Order selection with confidence for finite mixture models

Hien D. Nguyen[1] · **Daniel Fryer**[1] · **Geoffrey J. McLachlan**[1]

Abstract

The determination of the number of mixture components (the order) of a finite mixture model has been an enduring problem in statistical inference. We prove that the closed testing principle leads to a sequential testing procedure (STP) that allows for confidence statements to be made regarding the order of a finite mixture model. We construct finite sample tests, via data splitting and data swapping, for use in the STP, and we prove that such tests are consistent against fixed alternatives. Simulation studies and real data examples are used to demonstrate the performance of the finite sample tests-based STP, yielding practical recommendations of their use as confidence estimators in combination with point estimates such as the Akaike information or Bayesian information criteria. In addition, we demonstrate that a modification of the STP yields a method that consistently selects the order of a finite mixture model, in the asymptotic sense. Our STP is not only applicable for order selection of finite mixture models, but is also useful for making confidence statements regarding any sequence of nested models.

Keywords Order selection · Data splitting · Confidence sets · Hypothesis tests · Mixture models

1 Introduction

Let $X \in \mathbb{X}$ be a random variable. Let $\mathcal{K}(\mathbb{X})$ be a class of probability density functions (PDFs), defined on the set $\mathbb{X}$, which we shall refer to as components. We say that X arises from a g component mixture model of class $\mathcal{K}$ if the PDF f_0 of X belongs in the convex class

$$\mathcal{M}_g(\mathbb{X}) = \left\{ f(\boldsymbol{x}) : f(\boldsymbol{x}) = \sum_{z=1}^{g} \pi_z f_z(\boldsymbol{x}); \; \pi_z \geq 0, \sum_{z=1}^{g} \pi_z = 1, f_z \in \mathcal{K}(\mathbb{X}), z \in [g] \right\},$$

✉ Hien D. Nguyen
  h.nguyen5@latrobe.edu.au

1   School of Mathematics and Physics, University of Queensland, St. Lucia, Australia

where $g \in \mathbb{N}$ and $[g] = \{1, \ldots, g\}$.

Suppose that we observe an independent and identically distributed (IID) sample sequence of data $\mathbf{X}_n = (X_i)_{i=1}^{n}$, where each X_i has the same data generating process (DGP) as X, which is unknown. Under the assumption that $f_0 \in \mathcal{M}_{g_0}(\mathbb{X})$ for some $g_0 \in \mathbb{N}$, we wish to use the data $\mathbf{X}_n$ in order to determine the possible values of g_0. This problem is generally referred to as order selection in the mixture modeling literature, and reviews regarding the problem can be found in McLachlan and Peel (2000, Ch. 6) and McLachlan and Rathnayake (2014), for example.

Notice that the sequence $(\mathcal{M}_g)_{g=1}^{\infty}$ is nested, in the sense that $\mathcal{M}_g \subset \mathcal{M}_{g+1}$, for each g, and that $g_0 \in \mathbb{N}$ is equivalent to $f_0 \in \mathcal{M} = \bigcup_{g=1}^{\infty} \mathcal{M}_g$. We shall write the null hypothesis that $f_0 \in \mathcal{M}_g$ (or equivalently, $g_0 \leq g$) as H_g, and we assume that we have available a p-value $P_g(\mathbf{X}_n)$ for each hypothesis, and that $P_g(\mathbf{X}_n)$ correctly controls the size of the hypothesis test, in the sense that

$$\sup_{f \in \mathcal{M}_g} \mathrm{Pr}_f \left(P_g(\mathbf{X}_n) \leq \alpha \right) \leq \alpha, \tag{1}$$

for any $\alpha \in (0, 1)$. Here, Pr_f is the probability measure corresponding to the PDF f. In Wasserman et al. (2020), the following simple sequential testing procedure (STP) is proposed for determining the value of g_0 (for general nested models, not necessarily mixtures):

1. Choose some significance level $\alpha \in (0, 1)$ and initialize $\hat{g} = 0$;
2. Set $\hat{g} = \hat{g} + 1$;
3. Test the null hypothesis $\mathrm{H}_{\hat{g}}$ using the p-value $P_{\hat{g}}(\mathbf{X}_n)$;

    (a) If $P_{\hat{g}}(\mathbf{X}_n) \leq \alpha$, then go to Step 2.
    (b) If $P_{\hat{g}}(\mathbf{X}_n) > \alpha$, then go to Step 4.

4. Output the estimated number of components $\hat{g}_n = \hat{g}$.

It was argued informally in Wasserman et al. (2020) that, although the procedure above involves a sequence of multiple tests, each with local size α, it still correctly controls the Type I error in the sense that

$$\mathrm{Pr}_{f_0} \left(f_0 \in \mathcal{M}_{\hat{g}_n - 1} \right) \leq \alpha \tag{2}$$

for any $f_0 \in \mathcal{M}$. Here, we note that the complement of the event $\{f_0 \in \mathcal{M}_{\hat{g}_n - 1}\}$ is $\{f_0 \in \mathcal{M} \backslash \mathcal{M}_{\hat{g}_n - 1}\}$ or equivalently $\{g_0 \geq \hat{g}_n\}$. Thus, from (2), we can make the confidence statement that

$$\mathrm{Pr}_{f_0} \left(g_0 \geq \hat{g}_n \right) \geq 1 - \alpha, \tag{3}$$

for any $f_0 \in \mathcal{M}$.

In the present work, we shall provide a formal proof of result (2) using the closed testing principle of Marcus et al. (1976) (see also Dickhaus, 2014, Sec. 3.3). Using this result and the universal inference framework of Wasserman et al. (2020), we construct a sequence of tests for $\left(\mathrm{H}_g\right)_{g=1}^{\infty}$ with p-values satisfying (1) and prove that each of the tests is consistent under some regularity conditions. We then demonstrate the performance of our testing procedure for the problem of order selection for finite mixtures of normal distributions, and verify the empirical manifestation of the confidence result (3). Extensions of the STP are also considered, whereupon we construct a method that consistently estimates the order g_0, and consider the application of the STP to asymptotically valid tests.

We note that hypothesis testing for order selection in mixture models is a well-studied area of research. Difficulties in applying testing procedures to the order selection problem arise due to identifiability and boundary issues of the null hypothesis parameter spaces (see, e.g., Quinn et al., 1987, and references therein regarding parametric mixture models, and Andrews, 2001, more generally). Examples of testing methods proposed to overcome the problem include the parametric bootstrapping techniques of McLachlan (1987), Feng and McCulloch (1996), and Polymenis and Titterington (1998), whereupon bootstrapped distributions of test statistics are used to approximate finite sample distributions, in the absence of asymptotic results. Another approach is the penalization techniques of Chen (1998), Li and Chen (2010), and Chen et al. (2012), where asymptotically well-behaved penalized likelihood ratio statistics are proposed, with limiting distributions that are computable or simulatable. It is noteworthy that the bootstrap approaches provide only an approximate finite sample distribution of test statistics and thus the tests are not guaranteed to have the correct size. The penalization approach, on the other hand, provides asymptotic tests of the correct size, although the construction of the penalization of the test statistic must be specialized to every individual testing problem and is only suitable for parametric families of densities $\mathcal{K}$ that are characterized by a low-dimensional parameter.

In fact, the sequential procedure described above was also considered for order selection in the mixture model context by Windham and Cutler (1992) and Polymenis and Titterington (1998), although no establishment of the properties of the approach was provided. The possibility of constructing intervals of form (3) via bounding of discrete functionals of the underlying probability measure is discussed in Donoho (1988), although no implementation is suggested. Citing observations made by Donoho (1988) and Cutler and Windham (1994), it is suggested in McLachlan and Peel (2000, Sec. 6.1) that intervals of form (3) are sensible in practice, because reasonable functionals that characterize properties of f_0, such as for the number of components g_0, can be lower bounded with high probability from data, but often cannot be upper bounded.

As previously mentioned, we plan to prove that (2) holds by demonstrating that the sequential test is a closed testing procedure. However, we note that the procedure may also be considered under the sequential rejection principle of Goeman and Solari (2010), and if $\mathcal{M} = \bigcup_{g=1}^{G} \mathcal{M}_g$ for some fixed $G \in \mathbb{N}$, then we may also consider the procedure as a fixed sequence procedure, as considered by Maurer et al.

(1995). Another perspective regarding the sequential test is via the general procedures of Bauer and Kieser (1996), who consider the construction of confidence intervals using sequences of tests for nested and order sets of hypotheses. We also remark that the use of multiple testing procedures for model selection is well studied in the literature, as exemplified by the works of Finner and Giani (1996) and Hansen et al. (2011), who both consider the application of hypothesis testing schemes to generate confidence sets over model spaces.

