Ann Inst Stat Math (2015) 67:745–772
DOI 10.1007/s10463-014-0474-9

# Empirical identifiability in finite mixture models

**Daeyoung Kim · Bruce G. Lindsay**

Received: 27 November 2012 / Revised: 4 May 2014 / Published online: 31 July 2014


**Abstract** Although the parameters in a finite mixture model are unidentifiable, there is a form of local identifiability guaranteeing the existence of the identifiable parameter regions. To verify its existence, practitioners use the Fisher information on the estimated parameters. However, there exist model/data situations where local identifiability based on Fisher information does not correspond to that based on the likelihood. In this paper, we propose a method to empirically measure degree of local identifiability on the estimated parameters, *empirical identifiability*, based on one's ability to construct an identifiable likelihood set. From a detailed topological study of the likelihood region, we show that for any given data set and mixture model, there typically exists limited range of confidence levels where the likelihood region has a natural partition into identifiable subsets. At confidence levels that are too high, there is no natural way to use the likelihood to resolve the identifiability problem.

**Keywords** Asymptotic identifiability · Finite mixture model · Local identifiability · Likelihood topology · Nonidentifiability

## 1 Introduction

Parameter identifiability is very important if one wishes to make inferences in a statistical model. In some models, identifiability on the parameters does not exist formally,

D. Kim (✉)
Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, USA
e-mail: daeyoung@math.umass.edu

B. G. Lindsay
Department of Statistics, Penn State University, University Park, PA 16802, USA

although it might exist in some practical sense. In this paper, we will show how one can define identifiability on the estimated parameters and empirically measure it, based on one's ability to construct a reasonable likelihood confidence set. By a reasonable likelihood confidence set we mean a *locally identifiable* subset of the parameter space where every parameter in the constructed set generates a unique distribution. The finite mixture model is a perfect place to illustrate our methodologies.

There is no identifiability of the parameters in a finite mixture model, at least in a formal sense, due to the two types of nonidentifiability, *labeling nonidentifiability* (Redner and Walker 1984) and *degenerate nonidentifiability* (Crawford 1994; Lindsay 1995, p. 74). Nevertheless, when one bases inferences on the likelihood function, many statistical analyses have been carried out treating the parameters as identifiable. This is possible because the point estimators for the parameters can be defined uniquely by the output of the likelihood maximization. Moreover, there is a form of *asymptotic identifiability* which one can appeal to for the inference on identifiable parameters as the sample size increases (Redner and Walker 1984). Indeed, this theory becomes a justification for using the Fisher information matrix and the corresponding Wald sets for further inferences.

Asymptotic identifiability is related to *local identifiability* (Rothenberg 1971). Local identifiability means that even though there are more than one parameter values in the whole parameter space whose probability distributions are the same to the sample, one can still find an open neighborhood of the parameter such that every parameter in that neighborhood generates a unique distribution. Thus if we can be relatively certain that the estimated parameters lie in such a locally identifiable region, we can appeal to asymptotic identifiability. Since it is computationally difficult to determine a locally identifiable region in the whole parameter space, we often check if the Fisher information on the estimated parameters is positive definite to verify the existence of local identifiability on the estimated parameter (Rothenberg 1971; Goodman 1974; Huang and Bandeen-Roche 2004).

As we shall see later, however, there are data/model situations where local identifiability based on the positive definiteness of the Fisher information may not correspond to an identifiable likelihood confidence set for the parameter estimates. Moreover, if we try to construct likelihood regions for the estimated parameters, we find that there exists a limited range of confidence levels where one can create a locally identifiable likelihood set. If one chooses too high confidence level for the given data set, one can no longer create an identifiable confidence region using the likelihood and thus asymptotic identifiability does not hold. We interpret the bound on confidence leading to an identifiable likelihood confidence set as *empirical identifiability* that empirically measures degree of local identifiability with respect to the estimated parameters in a given data set. If the empirical identifiability is high enough, say 95 %, we are relatively confident that the sample contains enough information to confer meaningful interpretation on the estimated parameters.

The route to construct the likelihood regions that are consistent with asymptotic identifiability is not elementary but in many situations our methods are easy to implement. The idea behind our proposal is a topological decomposition of the likelihood confidence region using an *identifiable partition*. By identifiable partition we mean that the likelihood region consists of disjoint subsets, each of which is a connected and

identifiable subset. Although there are multiple such subsets, they are merely relabeled images of each other, and so any one of them can be chosen for equivalent inference on the identifiable parameters in them.

There are a few important aspects of the proposed methods in this paper. First, our constructed regions based on the likelihood topology share the same invariance property that standard likelihood regions have. That is, a smooth change in the parametrization yields an exactly equivalent change in the regions. Second, the empirical identifiability on the estimated parameters is obtained using the likelihood topology in the full-dimensional parameter space and so one can transmit information about the empirical identifiability to any lower dimensional parameters of interest. In Sect. 7 we will illustrate how to utilize the value of the empirical identifiability generated by the full dimensional likelihood topology for an important practical problem in using the likelihood, namely the visualization of the likelihood-based profile sets for the parameters of interest. Third, the proposed method designed to identify a range of confidence levels that guarantees identifiable partition in the likelihood can be implemented using the EM algorithm.

The structure of the paper is as follows. Section 2 reviews the nonidentifiabilites of the parameters in a finite mixture model. In Sect. 3, we formally introduce the *identifiable partition* and the *unimodal partition*, induced by the mixture likelihood. We then illustrate, using two simulated data sets, the identifiability issues that occur when one constructs the likelihood regions for the mixture parameters. Section 4 studies how the topology of the mixture likelihood determines existence of an identifiable unimodal partition. We then create data analytic tools to assess whether such a partition exists. In Sect. 5 we study the relationship between the parameters and the partition. We first show that the mixing weights can be ignored in creating an identifiable partition in the likelihood. We then show that when the component parameters are univariate, the partition yields order restricted confidence regions. We also show that no such simple rule exists for multivariate component parameters. For this case, we create a numerical diagnostic for the presence of the partition. In Sect. 6, we carry out a simulation study to evaluate the performance of our proposed diagnostic for a case of multivariate component parameter. In Sect. 7, we use four examples to show application of our proposed methods.

## 2 Background on nonidentifiabilities in finite mixture models

Suppose that $n$ observations $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ are randomly drawn from the $K$ component mixture density with mixing weights $\pi = (\pi_1, \ldots, \pi_K)$ ($0 \le \pi_j \le 1$ and $\sum_{j=1}^{K} \pi_j = 1$), component parameters $\xi = (\xi_1, \ldots, \xi_K)$ and a structural parameter $\omega$ for the density function $f$:

$$p(y \mid \theta) = \sum_{j=1}^{K} \pi_j f(y; \xi_j, \omega), \tag{1}$$

$$\theta = \left[ \begin{pmatrix} \pi_1 \\ \xi_1 \\ \omega \end{pmatrix}, \ldots, \begin{pmatrix} \pi_j \\ \xi_j \\ \omega \end{pmatrix}, \ldots, \begin{pmatrix} \pi_K \\ \xi_K \\ \omega \end{pmatrix} \right]. \tag{2}$$

Note that $\theta$ represents the set of parameters in the parameter space of dimension $p$, denoted by $\Omega$, which is the full product space (the simplex of $\pi_j$, and the cross product space of $\xi_j$ and $\omega$). One can associate $\theta$ in Eq. (2) with the *mixing distribution*, denoted by $Q^{<\theta>}$, which is the discrete distribution function with mass $\pi_j$ at $\xi_j : p(y \mid \theta) = \int f(y; \xi, \omega) dQ^{<\theta>}(\xi)$. The mixing distribution $Q^{<\theta>}$ is often identifiable (Teicher 1960, 1963; Yakowitz and Spragins 1968; Lindsay 1995). In this paper, we focus on the cases where $Q^{<\theta>}$ associated with $\theta$ in Eq. (1) is identifiable and the component density $f(y; \xi_j, \omega)$ indexed by the parameters $\xi_j$ and $\omega$ is also identifiable, but there exist serious identifiability problems if one is interested in $\theta$ and its interpretation.

Given $n$ independent data generated from Eq. (1), one can construct the mixture likelihood of a $p$-dimensional parameter $\theta$,

$$L(\theta) = \prod_{i=1}^{n} p(y_i \mid \theta). \tag{3}$$

We will denote the true parameter by $\theta_\tau$, and the Maximum Likelihood Estimator (MLE) for the parameter maximizing $L(\theta)$ in Eq. (3) by $\hat{\theta}$. We assume that the mixture likelihood is bounded.

In such a setting, the likelihood confidence region for $\theta$ can be written as

$$C_c^{\mathrm{LR}} = \{\theta : L(\theta) \geq c\}, \tag{4}$$

where $c$ is a tuning parameter that changes the size of the likelihood confidence set. We call $C_c^{\mathrm{LR}}$ in Eq. (4) as the *elevation c likelihood region*. For purposes of illustrating the methods in this paper, we use the limiting distribution of the likelihood ratio statistic, $T_1 = 2(\log L(\hat{\theta}) - \log L(\theta))$, to determine the value of $c : c = L(\hat{\theta})\mathrm{e}^{-q_{1-\alpha}/2}$ where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with $p$ degrees of freedom. Then $c$ is interpreted via the confidence level of $T_1$ in the $p$-dimensional parameter space which we denote by $\mathrm{Conf}_p(c)$. If one wishes to use elevations that provide more accurate confidence levels, one can do a parametric bootstrap adjustment (Efron and Tibshirani 1993; Davison and Hinkley 1997).

In our problem, interpretation of the likelihood confidence region presents some new challenges. There are two important nonidentifiabilities in $\theta$, the *labeling nonidentifiability* (Redner and Walker 1984) and the *degenerate nonidentifiability* (Crawford 1994; Lindsay 1995, p. 74). We describe these concepts next, and then return to their implications for confidence region estimation.

The degenerate nonidentifiability occurs at those parameters $\theta_0$ where $Q^{<\theta_0>}$ has fewer than $K$ mass points, so $\theta_0$ are non-identifiable in the $\theta$ space. We will call such $\theta_0$ a *degenerate* point. For example, the following three subsets of the boundary of the parameter space in a two-component mixture model without structural parameter, $\theta_0 = \left[\binom{\pi_1}{\xi_0}, \binom{1-\pi_1}{\xi_0}\right]$, $\left[\binom{0}{\xi}, \binom{1}{\xi_0}\right]$ and $\left[\binom{1}{\xi_0}, \binom{0}{\xi}\right]$, all generate a one-component density with parameter $\xi_0$ for arbitrary $\xi$ and $\pi_1$, and thus $\xi$ and $\pi_1$ are not identifiable. In this case $Q^{<\theta_0>}$ has just one support point. In this paper we assume that the number of components is at most $K$ (i.e., the support of $Q^{<\theta>}$ has at most $K$ elements). As illustrated in Sect. 3.3, we still face the problem that parameter values near this bound-

ary will create challenges to asymptotic inference methods. Moreover, the parameter points where $\xi_1 = \xi_2$ are in the interior of the cross product space and give the same distribution when permuted, which is related to the next nonidentifiability.

