Appeared in *CSDA* 54 (2010), 637-654.   DOI 10.1016/j.csda.2009.08.023.

Using combinatorial optimization in model–based trimmed clustering with cardinality constraints

María Teresa Gallegos and Gunter Ritter

Fakultät für Informatik und Mathematik
Universität Passau, Germany[1]

Abstract Statistical clustering criteria with free scale parameters and unknown cluster sizes are inclined to create small, spurious clusters. To mitigate this tendency a statistical model for cardinality–constrained clustering of data with gross outliers is established, its maximum likelihood and maximum a posteriori clustering criteria are derived, and their consistency and robustness are analyzed. The criteria lead to constrained optimization problems that can be solved by iterative, alternating trimming algorithms of k–means type. Each step in the algorithms requires the solution to a λ–assignment problem known from combinatorial optimization. The method allows to estimate the numbers of clusters and outliers. It is illustrated with a synthetic and a real data set.

Key words model–based clustering; classification model; outliers; size constraints; combinatorial optimization; λ–assignment problem; model selection

1 Introduction

Clustering means subdividing data sets into separated, cohesive subsets. Two statistical models are customary for this task: the *classification* or *clustering model* and the *mixture model*. While grouping of data is the main aim of the former which, to this end, makes explicit use of class labels, the latter serves primarily for estimating parameters and mixing rates and classification is secondary; for a comparison see [22], Remark (b) on pp. 354, 355. We apply here the maximum likelihood (ML) and maximum a posteriori (MAP) paradigms to the classification model. If class-conditional distributions are normal then *homoscedastic* and *heteroscedastic* models are customary. The former have equal scale parameters and the latter arbitrary ones. Moreover, both models can be endowed with spherical, diagonal, and full scale parameters so that we have at our disposal six normal models with spherical or elliptical cluster shapes. In these cases much progress towards an automatic data analysis has been made in the past 40 years. In the 1960's, the following cluster *criteria*, optimal for uncontaminated, normal data and applicable in the case of a known number of clusters of about equal size, had been established: 1. The *pooled trace criterion*, Ward Jr. [63]. It is the ML criterion for the homoscedastic, spherical normal model, see Scott and Symons [54]. 2. The *pooled determinant criterion*, Friedman and Rubin [19], the ML criterion for the homoscedastic, full normal model, Scott and Symons [54]. 3. The *heteroscedastic determinant criterion*, an ML criterion for general normal populations, Scott and Symons [54].

By that time it was well understood how criteria could be created based on other normal submodels, e.g. with independent features (the "diagonal" case), and on different cross–cluster constraints. Symons [57] noticed that subtracting the entropy of the cluster proportions from the log–likelihood could in some sense optimally handle unknown and unequal cluster sizes. His criterion may be derived from an MAP model with mixing rates as parameters.

Jancey [32] proposed an efficient *algorithm* for minimizing Ward's criterion. The algorithm was reinvented many times and had earlier been proposed by Steinhaus [56] for partitioning physical masses. It is now called the "k–means" algorithm, a name coined by MacQueen [38] who proposed and analyzed a similar algorithm in a somewhat different situation. The method consists essentially

[1] ritter@fim.uni-passau.de

of an alternating application of ML parameter estimation and ML discriminant analysis. Except for this method it took some time until dedicated algorithms suitable for efficiently reducing the other criteria mentioned above became known. Indeed, Schroeder [52] realized that all criteria could be optimized by similar, alternating algorithms. In the case of general normal populations, e.g., it is sufficient to replace the Euclidean distance used in k–means with the Mahalanobis distance.

The criteria and algorithms designed for the *homoscedastic* cases, in particular the pooled trace criterion and k–means, are popular for their stability, at least if applied to uncontaminated data sets. This is due to their known properties. MacQueen [38] obtained convergence of the pooled variance if the k–means algorithm is applied to a sequence of data points distributed iteratively amongst g clusters. Hartigan [29] proved for dimension $d = 1$ convergence of the cut points and their asymptotic normality. Pollard extended Hartigan's results to Euclidean data. In [46], he proved that Ward's criterion consistently estimates the optimal g cluster centers obtained from the related criterion for the underlying mixture if they are unique. However, these cluster centers differ from the means of the g original populations so that, contrary to the mixture model, the latter are not consistently estimated in the normal clustering case, neither with the ML nor with the MAP approach. The reason is the hard assignment intrinsic to the classification model: no matter how well the populations are separated, e.g. in the case of $g = 2$ clusters, the proportions in the tails on the opposite side of the separating hypersurface are assigned to the wrong cluster. Thus, the variances are under-estimated and the distance between the mean values is over-estimated in the limit as the size of the data set tends to infinity. If the data set is small, anything can happen. However, the better the populations are separated, the more closely the estimated means approach the population means since there is less overlapping. Pollard [47] gives conditions on the mixture under which its g cluster centers are asymptotically normal.

Now, the cluster structure of most real, grouped data sets is heterogeneous and application of homoscedastic models to such data sets may yield absurd results. This is the main reason why *heteroscedastic* models are of major interest. Some of their properties are elementary. It is easy to see that Scott and Symons's heteroscedastic determinant criterion is equivariant w.r.t. affine transformations. The heteroscedastic criteria do not consistently estimate normal population parameters just as described in the homoscedastic case above. Unfortunately, the heteroscedastic determinant criterion encounters some problems that make it unstable and that are of major concern. The first is the *non-existence* of an optimal solution since likelihood and posterior density are unbounded if no suitable constraints are applied. Therefore, the algorithms are inclined towards singular solutions with spurious, small clusters. Such solutions are undesirable and we know of three remedies in the normal cases:

(i) bounding the covariance matrices,

(ii) relaxed homoscedasticity as in Hathaway [30] in the case of the mixture model, and

(iii) bounding cluster sizes below so that small clusters cannot appear, see Rocke and Woodruff [48], Woodruff and Reiners [65].

The reason for the stable performance of k-means (with a possibly unacceptable solution) is that it implicitly assumes (ii) (even spherical homoscedasticity) and (iii) (even approximate equality of all cluster sizes).

Besides existence and consistency, also *robustness* and *algorithmic treatment* need close attention. The classical criteria and algorithms are not robust against outliers. It is well known that the asymptotic breakdown points of the classical methods stated at the beginning even vanish. An effective way of dealing with the related problem of robust *parameter estimation* is Rousseeuw's [50] *minimum–covariance–determinant criterion* (MCD). Rousseeuw and Van Driessen [51] proposed an efficient heuristic for its computation. Like k–means, this algorithm is alternating. Our paper can be viewed as an extension of MCD to robust clustering.

In the spirit of Pollard's [46] result, Cuesta-Albertos et al. [14] define a trimmed homoscedastic criterion and a trimmed extension of the k–means algorithm for g spherical clusters showing consistency in Pollard's sense. Rocke and Woodruff [49], see also Woodruff and Reiners [65], proposed on heuristic grounds a trimmed extension of Scott and Symons's heteroscedastic determinant criterion which they called MINO. In [22], we proposed a homoscedastic statistical model for clustering grouped data with outliers deriving from it a pooled cluster criterion and a related trimming algorithm. We also computed some breakdown points thus showing robustness of criterion and algorithm. Recently, García-Escudero et al. [24] designed a constrained heteroscedastic classification model, studied existence of solutions and convergence, and proposed an algorithm. The authors use eigenvalue constraints of the form (i) above that confine the deviation of the scale parameters from sphericity and from equality. This makes their model equivariant w.r.t. multiplication by a scalar but excludes full affine equivariance. However, the constraints guarantee existence of an optimal solution and consistency in Pollard's sense. Finally the authors design an iterative trimming algorithm that uses polar decomposition in order to rescue iterations that eventually violate the constraints. More recently, Gallegos and Ritter [23] designed and analyzed a robust, affine equivariant, heteroscedastic classification model with trimming based on constraints of the form (ii) above.

We present here a heteroscedastic classification model with arbitrary populations and outliers and an algorithmic method that uses lower bounds on cluster sizes as in (iii) above in order to avoid singular partitions in the heteroscedastic case from the outset. From this model we derive trimmed cluster criteria, Section 2. The size constraints on the clusters guarantee the existence of optimal solutions. They can also be used in order to implement a priori knowledge on cluster sizes. The model extends both Scott and Symons's heteroscedastic normal model and the statistical model with "spurious" outliers established in [22]. In the case of normal ML estimation we retrieve Rocke and Woodruff's MINO. Since we allow unconstrained parameters our model is affine equivariant. As with any classification model, the population parameters are not consistently estimated but we offer a relaxed version of consistency as separation of components increases, Theorems 2.1 and 2.2. The theorems imply also robustness in situations where they apply.

Optimizing the heteroscedastic criteria is not easy and we present monotone, alternating algorithms for this task, the single–point and the multipoint algorithms, Section 3 and Proposition 3.1. We show that the latter, interestingly, leads to a famous problem from combinatorial optimization: λ–assignment. Its solutions automatically satisfy the size constraints. We finally show how to use the algorithms for estimating the numbers of clusters and outliers, Section 4, and report on our experience with two data sets.

The present algorithmic method serves mainly for avoiding singularities in the course of the iterative process. If size constraints are adequately chosen it can produce reasonable solutions by itself. Otherwise, it can be combined to advantage with parameter constraints, e.g., of the form (i) or (ii) above.

2 Statistical model and constrained cluster criteria with trimming

We consider a very general parametric statistical clustering model of n data points with at least $r \leq n$ regular observations from $g \geq 1$ populations (classes) in an arbitrary sample space E. The remaining $n - r$ observations may, but do not have to, be (gross) *outliers*, observations that do not conform to the posited statistical model. In the framework of automatic clustering, only formal definitions are useful. We know of three: according to Barnett and Lewis [5], an outlier comes from a distribution different from the regular distribution. Davies and Gather [15] treat outliers as observations in the "tail" of a distribution. Both concepts are informative requiring a parent distribution. In [22], we introduced another concept: "spuriousness". Spuriousness handles observations that are unpredictable in the sense of obeying no statistical law in a statistical framework. It assumes that each outlier i comes from its own Bayesian statistical model with parameter space Ψ_i and a prior measure τ_i on it. Since each model is observed only once we cannot, and do not wish to, estimate their parameters $\psi_i \in \Psi_i$. We, therefore, consider them nuisances to be integrated out requiring the integral of the likelihood function to be constant. The density function of any random variable $X : \Omega \to E$ w.r.t. some reference measure on E, conditional on a parameter ψ, is denoted by $f_X[\,\cdot \mid \psi]$. (The reference measure on E is arbitrary but kept *fixed*.) The following is the main assumption on the "spurious outliers".

($\mathrm{SV_o}$) A *spurious outlier* $X_i : \Omega \to E$, $i \in 1..n$, obeys a parametric model with parameter $\psi_i \in \Psi_i$ such that the likelihood integrated w.r.t. some prior measure τ_i on Ψ_i satisfies

$$\int_{\Psi_i} f_{X_i}[x \mid \psi_i]\, \tau_i(\,\mathrm{d}\psi_i) = 1, \tag{1}$$

i.e., does not depend on x.

Opposed to the first two concepts above, spuriousness is not restrictive at all, we do not even assume that it contributes much prior statistical information. However, it allows to derive trimming algorithms and leaves the primary rôle to the regular populations which is its strength. In this respect, it is akin to Davies and Gather's [15] view but, instead of a density level, it uses the parameter r, the number of retained elements. The estimation of the outliers themselves and their number needs a more informative model and is postponed to a later stage of the analysis where various values of r have to be examined. There are two important and sufficiently general situations where ($\mathrm{SV_o}$) holds.

(A) The sample space $E = \mathbb{R}^d$ is Euclidean, $\Psi_i = E$, the outliers obey a *location model*

$$X_i = U_i + \psi_i,$$

with some (unknown) Lebesgue continuous, centered random noise $U_i : (\Omega, P) \to E$, and τ_i is flat on Ψ_i. Indeed, in this case, the conditional Lebesgue density is $f_{X_i}[x \mid \psi_i] = f_{U_i}(x - \psi_i)$ and, hence,

$$\int_{\Psi_i} f_{X_i}[x \mid \psi_i]\, \mathrm{d}\psi_i = 1.$$

(B) The parameter sets Ψ_i are singletons and the distribution of X_i is the reference measure on E and so $f_{X_i} = 1$. This case includes the idea of outliers "uniformly distributed" on some domain.

Each *regular* observation X_i comes from one of g populations represented by a density f_{γ_j}, $\gamma_j \in \Gamma_j$, $j \in 1..g$, in some dominated parametric statistical model with parameter set Γ_j.