For completeness, we note that apart from hypothesis testing, numerous solutions to the order selection problem for finite mixture models have been suggested. These related works include the use of information criteria, such as the Akaike information crtierion (AIC), Bayesian information criterion (BIC), and variants of such techniques (Biernacki et al., 2000; Keribin, 2000; Leroux, 1992), and parameter regularization, such as via the Lasso and elastic net, and penalization approaches (Chen & Khalili, 2009; Xu & Chen, 2015; Yin et al., 2019), among other techniques.

We note that the aforementioned order selection techniques are all, in a sense, point estimation procedures that each serve the purpose of consistently estimating the number of components of the DGP mixture model, in the sense that the estimate is close to the true number of components, for sufficiently large n. Our approach does not output a consistent estimator, but instead produces fixed-probability confidence set $\{g_0 \geq \hat{g}_n\}$, and should thus be viewed as an interval estimator. Although the lower-bound of the interval $\hat{g}_n$ can be an accurate estimator of the true number of components, it should not be considered as a competitor to proper point estimators and instead should be viewed as complementary to point estimation approaches. We finally note that outside of the multiple testing framework, the problem of model selection with confidence has also been addressed in the articles of Ferrari and Yang (2015) and Zheng et al. (2019).

The remainder of manuscript proceeds as follows. In Sect. 2, we recall the closed testing principle and use it to prove the inequality (2). In Sect. 3, we use the universal inference framework of Wasserman et al. (2020) to construct a class of likelihood ratio-based tests for the hypotheses $\left(H_g\right)_{g=1}^{\infty}$. In the context of normal mixture models, numerical simulations and real data examples are used to assess the performance of the sequential procedure using the constructed tests in Sect. 4. Extensions to the STP are discussed in Sect. 5. Finally, conclusions are provided in Sect. 6 and technical proofs are provided in the Appendix.

2 Confidence via the closed testing principle

Let $\mathbb{H} = \left\{H_g : g \in \mathbb{G}\right\}$ be a set of hypotheses that are indexed by some (possibly infinite) set $\mathbb{G}$, where each hypothesis H_g corresponds to the statement $\left\{\theta \in \mathbb{T}_g\right\}$ regarding the parameter of interest $\theta \in \mathbb{T}$, where $\mathbb{T}_g \subset \mathbb{T}$. We say that $\mathbb{H}$ is a $\cap$-closed system if for each $\mathbb{I} \subseteq \mathbb{G}$, either $\bigcap_{g \in \mathbb{I}} \mathbb{T}_g = \emptyset$ or $\bigcap_{g \in \mathbb{I}} \mathbb{T}_g \in \left\{\mathbb{T}_g : g \in \mathbb{G}\right\}$. That is, for

every set $\mathbb{I}$ of indices that yields a non-empty statement $\left\{ \boldsymbol{\theta} \in \bigcap_{g \in \mathbb{I}} \mathbb{T}_g \right\}$, there exists a hypothesis $H_g \in \mathbb{H}$, such that $g \in \mathbb{T}_g$.

Recalling the notation from Sect. 1, we say that H_g is rejected if $R_g(\mathbf{X}_n) = \mathbf{1}\{P_g(\mathbf{X}_n) \leq \alpha\}$ is equal to 1, and we say that H_g is not rejected, otherwise. Here, $\mathbf{1}\{\cdot\}$ is the indicator function. We further say that the familywise error rate (FWER) of a set rejections $\{R_g(\mathbf{X}_n)\}_{g \in \mathbb{G}}$ is strongly controlled at level $\alpha \in (0, 1)$ if for all $\boldsymbol{\theta} \in \mathbb{T}$,

$$\mathrm{Pr}_{\boldsymbol{\theta}} \left(\bigcup_{g \in \mathbb{G}_0(\boldsymbol{\theta})} \{R_g(\mathbf{X}_n) = 1\} \right) \leq \alpha,$$

where $\mathrm{Pr}_{\boldsymbol{\theta}}$ denotes the probability measure corresponding to parameter value $\boldsymbol{\theta}$, and $\mathbb{G}_0(\boldsymbol{\theta}) \subset \mathbb{G}$ is the set of indices with corresponding hypotheses that are true under $\mathrm{Pr}_{\boldsymbol{\theta}}$.

We note that the statement $\left\{ \bigcup_{g \in \mathbb{G}_0(\boldsymbol{\theta})} \{R_g(\mathbf{X}_n) = 1\} \right\}$ reads as: at least one true hypothesis has been rejected. The complement of the statement is therefore that no true hypotheses have been rejected and hence the strong control of the FWER implies that the true parameter value lies in the complement of union of the rejected subsets with probability $1 - \alpha$. That is, for all $\boldsymbol{\theta} \in \mathbb{T}$,

$$\mathrm{Pr}_{\boldsymbol{\theta}} \left(\boldsymbol{\theta} \in \bigcap_{g \in \mathbb{G}_1(X_n)} \mathbb{T}_g^{\complement} \right) \geq 1 - \alpha,$$

where $(\cdot)^{\complement}$ is the set complement operation and $\mathbb{G}_1(X_n) = \{g \in \mathbb{G} : R_g(X_n) = 1\}$ is the set of rejected hypotheses.

Define the set of closed tests corresponding to $\mathbb{H}$ as the rejection rules: $\left(\bar{R}_g(\mathbf{X}_n) \right)_{g \in \mathbb{G}}$, where for each $g \in \mathbb{G}$,

$$\bar{R}_g(\mathbf{X}_n) = \min_{\{j : \mathbb{T}_j \subseteq \mathbb{T}_g\}} R_j(\mathbf{X}_n). \tag{4}$$

all hypotheses $H_j, j \leq g$, are rejected, otherwise $\bar{R}_g(\mathbf{X}_n) = 0$. Then, we have the following result regarding the closed testing principle (cf. Dickhaus, 2014, Thm. 3.4).

Theorem 1 *For an $\cap$-closed* system of hypotheses $\mathbb{H}$ with corresponding α level local tests $\left(R_g(\mathbf{X}_n) \right)_{g \in \mathbb{G}}$, the closed testing procedure defined by $\left(\bar{R}_g(\mathbf{X}_n) \right)_{g \in \mathbb{G}}$ strongly controls the FWER at level α in the sense that

$$\mathrm{Pr}_{\boldsymbol{\theta}} \left(\bigcup_{g \in \mathbb{G}_0(\boldsymbol{\theta})} \{\bar{R}_g(\mathbf{X}_n) = 1\} \right) \leq \alpha,$$

for each $\boldsymbol{\theta} \in \mathbb{T}$.

We now demonstrate that the sequential procedure constitutes a set of closed tests of the form (4) and thus permits the conclusion of Theorem 1, which in turn implies (2) and thus (3). That is, we show that the sequence of hypotheses $\left(\mathrm{H}_g\right)_{g=1}^{\infty}$ corresponds to a $\cap$-closed system, where each H_g is defined by $f_0 \in \mathcal{M}_g$, and that the STP corresponds to a sequence of tests of form (4).

Theorem 2 *The hypotheses $\left(\mathrm{H}_g\right)_{g=1}^{\infty}$ and the STP from Sect. 1 constitute a $\cap$-closed system and a closed testing procedure, respectively, when testing using p-values $\left(P_g\left(\mathbf{X}_n\right)\right)_{g=1}^{\infty}$, satisfying (1). The sequential test therefore permit conclusions (2) and (3).*

Proof The proof of this result appears in the Appendix. $\qquad\square$

Thus, under the assumption that the data $\mathbf{X}_n$ arises from a DGP with density function f_0, corresponding to a g_0 component mixture model, the STP outputs a point estimator $\hat{g}_n$, where the event $\left\{g_0 \geq \hat{g}_n\right\}$ occurs with probability at least $1 - \alpha$.

3 Test of order via universal inference

Let $\mathbf{X}_n$ be split into two subsequences of lengths n_1 and n_2, where $\mathbf{X}_n^1 = \left(X_i\right)_{i=1}^{n_1}$ and $\mathbf{X}_n^2 = \left(X_i\right)_{i=n_1+1}^{n}$, and $n_1 + n_2 = n$. Assume that X has DGP characterized by the PDF f_0 and for each $g \in \mathbb{N}$, let $\hat{f}_g^1 \in \bar{\mathcal{M}}_g$ and $\hat{f}_g^2 \in \bar{\mathcal{M}}_g$ be estimators of f_0 (not necessarily maximum likelihood estimators), based on $\mathbf{X}_n^1$ and $\mathbf{X}_n^2$, respectively, where $\bar{\mathcal{M}}_g \subseteq \mathcal{M}$ is a class that characterizes an alternative to the null hypothesis that $f_0 \in \mathcal{M}_g$, with $\mathcal{M}_g \subset \bar{\mathcal{M}}_g$.