The labeling nonidentifiability means that for any particular $\theta$ in Eq. (2), one can rearrange its columns in an arbitrary fashion without changing the mixture density in Eq. (1). In other words, if $\theta^\sigma$ is a copy of $\theta$ with columns permuted according to any permutation $\sigma$ of the identity permutation $(1, \ldots, K)$, then the mixture density and the likelihood are invariant : $p(y \mid \theta) = p(y \mid \theta^\sigma)$ and $L(\theta) = L(\theta^\sigma)$. For example, when $K = 2$ and there is no structural parameter, $\theta = \left[\binom{\pi_1}{\xi_1}, \binom{\pi_2}{\xi_2}\right] = \left[\binom{0.4}{1}, \binom{0.6}{2}\right]$ has the same distribution as $\theta^\sigma = \left[\binom{0.6}{2}, \binom{0.4}{1}\right]$, and so the model (and data) provide no information as to which column in $\theta$ should be called the first component and which the second component.

For this reason, we will say that the labels on $\theta$ are not identifiable. In particular, in a $K$ component mixture, there are $K!$ different true values corresponding to all possible column permutations of $\theta_\tau$. Moreover, if $\hat{\theta}$ is a mode of the mixture likelihood, the likelihood surface has (at least) $K!$ MLE modes, corresponding to the column permutations of $\hat{\theta}$. Since the modes of the mixture likelihood come in sets of $K!$ points, we will call each such set a *modal group*. If there is only one group, corresponding to the MLE mode, we will say that *there are no secondary modes*. Notice that when $c = L(\hat{\theta})$ in $C_c^{\mathrm{LR}}$ of Eq. (4), $C_c^{\mathrm{LR}}$ will contains exactly $K!$ parameter values, namely the $K!$ MLE modes.

## 3 Challenging issues in constructing the mixture likelihood region

Although there are two types of nonidentifiability in $\theta$ of Eq. (1), we can create identifiable parameters by restricting the model parameters to come from a subset of the parameter space that includes neither permuted parameters nor degenerate parameters.

**Definition 1** A subset $S$ of the parameter space $\Omega$ is said to be an identifiable subset if, for any parameter $\theta$ in $S$, there exists no parameter $\theta^{'}$ in $S$ such that $p(y \mid \theta) \equiv p(y \mid \theta^{'})$ where $\equiv$ means that densities equal for almost all $y$.

One method to create identifiable subsets is to use parameter constraints. For example, when the component parameter $\xi_j$ is univariate, the order-restricted subset based on $\xi_1 < \cdots < \xi_K$ is a commonly used candidate. In this paper, we will use the topology of the likelihood to create an identifiable region that can be used for labeled inference.

### 3.1 Asymptotic identifiability

Our chosen route to creating identifiable subsets for statistical inference is based on an asymptotic identifiability theory for the MLE found in Redner and Walker (1984). Let $\theta_\tau$ be any one of the $K!$ true values. Suppose that it is associated with Fisher information $\mathcal{I}$. Let $\hat{\theta}_n$ be an arbitrary element in the modal group of $K!$ MLE's for each $n$. Then there exists a way to choose a column permutation of $\hat{\theta}_n$, $\sigma_n$, such that $\hat{\theta}_n^{\sigma_n}$ is consistent for $\theta_\tau$, and

$$\sqrt{n}(\hat{\theta}_n^{\sigma_n} - \theta_\tau) \xrightarrow{D} N(0, \mathcal{I}^{-1}). \tag{5}$$

The asymptotic result described above seems to imply that for large samples one can treat labels as identified, and use standard asymptotic distribution theory. However, the existence of $\sigma_n$ does not say how one might determine it. We think it is helpful to pretend that one does not know $\sigma_n$, and consider the implications for the construction of confidence regions. To do so, we invert Eq. (5) around every possible permutation $\sigma$ of $\hat{\theta}_n$. There will then be $K!$ Wald ellipsoidal regions, one for each permutation $\sigma$, one surrounding each permuted mode $\hat{\theta}^\sigma$,

$$\{\theta : (\hat{\theta}^\sigma - \theta_\tau)^{\mathrm{T}} \mathcal{I}^\sigma (\hat{\theta}^\sigma - \theta_\tau) \leq w\}$$

where $w$ is chosen to achieve a desired asymptotic confidence region. We can think of the $K!$ ellipsoids as describing the most likely location of the true value $\theta_\tau$ as well as the $K! - 1$ equivalent representations $\theta_\tau^\sigma$. Note that these $K!$ Wald ellipsoids are all just permuted copies of each other. It follows that $\theta_\tau$ is in any one of the $K!$ ellipsoids if and only if all of the permuted versions of $\theta_\tau$ are found in corresponding permuted ellipsoids.

For sufficiently small $w$, the ellipsoids are separated from the each other and are identifiable subsets. This implies that the labels are meaningful within each ellipsoid. This is a very important point. One can then pick any one ellipsoid and use it for set estimation. All the other ellipsoidal regions are just permuted copies and so give the same basic inference about $\theta_\tau$, just with different column labels.

If we do choose one ellipsoid and redefine *coverage probability* to mean the probability the chosen ellipsoid covers one of the $K!$ true values $\theta_\tau^\sigma$, we know that the coverage probabilities are asymptotically correct. Our conclusion is that in an asymptotic sense we can select any one of the ellipsoids as a descriptor of the full confidence region, knowing that all other possible regions can be obtained by permutations of indices, also known as relabeling.

Using this setup, Wald methods can be applied to the mixture problems we are considering. However, one might prefer to use likelihood regions because Wald regions are based on local quadratic approximations to the logarithm of likelihood ratios and so they need not capture global behavior of likelihood regions. There is also a substantial literature that suggests that the Wald sets are inferior to the likelihood sets for finite samples (Cox and Hinkley 2002; Kalbfleisch and Prentice 1980; Meeker and Escobar 1995; Agresti 2002). We will illustrate several issues one faces in using Wald sets for the parameters of finite mixture models in Sect. 3.3.

### 3.2 Partition of mixture likelihood region

Suppose that the mixture likelihood $L(\theta)$ in Eq. (3) is bounded and there exists a modal group of $K!$ MLE modes maximizing $L(\theta)$. To assist in our description of the mixture likelihood region, we define the *modal region* determined by the elevation of interest $c$ in Eq. (4) and the MLE mode $\hat{\theta}$.

**Definition 2** The *modal region* for $\hat{\theta}$ at the elevation $c$, as written as $C_c(\hat{\theta})$, is the set of all $\theta$ that are connected to $\hat{\theta}$ by a continuous path contained in the likelihood region with the elevation $c$, $C_c^{\mathrm{LR}} = \{\theta : L(\theta) \geq c\}$ of Eq. (4).

Note that this definition implies that the modal region around another mode of the modal group $\hat{\theta}^\sigma$, $C_c(\hat{\theta}^\sigma)$, can be found by permuting all elements of the original modal region $C_c(\hat{\theta}) : C_c(\hat{\theta}^\sigma) = C_c^\sigma(\hat{\theta})$ for any permutation $\sigma$. The set $C_c(\hat{\theta})$ and its permutations will play the role of the ellipsoids in the Wald analysis. Taylor expansion of the likelihood shows that the modal regions will be shaped elliptically for elevations $c$ near $L(\hat{\theta})$, but we will see that shapes deviate sharply from elliptical as the elevation $c$ is lowered.

The definition of the modal region given above leads to the definition of partition of the likelihood region:

**Definition 3** We say that there exists an *identifiable partition* in the elevation $c$ likelihood region, $C_c^{\mathrm{LR}}$, of Eq. (4) if there exist $K!$ modal regions $C_c(\hat{\theta}^\sigma)$, each for one of $K!$ MLE modes $\hat{\theta}^\sigma$, that satisfy the three properties,

(P1) they are disjoint,
(P2) each is an identifiable subset,
(P3) their union is equal to $C_c^{\mathrm{LR}}$.

In addition, we say that $C_c^{\mathrm{LR}}$ has a *unimodal partition* if there exists an identifiable partition that satisfies the additional property (P4).

(P4) each modal region $C_c(\hat{\theta}^\sigma)$ is a connected set containing just a single mode, $\hat{\theta}^\sigma$.

Existence of a unimodal partition is desirable because it is consistent with the large sample theory of the likelihood-based inference. Moreover, in a practical sense, the existence of such partition helps practitioners have a simple explanation for the data. We will focus our analysis on understanding conditions under which a unimodal partition exists in Sect. 4.

Note that if there exists a unimodal partition, then the labels on the parameters in each modal region are uniquely determined by its MLE mode and one can use any one of the $K!$ modal regions to describe $C_c^{\mathrm{LR}}$ with a locally identifiable set of parameters. Within each such modal region, there is no ambiguity about how to assign labels as any relabeled version of an element necessarily belongs in a permuted modal region. If we use the asymptotic distribution theory, any selected one of these regions has right confidence level provided that we define coverage probability to be the probability of covering one of the permutations of $\theta_\tau$.

However, if the elevation $c$ is much smaller than $L(\hat{\theta})$, the modal regions contain parameter values displaying the two types of nonidentifiability and thus the likelihood region no longer has an identifiable partition. We will illustrate these issues in the next subsection.

### 3.3 Two simulated examples

We next provide some data-based examples that will be useful to motivate our theoretical developments in the following sections. We consider construction of

the likelihood and Wald regions in a two-component normal mixture model with equal variances, $\theta_\tau = \left[\begin{pmatrix}\pi_1\\ \xi_1\\ \omega\end{pmatrix}, \begin{pmatrix}\pi_2\\ \xi_2\\ \omega\end{pmatrix}\right]$. We simulated 500 observations from $\theta = \left[\begin{pmatrix}0.4\\ -1\\ 1\end{pmatrix}, \begin{pmatrix}0.6\\ 1\\ 1\end{pmatrix}\right]$, and obtained the MLE for $\theta$ using the expectation–maximization(EM) algorithm (Dempster et al. 1977). The estimated MLE had parameters $\hat{\theta} = \left[\begin{pmatrix}0.588\\ 0.948\\ 0.89\end{pmatrix}, \begin{pmatrix}0.411\\ -1.003\\ 0.89\end{pmatrix}\right]$. Due to the labeling nonidentifiability, $\theta_\tau^\sigma = \left[\begin{pmatrix}0.6\\ 1\\ 1\end{pmatrix}, \begin{pmatrix}0.4\\ -1\\ 1\end{pmatrix}\right]$ and $\hat{\theta}^\sigma = \left[\begin{pmatrix}0.411\\ -1.003\\ 0.89\end{pmatrix}, \begin{pmatrix}0.588\\ 0.948\\ 0.89\end{pmatrix}\right]$ have the same density and likelihood as $\theta_\tau$ and $\hat{\theta}$, respectively, where $\theta^\sigma$ is the column permutation of $\theta$.

This means that if one wishes to construct the likelihood region for the mixture parameter, there should be two subsets, one around each MLE mode. Although the full parameter space is of dimension $p = 4$, we can partially view the full likelihood structure through 2-dimensional profile likelihood sets and Wald sets for $(\pi_1, \xi_1)$ (Meeker and Escobar 1995). This is possible because if one can see the two separated sets in the profiles at the targeted elevation $c$, then the full-dimensional likelihood region is also separated into two sets at the same elevation.