A *labelling* of the n objects is an array $\boldsymbol{\ell} = (\ell_1, \dots, \ell_n)$, $\ell_i \in 0..g$, where $\ell_i = j \in 1..g$ *retains* object i assigning it to class j and $\ell_i = 0$ *discards* it. We put $C_j = C_j(\boldsymbol{\ell}) = \{i \mid \ell_i = j\}$ and $n_j(\boldsymbol{\ell}) = \#C_j(\boldsymbol{\ell})$, the size of cluster $j \in 1..g$ w.r.t. $\boldsymbol{\ell}$. A labelling is *admissible* if each cluster $j \in 1..g$ contains at least b_j data points and if the set of discarded elements is of size $n - r$. The number $r \le n$ is fixed in advance; as to its choice see Section 4.1. Also the natural numbers $b_j \ge 1$, $\sum_j b_j \le r$, are fixed in advance and may serve two purposes. They may first reflect prior information on cluster sizes. Second, b_j has to be chosen in such a way as to secure ML estimation of the population parameters. Assume, e.g., the full normal model, N_{m_j,V_j}, and general position of the data, i.e. any $d+1$ observations are affine independent. Then, the m.l.e. of (m_j and) V_j exists if and only if $\#C_j \ge d+1$ for all $j \in 1..g$ so that we need $b_j \ge d+1$. Similar statements can be made for normal submodels. If the V_j's are assumed to be *diagonal* and if $x_{i_1,k} \ne x_{i_2,k}$ for any two observations $i_1 \ne i_2$ and all $k \in 1..d$ then it is sufficient that each cluster contains at least two elements. In the *spherical* normal model this minimum number even suffices if observations are only pairwise different. The same can be said about Lebesgue continuous elliptically contoured models. General position of the data is guaranteed with probability 1 if they arise from Lebesgue continuous populations.

Let $\Lambda_{r,\mathbf{b}}$ denote the set of all *admissible* labellings $\boldsymbol{\ell} : 1..n \to 0..g$ of the n objects. Since we have assumed that there are at least r regular observations there is an admissible labelling that retains no outlier. The parameter set of our complete model with g classes and at most $n - r$ outliers is the Cartesian product

$$\Lambda_{r,\mathbf{b}} \times \prod_{j=1}^{g} \Gamma_j \times \prod_{i=1}^{n} \Psi_i.$$

Since it retains exactly r elements some regular observations will be discarded if there are more than r of them. This is, however, a lesser problem than the opposite, retaining a bad outlier, which can even lead to the well–known masking effect, see Davies and Gather [15].

The density of X_i w.r.t. the parameters $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_n)$, $\gamma = (\gamma_1, \cdots, \gamma_g)$, and $\psi = (\psi_1, \cdots, \psi_n)$ is

$$f_{X_i}[x \mid \boldsymbol{\ell}, \gamma, \psi] = \begin{cases} f_{\gamma_j}(x), & \ell_i = j \in 1..g, \\ f_{X_i}[x \mid \psi_i] \quad \text{as in Eqn. (1)}, & \ell_i = 0. \end{cases}$$

We assume that the sequence of observations $(X_i)_{i=1}^n$ is statistically independent. By the product formula, the *classification likelihood* for the data set $\mathbf{x} = (x_1, \ldots, x_n)$ w.r.t. the product reference measure on E^n is thus

$$f_X[\mathbf{x} \mid \boldsymbol{\ell}, \gamma, \psi] = \prod_{j=1}^{g} \prod_{\ell_i = j} f_{\gamma_j}(x_i) \prod_{\ell_i = 0} f_{X_i}[x_i \mid \psi_i].$$

Considering the parameters ψ_i of the outliers nuisances, we deduce from Eqn. (1) the classification likelihood integrated w.r.t. all prior measures τ_i

$$f_X[\mathbf{x} \mid \boldsymbol{\ell}, \gamma] = \prod_{j=1}^{g} \prod_{\ell_i = j} f_{\gamma_j}(x_i). \tag{2}$$

Maximizing first w.r.t. all γ_j's and taking the logarithm, we infer that the ML estimate of $\boldsymbol{\ell}$ is given by the *(size–)constrained trimmed* ML *criterion*

$$\underset{\boldsymbol{\ell} \in \Lambda_{r,\mathbf{b}}}{\operatorname{argmax}} \sum_{j=1}^{g} \sum_{\ell_i = j} \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i), \tag{3}$$

where $\gamma_j(\boldsymbol{\ell})$ is the m.l.e. of the parameter $\gamma_j \in \Gamma_j$ w.r.t. cluster j. In the normal case, the double sum equals a constant minus $\frac{1}{2} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \ln \det S_j(\boldsymbol{\ell})$ with the scatter matrices $S_j(\boldsymbol{\ell})$ w.r.t. $\boldsymbol{\ell}$, so that the criterion becomes the *(size–)constrained trimmed* ML *determinant criterion*

$$\mathrm{ML}_{\boldsymbol{\ell}}(\mathbf{x}) = \underset{\boldsymbol{\ell} \in \Lambda_{r,\mathbf{b}}}{\operatorname{argmin}} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \ln \det S_j(\boldsymbol{\ell}). \tag{4}$$

We thus find the *heteroscedastic determinant criterion* MINO, Rocke and Woodruff [49]. These criteria are appropriate if it is a priori known that clusters are of (about) the same size. If not, empirical studies show that they tend to equalize cluster sizes. The reason may be that the assignment, although it is subject to estimation, is not a genuine parameter but rather a (hidden) variable since it grows with the size of the data set. Another reason is provided by Proposition 3.1(b) below which implies that the criteria (3) and (4) are improved by alternating parameter estimation and ML discriminant analysis. The latter is known to be biased. The following model removes this weakness. We shall see that it replaces ML with MAP discriminant analysis.

In order to account for *unequal* cluster sizes, we assume that the populations $j \in 1..g$ are "switched on" independently and with probability $0 < p_j < 1$, $\sum_{j=1}^g p_j = 1$, for sampling the regular observations. The formula of total probability implies that the regular observations are distributed according to the mixture with density $\sum_{j=1}^g p_j f_{\gamma_j}$. For this reason, the probabilities p_j are called the *mixture rates* or *mixture proportions.* We use a mixed ML and MAP estimator, ML w.r.t. the mixing rates $\mathbf{p} = (p_1, \cdots, p_g)$ and the population parameters γ and MAP w.r.t. the assignment $\boldsymbol{\ell} \in \Lambda_{r,\mathbf{b}}$, i.e., we maximize the conditional density $f[\mathbf{x}, \boldsymbol{\ell} \mid \mathbf{p}, \gamma]$ w.r.t. $\mathbf{p}$, γ, and $\boldsymbol{\ell}$. Now, the product formula, independence of $(\mathbf{p}, \boldsymbol{\ell})$ and γ, and independence of $\mathbf{x}$ and $\mathbf{p}$ conditional on $(\boldsymbol{\ell}, \gamma)$ imply

$$f[\mathbf{x}, \boldsymbol{\ell} \mid \mathbf{p}, \gamma] = P[\boldsymbol{\ell} \mid \mathbf{p}, \gamma] f_X[\mathbf{x} \mid \boldsymbol{\ell}, \mathbf{p}, \gamma] = P[\boldsymbol{\ell} \mid \mathbf{p}] f_X[\mathbf{x} \mid \boldsymbol{\ell}, \gamma]. \tag{5}$$

The set of all n–tuples in $(0..g)^n$ with exactly $n - r$ zeros carries the modified product $\Pi[\boldsymbol{\ell} \mid \mathbf{p}] = \binom{n}{r}^{-1} \prod_{j=1}^g p_j^{n_j(\boldsymbol{\ell})}$ as a probability measure. It is natural to use $\Pi[\cdot \mid \mathbf{p}]$ conditional on $\Lambda_{r,\mathbf{b}}$ as prior $P[\cdot \mid \mathbf{p}]$. Since $p_j > 0$ for all $j \in 1..g$, its normalizing constant $\Pi[\Lambda_{r,\mathbf{b}} \mid \mathbf{p}]$ converges to 1 as $r \to \infty$ by the law of large numbers and, thus, $P[\boldsymbol{\ell} \mid \mathbf{p}]$ is close to $\Pi[\boldsymbol{\ell} \mid \mathbf{p}]$ at least for large r. In order to avoid complex expressions, we use the latter and its optimizer $\mathbf{p}^* = (n_j(\boldsymbol{\ell})/r)_j$ instead of $P[\boldsymbol{\ell} \mid \mathbf{p}]$ in Eqn. (5). Taking also partial maxima w.r.t. γ we infer, similarly as in (3), the (size–) *constrained trimmed* MAP–*criterion*

$$\underset{\boldsymbol{\ell} \in \Lambda_{r,\mathbf{b}}}{\operatorname{argmax}} \Big\{ -r\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\Big) + \sum_{j=1}^g \sum_{\ell_i = j} \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i) \Big\}. \tag{6}$$

In the normal case, we derive in the same way as we derived criterion (4) from (3) the (size–) *constrained trimmed* MAP *determinant criterion*

$$\mathrm{MAP}_{\boldsymbol{\ell}}(\mathbf{x}) = \underset{\boldsymbol{\ell} \in \Lambda_{r,\mathbf{b}}}{\operatorname{argmin}} \Big\{ r\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\Big) + \frac{1}{2} \sum_{j=1}^g n_j(\boldsymbol{\ell}) \ln \det S_j(\boldsymbol{\ell}) \Big\}. \tag{7}$$

The ML criteria (3) and (4) and the MAP criteria differ in the entropy of the cluster proportions, $\mathrm{H}\big(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\big) = -\sum_{j=1}^g \frac{n_j(\boldsymbol{\ell})}{r} \ln \frac{n_j(\boldsymbol{\ell})}{r}$. It discourages equal cluster sizes in which case the entropy is maximal. Symons [57] was the first author to propose it in the uncontaminated normal case. On page 38, he notes that its influence may be slightly too large so that spurious, small clusters may arise; see also Section 4.2. We denote the ML and MAP estimates $\mathrm{ML}_{\boldsymbol{\ell}}(\mathbf{x})$ and $\mathrm{MAP}_{\boldsymbol{\ell}}(\mathbf{x})$ of the assignment also by $\boldsymbol{\ell}^*$. The associated estimates of the parameters are $\gamma_j^* = \gamma_j(\boldsymbol{\ell}^*)$, $m_j^* = m_j(\boldsymbol{\ell}^*) = \overline{x}_{C_j(\boldsymbol{\ell}^*)}$ = the sample mean of the cluster $C_j(\boldsymbol{\ell}^*)$ (a bar indicates mean values), and $V_j^* = S_j(\boldsymbol{\ell}^*) = S_{C_j(\boldsymbol{\ell}^*)}$ = the scatter matrix of the cluster $C_j(\boldsymbol{\ell}^*)$.

The criteria include a number of special cases. First, any state space E is allowed, be it discrete or continuous. Second, various statistical models on E may be used. It is not necessary that each population should belong to the same model. Third, there may be spurious outliers in which case one has to estimate their number, e.g., by varying the value of r, see Section 4.1. One may also collect noise elements in one cluster by choosing a particular model for them, e.g. a distribution uniform on the convex hull of the data, cf. Fraley and Raftery [18]. In this case choose $r = n$. The special case with $g = 1$ and the normal distribution is Rousseeuw's [50] robust parameter estimator MCD.

We conclude this section with two consistency and robustness properties for normal populations that shed some light on the behavior of the size–constrained, trimmed criteria. Contrary to Pollard [46] and García-Escudero et al. [24], we are here interested in the *population parameters* as opposed to the parameters obtained from "clustering" the mixture in g parts, see also the discussion in our introduction. Let $(U_i^{(k)})_{i \geq 1}$, $1 \leq k \leq g$, be g sequences of centered, standard spherical, normal random vectors, let m_k and V_k, $1 \leq k \leq g$, be vectors and positive definite matrices, respectively, and put $X_i^{(k)} = \sqrt{V_k} U_i^{(k)} + m_k$. Of course, $X_i^{(k)} \sim N_d(m_k, V_k)$. Furthermore, let $X_i^{(0)}$, $i \geq 1$, be outliers as in ($\mathrm{SV_o}$). All vectors $U_i^{(k)}$ and $X_i^{(0)}$ are supposed to be independent so that all $X_i^{(k)}$'s are independent, too.

Let the random cluster sizes,

$$\mathbf{N} = (N_1, \ldots, N_g) \colon (\Omega, P) \longrightarrow \{(n_1, n_2, \ldots, n_g) \in \mathbb{N}^g \mid n_1 + \ldots + n_g = r,\ n_k \geq b_k\},$$

be independent of the sequences $\big(X_i^{(0)}\big)_i, \ldots, \big(X_i^{(g)}\big)_i$ and multinomial with parameters r, g, and $(p_1, \cdots, p_g)$, $p_k > 0$, $\sum_k p_k = 1$, conditioned on $N_k \geq b_k$, $k \in 1..g$. We consider the data set $\mathbf{X} = (X_1, \ldots, X_n) = \big(X_1^{(1)}, \ldots, X_{N_1}^{(1)}, X_1^{(2)}, \ldots, X_{N_2}^{(2)}, \ldots, X_1^{(g)}, \ldots, X_{N_g}^{(g)}, X_1^{(0)}, \ldots, X_{n-r}^{(0)}\big)$ of r "regular" observations $(X_1, \ldots, X_r) = \big(X_1^{(1)}, \ldots, X_{N_g}^{(g)}\big)$ and $(n-r)$ "outliers" $(X_{r+1}, \ldots, X_n) = (X_1^{(0)}, \ldots, X_{n-r}^{(0)})$.