For notational convenience, for each $k \in \{1, 2\}$, we reindex the elements of $\mathbf{X}_n^k$ by inclusion of a superscript k, so that $\mathbf{X}_n^k = \left(X_i^k\right)_{i=1}^{n_k}$, and let

$$L_f\left(\mathbf{X}_n^k\right) = \prod_{i=1}^{n_k} f\left(X_i^k\right),$$

be the likelihood function corresponding to subsample $\mathbf{X}_n^k$, evaluated under PDF f. We wish to test the null hypothesis $\mathrm{H}_g\colon f_0 \in \mathcal{M}_g$ against the alternative $\bar{\mathrm{H}}_g\colon f_0 \in \bar{\mathcal{M}}_g$, using the Split test statistics

$$V_g^k\left(\mathbf{X}_n\right) = \frac{L_{\hat{f}_g^{3-k}}\left(\mathbf{X}_n^k\right)}{L_{\tilde{f}_g^k}\left(\mathbf{X}_n^k\right)},$$

for $k \in \{1, 2\}$, and the Swapped test statistic

$$\bar{V}_g\left(\mathbf{X}_n\right) = \frac{1}{2}\left\{V_g^1\left(\mathbf{X}_n\right) + V_g^2\left(\mathbf{X}_n\right)\right\},$$

as introduced in Wasserman et al. (2020). Here, the denominator estimator $\tilde{f}_g^k$ is the maximum likelihood estimator of f_0, based on $\mathbf{X}_n^k$ under the null hypothesis H_g, in the sense that

$$\tilde{f}_g^k \in \left\{ \tilde{f} \in \mathcal{M}_g \, : \, L_{\tilde{f}}\big(\mathbf{X}_n^k\big) = \max_{f \in \mathcal{M}_g} L_f\big(\mathbf{X}_n^k\big) \right\}.$$

We define the *p*-values for the Split and Swapped test statistics as $P_g^k(\mathbf{X}_n) = \min\{1/V_g^k(\mathbf{X}_n), 1\}$ and $\bar{P}_g(\mathbf{X}_n) = \min\{1/\bar{V}_g(\mathbf{X}_n), 1\}$, respectively. The adaptation of Wasserman et al. (2020, Thm. 3) demonstrates that the two tests have correct size for any sample size *n* (i.e., $P_g^k(\mathbf{X}_n)$ and $\bar{P}_g(\mathbf{X}_n)$ satisfy condition (1), for any *n*).

Theorem 3 *For any $n \in \mathbb{N}$ and $\alpha \in (0, 1)$,*

$$\sup_{f \in \mathcal{M}_g} \Pr_f\left(P_g^k(\mathbf{X}_n) \leq \alpha \right) \leq \alpha$$

and

$$\sup_{f \in \mathcal{M}_g} \Pr_f\big(\bar{P}_g(\mathbf{X}_n) \leq \alpha\big) \leq \alpha.$$

Theorem 3 implies that for each $g \in \mathbb{N}$ and $k \in \{1, 2\}$, and for any sample size $n \in \mathbb{N}$, if $f_0 \in \mathcal{M}_g$ is the DGP of $\mathbf{X}_n$, then events $\left\{ P_g^k(\mathbf{X}_n) \leq \alpha \right\}$ and $\{\bar{P}_g(\mathbf{X}_n) \leq \alpha\}$, corresponding to a rejection of the null hypothesis H_g, occur with probability no greater than α, as required for a test of size α.

It is suggested by Windham and Cutler (1992), Polymenis and Titterington (1998), and Wasserman et al. (2020) that the alternative hypothesis for each H_g should be that $f_0 \in \bar{\mathcal{M}}_g = \mathcal{M}_{g+1}$. However, since we are only looking to reject H_g, rather than making conclusions regarding the alternative, we can take $\bar{\mathcal{M}}_g$ to be a richer class of PDFs that is still feasible to estimate. Thus, in the sequel, we shall consider the possibility that $\bar{\mathcal{M}}_g = \mathcal{M}_{g+l_g}$ for some $l_g \in \mathbb{N}$, for each $g \in \mathbb{N}$. Typically, we can let $l_g = l$ for all g, but we anticipate that there may be circumstances where one may wish for l_g to vary.

3.1 Consistency of order tests

Although Theorem 3 guarantees the control of the Type I error for each local test of H_g, it makes no statement regarding the power of the tests. For tests against alternatives of the form: $\bar{\mathcal{M}}_g = \mathcal{M}_{g+l_g}$, we shall consider the issue of power from an asymptotic perspective in the parametric context. That is, we suppose that

$$\mathcal{K}(\mathbb{X}) = \{f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) \, : \, \boldsymbol{\theta} \in \mathbb{T}\}, \tag{5}$$

where $\mathbb{T} \subseteq \mathbb{R}^p$ for some $p \in \mathbb{N}$, and thus

$$\mathcal{M}_g(\mathbb{X}) = \left\{ f\big(\boldsymbol{x};\boldsymbol{\vartheta}^{(g)}\big) : f\big(\boldsymbol{x};\boldsymbol{\vartheta}^{(g)}\big) = \sum_{z=1}^{g} \pi_z f\big(\boldsymbol{x};\boldsymbol{\theta}_z\big); \pi_z \geq 0, \right.$$
$$\left. \sum_{z=1}^{g} \pi_z = 1, \boldsymbol{\theta}_g \in \mathbb{T}, z \in [g] \right\}. \tag{6}$$

We put the pairs $\big((\pi_z,\boldsymbol{\theta}_z)\big)_{z=1}^{g}$ in the vector $\boldsymbol{\vartheta}^{(g)} \in ([0,1] \times \mathbb{T})^g = \mathbb{T}_g$. Here, we further replace $\hat{f}_g^2$ and $\tilde{f}_g^1$ by $f\left(\cdot;\hat{\boldsymbol{\vartheta}}_n^{(g+l_g)}\right)$ and $f\left(\cdot;\tilde{\boldsymbol{\vartheta}}_n^{(g)}\right)$, respectively, where $\hat{\boldsymbol{\vartheta}}_n^{(g+l_g)}$ is a function of $\mathbf{X}_n^2$ and $\tilde{\boldsymbol{\vartheta}}_n^{(g)}$ is a function of $\mathbf{X}_n^1$. Further, since $f\left(\cdot;\tilde{\boldsymbol{\vartheta}}_n^{(g)}\right)$ is the maximum likelihood estimator of $f_0 \in \mathcal{M}_g$, we also write

$$\tilde{\boldsymbol{\vartheta}}_n^{(g)} \in \left\{ \tilde{\boldsymbol{\vartheta}}^{(g)} \in \mathbb{T}_g : \prod_{i=1}^{n_1} f\big(X_i^1;\tilde{\boldsymbol{\vartheta}}^{(g)}\big) = \max_{\boldsymbol{\vartheta}^{(g)} \in \mathbb{T}_g} \prod_{i=1}^{n_1} f\big(X_i^1;\boldsymbol{\vartheta}^{(g)}\big) \right\}. \tag{7}$$

Following DasGupta (2008, Def. 23.1), we say that a sequence of tests $\big(R_g(\mathbf{X}_n)\big)_{n=1}^{\infty}$ for H_g is consistent if under the true DGP, characterized by $f_0 \notin \mathcal{M}_g$, it is true that $\Pr_{f_0}\big(R_g(\mathbf{X}_n) = 1\big) \to 1$, as $n \to \infty$. Let $\|\cdot\|$ denote the Euclidean norm and define the Kullback–Leibler divergence between two PDFs on $\mathbb{X}$: f_1 and f_2, as

$$\mathrm{D}\big(f_1,f_2\big) = \int_{\mathbb{X}} f_1(\boldsymbol{x}) \log \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}.$$

Further, say that a class of parametric mixture models $\mathcal{M}_g$ is identifiable if

$$\sum_{z=1}^{g} \pi_z f\big(\boldsymbol{x};\boldsymbol{\theta}_z\big) = \sum_{z=1}^{g} \pi_z' f\big(\boldsymbol{x};\boldsymbol{\theta}_z'\big)$$

if and only if $\sum_{z=1}^{g} \pi_z \mathbf{1}\big(\boldsymbol{\theta} = \boldsymbol{\theta}_z\big) = \sum_{z=1}^{g} \pi_z' \mathbf{1}\big(\boldsymbol{\theta} = \boldsymbol{\theta}_z'\big)$, where $\mathbf{1}(\cdot)$ is the usual indicator function. For $R_g(\mathbf{X}_n) = \mathbf{1}\big(P_g^1(\mathbf{X}_n) < \alpha\big)$, where $P_g^1(\mathbf{X}_n)$ is obtained from testing H_g against the alternative $\bar{\mathcal{M}}_g = \mathcal{M}_{g+l_g}$, we obtain the following result. The equivalent result regarding $\bar{P}_g(X_n)$ can be established analogously.

