Appeared in *ADAC* 12 (2018), 179-202. DOI 10.1007/s11634-016-0278-2.

Probabilistic clustering via Pareto solutions and significance tests

Abstract: The present paper proposes a new strategy for probabilistic (often called model-based) clustering. It is well known that local maxima of mixture likelihoods can be used to partition an underlying data set. However, local maxima are rarely unique. Therefore, it remains to select the reasonable solutions, and in particular the desired one. Credible partitions are usually recognized by separation (and cohesion) of their clusters. We use here the p-values provided by the classical tests of Wilks, Hotelling, and Behrens–Fisher to single out those solutions that are well *separated by location*.

It has been shown that reasonable solutions to a clustering problem are related to Pareto points in a plot of scale balance vs. model fit of all local maxima. We briefly review this theory and propose as solutions all well-fitting Pareto points in the set of local maxima separated by location in the above sense.

We also design a new iterative, parameter-free cutting plane algorithm for the multivariate Behrens–Fisher problem.

Keywords: Cluster analysis; Probabilistic models; Mixture model, Classification model, Pareto solutions, Behrens–Fisher problem, Hotelling's T^2 statistic, Wilks' lambda.

Mathematics Subject Classification: Primary 62H30, Secondary 62–07

1 Introduction: The classical setup.

Many data sets $\mathbf{x} = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^d$, occurring in practice originate from different causes or situations. Examples are provided by biology when data comes from organisms of different subspecies or under different conditions. When causes are known, discriminant analysis may be used for classifying new objects. If not, literature knows many methods based on the observed data to cluster the objects and to detect and describe the causes. This task is not easy, in particular when the data clusters that originate from the various causes substantially overlap. The present paper offers a strategy for resolving these cases.

Let $N_{m,V}$ stand for the d-variate normal distribution with mean vector m and covariance matrix V. A popular model for cluster analysis is the normal *mixture model* on the sample space $\mathbb{R}^d$,

$$\sum_{j=1}^{g} \pi_j N_{m_j,V_j}. \tag{1}$$

Here, $g \geq 2$ is its number of components, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, $\pi_j > 0$, $\sum \pi_j = 1$, stands for its mixing rates, $\mathbf{m} = (m_1, \ldots, m_g)$ is its g-tuple of mean vectors in $\mathbb{R}^d$, and $\mathbf{V} = (V_1, \ldots, V_g)$ that of its positive definite covariance matrices. We will assume throughout that the pairs (m_j, V_j) are pairwise distinct. There exists a number of monographs devoted to this subject. We mention Lindsay [26], McLachlan and Peel [29], Böhning [6], Frühwirth-Schnatter [16],

and Ritter [32]. The preferred tool for estimating the parameters of (1) is the joint likelihood function

$$f(\mathbf{x}; \boldsymbol{\pi}, \mathbf{m}, \mathbf{V}) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j N_{m_j, V_j}(x_i) \tag{2}$$

at the data $\mathbf{x}$. By a result of Yakowitz and Spragins [39], the normal mixture model (1) is identifiable up to label switching. This allows us in principle to estimate its parameters when n is large enough. The data set $\mathbf{x}$ can subsequently be partitioned in clusters by Bayesian discriminant analysis based on the estimated parameters $\pi_j,\ m_j,\ V_j$ of the mixture.

However, the standard maximum likelihood paradigm cannot be applied right away – the likelihood function (2) is unbounded as noted by Kiefer and Wolfowitz [22]. It is sufficient to set $m_1 = x_1$ and to let $V_1 = \varepsilon I_d$ with $\varepsilon \to 0$. Eq. (2) has, thus, no maximum. As a resort, *local* likelihood maxima are used. In fact, there is the following result originally due to Kiefer [23] and Peters and Walker [31]. Their theorems refer to the univariate case. A proof in the multivariate case is contained in Ritter [32]. To avoid degeneracies, we will assume throughout that the data set has at least $d+1$ points and is in general position, that is, the affine space spanned by any $d+1$ data points is the whole $\mathbb{R}^d$. This is almost surely the case when data is sampled independently from a distribution with a Lebesgue density such as a normal mixture.

1.1 Theorem *Let the sampling distribution be a normal mixture* (1) *with parameters* $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$.

(a) There exists an open neighborhood U of $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$ such that the n-fold likelihood function (2) *of the normal mixture model possesses a.s. exactly one local maximum T_n in U for eventually all n.*

(b) The sequence $(T_n)_n$ is strongly consistent at $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$.

Thus, if we are given a sequence $x_1, x_2, \ldots$ sampled from (1), then Part (b) says that the sequence of local maxima $T_n \in U$ will a.s. converge to the original parameters $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$ as $n \to \infty$. Therefore, *local* likelihood maxima are interesting candidates for a solution. However, Day [10] noted that local maxima are not unique. Unfortunately, Theorem 1.1 does not tell us which one to choose. Literature knows several remedies to this problem.

(i) Generating all, or at least many, local maxima with repeated runs of the EM algorithm (see McLachlan and Peel [28]) and using the largest one, excepting "spurious" solutions, that is, solutions with an unusually small eigenvalue of one of its covariance matrices. But there is no generally accepted definition of what "unusually small" should mean.

(ii) Starting the EM iteration from an initial solution obtained with another, usually simpler, algorithm. One then obtains a single solution. Popular initial algorithms are hierarchical methods or k-means. This is fast, and acceptable if the initial algorithm leads us close to the desired solution. The EM algorithm serves here as a refinement of the initial solution. This is the viewpoint of Fraley and Raftery's [14] MCLUST.

(iii) Introducing a bound on all eigenvalue ratios of the covariance matrices and maximizing the likelihood function under these constraints. This is the TCLUST approach taken by Fritz et al. [15], who also design an algorithm for constrained optimization. This has the disadvantage that the bound is a priori unknown and has to be guessed.

(iv) Gallegos and Ritter [17, 18] select as solutions extreme points near the left lower corner of the convex hull of the so-called SBF plot (scale balance vs. fit) of local likelihood maxima; see also Section 4. However, there are sometimes other interesting solutions.

We will improve in Section 4 strategy (iv). The new strategy has the advantage of needing no prior assumptions. Cluster analysts face, of course, more problems than unboundedness of the likelihood function and multiplicity of solutions. Two are robustness in the presence of contaminated data and the so-called $p \gg n$-problem. We do not deal with these problems here. There is, however, a large body of literature on both subjects and we refer the interested reader to Ritter [32].

A probabilistic model competing with mixtures is the so-called normal *MAP classification model*

$$\log f(\mathbf{x}; \mathcal{C}, \boldsymbol{\pi}, \mathbf{m}, \mathbf{V}) = \sum_j \sum_{i \in C_j} \log\big(\pi_j N_{m_j, V_j}(x_i)\big). \tag{3}$$

Here, $\mathcal{C} = \{C_1, \ldots, C_g\}$ is some partition of the index set $1\,..\,n$ in g subsets (called clusters) and the other parameters are as defined in Eqs. (1) and (2). If $n_j = |C_j| \geq d+1$ for all j, $1 \leq j \leq g$, then partial unconstrained maximization w.r.t. π_j, m_j, and V_j is possible, leading to the *MAP determinant criterion*,

$$-\tfrac{1}{2} \sum_{j=1}^{g} n_j \log\det S_j - n\mathrm{H}\big(\tfrac{n_1}{n}, \ldots, \tfrac{n_g}{n}\big). \tag{4}$$

It depends only on the partition, namely via the scatter matrices $S_1, \ldots, S_g$ of its g clusters $C_1, \ldots, C_g$. It contains the entropy $\mathrm{H}\big(\frac{n_1}{n}, \ldots, \frac{n_g}{n}\big) = -\sum_{j=1}^{g} \frac{n_j}{n} \log \frac{n_j}{n}$ of the cluster proportions n_j/n which makes up for unequal cluster sizes. The idea is to produce partitions with large values of the determinant criterion, (4). However, its maximization is often not useful and may miss the desired solution. A strategy similar to that announced for the mixture model works in this case, too. Like the mixture model, the MAP criterion (4) is sensitive, having the power to adapt to arbitrary locations, scales, and cluster sizes.

A partition is called *steady* (w.r.t. the classification model) if it results from the Bayesian (or MAP) discriminant rule w.r.t. the (unconstrained, normal) ML parameter estimates of its own clusters. Steady partitions are classification analogs of local likelihood maxima in the mixture model. They are usually computed with an algorithm that alternates between parameter estimation and cluster assignment and is an extension of the k-means algorithm to arbitrary locations, scales, and mixing rates. Therefore, Gallegos and Ritter [20] named it *k-parameters algorithm*. Like the EM algorithm, it converges. Since there are only finitely many partitions, this means that its output becomes even stationary (or constant) in the end. The limit of this iterative algorithm is a steady partition.

Although classification and mixture model are different, both bear some common traits. Just as the likelihood function of the mixture model is unbounded, the determinant criterion becomes ∞ if one cluster is deficient, that is, contains $\leq d$ elements. Moreover, there are multiple steady solutions. Our experience even teaches us that their number typically exceeds that of local likelihood maxima of the mixture model. Items (i) – (iv) above apply *mutatis mutandis*.

