We ran each method on $X \times Y$ with a decision boundary H. To investigate the influence of the leakage level on the clustering effects of C3L, six leakage levels were considered, $\alpha \in \{0.01, 0.05, 0.15, 0.25, 0.35, 0.5\}$. Since C3L and CEC$_H$ internally

Table 2: UCI data sets used in subgroups detection experiment. Last two rows show which reference groups were used for creating classes Y_- and Y_+.

	Balance	Segmentation$^+$	User	Wine
# Instances	625	210	258	178
# Features	4	5	5	13
# Clusters	3	7	4	3
Y_-	{1,2}	{1,3,4,5,7}	{1,2}	{1,2}
Y_+	{3}	{2,6}	{3,4}	{3}

$^+$ PCA was used to reduce a dimension of data

find the final number of clusters, we ran them with 10 groups, while other methods used the median number of groups returned by C3L calculated over these six leakage levels.

The results presented in the Figure 5 prove that the quality of C3L model improves as the leakage level is increased. This is a natural behavior, because lower values of α indicate higher restrictions on the clusters models. Since other methods do not control the inconsistency with classification, we measured their resulting leakage levels and marked returned BIC values. One can observe that in most cases CEC_H and GMM_H gave worse BIC than C3L method for corresponding leakage levels. It follows from the fact that C3L optimizes the model on the entire data, while CEC_H and GMM_H search for the optimal solutions in each half space individually. Most importantly, since we are not able to directly control the inconsistency level of these methods, they might lead to high inconsistency with initial classification, even if each model is optimized on a separate class (see Forest and Stock data sets). Similar argument holds for classical GMM – although it should allow for optimal fit to the data, it does not take into account the decision boundary between classes.

6.2 Subgroups detection

Decision boundary usually delivers some meaningful information about true structure of clusters. For example, the User Knowledge Modeling data set [13] distinguishes users with very low, low, middle and high knowledge about a given subject. If we knew a coarse separation of the users into two basic classes, {very low, low} and {middle, high}, it should be easier to detect their exact level of knowledge. In this experiment, we want to verify how the information of binary classification influences the clustering results.

To simulate the above scenario, where a decision boundary is closely related with

Table 3: Normalized mutual information for UCI datasets.

Method	Balance	Segmentation	User	Wine
$C3L_{0.01}$	**0.50**	**0.62**	0.53	0.50
$C3L_{0.05}$	0.44	0.60	0.49	**0.51**
CEC_H	0.48	0.59	0.47	0.34
CEC	0.03	0.58	0.25	0.46
c-GMM	0.20	0.54	0.56	0.47
GMM_H	0.49	0.56	**0.66**	0.40
GMM	0.07	0.58	0.36	0.45

the expected clustering structure, we considered four UCI data sets. For each one we applied the following procedure:

- Given k reference groups $Y_1, \ldots, Y_k$ of a data set $X \subset \mathbb{R}^N$ we created two classes Y_-, Y_+ by merging selected groups together.
- We trained SVM classifier on 15% elements drawn randomly from Y_- and Y_+, which induced a linear decision boundary H dividing X into two classes X_- and X_+.

The goal is to discover the reference grouping $Y_1, \ldots, Y_k$. Table 2 contains detailed information about data sets and classes Y_-, Y_+.

Given such prepared data sets, we ran all the methods applied in previous experiment. Additionally, we used a version of GMM enhanced with pairwise cannot-link constraints, which is referred as c-GMM [29]. Cannot-link constraints specify the pairs of elements that should not be included into the same group, which suits perfectly to this clustering task[5]. To generate a set of pairwise constraints containing similar knowledge to the decision boundary H, we went over all pairs of labeled data points and generated a cannot-link constraint, if one element belonged to Y_- and the second belonged to Y_+.

To compare C3L with other methods we used only two leakage levels $\alpha = 0.01$ and $\alpha = 0.05$. This choice was motivated by a typical approach used in hypothesis testing, where the significance level is commonly set to 0.01 or 0.05. C3L, CEC and CEC_H were initialized with twice the correct number of clusters (and they were allowed to reduce redundant groups). GMM-based methods were run with the correct numbers of clusters and can thus be expected to perform better than C3L, especially that GMMs describe the clusters by arbitrary Gaussian distributions. The

[5]The introduction of must-link constraints is not suitable in this case.