Theorem 4 *Make the following assumptions*:

(A1)　*for each $g \in \mathbb{N}$, the class $\mathcal{M}_g$ is identifiable;*

(A2)　*the PDF $f(\boldsymbol{x};\boldsymbol{\theta}) > 0$ is everywhere positive and continuous for all $(\boldsymbol{x},\boldsymbol{\theta}) \in \mathbb{X} \times \mathbb{T}$, where $\mathbb{X}$ and $\mathbb{T}$ are Euclidean spaces and $\mathbb{T}$ is compact;*

(A3)　*for all $\boldsymbol{x} \in \mathbb{X}$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{T}, \left|\log f\big(\boldsymbol{x};\boldsymbol{\theta}_1\big)\right| \leq M_1(\boldsymbol{x})$ and*

$$\left|\log f\big(\boldsymbol{x};\boldsymbol{\theta}_1\big) - \log f\big(\boldsymbol{x};\boldsymbol{\theta}_2\big)\right| \leq M_2(\boldsymbol{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|,$$

where $E_{f_0} M_1(X) < \infty$ *and* $E_{f_0} M_2(X) < \infty$;

(A4)   *the estimator* $\hat{\boldsymbol{\vartheta}}_n^{(g+l_g)} \to \boldsymbol{\vartheta}_0^{(g+l_g)}$, *in probability, as* $n_2 \to \infty$, *where*

$$\boldsymbol{\vartheta}_0^{(g+l_g)} \in \left\{ \hat{\boldsymbol{\vartheta}}^{(g+l_g)} \in \mathbb{T}_{g+l_g} : E_{f_0} \log f\left(X; \hat{\boldsymbol{\vartheta}}^{(g+l_g)}\right) \right.$$

$$\left. = \max_{\boldsymbol{\vartheta}^{(g+l_g)} \in \mathbb{T}_{g+l_g}} E_{f_0} \log f\left(X; \boldsymbol{\vartheta}^{(g+l_g)}\right) \right\}.$$

Under Assumptions (A1)–(A4), *if* $f_0 \in \mathcal{M} \backslash \mathcal{M}_g$, *and* $n_1, n_2 \to \infty$, *then* $R_g(\mathbf{X}_n) = \mathbf{1}\left(P_g^1(\mathbf{X}_n) < \alpha\right)$ *is a consistent test for* H_g.

Proof   The proof of this result appears in the Appendix.   □

Assumption A1 ensures that the elements of (6) [i.e., the g component mixtures of densities of class (5)] are distinct (as noted in Titterington et al., 1985, Sec. 3.1.1), and A2 implies that the log-likelihood cannot take infinitely negative values and that it is continuous for any x and θ, where the compactness of $\mathbb{T}$ ensures that $f(x; \theta)$ is bounded for each fixed $x \in \mathbb{X}$. Assumption A3 then implies that the expected log-likelihood $E_{f_0} \log f(X; \theta)$ is bounded for each θ, and since $f(x; \theta)$ is continuous, $E_{f_0} \log f(X; \theta)$ is also continuous and thus has global optima within the compact set $\mathbb{T}$. Assumption A3 also implies that $E_{f_0} \log f(X; \theta)$ is Lipschitz continuous, with respect to $\theta \in \mathbb{T}$ equipped with the Euclidean norm, and A4 implies that $\hat{\boldsymbol{\vartheta}}_n^{(g+l_g)}$, characterizing $\hat{f}_g^2$, behaves asymptotically (with respect to convergence in probability) like a parametric maximum likelihood estimator, under the potentially misspecified supposition that $f_0 \in \mathcal{M}_{g+l_g}$.

Assumptions A1–A3 are required for the application of Leroux (1992, Lem. 1), with A2 and A3 also required for establishing the consistency of the estimators $\tilde{\boldsymbol{\vartheta}}_n^{(g)}$. The Lipschitz condition of A3 and A4 are further required to show that the logarithm of the split test statistic $V_g^1(\mathbf{X}_n)$ is a consistent estimator of an difference in divergence expression required in the proof.

Theorem 4 states that for each $g < g_0$ and for any significance level $\alpha \in (0, 1)$, the rejection probability of the test of H_g, based on $P_g^1(\mathbf{X}_n)$, converges to 1, as n gets large. We note that Assumptions A1–A4 are verifiable for typical models of interest. For example, when $\mathcal{M}_g$ is the class of g component normal mixture models (see Sect. 4), A1 is verified due to Yakowitz and Spragins (1968), A2 is satisfied by the usual compact restrictions on the parameter space (see, e.g., Ritter, 2014, Sec. B.6.2), and A3 is satisfied under A2.

4 Normal mixture models

We apply the STP with the Split and Swapped tests from Sect. 3 to the classic problem of order selection for normal mixture models, whereby $\mathbb{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and

$$\mathcal{K}(\mathbb{X}) = \mathcal{K}(\mathbb{R}^d) = \left\{ f(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) : \phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \right.$$
$$\left. \times \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \right] \right\},$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is symmetric positive definite.

To assess the performance of the STP, we conduct a thorough simulation study, within the R programming environment (R Core Team, 2020). For each $d \in \{2, 4\}$, we generate data sets $\mathbf{X}_n$, with $n_1 = n_2 \in \{300, 500, 1000, 2000, 5000, 10{,}000\}$ observations (recall that $n = n_1 + n_2$), where each $X_i \in \mathbb{R}^d$, from a multivariate normal mixture model in $\mathcal{M}_{g_0}(\mathbb{R}^d)$ for $g_0 \in \{5, 10\}$, with parameter elements $(\pi_z, \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)_{z=1}^{g_0}$ of $\mathcal{M}_{g_0}(\mathbb{R}^d)$ generated using the MixSim package (Melnykov et al., 2012), using the setting $\bar{\omega} \in \{0.01, 0.05, 0.1\}$ and $\min_{z \in [g_0]} \pi_z \geq (2g_0)^{-1}$. Here, the $\bar{\omega}$ parameter is described in Melnykov et al. (2012), and controls the level of overlap between the normal components of the mixture model. Four examples of data sets generated using various combinations of simulation parameters $(g_0, \bar{\omega})$, with $d = 2$ and $n_1 = 1000$, are provided in Fig. 1.

For each set of simulation parameters $(g_0, \bar{\omega}, d, n_1)$, we simulate $r = 100$ replicate data sets, whereupon we apply the STP at the $\alpha = 0.05$ level, using the Split and Swapped test p-values of the forms P_g^1 and $\bar{P}_g$, for each of the r data sets. To compute the maximum likelihood estimators $\tilde{f}_g^k = f\left(\cdot; \tilde{\boldsymbol{\vartheta}}_n^{(g)}\right)$, under the null hypotheses that $f_0 \in \mathcal{M}_g$, we use the gmm_full function from the Armadillo C++ library, implemented in R using the RcppArmadillo package (Eddelbuettel & Sanderson, 2014). We also use the maximum likelihood estimator as $\hat{f}_g^{3-k} = f\left(\cdot; \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)}\right)$, under the alternative hypotheses $f_0 \in \bar{\mathcal{M}}_g = \mathcal{M}_{g+l_g}$, where we set $l_g = l \in \{1, 2\}$, for all $g \in \mathbb{N}$. From each of the r STP results, we compute the coverage proportion (CovProp; proportion of r for which $g_0 \geq \hat{g}_n$), the mean estimated number of components (MeanComp; the average of $\hat{g}_n$ over the r repetitions), and the proportion of times that the estimated number of components corresponded with the g_0 (CorrProp; the proportion of times the event $\hat{g}_n = g_0$ occurs out of the r repetitions).