For $\boldsymbol{\ell} \in \Lambda_{r,\mathbf{b}}$, we denote the (random) mean vector, SSP matrix, and scatter matrix of $C_j(\boldsymbol{\ell})$ by $\overline{X}_j(\boldsymbol{\ell})$, $W_j(\boldsymbol{\ell})$, and $S_j(\boldsymbol{\ell})$, respectively, $j \in 1..g$.

2.1 Theorem

Let the statistical setup be normal as described above. Assume $b_j \geq (g+1)d + 1$ (and $r \geq \sum b_j$).

(a) P–a.s., the estimates $\boldsymbol{\ell}^*_{(r)} = ML_{\boldsymbol{\ell}}(\mathbf{X})$ (or $\boldsymbol{\ell}^*_{(r)} = MAP_{\boldsymbol{\ell}}(\mathbf{X})$, cf. (4) and (7)) of the assignment for fixed r are eventually correct as

(i) $\|m_j - m_k\| \to \infty, \quad 1 \leq j < k \leq g$,
(ii) $\|m_j - X_i^{(0)}\| \to \infty, \quad j \in 1..g,\ i \in 1..(n-r)$, and
(iii) the d–dimensional volume spanned by any $d+1$ outliers diverges to infinity.

(b) For all $j \in 1..g$, we have P–a.s.

$$\lim_{r\to\infty} \lim_{\text{(i),(ii),(iii)}} (m_j(\boldsymbol{\ell}^*_{(r)}) - m_j) = 0, \quad \lim_{r\to\infty} \lim_{\text{(i),(ii),(iii)}} V_j(\boldsymbol{\ell}^*_{(r)}) = V_j.$$

(c) If there are no outliers, $r = n$, then it is sufficient to require $b_j \geq gd + 1$ and (ii) and (iii) are dropped.

Proof. For all $j \in 1..g$, $k \in 0..g$, let $C_{jk} = C_{jk}(\boldsymbol{\ell})$ denote the subset of $C_j(\boldsymbol{\ell})$ consisting of observations of the form $X_1^{(k)}, \ldots, X_{N_k}^{(k)}$, $k \geq 1$, or of outliers if $k = 0$. By $\overline{X}_{jk}(\boldsymbol{\ell})$ and $S_{jk}(\boldsymbol{\ell})$ we denote the (random) mean vector and scatter matrix, respectively, of $C_{jk}(\boldsymbol{\ell})$ if this set is not empty. The proof of the theorem is based on the following identity of Steiner's type, the special case of [22], Lemma A.3, applied to the partition $\big\{C_{j1}, \ldots, C_{jg}, \{i\}_{i \in C_{j0}}\big\}$ of C_j.

$$\begin{aligned} W_j(\boldsymbol{\ell}) &= \sum_{k=0}^{g} \sum_{i \in C_{jk}} \big(X_i^{(k)} - \overline{X}_j(\boldsymbol{\ell})\big)\big(X_i^{(k)} - \overline{X}_j(\boldsymbol{\ell})\big)^{\mathrm{T}} \\ &= \sum_{k \in 1..g : \#C_{jk} > 0} \#C_{jk} S_{jk}(\boldsymbol{\ell}) \\ &\quad + \sum_{\substack{1 \leq k < l \leq g \\ \#C_{jk}(\boldsymbol{\ell}) \cdot \#C_{j\ell}(\boldsymbol{\ell}) > 0}} \frac{\#C_{jk} \cdot \#C_{j\ell}}{\#C_j} (\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}))(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}))^{\mathrm{T}} \qquad (8) \\ &\quad + \sum_{\substack{1 \leq k \leq g : \#C_{jk}(\boldsymbol{\ell}) > 0 \\ i \in C_{j0}}} \frac{\#C_{jk}}{\#C_j} \big(\overline{X}_{jk}(\boldsymbol{\ell}) - X_i^{(0)}\big)\big(\overline{X}_{jk}(\boldsymbol{\ell}) - X_i^{(0)}\big)^{\mathrm{T}} \\ &\quad + \sum_{\{h,i\} \in \binom{C_{j0}}{2}} \frac{1}{\#C_j} \big(X_h^{(0)} - X_i^{(0)}\big)\big(X_h^{(0)} - X_i^{(0)}\big)^{\mathrm{T}}. \end{aligned}$$

(a) If the assignment $\boldsymbol{\ell}$ generates the "natural" partition then the criteria to be maximized in (4) and (7) are finite and remain unchanged as (i), (ii), and (iii). It is therefore sufficient to show that the criteria of all other partitions diverge to ∞ as (i), (ii) and (iii). From the assumption on

b_j, we first infer that *each* cluster contains at least $d+1$ objects from some population or $d+1$ outliers. From this fact and the distributional assumptions, it follows by a standard argument that the determinants of the scatter matrices of all clusters are bounded away from zero, P–a.s.. By the heteroscedastic determinant criterion (4) it remains, thus, to show that, for each non–natural partition, the determinant of some cluster diverges to ∞ as (i), (ii), and (iii). By what we mentioned before, these partitions are of three types:

- some cluster contains at least $d+1$ elements of some population together with at least one element of another;
- some cluster contains at least $d+1$ elements of some population together with at least one outlier;
- some cluster contains at least $d+1$ outliers.

In the first case, let C_j be a cluster such that $\#C_{jk} \geq d+1$ for some population $k \in 1..g$ and at least one element from population $\ell \neq k$. By definition of $X_i^{(j)}$, we have

$$\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}) = \sqrt{V_k}\, \overline{U}_{jk}(\boldsymbol{\ell}) + m_k - \sqrt{V_\ell}\, \overline{U}_{j\ell}(\boldsymbol{\ell}) - m_l, \tag{9}$$

where $\overline{U}_{jk}(\boldsymbol{\ell})$ is the mean vector of all $U_i^{(j)}$'s with $X_i^{(j)} \in C_{jk}$. The distributional assumptions imply that $S_{jk}(\boldsymbol{\ell})$ is positive definite, P-a.s.. Note that $S_{jk}(\boldsymbol{\ell})$, $\overline{U}_{jk}(\boldsymbol{\ell})$, and $\overline{U}_{j\ell}(\boldsymbol{\ell})$ remain in a bounded set independent of the location parameters $\mathbf{m}$. As a first consequence, if the differences $m_k - m_l$ tend to infinity then so do the differences $\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell})$, P-a.s.. Moreover, the first two sums on the right of Eq. (8) imply (the symbol $\succeq$ denotes the Löwner or positive semi–definite ordering on the space of symmetric matrices)

$$\begin{aligned} W_j(\boldsymbol{\ell}) &\succeq \#C_{jk}\Big(S_{jk}(\boldsymbol{\ell}) + \frac{\#C_{j\ell}}{\#C_j}(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}))(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}))^{\mathrm{T}}\Big) \\ &\succeq S_{jk}(\boldsymbol{\ell}) + \frac{1}{r}(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}))(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell}))^{\mathrm{T}}, \end{aligned}$$

so that the estimate $\det(A + yy^T) \geq (\det A)(1 + y^T A^{-1} y)$ yields

$$\det W_j(\boldsymbol{\ell}) \geq \det S_{jk}(\boldsymbol{\ell})\Big(1 + \frac{1}{r}\big(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell})\big)^{\mathrm{T}} S_{jk}(\boldsymbol{\ell})^{-1}\big(\overline{X}_{jk}(\boldsymbol{\ell}) - \overline{X}_{j\ell}(\boldsymbol{\ell})\big)\Big), \quad P\text{–a.s..}$$

Hence, from (9)

$$\lim_{\substack{\|m_l - m_k\| \to \infty \\ 1 \leq l < k \leq g}} \det S_j(\boldsymbol{\ell}) = \infty, \quad P\text{–a.s..}$$

This is the claim in the first case.

In the second case, let C_j be a cluster that contains $\geq d+1$ elements from population k and at least one outlier $X_i^{(0)}$. The first and third sums on the right of (8) show

$$W_j(\boldsymbol{\ell}) \succeq \#C_{jk}\Big(S_{jk}(\boldsymbol{\ell}) + \frac{1}{\#C_j}(\overline{X}_{jk}(\boldsymbol{\ell}) - X_i^{(0)})(\overline{X}_{jk}(\boldsymbol{\ell}) - X_i^{(0)})^{\mathrm{T}}\Big).$$

By (ii), the difference $m_k - X_i^{(0)}$ is unbounded. As in the first case, one shows that $\overline{X}_{jk}(\boldsymbol{\ell}) - X_i^{(0)}$, too, is unbounded. This fact implies again $\det S_j(\boldsymbol{\ell}) \to \infty$.

In the third case, let cluster C_j contain $d+1$ outliers. From (iii) and [22], Lemma 4.1, we infer that the determinant of their scatter matrix diverges to ∞ and, by the last sum on the right of (8), we have again $\det S_j(\boldsymbol{\ell}) \to \infty$. This proves Claim (a).

(b) Part (a) implies that, for fixed $r \geq \sum_j b_j$ and all $j \in 1..g$, the optimal assignment $\boldsymbol{\ell}^*_{(r)}$ satisfies P–a.s. eventually as (i), (ii), and (iii)

$$m_j(\boldsymbol{\ell}^*_{(r)}) = \overline{X}_j(\boldsymbol{\ell}^*_{(r)}) = \frac{1}{N_j}\sum_{i=1}^{N_j} X_i^{(j)}, \tag{10}$$

$$V_j(\boldsymbol{\ell}^*_{(r)}) = S_j(\boldsymbol{\ell}^*_{(r)}) = \frac{1}{N_j}\sum_{i=1}^{N_j} \big(X_i^{(j)} - \overline{X}_j(\boldsymbol{\ell}^*_{(r)})\big)\big(X_i^{(j)} - \overline{X}_j(\boldsymbol{\ell}^*_{(r)})\big)^{\mathrm{T}}. \tag{11}$$

Let $j \in 1..g$. The strong law implies $\lim_{r\to\infty} \frac{N_j}{r} = p_j$ (> 0) and, hence, $\lim_{r\to\infty} N_j = \infty$, P–a.s.. From Eqns. (10) and (11), it follows again by the strong law

$$\lim_{r\to\infty} (m_j(\boldsymbol{\ell}^*_{(r)}) - m_j) = \lim_{r\to\infty} (\overline{X}_j(\boldsymbol{\ell}^*_{(r)}) - m_j) = 0 \qquad \text{and}$$
$$\lim_{r\to\infty} V_j(\boldsymbol{\ell}^*_{(r)}) = \lim_{r\to\infty} S_j(\boldsymbol{\ell}^*_{(r)}) = V_j,$$

P–a.s.. This is Claim (b). □

Condition 2.1(a)(iii) says that we have actually outliers and not "inliers".

In the normal case, the criteria (4) and (7) work if $b_j \geq d + 1$. In the theorem we need, however, the more stringent condition $b_j \geq (g + 1)d + 1$. The reason is that under the milder condition spurious clusters composed of $d + 1$ elements of different populations in or close to a hyperplane cannot be avoided under (i), (ii), and (iii). These would appear in an optimal partition rendering Parts (a) and (b) of the theorem incorrect. However, the requirement on b_j may be relaxed at the cost of restricting the statistical model. This is the content of the following theorem.

2.2 Theorem

Assume $b_j \geq d+1$, $\kappa > 0$, $U_i^{(j)}$ as above, m_j pairwise different, $X_i^{(j)} = \sqrt{V_j}U_i^{(j)} + \kappa m_j$, $1 \leq j \leq g$, and let $X_i^{(0)} = \kappa U_i^{(0)} + \psi_i$ with Lebesgue continuous random vectors $U_i^{(0)}$.

(a) P–a.s., the estimates $\boldsymbol{\ell}^*_{(r)} = ML_{\boldsymbol{\ell}}(\mathbf{X})$ (or $\boldsymbol{\ell}^*_{(r)} = MAP_{\boldsymbol{\ell}}(\mathbf{X})$, cf. (4) and (7)) for fixed r are correct if κ is large enough.