Table 4: Correlation between the accuracy of decision boundary and clustering results.

Method	Balance	Segmentation	User	Wine
$C3L_{0.01}$	0.95	0.64	0.98	0.45
$C3L_{0.05}$	0.90	0.34	0.94	0.36
CEC_H	0.99	0.68	0.96	0.76
GMM_H	0.99	0.98	0.88	0.80

similarity between the obtained clusterings and the ground truth partition $Y_1, \ldots, Y_k$ was evaluated using Normalized Mutual Information (NMI) [3]. NMI is bounded from the above by the value 1, which is attained for identical partitions.

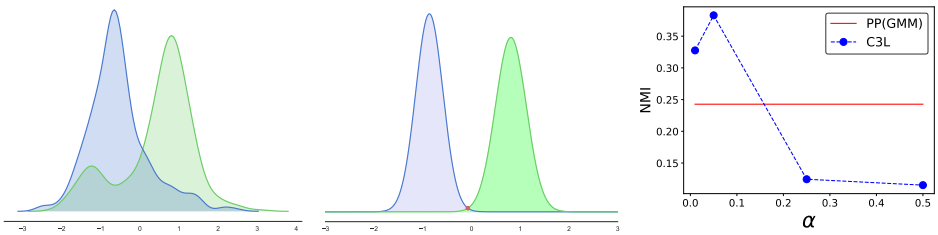
The results presented in Table 3 show that C3L performed better than other methods except the User Knowledge Modeling data set, where GMM_H gave very good result. This confirms that working with the entire data set is usually more profitable than finding subgroups in each half space individually (as GMM_H and CEC_H do). Moreover, lower leakage level $\alpha = 0.01$ usually led to higher NMI than $\alpha = 0.05$. It might follow from the fact that a decision boundary was constructed based on correctly labeled data and, therefore, it was very accurate.

To investigate the influence of the accuracy of decision boundary on the clustering results, we used 15% of data drawn from Y_-, Y_+ and assigned incorrect labels to a fixed percentage of them (we considered 0%, 10%, 20% and 30% of erroneous labels). The more labels were misspecified the worse a decision boundary should be.

Table 4 presents the correlation between the accuracy of decision boundary and the normalized mutual information of clustering. One can observe that CEC_H and GMM_H are more sensitive to incorrect decision boundary than both parameterizations of C3L. In consequence, these methods should not be used if there is a risk of unreliable information of class labels. Higher robustness of C3L could be explained by the fact that this model has an access to the entire data set, not only to its part (as GMM_H and CEC_H).

6.3 Improving clusters boundaries

Projection pursuit (PP) is a technique used for analyzing multivariate data by finding its interesting low dimensional views. One dimensional projection is a common choice, which was widely analyzed in the literature [11, 18]. Projections are usually determined by maximizing non-gaussianity. Pena and Prieto [26] showed that minimizing the kurtosis coefficient implies maximizing the bimodality of the projections, which in consequence is useful for detecting clusters. Since clustering in one dimen-



(a) (b) (c)

Figure 6: C3L with a decision boundary obtained by projection pursuit. Density estimated from true classes in one-dimensional subspace minimizing kurtosis coefficient 6(a). Clusters detected by GMM in projected space 6(b). C3L clustering with decision boundary found by one-dimensional GMM 6(c).

sional subspaces can only generate linear decision boundaries between clusters, PP cannot discover complex data patterns. We will show that given linear separation of data obtained by PP, the use of C3L allows to detect more accurate shapes of clusters.