It is worth recalling that our estimators $\hat{g}_n$ are not in fact point estimators of g_0, but are actually lower bounds of the STP interval estimators $\{g_0 \geq \hat{g}_n\}$ and are thus only expected to be close g_0, with $\hat{g}_n = g_0$ indicating that the interval is efficient. We can complement the output of the interval estimator with a point estimator of g_0, such as via the AIC or BIC procedures, whereupon for data set $\mathbf{X}_n$, and for each $g \in \mathbb{N}$, the AIC or BIC values:
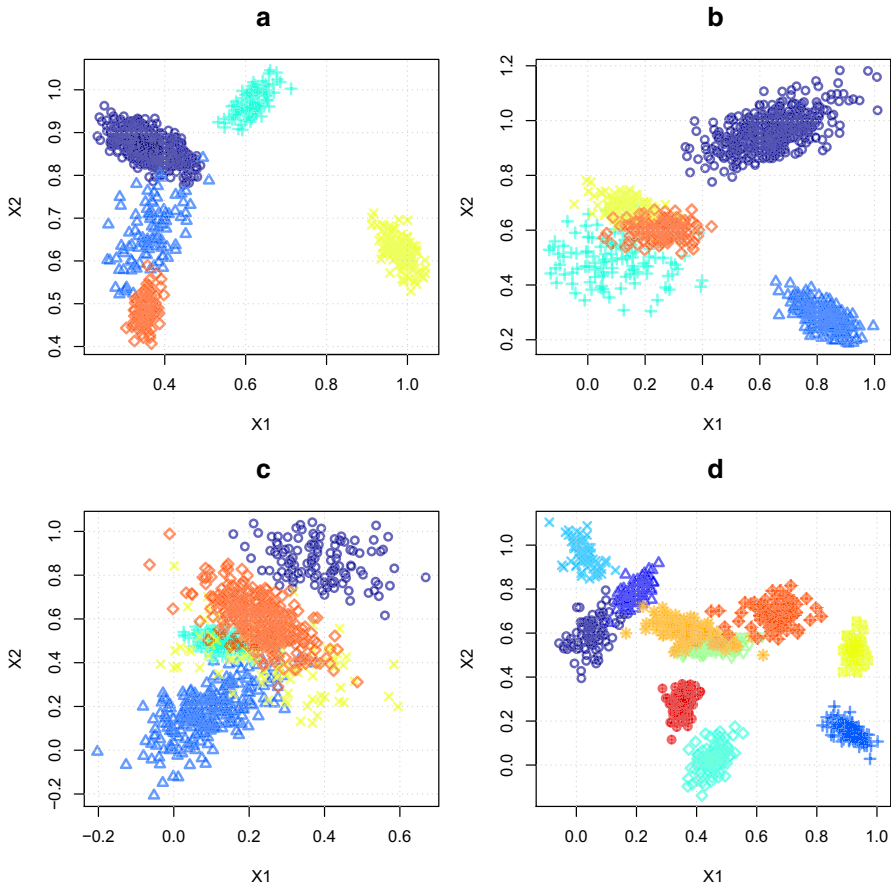
Fig. 1 Example data sets of $n_1 = 1000$ random observations from a $d = 2$ dimensional g_0 component normal mixture model, with parameters determined via parameter $\bar{\omega}$. Here, the pairs $(g_0, \bar{\omega})$ visualized in subplots a, b, c, and d are $(5, 0.01)$, $(5, 0.05)$, $(5, 0.1)$, and $(10, 0.01)$, respectively

$$\text{AIC}_g = \frac{2}{n}\dim_g - \frac{2}{n}\sum_{i=1}^{n}\log f\left(X_i; \boldsymbol{\vartheta}_n^{(g)}\right), \text{ and}$$

$$\text{BIC}_g = \frac{\log n}{n}\dim_g - \frac{2}{n}\sum_{i=1}^{n}\log f\left(X_i; \boldsymbol{\vartheta}_n^{(g)}\right), \tag{8}$$

respectively, are computed. Here $\dim_g$ is the dimensionality of $\boldsymbol{\vartheta}^{(g)}$ and $\boldsymbol{\vartheta}_n^{(g)}$ is a maximum likelihood estimator of form (7), computed using X_n instead of X_n^1. The AIC and BIC procedures then estimate g_0 via $\arg\min_g \text{AIC}_g$ or $\arg\min_g \text{BIC}_g$, respectively. To complement our results regarding $\{g_0 \geq \hat{g}_n\}$, we also provide the Mean-Comp and CorrProp values for the AIC and BIC procedures. All of our R scripts are made available at https://github.com/ex2o/oscfmm.

4.1 Simulation results

For all scenario combinations $(g_0, \bar{\omega}, d, n_1, l)$, the CovProp was 100% over the r repetitions. This confirms the conclusions of Theorems 2 and 3. This also implies that the tests are underpowered, which is conforming to the observations from the simulations of Wasserman et al. (2020). This result is unsurprising since the tests are constructed via a Markov inequality argument, which makes no use of the topological features of the sets $\mathcal{M}_g$ and $\bar{\mathcal{M}}_g$ that can be used to derive more specific results.

We report the STP interval estimator, and AIC and BIC point estimator results for all of the combinations $(g_0, \bar{\omega}, d, n_1, l)$, partitioned by $(g_0, \bar{\omega})$ in Tables 1, 2, 3, 4, 5 and 6. Here, Tables 1, 2, 3, 4, 5 and 6 contain results for pairs (5, 0.01), (5, 0.05), (5, 0.1), (10, 0.01), (10, 0.05), and (10, 0.1), respectively. In the (5, 0.01) case, we observe that both the Split and Swapped test-based STPs were able to identify the generative value of g in over 90% of the cases, except when $n_1 < 1000$ and for the case $(d, n_1, l) = (4, 1000, 2)$. There is some evidence that the Swapped test is more

Table 1 MeanComp and CorrProp results for different values of (d, n_1, l), when $(g_0, \bar{\omega}) = (5, 0.01)$

d	n_1	l	MeanComp				CorrProp			
			Split	Swapped	AIC	BIC	Split	Swapped	AIC	BIC
2	300	1	4.25	4.52	4.99	4.49	0.50	0.65	0.49	0.50
		2	4.53	4.63			0.56	0.63		
	500	1	4.69	4.89	4.95	4.50	0.82	0.92	0.52	0.50
		2	4.81	4.93			0.81	0.93		
	1000	1	4.88	4.95	5.20	5.00	0.95	0.98	0.87	1.00
		2	4.92	4.96			0.92	0.96		
	2000	1	4.89	4.94	5.13	5.00	0.96	0.98	0.88	1.00
		2	5.00	5.00			1.00	1.00		
	5000	1	4.94	4.98	5.05	5.00	0.97	0.99	0.96	1.00
		2	5.00	5.00			1.00	1.00		
	10000	1	4.93	4.96	5.05	5.00	0.97	0.98	0.95	1.00
		2	4.99	4.99			0.99	0.99		
4	300	1	3.45	3.87	5.38	4.46	0.08	0.18	0.42	0.47
		2	4.03	4.13			0.11	0.16		
	500	1	4.31	4.47	5.22	4.50	0.43	0.53	0.40	0.49
		2	4.34	4.41			0.34	0.41		
	1000	1	4.87	4.97	5.25	5.00	0.92	0.97	0.79	1.00
		2	4.85	4.86			0.85	0.86		
	2000	1	4.94	4.98	5.16	5.00	0.97	0.99	0.88	1.00
		2	5.00	5.00			1.00	1.00		
	5000	1	4.98	5.00	5.03	5.00	0.99	1.00	0.98	1.00
		2	5.00	5.00			1.00	1.00		
	10000	1	4.98	4.98	5.07	5.00	0.99	0.99	0.93	1.00
		2	5.00	5.00			1.00	1.00		

Table 2 MeanComp and CorrProp results for different values of (d, n_1, l), when $(g_0, \bar{\omega}) = (5, 0.05)$

d	n_1	l	MeanComp				CorrProp			
			Split	Swapped	AIC	BIC	Split	Swapped	AIC	BIC
2	300	1	3.60	3.77	5.16	4.17	0.11	0.14	0.40	0.36
		2	3.87	4.02			0.09	0.13		
	500	1	3.89	4.09	5.08	4.30	0.26	0.35	0.47	0.41
		2	4.20	4.32			0.23	0.35		
	1000	1	4.41	4.52	5.13	4.92	0.57	0.62	0.89	0.92
		2	4.51	4.60			0.53	0.60		
	2000	1	4.82	4.92	5.10	4.96	0.88	0.94	0.91	0.96
		2	4.85	4.88			0.85	0.88		
	5000	1	4.96	4.96	5.13	5.00	0.98	0.98	0.89	1.00
		2	4.90	4.93			0.90	0.93		
	10000	1	4.92	4.92	5.02	5.00	0.96	0.96	0.99	1.00
		2	4.95	4.98			0.97	0.98		
4	300	1	2.73	2.83	5.63	3.89	0.00	0.00	0.40	0.25
		2	3.00	3.12			0.00	0.00		
	500	1	3.09	3.42	5.55	4.21	0.02	0.04	0.38	0.36
		2	3.72	3.84			0.04	0.03		
	1000	1	3.98	4.25	5.41	4.95	0.22	0.37	0.72	0.95
		2	4.20	4.27			0.21	0.27		
	2000	1	4.71	4.85	5.19	4.98	0.77	0.87	0.85	0.98
		2	4.73	4.79			0.73	0.79		
	5000	1	4.96	5.00	5.10	5.00	0.98	1.00	0.91	1.00
		2	4.96	4.98			0.96	0.98		
	10000	1	4.98	4.98	5.06	5.00	0.99	0.99	0.94	1.00
		2	4.99	5.00			0.99	1.00		

powerful than Split test in all cases, as indicated by the higher values of MeanComp and CorrProp. Furthermore, the $l = 2$ alternative appears to be more powerful than the $l = 1$ alternative in all cases except when $(d, n_1) = (4, 1000)$.