(b) For all $j \in 1..g$, we have P–a.s.

$$\lim_{r\to\infty} \lim_{\kappa\to\infty} \frac{1}{\kappa} m_j(\boldsymbol{\ell}^*_{(r)}) = m_j, \quad \lim_{r\to\infty} \lim_{\kappa\to\infty} V_j(\boldsymbol{\ell}^*_{(r)}) = V_j.$$

Proof. A simple geometric argument based on the P–a.s. affine independence of $d+1$ of the $U_i^{(j)}$'s shows that the d–dimensional volume of $d + 1$ mixed observations or of $d + 1$ outliers diverges to ∞ as $\kappa \to \infty$. Therefore, an application of [22], Lemma 4.1, shows that the determinants of the scatter matrices of *mixed* clusters or of clusters composed solely of outliers diverge P–a.s. to ∞ as $\kappa \to \infty$. On the other hand, those of clusters composed of a *single* population are independent of κ and P–a.s. > 0. Therefore, an optimal partition must P–a.s. be natural if κ is large enough. This proves Part (a) and the proof of Part (b) is similar to that of Theorem 2.1(b).□

Note that Parts (a) of the theorems make a statement on the robustness of the constrained trimmed criteria: if separation is good and if the outliers are sufficiently spread out then they are detected and discarded by the criteria.

3 Clustering algorithms

In some rare cases there are simple algorithms for computing the ML estimates of the parameters γ_j given the admissible assignment $\boldsymbol{\ell}$. Popular examples are the normal and coin tossing models where ML estimation reduces to simple summation, see e.g. Criterion (4). Even then, maximizing the Criteria (3) and (6) w.r.t. the combinatorial structure of all admissible assignments $\Lambda_{r,\mathbf{b}}$ is not a simple task. Only for the MCD estimator was it recently shown that optimization can be carried out in polynomial time. In fact, Bernholt and Fischer [7] designed an algorithm of complexity $\mathcal{O}(n^v)$, $v = d(d{+}3)/2{+}1$, based on elliptical separability of regular objects and outliers. In general, one has recourse to general optimization schemes such as local descent methods combined with multistart or MCMC algorithms, cf. the discussion in [49], Section 3.1 and [65]. A shortcoming of these methods is the need to update the parameters with each move, even the unsuccessful ones. More efficient dedicated algorithms detect whether a move will be successful *before* the parameters are updated. They consist of the iteration of so–called *reduction steps*. Each reduction step combines ML estimation of the parameters γ_j based on the current assignment with the subsequent removal of misfits based on the new parameters. The following proposition states a practicable condition for a new assignment to improve the criterion. It extends a result of Schroeder's [52] to (spurious) outliers and arbitrary cluster sizes. A version for the m.l.e. in the pooled normal case appears in [22].

3.1 Proposition

Let $\boldsymbol{\ell}$ and $\boldsymbol{\ell}_{\rm new}$ be two admissible labellings such that

$$\sum_{j=1}^{g}\sum_{i:\ell_{{\rm new},i}=j}\big(\ln n_j(\boldsymbol{\ell}) + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\big) > \sum_{j=1}^{g}\sum_{i:\ell_i=j}\big(\ln n_j(\boldsymbol{\ell}) + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\big). \tag{12}$$

(a) Then $\boldsymbol{\ell}_{\rm new}$ strictly improves the *MAP*–criterion (6) w.r.t. $\boldsymbol{\ell}$.

(b) The same holds for the *ML* criterion (3) after dropping the summand $\ln n_j(\boldsymbol{\ell})$ on both sides of Estimate (12).

Proof. We give the proof for Case (a), Case (b) is similar. Applying (12), the entropy inequality, and ML estimation in this order, we have the following chain of estimates.

$$\begin{aligned}
&-r\mathrm{H}\left(\left(\frac{n_j(\boldsymbol{\ell})}{r}\right)_j\right) + \sum_j\sum_{i:\ell_i=j}\ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\\
&= \sum_j\sum_{i:\ell_i=j}\Big(\ln\frac{n_j(\boldsymbol{\ell})}{r} + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\Big)\\
&< \sum_j\sum_{i:\ell_{{\rm new},i}=j}\Big(\ln\frac{n_j(\boldsymbol{\ell})}{r} + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\Big)\\
&= \sum_j n_j(\boldsymbol{\ell}_{\rm new})\ln\frac{n_j(\boldsymbol{\ell})}{r} + \sum_j\sum_{i:\ell_{{\rm new},i}=j}\ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\\
&\le \sum_j n_j(\boldsymbol{\ell}_{\rm new})\ln\frac{n_j(\boldsymbol{\ell}_{\rm new})}{r} + \sum_j\sum_{i:\ell_{{\rm new},i}=j}\ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)\\
&\le \sum_j n_j(\boldsymbol{\ell}_{\rm new})\ln\frac{n_j(\boldsymbol{\ell}_{\rm new})}{r} + \sum_j\sum_{i:\ell_{{\rm new},i}=j}\ln f_{\gamma_j(\boldsymbol{\ell}_{\rm new})}(x_i)\\
&= -r\mathrm{H}\left(\left(\frac{n_j(\boldsymbol{\ell}_{\rm new})}{r}\right)_j\right) + \sum_j\sum_{i:\ell_{{\rm new},i}=j}\ln f_{\gamma_j(\boldsymbol{\ell}_{\rm new})}(x_i).
\end{aligned}$$

□

The fact that both sides in the hypothesis of Proposition 3.1 contain the *current* population parameters $\gamma_j(\boldsymbol{\ell})$ substantially reduces the complexity of the optimization. The proposition may be exploited for designing several reduction steps depending on the transitions from $\boldsymbol{\ell}$ to $\boldsymbol{\ell}_{\rm new}$

employed. The simplest one is defined by elementary moves, the change of an object from a cluster with a surplus of elements to another or the swap of two objects,

(c) $i : \ell_i \to j$, $\quad \ell_i, j \in 1..g,\ l_i \neq j$, if the size of cluster ℓ_i is $> b_{\ell_i}$;
(s) $i : l_i \to j,\ k : j \to l_i$, $\quad l_i \in 0..g,\ l_i \neq j = \ell_k \in 1..g$.

Both moves conserve admissibility. Combined they make the configuration space $\Lambda_{r,\mathbf{b}}$ a connected graph. Denote the assignment resulting from move (c) or (s) by $\boldsymbol{\ell}_{\text{new}}$. The difference between the left and right sides of (12) is easily computed. We have

$$u^{(c)}(i,j) = \ln \frac{n_j(\boldsymbol{\ell})}{n_{\ell_i}(\boldsymbol{\ell})} + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i) - \ln f_{\gamma_{\ell_i}(\boldsymbol{\ell})}(x_i), \qquad \text{and}$$

$$u^{(s)}(i,j,k) = \begin{cases} \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i) - \ln f_{\gamma_j(\boldsymbol{\ell})}(x_k), & l_i = 0, \\ \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i) - \ln f_{\gamma_{\ell_i}(\boldsymbol{\ell})}(x_i) + \ln f_{\gamma_{\ell_i}(\boldsymbol{\ell})}(x_k) - \ln f_{\gamma_j(\boldsymbol{\ell})}(x_k), & l_i \neq 0. \end{cases}$$

If the difference is strictly positive for some admissible move then the proposition asserts that the move improves the MAP–criterion (6). For the ML criterion (3) the term involving the cluster sizes is omitted. This leads to the following *reduction step.*

3.2 The single–point reduction step

```
// Input: An admissible labelling $\boldsymbol{\ell}$;
// Output: an admissible labelling $\boldsymbol{\ell}_{\text{new}}$ with larger criterion *or* the response "stop".

1. Compute the parameters $\gamma_j(\boldsymbol{\ell})$ w.r.t. $\boldsymbol{\ell}$, using update formulae if possible;
2. determine the (admissible) move (c) or (s) with maximal value of $u^{(c)}$ or $u^{(s)}$;
3. *if* this maximum is $> 0$ then update $\boldsymbol{\ell}$ accordingly and return the updated labelling;
   *else* respond "stop".
```

It is possible to stop the maximization in step 2 as soon as some value $u^{(c)} > 0$ or $u^{(s)} > 0$ is found.

The proposition suggests also a *multipoint* reduction step for improving the MAP–criterion. It often outperforms the single–point reduction step w.r.t. speed. A naive version is this: Given a labelling $\boldsymbol{\ell}$, compute the sizes $n_j(\boldsymbol{\ell})$ and the parameters $\gamma_j(\boldsymbol{\ell})$ and assign each observation x_i to the cluster $j = \ell_{\text{new},i}$ with maximum value $\ln n_j(\boldsymbol{\ell}) + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i)$. If the sum of the largest r of these numbers exceeds the sum for the original labelling then the proposition assures us that the new labelling $\boldsymbol{\ell}_{\text{new}}$ has larger MAP–criterion (6) *provided* it is admissible. A similar statement is true for the ML criterion (3) (drop the numbers $n_j(\boldsymbol{\ell})$). This is the main scheme of an efficient estimation procedure. As an advantage over a single–point reduction step it allows to explore the whole data set without updating the parameters after each successful reassignment of a single point. Since $\frac{n_j(\boldsymbol{\ell})}{r} f_{\gamma_j(\boldsymbol{\ell})}(x_i)$ is just the posterior density given the parameters of the current labelling used in discriminant analysis, the step is intuitively appealing since it assigns each observation to the class determined by the MAP (or ML) discriminant rule w.r.t the current parameters.

Unfortunately, in the present heteroscedastic case, the naive procedure just described does not guarantee admissibility of the new assignment, in particular if there is a small cluster or if the constraints b_j are large. This is contrary to the homoscedastic case where deficient or even empty clusters do not prevent a partition from being admissible. This fact renders the above procedure often unstable. In order to ensure admissibility, the multipoint reduction step actually requires solving the following *constrained* optimization problem.

3.3 The multipoint optimization problem

For given numbers n_j and parameters γ_j, maximize $\sum_{i:\ell_i\neq 0}\big(\ln n_{\ell_i}+\ln f_{\gamma_{\ell_i}}(x_i)\big)$ over all admissible labellings $\boldsymbol{\ell}=(\ell_1,\dots,\ell_n)$, i.e., subject to the constraints

$$\#\{i\in 1..n \mid \ell_i=j\}\ge b_j,\quad j\in 1..g,\qquad\text{and}$$
$$\#\{i\in 1..n \mid \ell_i=0\}=n-r.$$

This looks like a difficult problem at first sight. Yet it belongs to the class of problems known to be tractable in computer science. In fact, we will show in Appendix A that finding the labelling $\boldsymbol{\ell}$ may be transformed to finding some binary assignment matrix $\mathbf{z}$ by means of a certain λ–assignment problem which we now describe. The parameters of the problem are the *weights*

$$u_{i,j}=\ln n_j+\ln f_{\gamma_j}(x_i),\quad i\in 1..n,\ j\in 1..g, \tag{13}$$

and the constraints b_j, $j\in 1..g$. The transformed problem uses an artificial $(g+1)$th class with associated weights

$$u_{i,g+1}=\max_{j\in 1..g}u_{i,j}.$$

This class serves to accommodate the excess members w.r.t. the constraints $\ge b_j$ of the g natural classes. The discarded elements do not need weights $u_{i,0}$; they could be put to any constant value if necessary, e.g. 0. The announced λ*–assignment problem* is

$$(\lambda\text{A})\ \sum_{i=1}^{n}\sum_{j=1}^{g+1}u_{i,j}z_{i,j}\ \text{maximal over all matrices } \mathbf{z}\in\mathbb{R}^{n\times(g+2)} \text{ subject to the constraints}$$

$$\begin{cases}\sum_j z_{i,j}=1, & i\in 1..n,\\ \sum_i z_{i,0}=n-r, & \\ \sum_i z_{i,j}=b_j, & j\in 1..g,\\ \sum_i z_{i,g+1}=r-\sum b_j, & \\ z_{i,j}\ge 0, & i\in 1..n,\ j\in 0..(g+1).\end{cases}$$

A matrix $\mathbf{z}$ that satisfies the constraints of (λA) will be called *feasible*. Contrary to an admissible solution, a feasible solution is defined by equalities instead of inequalities.

According to Appendix A, (λA) has a binary solution $\mathbf{z}^*$ with exactly one entry 1 in each line, a fact that justifies the name *assignment matrix*. Moreover, we show in Appendix A that the assignment matrix $\mathbf{z}^*$ induces a solution $\boldsymbol{\ell}^*$ to the multipoint optimization problem, namely

$$\ell_i^*=\begin{cases}j, & \text{if } \mathbf{z}_{i,j}^*=1 \text{ and } j\le g,\\ \text{the natural class of } i, & \text{if } \mathbf{z}_{i,g+1}^*=1;\end{cases} \tag{14}$$

here, the "natural class" of object i is $\operatorname{argmax}_{j\in 1..g}u_{i,j}$.

Appendix B is devoted to a survey on efficient algorithms for solving (λA). We note that simultaneous lower and upper constraints $b_j\le n_j\le c_j$, and hence equality constraints, can be treated in a similar way.

Using $\boldsymbol{\ell}^*=(\ell_1^*,\dots,\ell_n^*)$ as $\boldsymbol{\ell}_{\text{new}}$ in Proposition 3.1, we obtain the following multipoint reduction step.

3.4 The multipoint reduction step

// Input: An admissible labelling, its parameters $\gamma_j(\boldsymbol{\ell})$, and its criterion;
// Output: an admissible labelling $\boldsymbol{\ell}_{\text{new}}$ with its parameters and (larger) criterion
or the response "stop".