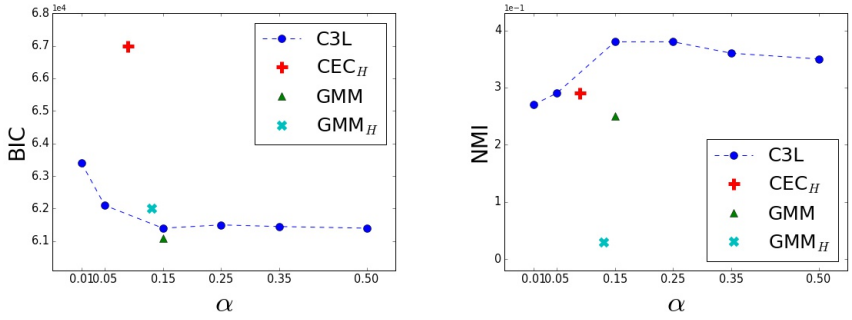
We considered Statlog data set retrieved from UCI repository concerning credit card applications [15]. Each example is represented by 14 attributes and belongs to one of two classes. First class contains 307 instances, while the second one has 383 objects.

We used R package REPPlab[6], which implements several indices for projecting the data on the associated one-dimensional directions. Since we aim at finding clusters, we chose such a direction, which minimizes the kurtosis coefficient of projection. Figure 6(a) presents density estimated from two underlying classes in an optimal one-dimensional subspace. Let us observe that these classes cannot be separated in the reduced space. Given one-dimensional view we applied classical GMM to detect two clusters (see Figure 6(b)). The agreement between true classification and GMM clustering was measured by NMI, which gave the score of 0.24.

GMM clustering in one dimensional space generated linear decision boundary H. In order to improve this clustering, we passed H to C3L, which was run with two clusters for four different leakage levels $\alpha \in \{0.01, 0.05, 0.25, 0.5\}$.

The results presented in Figure 6(c) show that applying C3L with low leakage levels led to the improvement of linear decision boundary produced by GMM in projected space. High values of NMI for $\alpha \in \{0.01, 0.05\}$ and its low values for $\alpha \in \{0.25, 0.5\}$ prove that the information retrieved by applying PP was meaningful. In other words, strictly unsupervised model-based clustering could not find true structure of classes, while combining one-dimensional projection with constrained

[6] `https://github.com/cran/REPPlab/`



(a) (b)

Figure 7: Results for chemical data set.

clustering gave significantly better results.

6.4 Detection of chemical classes

Finally, we considered a real data set of 2497 chemical compounds which was manually clustered into four categories by the experts in the field [39]. Each compound was characterized by its structural features using Klekota-Roth fingerprint (4860 attributes) [14, 31]. To reduce a dimensionality of the space, PCA was applied to attribute vectors and only five principle components were used.

Additionally, every compound was assigned to active or inactive class based on its binding constant $K_i \geq 0$ measured for 5-HT$_{1A}$ receptor, one of the proteins responsible for the regulation of central nervous system: compounds with $K_i < 50$ were considered active while those with $K_i \geq 50$ were treated as inactives [23]. Summarizing, this data set is contained in $\mathbb{R} \times \mathbb{R}^5$ space and a decision boundary is given by $H = \{50\} \times \mathbb{R}^5$. We investigate whether the information of compounds activity allows to obtain a partition which is more similar to the expert reference grouping with four chemical categories. Observe that this case study represents more realistic scenario than previously prepared experiments.

Figures 7(a) and 7(b) show that C3L gave the second best model in terms of BIC (worse than GMM). Moreover, the partition obtained by C3L was the most similar to the reference grouping. Observe that high NMI values coincide with the stabilization of BIC (compare Figure 7(a) with Figure 7(b)). The highest similarity was achieved for $\alpha = 0.15$. C3L with lower leakage levels as well as CEC$_H$ and GMM$_H$ gave worse results, because the binding constant does not reflect exactly the chemical classes (as it was prepared in previous experiment). Since the activity barrier represents a kind of noisy decision boundary it should be taken into account with lower confidence level (higher leakage). This case study demonstrated that

C3L can be successfully applied in natural machine learning problems.

6.5 Summary of the experiments

Experimental study showed that C3L can be successfully applied in a wide range of problems. If we have a coarse categorization of data into two basic classes, C3L can be used to construct clusters, which agree with both expert categorization and true data distribution. In other words, C3L is capable of detecting interesting subgroups, which belong to one of two classes with a fixed probability of error. The quality of such model was verified by applying internal clustering measures such as BIC (Section 6.1) as well as by comparing the results with reference grouping created by a domain expert on various examples of data (Section 6.2).