For the other pairs of $(g_0, \bar{w})$, we observe the same relationships between the values of l and the Split and Swapped tests. That is, $l = 2$ tends to be more powerful than $l = 1$ (except when n_1 is relatively small, i.e. $n_1 \in \{300, 500, 1000, 2000\}$), and the Swapped test tends to be more powerful than the Split test. In addition, we also observe that the STP becomes more powerful as n_1 increases, which supports the conclusions of Theorem 4, which applies to the normal mixture model that is under study.

For smaller sample sizes, we observe that the STP tended to be more powerful when $d = 2$ in almost all cases, and for larger sample sizes, the opposite appears to be true. This is likely due to a combination of the variability of the maximum likelihood estimator and the increase in separability of higher dimensional spaces.

Table 3 MeanComp and CorrProp results for different values of (d, n_1, l), when $(g_0, \bar{\omega}) = (5, 0.1)$

d	n_1	l	MeanComp				CorrProp			
			Split	Swapped	AIC	BIC	Split	Swapped	AIC	BIC
2	300	1	2.85	3.16	5.08	3.63	0.00	0.02	0.43	0.17
		2	3.19	3.35			0.01	0.01		
	500	1	3.33	3.65	5.01	4.02	0.06	0.09	0.43	0.24
		2	3.62	3.78			0.05	0.09		
	1000	1	3.86	3.98	5.13	4.60	0.18	0.22	0.83	0.60
		2	4.13	4.22			0.20	0.26		
	2000	1	4.41	4.60	5.11	4.88	0.57	0.67	0.87	0.88
		2	4.51	4.58			0.51	0.58		
	5000	1	4.71	4.75	5.03	4.95	0.77	0.80	0.97	0.95
		2	4.83	4.84			0.83	0.84		
	10000	1	4.87	4.88	5.02	4.98	0.90	0.91	0.98	0.98
		2	4.88	4.90			0.88	0.90		
4	300	1	2.13	2.33	5.69	2.85	0.00	0.00	0.34	0.01
		2	2.39	2.54			0.00	0.00		
	500	1	2.58	2.75	5.63	3.61	0.00	0.00	0.35	0.16
		2	3.01	3.22			0.00	0.00		
	1000	1	3.44	3.70	5.61	4.75	0.03	0.09	0.61	0.76
		2	3.82	3.94			0.01	0.03		
	2000	1	4.05	4.35	5.26	4.95	0.32	0.44	0.82	0.95
		2	4.35	4.48			0.35	0.48		
	5000	1	4.96	4.97	5.09	5.00	0.96	0.97	0.92	1.00
		2	4.87	4.91			0.87	0.91		
	10000	1	4.98	5.00	5.03	5.00	0.99	1.00	0.97	1.00
		2	5.00	5.00			1.00	1.00		

Finally, we notice that the STP was more powerful when the data were more separable (i.e., for smaller values of $\bar{\omega}$). Here, we can see that for $n_1 = 10,000$, the STP can identify the generative value of g in the $g_0 = 5$ scenarios, in a large proportion of cases. However, when $g_0 = 10$, the STP becomes less powerful. It is particularly remarkable that even when $n_1 = 10,000$, the highest detection proportion was 7% in the $(g_0, \bar{\omega}) = (10, 0.1)$ scenarios. This again implies that the STP lacks power, when applied with the Split or Swapped tests, especially when component densities of the generative mixture model are not well separated.

Regarding the AIC and BIC point estimators, we observe firstly that across all scenarios, the AIC procedure produces a larger estimate of g_0 than the BIC procedure, when observing the MeanComp values. We also observe that the AIC estimator is often larger than g_0, even for larger sample sizes. This observation is in concordance with the theory of Leroux (1992) and Keribin (2000) who show that the AIC procedure does not underestimate g_0, asymptotically, but is also not consistent. On the other hand, we observe that the BIC procedure tends

Table 4 MeanComp and CorrProp results for different values of (d, n_1, l), when $(g_0, \bar{\omega}) = (10, 0.01)$

d	n_1	l	MeanComp				CorrProp			
			Split	Swapped	AIC	BIC	Split	Swapped	AIC	BIC
2	300	1	4.80	6.12	10.13	9.12	0.01	0.01	0.45	0.37
		2	6.71	7.48			0.03	0.06		
	500	1	5.90	6.81	10.14	9.33	0.01	0.03	0.44	0.43
		2	7.98	8.33			0.12	0.14		
	1000	1	6.46	7.50	10.23	9.92	0.04	0.13	0.82	0.92
		2	8.69	9.05			0.25	0.29		
	2000	1	7.59	8.42	10.17	9.98	0.29	0.40	0.86	0.98
		2	9.31	9.52			0.47	0.58		
	5000	1	8.76	9.14	10.07	10.00	0.69	0.78	0.94	1.00
		2	9.50	9.79			0.76	0.84		
	10000	1	8.81	9.19	10.03	10.00	0.75	0.82	0.97	1.00
		2	9.60	9.81			0.82	0.90		
4	300	1	3.52	3.89	10.43	8.01	0.00	0.00	0.36	0.12
		2	4.35	4.84			0.00	0.00		
	500	1	4.66	5.26	10.35	8.84	0.00	0.01	0.38	0.27
		2	5.60	6.17			0.00	0.01		
	1000	1	5.83	6.50	10.47	9.85	0.00	0.00	0.68	0.85
		2	7.17	7.77			0.00	0.00		
	2000	1	7.04	7.82	10.34	9.99	0.07	0.12	0.75	0.99
		2	8.81	9.14			0.16	0.26		
	5000	1	8.54	9.25	10.08	9.99	0.47	0.60	0.93	0.99
		2	9.65	9.73			0.73	0.76		
	10000	1	9.37	9.68	10.06	10.00	0.84	0.91	0.95	1.00
		2	9.87	9.93			0.90	0.93		

to underestimate g_0, for small sample sizes, but becomes more accurate, on average, as n_1 increases. Again, this is in concordance with the consistency results regarding the BIC estimator of Keribin (2000). Regarding the CorrPro values, we observe that in smaller sample sizes ($n_1 \in \{300, 500\}$), the AIC procedure outperforms the BIC procedure in all cases other than those reported in Table 1. This is likely due to the downward bias of the BIC estimates, as observed via the MeanComp values. This downward bias appears to be most apparent in situations where the data are less separable and for larger g_0, as is evident by the results of Table 6, where the AIC procedure outperforms the BIC procedure with respect to CorrProp, across all sample sizes. In comparison to the STP estimates $\hat{g}_n$, as expected, it is notable that the AIC and BIC procedures provide point estimates of g_0 that are as accurate or more accurate in all simulation scenarios.