1. solve the λ–assignment problem (λA) with weights

$$\begin{cases} u_{i,j} = \ln n_j(\boldsymbol{\ell}) + \ln f_{\gamma_j(\boldsymbol{\ell})}(x_i), & j \in 1..g, \\ u_{i,g+1} = \max_{j\in 1..g} u_{i,j}, \end{cases} \qquad i \in 1..n;$$

2. determine the optimal assignment according to Eqn. (14);
3. compute parameters and criterion of this assignment;
4. *if* the criterion has improved, return it together with this solution and its parameters; *else* "stop".

A reduction step is the combination of parameter estimation and discriminant analysis. In this sense, the multipoint reduction step may be viewed as a combination of parameter estimation and *constrained* discriminant analysis. The transportation problem in connection with constrained discriminant analysis appears already in Tso et al. [62]. These authors deal with automatic chromosome classification using the constraints to ensure the correct number of chromosomes of each class in a biological cell. An application of λ–assignment to a constrained, least–squares clustering problem with *fixed* cluster centers and sizes appears in Aurenhammer et al. [4].

3.5 Overall algorithms

Reduction steps receive a labelling and improve it according to Proposition 3.1. Their iteration thus gradually increases the criterion. Since there are only finitely many labellings the iteration must stall after a finite number of steps with the "stop" signal. This process takes typically a few or a few tens of reduction steps. The returned partition is *one* proposal for the requested solution. It is self–consistent in the sense that it reproduces its parental parameters. While an optimum of the *criterion* shares this property, the solution obtained from a single iteration does, in general, not optimize the criterion. In fact, clustering is known to be NP–hard, see Garey and Johnson [25], and we cannot expect to find an optimal solution at all except in simple cases. All we can seek is a solution with a large value of the criterion in the time available. In many cases this is even sufficient. It can be achieved by application of the multistart method, e.g. with random initial labellings. Of course, the number of replications that lead to a solution near the optimum depends heavily on the size and structure of the data set, on the initial labellings, on the parameters g and r, and on the statistical model chosen. As a rule of thumb, we begin with a minimum number of replications, say 1000, dynamically extending this number to at least ten times the number needed for the last record. So, if the last record among the first 1000 replications was at replication 30, we stop at 1000 hoping that no new record will appear later. If it was at 900, we run the algorithm for at least 9000 replications, and so on. If this procedure does not stop in the time available then use the last record.

A word is in order concerning the choice of the lower bounds b_j which appear as additional parameters of the algorithm. In order to reduce the number of parameters, we set all b_j's to the same value, b. The value for b itself must be detected by experiment. Starting with a low value, say $d + 1$, we raise b until the size of the smallest cluster is substantially larger than b. Solutions $\boldsymbol{\ell}$ satisfying $\min_j n_j(\boldsymbol{\ell}) = b$ are in most cases forced and undesirable. This method will, of course, not detect small clusters which are then found among the discarded elements.

Reduction steps can also be based on heuristics, cf. App. B.2. However, since they need not improve the criterion, they do not necessarily terminate at a self–consistent solution. Iteration of single–point or (heuristic or exact) multipoint reduction steps thus gives rise to three stable optimization methods for *fixed* numbers of clusters and discarded elements. They produce reasonable solutions even if a cluster tends towards becoming too small by the end of an iteration. The multipoint is superior to the single–point algorithm w.r.t. speed. In complex situations, it is useful to refine the result of a multipoint search by single–point steps or local search.

Each series of reduction steps needs an initial solution to begin with. We pick a subset of size r uniformly at random, the retained elements. There exists a simple but sophisticated algorithm that accomplishes just this in one sweep through the data, see Knuth [35], p.136 ff. Moreover, we assign the objects of this set to g clusters again uniformly at random. The clusters must be large enough to allow estimation of their parameters. In general, this requirement does not pose a problem unless g is large. In this way, the clusters will be of about equal size with very high probability. It follows that the entropy of the mixing rates is large at the beginning making it easier to fulfill Hypothesis (12) of Proposition 3.1. Coleman et al. [12] compare the effectiveness of some initial solutions.

4 Number of clusters and number of outliers

4.1 Methods

Application of our method needs the parameters g and r. The first is the "number of clusters" and the second the number of retained elements which we are now going to use in order to estimate the "number of outliers." There is no unanimous concept of "cluster" or "outlier" and these notions are even overlapping – outliers may be traded for clusters: a sufficient number of "outliers" may give rise to a cluster and it may sometimes even be reasonable to assume the absence of outliers. In the present statistical model, outliers either originate from a rare class of which b_j elements have not been observed (Rocke and Woodruff [49], Section 4) or they constitute a set whose structure, e.g. shape, is not in harmony with the posited populations. In the first case, the outliers may find some extreme cluster members that complement them to a sufficiently large cluster. The model often accommodates also clusters whose structures do not conform with its populations. Therefore, in both cases, the numbers of clusters and outliers are subject to interpretation.

Nevertheless, *intervals* containing the two numbers can be a priori given. The parameter estimates may break down under the influence of a single remaining gross outlier and a missing cluster generally forces the algorithm to unite two clusters. It is therefore important to explore sufficiently small values of r and sufficiently large values of g. As a first step, establish a table of the optimal solutions w.r.t. the constrained trimmed ML or MAP criterion, see (3) and (6), for all (reasonable) numbers of clusters, g, and all (reasonable) numbers of discarded elements, $n - r$. In subsequent steps, reduce these solutions to the one or the few that seem most promising. If, for a given value of g, the number r is only inessentially smaller than the optimal then each cluster will just loose a few extreme members but the parameters will not be too much affected. It is therefore justified to use a lacunary subset of values r for each g; see also Rousseeuw and Van Driessen's [51], Section 5, discussion about the choice of r. If one assumes one or two more clusters than there actually exist then empirical observations show that clusters of minimal sizes b_j are split off the natural ones by the MAP method without much changing the estimated parameters of the latter.

We have so far reduced the original clustering problem to analyzing a table of potential solutions indexed by the chosen pairs $(g, n - r)$. This is a substantial reduction of the complexity of the problem but not yet its solution. If the optimal partition $\boldsymbol{\ell}^*$ for a pair $(g, n - r)$ contains outliers then they must show up in at least one cluster. In order to single out those pairs that do not contain outliers we use a multiple-testing procedure. Let $X^{(j)}$ be the random variable from which $C_j(\boldsymbol{\ell}^*)$ is sampled and consider a simple test for the hypotheses $H_0^j \colon X^{(j)} \sim \gamma_{C_j(\boldsymbol{\ell}^*)} =$ the ML estimate of γ for $C_j(\boldsymbol{\ell}^*)$, $j \in 1..g$. It is natural to delete all pairs $(g, n-r)$ for which the composite hypothesis $H_0 = \bigcap_j H_0^j$ is rejected. According to Roy's *union intersection test*, see Mardia et al. [39], Section 5.2.2, this means rejecting H_0^j for some j at some level of significance, say $\alpha = 0.1$. In other words, a pair $(g, n - r)$ and its optimal solution $\boldsymbol{\ell}^*$ is kept in this phase if the test rejects none of its clusters. There is a plethora of tests available for the simple hypotheses H_0^j. We mention tests for *goodness of fit* of the densities $f_{\gamma_j(\boldsymbol{\ell}^*)}$ with the clusters $C_j(\boldsymbol{\ell}^*)$, $j \in 1..g$, *normality tests*, see Mecklin and Mundfrom's [41] extensive survey article, and methods for *outlier detection* or *identification*, see Becker and Gather [6]. If g admits an acceptable pair $(g, n-r)$, keep the one with maximum r as a candidate. There now remains at most one entry per line in the table so that the complexity of the problem is again reduced.

If this procedure leaves no solution at all then the assumptions on the regular populations or the basic assumption that the data set in hand is composed of clusters and outliers is questionable. Otherwise, there are essentially three approaches for selecting the favorite number g, cf. [42, 27], *cluster validation*, the so-called *elbow criterion*, and *model selection criteria*. Cluster validation may be divided in two branches: tests and validity measures. The classical test, due to Wolfe [64], is a likelihood ratio test for the hypothesis of k clusters against $(k-1)$ clusters. Bock [9] discusses some significance tests for distinguishing between the hypothesis of a homogeneous population vs. the alternative of heterogeneity. Chen et al. [11] propose a modified likelihood ratio test for $g = 2$ vs. $g \geq 3$. Validity measures are functionals of partitions and usually measure the quality of cluster separation and of cluster cohesion (or "compactness"); see, e.g., Bezdek et al. [8]. Often, the total within–cluster sum of squared distances about the centroids is used as a measure of cohesion and the total between–cluster sum of squared distances for separation; cf. Milligan and Cooper [42] and the abridged presentation of their work by Gordon [27]. The elbow criterion identifies the number of clusters as the location where the decrease of some cluster criterion flattens markedly. For a recent refinement of this method we refer the reader to Tibshirani et al. [58].

Maximum likelihood and maximum a posteriori estimation tend towards a large number of clusters. A *model selection criterion* counteracts this tendency by subtracting a penalty term from the maximum of the log–likelihood or of the posterior log–density. Schwarz [53] proposed his popular Bayesian Information Criterion (BIC) for exponential families. In the uncontaminated case, its penalty term is $\frac{q}{2} \cdot \ln n$, q being the total dimension of the parametric model. There is some practical evidence that supports BIC as a means for estimating the number of clusters of *mixture models*, too; see the discussion in McLachlan and Peel [40], Ch. 6. Moreover, Kéribin [33] described a family of penalty terms, among them BIC, which *asymptotically* as $n \to \infty$ neither over– nor underestimate the correct number of components of a mixture model $\sum_i \sum_{j=1}^{g} p_j \ln f_{\gamma_j}$ if the class–conditional populations satisfy certain regularity conditions and the parameters certain constraints. Her interesting result is applicable, e.g., to Gaussian families if the mean values are bounded and if the covariance matrices are bounded below in the Löwner ordering by a positive multiple of the identity matrix. In the case of a mixture, $q = q(g)$ is $g - 1$ (for the mixing rates) plus the sum of the dimensions of the g population models.

BIC with this value of q may be applied also to the MAP Criterion (6) for normal classification if separation is sufficiently good. Indeed, let $\boldsymbol{\ell}^*$ be the optimal MAP–assignment and let p^* and γ^* be the optimal mixing rates and population parameters of a mixture model under suitable constraints as in Kéribin's theorem. For any g, the optimal value of Criterion (6) is no larger than that of the mixture model: Assuming without loss $r = n$, we have

$$\begin{aligned}
&-n\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell}^*)}{n}, \ldots, \frac{n_g(\boldsymbol{\ell}^*)}{n}\Big) + \sum_{j=1}^{g} \sum_{\ell_i^*=j} \ln f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i) = \sum_i \Big\{ \ln \frac{n_{\ell_i^*}(\boldsymbol{\ell}^*)}{n} + \ln f_{\gamma_{\ell_i^*}(\boldsymbol{\ell}^*)}(x_i) \Big\} \\
&= \ln \prod_i \frac{n_{\ell_i^*}(\boldsymbol{\ell}^*)}{n} f_{\gamma_{\ell_i^*}(\boldsymbol{\ell}^*)}(x_i) \leq \ln \prod_i \sum_j \frac{n_j(\boldsymbol{\ell}^*)}{n} f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i) \leq \max_{\mathbf{p},\gamma} \ln \prod_i \sum_j p_j f_{\gamma_j}(x_i) \\
&= \ln \prod_i \sum_j p_j^* f_{\gamma_j^*}(x_i). \qquad (15)
\end{aligned}$$

On the other hand, if the data set is well separated in g clusters then, at least in the normal case, $f_{\gamma_j^*}(x_i) \ll f_{\gamma_{\ell_i^*}^*}(x_i)$ for all $j \neq \ell_i^*$, $1 \leq i \leq n$, $f_{\gamma_{\ell_i^*}^*}(x_i) \approx f_{\gamma_{\ell_i^*}(\boldsymbol{\ell}^*)}(x_i)$, and $p_j^* \approx \frac{n_j(\boldsymbol{\ell}^*)}{n}$ for all $j \in 1..g$, cf. Theorems 2.1 and 2.2. Hence, the third and the last terms in the above chain almost meet so that we have, for this g and with $\gamma_j = (m_j, V_j)$,

$$-n\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell}^*)}{n},\dots,\frac{n_g(\boldsymbol{\ell}^*)}{n}\Big)+\sum_{j=1}^{g}\sum_{\ell_i^*=j}\ln f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i)\approx \ln\prod_i\sum_j p_j^* f_{\gamma_j^*}(x_i). \tag{16}$$

The combination of Kéribin's result with the estimate (15) and the aproximation (16) supports BIC as a penalty term also for MAP–partitioning in the case of large data sets and good separation.

