Appeared in JMVA 117 (2013) 14-31. DOI 10.1016/j.jmva.2013.01.013

Strong consistency of k-parameters clustering

María Teresa Gallegos and Gunter Ritter
Faculty of Informatics and Mathematics, University of Passau, D-94030 Passau, Germany
ritter@fim.uni-passau.de

Abstract: Pollard showed for k-means clustering and a very broad class of sampling distributions that the optimal cluster means converge to the solution of the related population criterion as the size of the data set increases. We extend this consistency result to k-parameters clustering, a method derived from the *heteroscedastic, elliptical* classification model. It allows a more sensitive data analysis and has the advantage of being affine equivariant. Moreover, the present theory yields a consistent criterion for selecting the number of clusters in such models.

Keywords: Cluster analysis, Classification models, Elliptical models Maximum likelihood estimation, Strong consistency

1 Introduction

The k–means algorithm, Steinhaus [32], enjoys great popularity in data analysis, knowledge discovery, and vector quantization in order to partition a data set $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^d$ in a given number $g \geq 2$ of clusters. Let $x_1, \dots, x_n$ be the data set to be clustered, let ℓ_i be the label of the cluster of data point x_i, $1 \leq i \leq n$, and write $\boldsymbol{\ell} = (\ell_1, \dots, \ell_n)$. If $\overline{x}_j(\boldsymbol{\ell})$ stands for the cluster mean and $W_j(\boldsymbol{\ell}) = \sum_{i:\ell_i=j}(x_i - \overline{x}_j(\boldsymbol{\ell}))(x_i - \overline{x}_j(\boldsymbol{\ell}))^\top$ for the SSP matrix ("sum of squares and products") of cluster j w.r.t. $\boldsymbol{\ell}$, this algorithm computes the *Pooled Trace* (or *Ward's*) *criterion*

$$\underset{\boldsymbol{\ell}}{\operatorname{argmin}} \operatorname{tr} \sum_{j=1}^{g} W_j(\boldsymbol{\ell}) = \underset{\boldsymbol{\ell}}{\operatorname{argmin}} \sum_{j=1}^{g} \sum_{\ell_i=j} \|x_i - \overline{x}_j(\boldsymbol{\ell})\|^2.$$

The partition created by it tends to produce spherical clusters of about equal size and about equal scatter, if the data set allows this. In fact, Bock [5] revealed the criterion as the ML estimator of a homoscedastic, isobaric, normal clustering model with spherical covariance matrices, see also Bock [6]. Properties of estimator and algorithm are well known. In particular, MacQueen [25] showed that the k-means algorithm reduces the criterion (and coined its name). Bryant and Williamson [8] studied the asymptotic behavior of a class of classification ML estimators and applied their result to a univariate, homoscedastic mixture of normal populations. Pollard [29, 30] proved for a very broad class of sampling distributions and *homoscedastic, isobaric, spherical* statistical models that the optimal means converge as the size of the data set increases. He also identified the limit as the solution to the related population criterion. This means that the *global* maximum is the favorite solution if the data set is large. His result is remarkable inasmuch as the sampling distribution may be very general and very different from the model. This property is, however, not specific to the classification model. In fact, White [35] proved consistency of ML estimators for independent observations coming from an unspecified parent distribution.

In vector quantization, application of Ward's criterion and the k-means algorithm, here ascribed to Lloyd [24], is justified and standard. The engineer using optimal quantization takes a geometric standpoint and decomposes a data set in subsets that unite nearby points and separate distant ones w.r.t. some *given* metric. Their application is less justified in cluster analysis where we search for causes that generate the data. To this end, we assume that the causes manifest themselves in different populations that induce in the data set *cohesive* (compact) clusters of possibly different sizes and extents and *separated* by location or sometimes by scale. The engineer even decomposes a data set uniform on a square, the cluster analyst finds that this data set bears no cluster structure, it originates from a single source. It is a matter between quantity and quality. Generally accepted, logical, mathematical definitions of the concepts of "cohesion" and "separation" based on the data set do not exist although there are validation methods and tests that are useful in this respect. This is contrary to mathematical topology where the analogous notion of a "connected component" is clearly and logically defined. Both concepts appeal also to intuition.

In the event of elliptical, non–spherical clusters, Ward's criterion (and, hence, the k-means algorithm) may lead to a result unacceptable in cluster analysis. A typical example is presented in Figure 1. The two-dimensional data set was sampled from two normal populations of equal scales elongated in horizontal direction and lying side by side. Up to small probability, the populations are separated by a horizontal line between them, but Ward's criterion traces a separator perpendicular to it as shown in Figure 1. The clusters obtained are neither isolated nor cohesive as visual inspection shows. Moreover, the solution created by Ward's criterion is not only inappropriate in the above sense, it also does not reflect the partition induced by the two original populations. The reason for the failure of Ward's criterion in the sense of cluster analysis is that the underlying populations are not spherical. Generally, an inappropriate, narrow model may "create" a wrong structure in the data set. It is therefore important to base cluster criteria on more general statistical models.

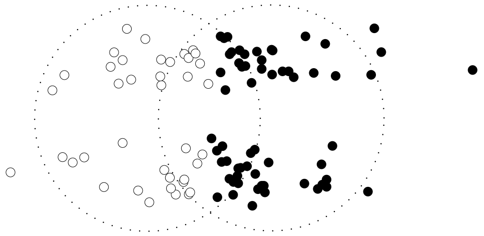

Figure 1: Example of a partition obtained from improper use of Ward's criterion. The estimated means and scales are indicated by the two circles.

Such a model was proposed by Scott and Symons [31]. They used the likelihood paradigm to derive a criterion for the heteroscedastic, isobaric, normal clustering model with arbitrary covariance matrices applicable to such a more general situation, the *heteroscedastic ML Determinant criterion*

$$\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log \det S_j(\boldsymbol{\ell}) \tag{1}$$

to be minimized w.r.t. $\boldsymbol{\ell}$. Here, $n_j(\boldsymbol{\ell})$ denotes the size of cluster j w.r.t. $\boldsymbol{\ell}$ and $S_j(\boldsymbol{\ell}) = W_j(\boldsymbol{\ell})/n_j$

is its scatter matrix.

This criterion works nicely in the case of elliptical clusters of about equal sizes but may otherwise run into trouble. Symons [33] corrected this shortcoming in considering the labeling $\boldsymbol{\ell}$ not as a parameter (which it is not since its length increases with the data set) but as a *hidden variable* drawn from the n-fold product $\boldsymbol{\pi}\otimes\cdots\otimes\boldsymbol{\pi}$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)$, on $(1\,..\,g)^{(1\,..\,n)}$. It acts as a prior probability and, since the number g is small, can be estimated by an empirical Bayesian procedure. Symons arrives at the *heteroscedastic MAP Determinant criterion*

$$n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n}, \dots, \tfrac{n_g(\boldsymbol{\ell})}{n}\big) + \tfrac{1}{2}\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \log\det S_j(\boldsymbol{\ell}), \tag{2}$$

again to be minimized w.r.t. $\boldsymbol{\ell}$. The criterion differs from the heteroscedastic ML Determinant criterion (1) in the entropy $\mathrm{H}(p_1, \dots, p_g) = -\sum_j p_j \log p_j$ of the cluster proportions $p_j = n_j(\boldsymbol{\ell})/n$ which counteracts the tendency of the ML criterion to create clusters of about equal sizes. This is the state of the art concerning the normal classification model of clustering. The related iterative relocation algorithm alternates between clustering and parameter estimation. We call it the *k-parameters* algorithm. This name reminds of the k-means algorithm but expresses the fact that it is not only means that are estimated but also other parameters such as scale matrices and weights. The criterion can be extended to elliptical basic models $E_{\phi,m,V}(x) = c_\phi\sqrt{\det V^{-1}} e^{-\phi((x-m)^{\mathrm{T}}V^{-1}(x-m))}$ with mean m, scale matrix V, and a normalizing factor c_ϕ that depends only on the fixed *radial function* ϕ. The conditional density becomes

$$f(\boldsymbol{\ell}, \mathbf{x} \mid \boldsymbol{\pi}, \mathbf{m}, \mathbf{V}) = \prod_{i=1}^{n} \pi_{\ell_i} E_{\phi, m_{\ell_i}, V_{\ell_i}}(x_i) \tag{3}$$

and its partial maximum w.r.t. $\boldsymbol{\pi}$, $\mathbf{m} = (m_1, \dots, m_g)$, and $\mathbf{V} = (V_1, \dots, V_g)$ for fixed $\boldsymbol{\ell}$ yields the heteroscedastic *Elliptical MAP* criterion

$$\begin{aligned} &n\mathrm{H}\big(\tfrac{n_1(\boldsymbol{\ell})}{n}, \dots, \tfrac{n_g(\boldsymbol{\ell})}{n}\big) \\ &+\sum_{j=1}^{g}\Big\{\tfrac{n_j(\boldsymbol{\ell})}{2}\log\det V_j(\boldsymbol{\ell}) + \sum_{\ell_i=j}\phi\big((x_i - m_j(\boldsymbol{\ell}))^{\mathrm{T}} V_j(\boldsymbol{\ell})^{-1}(x_i - m_j(\boldsymbol{\ell}))\big)\Big\} \end{aligned} \tag{4}$$

to be minimized w.r.t. $\boldsymbol{\ell}$. Here, $m_j(\boldsymbol{\ell})$ and $V_j(\boldsymbol{\ell})$ are the ML estimates of mean and scale matrix of cluster j w.r.t. $\boldsymbol{\ell}$ and H stands again for the entropy. The heteroscedastic MAP Determinant criterion (2) is easily retrieved for $\phi(t) = t/2$ up to the constant $nd/2$.

Some care is, however, needed with the "minima" of these criteria. Strictly speaking, they do not exist. First we have to assume that the data are in general position since, otherwise, some determinant may vanish. It still vanishes when some cluster has a deficient size $\leq d$. A remedy would here be to use only those labelings that generate clusters of size $\geq d+1$, see Gallegos and Ritter [13]. Some authors claim that it is the largest "local" maximum that should be considered the solution. But this is often spurious. We will follow here another way that was extensively used by Hathaway [16] in the context of mixture models and by Gallegos and Ritter [11, 12] with mixture and classification models: HDBT constraints. This acronym was coined by us to acknowledge the creation of the concept by Thompson, Beale, Dennis, and Hathaway, in this chronological order. If we denote the Löwner (or semi–definite) ordering on the set of symmetric matrices by $\preceq$ then a set or tuple of positive definite matrices V_j satisfies the HDBT constraints with HDBT constant $0 < c \leq 1$ if

$$V_j \succeq cV_\ell \tag{5}$$

for all indices j, ℓ. In the univariate case, $\succeq$ just means the usual ordering of real numbers. The constraints are affine equivariant and mean that the matrices V_j are not too different. They generalize homoscedasticity, i.e., equality of all scale matrices, which they contain as the special case $c = 1$. The smaller c is, the more different the V_j's may be. Let $\mathcal{V}_c$ denote the set of all finite collections of positive definite matrices that satisfy the HDBT constraints with constant c. Gallegos and Ritter [12] showed that the constraints guarantee a minimum $\boldsymbol{\ell} \in (1\,..\,g)^{(1\,..\,n)}$ in (2) and (4) if $n \geq gd+1$ without having to restrict cluster sizes. In combination with trimming, they lead to a robust criterion with a positive asymptotic breakdown point and they are the basis for the method of "balanced scales" that offers some guidance towards the estimation of a credible solution.

Bryant [7] proves consistency and asymptotic normality in a unified framework for some mixture and classification models. He assumes that the parameter space is a closed subspace of some Euclidean space excluding interesting examples. Cuesta–Albertos et al. [9] and García–Escudero et al. [14] investigate consistency of a likelihood based estimator for a normal clustering model with trimming. They use constraints on the scale structure that are more restrictive than the HDBT constraints since they do not only restrict the deviation from homoscedasticity but also that from sphericity. As a consequence the estimator lacks equivariance w.r.t. variable rescaling, let alone affine equivariance.

In the present paper, we generalize Pollard's consistency to the important (affine equivariant) model (3) under HDBT constraints, that is, the heteroscedastic HDBT constrained Elliptical MAP criterion (4) in a parametric framework with variable number of clusters. We show that, for a wide class of sampling distributions and radial functions, the estimates of $\boldsymbol{\pi}$, $\mathbf{m}$, and $\mathbf{V}$ converge to the solution of the related population criterion as sample size tends to infinity, Theorem 1. Besides the means m_j, this makes it necessary to also consider the mixing rates π_j and the constrained scale matrices V_j. Finally, we apply the theory to shed some light on model selection in the framework of the classification model, Corollary 1.