Components of normal mixtures are separated by location, scale, or both. In most real examples, they are separated by location, which means that their mean vectors do not lie

too close together. Partitions derived from such mixtures will consist of clusters separated in the same way. This does not necessarily mean non-overlapping of cluster pairs, that is, disjointness of their convex hulls. The latter represents an extreme case of separation by location. In such a case, there will be in general only one reasonable solution for a given number of clusters and many methods will return it. In this paper, we are rather interested in strategies that allow us to analyze clustered data sets with *overlapping* clusters. Such cases occur very often in practice and we call them complex. An example is the well-known CRABS data set which will, among others, be used in Section 5 to illustrate the present method.

2 Using the Wilks, Hotelling, and Behrens–Fisher test statistics

Properties that speak for the credibility of a partition are separation and cohesion of its clusters. Cohesion may be viewed as missing separability and so, we concentrate on separation, here by location. There exist several criteria for assessing cluster separation. One was proposed by Bailey and Dubes [2]. A popular criterion these days is Rousseeuw's [34] silhouette. Both produce graphics that are not easily interpreted in general. They depend on prior metrics and will work decently when component scales comply with them or when separation is good. In other cases they yield ambiguous results. The reason is that they utilize distances of points within and between clusters. These are not generally characteristic for membership of points in the same or in different clusters. It is sufficient to take a look at two close, parallel, elongated clusters, a standard counterexample in cluster analysis. We are not the first authors to criticize the practice of using prior metrics. For instance, Bock [5], p. 95, raises doubts about the power of a distance-based gap test.

An important principle of statistics is equivariance w.r.t. certain sample space transformations, see for instance Muirhead [30], Ch. 6. Clustering results should not depend on conventional quantities such as units of measurement. Sometimes it is a priori known that components are spherical. Then we wish that the estimator be rotationally equivariant. In other cases, it may be known that the variables are independent. Then the estimator should be chosen scale equivariant. In general, nothing is known a priori and the method of choice is the affine equivariant determinant criterion (4). Separation criteria should meet the same requirements. Affine equivariant criteria and tests for cluster separation are well known. Examples are the Kullback–Leibler divergence (or relative entropy) and the Hellinger, Bhattachariya, and Chernov distances of the estimated distributions; see Devroye et al. [11] and Ritter [32]. Moreover, there are several classical tests for the equality of location parameters based on samples. Let us consider the null hypothesis

$$\mathrm{H}_0 \colon m_1 = \cdots = m_g$$

of equality of all means of g normal populations against the alternative of disparity (or generality). The less likely H_0 is, that is, the smaller the p-value obtained from a statistical test is, the better the separation by location. We prefer p-values of statistical tests to other separation measures such as the ones above for ranking different solutions since they offer us an impartial criterion over different partitions, independent of cluster sizes or scales. It is often not necessary to compute the p-values themselves since test statistics and p-values

Table 1: Median p-values for two-dimensional data sets sampled from $0.5N_{m_1,I_2}+0.5N_{m_2,I_2}$ partitioned in two clusters with k-means; see also the text.

$\|m_2 - m_1\|$	8	6	4	2	1	0
Hotelling (Wilks)	$2.6{\cdot}10^{-134}$	$9.8{\cdot}10^{-105}$	$1.7{\cdot}10^{-69}$	$9.8{\cdot}10^{-39}$	$3.5{\cdot}10^{-28}$	$1.3{\cdot}10^{-26}$
Behrens–Fisher	$4.0{\cdot}10^{-105}$	$2.0{\cdot}10^{-87}$	$5.3{\cdot}10^{-64}$	$5.8{\cdot}10^{-39}$	$2.2{\cdot}10^{-28}$	$1.4{\cdot}10^{-26}$

often produce the same ranking. Special tests derive from the likelihood ratio (LR) and union intersection paradigms. We mention Wilks' [37] Λ test, Hotelling's T^2 test, and Behrens' [3] and Fisher's [12, 13] test. Their test statistics and p-values can actually be considered as measures of separation by location.

Before continuing, we have to interject an important remark. When a partition is obtained from a clustering algorithm then the null hypothesis H_0 of equality of means will in general be strongly rejected, even if the parent means are the same. Nevertheless, the strength of the rejection, that is, the size of the p-value, contains information on the separation of the mixture components. Table 1 shows the small p-values for the Hotelling and Behrens–Fisher tests for partitions of six two-dimensional data sets in two clusters, each, obtained with k-means. The data consists of 140 samples, each, from six mixtures $\frac{1}{2}N_{m_1,I_2} + \frac{1}{2}N_{m_2,I_2}$ with equal weights $1/2$ of two bivariate, standard normals with $\|m_2-m_1\| = 8$, 6, 4, 2, 1, and 0. The first mixture is very well separated in two components, the last one is actually a two-dimensional standard normal (no separation). The table shows median p-values among 21 random replications. If we had tested samples from the two components, we would have obtained much larger p-values, in particular for the small values of $\|m_2-m_1\|$. Nevertheless, the table shows that the strength of the rejection contains information on the degree of separation of the underlying mixture components. The p-value 10^{-28} is after all one hundred billion times larger than 10^{-39}. We may therefore conclude that a p-value of 10^{-39} indicates stronger separation than 10^{-28} does, although both mean strong rejection of the null hypothesis. This reasoning allows us to rank all solutions w.r.t. separation based on p-values obtained from the tests. We next elaborate on the tests.

(a) **Wilks' test**. We first consider the *homoscedastic* case. For all $1 \le j \le g$, let the cluster C_j consist of $n_j \ge d+1$ draws from some d-variate normal population $N_{m_j,V}$ with the same covariance matrix V. Put $n = \sum_j n_j$, $\mathcal{C} = \{C_1, \ldots, C_g\}$, and let $\overline{x}_j$ be the mean vector of C_j. The likelihood ratio test (LRT) for equality of means based on $C_1, \ldots, C_g$ leads to Wilks' [37] classical Λ test. Indeed, the estimate of the common mean vector under the hypothesis is $m^* = \frac{1}{n}\sum n_j\overline{x}_j$, the overall mean vector. The estimate of the common covariance matrix under hypothesis H_0 and alternative are the total scatter matrix $\overline{S}$ and the pooled scatter matrix S, respectively. Therefore, with $\mathbf{m}^* = (\overline{x}_1, \ldots, \overline{x}_g)$, the LRT statistic is

$$f_1(\mathbf{x}; \mathcal{C}, \mathbf{m}^*, S)/f_0(\mathbf{x}; \mathcal{C}, m^*, \overline{S}) = (\det \overline{S}/\det S)^{n/2}. \tag{5}$$

Its distribution under H_0 is a negative power of Wilks' Λ with parameters d, $n-g$, and $g-1$, $\Lambda(d, n-g, g-1)^{-n/2}$; see Mardia et al. [27], where all basic details on this and other tests are found. Since the distribution is independent of the specific partition, the order of the statistic is inverse to that of the p-values and so, the latter are not needed here.

Most test statistics used here depend on some Mahalanobis distance. It is, therefore, comfortable to abbreviate the squared Mahalanobis distance of x and y w.r.t. a positive

definite matrix B, $(x-y)^\top B^{-1}(x-y)$, by $M(x,y,B)$. It is known that Wilks' test for $g=2$ clusters is equivalent with Hotelling's two-sample T^2 test

$$\tfrac{n_1 n_2 (n-2)}{n^2} M(\overline{x}_1, \overline{x}_2, S) \sim T^2(d, n-2) \qquad \text{(under } \mathrm{H}_0\text{)}. \tag{6}$$

It is also known that T^2 is a function of the F distribution; see Mardia et al. [27], 5.3.3 and Theorem 3.6.1. Hotelling's two-sample test can also be interpreted as a union intersection test.

(b) **Hotelling test for cluster pairs**. The Wilks test for $g \geq 3$ clusters will reject equality if *one* mean deviates from the others. This may pose problems since a reasonable partition is characterized by separation of *all cluster pairs*. We therefore assess a g-cluster partition $\{C_1, \ldots, C_g\}$ $(g \geq 3)$ also by its least separated cluster pair w.r.t. Hotelling's test, characterized by the largest p-value of the squared Mahalanobis distance

$$\tfrac{n_j n_\ell (nj + n_\ell - 2)}{(nj + n_\ell)^2} M(\overline{x}_j, \overline{x}_\ell, S_{j,\ell}) \tag{7}$$

w.r.t. $T^2(d, n_j + n_\ell - 2)$ for all pairs C_j, C_ℓ, $j \neq \ell$. Here, $S_{j,\ell}$ denotes the pooled scatter matrix of C_j and C_ℓ.

Wilks' and Hotelling's tests are favorably applied when the true model is indeed homoscedastic. We next turn to tests applicable to the more general *heteroscedastic* normal case. They are somewhat more involved.