Since the number r_g of observations retained in the favorite partition for g clusters depends on g, it needs to be normalized, e.g. to n. The optimal value of the MAP–criterion (6) increases approximately linearly with the number r, asymptotically, at least in the normal case if there is sufficient separation. Indeed, if $p_1, \dots, p_g$ are the mixing proportions then, P–a.s., by Theorem 2.1 or 2.2 and by the law of large numbers,

$$\begin{aligned}&-\mathrm{H}\big((n_j(\boldsymbol{\ell}^*)/r)_j\big)+\sum_{j=1}^{g}\frac{1}{r}\sum_{\ell_i^*=j}\ln f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i)\underset{r\to\infty}{\longrightarrow}-\mathrm{H}\big((p_j)_j\big)+\sum_{j=1}^{g}\lim_{r\to\infty}\frac{1}{r}\sum_{\ell_i^*=j}\ln f_{\gamma_j}(x_i)\\&=-\mathrm{H}\big((p_j)_j\big)+\sum_{j=1}^{g}p_j\int f_{\gamma_j}(x)\ln f_{\gamma_j}(x)\,\mathrm{d}x=-\mathrm{H}\big((p_j)_j\big)-\frac{d}{2}-\sum_{j=1}^{g}\frac{p_j}{2}\ln\det 2\pi\, V_j.\end{aligned}$$

Therefore, we propose as model selection criterion with trimming the *corrected BIC*

$$\underset{g}{\operatorname{argmax}}\Big\{-n\,\mathrm{H}\Big(\frac{n_1(\boldsymbol{\ell}^*)}{n},\dots,\frac{n_g(\boldsymbol{\ell}^*)}{n}\Big)+\frac{n}{r_g}\sum_{j=1}^{g}\sum_{\ell_i^*=j}\ln f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i)-\frac{q(g)}{2}\ln n\Big\}. \tag{17}$$

We finally note that Neykov et al. [43] recently proposed a simple method that estimates both parameters at a time, the *trimmed BIC*. They establish a table of BIC values indexed by g and r proposing to use the parameter values where the minima w.r.t. g stabilize. Although their criterion is proposed for the mixture model it can be applied to the MAP–criterion as well.

4.2 Experimental studies

We finally illustrate the methods presented in Sections 3.5 and 4.1 with a synthetic, contaminated data set, MLNG, and Anderson's [3] famous four–dimensional Iris Data Set.

Means	$(0,0,0)$	$(-6,3,6)$	$(6,6,4)$
Covariance matrices	diagonal $(9.0, 4.0, 1.0)$	$\begin{pmatrix}4.0 & & \\ -3.2 & 4.0 & \\ -0.2 & 0.0 & 1.0\end{pmatrix}$	$\begin{pmatrix}4.0 & & \\ 3.2 & 4.0 & \\ 2.8 & 2.4 & 2.0\end{pmatrix}$
Eigenvalues	9.0 4.0 1.0	7.20 1.07 0.73	9.11 0.88 0.016
Cardinalities	200	50	50

Table 1: Structures of the three 3D normal populations from which Data Set MLNG is sampled. The eigenvalues of the covariance matrices are also shown. The data set contains 30 additional "outliers" uniformly distributed in the cube $[-15, 15]^3$.

The three-dimensional *Data Set MLNG* is sampled from the three normal populations with parameters specified in Table 1. The corresponding mixture appears in McLachlan and Peel [40], p. 218. To this basic data set we add 30 outliers uniformly distributed in the cube $[-15, 15]^3$. We set the minimal size constraints $b_j = d + 1 = 4$. The clustering according to the trimmed MAP determinant criterion (7) with *known* numbers of classes and outliers is essentially the original one. The estimated parameters for these input values are shown in Table 2.

Means	$(0.09, 0.13, -0.02)$	$(-5.94, 2.94, 5.95)$	$(6.38, 6.40, 4.28)$
Covariance matrices	$\begin{pmatrix} 9.1 & & \\ 0.4 & 3.7 & \\ 0.2 & 0.0 & 1.0 \end{pmatrix}$	$\begin{pmatrix} 4.0 & & \\ -2.3 & 2.7 & \\ -0.5 & -0.0 & 1.2 \end{pmatrix}$	$\begin{pmatrix} 3.4 & & \\ 2.6 & 3.0 & \\ 2.5 & 2.0 & 1.8 \end{pmatrix}$
Eigenvalues	9.1 3.7 1.0	4.0 1.4 1.0	3.4 1.1 0.02
Cardinalities	202	51	47

Table 2: Data Set MLNG: parameters estimated with known numbers of clusters and outliers (30). This is also the partition obtained with the χ^2–goodness-of-fit test and model selection criterion BIC, Eqn. (17). See also the text in Section 4.2.

We next estimate the numbers of outliers and clusters using the optimal solutions $\boldsymbol{\ell}^*$ of the trimmed MAP determinant criterion (7) for various parameters g and r with the multiple-testing procedure explained in in Section 4.1. We have to specify a simple test for the hypotheses $H_0^j\colon X^{(j)} \sim \mathrm{N}_{m_j^*,V_j^*}$ with the estimated mean vectors $m_j^* = \overline{x}_{C_j(\boldsymbol{\ell}^*)}$ and covariance matrices $V_j^* = S_{C_j(\boldsymbol{\ell}^*)}$. We use a χ^2–goodness–of–fit test with Yates' correction. Its ten cells are the elliptical layers defined by the Mahalanobis distance square $\varphi_j(x) = (x - m_j^*)^{\mathrm{T}} \left(V_j^*\right)^{-1}(x - m_j^*)$, $x \in \mathbb{R}^d$, and the $k/10$-quantiles q_k, $1 \le k < 10$, of the χ^2_d–distribution as division points. Under the hypothesis H_0^j, the random variable $X^{(j)}$ is approximately χ^2_d–distributed so that the cells $\{x \in \mathbb{R}^d \mid q_{k-1} \le \varphi_j(x) < q_k\}$, $k \in 1..10$, $q_0 = 0$, $q_{10} = \infty$, are about equiprobable. Since there are nine degrees of freedom the level of significance $\alpha = 0.1$ means a value of the χ^2–test statistic of ≥ 14.7. The test does not reject the solutions with 3 clusters and between 30 and 40 discarded elements and so we accept the one with 30 as the solution for $g = 3$, cf. Table 3. Its sizes are 202, 51, and 47. Two solutions for $g = 4$ are also not rejected, the one which retains the whole data set and the one which discards 15 elements. The first solution accepts the outliers as a cluster and is the proposed partition for $g = 4$ since it retains more objects. The second solution removes extreme elements in the corners of the cube and accepts the rest as a cluster. The values of the corrected BIC (17) are -2150 ($g = 3$) and -2418 ($g = 4$). We, thus, accept the former solution. It is close to the original partition.

Between five and fifty replications of (exact) multipoint iterations 3.4 with randomly selected initial partitions were sufficient to reach the final result for one pair $(g, n - r)$. Each iteration consisted of about ten reduction steps. Our C++ implementation takes about $3 \cdot 10^{-4}$ sec on a 2 GHz processor for one reduction step.

The *Iris Data Set* has served for demonstrating the performance of many clustering algorithms, beginning with Fisher [17]; for a recent study see Li [37]. It consists of 50 observations of each

g	$n-r$	MLNG		
3	25	20.1	14.3	10.9
3	**30**	12.5	11.6	11.0
3	**35**	11.1	7.5	3.5
3	**40**	11.3	7.5	3.1
3	45	15.9	15.0	2.5
3	50	19.1	15.1	2.5
3	55	22.8	19.1	2.5
3	60	37.7	20.8	3.7

g	$n-r$	MLNG			
4	**0**	12.5	11.0	10.7	9.3
4	5	23.2	12.5	11.0	9.3
4	10	21.9	12.5	11.0	9.3
4	**15**	14.2	13.2	12.1	3.5
4	20	15.8	12.5	11.0	8.5

g	$n-r$	Iris		
3	**0**	13.3	3.7	5.0
3	**5**	11.9	7.8	5.0
3	**10**	6.2	8.5	1.8
3	**15**	9.1	8.5	4.2
3	**20**	11.4	7.6	4.2
3	25	16.5	9.1	3.8
3	30	20.7	11.6	7.1

Table 3: Yates corrected χ^2–goodness–of–fit test with nine degrees of freedom for estimating the number of outliers. The rows show the values of the test statistic for the Data Sets MLNG and Iris and clustering results with g clusters and $n - r$ discarded elements. The minimum constraints $b_j = 4$ were applied to MLNG and the constraints $b_j = 20$ to Iris. See also the text.

of the three subspecies setosa, versicolour, and virginica with at most few outliers. Its entries are of the form *.* with only two digits of precision preventing general position of the data. The observations 102 and 143 are even equal. Since this is a deficiency of the data we add a number uniformly distributed in the interval $[-0.05, 0.05]$ to each entry, thus trying to "restore" their numerical character, although not faithfully. But the noise added is negligable compared with the natural variation of the data. Since the entries are positive, we take their logarithms.

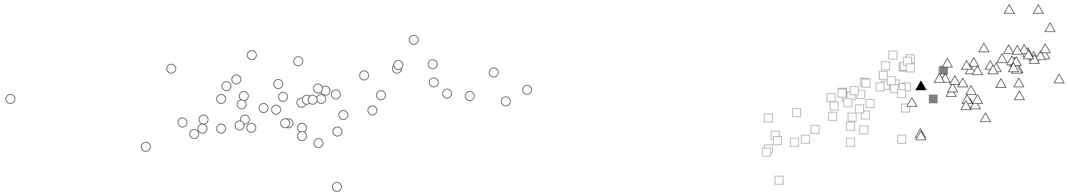


Figure 1: The Iris data projected to the plane spanned by the three cluster centers estimated with ML. Left I. setosa, center I. versicolour. The errors are shown solid.

The known, correct partition of Iris is essentially reproduced by the ML determinant criterion (4) with three clusters, no discarded elements, and minimal cardinalities $b_j = b = d + 1 = 5$, just two versicolour plants are placed in the virginica cluster and one virginica plant is falsely assigned to the versicolour cluster. The same result was obtained by Scott and Symons [54] with their homoscedastic model and by Li [37]. Fig. 1 presents a scatter plot of the projection of the data onto the plane spanned by the three cluster centers. The optimal clustering according to the MAP determinant criterion (7) with three clusters, no discarded elements, and $b = 5$ unites I. versicolour and I. virginica identifying only two major clusters and a small spurious cluster of minimal size five close to some hyperplane. Up to $b = 19$, the optimal solution still contains a cluster of or close to the minimum size b. But from $b = 20$ on, up to 46, it returns a solution with cardinalities 50, 46, 54 close to the correct one. The χ^2–goodness–of–fit test presented on the right of Table 3 indicates that Iris does not contain many outliers, if any.

Here, it takes on the average 16 000 replications of (exact) reduction step iterations 3.4 (of length about ten, each) with random initial labellings in order to reach the (conjectured) optimum. Standard deviation of the number of replications is 12 000. Our C++ implementation needs 1.5 seconds for 1000 replications on a 2 GHz processor.

Appendices

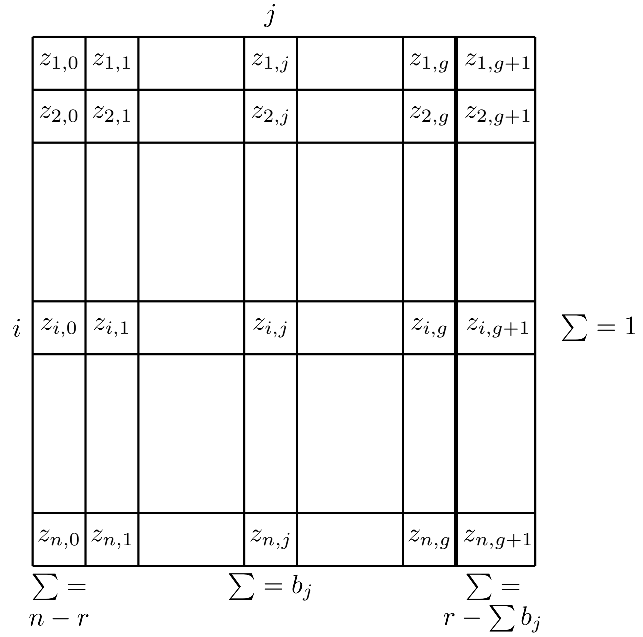
A Combinatorial optimization and λ–assignment problem

In this appendix we transform the multipoint optimization problem 3.3 to the λ–assignment problem (λA). A linear optimization problem of the form

$$\text{(TP)}\quad \sum_{i,j} u_{ij} z_{ij} \text{ maximal over all matrices } \mathbf{z} \in \mathbb{R}^{n\times m} \text{ subject to the constraints}$$

$$\begin{cases} \sum_j z_{ij} = a_i, & i \in 1..n, \\ \sum_i z_{ij} = b_j, & j \in 1..m, \\ z_{i,j} \geq 0, \end{cases}$$

is called a *transportation* or *Hitchcock problem*, a problem surprisingly equivalent to the *circulation* and to the *min–cost flow* problem (for minimization instead of maximization), see [45]. Here, $(u_{i,j})$ is a real n by m matrix of weights and the "supplies" a_i and "demands" b_j are real numbers ≥ 0 such that $\sum a_i = \sum b_j$. Plainly this condition is necessary and sufficient for a solution to exist.



				j				
	$z_{1,0}$	$z_{1,1}$		$z_{1,j}$		$z_{1,g}$	$z_{1,g+1}$	
	$z_{2,0}$	$z_{2,1}$		$z_{2,j}$		$z_{2,g}$	$z_{2,g+1}$	
i	$z_{i,0}$	$z_{i,1}$		$z_{i,j}$		$z_{i,g}$	$z_{i,g+1}$	$\sum = 1$
	$z_{n,0}$	$z_{n,1}$		$z_{n,j}$		$z_{n,g}$	$z_{n,g+1}$	
	$\sum =$ $n - r$			$\sum = b_j$			$\sum =$ $r - \sum b_j$	

Figure 2: Table of assignments of objects to classes in the assignment problem associated with the binary linear optimization problem (BLO).