2 Asymptotic behavior

Let $X = (X_i)_{i\geq 1}$ be an i.i.d. sequence of $\mathbb{R}^d$-valued random variables with common distribution μ. The most interesting parent μ's are here mixtures of unimodal distributions. Taking the partial maximum w.r.t. $\boldsymbol{\ell}$ of the logarithm of (3) (parameters $\boldsymbol{\pi}$, $\mathbf{m}$, and $\mathbf{V} \in \mathcal{V}_c$ fixed) and dividing by n we obtain

$$\tfrac{1}{n} \sum_{i=1}^{n} \max_{1\leq j\leq g} \log(\pi_j E_{\phi,m_j,V_j}(X_i)). \tag{6}$$

The question arises whether the maximum of this expression w.r.t. the three parameters is stable as n increases, that is, whether there is convergence of the optimal parameters $\boldsymbol{\pi}$, $\mathbf{m}$, and $\mathbf{V} \in \mathcal{V}_c$ and of the whole expression as $n \to \infty$. In the consistency theorems for the mixture model, see Peters and Walker [28] and Kiefer [21], the limit is the parent distribution μ of the data if it is a member of the given model. This cannot be expected here, see Remark 1(d) below. But if we tentatively assume that the limit w.r.t. n and the maximum w.r.t. $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$ in (6) commute, then the strong law suggests that the limit might be

$$\max_{\boldsymbol{\pi},\mathbf{m},\mathbf{V}} \mathrm{E} \max_{1\le j\le g} \log\big(\pi_j E_{\phi,m_j,V_j}(X_1)\big).$$

To confirm this, Pollard [29] proposed to use *subsets* $A \subseteq \mathbb{R}^d$ of size $\le g$ instead of g-tuples $(a_1,\dots,a_g) \in \mathbb{R}^{gd}$ of parameters (means, in his case). The subsets take account of the unordered structure of mixture components. They also enable easy transition between different numbers of components. It turns out that Pollard's idea is favorable in the present context of more general basic models and mixing rates, too. Thus, let Θ be the collection of all nonempty, compact subsets of the Cartesian product $]0,1] \times \mathbb{R}^d \times \mathrm{PD}(d)$ endowed with the Hausdorff metric. Denote the three entries of $a \in]0,1] \times \mathbb{R}^d \times \mathrm{PD}(d)$ by π_a, m_a, and V_a, respectively. For an integer $g \ge 1$, define the subset

$$\Theta_{\le g,c} = \Big\{A \in \Theta \mid 1 \le |A| \le g,\ \sum_{a\in A} \pi_a \le 1,\ (V_a)_{a\in A} \in \mathcal{V}_c\Big\} \subseteq \Theta.$$

The relation "≤ 1" (instead of "$= 1$") makes it an extension of the (HDBT constrained) solution space. The reader may wonder why the definition of $\Theta_{\le g,c}$ requires the sum to be ≤ 1 and not $= 1$. The reason is that the modification of the space $\Theta_{\le g,c}$ defined with $\sum_{a\in A}\pi_a = 1$ lacks the important compactness property that will be described in Lemma 1(b). To see this, consider for simplicity a pure location model (no scales V). Let $a_m = (1/2, m)$, $m \in \mathbb{R}$. For $m \ne 0$, the set $A_m = \{a_0, a_m\} \in \Theta_{\le 2,c}$ has the property $\pi_{a_0} + \pi_{a_m} = 1$ and converges w.r.t. the Hausdorff metric to the singleton $A_0 = \{a_0\}$ as $m \to 0$. But $\pi_{a_0} = 1/2$ and so the limit is not a member of the modified space. To counteract this defect we complete the parameter space for deficient mixing rates.

Now, define the functional

$$t_a(x) = -\log \pi_a + \tfrac{1}{2}\log\det V_a + \phi\big((x-m_a)^{\mathrm{T}} V_a^{-1}(x-m_a)\big),$$

and denote the *sampling criterion* by

$$\Phi_n(A) = \tfrac{1}{n}\sum_{i=1}^{n} \min_{a\in A} t_a(X_i), \quad A \in \Theta_{\le g,c}. \tag{7}$$

Up to the opposite sign this is the criterion introduced in (6). With the empirical probability $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ it has the representation $\Phi_n(A) = \int \min_{a\in A} t_a(x)\mu_n(\,\mathrm{d}x)$. Its constrained minimum w.r.t. the parameters A is the same as that of the heteroscedastic *Elliptical MAP* criterion (4) w.r.t. $\boldsymbol{\ell}$. We also need the related *population criterion*

$$\Phi(A) = \mathrm{E}\min_{a\in A} t_a(X_1), \quad A \in \Theta_{\le g,c}. \tag{8}$$

It is well defined when $\log E_{\phi,m,V}(X_1)$ is integrable for all (m,V). Although the definition of $\Theta_{\le g,c}$ allows deficient mixing rates, it is reassuring to observe that optimal solutions A^* for Φ_n and Φ do enjoy the property $\sum_{a\in A^*}\pi_a = 1$. Moreover, each parameter m appears only once in A^*. The reason is the strict increase of the logarithm: As a typical example, let $A =$

$\{(\pi_{11}, m_1), (\pi_{12}, m_1), (\pi_2, m_2)\}$ with $\pi_{11} \neq \pi_{12}$ and $p = \pi_{11} + \pi_{12} + \pi_2 \leq 1$. The modification $A' = \{(\pi'_1, m_1), (\pi'_2, m_2)\}$ with $\pi'_1 = (\pi_{11} + \pi_{12})/p$ and $\pi'_2 = \pi_2/p$ satisfies $\pi'_1 + \pi'_2 = 1$. Moreover, $\log \pi'_1 > \log \pi_{11} \vee \log \pi_{12}$ and $\log \pi'_2 \geq \log \pi_2$ and, hence, $\Phi_n(A') < \Phi_n(A)$ and $\Phi(A') < \Phi(A)$:

$$\begin{aligned}
&\min_{a' \in A'} t_{a'} = \min_{a' \in A'} (-\log \pi_{a'} - \log f_{m_{a'}}) \\
&= \big(-\log \pi'_1 - \log f_{m_1} \big) \wedge \big(-\log \pi'_2 - \log f_{m_2} \big) \\
&< \big(-\log \pi_{11} - \log f_{m_1} \big) \wedge \big(-\log \pi_{12} - \log f_{m_1} \big) \wedge \big(-\log \pi_2 - \log f_{m_2} \big) \\
&= \min_{a \in A} t_a.
\end{aligned}$$

If the scale matrices satisfy the HDBT constraints, if ϕ is large enough, and if the data are in general position and $n \geq gd + 1$, then the sampling criterion (7), Φ_n, has a minimum on $\Theta_{\leq g,c}$, that is, under HDBT constraints.

The remainder of this section is devoted to the question under which conditions the constrained minimizers $(\boldsymbol{\pi}, \mathbf{m}, \mathbf{V})$ of the sampling criterion Φ_n converge as $n \to \infty$. The related population criterion (8), Φ, is well defined when $\phi(\beta \|X_1\|^2)$ is integrable for all $\beta > 0$. The following lemma collects some properties of the quantities just defined.

Lemma 1 *(a) The set $\Theta_{\leq g,c}$ is closed in Θ w.r.t. the Hausdorff metric.*

(b) Let $0 < \pi_0 \leq 1$ and $R, \varepsilon > 0$. The sub-collection of $\Theta_{\leq g,c}$ consisting of all sets of elements a such that $\pi_a \geq \pi_0$, $\|m_a\| \leq R$, and $\varepsilon I_d \preceq V_a \preceq \frac{1}{\varepsilon} I_d$ is compact.

(c) If ϕ is continuous then the function $A \mapsto \min_{a \in A} t_a(x)$ is continuous on $\Theta_{\leq g,c}$ for all $x \in \mathbb{R}^d$.

The proof of the following lemma needs the uniform strong law of large numbers (USLLN), see for instance Bierens [4]: If $g : T \times K \to \mathbb{R}^d$ is a function defined on the product of a measurable space T and a compact space K such that (i) $g(x, \cdot)$ is continuous for all $x \in T$; (ii) $g(\cdot, \vartheta)$ is measurable for all $\vartheta \in K$; (iii) $\|g(x, \vartheta)\| \leq h(x)$ for all $\vartheta \in K$ and all $x \in T$ with some measurable, μ-integrable function h on T, and if $Z_1, Z_2, \ldots$ is an i.i.d. sequence of random variables on (Ω, P) with values in T and a common distribution μ, then

(a) P-a.s., we have $\frac{1}{n} \sum_{i=1}^n g(Z_i, \vartheta) \underset{n\to\infty}{\longrightarrow} \mathrm{E} g(Z_1, \vartheta)$ uniformly for all $\vartheta \in K$.

(b) In particular, $\mathrm{E} g(Z_1, \cdot) = \int_T g(x, \cdot) \mu(\mathrm{d}x)$ is continuous.

We will often replace V^{-1} with Λ.

Lemma 2 *Assume that*

(i) ϕ is increasing and continuous;

(ii) $\phi\big(\beta \|X_1\|^2\big)$ is integrable for all $\beta \geq 1$.

Then, P-a.s., the sampling criterion (7), Φ_n, converges to the population criterion (8), Φ, locally uniformly on $\Theta_{\leq g,c}$. In particular, Φ is continuous.

Proof 1 *We verify the assumptions of the USLLN with* $g(x,A)=\min_{a\in A}t_a(x)$*. Continuity (i) is just Lemma 1*(c)*. In order to obtain the integrable upper bound (iii) it is sufficient to consider elements* a *in the compact set* $K_{\varepsilon,R}=[\varepsilon,1]\times\overline{B}_R(0)\times\{V\in PD(d)\mid \varepsilon I_d\preceq V\preceq \frac{1}{\varepsilon}I_d\}$ *for* $0<\varepsilon\le 1$ *and* $R>0$*. Since* ϕ *is increasing, we estimate for* $\|x\|\ge R$

$$\begin{aligned}&t_a(x)=-\log\pi_a-\tfrac12\log\det\Lambda_a+\phi\big((x-m_a)^{\mathrm T}\Lambda_a(x-m_a)\big)\\&\le-\log\varepsilon-\tfrac{d}{2}\log\varepsilon+\phi\big(\|x-m_a\|^2/\varepsilon\big)\\&\le\text{const}+\phi\Big(\tfrac{2}{\varepsilon}\big(\|x\|^2+\|m_a\|^2\big)\Big)\le\text{const}+\phi\Big(\tfrac{4}{\varepsilon}\|x\|^2\Big).\end{aligned}$$

By assumption, the right–hand side is the requested μ*-integrable upper bound.* □

Assumption (ii) of Lemma 2 requires that the density generator $\varphi=e^{-\phi}$ should have a heavy tail if μ does. It is satisfied if the radial function ϕ grows at most polynomially and $\phi\big(\|X_1\|^2\big)$ is integrable. The main idea of the consistency proof is to show that all optimal sets remain in a compact subset of $\Theta_{\le g,c}$ independent of n. The following lemma is a first step towards this goal and prepares Lemma 4 which shows that the optimal scale parameters are bounded.

Lemma 3 *Let* $\alpha>0$ *and assume that* $\|X_1\|^{2\alpha}$ *is integrable. Then the minimum of*

$$\mathrm{E}\min_{a\in A}\big|v^{\mathrm T}(X_1-m_a)\big|^{2\alpha}$$

w.r.t. v*,* $\|v\|=1$*, and* $A\in\Theta_{\le g,c}$ *exists. It is zero if and only if* μ *is supported by* g *parallel hyperplanes.*

Proof 2 *Let* $\Theta'_{\le g}$ *stand for the collection of all nonempty subsets of* $\mathbb{R}$ *with at most* g *elements and let* μ_v *be the projection of* μ *to the real line via* $x\mapsto v^{\mathrm T}x$*. We prove that the minimum of the finite functional*

$$J(v,B)=\int_{\mathbb{R}}\min_{b\in B}|y-b|^{2\alpha}\mu_v(\,\mathrm{d}y)$$

w.r.t. v*,* $\|v\|=1$*, and* $B\in\Theta'_{\le g}$ *exists. This is equivalent to the claim. Indeed, if* B *is related to* A *by* $b=v^{\mathrm T}m_a$ *or* $m_a=bv$ *then*

$$\int_{\mathbb{R}}\min_{b\in B}|y-b|^{2\alpha}\mu_v(\,\mathrm{d}y)=\mathrm{E}\min_{a\in A}\big|v^{\mathrm T}(X_1-m_a)\big|^{2\alpha}.$$

The proof proceeds in three steps. In step (α*), it is shown that sets* B *that contain only remote points cannot have small values of the criterion. In the main step (*β*), the same is shown for a mixture of nearby and remote points. Thus, sets of nearby points remain to consider. The lemma follows in a third step. Let* $B_r(0)=\{x\in\mathbb{R}^d\mid\|x\|\le r\}$ *be the centered unit ball of radius* r *and let* $r>0$ *be such that* $\mu(B_r(0))>0$*.*