(c) **Behrens–Fisher's test**: Behrens [3] and R.A. Fisher [12, 13] test the hypothesis H_0 under no assumptions on the covariance matrices in hypothesis and alternative. Their original test is univariate. Its multivariate extension is well studied, too, and called the *multivariate Behrens–Fisher problem*; see Belloni and Didier [4], where an account of the extensive literature on this subject is found. The related p-value can again be considered as a measure of separation by location. Under the alternative of general parameters, the heteroscedastic normal classification likelihood w.r.t. $\mathcal{C} = \{C_1, \ldots, C_g\}$ is

$$-2 \log f_1(\mathbf{x}; \mathcal{C}, \mathbf{m}, \mathbf{V}) = \sum_{1 \leq j \leq g} \big[n_j \log \det(2\pi V_j) + \sum_{i \in C_j} (x_i - m_j)^\top V_j^{-1} (x_i - m_j) \big].$$

Its minimum w.r.t. $\mathbf{m} = (m_1, \ldots, m_g)$ and $\mathbf{V} = (V_1, \ldots, V_g)$ is obtained by considering each $j = 1, \ldots, g$ separately. Let $\overline{x}_j$ be the mean vector of group C_j, let $S_j(m) = \frac{1}{n_j} \sum_{i \in C_j} (x_i - m)(x_i - m)^\top$ be its scatter matrix about an arbitrary $m \in \mathbb{R}^d$, and let $S_j = S_j(\overline{x}_j)$ be its usual scatter matrix. Denoting the minimizers of $\mathbf{m}$ and $\mathbf{V}$ by $\mathbf{m}^*$ and $\mathbf{V}^*$, respectively, we obtain $m_j^* = \overline{x}_j$, $V_j^* = S_j$, and

$$-2 \log f_1(\mathbf{x}; \mathcal{C}, \mathbf{m}^*, \mathbf{V}^*) = nd(1 + \log 2\pi) + \sum_{1 \leq j \leq g} n_j \log \det S_j.$$

By assumption, the data is in general position and the group C_j is not deficient, that is, it has at least $d+1$ elements. Therefore, S_j is regular and $\log \det S_j$ is well defined.

Twice the negative classification log-likelihood with common mean vector m and general covariance matrices $\mathbf{V} = (V_1, \ldots, V_g)$ is

$$-2 \log f_0(\mathbf{x}; \mathcal{C}, m, \mathbf{V}) = \sum_{1 \leq j \leq g} \big[n_j \log \det 2\pi V_j + \sum_{i \in C_j} (x_i - m)^\top V_j^{-1} (x_i - m) \big]. \tag{8}$$

Its minimum w.r.t. m and $\mathbf{V}$ exists. Indeed, we have the lower bound

$$\begin{aligned}-2\log f_0(\mathbf{x};\mathcal{C},m,\mathbf{V}) &= nd\log 2\pi + \sum_{1\le j\le g} n_j\big[\log\det V_j + \operatorname{tr} S_j(m)V_j^{-1}\big]\\ &\ge nd\log 2\pi + \sum_{1\le j\le g} n_j\big[\log\det V_j + \operatorname{tr} S_j V_j^{-1}\big].\end{aligned}$$

It is well known from classical normal analysis that the jth summand on the right-hand side increases to ∞ as V_j approaches the boundary of the set of all positive definite matrices (that is, an eigenvalue vanishes or approaches ∞). It follows that (8) has a minimum which can, however, not be represented in closed form, not even in the univariate case. In Section 3, we will resort to an iteration.

Let m^* and $\widetilde{\mathbf{V}}^* = \big(\widetilde{V}_1^*,\dots,\widetilde{V}_g^*\big)$ minimize Eq. (8). Again by classical normal analysis applied to the jth summand in (8), $\widetilde{V}_j^* = S_j(m^*)$. We thus obtain the LRT statistic of the multivariate Behrens–Fisher problem,

$$\begin{aligned}&\mathrm{BF}(\mathcal{C}) = 2\big(\log f_1(\mathbf{x};\mathcal{C},\mathbf{m}^*,\mathbf{V}^*) - \log f_0(\mathbf{x};\mathcal{C},m^*,\widetilde{\mathbf{V}}^*)\big)\\ &= \sum_{1\le j\le g} n_j\big[\log\det S_j(m^*) - \log\det S_j\big] = \sum_{1\le j\le g} n_j\log\big(1+(\overline{x}_j-m^*)^\top S_j^{-1}(\overline{x}_j-m^*)\big)\\ &= \sum_{1\le j\le g} n_j\log\big(1+M(\overline{x}_j,m^*,S_j)\big).\end{aligned} \tag{9}$$

The last but one equality follows from $S_j(m^*) = S_j + (\overline{x}_j-m^*)(\overline{x}_j-m^*)^\top$ along with a well-known identity for rank-one perturbed determinants; see Muirhead [30], Theorem A3.5, or Ritter [32], Lemma A6(c). By a theorem due to Wilks [38] and Aitchison and Silvey [1] (see also Cox and Hinkley [9] or Silvey [35]), $\mathrm{BF}(\mathcal{C})$ is asymptotically distributed under H_0 as $\chi^2_{(g-1)d}$ with $(g-1)d$ degrees of freedom. If all clusters are large enough, the latter distribution can, therefore, be used to compute approximate p-values. Since χ^2 depends only on the number of clusters and not on the partition itself, $\mathrm{BF}(\mathcal{C})$ creates among several partitions the inverse ranking of the approximate p-values. The degree of separation of g clusters thus turns to the weighted sum (9) of g logarithms of squared Mahalanobis distances.

It remains to compute the optimal common mean vector m^*. We will propose a cutting plane algorithm for this problem in Section 3.

(d) **Behrens–Fisher test for cluster pairs**: When $g \ge 3$, we may again consider separation of all cluster pairs for a ranking. Thus, let C_j and C_ℓ be two clusters of $\mathcal{C}$. By Eq. (9), the Behrens–Fisher LRT statistic for equality of their mean vectors is

$$\mathrm{BF}(\{C_j,C_\ell\}) = n_j\log\big(1+M(\overline{x}_j,m^*_{j,\ell},S_j)\big) + n_\ell\log\big(1+M(\overline{x}_\ell,m^*_{j,\ell},S_\ell)\big),$$

where $m^*_{j,\ell}$ is the optimal common mean vector in Eq. (9) computed for the pair $\mathcal{C} = \{C_j,C_\ell\},\ j\ne\ell$. Again, the asymptotic χ^2 distribution is independent of clusters and sizes if they are large enough. Therefore, p-values under H_0 are ordered inversely to the test statistic and partitions $\mathcal{C} = \{C_1,\dots,C_g\}$ may be ranked according to the separation criterion

$$\min_{1\le j<\ell\le g} \mathrm{BF}(\{C_j,C_\ell\}). \tag{10}$$

We propose the Wilks (5), Hotelling for cluster pairs (7), Behrens–Fisher (9), and pairwise Behrens–Fisher (10) tests for ranking all local likelihood maxima (EM algorithm) or steady solutions (k-parameters algorithm). The pairwise Behrens–Fisher test utilizes Procedure 3.1 below in order to compute the vectors $m^*_{j,\ell}$ and the pairwise LRT statistic (10). All present ranking criteria are affine equivariant. In order to handle extremely small p-values in connection with large data sets, we use their logarithms. They can be computed directly from the statistics.

3 Computing the Behrens–Fisher test statistic

Despite the simple appearance of Eqs. (8), (9), and (10), computing their minima is not easy because of possible local minima. It has a long history. An early proposal is due to Mardia et al. [27]. They use the iteration

$$m \;\leftarrow\; \big(\textstyle\sum_{j=1,2} n_j S_j(m)^{-1}\big)^{-1}\big(\sum_{j=1,2} n_j S_j(m)^{-1}\overline{x}_j\big).$$

As above, $S_j(m) = S_j + (\overline{x}_j - m)(\overline{x}_j - m)^\top$ is the scatter matrix of C_j about m. The iteration is repeated until convergence. It can be shown that this occurs at a critical point m^* of (9), in most cases a local minimum. However, if the initial m is a local maximum, then the iteration stays there. Since local minima are not unique, the process has to be repeated from several initial values so that we can be (almost) sure to have finally attained the global minimum. This is the main disadvantage of Mardia's method.

An effective and efficient more recent approach in one sweep for $g = 2$ by Belloni and Didier [4] uses convex analysis. Their algorithm needs a computational parameter. We modify it here so that no parameter is needed. Instead of Eq. (8) it is possible to minimize Eq. (9) w.r.t. m since both differ just by the constant $2\log f_1(\mathbf{x}; \{C_1, C_2\}, \mathbf{m}^*, \mathbf{V}^*)$. It is a simple and well-known trick to reduce minimization of Eq. (9) for m to that of

$$h(u_1, u_2) = n_1 \log(1 + u_1) + n_2 \log(1 + u_2) \tag{11}$$

for (u_1, u_2) subject to the constraints

$$u_1 \geq M(\overline{x}_1, m, S_1),\;\; u_2 \geq M(\overline{x}_2, m, S_2).$$

Let Q_m be the right, upper quadrant with the left lower vertex $\big(M(\overline{x}_1, m, S_1), M(\overline{x}_2, m, S_2)\big) \in \mathbb{R}^2_+$ in the plane. The following subset of the plane plays a key role in theory and algorithm:

$$\mathcal{K} = \bigcup_{m \in \mathbb{R}^d} Q_m. \tag{12}$$

In Lemma 7.1 of the Appendix it will be shown that it is closed in $\mathbb{R}^2$ and convex. Since h increases with u_1 and u_2, the minimum of h on $\mathcal{K}$ is attained on its lower Pareto boundary

$$\{(u_1, u_2) \,|\, 0 \leq u_1 \leq M(\overline{x}_2, \overline{x}_1, S_2), u_2 = \min\{v \,|\, (u_1, v) \in \mathcal{K}\}.$$