Overall, we observe that the conclusions of Theorems 2–4 appear to hold over the assessed simulation scenarios. From a practical perspective we can make the

Table 5 MeanComp and CorrProp results for different values of (d, n_1, l), when $(g_0, \bar{\omega}) = (10, 0.05)$

d	n_1	l	MeanComp				CorrProp			
			Split	Swapped	AIC	BIC	Split	Swapped	AIC	BIC
2	300	1	3.46	4.18	9.72	6.46	0.00	0.00	0.24	0.01
		2	4.07	4.60			0.00	0.00		
	500	1	4.10	4.91	9.88	7.15	0.00	0.00	0.35	0.03
		2	4.95	5.46			0.00	0.01		
	1000	1	4.78	5.44	10.06	8.87	0.01	0.01	0.74	0.27
		2	6.11	6.68			0.00	0.00		
	2000	1	5.74	6.56	9.96	9.33	0.01	0.01	0.88	0.52
		2	6.96	7.85			0.02	0.11		
	5000	1	7.21	7.82	10.02	9.74	0.06	0.11	0.98	0.78
		2	8.59	8.91			0.14	0.19		
	10000	1	7.62	8.38	9.99	9.88	0.17	0.26	0.95	0.88
		2	8.86	9.04			0.24	0.33		
4	300	1	2.45	2.64	10.54	4.36	0.00	0.00	0.30	0.00
		2	2.75	2.97			0.00	0.00		
	500	1	2.96	3.30	10.66	5.54	0.00	0.00	0.32	0.00
		2	3.41	3.71			0.00	0.00		
	1000	1	3.89	4.32	10.45	7.89	0.00	0.00	0.68	0.05
		2	4.98	5.30			0.00	0.00		
	2000	1	5.07	5.63	10.41	9.28	0.00	0.00	0.75	0.45
		2	6.30	6.63			0.00	0.00		
	5000	1	6.63	7.45	10.11	9.87	0.00	0.04	0.89	0.87
		2	8.35	8.71			0.01	0.05		
	10000	1	8.25	8.57	10.04	9.97	0.13	0.18	0.96	0.97
		2	9.23	9.36			0.32	0.42		

following recommendations. Firstly, the STP based on the Swapped test is preferred over the Split test. Secondly, the alternative based on $l = 2$ is preferred over $l = 1$. Thirdly, to obtain intervals of a fixed level of efficiency, larger sample sizes are necessary when data arise from mixture models with larger numbers of mixture components and when the mixture components are not well separated. Finally, we note that the AIC and BIC procedures both provide accurate point estimation of g_0 and are both complementary to the interval estimator $\{g_0 \geq \hat{g}_n\}$ obtained from the STP.

4.2 Example applications

We procedure to demonstrate the utility of the STP via example applications of varying complexity.

Table 6 MeanComp and CorrProp results for different values of (d, n_1, l), when $(g_0, \bar{\omega}) = (10, 0.1)$

d	n_1	l	MeanComp				CorrProp			
			Split	Swapped	AIC	BIC	Split	Swapped	AIC	BIC
2	300	1	2.76	3.09	9.08	4.61	0.00	0.00	0.16	0.00
		2	3.13	3.57			0.00	0.00		
	500	1	3.10	3.51	9.28	5.24	0.00	0.00	0.21	0.00
		2	3.98	4.56			0.00	0.00		
	1000	1	4.14	4.74	9.61	6.72	0.00	0.00	0.52	0.03
		2	5.02	5.28			0.00	0.00		
	2000	1	4.99	5.55	9.79	7.93	0.00	0.00	0.77	0.08
		2	5.87	6.30			0.00	0.00		
	5000	1	6.02	6.51	9.86	8.82	0.00	0.00	0.85	0.25
		2	7.16	7.60			0.00	0.01		
	10000	1	6.75	7.45	9.95	9.41	0.02	0.07	0.93	0.53
		2	7.75	8.06			0.04	0.06		
4	300	1	1.89	2.01	10.44	2.77	0.00	0.00	0.26	0.00
		2	1.92	2.17			0.00	0.00		
	500	1	2.27	2.53	10.53	3.66	0.00	0.00	0.28	0.00
		2	2.57	2.73			0.00	0.00		
	1000	1	3.07	3.37	10.32	5.64	0.00	0.00	0.69	0.00
		2	3.61	3.82			0.00	0.00		
	2000	1	4.12	4.52	10.39	7.54	0.00	0.00	0.74	0.04
		2	4.99	5.22			0.00	0.00		
	5000	1	5.40	6.09	10.08	9.23	0.00	0.00	0.89	0.40
		2	6.95	7.32			0.00	0.00		
	10000	1	6.90	7.60	10.06	9.77	0.01	0.03	0.97	0.79
		2	8.38	8.54			0.02	0.02		

4.2.1 Old faithful data

Our first example is to assess the number of Gaussian mixture components that are present in the faithful data set from R, which was originally studied in Hardle (1991). The data set consists of a length $n = 272$ realizations of a sequence $\mathbf{X}_n$, where $X_i \in \mathbb{R}^2$ for each $i \in [n]$. Here, each observation $X_i = (X_{i1}, X_{i2})$ contains measurements regarding the eruption length of time X_{i1} and the waiting time until the next eruption X_{i2}, in minutes, of eruption event i, for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. A visualization of the data appears in Fig. 2.

We apply the STP using a $n_1 = n_2 = 136$ split. The p-values obtained from the Split tests of hypotheses H_g versus $\bar{H}_g$ with $l_g = 2$, for $g = 1, 2$ are 3.40×10^{-32} and 1. Respectively, the p-values for the Swapped test are 6.80×10^{-32} and 1. Thus, using either the Split or the Swapped test variants of the STP, for
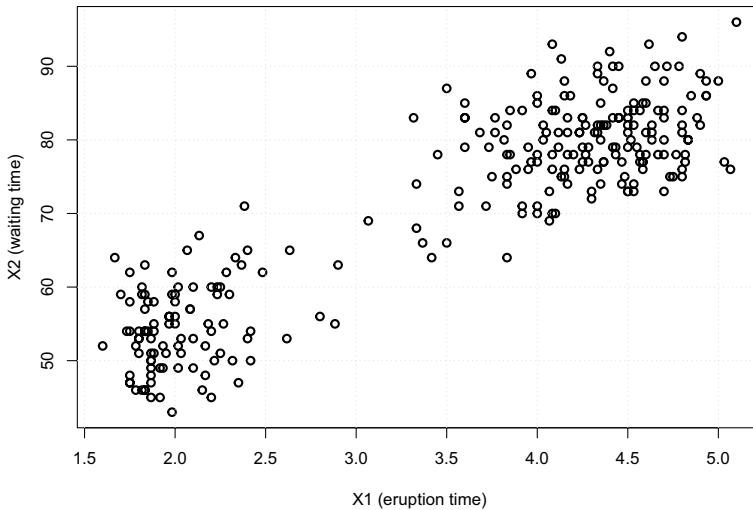
Fig. 2 Scatter plot of the `faithful` data set

$\alpha > 6.80 \times 10^{-32}$, we can conclude that the event $\{g_0 \geq 2\}$ occurs with a probability of at least $1 - \alpha$. We also obtain the AIC_g values for $g = 1, 2, 3$: 9.53, 8.40, and 8.42, and the respective BIC_g values: 9.61, 8.56, and 8.66. Thus, both procedures estimate the order of the underlying mixture distribution to be 2. Although there is no ground truth regarding the `faithful` data set, a visual inspection of Fig. 2 suggests that both the STP interval estimator and the point estimation provided by the AIC and BIC procedures are reasonable.

4.2.2 Palmer penguins data

Our second example is to estimate the Gaussian mixture order of the `penguins` data set from the R package palmerpenguins, originally considered by Gorman et al. (2014). After removing rows with missing data, the data set contains a length $n = 342$ realization of a sequence $\mathbf{X}_n$, where $X_i \in \mathbb{R}^4$ for each $i \in [n]$. Here, each observation $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ contains measures regarding penguins of the Adelie, Gentoo, and Chinstrap species. Specifically, for each i, the measurements are the bill length X_{i1}, bill depth X_{i2}, and flipper length X_{i3}, all in millimeters, along with the body mass X_{i4}, in grams. A visualization of the data, with separate symbols for the different penguin species, is provided in Fig. 3.