(α*) If* $R>r$ *satisfies*

$$|R-r|^{2\alpha}\mu(B_r(0))\ge\int_{\mathbb{R}^d}\|x\|^{2\alpha}\mu(\,\mathrm{d}x)\tag{9}$$

then no set $B \in \Theta'_{\le g}$ contained in the complement of the interval $[-R, R]$ can be optimal for J:
Note $\mu_v(]-r,r[) \ge \mu(B_r(0))$ and use the Cauchy–Schwarz inequality to estimate

$$\begin{aligned}
&J(v,B) \ge \int_{]-r,r[} \min_{b\in B} |y-b|^{2\alpha}\mu_v(\,\mathrm{d}y) > |R-r|^{2\alpha}\mu_v(]-r,r[)\\
&\ge |R-r|^{2\alpha}\mu(B_r(0)) \ge \int_{\mathbb{R}^d} \|x\|^{2\alpha}\mu(\,\mathrm{d}x)\\
&\ge \int_{\mathbb{R}^d} |w^{\mathrm{T}}x|^{2\alpha}\mu(\,\mathrm{d}x) = J(w,\{0\}), \quad \|w\|=1.
\end{aligned}$$

The rest of the proof proceeds by induction over g. The claim for $g = 1$ immediately follows from (α). Let $g \ge 2$ and assume that the claim is true for $g-1$. We distinguish between two cases. If $\inf_{v,B\in\Theta'_{\le g}} J(v,B) = \min_{v,B\in\Theta'_{\le g-1}} J(v,B)$ then nothing has to be proved since the largest lower bounds decrease with g. In the opposite case, let

$$\varepsilon = \min_{v,B\in\Theta'_{\le g-1}} J(v,B) - \inf_{v,B\in\Theta'_{\le g}} J(v,B) \quad (>0)$$

and let $R > 0$ satisfy

$$\int_{\|x\|\ge 2R} \|2x\|^{2\alpha}\mu(\,\mathrm{d}x) < \varepsilon/2. \tag{10}$$

(β) If B contains elements in both $[-R,R]$ and the complement of $[-5R,5R]$ then its J-value cannot be arbitrarily close to the largest lower bound:
We argue by contradiction. Let $S_1(0) = \{x \in \mathbb{R}^d \mid \|x\| = 1\}$ be the centered unit sphere in $\mathbb{R}^d$ and assume that there is $(v,B) \in S_1(0) \times \Theta'_{\le g}$ such that

(i) B contains a point b_-, $|b_-| \le R$, and a point b_+, $|b_+| > 5R$, and

(ii) $J(v,B) - \inf_{\|v'\|=1,B'\in\Theta'_{\le g}} J(v',B') \le \varepsilon/2$.

We compare the J-value of B with that of its subset $\widetilde{B} = B \cap [-5R,5R] \in \Theta'_{\le g-1}$. As a consequence of removing points from B, some set C of points have to be reassigned. If $|y| < 2R$ then $|y-b_-| < 3R < |y-b_+|$. Thus, the deletion of points in the complement of $[-5R,5R]$ from B does not affect the assignment of points $|y| < 2R$, that is $C \subseteq \{y \in \mathbb{R} \mid |y| \ge 2R\}$. If $|y| \ge 2R$ then $|y-b_-| \le |2y|$. It follows

$$\begin{aligned}
&J(v,\widetilde{B}) - J(v,B) = \int_C \min_{b\in\widetilde{B}} |y-b|^{2\alpha}\mu_v(\,\mathrm{d}y) - \int_C \min_{b\in B} |y-b|^{2\alpha}\mu_v(\,\mathrm{d}y)\\
&\le \int_C \min_{b\in\widetilde{B}} |y-b|^{2\alpha}\mu_v(\,\mathrm{d}y) \le \int_{|y|\ge 2R} \min_{b\in\widetilde{B}} |y-b|^{2\alpha}\mu_v(\,\mathrm{d}y)\\
&\le \int_{|y|\ge 2R} |y-b_-|^{2\alpha}\mu_v(\,\mathrm{d}y) \le \int_{|y|\ge 2R} |2y|^{2\alpha}\mu_v(\,\mathrm{d}y)\\
&\le \int_{\|x\|\ge 2R} \|2x\|^{2\alpha}\mu(\,\mathrm{d}x) < \varepsilon/2.
\end{aligned}$$

Recalling $\widetilde{B} \in \Theta'_{\le g-1}$, we have found a contradiction to the definition of ε and (ii) above. This proves (β).

Let R satisfy (9) *and* (10). *Assertions (α) and (β) show that small values of J are attained on $\Theta'_{\le g}$ only if $B \subseteq [-5R, 5R]$. The collection $\mathcal{K}$ of these sets B is compact for the Hausdorff metric. The first claim finally follows from continuity of J on the compact space $S_1(0) \times \mathcal{K}$.*

The expectation is zero if and only if the function $\min_{a\in A} v^{\mathrm{T}}(\,\boldsymbol{\cdot} - m_a)$ vanishes on the support of μ for some pair (v, A). This means that the support of μ is contained in some set of the form $\bigcup_{a\in A}\{x \mid v^{\mathrm{T}}(x - m_a) = 0\}$. □

Lemma 4 *Assume that*

(i) μ is not supported by g parallel hyperplanes;

(ii) $\phi(t) \ge b_0 + b_1 t^\alpha$ for some numbers $b_0 \in \mathbb{R},\ b_1, \alpha > 0$;

(iii) $\phi(\|X_1\|^2)$ is integrable.

Then there are numbers $0 < c_1 \le c_2$ such that, P-a.s., all scale parameters V^* of sets optimal for Φ_n on $\Theta_{\le g,c}$, see (7), *satisfy $I_d/c_2 \preceq V^* \preceq I_d/c_1$ for eventually all n.*

Proof 3 *First note that (iii) along with (ii) implies integrability of $\|X_1\|^{2\alpha}$. Let us show that $\Phi_n(A)$ is large for eventually all n if an eigenvalue of the scale matrices of A is small or large. Let $A \in \Theta_{\le g,c}$, and let $a_0 \in A$ such that $\det \Lambda_{a_0}$ is maximal. Using the SLLN, we estimate*

$$\begin{aligned}
&\Phi_n(A) = \tfrac{1}{n}\sum_{i=1}^n \min_{a\in A} t_a(X_i)\\
&= \tfrac{1}{n}\sum_{i=1}^n \min_{a\in A}\big\{-\log \pi_a - \tfrac12 \log\det\Lambda_a + \phi\big((X_i - m_a)^{\mathrm{T}}\Lambda_a(X_i - m_a)\big)\big\}\\
&\ge b_0 - \tfrac12\log\det\Lambda_{a_0} + \tfrac{b_1}{n}\sum_{i=1}^n \min_{a\in A}\big(c(X_i - m_a)^{\mathrm{T}}\Lambda_{a_0}(X_i - m_a)\big)^\alpha\\
&\underset{n\to\infty}{\longrightarrow} b_0 - \tfrac12\log\det\Lambda_{a_0} + b_1 c^\alpha \mathrm{E}\min_{a\in A}\big((X_1 - m_a)^{\mathrm{T}}\Lambda_{a_0}(X_1 - m_a)\big)^\alpha, \quad P\text{-a.s.}.
\end{aligned}$$

Let $\Lambda_{a_0} = \sum_k \lambda_k v_k v_k^{\mathrm{T}}$ be the spectral decomposition. Since $t \mapsto t^\alpha$ is increasing and multiplicative,

$$\begin{aligned}
&\big((x - m_a)^{\mathrm{T}}\Lambda_{a_0}(x - m_a)\big)^\alpha = \Big(\sum_k \lambda_k\big(v_k^{\mathrm{T}}(x - m_a)\big)^2\Big)^\alpha\\
&\ge \tfrac1d\sum_k\Big(\lambda_k\big(v_k^{\mathrm{T}}(x - m_a)\big)^2\Big)^\alpha = \tfrac1d\sum_k \lambda_k^\alpha\big|v_k^{\mathrm{T}}(x - m_a)\big|^{2\alpha}.
\end{aligned}$$

Inserting, we find

$$\begin{aligned}
&\lim_n \Phi_n(A) \ge b_0 + \sum_k\Big\{-\tfrac12\log\lambda_k + \frac{b_1c^\alpha}{d}\lambda_k^\alpha\,\mathrm{E}\min_{a\in A}\big|v_k^{\mathrm{T}}(X_1 - m_a)\big|^{2\alpha}\Big\}\\
&\ge b_0 + \sum_k\Big\{-\tfrac12\log\lambda_k + \kappa\frac{b_1c^\alpha}{d}\lambda_k^\alpha\Big\} \qquad (11)
\end{aligned}$$

with the constant $\kappa = \min_{v \in S_1(0), A} \mathrm{E}\min_{a\in A} \left|v^{\mathrm{T}}(X_1 - m_a)\right|^{2\alpha} > 0$, *see Lemma 3. Note that the* k*th summand in* (11) *converges to* ∞ *as* $\lambda_k \to 0$ *and, since* λ_k^α *beats* $\log \lambda_k$, *also as* $\lambda_k \to \infty$. *It has, thus, a minimum for* $\lambda_k > 0$. *Therefore, if one summand converges to* ∞ *then so does the whole sum. But one summand tends to* ∞ *under any of the two cases: (1) the largest of all eigenvalues* λ_k *exceeds a certain value* c_2 *and (2)* $\lambda_k < c_2$ *for all* k *and the smallest eigenvalue drops below another value* c_1. *We conclude that*

$$\Phi_n(A) \geq \Phi(\{(1, 0, I_d)\}) + 1 \tag{12}$$

for eventually all n *if (1) or (2) is satisfied. Finally, a set* A_n^* *minimizing* Φ_n *satisfies*

$$\Phi_n(A_n^*) \leq \Phi_n(\{(1, 0, I_d)\}) \underset{n\to\infty}{\longrightarrow} \Phi(\{(1, 0, I_d)\})$$

by (iii) and the SLLN. Combined with (12), *this is the claim.* □

For $A \in \Theta_{\leq g,c}$ and $a \in A$, let $C_a(A)$ be the set $\{x \in \mathbb{R}^d \mid t_a(x) \leq t_b(x) \text{ for all } b \in A\}$. In order to have disjoint, measurable sets, ties are broken by ordering A and favoring the smallest a. The decomposition $\{C_a(A) \mid a \in A\}$ of $\mathbb{R}^d$ in at most g measurable "cells" is defined by the Bayes discriminant rule. The population criterion $\Phi(A)$ has the representation

$$\Phi(A) = \int_{\mathbb{R}^d} \min_{a \in A} t_a \,\mathrm{d}\mu = \sum_{a\in A} \int_{C_a(A)} t_a \,\mathrm{d}\mu.$$

The entropy inequality shows

$$\begin{aligned}\Phi(A) = &\sum_{a \in A} \Big\{ -\mu(C_a(A)) \log \pi_a - \frac{\mu(C_a(A))}{2} \log\det\Lambda_a \\ &\qquad + \int_{C_a(A)} \phi\big((x-m_a)^{\mathrm{T}}\Lambda_a(x-m_a)\big)\mu(\,\mathrm{d}x)\Big\} \\ \geq\ &\mathrm{H}(\mu(C_a(A)) \mid a \in A) + \sum_{a\in A} \Big\{ -\frac{\mu(C_a(A))}{2} \log\det\Lambda_a \\ &\qquad\qquad + \int_{C_a(A)} \phi\big((x-m_a)^{\mathrm{T}}\Lambda_a(x-m_a)\big)\mu(\,\mathrm{d}x)\Big\}.\end{aligned} \tag{13}$$

In particular, an optimal set A^* satisfies $\pi_a = \mu(C_a(A^*))$ for all $a \in A^*$. Moreover, equality obtains in (13) if and only if $\pi_a = \mu(C_a(A^*))$ for all $a \in A^*$. The following lemma states conditions ensuring that optimal means and mixing rates are bounded. In view of its part (b) note that, under the assumptions of Lemma 4, $\liminf_n \min_{A\in\theta_{\leq g,c}} \Phi_n(A) > -\infty$ and $\limsup_n \min_{A\in\theta_{\leq g,c}} \Phi_n(A) < \infty$ for eventually all n. Indeed, $\min_{A\in\theta_{\leq g,c}} \Phi_n(A) \geq -\frac{d}{2}\log c_2 + b_0$ and $\min_{A\in\theta_{\leq g,c}} \Phi_n(A) \leq \Phi_n(\{(1,0,I_d)\}) \to \Phi(\{(1,0,I_d)\})$ by the SLLN.