This is also the set of extreme points of $\mathcal{K}$ and also the graph of the function g defined by

$$g(u_1) = \min\{u_2 \,|\, (u_1, u_2) \in \mathcal{K}\} = \min_{m:\; M(\overline{x}_1, m, S_1) = u_1} M(\overline{x}_2, m, S_2), \tag{13}$$

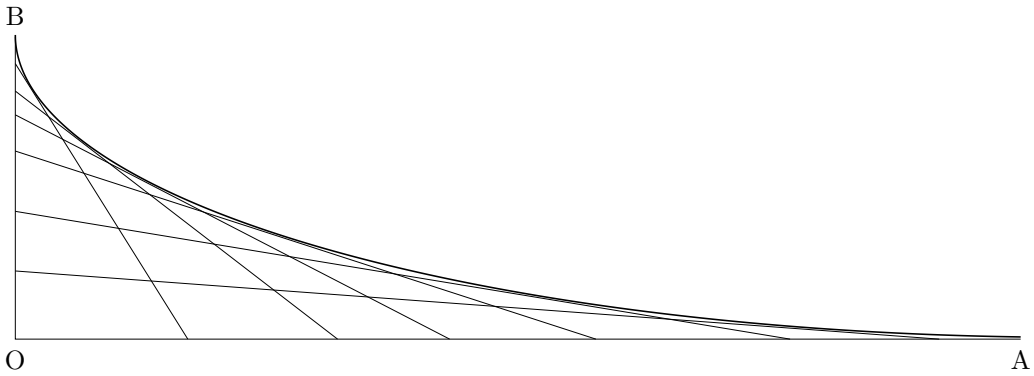


Figure 1: Schematic view of the graph of g and some tangents.

$u_1 \in [0, M(\overline{x}_2, \overline{x}_1, S_2)]$. By convexity of $\mathcal{K}$, the function $g\colon [0, M(\overline{x}_2, \overline{x}_1, S_2)] \to \mathbb{R}_+$ is convex. Using a Lagrange multiplier λ, it will be shown in Lemma 7.2 that g is even strictly convex and real-analytic, and that $g' = -\lambda$. These three properties open a way to approximating the constrained minimum of the function h defined by Eq. (11). The problem is known as convex programming and favorably solved by the cutting plane method; see, for instance, Bonnans et al. [7]. The cutting plane method assumes that the target function is defined on a neighborhood of a compact, convex set. It surrounds the set with a sequence of polygons whose edges touch the set. The minimum of h on each polygon is assumed at a vertex. Each polygon is constructed as a refinement of its predecessor. The refinement takes place at the current minimal vertex which is replaced with two new ones. More precisely, the algorithm iterates the following procedure.

3.1 Cutting plane procedure

// Input: The ordered list $\mathcal{V}$ of the vertices of a convex polygon and the vertex $v \in \mathcal{V}$ that minimizes h on $\mathcal{V}$.
// Output: Updated list and minimizing vertex.

1. Let e_1 and e_2 be the edges of the polygon adjacent to v and let s_1 and s_2 be the abscissas of the points where they touch the graph of g. Construct the tangent l to the graph of g at the midpoint of s_1 and s_2;
2. intersect l with e_1 and e_2 to obtain two points v_1 and v_2;
3. insert v_1 and v_2 in place of v into $\mathcal{V}$ to obtain $\mathcal{V}_{\text{new}}$;
4. determine the minimizer $v_{\text{new}} \in \mathcal{V}_{\text{new}}$ of h on $\mathcal{V}_{\text{new}}$.

A computer program favorably stores more information on the tangents in order to avoid multiple recomputations. It is also favorable to replace the midpoint in step 1 with a convex combination in the direction of the smaller touching point, for instance,

$$[h(s, g(s))t + h(t, g(t))s]/[h(s, g(s)) + h(t, g(t))].$$

The *cutting plane algorithm* iterates the procedure starting from the polygon $\mathcal{V} = \{B, O, A\}$ (see Figure 1) and the origin O as the current minimizing vertex. It produces a sequence of polygons and minimizing vertices of h. The minima increase to the minimum value on the graph of g and we have the following theorem. Its proof is deferred to the appendix.

3.2 Theorem *(a) Any cluster point of the sequence of minimizing vertices solves the constrained minimization problem* (11)*.*

(b) If the minimizer is unique then it is the limit of the cutting plane algorithm.

We finally note that the constrained minimization problem (11) is equivalent to minimizing the function $u_1 \mapsto n_1 \log(1+u_1) + n_2 \log(g(1+u_1))$, $u_1 \geq 0$. Since g is analytic, this can be carried out by the Newton–Raphson algorithm. Since $g' = -\lambda$, the second derivative of g w.r.t. u_1 can be computed with the aid of Eq. (16) in the appendix. Convergence will, however, take place to any critical point close to the initial point and so, Newton–Raphson has to be replicated from many random initial points. It has thus the same disadvantage as the algorithm of Mardia et al. mentioned above.

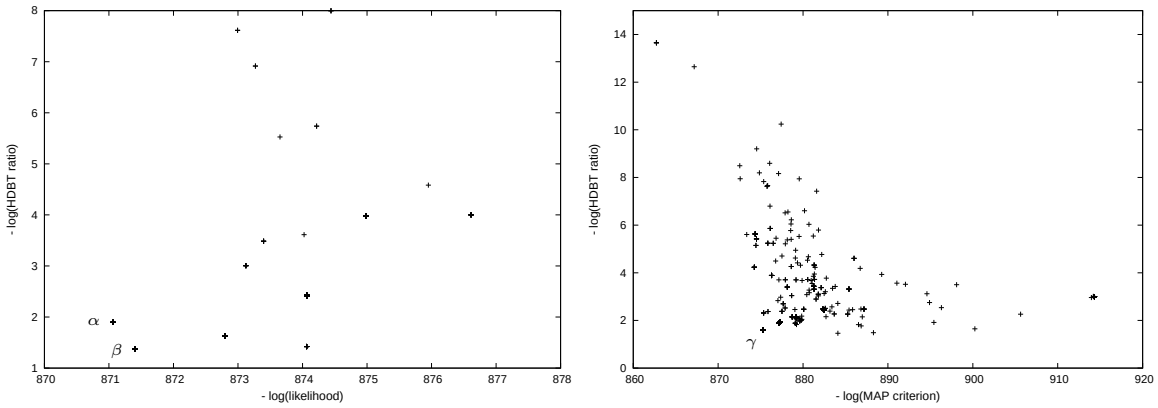
4 Scale balance, fit, and Pareto solutions

In [17] and [18], we introduced the concept of SBF plot of local likelihood maxima and steady solutions, respectively, as an aid to select the desired solution in mixture and cluster analysis; see also item (iv) in Section 1. We will use it again as a major tool in the sequel. Its definition needs the HDBT ratio of g scale matrices. (The acronym HDBT stands for Hathaway, Dennis, Beale, and Thompson; see Gallegos and Ritter [18] or Ritter [32].) The positive semi-definite (or Löwner) order of two symmetric matrices $A, B \in \mathbb{R}^{d\times d}$ is defined by "$A \preceq B$ if and only if $B - A$ is positive semi-definite." The order is total if and only if $d = 1$, in which case the matrices are numbers and the order coincides with the usual order on the real line. An HDBT constant for g positive definite matrices V_j is any number c such that the relations

$$cV_j \preceq V_\ell$$

are satisfied for all indices $j, \ell = 1, \ldots, g$. Necessarily, $c \leq 1$. In order to asses the heterogeneity of the g matrices, we use the HDBT ratio r_{HDBT}, the largest HDBT constant. It can be easily computed as $r_{\mathrm{HDBT}}(\mathbf{V}) = \min_{j,\ell,k} \lambda_k\big(V_\ell^{-1/2} V_j V_\ell^{-1/2}\big)$, where $\lambda_1(A), \ldots, \lambda_d(A)$ denote the d eigenvalues of a $d \times d$ matrix A; see Ritter [32], A.10. An alternative formula is $r_{\mathrm{HDBT}}(\mathbf{V}) = \min_{j,\ell,k} \lambda_k\big(V_\ell^{-1} V_j\big)$. Although the matrix $V_\ell^{-1} V_j$ is not necessarily symmetric, its eigenvalues are all real and equal to those of $V_\ell^{-1/2} V_j V_\ell^{-1/2}$. HDBT constants and ratios are bounded above by 1. The HDBT ratio is a measure of balance of the component scales. The larger the HDBT ratio is, the larger the balance and the less the heterogeneity. The homoscedastic case $V_1 = \cdots = V_g$ is easily characterized by maximum scale balance, $r_{\mathrm{HDBT}}(\mathbf{V}) = 1$.

We apply the HDBT ratio to the scale matrices of mixture and classification models. A plot of $-\log r_{\mathrm{HDBT}}$ vs. the negative logged values of all local likelihood maxima (mixture model) or negative log-criteria of all steady solutions (classification model) for a data set is called SBF plot. It depends on the data and on the number of clusters, g. In practice, the plot points are generated by repeated runs of the algorithm from random starting points. We use the SBF plot as a major tool for selecting the favorite solution. Two SBF plots for the synthetic data of Section 5.1 are presented in Figure 2. Part (a) shows the SBF plot for 1500 runs of the EM algorithm and Part (b) presents the left half of the SBF plot of all steady solutions obtained with 1500 runs of the k-parameters algorithm. Note the larger number of steady solutions for the same number of runs. These plots contain often hundreds or thousands of points leaving



(a) (b)

Figure 2: (a) SBF plot for the synthetic data of Section 5.1 w.r.t. the mixture model and (b) the left half of that for the classification model.

us at a loss which solution to choose. The problem occurs in particular with sensitive and flexible models such as normal mixtures or the classification model with full scales.