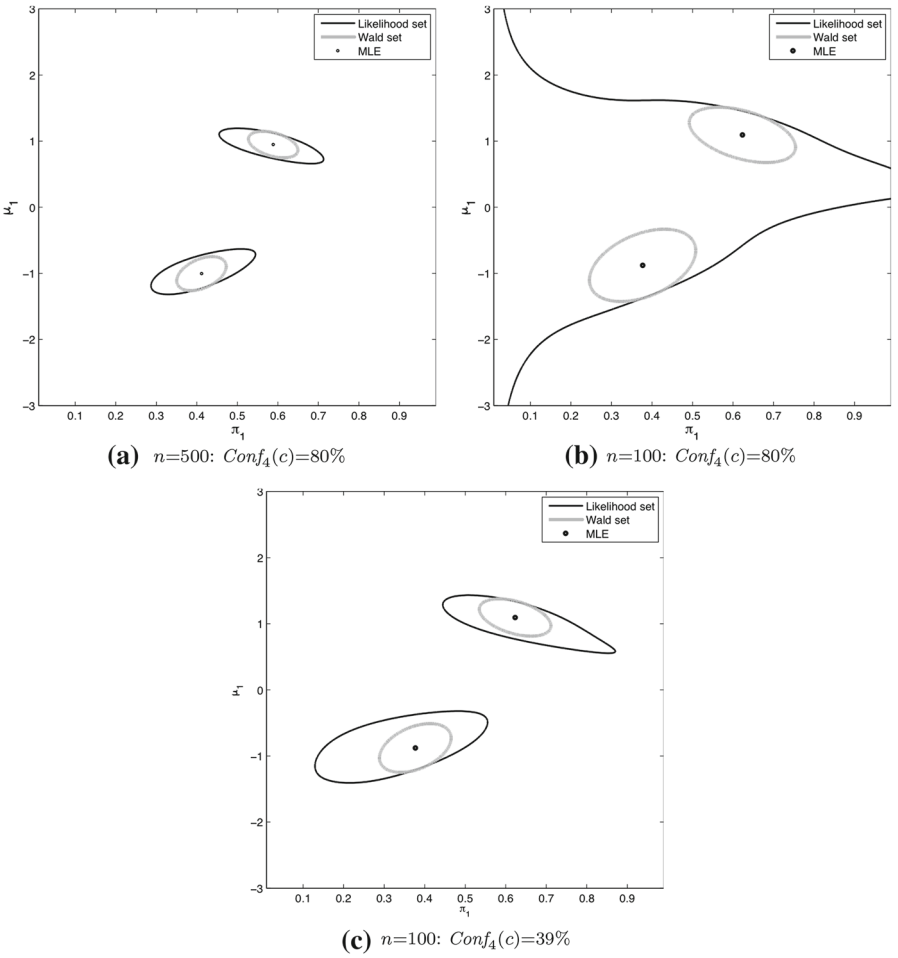
Figure 1a shows the numerical profile likelihood (black) and Wald (gray) contours for $(\pi_1, \xi_1)$ with the elevation $c$ corresponding to $\mathrm{Conf}_4(c) = 80\,\%$, the 80 % confidence level in the full-dimensional parameter space. Here we can see that both likelihood and Wald regions corresponding to the MLE and its permutation have appeared in the profile plot: for the given MLE $\hat{\theta}$, one mode had $\binom{\pi_1}{\xi_1} = \binom{0.588}{0.948}$ and $\hat{\theta}^\sigma$ had $\binom{\pi_1}{\xi_1} = \binom{0.411}{-1.003}$.

We observe from Fig. 1a that asymptotic identifiability at $n = 500$ appears appropriate. This is because the likelihood region has a unimodal partition, say $C_1$(upper region) and $C_2$ (lower region), one for each of the two modes of the likelihood. Thus, one can use $C_1$ and $C_2$ together to describe the two-mode profile likelihood region for $(\pi_1, \xi_1)$. Or one can pick labels, for example calling $C_1$ the likelihood set for the first component $(\pi_1, \xi_1)$ and $C_2$ the set for the second component $(\pi_2, \xi_2)$.

Note that results of the Wald regions are also consistent with those of asymptotic identifiability, as there were two identifiable and disjoint subsets, one for each of the two modes. However, the subsets in the Wald regions appear to be much smaller than those generated by the likelihood, even though their orientation matches up and their shape seems to be similar to an ellipse.

Instead of using the likelihood region, one might force identifiability by constraining the parameter space. In this example, using $\xi_1 < \xi_2$ would give the same partition as the likelihood. On the other hand, using $\pi_1 < 0.5 < \pi_2$ would not give a partition into identifiable subsets consistent with those induced by the likelihood. The theory behind this will be in Sect. 5.

If we reduce our sample size from $n = 500$ to $n = 100$ and hold the confidence level fixed, however, the structure of the likelihood dramatically changes. To illustrate this issue, we first simulated 100 observations from the same simulation model with the same true value $\theta_\tau$ used above, and obtained the MLE for $\theta$. Note that the estimated MLE was $\hat{\theta} = \left[\begin{pmatrix}0.623\\ 1.094\\ 0.83\end{pmatrix}, \begin{pmatrix}0.377\\ -0.879\\ 0.83\end{pmatrix}\right]$. We then constructed the profile likelihood (black) set for $(\pi_1, \xi_1)$ with the elevation $c$ corresponding to $\mathrm{Conf}_4(c) = 80\,\%$. From




**(a)** $n$=500: $Conf_4(c)$=80%

**(b)** $n$=100: $Conf_4(c)$=80%

**(c)** $n$=100: $Conf_4(c)$=39%

**Fig. 1** **a**–**c** are numerical profile likelihood (*black*) and Wald (*gray*) *contours* for $(\pi_1, \xi_1)$

Fig. 1b, we can clearly see that there is no sign of two separable subsets: the two modes at $\binom{0.623}{1.094}$ and $\binom{0.377}{-0.879}$ are both in one connected region. This region also contains degenerate parameter values (e.g., $\pi_1 = 0$), so the likelihood does not provide a way to partition this region into two identifiable subsets, one for each mode. In particular, using order constraints such as $\pi_1 < 0.5$ and $\xi_1 < \xi_2$ appears to create partitions that have no relationship with the structure of the likelihood. We also observe that the union of Wald regions (gray) gives misleading information on the structure of the likelihood region in the sense that the two subsets in the Wald regions are still separated, unlike the likelihood regions.

Using the same likelihood, one can also create an identifiable partition by decreasing the confidence level (i.e., increasing the elevation) of interest. Figure 1c shows the numerical profile likelihood (black) and Wald (gray) contour for $(\pi_1, \xi_1)$ with the elevation $c$ corresponding to $\text{Conf}_4(c) = 39\ \%$ when $n = 100$. We observe that the

profile, and hence the full likelihood regions, again have separable subsets, one for each MLE mode.

These two examples lead to several important observations. First, asymptotic identifiability may or may not be relevant in a finite sample. In these examples, we simulated data where the number of components $K$ was fixed and known, so technically we were not in a degenerate situation. However, the results based on the likelihood regions often bore little resemblance to those predicted by asymptotic identifiability. Second, the structure of the likelihood regions clearly depended on the elevation of interest. An identifiable unimodal partition might not exist at a chosen confidence level, but it is almost always possible to find a level small enough that such a partition exists. We also observed the different roles that the mixing weights and component parameters seemed to play in determining the identifiable subsets. These observations become the basis for our theory in Sect. 5.

## 4 Landmark elevations and topology of likelihood regions

In Sect. 3, we illustrated that the existence of an ideal partition at any particular elevation/confidence level depends on the topology of the mixture likelihood surface and on the elevation/confidence level chosen. In this section, we identify the landmark elevations that are critical in determining the existence of the unimodal partition.

For regularity we will assume that the likelihood for the parameter $\theta$ of a $K$ component mixture model in Eq. (1) is bounded, twice continuously differentiable and has no critical points with singular Hessians on the domain $\{\theta : L(\theta) > \psi_{\text{deg}}\}$ where $\psi_{\text{deg}}$ is the highest elevation in the degenerate class of parameters (i.e., the likelihood of the MLE for a $(K-1)$ component mixture model). The degenerate class of parameters introduces some pathologies into the topological structure of the regions, as will be clear shortly. For this reason, we will restrict attention to the elevations $c$ above $\psi_{\text{deg}}$. This assumption is important because one can appeal to Morse theory on that domain (Matsumoto 2002), which relates the topology of a manifold to critical points of functions. Among other things, this assumption has a few important implications. First, the mixture likelihood $L(\theta)$ on the domain $\{\theta : L(\theta) > \psi_{\text{deg}}\}$ has only finitely many critical points and these critical points are isolated. Second, the topology of $C_c^{\text{LR}}$ does not change except when $c$ passes the elevation of a critical point and thus we can focus $K!$ MLE modes, secondary modes and degeneracy in the likelihood to get information about the topology of the domain $\{\theta : L(\theta) > \psi_{\text{deg}}\}$. Last, the topology of $C_c^{\text{LR}}$ is invariant under smooth parameter transformations, and so our method will yield invariant confidence regions.

### 4.1 Existence of a unimodal partition

There are a number of topological landmarks whose elevations play an important role in the theory. Let $\psi_{\text{MLE}}$ be the elevation of the MLE modes for a $K$ component mixture model in Eq. (1) and let $\psi_{\text{deg}}$ be the elevation of the MLE for a $(K-1)$ component mixture model, as we defined above. If there exists a mode next highest to the MLE between $\psi_{\text{MLE}}$ and $\psi_{\text{deg}}$, we let its elevation be $\psi_{\text{2nd}}$.

Suppose that, at elevation $c$, the intersection of $C_c(\hat{\theta})$ and $C_c(\hat{\theta}^\sigma)$ is non-empty for some permutation $\sigma$ and therefore the two sets are equal. Then there exists a continuous path that has the following characteristics

1) it lies entirely in the elevation $c$ likelihood region, $C_c^{\mathrm{LR}}$ in Eq. (4),
2) it contains the saddle point connecting the two MLE modes, $\hat{\theta}$ and $\hat{\theta}^\sigma$,
3) among all paths satisfying 1) and 2), it maximizes the minimal elevation attained by the likelihood along the path.

We will say such a path is *maximin*, and denote its minimal elevation by $\psi_{\mathrm{mm}}$. This is also the elevation of the saddle point connecting the two modal regions.

We can now describe quite precisely the elevations that will give us a unimodal partition of the likelihood.

**Theorem 1** *A unimodal partition exists at elevation c if and only if* $\max\{\psi_{\mathrm{mm}}, \psi_{2\mathrm{nd}}\} < c \leq \psi_{\mathrm{MLE}}$.

*Proof* See the Appendix. □

From the proof of Theorem 1, we also have three lessons helpful for further analysis.

**Corollary 1** *An identifiable partition in the likelihood cannot exist for elevations c below* $\psi_{\mathrm{deg}}$.

**Corollary 2** *The likelihood region* $C_c^{\mathrm{LR}}$ *cannot display a unimodal partition for elevations* $c \leq \psi_{\mathrm{fail}} = \max\{\psi_{\mathrm{deg}}, \psi_{2\mathrm{nd}}\}$.

**Corollary 3** *The modal regions* $C_c(\hat{\theta})$ *are identifiable subsets if they are disjoint.*

A standard mixture analysis involves a search for the local maxima of the likelihood when one uses the EM algorithm for finding the MLE and so $\psi_{2\mathrm{nd}}$ in Theorem 1 is often available (Lindsay 1995; McLachlan and Peel 2000). Note that $\psi_{\mathrm{deg}}$ in Corollary 1 and 2 can also be easily computed via the EM algorithm because it corresponds to the likelihood of the MLE for $(K-1)$ component mixture model.