In the case of partial labeling, where only a small sample of data is categorized, we can first apply any binary classifier to construct approximated decision boundary on the entire data space (Section 6.2). If constructed classification is meaningful (accurate), then subgroups can be found by applying C3L with low leakage levels, while in more uncertain situations the leakage levels should be higher. In particular, we showed that compounds activity delivers small amount of information about structural division of chemical space (Section 6.4).

Finally, we demonstrated that C3L can also be used in strictly unsupervised cases. C3L can improve the results of simpler clustering techniques by introducing nonlinearities in clusters descriptions. In particular, we showed that combining C3L with projection pursuit allows to construct better clustering structure than using projection pursuit or model-based clustering individually (Section 6.3).

7 Conclusion and future work

The paper presented a clustering model, C3L, which integrates information coming from the initial classification with the true structure of data. The idea is based on retrieving Gaussian-like clusters which are contained in one of initial classes within a predefined confidence level. Experimental results prove that our algorithm provides high quality model, which can be used for discovering natural subgroups in partially classified data spaces. In the optimization procedure we restricted the problem to special type of Gaussian densities, which is the main limitation of our model. It is worth to eliminate this assumption in future and extend C3L to arbitrary Gaussian distributions using either analytical calculations or by applying some numerical procedures.

We also plan to examine its practical usefulness in real-life problems. In particular, we will focus on applying C3L in text translation systems to find clusters,

which combine domain knowledge with a distribution of data (see introduction for details). Since Gaussian mixture model does not fit well to high dimensional data, such as text, it might be needed to extend the model to other probability models or to use projections to lower dimensional spaces. In our opinion, combining C3L with projection pursuit is an approach of great practical potential, which needs further studies. While projection pursuit allows to select optimal low dimensional view of data for finding general clusters regions, C3L is able to use this knowledge to discover detailed clusters description.

A Algorithm

Given a density model $\mathcal{G}_{1,N-1}$ and a linear constraint (H, α), where $H = \{0\} \times \mathbb{R}^{N-1}$, the one cluster C3L cost function can be evaluated, as a sum of partial costs (see Corollary 1):

$$E_H^\alpha(X \| \mathcal{G}_{1,N-1}) = E_{\{0\}}^\alpha(X[1] \| \mathcal{G}_1) + h^\times(X[2:N] \| \mathcal{G}_{N-1}).$$

The optimization of $(N-1)$-dimensional density $g_{N-1} \in \mathcal{G}_{N-1}$ is independent of the constraint. Its optimal parameters are the maximum likelihood estimators (MLE) of a mean and covariance of a cluster, i.e.,

$$g_{N-1} = \mathcal{N}(\hat{m}_{X[2:N]}, \hat{\Sigma}_{X[2:N]}). \tag{19}$$

The constraint only affects the form of remaining one dimensional density $g_1 \in \mathcal{G}_1$. Making use of the results and the notations of Theorem 1 it is calculated as:

$$g_1 = \mathcal{N}(m_{X[1]}^\alpha, \sigma_{X[1]}^\alpha). \tag{20}$$

The one cluster cost functions of every group are plugged into the expression (8) and determine the overall C3L cost. Its minimization can be performed in an iterative procedure which is a modified online Hartigan algorithm employed in k-means method [10]. Basically, the procedure consists of two steps: initialization and iteration. In the initialization stage, $k \geq 2$ nonempty groups are formed randomly. Then the elements are reassigned between clusters in order to minimize the criterion function. Presented clustering procedure is non-deterministic and one of the local minima is found [12]. To provide more stable and accurate results, the algorithm has to be run a couple of times and a partition with a minimal cost should be chosen. A pseudocode of C3L is given below:

1: **INPUT:**

$X \subset \mathbb{R}^N$
k - number of clusters
(H, α) - linear constraint
OUTPUT:
Final partition $\mathcal{Y}$ of X
INITIALIZATION:
$X = T_H(X)$ // data transformation such that H // is mapped into $(\{0\} \times \mathbb{R}^{N-1})$
$\mathcal{Y} \leftarrow$ random partition of X into k groups
for all $Y \in \mathcal{Y}$ **do**
  $g^Y_{N-1} \leftarrow MLE(Y[2:N] \| \mathcal{G}_{N-1})$ // use standard MLE to find optimal density (19)
  $g^Y_1 \leftarrow ConstrMLE(Y[1] \| \mathcal{G}_1$ s.t. $(\{0\} \times \mathbb{R}^{N-1}, \alpha))$ // use Theorem 1 for density estimation (20)
  $g^Y \leftarrow g^Y_1 \cdot g^Y_{N-1}$
end for
ITERATION:
while NOT Done **do**
  $Done \leftarrow True$
  for all $x \in X$ **do**
    $Y_{new} \leftarrow arg \max_{Y \in \mathcal{Y}} \Delta E^{\alpha}_{\{0\} \times \mathbb{R}^{N-1}}(x, Y)$ // find a membership of x maximizing the decrease of cost
    if $Y_{new} \neq x.cluster$ **then**
      $Done \leftarrow False$
      $Reassign(x, Y_{new}, Y_{old})$ // reassign x from Y_{old} to Y_{new}
      $Update((Y_{new}, g^{new}), (Y_{old}, g^{old}), x, \alpha)$ // recalculate clusters parameters after the change
    end if
  end for
end while

B Proof of Lemma 1

We consider the limiting case of $\alpha \to 0$. Therefore, without loss of generality we may assume that there exists $\alpha_0 > 0$ such m^{α}_X and σ^{α}_X, for $\alpha < \alpha_0$, are given by (15), i.e.,

$$\begin{aligned} m^{\alpha}_X &:= \frac{-(p^{\alpha})^2 \hat{m}_X + \text{sign}(\hat{m}_X) p^{\alpha} \sqrt{((p^{\alpha})^2+4)\hat{m}^2_X + 4\hat{\sigma}^2_X}}{2}, \\ \sigma^{\alpha}_X &:= \frac{|\hat{m}^{\alpha}_X|}{p^{\alpha}}, \end{aligned}$$

where

$$p^\alpha = \Phi^{-1}_{\mathcal{N}(0,1)}(1-\alpha).$$

Let us calculate the limiting value of m_X^α:

$$\begin{aligned}
&2m_X^\alpha = \\
&-(p^\alpha)^2\hat{m}_X + \operatorname{sign}(\hat{m}_X)p^\alpha\sqrt{((p^\alpha)^2+4)\hat{m}_X^2 + 4\hat{\sigma}_X^2} = \\
&-(p^\alpha)^2\hat{m}_X + (p^\alpha)^2|\hat{m}_X|\operatorname{sign}(\hat{m}_X)\sqrt{1+\tfrac{4}{(p^\alpha)^2}+\tfrac{4\hat{\sigma}_X^2}{(p^\alpha)^2\hat{m}_X^2}} = \\
&-(p^\alpha)^2\hat{m}_X + (p^\alpha)^2\hat{m}_X\sqrt{1+\tfrac{4}{(p^\alpha)^2}\left(1+\tfrac{\hat{\sigma}_X^2}{\hat{m}_X^2}\right)}.
\end{aligned}$$

Making use of Taylor expansion we have

$$\sqrt{1+x} = 1 + \frac{x}{2} + \varepsilon_x \text{ , for } \varepsilon_x = -\frac{1}{8}(1+\xi_x)^{-\frac{3}{2}}\xi_x^2,$$

where $\xi_x \in (0,x)$. Consequently, in our situation there exists $\xi_\alpha \in (0, \frac{4}{(p^\alpha)^2}(1+\frac{\hat{\sigma}_X^2}{\hat{m}_X^2}))$ such that for $\varepsilon_\alpha = -\frac{1}{8}(1+\xi_\alpha)^{-\frac{3}{2}}\xi_\alpha^2$ we get