In order to appreciate the importance of HDBT ratios and SBF plots, we recall the consistency theorem for mixtures, Theorem 1.1. Hathaway [21] stated and proved another theorem for *scale constrained* mixture models. Its presentation needs some preparation. Let

$$\overset{\circ}{\Delta}_{g-1} = \big\{\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g) \,|\, \pi_j > 0,\ \textstyle\sum_j \pi_j = 1\big\}$$

be the open, g–1-dimensional unit simplex. Let $\mathcal{V}_c$ stand for the set of all finite collections (tuples or sets) of positive definite d by d matrices with HDBT ratio at least c. The set $\mathcal{V}_c$ increases with decreasing c. We consider a constrained, normal, g-component mixture model with HDBT constraint (or constant) c and parameter space

$$\Theta_{g,c} = \overset{\circ}{\Delta}_{g-1} \times \big\{(\mathbf{m}, \mathbf{V}) \in \mathbb{R}^{gd} \times \mathcal{V}_c \,|\, (m_j, V_j),\ j \in 1\,..\,g,\ \text{pairwise distinct}\big\}.$$

Its covariance matrices do not deviate too much from each other, depending on c. The advantage of HDBT constraints over other scale constraints is their affine equivariance. For example, the popular constraint "all eigenvalues of all V_j's $\geq \varepsilon$ for some $\varepsilon > 0$" does not enjoy this property. After these preliminaries, we can state a multivariate version of Hathaway's theorem on the HDBT constrained MLE. It is a corollary of Ritter [32], Theorem 1.27, which treats the more general case of constrained, multivariate, *elliptical* mixtures.

4.1 Theorem *Let the sampling distribution be a g-component, normal mixture with parameters $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$, (m_j, V_j) pairwise distinct.*

(a) For any $0 < c \leq 1$, a constrained ML estimate T_n exists in $\Theta_{g,c}$ for eventually all n, a.s..

(b) Let $c \leq r_{\mathrm{HDBT}}(\mathbf{V})$ and let $(T_n)_n$ be any sequence of ML estimates in $\Theta_{g,c}$. A representative $\widetilde{T}_n$ of T_n w.r.t. label switching can be chosen for each n in such a way that the sequence $(\widetilde{T}_n)$ is strongly consistent at $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$ w.r.t. the Euclidean metric on $\Theta_{g,c}$.

Theorem 1.1 in Section 1 suffers from the problem that it doesn't betray us the desired local maximum. Hathaway's theorem 4.1, too, cannot be applied right away. While its Part (b) tells us that an arbitrarily small HDBT constant c is possible *asymptotically* as $n \to \infty$, the solution for a *finite* data set depends on the constant, in particular when the data set is small. The theorem does not reveal a suitable constant c, which appears as another unwanted parameter in the estimation process. As a result, neither of the two theorems determines a single solution to the clustering problem, even if there is a true one.

As a way out of the dilemma, Ritter [32], Section 4.1, combined the two consistency theorems. Theorem 1.1 shows that the mixture model contains information on a consistent solution in the form of a local maximum without recurring to constraints. Combined with Theorem 4.1, it teaches us that, in our search for a consistent solution, we may confine ourselves to (unconstrained) local likelihood maxima with an HDBT ratio above *some* constraint c. They are the candidates for the consistent solution. It is shown in [32] and easy to see that they just belong to the minima w.r.t. the product order in the SBF plot, the so-called *lower Pareto points*. They can also be defined as the collection of all points such that the left lower quadrants emanating from them do not intersect any other point in the plot or, again, as the minimum set of points such that the right upper quadrants emanating from them cover all points. Pareto points are solutions to biobjective optimization problems, here the joint optimization of scale balance and fit represented by HDBT ratio and likelihood, respectively. They are rarely unique and the present situation is no different. In fact, there often exist multiple reasonable solutions to clustering problems. This is one item that makes cluster analysis difficult in practice. The solutions corresponding to Pareto points are called *Pareto solutions*. They render constrained optimization as required by Hathaway's theorem unnecessary. The above reasoning establishes large HDBT ratio (scale balance) as a second principle complementary to large likelihood 2 or determinant criterion (4) (fit). Both assets of the solutions are equally important. Generally, algorithms that monitor exclusively the likelihood function should be refused in cluster analysis.

The two consistency theorems are asymptotic statements. There is a risk that, due to randomness of the finite data, the desired solution is masked by an undesired one which happens to have larger scale balance and likelihood. As an additional idea, we counteract this risk by performing the tests of Section 2 before extracting Pareto points, thus, first cleaning the SBF plot from solutions that are unsuitable because of insufficient separation. We obtain the following overall clustering strategy w.r.t. the classification model (3) and the k-parameters algorithm with a fixed number of components, g.

4.2 Clustering strategy

// Input: A data set and a natural number $g \geq 2$.
// Output: A set of solutions.

1. Use the k-parameters algorithm to generate as many steady partitions as possible;
2. extract all sufficiently well separated partitions using the p-values under H_0 of one or all statistics (5), (7), (9), (10); if no partition passes the test, this indicates that the number of clusters, g, is too large.
3. determine the Pareto points in the SBF plot established with these solutions;
4. return all Pareto solutions with sufficiently large determinant criterion (4) and rank them by p-values.

The strategy is also applicable to mixture models. The k-parameters algorithm in 1 is then replaced with the EM algorithm followed by the Bayesian discriminant rule to partition the data set. Strategy 4.2 is necessarily fuzzy and not an algorithm, as will now be explained.

(i) In item 1, "as many ... as possible" refers to the available computing time and memory.

(ii) "Well separated" in item 2 means in general p-values below 10^{-15} or even 10^{-20}. It depends on the size of the data. The solution with the least p-value is not necessarily favored by the above strategy nor does it always represent the desired solution.

(iii) Pareto points are derived from (asymptotic) consistency theorems. For a *finite* data set, log-HDBT ratio and log-likelihood are subject to random fluctuations which may be controlled by confidence regions. Therefore, the analyst may also consider in item 3 separated solutions that become Pareto after a slight left or downward shift ("*nearly* Pareto solutions"). It turns out that several (nearly) Pareto points could in principle represent the true solution.

(iv) Computation of Pareto points in item 3 can be done online during the generation of local likelihood maxima by a general algorithm called *list updating* in computer science; see Ritter [32], Procedure 4.4.

(v) The solution with the maximum criterion (maximum local likelihood) is always Pareto. Sometimes it has a very small HDBT ratio which usually indicates a spurious solution. Nevertheless, it could be the desired solution since the data set could, in principle, have been sampled from it (for instance, in a simulated data set). In real-world applications, we would rather doubt that it is reasonable. However, a comprehensive strategy has to return this solution, too.

(vi) On the other hand, the steady solution with the largest HDBT ratio (scale balance), too, is always a Pareto solution. If its determinant criterion is poor, it is again not likely to be reasonable. Partitions with poor criteria should be rejected.

To summarize, in our search for the desired solution, it is sufficient to focus on (nearly) Pareto solutions in the set of separated solutions that fit the data well. On the other hand, each of these solutions could be the sought one. Note, however, that the best solutions might not be good. It is, therefore, necessary to validate all partitions obtained. There are some refined methods such as visualization, other separation indices, normality tests, and cluster stability; see, for instance, Chapter 4 of Ritter [32].

When the number of clusters, g, is unknown, the procedure is repeated for all reasonable numbers. Again, validation indicates reasonable numbers of mixture components. Another simple and convenient although crude way for assessing this number is the BIC. The reader should bear in mind that data sets with overlapping clusters often admit neither a unique solution, even for a fixed number of components, nor do they in general disclose unambiguously their parental number of components. What's more, the data in our hands might not bear any cluster structure at all.

5 Three examples

In order to illustrate the methods of Sections 2 and 4 we will now analyze a synthetic and two real data sets.