However, it is not standard to search for $\psi_{\mathrm{mm}}$ and so Theorem 1 is not directly useful in a mixture likelihood analysis. Since it would be a major undertaking to add such a step to a conventional mixture analysis, we will develop a substitute method that is not as precise, but does guarantee the existence of a unimodal partition.

Our first lesson is that every hyperplane that separates $\hat{\theta}$ from $\hat{\theta}^\sigma$ can be used to gain information about $\psi_{\mathrm{mm}}$ in the maximin path. Consider a hyperplane that separates the two MLE modes, $\hat{\theta}$ and $\hat{\theta}^\sigma$,

$$H_{r,a} = \{\theta \in \Omega : r^{\mathrm{T}}\theta = a\} \tag{6}$$

where $r \in \mathbb{R}^p$ is a nonzero vector and $a \in \mathbb{R}$. Let $\psi_{\mathrm{hyp}}$ be the largest likelihood value attained in the hyperplane of Eq. (6).

**Theorem 2** *For any separating hyperplane in Eq. (6),* $\psi_{\mathrm{hyp}} \geq \psi_{\mathrm{mm}}$.

*Proof* See the Appendix. □

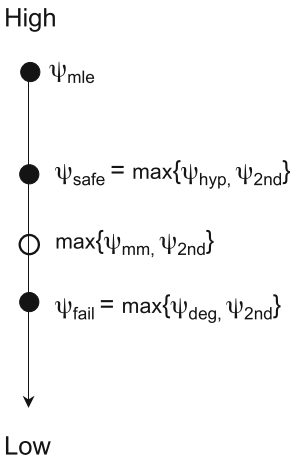



**Fig. 2** Landmark elevations for unimodal partition in a $K$ component mixture model: $\psi_{\text{MLE}}$ is the elevation of the MLE modes for a $K$ component mixture model, $\psi_{\text{2nd}}$ is the elevation of the secondary mode next highest to the MLE modes, $\psi_{\text{hyp}}$ is the highest elevation in the hyperplane of Eq. (6), $\psi_{\text{mm}}$ is the minimum elevation of the maximin path, and $\psi_{\text{deg}}$ is the elevation of the MLE for a $(K-1)$ component mixture model

We now can describe how one can use the value of $\psi_{\text{hyp}}$ in a data analysis to create a safe elevation for the construction of a unimodal partition.

**Corollary 4** *A unimodal partition is sure to exist if* $\psi_{\text{safe}} = \max\{\psi_{\text{hyp}}, \psi_{\text{2nd}}\} < c \leq \psi_{\text{MLE}}$.

In Sect. 4.2 we will discuss determination of $r$ and $a$ in the hyperplane of Eq. (6) for using Corollary 2 and 4 in an optimal fashion.

Figure 2 shows the landmark elevations that we have proposed in identifying a unimodal partition in a data analysis. If one determines the three landmark elevations, $\psi_{\text{MLE}}$, $\psi_{\text{safe}}$ in Corollary 4 and $\psi_{\text{fail}}$ in Corollary 2, one can gain considerable knowledge about whether or not one can create a unimodal partition of the likelihood region. One can also transform the computed $\psi_{\text{fail}}$ and $\psi_{\text{safe}}$ into confidence levels in the full-dimensional parameter space, $\text{Conf}_p(\psi_{\text{fail}})$ and $\text{Conf}_p(\psi_{\text{safe}})$, and thus employ the confidence bound on unimodal inference as the empirical identifiability in a given data set which measures degree of local identifiability at the MLE modes, $\hat{\theta}$. We will learn more about this in the following results. Note that all these landmark elevations can be easily computed using the EM algorithm.

*Remark 1* We think that using the likelihood with identifiable confidence sets is quite defensible based on the asymptotic theory, but only when the likelihood displays a unimodal partition. Suppose one wished to construct labeled regions at the elevations $c$ where unimodal partition does not hold. If there exists a secondary modal group with $\psi_{\text{2nd}}$ above $\psi_{\text{deg}}$, there will be a range of elevations where there exist more than $K!$ regions. If there was one secondary modal group, there would be $2K!$ unimodal regions. One could match regions into $K!$ identifiable pairs, but clearly this would be a manmade construction, not one dictated by the likelihood. Clearly the asymptotic theory is not working perfectly in this case.

For a case of the elevation $c$ below $\psi_{\text{deg}}$, one can imagine trying to create artificial partitions of the likelihood region into identifiable and unimodal subsets. However,

if the likelihood region contains any points displaying the either of the two nonidentifiabilities, the range of the parameters in this set is extremely wide. We conclude that creation of a partition of the likelihood region under these circumstances is quite artificial and misleading.

*Remark 2* It should be noted that the difference in log likelihood between $\psi_{\text{MLE}}$ and $\psi_{\text{deg}}$ corresponds to the likelihood ratio statistic for testing the null hypothesis of $K-1$ components versus the alternative of $K$ components (Titterington et al. 1985; Lindsay 1995). We have shown here that under the simplest likelihood topology where there is no secondary mode above $\psi_{\text{deg}}$, this difference will also determine whether or not we can construct labeled confidence regions at a specific confidence level. That is, if there is an "insignificant difference from degeneracy" then it is likely that one cannot produce labeled confidence regions with high confidence level. There is one subtlety in this relationship. The distribution of the hypothesis test statistic would be done under the null hypothesis (and is rather complicated; see Chen and Chen (2001); Liu and Shao (2003)). We are using the distribution of this statistic under the alternative hypothesis of K components, either asymptotic or simulated, to make judgments on confidence sets (One could not make inference under the null hypothesis because the parameters would be non-identifiable).

### 4.2 The default hyperplane

In Sect. 4.1, we obtained the two elevations for existence of a unimodal partition, the "sure-to-exist elevation", $\psi_{\text{safe}} = \max\{\psi_{\text{hyp}}, \psi_{\text{2nd}}\}$ in Corollary 4, and "sure-to-fail elevation", $\psi_{\text{fail}} = \max\{\psi_{\text{deg}}, \psi_{\text{2nd}}\}$ in Corollary 2. Since $\psi_{\text{hyp}}$ gives an upper bound on $\psi_{\text{mm}}$, we would like to choose our hyperplane(s) so as to make the constant $\psi_{\text{hyp}}$ as small as possible. We know in advance that the ideal value for $\psi_{\text{hyp}}$ is $\psi_{\text{deg}}$ as then $\psi_{\text{fail}}$ and $\psi_{\text{safe}}$ are equal (see Fig. 2), and there remains no ambiguity about the elevations for a unimodal partition. Our strategy is therefore to focus on using separating hyperplanes of Eq. (6) that contain one or more maximal degenerate parameter values. In such a hyperplane, the degenerate parameter values are local maxima. If they are global maxima, then we have the tight relationship, $\psi_{\text{fail}} = \psi_{\text{safe}}$.

For simplicity consider a two-component ($K = 2$) mixture,

$$\theta = (\pi, \xi_1, \xi_2, \omega)^{\text{T}} \tag{7}$$

where $\pi$ is a mixing weight for the component 1, $\xi_1$ and $\xi_2$ are the $m$-dimensional component parameter vectors for the component 1 and 2, respectively, and $\omega$ is a structural parameter vector. Note that the permuted version of $\theta$ is $\theta^{\sigma} = (1-\pi, \xi_2, \xi_1, \omega)^{\text{T}}$.

We then construct the following separating hyperplane, $H_{r,a} = \{\theta : r^{\text{T}}\theta = a\}$ with

$$r = (0, \hat{\xi}_1 - \hat{\xi}_2, -\hat{\xi}_1 + \hat{\xi}_2, 0)^{\text{T}} \quad \text{and } a = 0. \tag{8}$$

This hyperplane, equivalent to $\{\theta : (\hat{\xi}_1 - \hat{\xi}_2)^{\text{T}}(\xi_1 - \xi_2) = 0\}$, will be called the *default hyperplane*. It has the two interesting properties. First, the default hyperplane separates

$\hat{\theta}$ from $\hat{\theta}^{\sigma}$. That is, the MLE $\hat{\theta}$ and the permuted MLE $\hat{\theta}^{\sigma}$ are on opposite sides of the hyperplane, as $(\hat{\xi}_1 - \hat{\xi}_2)^{\mathrm{T}}(\xi_1 - \xi_2)$ is positive for the former and negative for the latter. Secondly, the default hyperplane allows $\pi$ and $\omega$ to take on arbitrary values. In other words, $\pi$ does not play a role in separating modes, which we will prove in Sect. 5.1, and we suspect $\omega$ will not, and so we leave them unconstrained in this hyperplane.

Therefore, the largest likelihood value attained in the default hyperplane of Eq. (8) is the value of $\psi_{\mathrm{hyp}}$ in the safe elevation $\psi_{\mathrm{safe}}$ of Corollary 4. We will describe how to determine $\psi_{\mathrm{hyp}}$ by modifying the EM algorithm in Sect. 5.4.

When the number of components are more than two, let us say three, there are six MLE modes and corresponding modal regions. There exist at least one path through a pair of the MLE modes whose lowest elevation is maximal among all paths. To determine this elevation one can apply the proposed hyperplane method to each pair of the MLE modes and use the highest elevation among all possible estimated elevations in the default hyperplanes as an estimate of $\psi_{\mathrm{hyp}}$.

## 5 Relationship between order restrictions and likelihood partitions

We next use our study of landmark elevations developed in Sect. 4 to create greater understanding of the role of the mixture parameters in the identifiable partition. We start with the mixing weight parameters.

### 5.1 Ignorable parameters

In the two simulated examples of Sect. 3.3 where there existed an identifiable unimodal partition, it seemed that the mixing weights were poorly related to the partitions. In fact, this is a theoretical feature of the weight parameters, as they are ignorable in the following sense.

**Definition 4** Suppose that the parameters $\theta$ in Eq. (1) partition into $(\gamma, \phi)$, and there exists a unique maximum over $\phi$ for each fixed $\gamma$ in the likelihood, say $\hat{\phi}(\gamma)$, such that $\hat{\phi}(\gamma)$ is continuous in $\gamma$. We will then say that $\phi$ is *ignorable for partitioning*.

**Theorem 3** *Assume that $\phi$ is ignorable for partitioning. If two modal regions in the mixture likelihood are connected in the $\gamma$ profile space at elevation $c$, they are connected in the full likelihood space at elevation $c$. It follows that modal regions form disconnected sets in the full parameter space if and only if they do so in the $\gamma$ profile space.*

*Proof* See the Appendix. □

If we let $\phi$ be the mixing weights $\pi = (\pi_1, \ldots, \pi_K)$ in the above theorem, then under identifiability conditions on $Q^{<\theta>}$ (the mixing distribution associated with $\theta$ in Eq. (1)), the negative Hessian matrix is positive definite for every fixed value of the other parameters, $\xi = (\xi_1, \ldots, \xi_K)$ and $\omega$. It follows that the mixture likelihood is strictly log-concave in $\pi$ and so has a unique maximum. Under regularity conditions on $\xi$ and $\omega$, the maximizing $\hat{\pi}(\xi, \omega)$ will be continuous in its arguments. We can conclude,

by Theorem 3, that the identifiable confidence sets are determined by the component parameters $\xi$ and the structural parameters $\omega$, not the weights $\pi$. This result does not depend on the component models that are used in the inference and so we have:

**Corollary 5** *Under regularity of the model, existence of an identifiable partition in the full-dimensional likelihood regions implies existence of an identifiable partition in the profiles of the component parameters $\xi$ and the structural parameters $\omega$, regardless of parameter dimension and the number of component.*

*Remark 3* If one wanted to apply Corollary 5 to a posterior density, one would have to show that the posterior density is log concave in $\pi$ for a given prior distribution on $\theta$. For example, Corollary 5 would hold for a flat Dirichlet prior.