If the supplies are unitary, (TP) is called a λ*–assignment problem*. Note that Problem (λA) in Section 3.3 is of this kind. Its demands are $n - r$, $b_1, \cdots, b_g$, and $r - \sum b_j$ as illustrated in Fig. 2. Note also that the hypothesis on the supplies and demands for a transportation problem to be solvable is satisfied so that (λA) possesses an optimal solution.

In order to explain the transformation, it is suitable to first cast Problem 3.3 in a form common in combinatorial optimization. A labelling $\boldsymbol{\ell}$ may be represented by a zero–one matrix $\mathbf{y}$ of size $n \times (g+1)$ by putting $y_{i,j} = 1$ if and only if $\ell_i = j$, i.e., if $\boldsymbol{\ell}$ assigns object i to cluster j. A zero–one matrix $\mathbf{y}$ is *admissible*, i.e., corresponds to an admissible labelling, if it satisfies the constraints $\sum_j y_{i,j} = 1$ for each i (each object has exactly one label), $\sum_i y_{i,0} = n - r$ (there are $n - r$ discarded elements) and $\sum_i y_{i,j} \geq b_j$ for all $j \geq 1$ (each cluster j contains at least b_j elements). Using this matrix and the weights u_{ij} defined by (13), we may reformulate the multipoint optimization problem 3.3 as a *binary linear optimization problem*, cf. Papadimitriou and Steiglitz [45], in the following way.

$$\text{(BLO)} \quad \sum_{i=1}^{n} \sum_{j=1}^{g} u_{i,j} y_{i,j} \text{ maximal over all matrices } \mathbf{y} \in \mathbb{R}^{n \times (g+1)} \text{ subject to the constraints}$$

$$\begin{cases} \sum_j y_{i,j} = 1, & i \in 1..n, \\ \sum_i y_{i,0} = n - r, & \\ \sum_i y_{i,j} \geq b_j, & j \in 1..g, \\ y_{i,j} \in \{0, 1\}. & \end{cases}$$

Our (BLO) problem is not yet a λ–assignment problem for two reasons: first, the constraints contain also an inequality and, second, λ–assignment is not restricted to binary solutions. The introduction of the dummy class $g+1$ in (λA), Section 3.3, is a trick to overcome the first problem. Fortunately, the second turns out to be a free lunch: the constraints of (λA) are integral. By the Integral Circulation Theorem, Lawler [36], Theorem 12.1, or Cook et al. [13], Theorem 4.5, there is an optimal solution $\mathbf{z}^*$ to (λA) with integral entries. The first constraint then implies that it is even binary thus representing an assignment.

With the optimal solution $\mathbf{z}^*$ we associate a solution $\mathbf{y}^* \in \mathbb{R}^{n\times(g+1)}$ to (BLO) as in (14): in each line i of $\mathbf{z}^*$, move the excess members collected in cluster $g+1$ to their natural class. We claim that $\mathbf{y}^*$ optimizes (BLO). Indeed, let $\mathbf{y}$ be admissible for (BLO). Define a feasible matrix $\mathbf{z}$ by moving excess members i from the classes to the artificial class $g+1$. By definition of $u_{i,g+1}$, the value of $\mathbf{y}$ w.r.t. (BLO) is smaller than that of $\mathbf{z}$ w.r.t. (λA) and, by optimality, the latter value is smaller than that of $\mathbf{z}^*$ w.r.t. (λA) which, again by definition of $u_{i,g+1}$, equals that of $\mathbf{y}^*$ w.r.t. (BLO). Hence, $\mathbf{y}$ is inferior to $\mathbf{y}^*$ and $\mathbf{y}^*$ is an optimal solution to the original problem (BLO).

Optimal solutions to (λA) and (BLO) are actually equivalent. Given an optimal solution to (BLO), move any excess elements in classes $1,\dots,g$ to class $g+1$. Note that any class that contains excess elements contains no forced elements since they could choose a better class. Therefore, the new assignment creates the same total weight.

It is easy to construct a solution to (BLO) that satisfies all but the third constraint. Just assign object i to its natural class, $\mathrm{argmax}_{j\in 1..g}\, u_{i,j}$. If this solution happens to satisfy also the third constraint then it is plainly optimal. The opposite case needs special attention. The deficient clusters j in such a solution contain exactly b_j objects in an optimal solution. Indeed, if the size of an originally deficient cluster j were $> b_j$ in an optimal solution then at least one of the forced elements would be free to go to its natural cluster, thus reducing the target function. Although binary linear optimization is NP–hard, in general, the present problem (BLO) has an efficient solution.

B Algorithmic considerations

B.1 Algorithms for the λ–assignment problem

The multipoint reduction step requires the solution to a λ–assignment problem. Since reduction steps are executed many times during a program run, the reader who wishes to implement the algorithm may be interested in known algorithms for its solution and in their complexities. Both have been extensively studied in operations research and combinatorial optimization. As a *linear optimization* problem and a special case of *minimum–cost flow*, λ–assignment may be solved by the simplex method. An instance of the min–cost flow problem is specified by a directed graph, net flows in its nodes, arc capacities, and cost coefficients. In our application the cost coefficients are the negative weights. The aim is to determine a flow with minimum overall cost satisfying the required net flows in all nodes without violating given capacities. Classical adaptations of the simplex method tailored to the particularities of min–cost flow are the (primal) Augmenting Circuit Algorithm, cf. Cook et al. [13], 4.2, the *Network Simplex* Method, Sedgewick [55], and *primal–dual* algorithms, cf. Cook et al. [13], 4.2 and 4.3, such as the *Out–of–Kilter* Method [20, 36]. The Network Simplex Method seems to be most popular in applications although these algorithms are exponential in the worst case.

Polynomial algorithms for min–cost flow have also existed for some time. It is common to represent the weight matrix as the adjacency matrix of a weighted graph (V,E) with node set V of size n and edge set E of size m. The first (weakly) polynomial algorithm is due to Edmonds and Karp [16] who introduced the concept of weight–*scaling* thus showing that min–cost flow was a low–complexity problem. Scaling solves a series of subproblems with approximated instance parameters, capacities or weights or both. Orlin [44] refined their method to obtain a strongly polynomial algorithm of complexity $\mathcal{O}\big(n(m+n\log n)\log n\big)$ for the uncapacitated min–cost flow problem. In the context of λ–assignment all row sums in the assignment matrix are *equal*, and equal to 1. This implies that the capacities may be chosen unitary so that capacity scaling becomes trivial. An example is Gabow and Tarjan's [21] $\mathcal{O}\big(n(m+n\log n)\log U\big)$ algorithm, $\log U$ being the bit length used to represent network capacities. In the case of λ–assignment, put $\log U = 1$.

Now, the λ–assignment problem has some special features that reduce its complexity compared with the general min–cost flow problem. First, its graph is *bipartite*, its node set being partitioned in two subsets, the objects i and the clusters j, such that each edge connects an object with a cluster. For bipartite network flow algorithms it is often possible to sharpen the complexity bounds by using the sizes of the smaller (k) and the larger (n) node subsets as parameters, Gusfield et al. [28]. Second, the bipartite network is *unbalanced* in the sense that the number of clusters is (at least in our case) much smaller than the number of objects. Bounds based on the two subsets are particularly effective for unbalanced networks. Third, the capacities may be defined as 1. As a consequence the algorithms for the lambda assignment problem mentioned at the beginning become polynomial. E.g., the time complexity of the out–of–kilter method becomes $\mathcal{O}(nm) = \mathcal{O}(kn^2)$ since $m = kn$ here, see Lawler [36], p. 157.

Algorithms dedicated to the λ–assignment problem are due to Kleinschmidt et al. [34] and Achatz et al. [1]. Both algorithms have an asymptotic run time of $\mathcal{O}\big(kn^2\big)$. The former algorithm uses Hirsch paths for the dual assignment problem and is related to an algorithm of Hung and Rom's [31]. The latter is an interior point method.

The algorithms mentioned so far are at least quadratic in the size of the larger node set, n. By contrast, two *weight*–scaling min–cost flow algorithms are linear in n: Goldberg and Tarjan's [26], Theorem 6.5, algorithm solves the min–cost flow problem on a bipartite network asymptotically in $\mathcal{O}\big(k^2 n \log(kC)\big)\big)$ time, see Ahuja et al. [2], and the "two–edge push" algorithm of the latter authors needs $\mathcal{O}\big((km + k^3)\log(kC)\big)$ time. In our application, $m = kn$. Both estimates contain the bit length, $\log C$, for representing weights.

A different low–complexity approach is due to Tokuyama and Nakano [59, 61, 60]. These authors state and prove a *geometric* characterization of optimal solutions to the λ–assignment and transportation problems by a so–called *splitter*, a k–vector that partitions Euclidean k–space into k closed cones. The corresponding subdivision of the lines of the weight matrix describes an optimal assignment. Tokuyama and Nakano design a deterministic and a randomized strategy for splitter finding that run in $\mathcal{O}(k^2 n \ln n)$ time and $\mathcal{O}\big(kn + k^{5/2}\sqrt{n}\ln^{3/2} n\big)$ expected time, respectively. Their algorithms are almost linear in n and close to the absolute lower bound $\mathcal{O}\big(kn\big)$ if k is small, the case of interest for (λA).

B.2 Heuristic methods for feasible solutions

Besides exact solutions, there are reasons to say a word about heuristic feasible solutions to the λ–assignment problem. First, they may be used for multipoint reduction steps on their own. Moreover, some of the graphical methods presented in Section B.1 need initial feasible solutions for starting or at least profit from good ones. The network simplex method, e.g., needs a primal feasible solution and the method presented in Achatz et al. [1] needs a dual feasible solution. While arbitrary feasible solutions are easily produced, *good* initial feasible solutions can be constructed by means of greedy heuristics. We propose here two. If the bounds b_j are small enough, the heuristics often produce even optimal solutions.

Each reduction step receives a set of parameters γ_j from which all weights $u_{i,j}$, $i \in 1..n$, $j \in 1..g$, are computed. The first two heuristics construct primal feasible solutions. Both start from the best *unconstrained* assignment of the clustering problem which can be easily attained by sorting the numbers $u_i = \max_{1 \le j \le g} u_{i,j}$. More precisely:

Basic primal heuristic

1. sort the numbers $u_i = \max_{1 \le j \le g} u_{i,j}$ in decreasing order, $i \in 1..n$;

2. assign the first r objects in the ordered list to the class $1, \dots, g$ where the maximum is attained;

3. attach label 0 to the last $n - r$ objects;

4. *if* this assignment is admissible then *stop* (it is optimal);
   else
   - (α) starting from element r in the ordered list and going downwards, reassign surplus elements to arbitrary deficient classes until they contain exactly b_j elements;
   - (β) assign any remaining surplus elements in the classes to class $g+1$.

If an admissible instead of a feasible solution is required, only, then we drop step 4(β). We next refine the "else" part of step 4. Steps 1 and 2 are as before.

Refined primal heuristic

1. sort the numbers $u_i = \max_{1\le j\le g} u_{i,j}$ in decreasing order, $i \in 1..n$;
2. assign each of the first r objects to the class $1..g$ where the maximum is attained;
3. denote the set of the last $n-r$ objects in the ordered list by L;
4. *if* this assignment is admissible then *stop* (it is optimal);
   else
   - (α) let $\mathcal{D}$ be the set of deficient classes in $1..g$ and let δ be their total deficiency;
   - (β) starting from element r in the ordered list and going downwards, move the first δ elements in surplus classes to L;
   - (γ) sort the object–class pairs $(i,j) \in L \times \mathcal{D}$ in decreasing order according to their weights $u_{i,j}$ to obtain an array $(i_1, j_1),\ (i_2, j_2), \cdots, (i_{\#L\cdot\#\mathcal{D}}, j_{\#L\cdot\#\mathcal{D}})$;
   - (δ) scan all pairs (i_k, j_k) in this list starting from $k=1$ assigning object i_k to class j_k unless i_k has already been assigned or j_k is saturated, until all classes are saturated;
   - (ϵ) discard the yet unassigned elements of L;
   - (ζ) assign the smallest remaining surplus elements in classes $(1..g) \setminus \mathcal{D}$ to class $g+1$.