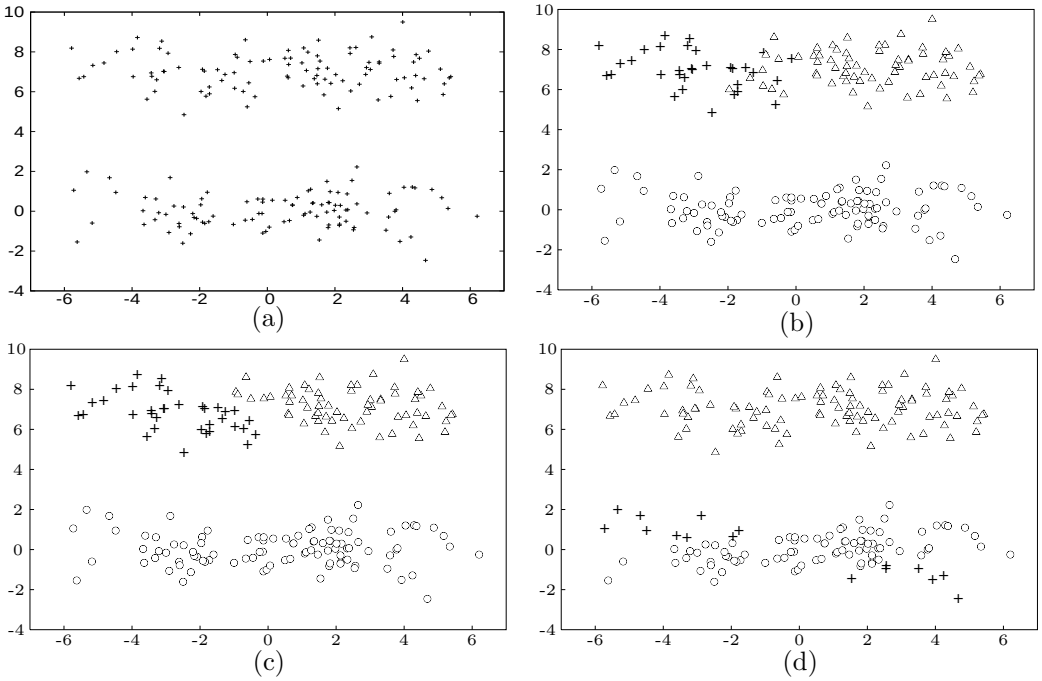


Figure 3: (a) Synthetic data set; (b) true partition in three clusters; (c) Wilks, pairwise Hotelling, and pairwise Behrens–Fisher (Pareto) solution recovered by the mixture model (β in Figure 2(a)) and Behrens–Fisher (Pareto) solution of the classification model (γ); (d) second Pareto solution for the mixture model (α).

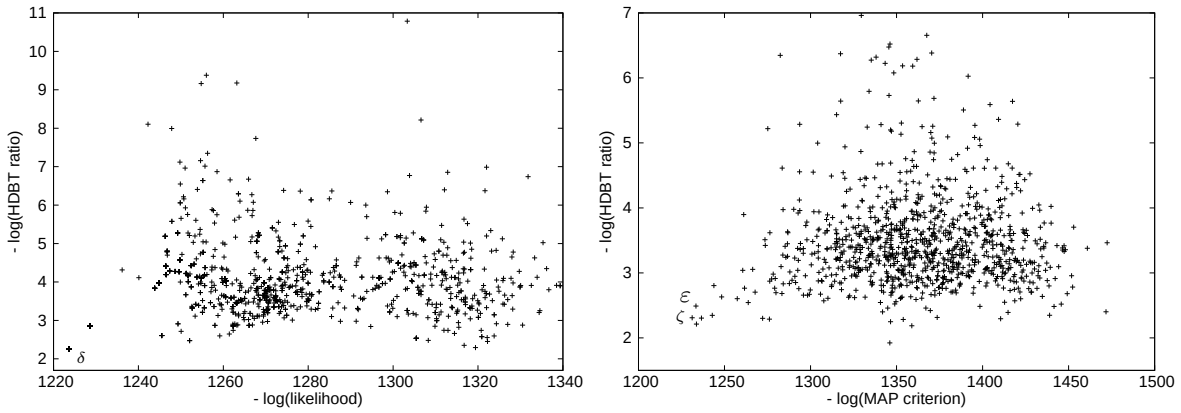
5.1 Synthetic data set

Our first example is the two-dimensional synthetic data set shown in Figure 3(a). At first glance, it might seem to be composed of two or, maybe, four components. It was however sampled from three normal components with mean vectors (0,0), (-2.6,7), and (2.2,7), respectively. The lower component is diagonal with variances 9 and 1, the covariance matrices of the other two components are $\begin{bmatrix} 2.2 & -0.37 \\ -0.37 & 0.79 \end{bmatrix}$ and $\begin{bmatrix} 4 & 0.04 \\ 0.04 & 0.68 \end{bmatrix}$, respectively. The original partition in clusters of sizes 99, 32, and 67 is presented in Figure 3(b). The plot exhibits a small overlap of the two upper clusters. We wish to estimate the original components.

The SBF plot of the fifteen local maxima of the mixture likelihood shown in Figure 2(a) results from thousand replications of the normal mixture EM algorithm for three components. Figure 3(c) shows the local maximum with the best Hotelling (7) and Behrens–Fisher (10) separation criteria for cluster pairs. This solution belongs to the Pareto point (β) in Figure 2(a). It is, therefore, our favored solution. It belongs indeed to the solution closest to the original partition among all local likelihood maxima, enjoying the largest adjusted Rand index. Note that it has poorer fit (likelihood) but larger scale balance (HDBT ratio) compared with the solution associated with Pareto point (α). The latter is shown in Figure 3(d). Two of its components are separated by scale, only, forming a cross. The example shows that it is not always fit that matters most. Sometimes it is even the Pareto solution with the least fit that turns out to be the desired one.

The more interesting (left) half of the SBF plot for the classification model and the k-parameters algorithm is shown in Figure 2(b). We obtain a common optimal solution for the Hotelling and Behrens–Fisher criteria for cluster pairs and another one for the Wilks criterion. The negative MAP determinant criterion of the former is 972 (in the right-hand half, not shown) and that of the latter 875.3. For reasons of fit we accept the Wilks solution. It is again the partition shown in Figure 3(c) and belongs to the (extreme) Pareto point γ in Figure 2(b).

5.2 CRABS data



(a) (b)

Figure 4: CRABS: SBF plots for (a) the mixture model and (b) the MAP classification model with four components.

CRABS is a real data set contained in the MASS package of the R documentation. It was compiled at Fremantle, W. Australia, and first cluster analyzed by Campbell and Mahon [8]. The data describes five morphological measurements, "frontal lobe size," "rear width," "carapace length," "carapace width," and "body depth," all in mm, on 50 crabs, each, of four classes of Leptograpsus variegatus. The classes are all four combinations of the two species "blue" and "orange" and the two sexes. Visual inspection of pairs of features indicates a voluminous data set. This suggests application of the present methods.

The SBF plot for the mixture model after 1200 replications of the EM algorithm (Figure 4(a)) exhibits a multitude of local likelihood maxima. We make the only Pareto point (δ) our favorite solution. Its fit, scale balance, and pairwise Behrens–Fisher p-value ($1.37 \cdot 10^{-19}$) are superior to all others. Comparison with the known true solution shows 15 misclassifications.

The results for the classification model are as follows. The SBF plot after 1000 replications of the k-parameters algorithm is shown in Figure 4(b). The best pairwise Behrens–Fisher solution ($p = 8.5 \cdot 10^{-28}$) after 100 000 replications is marked by the letter ε (1233,2.49). It is not a Pareto point, but nearly. The best pairwise Hotelling solution is at (1260,2.51). Its likelihood is smaller and the point is located farther away from Pareto points and therefore rejected. The best Wilks solution ($p = 3.2 \cdot 10^{-160}$) is marked with ζ (1229,2.23). This is a

Pareto point. We therefore select the first and the third solution as our favorites. Comparison with the known true solution results in 9 and 17 errors, respectively. The contingency tables for the three favorite solutions w.r.t. the true one are displayed in Table 2. Cluster order is BM, BF, OM, OF.

5.3 SPIKE SORTING

Understanding the computational mechanism of the mammalian brain has been greatly promoted by the advent of dense microfabricated electrode arrays with many closely spaced recording sites. They allow simultaneous recording of the signals issued by large numbers of neurons. The signal of each electrode consists of noise superposed by firing events, so-called *spikes*. These local spatiotemporal events indicate mental activity. The task of assigning the different spikes to neurons is called *spike sorting* in the diction of neurophysiologists.

Spike sorting proceeds in three steps. The firing pattern of each neuron is first separated from noise by methods of signal processing and pattern recognition. The extracted waveforms of all spikes are then summarized by feature vectors, typically using principal component analysis. In a third step, feature vectors are divided into groups corresponding to putative neurons. It is here, where cluster analysis comes in.

We apply our clustering strategy to a subsample of data recorded from rat neocortex with 32-channel shank electrodes compiled by Kenneth Harris and his group at University College London; see Rossant et al. [33]. The SPIKE DATA used here consists of 19 756 data points in $32 \cdot 3 = 96$ variables (three principal components were extracted for each channel). We visually removed the 65 variables with elliptical pairwise scatter plots. Visual inspection shows that the data contains many outliers. These are known to heavily distort clustering results. We, therefore, applied the k-parameters clustering algorithm with combinatorial trimming for outlier protection proposed by Gallegos and Ritter [19]. It contains as parameters the numbers of clusters and of discarded elements. We combined it with the present pairwise Behrens–Fisher and Hotelling tests.

The plot w.r.t. variables 16 and 79 suggests that there are at least six clusters; see Figure 5(a). The algorithm was, therefore, run with $g = 6, 7, 8, \cdots$ clusters up to the first number where all steady solutions found failed the separation test, and with 5, 10, 15, and 20 percent of discarded elements. The number $g = 18$ was the first number where a cluster was split. Hence our estimate of the number of clusters is 17. Of the four numbers of discards, 15% was the smallest one where the clusters obtained appeared sufficiently clean (visual inspection). This is our estimated percentage of outliers. The SBF plot for 17 clusters and 15% of discards is shown in Figure 5(b). We chose as final solution the Pareto partition with best separation

Table 2: CRABS: Contingency tables of the favorite solutions for the mixture (left) and the MAP classification model.