$$\begin{aligned}
&2m_X^\alpha = \\
&-(p^\alpha)^2\hat{m}_X + (p^\alpha)^2\hat{m}_X\sqrt{1+\tfrac{4}{(p^\alpha)^2}(1+\tfrac{\hat{\sigma}_X^2}{\hat{m}_X^2})} = \\
&-(p^\alpha)^2\hat{m}_X + (p^\alpha)^2\hat{m}_X\left(1+\tfrac{1}{2}\tfrac{4}{(p^\alpha)^2}\left(1+\tfrac{\hat{\sigma}_X^2}{\hat{m}_X^2}\right)+\varepsilon_\alpha\right) = \\
&\underbrace{(p^\alpha)^2\hat{m}_X\frac{2}{(p^\alpha)^2}\left(1+\frac{\hat{\sigma}_X^2}{\hat{m}_X^2}\right)}_{\text{(I)}} + \underbrace{(p^\alpha)^2\hat{m}_X\varepsilon_\alpha}_{\text{(II)}},
\end{aligned}$$

Clearly, (I) converges to $2(\hat{m}_X + \frac{\hat{\sigma}_X^2}{\hat{m}_X})$, as $\alpha \to 0$. We consider the term (II). Observe that for sufficiently small $\alpha > 0$, we have:

$$\begin{aligned}
|(p^\alpha)^2\hat{m}_X\varepsilon_\alpha| &= \tfrac{1}{8}(p^\alpha)^2|\hat{m}_X|(1+\xi_\alpha)^{-\frac{3}{2}}\xi_\alpha^2 \\
&\leq \tfrac{1}{8}(p^\alpha)^2|\hat{m}_X|\cdot 1\cdot\tfrac{16}{(p^\alpha)^4}(1+\tfrac{\hat{\sigma}_X^2}{\hat{m}_X^2})^2 \\
&\leq 2|\hat{m}_X|\tfrac{1}{(p^\alpha)^2}(1+\tfrac{\hat{\sigma}_X^2}{\hat{m}_X^2})^2 \xrightarrow{\alpha\to 0} 0.
\end{aligned}$$

Concluding $m_X^\alpha \to \hat{m}_X + \frac{\hat{\sigma}_X^2}{\hat{m}_X}$, as $\alpha \to 0$.

From the above calculation we directly get,

$$\sigma_X^\alpha = \frac{m_X^\alpha}{p^\alpha} \xrightarrow{\alpha\to 0} 0,$$

which completes the proof.

C Proof of Lemma 2

Let $\alpha > 0$. The violation of the linear constraint $(\{0\}, \alpha)$ by a density $\mathcal{N}(m, \sigma)$ is verified by the condition:

$$|m| \geq p^\alpha \sigma, \text{ where } p^\alpha = \Phi^{-1}_{\mathcal{N}(0,1)}(1-\alpha).$$

Theorem 1 states that the parameters of optimal one dimensional Gaussian density are calculated as MLE within the cluster if only it does not violate the linear constraint. However, in the limiting case there exists α_0 such that for all $\alpha < \alpha_0$ the constraint is violated and consequently the formulas (15) for $m^\alpha_\pm, \sigma^\alpha_\pm$ give optimal solutions.

Let $x > 0$. For normal distributions $g^\alpha_\pm = \mathcal{N}(m^\alpha_\pm, \sigma^\alpha_\pm)$ we have

$$\begin{array}{l} \ln g^\alpha_\pm(x) = \\ -\frac{1}{2}\ln(2\pi) - \ln(\sigma^\alpha_\pm) - \frac{(x-m^\alpha_\pm)^2}{2(\sigma^\alpha_\pm)^2} = \\ -\frac{x^2}{2(\sigma^\alpha_\pm)^2} + \frac{2xm^\alpha_\pm}{2(\sigma^\alpha_\pm)^2} - \frac{(m^\alpha_\pm)^2}{2(\sigma^\alpha_\pm)^2} - \ln(\sigma^\alpha_\pm) - \frac{1}{2}\ln(2\pi). \end{array}$$