### 5.2 Unimodal partition in a case of univariate component parameter

We observed from the two simulated examples of Sect. 3.3 that when the component parameter $\xi_j$ was univariate and a unimodal partition existed, the two identifiable modal regions induced by the likelihood were identical to those determined by the order restriction on $\xi = (\xi_1, \xi_2)$. In this subsection, we consider the simplifications of the likelihood topology that are possible when the component parameter $\xi_j$ is univariate.

For any specific model/data situation where there are no modes between $\psi_{\text{deg}}$ and $\psi_{\text{MLE}}$ in the univariate case, we have very neat necessary and sufficient conditions for a unimodal partition.

**Theorem 4** *Suppose the component parameter $\xi_j$ is univariate in the $K$ component mixture model. Then*

1) *$\psi_{\text{hyp}} = \psi_{\text{deg}}$ in the default hyperplane of Eq. (8) and every continuous maximin path connecting two MLE modes must pass through the degenerate set,*
2) *when no secondary modes exist, we have a unimodal partition for all elevations $c > \psi_{\text{deg}}$,*
3) *the labels on the parameters can be determined by order restriction on $\xi$ when a unimodal partition exists at elevation $c$.*

*Proof* The proof is based on the default hyperplane. See the Appendix. □

Note Theorem 4 shows that if $\xi_j$ is univariate, both $\omega$ and $\pi$ are ignorable when determining if there exists a unimodal partition, and the unimodal partition is sure to exist if and only if $\psi_{\text{safe}} = \psi_{\text{fail}} = \max\{\psi_{\text{deg}}, \psi_{\text{2nd}}\} < c \leq \psi_{\text{MLE}}$ (see Fig. 2).

Referring back to the two simulated examples in Sect. 3.3, the component parameter was univariate. When we used the EM algorithm for finding the MLE in both examples, multiple starting values were employed and there was no secondary mode. Therefore, by Theorem 4, $\psi_{\text{deg}}$ is the safe elevation for existence of a unimodal partition in both examples. Note that $\psi_{\text{deg}}$ here is the maximum of the likelihood among the degenerate parameters which here correspond to having a single component. In other words, $\text{Conf}_4(\psi_{\text{safe}}) = \text{Conf}_4(\psi_{\text{deg}}) = 98.6\,\%$ at $n = 500$ and $\text{Conf}_4(\psi_{\text{safe}}) = \text{Conf}_4(\psi_{\text{deg}}) = 50.5\,\%$ at $n = 100$ represent the exact upper bound on the confidence levels one can use when constructing an identifiable unimodal confidence set.

### 5.3 Unimodal partition in a case of multivariate component parameter

In this subsection we show by an example that when $\xi_j$ is multivariate, one cannot use Theorem 4 to determine the proper elevation of a unimodal partition. While this example is simple and artificial, it nevertheless provides us with some insights into more complicated problems as well.

When $\xi_j$ is multivariate, it is technically feasible that there exist continuous maximin paths that connect MLE modal regions while staying at higher elevations than the degenerate parameter set, so that $\psi_{\text{deg}} < \psi_{\text{mm}}$. In the univariate examples, $\psi_{\text{hyp}}$ was $\psi_{\text{deg}}$, and so there was not really a new landmark quantity to calculate. It now becomes a crucial diagnostic.

We illustrate this point with the following example. The data set will be 99 data points equally spaced around the circle of radius 2 around the origin, $\{2\cos(2k\pi/99), 2\sin(2k\pi/99) : k = 1, \ldots, 99\}$. We then fit them with a two-component bivariate normal mixture model with a fixed weight $\pi = .5$ and a fixed common covariance matrix $0.5\mathcal{J}$ where $\mathcal{J}$ is an identity matrix. This model has certain equivariance properties that make the analysis simpler. In particular, a rotation about the origin of a random variable generated from this model gives a new random variable that is also from the two-component mixture model, but now with new mean parameters that have also been rotated by the same angle about the origin. Since the 99 points in the data set stay unchanged when they are simultaneously rotated through angles of $2k\pi/99$, it follows that the critical points of the likelihood will come in sets of 99, corresponding to the same rotations of the means of any one MLE solution element about the origin. Moreover, all these critical points will have the same elevation. That is, unless the degenerate point $(0, 0)$ is the MLE, there are at least 99 elements of the MLE modal group.

Intuitively, for any one mode, the MLE means $\hat{\xi}_1$ and $\hat{\xi}_2$ will be diametrically opposite to each other due to the symmetry of the problem. We numerically verified this, finding that when one component of the MLE set was $\hat{\xi}_1 = (-1.25, 0.07)$, the second was $\hat{\xi}_2 = (1.25, -0.07)$. We also constructed a numerical profile confidence set for $\xi_1$ with the elevation $c$ corresponding to $\text{Conf}_2(c) = 0.1\ \%$, the black circle, and 99 %, the region between the blue circles (see Fig. 3). From the 0.1 % profile set we can see that the 99 MLE solutions in $\xi_1$ are on a circular ridge of high likelihood. We also observe that the 99 % profile confidence set excludes the degenerate solution, which has mass 1 at $\xi_1 = (0, 0)$. This means that the lowest elevation, $\psi_{\text{mm}}$, of the continuous, four-dimensional maximin path that rotates the pair $(\xi_1, \xi_2)$ around the circle is substantially greater than $\psi_{\text{deg}}$, so that the MLEs do not generate identifiable subsets at any elevation below $\psi_{\text{mm}}$. Thus, Theorem 4 cannot hold in this multivariate example.

*Remark 4* The preceding example could be viewed as artificial. One of our reviewers raised the question of whether the third part of Theorem 4 would hold under some weak assumptions, such as having a likelihood function with just $K!$ MLE's. That is, "can the labels on the parameters be determined by an appropriate order on the coordinates in $\xi$ when there are just $K!$ MLE's and a unimodal partition exists at elevation $c$?". This is a delicate geometric question. In essence we need to find a function of the parameters that is certain to separate the different modal regions. Our example shows

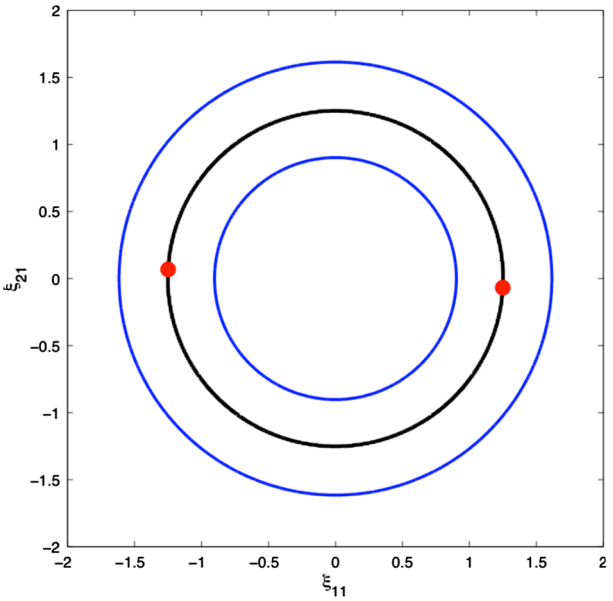



**Fig. 3** Numerical profile set for $\xi_1 = (\xi_{11}, \xi_{21})$ with 0.1 % (*black circle*) and 99 % (the region between two *blue circles*) : *red dots* represent one element in the MLE set

the dangers of assuming that the separation only depends on the regions of degenerate solutions. Since there is, as yet, no theory guaranteeing separation based on some set of order restrictions, we would advise the use of diagnostic plots such as found in Yao and Lindsay (2009). They devised linear and quadratic discriminant functions to identify separation of posterior modes in a Bayesian MCMC plot. If one could find a discriminant function that always works perfectly then one would, in effect, have identified the "parameter restrictions" that give separation. However, the development of these tools in our likelihood context is beyond the scope of this paper.

### 5.4 Determining $\psi_{\text{hyp}}$

In Sect. 5.3, we showed by example that if the component parameter is multivariate, then the two modal regions could be connected with each other at the elevation $c$ above $\psi_{\text{deg}}$ (so that $\psi_{\text{mm}} > \psi_{\text{deg}}$), even when there exists no secondary mode of the likelihood.

To do mixture inference based on unimodal partition in a case of multivariate component parameter, thus, one needs to calculate $\psi_{\text{hyp}}$ of the safe elevation in Corollary 4, $\psi_{\text{safe}} = \max\{\psi_{\text{hyp}}, \psi_{\text{2nd}}\}$. We propose to do this using the restricted EM algorithm (Kim and Taylor 1995). This is because estimation of $\psi_{\text{hyp}}$ is equivalent to the maximization of the mixture likelihood under the linear restriction on $\theta$ in the default hyperplane of Eq. (8), $r^{\text{T}}\theta = a$ with $r = (0, \hat{\xi}_1 - \hat{\xi}_2, -\hat{\xi}_1 + \hat{\xi}_2, 0)^{\text{T}}$ and $a = 0$.

One needs to employ a strategy of multiple starting values for the restricted EM that starts well away from the degenerate set. Starting values near the degenerate

set will simply create paths back to the maximum degenerate point. To speed up the calculation of $\psi_{\text{hyp}}$, we propose using the predicted final likelihood using an the Aitken acceleration (Böhning et al. 1994; Lindsay 1995). With this device one can produce a reasonable estimator of the final log likelihood of a EM sequence based on a smaller set of iterations. Note that we are not interested in the parameter estimates, just the maximal elevation attained. To ensure accuracy in the Aitken predictions, we will employ an Aitken acceleration-based stopping rule in the algorithm.

## 6 Simulation study

In this section, we now examine by simulation our landmark elevations proposed in Sect. 4 for constructing a unimodal partitions when the component parameter is multivariate. As a simulation model we consider a two component $m$-variate normal mixture model with equal covariance:

$$p(y \mid \theta) = \pi N_m(x; \xi_1, \Sigma) + (1 - \pi) N_m(x; \xi_2, \Sigma)$$

where $\xi_1 = (\xi_{11}, \ldots, \xi_{m1})$ and $\xi_2 = (\xi_{12}, \ldots, \xi_{m2})$.

In equal covariance case one can standardize the data vectors by $\Sigma^{-1/2}$ and turn to a case of an identity matrix, $\mathcal{J}$. So, we here set $\Sigma = \mathcal{J}$. In the simulation study, we consider four factors, dimension of data, sample size, mixing weight and separation of the components. Regarding the dimension of the data, two values are considered: $m = 2$ and 5. For the sample size we use the two levels of $n$: (100, 200) when $m = 2$, and (200, 400) when $m = 5$. We also consider two levels of mixing weight, $\pi = 0.5$ and 0.2.