39	11	0	0
0	49	0	1
0	0	50	0
0	0	3	47

(δ)

49	1	0	0
5	45	0	0
0	0	50	0
0	0	3	47

(ε)

37	13	0	0
0	50	0	0
0	0	50	0
0	0	4	46

(ζ)

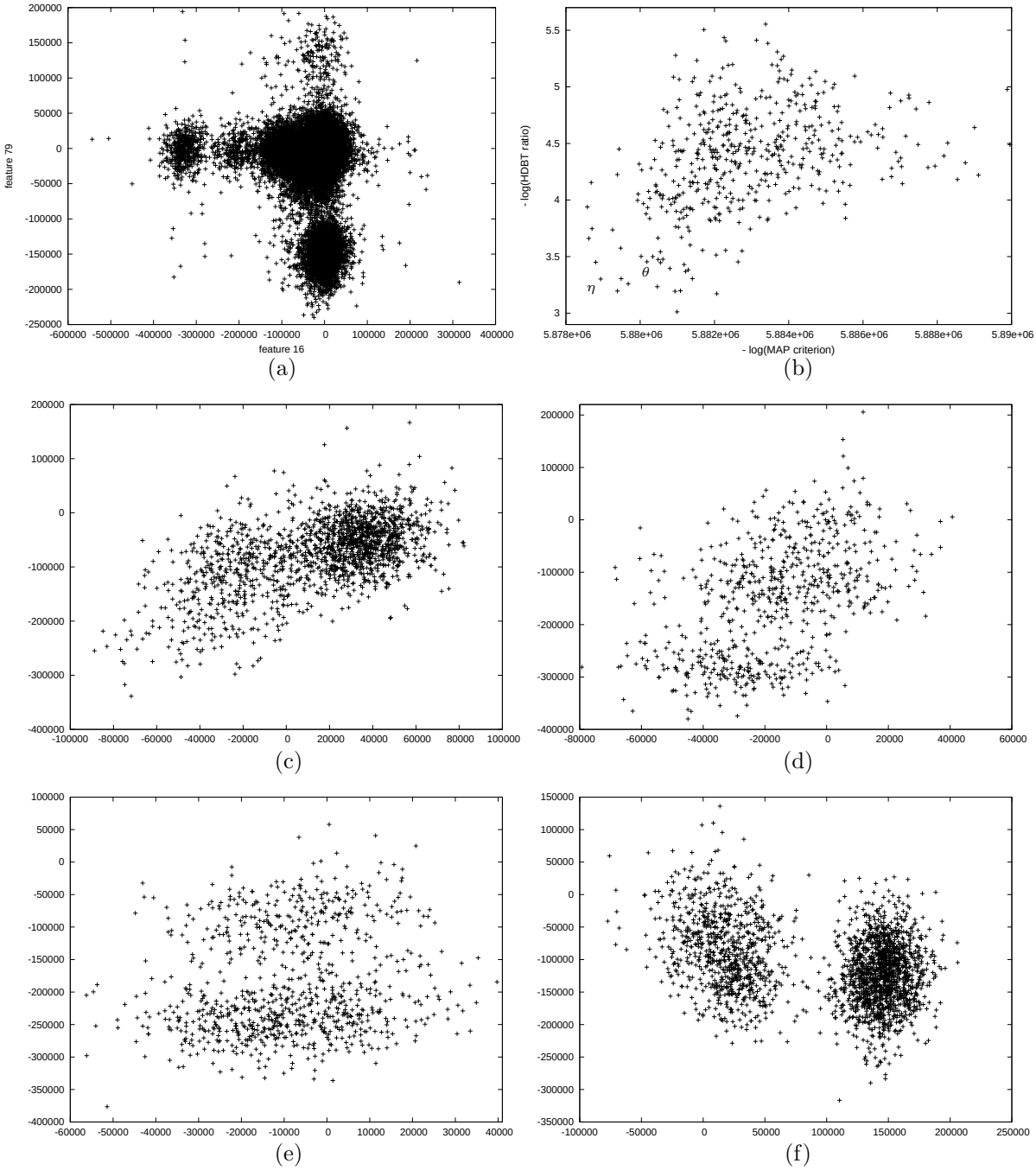


Figure 5: SPIKES: (a) Projection of the data set to two coordinates displaying structure; (b) SBF plot of steady solutions for the MAP classification model with 17 clusters; (c)-(e) projections of the three cluster pairs with poorest separation in directions that optimally exhibit the separation; (f) good separation. The SBF plot (b) and the data plots (c)–(f) refer to the final partition with 17 clusters after removing 15% of outliers.

w.r.t. the pairwise Behrens–Fisher test, η. The sizes of its clusters are 3019, 2641, 1833, 1489, 1290, 923, 815, 792, 645, 611, 441, 429, 410, 377, 366, 348, and 327. Figure 5(c)-(f) displays the separation of some cluster pairs by projections to pair-specific planes which contain the

vector from Fisher's linear discriminant analysis; see also Ritter [32], Section 4.4.3. This projection exhibits their separation best. Plots (c)-(e) display the pairs with poorest separation; most cluster pairs look like plot (f). The maximum Behrens–Fisher p-value for any two clusters to come from two normal populations with the same mean vector is $6.39 \cdot 10^{-26}$ (Figure 5(e)).

Visual inspection shows that the assumption of normality of components was appropriate. We finally mention that partition θ in Figure 5(b) enjoys better minimum separation having a maximum p-value of $5.54 \cdot 10^{-110}$ (pairwise Behrens–Fisher) and $1.97 \cdot 10^{-182}$ (Hotelling) over all cluster pairs. It is not nearly Pareto and therefore not acceptable as a solution.

6 Discussion

We have proposed strategies for reducing the large number of locally optimal partitions of the mixture and classification models. Their generation with the related EM and k-parameters algorithms needs many replications. This computational load may be considered a disadvantage. We claim that, in general, there is no simple algorithm with comparable reliability. Moreover, there is a straightforward parallel algorithm. It is sufficient to start the EM (or k-parameters) algorithm from random partitions on m processors in parallel. The solutions are fed to another processor for statistical testing and, finally, extraction of Pareto solutions. This method reduces the processing time by a factor of m.

All present methods are based on normal assumptions. This seems to be rather strict. Two remedies in literature are more general distributions, such as elliptical symmetries or skewed distributions, Lee and McLachlan [24, 25], as well as nonparametric setups. All approaches bear certain risks when applied to cluster analysis. Generally, cluster analysis depends on the cluster models sought. If classes are actually asymmetric and normal mixtures are used, then too large a number of components could result. If data is sampled from a mixture of normals and the model is skew-normal, then some overlapped symmetric components might be merged. Some nonparametric methods hinge on prior metrics or bandwidths. Their choice depends on the solution sought.

The proposed testing methods belong to the realm of cluster validation. Some authors claim that this is the most difficult part of cluster analysis. We agree with this opinion and note that there are some acceptable, equivariant methods. Some are found in Chapter 4 of Ritter [32]. Finally, it is a good idea to apply exploratory methods to cluster validation in order to avoid the deepest pitfalls. According to Tukey [36], *"there is no excuse for failing to plot and look."*

References

[1] J. Aitchison and S.D. Silvey. Maximum-likelihood estimation procedures and associated tests of significance. *J. Royal Statist. Soc., Series B*, 22:154–171, 1960.

[2] T. A. Bailey, Jr. and R. C. Dubes. Cluster validity profiles. *Patt. Rec.*, 15:61–83, 1982.

[3] W. U. Behrens. Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher. Zeitschrift für wissenschaftliche Landwirtschaft und Archiv des Königlich Preussischen Landes-Oekonomie-Kollegiums*, 68:807–837, 1929. Original in Hathi Trust Digital Library.

[4] A. Belloni and G. Didier. On the Behrens–Fisher problem: A globally convergent algorithm and a finite-sample study of the Wald, LR and LM tests. *Ann. Statist.*, 36:2377–2408, 2008.

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

[6] D. Böhning. *Computer-Assisted Analysis of Mixtures and Applications.* Chapman & Hall/CRC, Boca Raton, 2000.

[7] J.-F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal. *Numerical Optimization. Theoretical and Practical Aspects.* Springer, Berlin, 2nd edition, 2006.

[8] N. A. Campbell and R. J. Mahon. A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Austral. J. Zool.*, 22:417–425, 1974.

[9] D. R. Cox and D. V. Hinkley. *Theoretical Statistics.* Chapman & Hall, London, 1974.

[10] N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

[11] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York etc., 1996.

[12] R. A. Fisher. The comparison of samples with possibly unequal variances. *Ann. Eugenics*, 9:174–180, 1939.

[13] R. A. Fisher. The asymptotic approach to Behrens' integral with further tables for the d test of significance. *Ann. Eugenics*, 11:141–172, 1941.

[14] C. Fraley and A. E. Raftery. MCLUST: Software for model-based cluster analysis. *J. Classification*, 16:297–306, 1999.

[15] H. Fritz, L. A. García-Escudero, and A. Mayo-Iscar. A fast algorithm for robust constrained clustering. *Computational Statistics and Data Analysis*, 61:124–136, 2013.

[16] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models.* Springer, Heidelberg, 2006.