Lemma 5 *Assume that*

(i) the parent distribution μ *is not supported by* g *parallel hyperplanes;*

(ii) ϕ *is increasing and* $\phi(t) \geq b_0 + b_1 t^\alpha$ *for some numbers* $b_0 \in \mathbb{R},\ b_1, \alpha > 0$*;*

(iii) $\phi(\|X_1\|^2)$ *is integrable.*

Let $g \geq 1$ *and let* $A_n^*,\ n \geq gd+1$*, be optimal for* Φ_n *on* $\Theta_{\leq g,c}$*. Then,* P*-a.s., we have:*

(a) If $a_n^* \in A_n^*$ *is such that* $\pi_{a_n^*} = \mu_n\big(C_{a_n^*}(A_n^*)\big) \geq \varepsilon > 0$ *then the sequence of means* $(m_{a_n^*})_n$ *is bounded.*

(b) If $g \geq 2$ *and*

$$\limsup_n \min_{A \in \Theta_{\leq g,c}} \Phi_n(A) < \limsup_n \min_{A \in \Theta_{\leq g-1,c}} \Phi_n(A)$$

then the mixing rates $\pi_{a_n^*}$ *of* all $a_n^* \in A_n^*$*,* $n \geq gd+1$*, are bounded away from zero (and all means* $m_{a_n^*}$ *are bounded).*

Proof 4 *Let* c_1 *and* c_2 *be as in Lemma 4.*
(a) This part claims that means of sets A_n^* *optimal for* Φ_n *with lower bounded probabilities of their cells are small. Let* $r > 0$ *be so large that* $\mu(B_r(0)) > 1 - \varepsilon/2$*. Using the estimates* $c_1 I_d \preceq \Lambda_a \preceq c_2 I_d$*, see Lemma 4, we infer*

$$\begin{aligned}
&\Phi_n(\{(1,0,I_d)\}) \geq \Phi_n(A_n^*) = \int \min_{a \in A_n^*} t_a \,\mathrm{d}\mu_n = \sum_{a \in A_n^*} \int_{C_a(A_n^*)} t_a \,\mathrm{d}\mu_n \\
&\geq \sum_{a \in A_n^*} \int_{C_a(A_n^*)} \Big\{ -\tfrac{1}{2} \log\det \Lambda_a + \phi\big((x-m_a)^{\mathrm{T}} \Lambda_a (x-m_a)\big) \Big\} \mu_n(\,\mathrm{d}x) \\
&\geq -\tfrac{d}{2} \log c_2 + b_0 + b_1 \sum_{a \in A_n^*} \int_{C_a(A_n^*)} \big(c_1 \|x-m_a\|^2\big)^\alpha \mu_n(\,\mathrm{d}x) \\
&\geq -\tfrac{d}{2} \log c_2 + b_0 + b_1 c_1^\alpha \int_{C_{a_n^*}(A_n^*) \cap B_r(0)} \big(\|x - m_{a_n^*}\|^{2\alpha}\big) \mu_n(\,\mathrm{d}x) \\
&\geq -\tfrac{d}{2} \log c_2 + b_0 + b_1 c_1^\alpha \big|\, \|m_{a_n^*}\| - r\big|^{2\alpha} \mu_n\big(C_{a_n^*}(A_n^*) \cap B_r(0)\big)
\end{aligned}$$

if $\|m_{a_n^*}\| \geq r$*. By the SLLN, we have* $\mu_n(B_r(0)) > 1 - \varepsilon/2$ *for eventually all* n *and, hence,* $\mu_n\big(C_{a_n^*}(A_n^*) \cap B_r(0)\big) \geq \varepsilon/2$ *for these* n*. Therefore, an application of (iii) and the SLLN imply*

$$\begin{aligned}
&\Phi(\{(1,0,I_d)\}) = \lim_n \Phi_n(\{(1,0,I_d)\}) \\
&\geq -\tfrac{d}{2} \log c_2 + b_0 + \frac{b_1 c_1^\alpha \varepsilon}{2} \limsup_n \big|\, \|m_{a_n^*}\| - r\big|^{2\alpha}.
\end{aligned}$$

This defines a bound for all $m_{a_n^*}$*.*

(b) Since $\sum_{a \in A_n^*} \pi_a = 1$ *there is* $a_n \in A_n^*$ *such that* $\pi_{a_n} \geq 1/g$*,* $n \geq gd+1$*. Let* $R > 0$*,* $R' \geq 2R$*, and* $u' < 1/g$ *be three constants to be specified later. According to part (a) we may and do assume* $R \geq \|m_{a_n}\|$*,* $n \geq gd+1$*. Also assume that there is an element* $a' \in A_n^*$ *with the property* $\|m_{a'}\| > R'$ *or* $\pi_{a'} < u'$ *and delete all such elements from* A_n^* *to obtain a set* $\widetilde{A}_n \in \Theta_{\leq g-1,c}$*. Of course,* $a_n \in \widetilde{A}_n$*. Note that any* x *assigned to* $a \in \widetilde{A}_n$ *w.r.t.* A_n^* *is also assigned to* a *w.r.t.* $\widetilde{A}_n$*. Therefore, the sample space splits in two parts: The set* $\bigcup_{a \in \widetilde{A}_n} C_a(A_n^*)$

of points assigned to elements in $\widetilde{A}_n$ w.r.t. both A_n^* and $\widetilde{A}_n$ and the set $C = \bigcup_{a\in A_n\setminus\widetilde{A}_n} C_a(A_n^*)$ of points reassigned w.r.t. $\widetilde{A}_n$ because they were originally assigned to points deleted from A_n^*.

We first show that the centered ball with radius $2R$ is contained in the complement of C. So let $\|x\| < 2R$ and let $a' \in A_n^* \setminus \widetilde{A}_n$. We have

$$\begin{aligned} t_{a_n}(x) &= -\log\pi_{a_n} - \tfrac{1}{2}\log\det\Lambda_{a_n} + \phi\big((x-m_{a_n})^{\mathrm{T}}\Lambda_{a_n}(x-m_{a_n})\big)\\ &\le \log g - \tfrac{d}{2}\log c - \tfrac{1}{2}\log\det\Lambda_{a'} + \phi\big(c_2(\|x\|+R)^2\big)\\ &\le \log g - \tfrac{d}{2}\log c - \tfrac{1}{2}\log\det\Lambda_{a'} + \phi\big(9c_2R^2\big). \end{aligned} \tag{14}$$

Now fix u' and R' in such a way that

$$\log g - \tfrac{d}{2}\log c + \phi\big(9c_2R^2\big) < (b_0 - \log u') \wedge \phi\big(c_1(R'-2R)^2\big).$$

The element a' has one of two properties. If $\pi_{a'} < u'$ then

$$(14) < b_0 - \log u' - \tfrac{1}{2}\log\det\Lambda_{a'} \le b_0 - \log\pi_{a'} - \tfrac{1}{2}\log\det\Lambda_{a'}.$$

If $\|m_{a'}\| > R'$ then $R' - 2R \le \|m_{a'}\| - \|x\| \le \|x - m_{a'}\|$ and

$$(14) < -\tfrac{1}{2}\log\det\Lambda_{a'} + \phi\big(c_1(R'-2R)^2\big) \le -\tfrac{1}{2}\log\det\Lambda_{a'} + \phi\big(c_1\|x-m_{a'}\|^2\big).$$

Hence, in both cases, $t_{a_n}(x) < t_{a'}(x)$, that is, x is not assigned to a' and $B_{2R}(0) \subseteq \complement C$ (the complement of C) as claimed.

Observing the properties of the set C explained above, we obtain

$$\begin{aligned} \Phi_n(\widetilde{A}_n) - \Phi_n(A_n^*) &= \int_{\mathbb{R}^d}\Big(\min_{a\in\widetilde{A}_n} t_a - \min_{a\in A_n^*} t_a\Big)\,\mathrm{d}\mu_n\\ &= \int_C \Big(\min_{a\in\widetilde{A}_n} t_a - \min_{a\in A_n^*} t_a\Big)\,\mathrm{d}\mu_n \le \int_C t_{a_n}\,\mathrm{d}\mu_n - \int_C \min_{a\in A_n^*\setminus\widetilde{A}_n} t_a\,\mathrm{d}\mu_n. \end{aligned}$$

Now we have

$$\begin{aligned} t_{a_n}(x) &= -\log\pi_{a_n} - \tfrac{1}{2}\log\det\Lambda_{a_n} + \phi\big((x-m_{a_n})^{\mathrm{T}}\Lambda_{a_n}(x-m_{a_n})\big)\\ &\le \log g - \tfrac{d}{2}\log c_1 + \phi\big(c_2\|x-m_{a_n}\|^2\big) \end{aligned}$$

and $t_a(x) \ge -\frac{d}{2}\log c_2 + b_0$ for all a. Inserting and observing $C \subseteq \complement B_{2R}(0)$, we infer for all n

$$\begin{aligned} &\min_{A\in\Theta_{\le g-1,c}} \Phi_n(A) - \min_{A\in\Theta_{\le g,c}} \Phi_n(A) \le \Phi_n(\widetilde{A}_n) - \Phi_n(A_n^*)\\ &\le \int_C \Big\{\log g - \tfrac{d}{2}\log c_1 + \phi\big(c_2\|x-m_{a_n}\|^2\big)\Big\}\mu_n(\,\mathrm{d}x) + \int_C \Big\{\tfrac{d}{2}\log c_2 - b_0\Big\}\mu_n(\,\mathrm{d}x)\\ &\le \big\{\log g + \tfrac{d}{2}\log\tfrac{c_2}{c_1} - b_0\big\}\mu_n(\complement B_{2R}(0)) + \int_{\complement B_{2R}(0)} \phi\big(4c_2\|x\|^2\big)\mu_n(\,\mathrm{d}x). \end{aligned} \tag{15}$$

From $t_a(x) \ge -\frac{d}{2}\log c_2 + b_0$ and the assumption of part (b), we also obtain $\liminf_n \min_{A\in\Theta_{\le g,c}} \Phi_n(A) \in \mathbb{R}$. Passing to the $\limsup_n$ in (15) with the aid of the SLLN, *we therefore find*

$$\begin{aligned}
&\limsup_n \min_{A\in\Theta_{\le g-1,c}} \Phi_n(A) - \limsup_n \min_{A\in\Theta_{\le g,c}} \Phi_n(A)\\
&\le \limsup_n \Big(\min_{A\in\Theta_{\le g-1,c}} \Phi_n(A) - \min_{A\in\Theta_{\le g,c}} \Phi_n(A)\Big)\\
&\le \big\{\log g + \tfrac{d}{2}\log\tfrac{c_2}{c_1} - b_0\big\}\mu\big(\complement B_{2R}(0)\big) + \int_{\|x\|\ge 2R}\phi\big(4c_2\|x\|^2\big)\mu(\,\mathrm{d}x).
\end{aligned}$$

The assumption of part (b) says that the left–hand side is P–a.s. strictly positive. Since the right–hand side vanishes as $R\to\infty$, the assumption on the existence of a' made at the beginning cannot hold if R is large. This proves part (b). □

Lemma 6 *Let the assumptions of Lemma 2 be satisfied. If $\mathcal{K}\subseteq\Theta_{\le g,c}$ is compact and contains some minimizer of the sampling criterion* (7)*, Φ_n, for all $n\ge gd+1$, then*

(a) $\mathcal{K}$ contains a minimizer of the population criterion (8)*, Φ;*

(b) P-a.s., $\min\Phi_n \underset{n\to\infty}{\longrightarrow} \min\Phi$ on $\Theta_{\le g,c}$.

Proof 5 *Since Φ is continuous, the restriction $\Phi_{|\mathcal{K}}$ has a minimizer A^*. We have to show that A^* minimizes Φ on all of $\Theta_{\le g,c}$. Now, let $A_n^*\in\mathcal{K}$ be some minimizer of Φ_n. The uniform convergence $\Phi_n\to\Phi$ on $\mathcal{K}$, Lemma 2, implies*

$$\Phi(A^*)\le\Phi(A)\le\Phi_n(A)+\varepsilon,\ \ A\in\mathcal{K},$$

for eventually all n. Conversely, by optimality of A_n^*,

$$\Phi_n(A_n^*)\le\Phi_n(A^*)\le\Phi(A^*)+\varepsilon$$

if n is large. Hence, $\Phi_n(A_n^*)\to\Phi(A^*)$ as $n\to\infty$. Finally, the inequality $\Phi_n(A_n^*)\le\Phi_n(A)$ for all $A\in\Theta_{\le g,c}$ shows that A^* is a minimizer of Φ on all of $\Theta_{\le g,c}$. □

The remainder of our analysis depends on the following notion.

Definition 1 *An integer $g\ge 2$ is a* **drop point** *of the population criterion Φ (under the HDBT constant c, see* (8)*) if*

$$\inf_{A\in\Theta_{\le g,c}}\Phi(A) < \inf_{A\in\Theta_{\le g-1,c}}\Phi(A).$$

Also $g=1$ is defined as a drop point.

The number of drop points may be finite or infinite. The following lemma shows that all optima of all sampling criteria Φ_n, $n\ge gd+1$, remain in a compact set if g is a drop point. This is the basis for consistency. The lemma shows also that their minima behave reasonably.

Lemma 7 *Assume that*

(i) hyperplanes in $\mathbb{R}^d$ are μ-null sets;

(ii) ϕ is continuous and increasing;

(iii) $\phi(t) \ge b_0 + b_1 t^\alpha$ for some numbers $b_0 \in \mathbb{R},\ b_1, \alpha > 0$;

(iv) $\phi^p(\beta\|X_1\|^2)$ is integrable for some $p > 1$ and all $\beta \ge 1$.