Making use of equality $\sigma^2_\pm = \frac{m^2_\pm}{(p^\alpha)^2}$ we get

$$\begin{array}{l} \ln g^\alpha_\pm(x) = \\ \frac{1}{2}\left(-\frac{(p^\alpha)^2}{(m^\alpha_\pm)^2}x^2 + \frac{2(p^\alpha)^2}{m^\alpha_\pm}x - (p^\alpha)^2 - \ln\frac{(m^\alpha_\pm)^2}{(p^\alpha)^2} - \ln(2\pi)\right), \end{array}$$

and consequently

$$\begin{array}{l} \ln g^\alpha_+(x) - \ln g^\alpha_-(x) = \\ \frac{(p^\alpha)^2 x}{(m^\alpha_+)^2(m^\alpha_-)^2}\left(-x(m^\alpha_-)^2 + x(m^\alpha_+)^2 + \right. \\ \left. 2m^\alpha_+(m^\alpha_-)^2 - 2(m^\alpha_+)^2 m^\alpha_-\right) - \ln\frac{(m^\alpha_+)^2}{(m^\alpha_-)^2}. \end{array}$$

Since $m^\alpha_\pm \to \hat{m}_\pm + \frac{\hat{\sigma}^2_\pm}{\hat{m}_\pm}$, as $\alpha \to \infty$ (see Theorem 2) then

$$0 < \frac{x}{(m^\alpha_+)^2(m^\alpha_-)^2} < \infty \text{ and } 0 < \left|\ln\frac{(m^\alpha_+)^2}{(m^\alpha_-)^2}\right| < \infty, \text{ as } \alpha \to 0.$$

Therefore,

$$\ln g^\alpha_+(x) - \ln g^\alpha_-(x) \to +\infty,$$

iff

$$-x(m^\alpha_-)^2 + x(m^\alpha_+)^2 + 2m^\alpha_+(m^\alpha_-)^2 - 2(m^\alpha_+)^2 m^\alpha_- > 0, \text{ as } \alpha \to 0. \tag{21}$$

The inequality (21) holds iff

$$\begin{array}{ll} \text{either:} & x > 2m_+^\alpha \text{ and } m_-^\alpha > \frac{m_+^\alpha x}{2m_+^\alpha - x} \\ \text{or:} & x \leq 2m_+^\alpha. \end{array}$$

In the limiting case the last inequality expands to:

$$x \leq 2m_+^\alpha \xrightarrow{\alpha \to 0} 2\left(\hat{m}_+ + \frac{\hat{\sigma}_+^2}{\hat{m}_+}\right),$$

which completes the proof.

Acknowledgement

This research was partially supported by the National Science Centre (Poland) grant no. 2016/21/D/ST6/00980 and grant no. 2015/19/B/ST6/01819.

References

[1] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[2] Christophe Ambroise, Thierry Denoeux, Gérard Govaert, and Philippe Smets. Learning from an imprecise teacher: probabilistic and evidential approaches. *Applied Stochastic Models and Data Analysis*, 1:100–105, 2001.

[3] L N F Ana and Anil K Jain. Robust data clustering. In *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–128, 2003.

[4] S. Asafi and D. Cohen-Or. Constraints as features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1634–1641, Portland, OR, June 2013.

[5] S. Basu, I. Davidson, and K. Wagstaff. *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

[6] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 59–68. ACM, 2004.

[7] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, 6(Jan):165–188, 2005.

[8] Chris Fraley and Adrian E Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[9] D. Gondek and T. Hofmann. Non-redundant data clustering. *Knowl Inf Syst*, 12(1):1–24, 2007.

[10] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[11] Siyuan Hou and Peter D Wentzell. Re-centered kurtosis as a projection pursuit index for multivariate data analysis. *Journal of Chemometrics*, 28(5):370–384, 2014.

[12] A. K. Jain and P. J. Murty, M. N.and Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, 1999.

[13] H Tolga Kahraman, Seref Sagiroglu, and Ilhami Colak. The development of intuitive knowledge classifier and the modeling of domain dependent data. *Knowledge-Based Systems*, 37:283–295, 2013.

[14] Justin Klekota and Frederick P Roth. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21):2518–2525, 2008.

[15] M. Lichman. UCI machine learning repository, 2013.