As to separation of the components, we use invariance properties of the likelihood. That is, when the covariances are fixed to be identity matrices and the mixing weights are fixed, all matters is the Mahalanobis distance between the two-component mean vectors. For example, when $m = 2$ and $(\xi_1, \xi_2) = (\xi_{11}, \xi_{21}, \xi_{12}, \xi_{22})$, $(\xi_1, \xi_2) =$ (0,0,0,0) and (1,1,1,1) will provide the same properties as $(\xi_1, \xi_2) =$ (0,0,0,0) and (4,0,0,0). Thus, we use the first coordinate in the second component for component separation, for example, $(\xi_1, \xi_2) =$ (0,0,$d$,0) when $m = 2$. Regarding the value of $d$ we use (2, 3, 4) for $\pi = 0.5$, and (2.5, 3.75, 5) for $\pi = 0.2$. Note that they are obtained by setting the standard deviation of the mixing distribution with mass $\pi$ and $1 - \pi$ at two support points, $\xi_1$ and $\xi_2$, respectively, $\sqrt{\pi(1 - \pi)d^2}$ to be 1, 1.5 and 2 for $\pi$.

We simulate $R = 500$ replicate samples of size $n$ from the simulation model at each combination of $(m, \pi, d)$, and then estimate the MLE for $\theta$ via the EM algorithm. Note that 50 starting values are used for computing the MLE for $\theta$ at each simulated data set.

To identify a range of confidence level that displays unimodal likelihood regions, we first compute the lower and upper bound for the elevation of unimodal partition, $\psi_{\text{fail}}$ in Corollary 2 and $\psi_{\text{safe}}$ in Corollary 4. Then we obtain the confidence bound on unimodal inference by transforming the computed $\psi_{\text{fail}}$ and $\psi_{\text{safe}}$ into confidence levels in the full dimensional parameter space, $\text{Conf}_p(\psi_{\text{fail}})$ and $\text{Conf}_p(\psi_{\text{safe}})$.

**Table 1** Percentage of categorization with respect to 95 % unimodal partition: $m$=2 and $R = 500$

| $n$ | Mode | Category | $\pi = 0.5$ | | | $\pi = 0.2$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | $d = 2$ | $d = 3$ | $d = 4$ | $d = 2.5$ | $d = 3.75$ | $d = 5$ |
| 100 | Unimode | GOOD | 0.6 | 12.6 | 38.4 | 5.0 | 38.2 | 43.6 |
| | | BETWEEN | 0.4 | 1.6 | 0.2 | 1.6 | 0.4 | 0.0 |
| | | BAD | 14.2 | 13.8 | 0.4 | 15.6 | 1.0 | 0.0 |
| | Multimodes | GOOD | 1.0 | 19.0 | 59.0 | 6.6 | 51.4 | 56.4 |
| | | BETWEEN | 0.0 | 1.0 | 0.0 | 0.2 | 0.0 | 0.0 |
| | | BAD | 83.8 | 52.0 | 2.0 | 71.0 | 9.0 | 0.0 |
| | | GAP>10 % | 6.2 | 5.2 | 0.0 | 6.2 | 0.6 | 0.0 |
| 200 | Unimode | GOOD | 1.4 | 31.6 | 42.6 | 18.0 | 44.4 | 52.6 |
| | | BETWEEN | 0.4 | 1.4 | 0.0 | 1.6 | 0.0 | 0.0 |
| | | BAD | 13.8 | 1.6 | 0.0 | 8.0 | 0.0 | 0.0 |
| | Multimodes | GOOD | 2.2 | 52.8 | 57.4 | 26.6 | 55.6 | 47.4 |
| | | BETWEEN | 0.4 | 1.2 | 0.0 | 1.6 | 0.0 | 0.0 |
| | | BAD | 81.8 | 11.4 | 0.0 | 44.2 | 0.0 | 0.0 |
| | | GAP>10 % | 10.2 | 0.8 | 0.0 | 3.2 | 0.0 | 0.0 |

Given the calculated confidence bounds over 500 simulated data sets, we classify them into three categories, "GOOD", "BAD" and "BETWEEN", depending on whether or not one can do 95 % identifiable unimodal inference. Note that we use 95 % because it is the most commonly used confidence level.

GOOD: $\mathrm{Conf}_p(\psi_{\mathrm{safe}}) \geq 95\ \%$ so that unimodal partition exists in the 95 % likelihood region.

BAD: 95 % likelihood region does not have a unimodal partition. That is, $\mathrm{Conf}_p(\psi_{\mathrm{fail}}) < 95\ \%$.

BETWEEN: it is not clear whether or not 95 % labeled modal inference is possible, as $\mathrm{Conf}_p(\psi_{\mathrm{safe}}) \leq 95\ \% < \mathrm{Conf}_p(\psi_{\mathrm{fail}})$.

Note that our goal in creating $\psi_{\mathrm{hyp}}$ in $\psi_{\mathrm{safe}}$ was to create greater certainty about the existence of a unimodal partition. That is, we wanted the "BETWEEN" category to be small. As another measure of our success, we calculated the difference between $\mathrm{Conf}_p(\psi_{\mathrm{fail}})$ and $\mathrm{Conf}_p(\psi_{\mathrm{safe}})$. The percentage of cases where this difference was bigger than 10 % is reported as "GAP>10 %". The success of our default hyperplane method rides on this fraction being small.

Table 1 shows the percentage of categorization for the confidence bound regarding 95 % unimodal partition in the mixture likelihood when the dimension of data is two (i.e., $m$=2). We observe that overall the cases, at most 3.2 % of the time there was an unresolved decision as to whether existence of a unimodal partition was possible at 95 % (i.e., "BETWEEN"). At all of the parameter settings the probability of a 10 % or worse gap in $\mathrm{Conf}_p(\psi_{\mathrm{fail}})$ and $\mathrm{Conf}_p(\psi_{\mathrm{safe}})$ was less than 0.102.

**Table 2** Percentage of categorization with respect to 95 % unimodal partition: $m$=5 and $R = 500$

| $n$ | Modes | Diagnostics | $\pi = 0.5$ | | | $\pi = 0.2$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | $d = 2$ | $d = 3$ | $d = 4$ | $d = 2.5$ | $d = 3.75$ | $d = 5$ |
| 200 | Unimode | GOOD | 0.0 | 0.2 | 3.2 | 0.0 | 2.2 | 3.4 |
| | | BETWEEN | 0.0 | 0.4 | 0.2 | 0.0 | 0.2 | 0.0 |
| | | BAD | 0.0 | 0.4 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Multimodes | GOOD | 0.0 | 2.2 | 90.2 | 0.4 | 79.8 | 96.6 |
| | | BETWEEN | 0.0 | 0.4 | 0.8 | 0.2 | 1.6 | 0.0 |
| | | BAD | 100 | 96.4 | 5.6 | 99.4 | 16.2 | 0.0 |
| | | GAP>10 % | 0.0 | 8.0 | 0.8 | 2.4 | 0.0 | 0.0 |
| 400 | Unimode | GOOD | 0.0 | 2.0 | 7.8 | 0.0 | 4.6 | 7.2 |
| | | BETWEEN | 0.0 | 0.8 | 0.0 | 0.8 | 0.0 | 0.0 |
| | | BAD | 0.6 | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 |
| | Multimodes | GOOD | 0.0 | 58.4 | 92.2 | 6.6 | 95.4 | 92.8 |
| | | BETWEEN | 0.0 | 3.2 | 0.0 | 2.6 | 0.0 | 0.0 |
| | | BAD | 99.4 | 35.6 | 0.0 | 89.4 | 0.0 | 0.0 |
| | | GAP>10 % | 1.6 | 4.6 | 0.0 | 7.6 | 0.0 | 0.0 |

Our table also shows that secondary modes in the mixture likelihood exist at least 50 % of the time, regardless of parameter setting. As we expected, multiple modal groups occurred more frequently as component separation and sample size decreased. If we compare the settings where the standard deviation of the mixing distribution was the same, then a unimodal partition was more likely to occur for the case where the weights were unequal. For example, when $n = 100$, the percentage of cases where 95 % unimodal partition existed at $(\pi, d) = (0.2,3.75)$ was 2.8 times higher than at $(\pi, d) = (0.5,3)$. Note that when the smallest mixing weight gets smaller, the locations of the components are pushed apart. This forced the components to be more obviously different, and thus it was easier to construct identifiable partition.

Table 2 shows the percentage of categorization for the confidence bound regarding 95 % unimodal partition in the mixture likelihood when the dimension was $m = 5$. We see that $\mathrm{Conf}_p(\psi_{\mathrm{safe}})$ proved to be a highly useful lower bound because the BETWEEN cases were at most 4 %. Note that the frequency of the cases where "GAP>10 %" was never larger than 8 %.

We also observe that there were more than one modal group in the mixture likelihood at least 92 % of the time. In the most extreme case, when the sample size was small relative to component separation, all simulated data sets had more than one modal group. If we compare with $m = 2$ in Table 1, we can see that, for a given sample size and standard deviation of the mixing distribution, the chances of a unimodal partition were smaller for $m = 5$ than for $m = 2$. One needs both larger component separation and unequal weights under the same standard deviation of a mixing distribution to guarantee a 95 % unimodal partition.

## 7 Data analysis

In this section we will present a few examples to show the use of the proposed methods designed to detect the confidence level for a unimodal partition (i.e., the empirical identifiability) at finite samples. Note that in each data set we used a strategy of multiple starting values in the EM algorithm to search solutions to the mixture likelihood equations. We will also illustrate how to use the value of the empirical identifiability to visualize the likelihood-based profile sets for the parameters of interest by using the *modal simulation* developed in Kim and Lindsay (2011a, b).

### 7.1 Empirical identifiability using landmark elevations

*Example 1* The first example is a case where there is a significant secondary mode in the likelihood. That is, its elevation is lower than that of the MLE mode, but higher than $\psi_{\text{deg}}$. We generated 75 observations from a three-component normal mixtures, 0.33 N(−3,1) + 0.34 N(0,1) + 0.33 N(3,1). Then we fitted a three-component normal mixture with equal component variance. Note that we fixed the component variance to be .25, which was smaller than the true value 1 to create more than one mode in the likelihood. The two modes were $\hat{\theta}_{\text{MLE}} = [\binom{0.257}{-3.166}, \binom{0.350}{-0.782}, \binom{0.392}{2.430}]$ with $\ell(\hat{\theta}_{\text{MLE}}) = -211.7$ and $\hat{\theta}_{\text{2nd}} = [\binom{0.367}{-2.711}, \binom{0.351}{0.106}, \binom{0.282}{2.917}]$ with $\ell(\hat{\theta}_{\text{2nd}}) = -213.54$. The maximum log likelihood in the degenerate class was −298.11. Based on that information alone, the maximum possible confidence for an identifiable partition was $\text{Conf}_5(\psi_{\text{deg}}) = 100\ \%$. However, in this example, $\psi_{\text{safe}} = \psi_{\text{2nd}}$ because the component parameter is univariate and $\psi_{\text{hyp}} = \psi_{\text{deg}}$. Thus, a unimodal partition exists only at the levels below $\text{Conf}_5(\psi_{\text{safe}}) = \text{Conf}_5(\psi_{\text{2nd}}) = 40\ \%$.