In Section A, we exploited the fact that any cluster j with more than b_j members in an optimal solution contained no forced elements. Plainly, both heuristics share this property since such members could be freely relabelled.

Both heuristics are much faster than any of the solution algorithms while yielding often a large value of the criterion, the refined heuristic larger than the basic. However, contrary to the exact solution, the improvements in the sense of Proposition 3.1 they provide are not optimal, in general. In most cases, the criterion increases although one may construct examples where this fails: Consider the data set $\{-40, -8, -6, 0, 1, 2, 3, 40\}$, let $g=4$ and $b_1 = b_2 = b_3 = b_4 = 2$, and assume that there are no outliers. Suppose that the parameters are generated from the initial partition $C_1 = \{-40, 3\}$, $C_2 = \{-8, 1\}$, $C_3 = \{-6, 0\}$, $C_4 = \{2, 40\}$. They are $m_1 = -18.5$, $m_2 = -3.5$, $m_3 = -3.0$, $m_4 = 21.0$, and $v_1 = 462.25$, $v_2 = 20.25$, $v_3 = 9.0$, $v_4 = 361.0$. The matrix of negative weights,

$$(-u_{i,j}) = \begin{pmatrix} 9.91 & 9.15 & 9.25 & 9.65 & 9.73 & 9.82 & 9.91 & 16.31 \\ 71.57 & 6.78 & 6.09 & 6.39 & 6.78 & 7.27 & 7.87 & 99.23 \\ 157.08 & 7.75 & 5.97 & 5.97 & 6.75 & 7.75 & 8.97 & 210.41 \\ 18.97 & 10.99 & 10.68 & 9.88 & 9.77 & 9.66 & 9.56 & 9.66 \end{pmatrix}^T,$$

generates the free partition $\{-40\}$, $\{-8, 2, 3\}$, $\{-6, 0, 1\}$, $\{40\}$ and the refined heuristic modifies it to $\{-40, 1\}$, $\{-8, 2\}$, $\{-6, 0\}$, $\{3, 40\}$. The score of the latter is -39.73 whereas the initial partition has the larger score -39.67.

The problem dual to (λA) reads

$$(\lambda\text{A}^*)\ \sum_{i=1}^{n} p_i + \sum_{j=0}^{g+1} b_j q_j \text{ minimal over } (\mathbf{p},\mathbf{q}) \in \mathbb{R}^{n+g+2} \text{ subject to the constraints}$$

$$p_i + q_j \geq u_{i,j},\ i \in 1..n,\ j \in 0..(g+1).$$

A simple initial heuristic for this problem is found in Carpaneto et al. [10]:

Dual heuristic $q_j = \max_i u_{i,j}$ and $p_i = \max_j (u_{i,j} - q_j)$.

Acknowledgments. We thank Prof. Peter Kleinschmidt and Dr. Hans Achatz for providing us an implementation of their Dual Forest Algorithm and for discussing the algorithm with us. We thank Frau Katharina Soller for her implementation of Goldberg and Tarjan's scaling algorithm [26]. We also thank an anonymous referee for pointing out Reference [15] to us.

References

[1] Achatz, H., Kleinschmidt, P., and Paparrizos, K. A dual forest algorithm for the assignment problem. In *Applied Geometry and Discrete Mathematics*, P. Gritzmann and B. Sturmfels, Eds., vol. 4 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, RI, 1991, pp. 1–10.

[2] Ahuja, R. K., Orlin, J. B., Stein, C., and Tarjan, R. E. Improved algorithms for bipartite network flows. *SIAM J. Computing 23* (1994), 906–933.

[3] Anderson, E. The irises of the Gaspé Peninsula. *Bull. Amer. Iris Soc. 59* (1935), 2–5.

[4] Aurenhammer, F., Hoffmann, F., and Aronov, B. Minkowski–type theorems and least–squares clustering. *Algorithmica 20* (1998), 61–76.

[5] Barnett, V., and Lewis, T. *Outliers in Statistical Data.* Wiley, Chichester, UK, 1994.

[6] Becker, C., and Gather, U. The masking breakdown point of multivariate outlier identification rules. *JASA 94* (1999), 947–955.

[7] Bernholt, T., and Fischer, P. The complexity of computing the MCD–estimator. *Theor. Comp. Science 326* (2004), 383–398.

[8] Bezdek, J. C., Keller, J., Krisnapuram, R., and Pal, N. R. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* The Handbooks of Fuzzy Sets Series. Kluwer, Boston, London, Dordrecht, 1999.

[9] Bock, H.-H. On some significance tests in cluster analysis. *J. Classification 2* (1985), 77–108.

[10] Carpaneto, G., Martello, S., and Toth, P. Algorithms and codes for the assignment problem. *Ann. OR 13* (1988), 193–223.

[11] Chen, H., Chen, J., and Kalbfleisch, J. D. Testing for a finite mixture model with two components. *J. Royal Stat. Soc, Series B 66* (2004), 95–115.

[12] Coleman, D. A., Dong, X., Hardin, J., Rocke, D. M., and Woodruff, D. L. Some computational issues in cluster analysis with no a priori metric. *CSDA 31* (1999), 1–11.

[13] Cook, W. J., Cunningham, W. H., Pulleyblank, W. R., and Schrijver, A. *Combinatorial Optimization.* Wiley, New York etc., 1998.

[14] Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. Trimmed k–means: An attempt to robustify quantizers. *The Annals of Statistics 25* (1997), 553–576.

[15] Davies, L., and Gather, U. The identification of multiple outliers. *JASA 88* (1993), 782–792.

[16] Edmonds, J., and Karp, R. Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM 19* (1972), 248–264.

[17] Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugenics 7* (1936), 179–188.

[18] Fraley, C., and Raftery, A. E. Model–based clustering, discriminant analysis, and density estimation. *JASA 97* (2002), 611–631.

[19] Friedman, H., and Rubin, J. On some invariant criteria for grouping data. *JASA 62* (1967), 1159–1178.

[20] Fulkerson, D. An out-of-kilter method for minimal cost flow problems. *J. SIAM 9* (1961), 18–27.

[21] Gabow, H. N., and Tarjan, R. E. Faster scaling algorithms for network problems. *SIAM J. Computing 18* (1989), 1013–1036.

[22] Gallegos, M. T., and Ritter, G. A robust method for cluster analysis. *Annals of Statistics 33* (2005), 347–380.

[23] Gallegos, M. T., and Ritter, G. Trimming algorithms for clustering contaminated grouped data and their robustness. *Adv. Data Anal. Classif. 3* (2009), 135–167.

[24] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. A general trimming approach to robust cluster analysis. *Ann. Stat. 36* (2008), 1324–1345.

[25] Garey, M. R., and Johnson, D. S. *Computers and Intractibility.* Freeman, San Francisco, 1979.

[26] Goldberg, A. V., and Tarjan, R. E. Finding minimum–cost circulations by successive approximation. *Math. of OR 15* (1990), 430–466.

[27] Gordon, A. D. *Classification*, second ed., vol. 82 of *Monographs on Statistics and Applied Probability.* CRC Press, 1999.

[28] Gusfield, D., Martel, C., and Fernandez-Baca, D. Fast algorithms for bipartite network flow. *SIAM J. Comput. 16* (1987), 237–251.

[29] Hartigan, J. A. Asymptotic distributions for clustering criteria. *Ann. Stat. 6* (1978), 117–131.

[30] Hathaway, R. J. A constrained formulation of maximum–likelihood estimation for normal mixture distributions. *Ann. Stat. 13* (1985), 795–800.

[31] Hung, M. S., and Rom, W. O. Solving the assignment problem by relaxation. *Operations Research 28* (1980), 969–982.

[32] Jancey, R. Multidimensional group analysis. *Australian J. Botany 14* (1966), 127–130.

[33] Kéribin, C. Consistent estimation of the order of mixture models. *Sankhyā, Series A 62* (2000), 49–66.

[34] Kleinschmidt, P., Lee, C. W., and Schannath, H. Transportation problems which can be solved by use of Hirsch paths for the dual problem. *Math. Prog. 37* (1987), 153–168.

[35] Knuth, D. E. *The Art of Computer Programming*, 2nd ed., vol. 2. Addison–Wesley, Reading, Menlo Park, London, Amsterdam, Don Mills, Sydney, 1981.

[36] Lawler, E. *Combinatorial Optimization: Networks and Matroids.* Holt, Rinehart and Winston, New York, 1976.

[37] Li, B. A new approach to cluster analysis: the clustering–function–based method. *J. Royal Stat. Soc., Series B 68* (2006), 457–476.

[38] MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Statist. Probab.* (1967), pp. 281–297.

[39] Mardia, K., Kent, T., and Bibby, J. *Multivariate Analysis.* Academic Press, London, New York, Toronto, Sydney, San Francisco, 1979.

[40] McLachlan, G. J., and Peel, D. *Finite Mixture Models.* Wiley, New York etc., 2000.

[41] Mecklin, C. J., and Mundfrom, D. J. An appraisal and bibliography of tests for multivariate normality. *International Statistical Review 72*, 1 (2004), 123–138.

[42] Milligan, G., and Cooper, M. An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50* (1985), 159–179.

[43] Neykov, N., Filzmoser, P., Dimova, R., and Neytchev, P. Robust fitting of mixtures using the trimmed likelihood estimator. *CSDA 52* (2007), 299–308.

[44] Orlin, J. B. A faster strongly polynomial minimum cost flow algorithm. In *Proc. 20th ACM Symp. Theory of Computing* (1988), pp. 377–387.

[45] Papadimitriou, C. H., and Steiglitz, K. *Combinatorial Optimization.* Prentice–Hall, Englewood Cliffs, New Jersey, 1982.

[46] Pollard, D. Strong consistency of k–means clustering. *Ann. Stat. 9* (1981), 135–140.

[47] Pollard, D. A central limit theorem for k–means clustering. *Ann. Proba. 10* (1982), 919–926.

[48] Rocke, D. M., and Woodruff, D. L. Identification of outliers in multivariate data. *JASA 91* (1996), 1047–1061.

[49] Rocke, D. M., and Woodruff, D. L. A synthesis of outlier detection and cluster identification. Tech. rep., University of California, Davis, 1999. http://handel.cipic.ucdavis.edu/∼dmrocke/Synth5.pdf.

[50] Rousseeuw, P. J. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, W. Grossmann, G. C. Pflug, I. Vincze, and W. Wertz, Eds., vol. 8B. Reidel, Dordrecht–Boston–Lancaster–Tokyo, 1985, pp. 283–297.

[51] Rousseeuw, P. J., and Van Driessen, K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41* (1999), 212–223.

[52] Schroeder, A. Analyse d'un mélange de distributions de probabilités de même type. *Revue de Statistique Appliquée 24* (1976), 39–62.

[53] Schwarz, G. Estimating the dimension of a model. *Ann. Stat. 6* (1978), 461–464.

[54] Scott, A., and Symons, M. J. Clustering methods based on likelihood ratio criteria. *Biometrics 27* (1971), 387–397.

[55] Sedgewick, R. *Algorithms in C*, third ed., vol. 5 - Graph Algorithms. Addison-Wesley, Boston etc., 2002.

[56] Steinhaus, H. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci. 4* (1956), 801–804.

[57] Symons, M. J. Clustering criteria and multivariate normal mixtures. *Biometrics 37* (1981), 35–43.

[58] Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc., Series B 63* (2001), 411–423.

[59] Tokuyama, T., and Nakano, J. Geometric algorithms for a minimum cost assignment problem. In *Proc. 7th ACM Symp. on Computational Geometry* (1991), pp. 262–271.

[60] Tokuyama, T., and Nakano, J. Efficient algorithms for the Hitchcock transportation problem. *SIAM J. Comput. 24* (1995), 563–578.

[61] Tokuyama, T., and Nakano, J. Geometric algorithms for the minimum cost assignment problem. *Random Structures and Algorithms 6* (1995), 393–406.

[62] Tso, M., Kleinschmidt, P., Mitterreiter, I., and Graham, J. An efficient transportation algorithm for automatic chromosome karyotyping. *Patt. Rec. Lett. 12* (1991), 117–126.

[63] Ward, Jr., J. H. Hierarchical grouping to optimize an objective function. *JASA 58* (1963), 236–244.

[64] Wolfe, J. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Res. 5* (1970), 329–350.

[65] Woodruff, D. L., and Reiners, T. Experiments with, and on, algorithms for maximum likelihood clustering. *CSDA 47* (2004), 237–252.