[17] M. T. Gallegos and G. Ritter. Trimmed ML-estimation of contaminated mixtures. *Sankhyā, Series A*, 71:164–220, 2009.

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

[19] M. T. Gallegos and G. Ritter. Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. *Computational Statistics and Data Analysis*, 54:637–654, 2010. DOI 10.1016/j.csda.2009.08.023.

[20] M. T. Gallegos and G. Ritter. Strong consistency of k-parameters clustering. *J. Multivariate Anal.*, 117:14–31, 2013.

[21] R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, 13:795–800, 1985.

[22] J. Kiefer and J. Wolfowitz. Consistency of the maximum-likelihood estimation in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27:887–906, 1956.

[23] N. M. Kiefer. Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica*, 46:427–434, 1978.

[24] S. X. Lee and G. J. McLachlan. On mixtures of skew normal and skew t-distributions. *Adv. Data Anal. Classif.*, 7:241–266, 2013.

[25] S. X. Lee and G. J. McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Stat. Comput.*, 24:181–202, 2014.

[26] B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. IMS and ASA, Hayward, California and Alexandria, Virginia, 1995.

[27] K. V. Mardia, T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, New York, Toronto, Sydney, San Francisco, 6th edition, 1997.

[28] G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York etc., 2000.

[29] G. J. McLachlan and D. Peel. On computational aspects of clustering via mixtures of normal and t-components. In *Proceedings of the American Statistical Association*, Alexandria, Virginia, August 2000. American Statistical Association.

[30] R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, 1982.

[31] B. C. Peters, Jr. and H. F. Walker. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.*, 35:362–378, 1978.

[32] G. Ritter. *Robust Cluster Analysis and Variable Selection*, volume 137 of *Monographs in Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, London, New York, 2015.

[33] C. Rossant, S. Kadir, D. F. M. Goodman, and K. D. Harris. Spike sorting for large, dense electrode arrays. *Nature Neuroscience*, 19:624–641, 2016.

[34] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

[35] S. D. Silvey. *Statistical Inference*. Penguin, Baltimore, 1970.

[36] J. W. Tukey. *Exploratory Data Analysis*. Addison–Wesley, Reading/Mass., 1977.

[37] S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24:471–494, 1932.

[38] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9:60–62, 1938.

[39] S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *Ann. Statist.*, 39:209–214, 1968.

Appendix

7.1 Lemma *The set $\mathcal{K}$ introduced in Eq.* (12) *is closed in $\mathbb{R}^2$ and convex.*

Proof. Let $(u_1^{(k)}, u_2^{(k)})$ be a sequence of pairs in $\mathcal{K}$, convergent to $(u_1, u_2) \in \mathbb{R}^2$. There exists a sequence $m^{(k)}$ in $\mathbb{R}^d$ such that

$$u_j^{(k)} \geq M(\overline{x}_j, m^{(k)}, S_j) \tag{14}$$

for all k and $j = 1, 2$. It follows that $m^{(k)}$ is bounded. By Bolzano–Weierstrass, we may without loss of generality assume that it converges to $m \in \mathbb{R}^d$. Closedness of $\mathcal{K}$ follows after passing to the limit as $k \to \infty$ in Eq. (14).

In view of convexity of $\mathcal{K}$, let $u = (u_1, u_2)$, $v = (v_1, v_2) \in \mathcal{K}$ and let $0 < \lambda < 1$. By definition,

$$u_j \geq M(\overline{x}_j, m_u, S_j) \text{ and } v_j \geq M(\overline{x}_j, m_v, S_j)$$

for $j = 1, 2$ and two vectors $m_u, m_v \in \mathbb{R}^d$. By convexity of $m \mapsto M(x, m, S)$ it follows for $j = 1, 2$

$$(1-\lambda)u_j + \lambda v_j \geq (1-\lambda)M(\overline{x}_j, m_u, S_j) + \lambda M(\overline{x}_j, m_v, S_j) \geq M(\overline{x}_j, (1-\lambda)m_u + \lambda m_v, S_j).$$

Hence, $(1-\lambda)u + \lambda v \in \mathcal{K}$. This is the second claim. □

We next deal with the properties of the function g defined in Eq. (13).

7.2 Lemma *The function g defined in Eq.* (13) *is real-analytic and strictly convex; moreover $g' = -\lambda$.*

Proof. The function g arises from minimizing the parabolic function $M(\overline{x}_2, m, S_2)$ on the set of all $m \in \mathbb{R}^d$ such that $M(\overline{x}_1, m, S_1) = u_1$. Application of the Lagrange function

$$M(\overline{x}_2, m, S_2) + \lambda(M(\overline{x}_1, m, S_1) - u_1)$$

with multiplier λ leads to the system of equations

$$\begin{cases} \mathrm{D}_m(m-\overline{x}_2)^\top S_2^{-1}(m-\overline{x}_2) + \lambda \mathrm{D}_m(m-\overline{x}_1)^\top S_1^{-1}(m-\overline{x}_1) = 0, \\ (m-\overline{x}_1)^\top S_1^{-1}(m-\overline{x}_1) = u_1 \end{cases}$$

for m and the multiplier λ. The multiplier is positive since we require the minimum of a parabolic function on a convex set. The first equation reduces to

$$S_2^{-1}(m-\overline{x}_2) + \lambda S_1^{-1}(m-\overline{x}_1) = 0. \tag{15}$$

Without loss of generality we assume from here on that $S_1 = I_d$, the identity matrix, and $\overline{x}_1 = 0$. Eq. (15) yields $m = (I_d + \lambda S_2)^{-1}\overline{x}_2$. Inserting into the second equation above, $\|m\|^2 = u_1$, we find

$$\|(I_d + \lambda S_2)^{-1}\overline{x}_2\|^2 = u_1. \tag{16}$$

Since $\overline{x}_2 \neq 0$, the function $\lambda \mapsto \|(I_d + \lambda S_2)^{-1}\overline{x}_2\|^2$, $\lambda \geq 0$, is strictly positive and strictly decreasing from $\|\overline{x}_2\|^2$ to 0. Since $0 < u_1 = \|m\|^2 < \|\overline{x}_2\|^2$, Eq. (16) has a unique solution $\lambda > 0$ which determines the minimizer $m = (I_d + \lambda S_2)^{-1}\overline{x}_2$ and, hence, $g(u_1)$ $(= M(\overline{x}_2, m, S_2))$.

Now keep in mind that both solutions λ and m are functions of u_1. Equation (16) shows that λ is real-analytic. Therefore, so are both m and g. In view of the derivative of g, we deduce from Eq. (15) $m^\top S_2^{-1}(m - \overline{x}_2) = -\lambda\|m\|^2 = -\lambda u_1$ and, therefore,

$$\begin{aligned} g(u_1) &= (m - \overline{x}_2)^\top S_2^{-1}(m - \overline{x}_2) = m^\top S_2^{-1}(m - \overline{x}_2) - \overline{x}_2^\top S_2^{-1}(m - \overline{x}_2) \\ &= -\lambda u_1 - \overline{x}_2^\top S_2^{-1}(m - \overline{x}_2). \end{aligned}$$

Hence, $g'(u_1) = -\frac{\mathrm{d}\lambda}{\mathrm{d}u_1}u_1 - \lambda - \overline{x}_2^\top S_2^{-1}\frac{\mathrm{d}m}{\mathrm{d}u_1}$. Now, $m = (I_d + \lambda S_2)^{-1}\overline{x}_2$ implies

$$\frac{\mathrm{d}m}{\mathrm{d}u_1} = -\frac{\mathrm{d}\lambda}{\mathrm{d}u_1}(I_d + \lambda S_2)^{-2}S_2\overline{x}_2$$

and

$$-\overline{x}_2^\top S_2^{-1}\frac{\mathrm{d}m}{\mathrm{d}u_1} = \frac{\mathrm{d}\lambda}{\mathrm{d}u_1}\overline{x}_2^\top S_2^{-1}(I_d + \lambda S_2)^{-2}S_2\overline{x}_2 = \frac{\mathrm{d}\lambda}{\mathrm{d}u_1}\overline{x}_2^\top(I_d + \lambda S_2)^{-2}\overline{x}_2 = \frac{\mathrm{d}\lambda}{\mathrm{d}u_1}\|m\|^2 = \frac{\mathrm{d}\lambda}{\mathrm{d}u_1}u_1.$$

We conclude $g'(u_1) = -\lambda$. From Eq. (16), it follows that λ decreases as u_1 increases. Hence g is strictly convex. □

7.3 Proof of theorem 3.2. By Bolzano–Weierstrass, the sequence of minimizing vertices of the cutting plane algorithm has a cluster point in the compact area described by the vertices B, 0, and A in Figure 1. The construction of the polygons excludes every point u below the graph of g as a cluster point of the minimizers on the vertices. Hence all cluster points lie on the graph of g. Denoting the minimum of h there by h^* and by $(\widehat{u}_{t_k})$, a subsequence of minimizing vertices converging to a cluster point, we estimate

$$h^* \leq h(u_1, u_2) = h(\lim_k \widehat{u}_{t_k}) = \lim_k h(\widehat{u}_{t_k}) \leq h^*.$$

The last inequality follows from the fact that the concave function h assumes its minimum on the convex set described by the polygon at a vertex. Hence any cluster point (u_1, u_2) of minimizing vertices is minimal on g and we have also proved part (b). □