*Example 2* (SLC data) The second example concerns red blood cell sodium–lithium countertransport (SLC) activity data collected from 190 individuals (Dudley et al. 1991). The SLC is measured as the difference in lithium efflux rate from lithium-loaded cells into sodium chloride and sodium-free media. Geneticists are interested in the SLC because it may be an important cause of essential hypertension. Roeder (1994) fitted a three-component normal mixture model with equal variances to this data. We used the same mixture model and obtained the MLE for $\theta = \left[\begin{pmatrix}\pi_1\\ \xi_1\\ \omega\end{pmatrix}, \begin{pmatrix}\pi_2\\ \xi_2\\ \omega\end{pmatrix}, \begin{pmatrix}\pi_3\\ \xi_3\\ \omega\end{pmatrix}\right]$, $\hat{\theta} = \left[\begin{pmatrix}0.78\\ 0.22\\ 0.003\end{pmatrix}, \begin{pmatrix}0.20\\ 0.38\\ 0.003\end{pmatrix}, \begin{pmatrix}0.02\\ 0.58\\ 0.003\end{pmatrix}\right]$. Note that there was only the MLE modal group. Since the component parameter is univariate, the safe elevation for a unimodal partition in the likelihood is equal to $\psi_{\text{safe}} = \psi_{\text{deg}}$, the elevation of the MLE for a class of degenerate parameters (a two-component normal mixture with equal variances). We calculated the confidence level corresponding to $\psi_{\text{deg}}$, which was $\text{Conf}_6(\psi_{\text{deg}}) = 86\ \%$. Therefore, we have a unimodal partition at any confidence level below 86 % in the full-dimensional likelihood.

*Example 3* (Morbidity data) The third example concerns a cohort study in northeast Thailand (Schelp et al. 1990) where the health status of $n = 602$ preschool children checked every two weeks from June 1982 until September 1985. For each child it

was recorded whether the child showed one of the symptoms fever or cough, or both together. The data were the frequencies of these illness spells during the study period, which can be found in Böhning et al. (1992).

Böhning et al. (1992) and Schlattmann (2005) fitted four-component Poisson mixture model to this data : $g(y;\theta) = \sum_{j=1}^{4} \pi_j \left(\mathrm{e}^{-\xi_j}\xi_j^y/x!\right)$ where $\theta = [\binom{\pi_1}{\xi_1}, \binom{\pi_2}{\xi_2}, \binom{\pi_3}{\xi_3}, \binom{\pi_4}{\xi_4}]$. We computed the maximum likelihood estimator for $\theta$, $\hat{\theta} = \left[\binom{0.197}{0.14}, \binom{0.48}{2.82}, \binom{0.27}{8.16}, \binom{0.05}{16.16}\right]$. The log likelihood of the MLE for the class of degenerate parameters (i.e., a three-component Poisson mixture model) was -1568.28 and then $\mathrm{Conf}_7(\psi_{\mathrm{deg}})$ was larger than 99 %. Since $\xi_j$ was univariate and there was no secondary mode, $\psi_{\mathrm{safe}} = \psi_{\mathrm{fail}} = \psi_{\mathrm{deg}}$ and $\mathrm{Conf}_7(\psi_{\mathrm{deg}})$ was the exact upper bound on the confidence levels for a unimodal partition at this data. In other words, there are 4! disjoint unimodal regions, each identifiable subset for one of 4! MLE modes.

*Example 4* (Blue crab data) The fourth example, analyzed in Campbell and Mahon (1974) and McLachlan and Peel (2000), contains five variables measured from $n = 100$ blue crabs: the measurements (in mm) were on the width of the frontal lip, the rear width, the length along the midline and the maximum width of the carapace, and the body depth. In this data, there are 50 males and 50 females. We here fit the observations of the second and third variable with a two-component bivariate normal mixture model with equal covariances, ignoring the known classification.

In this data set, we found four modes with the following elevations and corresponding confidence levels, $\psi_{\mathrm{MLE}}$=−457.515 with $\mathrm{Conf}_8(\psi_{\mathrm{MLE}})$=0 %, $\psi_{2\mathrm{nd}} = -467.883$ with $\mathrm{Conf}_8(\psi_{2\mathrm{nd}}) = 99.2$ %, $\psi_{3rd} = -475.467$ with $\mathrm{Conf}_8(\psi_{3rd}) = 100$ % and $\psi_{4th} = -477.003$ with $\mathrm{Conf}_8(\psi_{4th}) = 100$ %. We also computed the elevation of the MLE for the degenerate class of parameters, $\psi_{\mathrm{deg}} = -477.006$ with $\mathrm{Conf}_8(\psi_{\mathrm{deg}}) = 100$ %. The maximum elevation of the default hyperplane in Eq. (8) estimated from the restricted EM algorithm of Sect. 5.4 was $\psi_{\mathrm{hyp}} = -469.116$ with $\mathrm{Conf}_8(\psi_{\mathrm{hyp}}) = 99.6$ %. Therefore, one can construct a unimodal partition for the MLE in the likelihood region at any confidence level below $\mathrm{Conf}_8(\psi_{\mathrm{safe}}) = \mathrm{Conf}_8(\psi_{2\mathrm{nd}}) = 99.2$ %.

### 7.2 Visual assessment using the modal simulation

An important practical problem in using the likelihood is to create a way to use the identifiability generated by the full-dimensional likelihood topology when constructing profile sets for various functions of parameters of interest at the targeted confidence level (elevation). A conventional numerical approach to find the boundaries of the targeted profile set that are locally identifiable can be computationally demanding.

To cope with this computational difficulty, Kim and Lindsay (2011a, b) developed a stochastic parameter sampling method, *modal simulation*, as a means to facilitate a visual analysis of (joint/profile) confidence sets generated by a likelihood. The basic idea is that, given the observed data and a MLE mode for the parameters in the assumed model, one creates a sampling distribution for the parameter values at the boundaries of the targeted likelihood regions in the full dimensional parameter space. That is, given a targeted elevation $c$, one generates a random direction from a multivariate standard

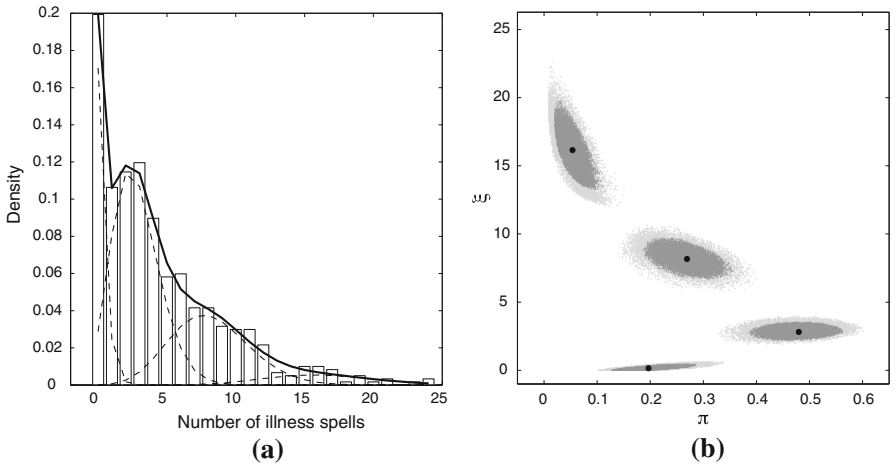



**Fig. 4** Example 3: **a** a histogram of the frequencies of illness spells and overlaid mixture density estimate; **b** a 45.93 % (*dark gray*) and 94.54 % (*light gray*) profile sampling plot for $(\boldsymbol{\pi}, \boldsymbol{\xi})$. Note that the *black circles* represents the MLEs for $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$

normal distribution and heads in that random direction from the MLE mode of interest until one hits the boundaries of the targeted region. Then this point becomes the sampled value of the parameters. Those authors showed by simulations and real data applications that modal simulation can successfully visualize numerically computed profile likelihood boundaries even when there exist multiple modes.

There are two important features of the modal simulation method. First, based on the sampled parameter values, every function of the parameters of interest can be computed and then mapped to create lower dimensional confidence regions of interest, all without further numerical optimization. Second, the simulated parameter values automatically have the same labels as that of the MLE mode when a unimodal partition exists in the likelihood.

In this subsection, we use the two examples (Example 3 and Example 4) analyzed in Sect. 7.1 to illustrate how to utilize the value of empirical identifiability via the modal simulation method for visually assessing a locally identifiable likelihood region around the major MLE mode.

Figure 4a shows a histogram of Morbidity data in Example 3 overlaid by a four-component Poisson mixture density estimate. The exact upper bound of empirical identifiability for a unimodal partition was $\text{Conf}_7(\psi_{\text{deg}})$, which is larger than 99 %. Figure 4b is a profile sampling plot for $(\pi, \xi)$ of each component with $\text{Conf}_7(c) = 45.93$ % (dark gray) and 94.54 % (light gray). From Fig. 4b, we see four connected sets of the MLE modes, one for each component, that are separated by ordered component parameters, not ordered mixing weights.

Figure 5a represents a contour plot of an estimated two-component bivariate normal mixture model with equal covariances, overlaid with scatter plot of real width (first coordinate) and length along the midline (second coordinate) in Example 4 (Blue crab data). Note that the estimated mean vectors for two components, $\boldsymbol{\xi}_1 = (\boldsymbol{\xi}_{11}, \boldsymbol{\xi}_{21})$ and

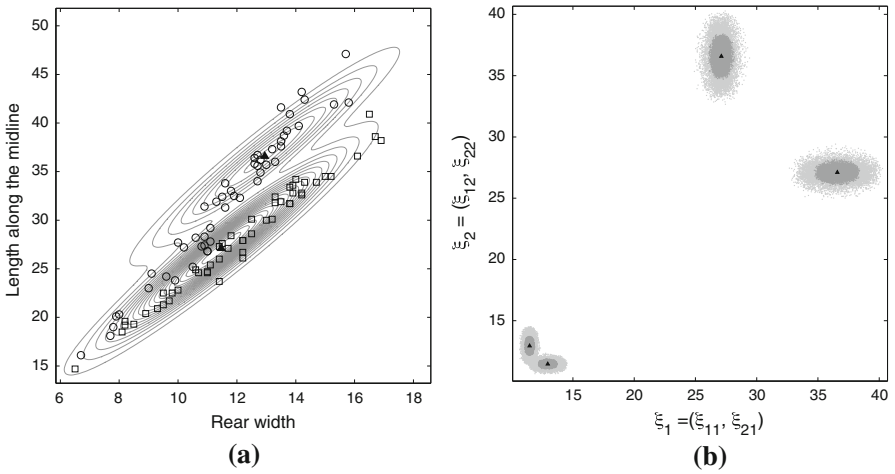



**Fig. 5** Example 4: **a** Contour plot of bivariate density estimate overlaid with scatter plot of bluecrab data (*circles* and *squares* represent male crabs and female crabs, respectively); **b** a 12.89 % (*dark gray*) and 95 % (*light gray*) profile sampling plot for two-component mean vectors, $\boldsymbol{\xi}_1 = (\xi_{11}, \xi_{21})$ and $\boldsymbol{\xi}_2 = (\xi_{12}, \xi_{22})$. Note that the *black triangles* represent the MLEs for the mean parameters

$\boldsymbol{\xi}_2 = (\xi_{12}, \xi_{22})$, are $\hat{\boldsymbol{\xi}}_1 = (12.95, 36.58)$ and $\hat{\boldsymbol{\xi}}_2 = (11.47, 27.11)$, and the computed empirical identifiability is 99.2 %. Figure 5b shows a profile sampling plot for two-component mean vectors, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, with $\text{Conf}_8(c) = 12.89$ % (dark gray) and 95 % (light gray). Due to labeling nonidentifiability, this confidence region is symmetric with respect to 180 degree rotations. We observe from Fig. 5b that when the confidence level is small, the two modal regions around two MLE modes for both coordinates are separated from each other. But, the two modal regions with a large confidence level are disjoint only for the second coordinate. This result indicates that the second coordinate in the data contains information that generates an identifiable unimodal partition with a large confidence level in the full parameter space (and guarantees a high value of empirical identifiability on the model parameters).