[16] H. Liu and Y. Fu. Clustering with partition level side information. In *Proc. IEEE Int. Conf. Data Mining*, pages 877–882. IEEE, 2015.

[17] Nicola Loperfido. Skewness and the linear discriminant function. *Statistics & Probability Letters*, 83(1):93–99, 2013.

[18] Nicola Loperfido. Vector-valued skewness for model-based clustering. *Statistics & Probability Letters*, 99:230–237, 2015.

[19] Zhengdong Lu and Todd K. Leen. Semi-supervised learning with penalized probabilistic clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, Vancouver, British Columbia, Canada, December 2005.

[20] Zhengdong Lu and Todd K Leen. Semi-supervised clustering with pairwise constraints: A discriminative approach. In *AISTATS*, pages 299–306, 2007.

[21] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, Inc., 2004.

[22] R. Nguyen, N.and Caruana. Consensus clusterings. In *Proc. IEEE Int. Conf. Data Mining*, pages 607–612. IEEE, 2007.

[23] Berend Olivier, Willem Soudijn, and Ineke van Wijngaarden. The 5-HT_{1A} receptor and its ligands: structure and function. In Ernst Jucker, editor, *Progress in Drug Research*, volume 52, pages 103–165. 1999.

[24] Witold Pedrycz, Alberto Amato, Vincenzo Di Lecce, and Vincenzo Piuri. Fuzzy clustering with partial supervision in organization and classification of digital images. *IEEE Transactions on Fuzzy Systems*, 16(4):1008–1026, 2008.

[25] Witold Pedrycz and James Waletzky. Fuzzy clustering with partial supervision. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 27(5):787–795, 1997.

[26] Daniel Peña and Francisco J Prieto. Cluster identification using projections. *Journal of the American Statistical Association*, 96(456):1433–1445, 2001.

[27] J. Rissanen. *Minimum description length principle*. Wiley Online Library, 1985.

[28] Aghiles Salah, Nicoleta Rogovschi, and Mohamed Nadif. Model-based co-clustering for high dimensional sparse data. In *Proc. Int. Conf. on Artificial Intelligence and Statistics*, pages 866–874, 2016.

[29] Noam Shental, Aharon Bar-hillel, Tomer Hertz, and Daphna Weinshall. Computing gaussian mixture models with em using equivalence constraints. In S. Thrun, L. K. Saul, and B. Scholkopf, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 465–472. MIT Press, 2004.

[30] Marek Śmieja and Jacek Tabor. Spherical wards clustering and generalized voronoi diagrams. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[31] Marek Śmieja and Dawid Warszycki. Average information content maximization—a new approach for fingerprint hybridization and reduction. *PloS one*, 11(1):e0146666, 2016.

[32] Marek Śmieja and Magdalena Wiercioch. Constrained clustering with a complex cluster structure. *Advances in Data Analysis and Classification*, pages 1–26, 2016.

[33] P Spurek, K Kamieniecki, J Tabor, K Misztal, and M Śmieja. R package cec. *Neurocomputing*, 237:410–413, 2017.

[34] Przemysław Spurek. General split gaussian cross–entropy clustering. *Expert Systems with Applications*, 68:58–68, 2017.

[35] J. Tabor and P. Spurek. Cross-entropy clustering. *Pattern Recognition*, 47(9):3046–3059, 2014.

[36] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. Allerton Conf. on Communication, Control, and Computing*, pages 368–377, Monticello, IL, September 1999.

[37] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584, 2001.

[38] Ziang Wang and Ian Davidson. Flexible constrained spectral clustering. In *Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, pages 563–572, Washington, DC, July 2010.

[39] D. Warszycki, S. Mordalski, K. Kristiansen, R. Kafel, I. Sylte, Z. Chilmonczyk, and A. J. Bojarski. A linear combination of pharmacophore hypotheses as a new tool in search of new active compounds–an application for 5-ht1a receptor ligands. *PloS ONE*, 8(12):e84510, 2013.

[40] Ron Wehrens, Lutgarde MC Buydens, Chris Fraley, and Adrian E Raftery. Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification*, 21(2):231–253, 2004.

[41] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.