Then, for all $g \ge 1$,

(a) P-a.s., each sampling criterion $\Phi_n,\ n \ge gd+1$, has a minimizer $A_n^* \in \Theta_{\le g,c}$;

(b) if g is a drop point then there is a compact subset of $\Theta_{\le g,c}$ that contains all minimizers A_n^* of Φ_n for all $n \ge gd+1$;

(c) the population criterion Φ has a minimizer in $\Theta_{\le g,c}$;

(d) P-a.s., $\min_{A\in\Theta_{\le g,c}} \Phi_n(A) \underset{n\to\infty}{\longrightarrow} \min_{A\in\Theta_{\le g,c}} \Phi(A)$.

Proof 6 *By assumption (i), the data are P-a.s. in general position. Therefore, claim (a) is proved in a similar way as Lemma 1 in Gallegos and Ritter [12]. For claims (b), (c), and (d) we use induction over $g \ge 1$. Let $g = 1$. By Lemmas 4, 5(a), and 1(b), there exists a compact subset $\mathcal{K}_1 \subseteq \Theta_{\le 1,c}$ that contains all minimizers of all sampling criteria Φ_n, that is claim (b) for $g = 1$. Claims (c) and (d) for $g = 1$ follow from Lemma 6.*

Now let $g \ge 2$. The following arguments need that $\inf_{A\in\Theta_{\le g,c}} \Phi(A)$ is finite. The proof is similar to that of (11)*. Indeed, let $a_0 \in A$ such that $\det \Lambda_{a_0}$ is maximal and let $\Lambda_{a_0} = \sum_k \lambda_k v_k v_k^{\mathrm{T}}$ be the spectral decomposition. We have*

$$\begin{aligned}\Phi(A) &\ge -\tfrac{1}{2}\log\det\Lambda_{a_0} + \int \min_{a\in A} \phi\big((x-m_a)^{\mathrm{T}}\Lambda_a(x-m_a)\big)\mu(\,\mathrm{d}x)\\ &\ge \tfrac{d}{2}\log c + b_0 + \sum_k \Big(-\tfrac{1}{2}\log(c\lambda_k) + \kappa\tfrac{b_1}{d}(c\lambda_k)^\alpha\Big)\end{aligned}$$

with the strictly positive constant $\kappa = \min\limits_{\|v\|=1,A} \mathrm{E}\min_{a\in A}\big|v^{\mathrm{T}}(X_1-m_a)\big|^{2\alpha}$. The claim follows from the fact that each summand on the right–hand side is bounded below as a function of λ_k.

In view of the induction step g –1 → g let A_n^* be minimal for Φ_n on $\Theta_{\le g,c}$, $n \ge gd+1$. We distinguish between two cases. First, assume $\pi_a \ge \varepsilon > 0$ for all $a \in A_n^*$ and all such n. By Lemmas 4, 5(a), and 1, there exists a compact subset $\mathcal{K}_g \subseteq \Theta_{\le g,c}$ which contains all minima A_n^*. This is one half of claim (b) and claims (c) and (d) follow again from Lemma 6.

In the second case we may and do assume that there are elements $a_n \in A_n^*$ such that $\pi_{a_n} = \mu_n(C_{a_n}(A_n^*)) \to 0$ as $n \to \infty$. Of course there is also at least one element $a_n' \in A_n^*$ such that $\mu_n(C_{a_n'}(A_n^*)) \ge 1/g$ and Lemma 5(a) implies $\|m_{a_n'}\| \le R$ for some R. By assumption (iv) *and by Hölder's inequality with* $\frac{1}{p}+\frac{1}{q}=1$*, we obtain with* c_2 *as in Lemma 4*

$$\begin{aligned}
&\int_{C_{a_n}(A_n^*)} \phi\big((x-m_{a_n'})^{\mathrm{T}}\Lambda_{a_n'}(x-m_{a_n'})\big)\mu_n(\,\mathrm{d}x)\\
&\le \int_{C_{a_n}(A_n^*)} \phi\big(c_2\|x-m_{a_n'}\|^2\big)\mu_n(\,\mathrm{d}x) \le \int_{C_{a_n}(A_n^*)} \phi\big(2c_2\big(\|x\|^2+R^2\big)\big)\mu_n(\,\mathrm{d}x)\\
&\le \mu_n\big(C_{a_n}(A_n^*)\big)^{1/q}\Big(\int |\phi|^p\big(2c_2\big(\|x\|^2+R^2\big)\big)\mu_n(\,\mathrm{d}x)\Big)^{1/p}\\
&\le \mu_n\big(C_{a_n}(A_n^*)\big)^{1/q}\Big(|\phi|^p\big(4c_2R^2\big)+\int_{\|x\|>R}|\phi|^p\big(4c_2\|x\|^2\big)\mu_n(\,\mathrm{d}x)\Big)^{1/p} \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Since $\mu_n\big(C_{a_n}(A_n^*)\big)\big\{\log\mu_n\big(C_{a_n'}(A_n^*)\big)+\frac{1}{2}\log\det\Lambda_{a_n'}\big\}\to 0$ *we have* $\int_{C_{a_n}(A_n^*)} t_{a_n'}\,\mathrm{d}\mu_n \to 0$ *as* $n\to\infty$*. Now write*

$$\begin{aligned}
\Phi_n(A_n^*) &= \sum_{a\in A_n^*}\int_{C_a(A_n^*)} t_a\,\mathrm{d}\mu_n\\
&= \sum_{a\ne a_n}\int_{C_a(A_n^*)} t_a\,\mathrm{d}\mu_n + \int_{C_{a_n}(A_n^*)} t_{a_n}\,\mathrm{d}\mu_n + \int_{C_{a_n}(A_n^*)} t_{a_n'}\,\mathrm{d}\mu_n - \int_{C_{a_n}(A_n^*)} t_{a_n'}\,\mathrm{d}\mu_n
\end{aligned}$$

and put $A_n' = \{a\in A_n^* \mid a\ne a_n\}\in\Theta_{\le g-1,c}$*. The sum of the first and the third term on the right is the* μ_n*-integral of a function pieced together from the functions* t_a*,* $a\in A_n'$*. It is thus* $\ge \int \min_{a\in A_n'} t_a\,\mathrm{d}\mu_n = \Phi_n(A_n')$*. Since* t_{a_n} *is lower bounded, the* $\liminf_n$ *of the second term is* ≥ 0 *and we have already seen that the* $\limsup_n$ *of the last term vanishes. Therefore,*

$$\begin{aligned}
&\liminf_n \min_{A\in\Theta_{\le g,c}}\Phi_n(A) = \liminf_n \Phi_n(A_n^*) \ge \liminf_n \Phi_n(A_n')\\
&\ge \liminf_n \min_{A\in\Theta_{\le g-1,c}}\Phi_n(A) = \min_{A\in\Theta_{\le g-1,c}}\Phi(A)
\end{aligned}$$

by the inductive hypotheses (c) and (d). Also, $\limsup_n \min_{A\in\Theta_{\le g,c}}\Phi_n(A) \le \limsup_n \Phi_n(A_0) = \Phi(A_0)$ *for all* $A_0\in\Theta_{\le g,c}$ *by the SLLN and, hence,* $\limsup_n \min_{A\in\Theta_{\le g,c}}\Phi_n(A) \le \inf_{A\in\Theta_{\le g,c}}\Phi(A)$*. We conclude*

$$\limsup_n \min_{A\in\Theta_{\le g,c}}\Phi_n(A) \le \inf_{A\in\Theta_{\le g,c}}\Phi(A) \le \min_{A\in\Theta_{\le g-1,c}}\Phi(A).$$

Both estimates combine to show $\inf_{A\in\Theta_{\le g,c}}\Phi(A) = \min_{A\in\Theta_{\le g-1,c}}\Phi(A)$*, that is, g is no drop point in this case, the other half of claim (b). Moreover, claims (c) and (d) follow for g.* □

We are now prepared to prove consistency of k-parameters clustering in the HDBT constrained Elliptical MAP classification model with the radial function ϕ.

Theorem 1 *Let* $0<c\le 1$*. Let* $(X_i)_i$ *be i.i.d. with common distribution* μ*. Assume that*

(i) hyperplanes in $\mathbb{R}^d$ *are* μ*-null sets;*

(ii) ϕ *is continuous and increasing;*

(iii) $\phi(t)\ge b_0+b_1t^\alpha$ *for some numbers* $b_0\in\mathbb{R}$*,* $b_1,\alpha>0$*;*

(iv) $\phi^p\big(\beta\|X_1\|^2\big)$ *is P-integrable for some $p > 1$ and all $\beta \geq 1$.*

Then the following claims hold true for all $g \geq 1$:

(a) P-a.s., each sampling criterion (7), Φ_n, $n \geq gd+1$, has a minimizer $A_n^* \in \Theta_{\leq g,c}$;

(b) the population criterion (8), Φ, has a minimizer $A^* \in \Theta_{\leq g,c}$;

(c) we have $\sum_{a\in A_n^*}\pi_a = 1$, $n \geq gd+1$, and $\sum_{a\in A^*}\pi_a = 1$.

Moreover, if g is a drop point then

(d) P-a.s., any sequence of minimizers A_n^* of Φ_n on $\Theta_{\leq g,c}$ converges to the set of minimizers of Φ on $\Theta_{\leq g,c}$.

(e) In particular: If the minimizer A^* of Φ on $\Theta_{\leq g,c}$ is unique then (A_n^*) converges P-a.s. to A^* for any choice of minimizers A_n^*.

Proof 7 *Claims (a) and (b) are just Lemma 7(a),(c) and claim (c) was discussed after the definitions of Φ_n and Φ at the beginning of this section. If g is a drop point then Lemma 7(b) says that the optimal sampling parameters remain in a compact subset $\mathcal{K}_g \subseteq \Theta_{\leq g,c}$. By Lemma 6, $\mathcal{K}_g$ contains at least one minimum of Φ. Denote the set of these minima by K $(\subseteq \mathcal{K}_g)$. For $\varepsilon > 0$ let $U_\varepsilon = \{A \in \mathcal{K}_g \mid \Phi(A) \leq \min\Phi + \varepsilon\}$. If U is any open neighborhood of the compact set K in $\mathcal{K}_g$ then $\bigcap_{\varepsilon>0} U_\varepsilon\setminus U = K\setminus U = \emptyset$. By compactness, $U_\varepsilon\setminus U = \emptyset$ for some $\varepsilon > 0$, that is $U_\varepsilon \subseteq U$. Hence, U_ε forms a neighborhood base of K. Because of Lemma 2, all minima of Φ_n lie in U_ε for eventually all n. This is the consistency (d) and the consistency (e) is a direct consequence.* □

Remarks 1 (a) Of course, analogues of Theorem 1 can be stated for normal and elliptical submodels such as ones with diagonal or spherical scale matrices.

(b) Since we use *sets A* as parameters (and not tuples) there is no label switching and so, it cannot cause the usual non-uniqueness of the minimizers of the population and sample criteria. Yet, there may be non-uniqueness whenever different sets A can generate the same minimum $\min_{a\in A} t_a$. Non-uniqueness occurs even in the contrary case when the parent μ bears symmetries. For instance, bivariate standard normals centered at the four vertices of a square allow two equivalent minima of the population criterion (8) on $\Theta_{\leq 2,c}$. For these reasons we state in Theorem 1(d) "converges to the set of minimizers of Φ."

(c) If there is an uninterrupted chain of drop points $1,\ldots,g_{\max}$ then the claims of Theorem 1 hold true for $g \in 1\,..\,g_{\max}$ even with assumption (iv) relaxed to the assumption (ii) of Lemma 2:

$\phi\big(\beta\|X_1\|^2\big)$ is integrable for all $\beta \geq 1$.

Indeed, in this case it is not necessary to resort to Lemma 7 in order to show that the optimal sampling parameters remain in a compact subset of $\Theta_{\leq g,c}$. The proof proceeds again by induction over $g \in 1\,..\,g_{\max}$: The case $g = 1$ is as in Lemma 7. For $g \geq 2$, we verify the assumption of Lemma 5(b). Let $A_0 \in \Theta_{\leq g,c}$ such that $\Phi(A_0) \leq \inf_{A\in\Theta_{\leq g,c}}\Phi(A) + \varepsilon$. By convergence $\Phi_n \to \Phi$, Lemma 2,

$$\min_{A\in\Theta_{\leq g,c}}\Phi_n(A) \leq \Phi_n(A_0) \leq \Phi(A_0) + \varepsilon \leq \inf_{A\in\Theta_{\leq g,c}}\Phi(A) + 2\varepsilon$$

if n is large. Since g is a drop point it follows

$$\begin{aligned}&\limsup_n \min_{A\in\Theta_{\le g,c}} \Phi_n(A) \le \inf_{A\in\Theta_{\le g,c}} \Phi(A)\\ &< \inf_{A\in\Theta_{\le g-1,c}} \Phi(A) = \lim_n \min_{A\in\Theta_{\le g-1,c}} \Phi_n(A).\end{aligned}$$

The last equality follows from the inductive hypothesis and from Lemma 6(b). This is the assumption of Lemma 5(b) for g. The rest of the proof proceeds as in the theorem.