## 8 Discussion

In this paper, we have shown that for any given data set and model, there will typically exist limited range of confidence levels at which one can define unimodal identifiable partition in the mixture likelihood. We have interpreted the bound on such confidence as the *empirical identifiability* that quantifies degree of local identifiability on the estimated parameters in a given data set. If the components are not well separated relative to the sample size, we can expect that this range of confidence levels will be small, and possibly unsatisfactory. Further, as indicated in Remark 2, one's ability to make identifiable confidence sets at a given level is closely related to the evidence in the data for the existence of K components.

We have proposed in Sect. 4 a few landmark elevations designed to determine the existence of an identifiable unimodal partition. For calculating them in simulation

studies and data analysis, we used the EM algorithm, one of the widely used local optimization technique, required to run from multiple starting points. If one is concerned with number of starting values and confidence in the numerical results as the dimension of the parameters increases, we suggest computing probability of finding a new local maximum (Finch et al. 1989). If the estimated probability is small enough, one can stop re-running the algorithm with new initial values.

We should point out that in this paper we have considered regular likelihoods for which the likelihood is bounded. For irregular likelihoods the proposed methods would apply after some suitable regularization. That is, our approaches can be applied to any similar inference function, such as a penalized likelihood or a posterior density, though some of our theoretical results might need additional conditions.

From an applied viewpoint, there are significances of the empirical identifiability and its measuring method proposed in this paper. First, for better understanding the relationships between the parameters in the assumed mixture model, it is often useful to construct the profile likelihood for the parameters or functions of the parameters of interest. A practical difficulty lies in the presentation of locally identifiable profile likelihood regions that can retain information on the empirical identifiability generated by the full-dimensional likelihood topology. The proposed method enables transmission of information about the empirical identifiability in the full-dimensional parameter space to any lower dimensional parameters of interest. Second, the empirical identifiability computed using our proposed method can be used to assess the need to specify priors on the parameters in Bayesian mixture modeling. The empirical identifiability defined in this paper represents the maximum size of the locally identifiable likelihood region around an estimated MLE mode in terms of the confidence level. The practitioners can compare the empirical identifiability computed using the proposed approach with a desired target confidence level predetermined by them. When one cannot make identifiable likelihood confidence sets at a target level, this may indicate that the data are not sufficient for precise identification of a given model. One possible solution to the problem is to assign a prior putting sufficient mass away from the parameters that are causing pathologies in the topology of the likelihood regions, in particular, the degenerate class of parameters. Third, the empirical identifiability computed from the proposed method can help understand the effect of the non-constant/non-flat priors on the Bayesian mixture inference. The likelihood confidence set around the MLE mode corresponds to the Highest Posterior Density (HPD) region around the maximum-a-posteriori estimate under flat and constant priors. Thus, one can compare the empirical identifiability computed in the likelihood with the counterpart in the posterior (with non-flat priors), for example, the HPD region-based labeling credibility proposed by Yao and Lindsay (2009).

## 9 Appendix: Proofs

*Proof of Theorem 1.* Suppose that one starts at the highest elevation in the mixture likelihood, $\psi_{\text{MLE}}$. Then the likelihood region at $\psi_{\text{MLE}}$ is equal to the set of $K!$ modes. If the elevation $c$ is just below $\psi_{\text{MLE}}$, the likelihood region consists of $K!$ disjoint modal regions $C_c(\hat{\theta}^{\sigma})$.

However, if the elevation $c$ goes too low, some properties in Definition 3 could fail so that a unimodal partition cannot exist. Two different problems could arise. The first case is where there exists a secondary mode in the likelihood, with the elevation $\psi_{2\text{nd}}$. In this case, for the elevation $c$ at or just below $\psi_{2\text{nd}}$, a secondary set of modal regions is formed containing the $K!$ secondary modes. Since these points are not path-connected to the MLE modes, the property (P3) would be violated. If the primary and secondary modal regions become reconnected at a lower level of $c$, then, even if identifiable, each element contains two modes, thereby violating property (P4).

A second case where a unimodal partition could fail is when $c \leq \psi_{\text{mm}}$, the minimal elevation of the maximin path connecting $C_c(\hat{\theta})$ to a permuted region $C_c(\hat{\theta}^\sigma)$. This causes a violation of property (P1).

As long as the elevation $c$ is larger than $\max\{\psi_{\text{mm}}, \psi_{2\text{nd}}\}$, however, then there are no secondary modal regions possible, and there cannot exist connections between the modal regions of the MLE group, as there exist no saddle points. Therefore the $K!$ modal regions $C_c(\hat{\theta}^\sigma)$ are disjoint and their union is the elevation $c$ likelihood region $C_c^{\text{LR}}$, so that we have verified properties (P1) and (P3).

We claim that the second property (P2) holds because the $K!$ modal regions $C_c(\hat{\theta}^\sigma)$ are disjoint by (P1) and each is connected (by definition). The argument for identifiability goes as follows: first, given any $\theta$ in $C_c(\hat{\theta})$, its permutation $\theta^\sigma$ cannot also be in the same modal region $C_c(\hat{\theta})$, as we know it lies in a disjoint $C_c(\hat{\theta}^\sigma)$. Secondly, we suppose, for purposes of contradiction, that $C_c(\hat{\theta})$ contains a degenerate point $\theta_0$, but is not path-connected to any $C_c(\hat{\theta}^\sigma)$, for any permutation $\sigma$. Then the permuted $\theta_0$, $\theta_0^\sigma$, is contained in only $C_c(\hat{\theta}^\sigma)$. Since the region of $\theta$ values that generate the same degenerate mixing distribution and hence the same likelihood as $\theta_0$ is a connected set, $\theta_0^\sigma$ should be in the same connected set. Thus, at the elevation $c \leq \psi_{\text{deg}}$, $C_c(\hat{\theta})$ and $C_c(\hat{\theta}^\sigma)$ are connected to each other by a path. But this is contradicted with the property (P1) that the $K!$ modal regions $C_c(\hat{\theta}^\sigma)$ are disjoint. □

*Proof of Theorem 2.* Suppose the two modal regions, $C_c(\hat{\theta})$ and $C_c(\hat{\theta}^\sigma)$, are connected to each other at the elevation $c$. Then a maximin path connecting them has $\psi_{\text{mm}} \geq c$. Any path connecting the modes necessarily intersects the hyperplane of Eq (6) in one or more points. Let $c^\star$ be the minimal value of the likelihood on the intersection set. Then $c^\star \geq \psi_{\text{mm}}$ by the definition of $\psi_{\text{mm}}$. However, we also have $c^\star \leq \psi_{\text{hyp}}$ by the definition of the latter. We therefore have $\psi_{\text{mm}} \leq \psi_{\text{hyp}}$. □

*Proof of Theorem 3.* Suppose there exists a continuous path of $\gamma(t)$ parameter values that connects the profile modes $\gamma(\theta_A)$ and $\gamma(\theta_B)$ such that the profile likelihood along the path stays above the elevation $c$. Then the path $(\gamma(t), \hat{\phi}(\gamma(t)))$ is a continuous path in the full parameter space that connects the two modes and whose likelihood stays above $c$. □

*Proof of Theorem 4.* Suppose $K = 2$. Then the default hyperplane generated by Eq.(8) consists of all mixtures with $\xi_1 = \xi_2$, and so $\psi_{\text{hyp}} = \psi_{\text{deg}} = \psi_{\text{mm}}$. Thus, if $c > \psi_{\text{deg}}$, $C_c(\hat{\theta})$ and $C_c(\hat{\theta}^\sigma)$ cannot have a connecting path. If this is the case, the labels in a selected modal region must satisfy $\xi_1 > \xi_2$ or $\xi_1 < \xi_2$, as the modal set cannot contain points satisfying $\xi_1 = \xi_2$. That is, the points in each modal region

must have order restriction labels. The argument for a two-component case extends to $K(>2)$ components. Given a set of ordered $\xi$'s, there is no way to move continuously to a permuted set without one pair becoming equal along the way. □

## References


Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: Wiley.

Böhning, D., Schlattmann, P., Lindsay, B. G. (1992). C.a.man-computer assisted analysis of mixtures: statistical algorithms. *Biometrics*, *48*, 283–303.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one parameter exponential family. *Annals of the Institute of Statistical Mathematics*, *46*, 373–388.

Campbell, N. A., Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of genus Leptograsus. *Australian Journal of Zoology*, *22*, 417–425.

Chen, H., Chen, J. (2001). The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics*, *29*, 201–215.

Cox, D. R., Hinkley, D. V. (2002). *Theoretical Statistics*. London: Chapman & Hall.

Crawford, S. L. (1994). An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, *89*, 259–267.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, *39*, 1–38.

Dudley, C. R. K., Giuffra, L. A., Raine, A. E. G., Reeders, S. T. (1991). Assessing the role of apnh, a gene encoding for a human amiloride-sensitive na+/h+ antiporter, on the interindividual variation in red cell na+/li+ countertransport. *Journal of the American Society of Nephrology*, *2*, 937–943.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Finch, S. J., Mendell, N. R., Thode, H. C. (1989). Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistical Association*, *84*, 1020–1023.

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231.

Huang, G. H., Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, *69*, 5–32.

Kalbfleisch, J. D., Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New Jersey: Wiley.

Kim, D., Lindsay, B. G. (2011a). Modal simulation and visualization in finite mixture models. *Canadian Journal of Statistics*, *39*, 421–437.

Kim, D., Lindsay, B. G. (2011b). Using confidence distribution sampling to visualize confidence sets. *Statistica Sinica*, *21*, 923–948.

Kim, D. K., Taylor, J. M. G. (1995). The restricted em algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association*, *430*, 708–716.

Lindsay, B. G. (1995). Mixture Models: Theory, Geometry, and Applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics* (Vol. 5). Hayward: Institute of Mathematical Statistics.

Liu, X., Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, *31*, 807–832.

Matsumoto, Y. (2002). *An Introduction to Morse Theory. Translations of Mathematical Monographs* (Vol. 208). Providence: American Mathematical Society.

McLachlan, J., Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Meeker, W., Escobar, L. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, *49*, 48–53.

Redner, R. A., Walker, H. C. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, *26*, 195–239.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, *89*, 487–495.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica*, *39*, 577–591.

Schelp, F. P., Vivatanasept, P., Sitaputra, P., Sormani, S., Pongpaew, P., Vudhivai, N., et al. (1990). Relationship of the morbidity of under-fives to anthropometric measurements and community health intervention. *Tropical Medicine and Parasitology*, *41*, 121–126.

Schlattmann, P. (2005). On bootstrapping the number of components in finite mixtures of poisson distributions. *Statistics and Computing*, *15*, 179–188.

Teicher, H. (1960). On the mixture of distributions. *The Annals of Mathematical Statistics*, *31*, 55–73.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *34*, 1265–1269.

Titterington, D. M., Smith, A. F. M., Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Models*. New York: Wiley.

Yakowitz, S. J., Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*, 209–214.

Yao, W., Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association*, *104*, 758–767.