(d) As in Pollard's theorem, the parent distribution μ in the consistency Theorem 1 does not have to be a member of the collection of elliptical mixtures represented by the sets in $\Theta_{\le g,c}$. On the other hand, even if μ is an elliptical mixture, it cannot be the mixture associated with the limit. This is in contrast to the mixture model. No matter how well the components are separated, think for instance of two, the proportions in the tails on the opposite side of the separating hypersurface are assigned to the wrong cluster. Thus, the variances are underestimated and the distance between the mean values is overestimated in the limit as $n \to \infty$, see Marriott [26] and Bryant and Williamson [8]. However, this bias disappears as cluster separation grows since there is less overlap.

(e) This and the next remark concern modified versions of drop points. In Definition 1, they have been defined w.r.t. the (MAP) population criterion (8). The criterion is accompanied by an ML version, just delete the term $-\log \pi_a$ from the definition of t_a just before (7). If μ has no discrete part, if $\phi(t) > \phi(0)$ for all $t > 0$, and if the assumptions of the theorem are satisfied ((iv) may again be relaxed to assumption (ii) of Lemma 2), it can be shown that the ML version strictly decreases with increasing number of components $g \ge 1$. This behavior is actually not desirable in cluster analysis. If μ is a mixture of g well–separated components then we would prefer that any solution with more than g components be rejected, at least up to an a priori given upper bound. The ML criterion does not comply with this wish. Contrary to the ML criterion, the MAP criterion may possess non–drop points. Examples are presented after the proposition below.

(f) Drop points are accompanied by *sample drop points* g of the sampling criterion (7), Φ_n,

$$\inf_{A\in\Theta_{\le g,c}} \Phi_n(A) < \inf_{A\in\Theta_{\le g-1,c}} \Phi_n(A).$$

Also $g = 1$ is a sample drop point. The present classification model (3) is related to the mixture model $\prod_{i=1}^n \sum_j \pi_j E_{\phi,m_j,V_j}(x_i)$ via an approximation in the case of good cluster separation. A reviewer therefore asked us to compare the decrease of the minimum sampling criterion (7) with the increase of the maximum log-likelihood values of the mixture model along increasing values of g. Everything depends heavily on the scale constraints used. Besides HDBT constraints, other popular scale constraints are compactness of the scale space or lower bounds on the scale parameters, for instance boundedness of all eigenvalues away from zero. A well–known theorem of Lindsay's [22, 23] hinges on compactness of the parameter space and states that the MLE of the mixing distribution can be represented by a discrete probability supported by at most n points. It does not hold true under HDBT constraints which guarantee the MLE to exist only for $n \ge gd + 1$, see Gallegos and Ritter [11]. Now, under compactness of the parameter space and general position of the data, minima of the heteroscedastic elliptical criterion (4) and of the related sampling criterion (7), Φ_n, exist even for all values of g up to n and not all have to be sample drop points. This is also true for the mixture model. Moreover, in both cases we do not have to look for g beyond n: Although the mixture model still exists, the maximum likelihood does not improve according to Lindsay's theorem, and the classification model looses its sense because there cannot be more than n nonempty clusters.

(g) The assumption (iii) of the theorem excludes radial functions ϕ of logarithmic growth. Such functions define, for instance, distributions of Pearson's type VII (or multivariate Student t). Lemma 4 needs the multiplicativity of $t \mapsto t^\alpha$ which the logarithm does not share. It is, however, possible to adapt the proof of Lemma 4 to radial functions $\phi(t) = \frac{\eta}{2}\log(1+t)+\mathrm{const}$, $\eta > (d+1)d$, at the cost of introducing the assumption

$$\inf_{B\in\Theta'_{\le g},\|v\|=1}\int \min_{b\in B}\log|v^{\mathrm{T}}x-b|\mu(\,\mathrm{d}x)\in\mathbb{R} \tag{16}$$

(the space $\Theta'_{\le g}$ is defined in the proof of Lemma 3) and of strengthening the condition (iii) to integrability of $\phi(\beta\|X_1\|^2)$ for all $\beta \ge 1$. Theorem 1 holds true for such ϕ if assumption (iii) is replaced with (16). The optimum of the sampling criterion (7) exists if $n \ge \frac{gd}{1-(d+1)d/\eta}+1$.

Theorem 1(b) states that, under certain assumptions, the population criterion Φ has a minimum. The following theorem provides a tool to compute it. We will use it to verify the subsequent examples but it is interesting in its own right since it is the population version of the Elliptical MAP criterion (4).

Proposition 1 *Let μ and ϕ satisfy the assumptions of Lemma 2 and the assumptions (i) and (iii) of Theorem 1. Denote partitions of the sample space $\mathbb{R}^d$ in at most g measurable subsets C such that $\mu(C) > 0$ by the letter $\mathbf{P}$. Let $\mathbf{m} = (m_C)_{C\in\mathbf{P}}$ and $\mathbf{V} = (V_C)_{C\in\mathbf{P}}$.*
(a) For each such partition $\mathbf{P}$ the minimum of

$$\sum_{C\in\mathbf{P}}\Big\{\tfrac{\mu(C)}{2}\log\det V_C + \int_C \phi\big((x-m_C)^{\mathrm{T}}V_C^{-1}(x-m_C)\big)\mu(\,\mathrm{d}x)\Big\}$$

w.r.t. $\mathbf{m}$ and $\mathbf{V}\in\mathcal{V}_c$ exists.
(b) The population criterion Φ has a minimum on $\Theta_{\le g,c}$ if and only if

$$\begin{aligned}&\mathrm{H}(\mu(C)\mid C\in\mathbf{P})\\&+\min_{\mathbf{m},\mathbf{V}\in\mathcal{V}_c}\sum_{C\in\mathbf{P}}\Big\{\tfrac{\mu(C)}{2}\log\det V_C + \int_C \phi\big((x-m_C)^{\mathrm{T}}V_C^{-1}(x-m_C)\big)\mu(\,\mathrm{d}x)\Big\}\end{aligned} \tag{17}$$

has a minimum w.r.t. all $\mathbf{P}$. In this case, the minima coincide.
(c) In the homoscedastic, *normal case $\phi(t) = \frac{t}{2}+\frac{d}{2}\log 2\pi$, denote the pooled covariance matrix of μ w.r.t. $\mathbf{P}$ by $V(\mathbf{P}) = \sum_{C\in\mathbf{P}} V[X_1; X_1 \in C]$. The minimum of Φ on $\Theta_{\le g,1}$ exists if and only if the minimum of*

$$\mathrm{H}(\mu(C)\mid C\in\mathbf{P}) + \tfrac{d}{2}(1+\log 2\pi) + \tfrac{1}{2}\log\det V(\mathbf{P})$$

exists w.r.t. all $\mathbf{P}$ and, in this case, the minima are equal.

Proof 8 *(a) By the assumptions on μ and ϕ, the sum is continuous as a function of $\mathbf{m} = (m_C)_C$ and $\mathbf{V} = (V_C)_C$. Let $\Lambda_C = \sum_k \lambda_k v_k v_k^{\mathrm{T}}$ be the spectral decomposition. We have $(x - m_C)^{\mathrm{T}} V_C^{-1}(x - m_C) = \sum_k \lambda_k (v_k^{\mathrm{T}}(x - m_C))^2$ and, by Theorem 1(iii) and the increase of $t \mapsto t^\alpha$,*

$$\begin{aligned}
&\int_C \phi\big((x - m_C)^{\mathrm{T}} \Lambda_C (x - m_C)\big)\mu(\,\mathrm{d}x) \\
&\ge b_0 \mu(C) + b_1 \int_C \Big(\sum_k \lambda_k (v_k^{\mathrm{T}}(x - m_C))^2\Big)^\alpha \mu(\,\mathrm{d}x) \\
&\ge b_0 \mu(C) + \tfrac{b_1}{d} \int_C \sum_k \big(\lambda_k (v_k^{\mathrm{T}}(x - m_C))^2\big)^\alpha \mu(\,\mathrm{d}x) \\
&= b_0 \mu(C) + \tfrac{b_1}{d} \sum_k \lambda_k^\alpha \int_C \big|v_k^{\mathrm{T}}(x - m_C)\big|^{2\alpha} \mu(\,\mathrm{d}x) \\
&\ge b_0 \mu(C) + \tfrac{b_1}{d} \sum_k \lambda_k^\alpha \min_{\|v\|=1, m} \int_C \big|v^{\mathrm{T}}(x - m)\big|^{2\alpha} \mu(\,\mathrm{d}x).
\end{aligned}$$

The minimum exists since the integral is continuous as a function of m and v and converges to ∞ as $\|m\| \to \infty$. Moreover, by assumption Theorem 1(i) and $\mu(C) > 0$ it is a strictly positive constant κ_C. It follows

$$\begin{aligned}
&\tfrac{\mu(C)}{2} \log\det V_C + \int_C \phi\big((x - m_C)^{\mathrm{T}} \Lambda_C (x - m_C)\big)\mu(\,\mathrm{d}x) \\
&\ge b_0 \mu(C) + \sum_k \Big\{ -\tfrac{\mu(C)}{2} \log \lambda_k + \kappa_C \tfrac{b_1}{d} \lambda_k^\alpha \Big\}.
\end{aligned}$$

This expression converges to ∞ as $\lambda_k \to 0$ or $\lambda_k \to \infty$. It is, thus, sufficient to consider matrices V_C such that $c_1 I_d \preceq \Lambda_C \preceq c_2 I_d$ for two numbers $0 < c_1 \le c_2$. Since $\phi(t) \to \infty$ as $t \to \infty$ by assumption (iii) of Theorem 1 and since $\mu(C) > 0$, it follows that each integral tends to ∞ as $\|m_C\| \to \infty$. It is therefore sufficient to restrict the range of each m_C to a compact subset of $\mathbb{R}^d$ and the claim follows from continuity.

(b) We have to show that, to each $A \in \Theta_{\le g,c}$, there corresponds some $\mathbf{P}$ for which (17) is no larger than $\Phi(A)$ and vice versa. The first claim follows from (13). For the converse let $\mathbf{P}$ be given and let $m(C)$ and $V(C)$ be the minimizers w.r.t. $m_C \in \mathbb{R}^d$ and $(V_C)_C \in \mathcal{V}_c$ in (a). The elements $a_C = (\mu(C), m(C), V(C))$, $C \in \mathbf{P}$, satisfy $\{a_C \mid C \in \mathbf{P}\} \in \Theta_{\le g,c}$ and we have

$$\begin{aligned}
&\mathrm{H}(\mu(C) \mid C \in \mathbf{P}) \\
&\quad + \min_{\mathbf{m}, \mathbf{V} \in \mathcal{V}_c} \sum_{C \in \mathbf{P}} \Big\{ \tfrac{\mu(C)}{2} \log\det V_C + \int_C \phi\big((x - m_C)^{\mathrm{T}} \Lambda_C (x - m_C)\big)\mu(\,\mathrm{d}x) \Big\} \\
&= \sum_{C \in \mathbf{P}} \int_C \Big\{ -\log \mu(C) + \tfrac{1}{2} \log\det V(C) \\
&\qquad\qquad + \phi\big((x - m(C))^{\mathrm{T}} \Lambda_C (x - m(C))\big) \Big\} \mu(\,\mathrm{d}x) \\
&= \sum_{C \in \mathbf{P}} \int_C t_{a_C} \,\mathrm{d}\mu \ge \int_{\mathbb{R}^d} \min_C t_{a_C} \,\mathrm{d}\mu.
\end{aligned}$$

This is the desired inequality.

(c) The proof in the homoscedastic, normal case is similar to that of the Pooled Determinant criterion: Let $PD(d)$ denote the convex cone of all positive definite, symmetric d by d matrices.

$$\begin{aligned}
&\min_{\mathbf{m},V\in PD(d)} \sum_{C\in\mathbf{P}} \Big\{ \tfrac{\mu(C)}{2}\log\det V + \int_C \phi\big((x-m_C)^{\mathrm{T}}\Lambda(x-m_C)\big)\mu(\,\mathrm{d}x)\Big\}\\
&= \tfrac12 \min_{\mathbf{m},V\in PD(d)} \Big\{ \log\det V + d\log 2\pi + \sum_{C\in\mathbf{P}}\int_C (x-m_C)^{\mathrm{T}}\Lambda(x-m_C)\mu(\,\mathrm{d}x)\Big\}\\
&= \tfrac12 \min_{V\in PD(d)} \big\{ \log\det V + d\log 2\pi + tr(\Lambda V(\mathbf{P}))\big\}\\
&= \tfrac12 \big\{ d(1+\log 2\pi) + \log\det V(\mathbf{P})\big\}.
\end{aligned}$$

This is claim (c). □

In parts (b) and (c) of the proposition it is obviously sufficient to take the minimum w.r.t. a collection of partitions $\mathbf{P}$ that is known to contain the optimal one. The following examples show that the minimum of the population criterion (8), Φ, regarded as a function on $\Theta_{\le g,c}$, does not always decrease as the number of components, g, increases. In other words, there are non–drop points.

Example 1 Let $\mu = N_{0,I_d}$ and let the approximating model be the normal location and scale family, that is, $\phi(t) = (d/2)\log 2\pi + t/2$. For $g = 1$, the entropy inequality shows that the optimal solution in $\Theta_{\le 1,c}$ is $\{(1,0,I_d)\}$. Let now $g \ge 2$, $A = \{(\pi_1,m_1,V_1),\dots,(\pi_g,m_g,V_g)\} \in \Theta_{\le g,c}$, (m_j,V_j) pairwise distinct, $\sum \pi_j = 1$, $\Lambda_j = V_j^{-1}$, $f_j(x) = (2\pi)^{-d/2}\sqrt{\det\Lambda_j}e^{-(x-m_j)^{\mathrm{T}}\Lambda_j(x-m_j)/2}$, and abbreviate $t_j = t_{(\pi_j,m_j,V_j)}$. We have $\sum_j \pi_j f_j > \max_j \pi_j f_j$ and, hence, $-\log\sum_j \pi_j f_j < \min_j t_j$. The entropy inequality shows

$$\Phi(A) = \int \min_j t_j \,\mathrm{d}\mu > -\int \log\sum_j \pi_j f_j \,\mathrm{d}\mu \ge -\int \log N_{0,I_d}\,\mathrm{d}\mu = \Phi(\{(1,0,I_d)\}).$$

Thus, the only optimal solution in $\Theta_{\le g,c}$ is the singleton $\{(1,0,I_d)\}$. No genuine mixture of normals is superior. This is true for any HDBT constant $c \le 1$.

Example 2 Example 1 raises the question whether the population criterion Φ can decrease after having been constant for at least two (consecutive) values of g. The answer is yes. Consider the homoscedastic normal classification model, that is $c = 1$, and the distribution μ on the real line with Lebesgue density

$$f_0(x) = \begin{cases} 1/(8\alpha), & |x\pm 1| < \alpha,\\ 1/(4\alpha), & |x| < \alpha,\\ 0, & \text{otherwise,}\end{cases}$$

for $0 < \alpha < 1/3$. The optimal solution for $g = 1$ w.r.t. the population criterion Φ is $\{(1,0,v)\}$ with $v = \frac12 + \frac{\alpha^2}{3}$ and $\Phi(\{(1,0,v)\}) = \frac12(\log 2\pi + 1 + \log v)$.

In order to see that there is no better solution in $\Theta_{\le 2,1}$ note that any solution $A^* = \{a_1,a_2\}$, $a_1 \ne a_2$, is specified by some cut $s^* \in]-1-\alpha, 1+\alpha[$ that separates $C_{a_1}(A^*)$ from $C_{a_2}(A^*)$.

Let F be the cumulative distribution function belonging to f_0 and let $R = 1 - F$ be its tail distribution. According to Proposition 1(c) it is sufficient to run across all cuts s and to compute entropy and pooled covariance

$$v(s) = \int_{-\infty}^{s} (x - m_1(s))^2 f_0(x)\,\mathrm{d}x + \int_{s}^{\infty} (x - m_2(s))^2 f_0(x)\,\mathrm{d}x,$$

with the conditional expectations

$$\begin{aligned} m_1(s) &= \mathrm{E}[X_1 \mid X_1 < s] = \tfrac{1}{F(s)} \int_{-\infty}^{s} x f_0(x)\,\mathrm{d}x, \\ m_2(s) &= \mathrm{E}[X_1 \mid X_1 > s] = \tfrac{1}{R(s)} \int_{s}^{\infty} x f_0(x)\,\mathrm{d}x. \end{aligned}$$

Omitting the addend $\frac{1}{2}\log(2\pi)$ in Φ, the integral version of Lemma A.3 in Gallegos and Ritter [10], a formula of Steiner's type, asserts that

$$\begin{aligned} \Phi(A_s) &= \mathrm{H}(F(s), R(s)) + \tfrac{1}{2}\big(1 + \log v(s)\big) \\ &= \mathrm{H}(F(s), R(s)) + \tfrac{1}{2} + \tfrac{1}{2}\log\big(v - F(s)R(s)(m_2(s) - m_1(s))^2\big), \end{aligned}$$

where $A_s = \{a_1(s), a_2(s)\}$, $a_1(s) = (F(s), m_1(s), v(s))$, $a_2(s) = (R(s), m_2(s), v(s))$, and where v is the total variance above. The difference between this value and the optimum for $g = 1$ is

$$\begin{aligned} &\Phi(A_s) - \Phi(\{(1, 0, v)\}) \\ &= \mathrm{H}(F(s), R(s)) + \tfrac{1}{2}\log\Big(1 - F(s)R(s)\frac{(m_2(s) - m_1(s))^2}{v}\Big). \end{aligned} \tag{18}$$

We have to show that this number is strictly positive for all $s \in]-1-\alpha, 1+\alpha[$ and begin with $s \in]-1-\alpha, -1+\alpha[$. The conditional expectations are $m_1(s) = \frac{1}{2}(s - 1 - \alpha)$ and $m_2(s) = \frac{(1+\alpha)^2 - s^2}{14\alpha - 2(s+1)}$. Hence,

$$m_2(s) - m_1(s) = 4\alpha \frac{1 + \alpha - s}{7\alpha - (s+1)} \le \tfrac{4}{3}$$

since $\alpha < 1/3$. Inserting in (18) and observing $v > 1/2$ yields

$$\Phi(A_s) - \Phi(\{(1, 0, v)\}) \ge \mathrm{H}(F(s), R(s)) + \tfrac{1}{2}\log\Big(1 - \tfrac{32}{9}F(s)R(s)\Big).$$

The derivatives w.r.t. F of the functions $-F\log F$ and $-R\log R$, $R = 1 - F$, are strictly decreasing on $]0, 1/4[$. The same is true for $\log\big(1 - \frac{32}{9}FR\big)$ since

$$\frac{\mathrm{d}}{\mathrm{d}F}\log\Big(1 - \tfrac{32}{9}(1 - F)F\Big) = -\tfrac{32}{9}\frac{1 - 2F}{1 - \frac{32}{9}(1 - F)F}$$

and since $\frac{32}{9}(1 - F) \ge 2$ for $F \le \frac{1}{4}$. Hence, the function $\mathrm{H}(F, R) + \frac{1}{2}\log\big(1 - \frac{32}{9}FR\big)$ is strictly concave. Since it vanishes at $F = 0$ and has the value $-\big(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\big) + \frac{1}{2}\log\frac{1}{3} = \log 4 - \frac{5}{4}\log 3 = 0.01302\ldots$ at $F = 1/4$, it is strictly positive for $0 < F \le 1/4$. That is, $\Phi(A_s) > \Phi(\{(1, 0, v)\})$ for $-1 - \alpha < s \le -1 + \alpha$, the first claim. The value obtained for $s = -1 + \alpha$ persists on the interval $[-1 + \alpha, -\alpha]$ since $F(s) = 1/4$, $m_1(s) = -1$ and $m_2(s) = 1/3$ do not depend on s.

For reasons of symmetry we are done if we verify the claim for $s \in]-\alpha, 0]$. In this case, $F(s) = \frac{1}{2} + \frac{s}{4\alpha}$, $R(s) = \frac{1}{2} - \frac{s}{4\alpha}$, $F(s)m_1(s) = \frac{s^2-\alpha^2}{8\alpha} - \frac{1}{4} = -R(s)m_2(s)$, and $m_2(s) - m_1(s) = \frac{R(s)m_2(s)}{F(s)R(s)}$. Hence

$$R(s)F(s)(m_2(s) - m_1(s))^2 = \frac{(R(s)m_2(s))^2}{R(s)F(s)} = \frac{(s^2 - \alpha^2 - 2\alpha)^2}{4(4\alpha^2 - s^2)}.$$

For $0 < \alpha \le 0.3$ the right–hand side is $\le 1/3$. Indeed, $3(\alpha^2 - s^2) + 12\alpha < 4$ and, hence,

$$\begin{aligned}&3(2\alpha + \alpha^2 - s^2)^2 = 12\alpha^2 + \big(3(\alpha^2 - s^2) + 12\alpha\big)(\alpha^2 - s^2)\\&< 12\alpha^2 + 4(\alpha^2 - s^2) = 4\big(4\alpha^2 - s^2\big).\end{aligned}$$

Inserting in (18) yields

$$\begin{aligned}&\Phi(A_s) - \Phi(\{(1, 0, v)\})\\&\ge \mathrm{H}(F(s), R(s)) + \tfrac{1}{2}\log\big(1 - 2F(s)R(s)(m_2(s) - m_1(s))^2\big)\\&\ge \mathrm{H}\big(\tfrac{1}{4}, \tfrac{3}{4}\big) + \tfrac{1}{2}\log\tfrac{1}{3}.\end{aligned}$$

This is again the strictly positive number computed above. We have shown that the optimum number of components up to two is one and that Φ is bounded below on $\Theta_{\le 2,c}$ by a number independent of α. The minimum of Φ on $\Theta_{\le 3,c}$ is smaller than that on $\Theta_{\le 2,c}$, at least for small α. It is in fact unbounded below as $\alpha \to 0$.

We finally remark that the situation changes completely as we consider the uniform sampling distribution on the set $[-1-\alpha, -1+\alpha] \cup [-\alpha, \alpha] \cup [1-\alpha, 1+\alpha]$. Here the optimal solution for $g = 1$ is no longer optimal for $g \le 2$. By weak continuity it is sufficient to consider the weak limit as $\alpha \to 0$, the discrete probability $\mu = \frac{1}{3}(\delta_{-1} + \delta_0 + \delta_1)$. The optimal solution for $g = 1$ is $\big\{\big(1, 0, \frac{2}{3}\big)\big\}$, its population criterion being (up to the constant $(\log 2\pi)/2$) $\Phi\big(\big\{\big(1, 0, \frac{2}{3}\big)\big\}\big) = \frac{1}{2}\big(1 + \log\frac{2}{3}\big) = 0.2972\ldots$. A solution for $g = 2$ is $A_2 = \{a_1, a_2\}$ with $a_1 = \big(\frac{1}{3}, -1, \frac{1}{6}\big)$ and $a_2 = \big(\frac{2}{3}, \frac{1}{2}, \frac{1}{6}\big)$. Its criterion is $\Phi(A_2) = \mathrm{H}\big(\frac{1}{3}, \frac{2}{3}\big) + \frac{1}{2}\big(1 + \log\frac{1}{6}\big) = 0.2406\ldots$.

Example 3 This example shows that g is a drop point of a homoscedastic mixture of $g \ge 2$ normal distributions on $\mathbb{R}^d$ if there is sufficient separation. It can be extended to more general mixtures but we want to use Proposition 1(c). The example holds for all dimensions d but our proof for $d \ge 2$ is somewhat technical and we confine ourselves to $d = 1$. For $v > 0$, $m_1 < m_2 < \cdots < m_g$, and $\pi_j > 0$ such that $\sum \pi_j = 1$, consider the homoscedastic, normal mixture $\mu_v = \sum_{j=1}^{g} \pi_j N_{m_j,v}$. We denote the population criterion (8) w.r.t. μ_v by $\Phi_v(A) = \int \min_{a\in A} t_a \, \mathrm{d}\mu_v$ and show that its minimum over $A \in \Theta_{\le g-1,1}$ remains bounded below, while that over $A \in \Theta_{\le g,1}$ becomes arbitrarily small as $v \to 0$.

By Proposition 1(c), $\min_{A\in\Theta_{\le h,1}} \Phi_v(A)$ is the minimum of

$$\mathrm{H}(\mu_t(C) \mid C \in \mathbf{P}) + \tfrac{1}{2}(1 + \log 2\pi) + \tfrac{1}{2}\log V_v(\mathbf{P}) \tag{19}$$

taken over all partitions $\mathbf{P}$ of $\mathbb{R}$ in at most h measurable subsets. Here, $V_v(\mathbf{P})$ is the pooled variance of $\mathbf{P}$ w.r.t. μ_v. Any partition $\mathbf{P}$ of $\mathbb{R}$ in $h < g$ subsets contains at least one subset C_v where two different components of μ_v contain probability $\ge 1/h$. Indeed, the stochastic matrix $(N_{m_j,v}(C) \mid 1 \le j \le g, C \in \mathbf{P})$ with indices j and C has g rows and h columns. So, each row contains an entry $\geq 1/h$ and, since there are more rows than columns, the pigeon hole principle shows that one column must contain two such entries. In other words, there are two different components that load some subset $C_v \in \mathbf{P}$ with probability at least $1/h$, each. Since all elements m_j are different, a moment of reflection shows that the conditional variance w.r.t. C_v and, hence, the pooled variance $V_v(\mathbf{P})$ remain bounded away from zero as $v \to 0$. By Proposition 1(c), (19) is bounded below and so is $\Phi_v(A)$ uniformly for $A \in \Theta_{\leq g-1,1}$, the first half of the claim.

Now let $h = g$. We construct a partition $\mathbf{P}$ with a small value of (19). Let $3r > 0$ be the minimum distance between any two means m_j. For $1 \leq j < h$, let $C_j = [m_j - r, m_j + r] \subseteq \mathbb{R}$, let C_h be the complement of $\bigcup_{j<h} C_j$ in $\mathbb{R}$, and let $\mathbf{P}$ be the partition $\{C_1, \dots, C_h\}$. As $v \to 0$, $N_{m_j,v}$ concentrates on the neighborhood of $m_j \in C_j$ and, therefore, the pooled variance $V_v(\mathbf{P})$ converges to zero. Hence, (19) diverges to $-\infty$ as $v \to 0$ and, again by Proposition 1(c), so does the minimum of Φ_v on $\Theta_{\leq g,1}$, the second half of the claim. We have shown that g is a drop point if v is small enough.

3 Model selection

The parameters subject to estimation in the foregoing theory are the HDBT constant c, the number of clusters, g, and the set A itself. Some authors claim that the solutions are often ambiguous, see Gondek [15], p. 245, "*Data contains many plausible clusterings*" and p. 249. Jain et al. [18] even feel that clustering is a subjective process. We, therefore, do not always expect a unique solution.

Besides the set A, the HDBT constant c, too, is unknown. Since set and constant depend on each other, it is reasonable to combine their estimation. For this purpose, Gallegos and Ritter [11, 12] introduced for the mixture and classification models the affine invariant method of "balanced scales." It is inspired by the HDBT constraints and postulates as an additional statistical assumption that a valid solution to a clustering problem requires not only good fit, that is a small criterion, but also sufficiently balanced cluster scales, expressed by not too small an HDBT constant c in order to avoid spurious clusters. The two objectives are often in conflict which has then to be resolved by biobjective optimization. For further information on this subject, we refer the interested reader to the two papers cited above.

The problem of estimating g continues to be a subject of discussion. Examples 1 and 3 indicate that the number of components of a clear mixture will often be a drop point. But the last paragraph of Example 2 shows that not every drop point can be considered a valid number of components. Given an upper bound $g_{\max} \geq 1$, we may consider the set A^* of smallest size that minimizes the population criterion (8), Φ, on $\Theta_{\leq g_{\max},c}$ a reasonable solution. Its size $g^* = |A^*|$, the largest drop point up to $g_{\max}$, is a candidate for the "number of components" that make up μ.

What we actually observe are not (population) drop points but *sample drop points*, see Remark 1(f). Because of random fluctuation, a population drop point may not be a sample drop point and vice versa. The following result says that a small (but unknown) distortion of Φ_n converts the largest sample drop point up to $g_{\max}$ to the desired estimate of the number of components, at least asymptotically. Denote the sequence of (population) drop points by $1 = g_1 < g_2 < \cdots$.

Corollary 1 *Let $g_{\max} \geq 2$ and let $(s_1, \ldots, s_{g_{\max}})$ be any strictly increasing $g_{\max}$-tuple of real numbers. If the assumptions of Theorem 1 are satisfied then there exists $\varepsilon_0 > 0$ such that, for any $0 < \varepsilon \leq \varepsilon_0$, the maximum (population) drop point up to $g_{\max}$ is given by*

$$\underset{1 \leq g \leq g_{\max}}{\operatorname{argmin}} \Big(\min_{A \in \Theta_{\leq g,c}} \Phi_n(A) + \varepsilon s_g \Big),$$

for eventually all n.

Proof 9 *Write $g^* = \max_{g_j \leq g_{\max}} g_j$ and $h_g = \min_{A \in \Theta_{\leq g,c}} \Phi(A)$. We have $h_1 \geq \cdots \geq h_{g^*-1} > h_{g^*} = h_{g^*+1} = \cdots = h_{g_{\max}}$. If $g^* = 1$ then we choose an arbitrary $\varepsilon_0 > 0$. If $g^* > 1$ then there is $\varepsilon_0 > 0$ such that*

$$h_g + \varepsilon_0 s_g > h_{g^*} + \varepsilon_0 s_{g^*}$$

for all $1 \leq g < g^*$. This relation continues to hold for all strictly positive $\varepsilon \leq \varepsilon_0$ and, of course, also for $g > g^*$. The corollary now follows from $\min_{A \in \Theta_{\leq g,c}} \Phi_n(A) \to h_g$ as $n \to \infty$, see Lemma 6(b). □

We have obtained a *penalized* (MAP) sampling criterion. The proof shows that the penalty term is needed because of the random fluctuation of the nth minimum $\min_{A \in \Theta_{\leq g,c}} \Phi_n(A)$ for g beyond the maximum drop point g^*. Being defined by the population criterion Φ, the point g^* is an asymptotic quantity that also depends on the choice of the HDBT constant c. If the constraints are omitted, then, beyond g^*, the sampling criterion usually splits some cluster or splits off clusters of small or even deficient size producing spurious solutions.

The corollary should be compared with the traditional elbow criterion and its refinement by Tibshirani et al. [34] and, in particular, with model selection criteria for finite mixture models such as AIC, see Akaike [1], and the Bayesian information criterion BIC for finite mixtures, see Keribin [19, 20] and Nishi [27] in a more general context. Akaike, Keribin, and Nishi require that the penalty term strictly increase with model dimension and that it converge to zero with n. (Note that, in their notation, the criterion is not divided by n.) Their first requirement corresponds to the increase of our finite sequence (s_g) and the way it strictly increases is not important since all authors consider only a finite range of models. Instead of convergence to zero with n we have the small quantity ε.

We finally illustrate theorem and corollary with a data set of size 3,000 sampled from the two-dimensional normal mixture $0.27N_{(-3,4),V_1} + 0.27N_{(3,4),V_2} + 0.46N_{(0,-3),I_2}$ with $V_1 = \begin{pmatrix} 4 & \\ -3 & 4 \end{pmatrix}$ and $V_2 = \begin{pmatrix} 4 & \\ 3 & 4 \end{pmatrix}$. In order to observe the behavior of the sampling criterion for increasing n we also draw random sub-samples of sizes $n = 30$, 100, 300, and 1,000. The sub-sample with $n = 100$ points is plotted in Figure 2. In order to handle the notorious non-uniqueness of solutions, even for a fixed number of clusters, partitions globally optimal according to the method of "balanced scales" indicated above were chosen under the HDBT constant $c = 0.4$. For this purpose, we used our C++ program that implements among others the HDBT constrained heteroscedastic MAP Determinant criterion (2).

Table 1 shows the first four digits of the minima of the sampling criteria $\min_{A \in \Theta_{\leq g,c}} \Phi_n(A)$ under normality for the five values of n and $g \in 1\,..\,4$. In the present case, the entries remain

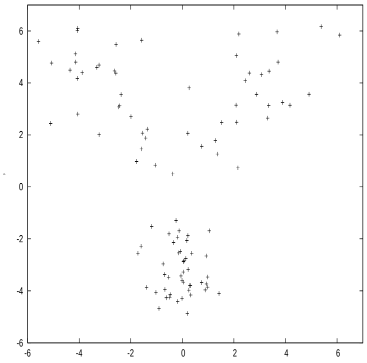

Figure 2: The sub-sample of 100 points.

n	30	100	300	1000	3000
g					
1	5.163	4.995	5.084	5.181	5.164
2	4.257	4.191	4.438	4.545	4.536
3	4.067	4.009	4.248	4.330	4.330
4	3.800	3.976	4.243	4.330	4.330

Table 1: The minimal values of the constrained sampling criteria Φ_n for various sub-sample sizes and numbers of clusters.

stable from n =1,000 and $g = 3$ on and so we conclude $g^* = 3$. This happens to be the number of components that were used to generate the data. The method can be applied to general data sets that are large enough, just draw sub-samples. It should also be repeated with different sub-samples to confirm the result.

The proof of the corollary says $\varepsilon_0 < \min_{g<g^*} \frac{h_g - h_{g^*}}{s_{g^*} - s_g}$. Letting $s_g = g$ and approximating the optimal population criterion Φ with the optimal sampling criterion Φ_n according to Lemma 2 and Lemma 7(d) we obtain for the five values of n the approximations 0.19, 0.182, 0.19, 0.214, and 0.205 to this minimum. They do not fluctuate too much. An appropriate value of ε_0 is therefore about 0.2, here. Applying the corollary with $\varepsilon = 0.1$ we see that three clusters are obtained for n = 100, 300, 1000, and 3000. The estimate obtained from the smallest subset is at least four. Note that the cluster sizes for three components are here about ten and a two-dimensional normal sample of such a small size has usually several interpretations as a clustered data set.

A caveat: Blind faith in the results of any clustering method is not advisable. All statistical methods face many opponents but they grow when we are dealing with clustering. The data set may be too small, the model chosen may be too far away from truth, there may be too few variables or too many irrelevant ones, it may contain outliers that will hamper the analysis. Most of these problems have been tackled in the past. Nevertheless, any clustering result should be validated. Are the clusters obtained cohesive ? Have different clusters been returned as a single one ? Is the data set clustered at all ? Goodness–of–fit and normality tests, for instance, can be used to answer these questions. Moreover, various validation methods are of benefit in checking the soundness of a proposed solution, for instance Bailey and Dubes's [2] Cluster Validity Profiles or Bertrand and Bel Mufti's [3] and Hennig's [17] cluster stability methods.

Acknowledgment We thank both referees for their constructive questions and comments. They clarified some paragraphs in our paper.

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19:716–723, 1974.

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

[3] P. Bertrand and G. Bel Mufti. Loevinger's measures of rule quality for assessing cluster stability. *Computational Statistics and Data Analysis*, 50:992–1015, 2006.

[4] Herman J. Bierens. *Introduction to the Mathematical and Statistical Foundations of Econometrics.* Cambridge University Press, Cambridge, 2004.

[5] Hans-Hermann Bock. *Statistische Modelle für die einfache und doppelte Klassifikation von normalverteilten Beobachtungen.* PhD thesis, University of Freiburg, Germany, 1968.

[6] Hans-Hermann Bock. Statistische Modelle und Bayessche Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren. *Metrika*, 18:120–132, 1972.

[7] Peter G. Bryant. Large-sample results for optimization-based clustering methods. *J. Classification*, 8:31–44, 1991.

[8] Peter G. Bryant and J.A Williamson. Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika*, 65:273–281, 1978.

[9] J. A. Cuesta-Albertos, Alfonso Gordaliza, and Carlos Matrán. Trimmed k-means: An attempt to robustify quantizers. *Ann. Statist.*, 25:553–576, 1997.

[10] María Teresa Gallegos and Gunter Ritter. A robust method for cluster analysis. *Ann. Statist.*, 33:347–380, 2005.

[11] María Teresa Gallegos and Gunter Ritter. Trimmed ML-estimation of contaminated mixtures. *Sankhyā, Series A*, 71:164–220, 2009.

[12] María Teresa Gallegos and Gunter Ritter. Trimming algorithms for clustering contaminated grouped data and their robustness. *Adv. Data Anal. Classif.*, 3:135–167, 2009.

[13] María Teresa Gallegos and Gunter 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.

[14] Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Agustín Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Statist.*, 36:1324–1345, 2008.

[15] David Gondek. Non-redundant data clustering. In Sugatu Basu, Ian Davidson, and Kiri L. Wagstaff, editors, *Constrained Clustering*, Data Mining and Knowledge Discovery Series, chapter 11, pages 245–283. Chapman & Hall/CRC, Boca Raton, London, New York, 2009.

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

[17] Christian Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, 52:258–271, 2007.

[18] Anil K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. *ACM Comput. Surveys*, 31:264–323, 1999.

[19] Christine Keribin. Estimation consistante de l'ordre de modèles de mélange. *C.R. Acad. Sc. Paris*, 326:243–248, 1998.

[20] Christine Keribin. Consistent estimation of the order of mixture models. *Sankhyā, Series A*, 62:49–66, 2000.

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

[22] Bruce G. Lindsay. Properties of the maximum likelihood estimator of a mixing distribution. In C.P. Taillie, G.P. Patil, and B.A. Baldessari, editors, *Statistical Distributions in Scientific Work*, volume 5, pages 95–109, Maryland, 1981. International Co-operative Publishing House.

[23] Bruce 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.

[24] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28:129–137, 1982. Originally a 1957 Bell Labs memorandum.

[25] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L.M. LeCam and J. Neyman, editors, *Proc. 5th Berkeley Symp. Math. Statist. Probab. 1965/66*, volume I, pages 281–297, Berkeley, 1967. Univ. of California Press.

[26] F.H.C. Marriott. Separating mixtures of normal distributions. *Biometrics*, 31:767–769, 1975.

[27] R. Nishi. Maximum likelihood principle and model selection when the true model is unspecified. *J. Multivariate Anal.*, 27:392–403, 1988.

[28] B. Charles Peters, Jr. and Homer 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.

[29] David Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9:135–140, 1981.

[30] David Pollard. Quantization and the method of k-means. *IEEE Trans. Inf. Theory*, 28:199–205, 1982.

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

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

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

[34] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Statist. Soc., Series B*, 63:411–423, 2001.

[35] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.