We apply the STP using a $n_1 = n_2 = 171$ split. The p-values obtained from the Split tests of hypotheses H_g versus $\bar{H}_g$ with $l_g = 2$, for $g = 1, 2$ are 3.78×10^{-61} and 1. Respectively, the p-vales for the Swapped test are 9.25×10^{-66} and 1. Thus, using either version of the STP, for $\alpha > 3.78 \times 10^{-61}$, we conclude that the event $\{g_0 \geq 2\}$ occurs with a probability of at least $1 - \alpha$. For these data, the AIC_g values for $g = 1, 2, 3, 4, 5$ are 32.37, 30.65, 30.38, 30.35, and 30.36.,
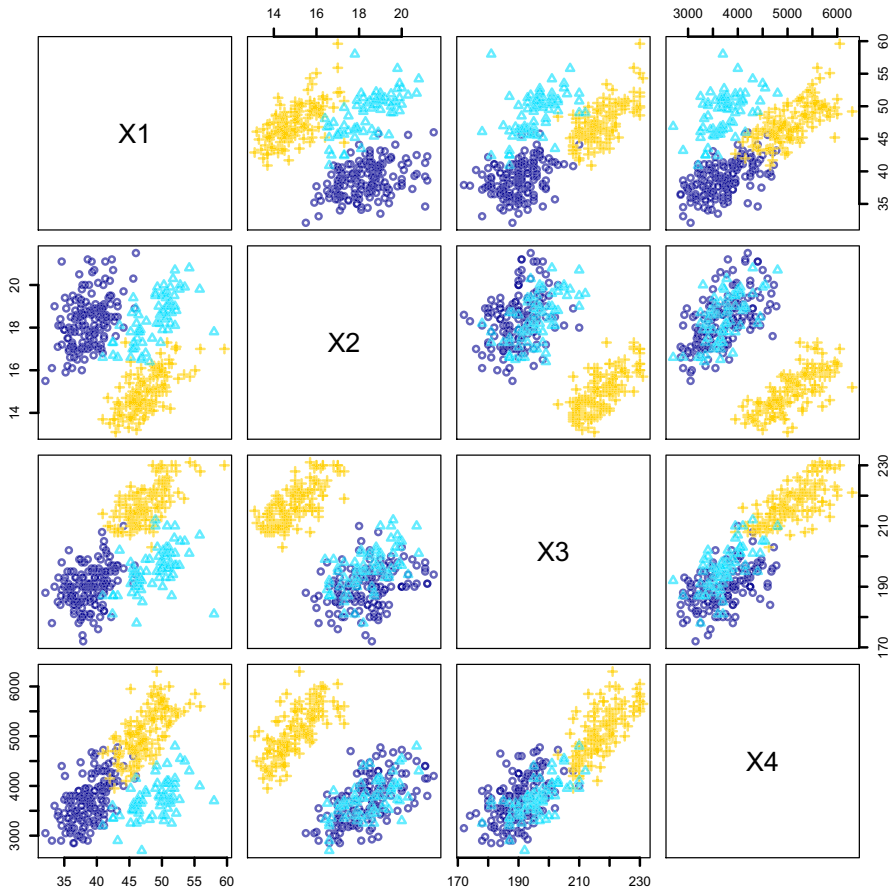
Fig. 3 Scatter plot of the `penguins` data set. Adelie, Chinstrap, and Gentoo data points are plotted as circles, triangles, and plus signs, respectively

and the respective BIC_g values are 32.54, 30.99, 30.89, 31.03, and 31.20. Thus, the AIC and BIC procedures estimate the true order g_0 to be 4 and 3, respectively. Compared to the ground truth of three penguin species, we observe that the AIC procedure is an over estimate, whereas the BIC is accurate. The inference obtained from the STP is also correct, with the assessment that there are at least 2 mixture components, with high probability.

4.2.3 Cell lines data set

Our final example is to identify the number of mixture components in the `cell_lines` data set from the `harmony` package of Korsunsky et al. (2019). As presented in https://portals.broadinstitute.org/harmony/articles/quickstart.html, the data set consists of $n = 2370$ rows consisting of a realization of the sequence $\mathbf{X}_n$ of random variable $X_i \in \mathbb{R}^2$, for each $i \in [n]$. Each observation $X_i = (X_{i1}, X_{i2})$
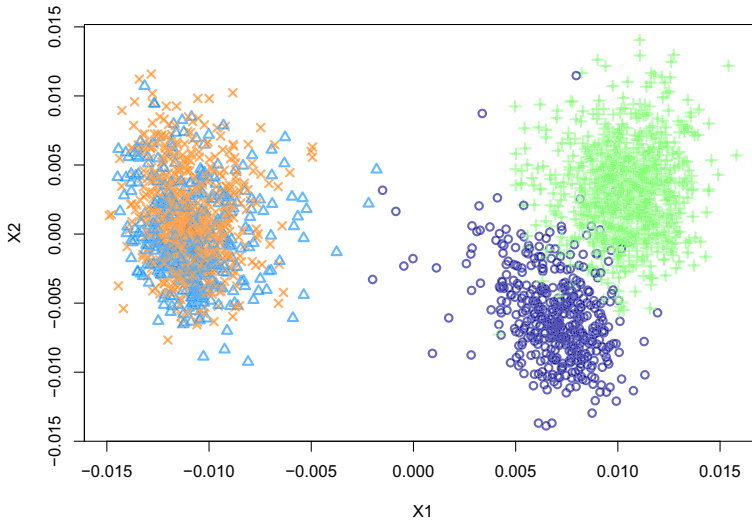
Fig. 4 Scatter plot of the `cell_lines` data set. The pure and mixed Jurkat cells data are plotted as plus signs and circles, respectively, and the pure and mixed HEK293T data are plotted as crosses and triangles, respectively

contains measurements of the first and second scaled principal components of single cell gene expression data. The data come from three sources, where the first source comes form a pure Jurkat cell lines, the second comes from a pure HEK293T cell lines, and the third source consists of a half-and-half mix of Jurkat cells and HEK293T cells. Since the data from the mixed sources are not registered to the pure sources data, there are in effect four separate subpopulations of observations. We plot the `cell_lines` data in Fig. 4.

We apply the STP using a $n_1 = n_2 = 1185$ split. The p-values obtained from the Split tests of hypotheses H_g versus $\bar{H}_g$ with $l_g = 2$, for $g = 1, 2, 3, 4$, are 0 (in double precision zero), 5.22×10^{-49}, 5.70×10^{-13}, and 0.21. Respectively, the Swapped tests yield p-values 0, 1.04×10^{-48}, 2.39×10^{-17}, and 0.42. Thus, for any $\alpha > 5.70 \times 10^{-13}$, the STP concludes that the event $\{g_0 \geq 4\}$ occurs with a probability of at least $1 - \alpha$. Again, we compute the AIC_g and BIC_g values. For each $g = 1, 2, 3, 4, 5, 6, 7, 8$, the AIC_g values are $-14.23, -16.32, -16.45, -16.52, -16.54,$ $-16.55, -16.56,$ and -16.55, respectively. Thus, The AIC procedure estimates g_0 as 7. For each $g = 1, 2, 3, 4, 5, 6$, the BIC_g values are $-14.22, -16.29, -16.41,$ $-16.46, -16.47,$ and -16.46. Thus, the BIC procedure estimates the mixture order to be 5. Compared to the ground truth, it appears that the confidence set $\{g_0 \geq 4\}$ provides sensible inference regarding the underlying number of Gaussian mixture components. It would appear that the AIC and BIC procedures both overestimate

the underling mixture order. However, it could also be true that the subpopulations corresponding to each of the cell lines cannot be adequately modeled via Gaussian mixture components.

5 Extensions

5.1 A consistent sequential testing procedure

Important criteria regarding the validity of an order selection method are the large sample properties of conservativeness and consistency. These properties are defined by Leeb and Potscher (2009), in the context of this work, as

$$\lim_{n_1,n_2\to\infty} \Pr_{f_0}\left(g_0 \geq \hat{g}_n\right) = 1$$

and

$$\lim_{n_1,n_2\to\infty} \Pr_{f_0}\left(g_0 = \hat{g}_n\right) = 1,$$

for all $f_0 \in \mathcal{M}$, respectively (see also Dickhaus, 2014, Sec. 7.1).

By Theorem 2, we have the fact that (3) holds for all n, and thus the STP, as stated in Sect. 1, cannot be conservative, nor consistent. However, if we replace α by a sequence $\left(\alpha_n\right)_{n=1}^\infty$, where $\alpha_n \to 0$ as $n_1, n_2 \to \infty$, then we can conclude that the modified procedure is conservative by taking the limits on both sides of inequality (3).

We now specialize our focus, again, to the parametric setting. To construct a procedure that is consistent requires further modification to the STP. Namely, we require additionally that the individual tests of H_g are consistent (i.e., that Theorem 4 holds for the sequence $\left(\alpha_n\right)_{n=1}^\infty$, replacing α in each test). Thus, to make (14) hold with probability approaching one, we require that the third term on the left-hand side converges to zero. We observe that the sequence $\left(\alpha_n\right)_{n=1}^\infty$ must simultaneously satisfy the conditions that $\alpha_n \to 0$ and $n_1^{-1} \log \alpha_n \to 0$, as $n_1, n_2 \to \infty$. For instance, we may choose to set $\alpha_n = n_1^{-\kappa}$, with $\kappa > 0$. We thus have the following result regarding the STP when applied using the sequence of p-values $\left(P_g^1\left(\mathbf{X}_n\right)\right)_{g=1}^\infty$.

Corollary 1 *Assume (A1)–(A4) from Theorem 4, and that $g_0 < \infty$. If $\alpha_n \to 0$ and $n_1^{-1} \log \alpha_n \to 0$, as $n_1, n_2 \to \infty$, then the STP for testing the sequence $\left(H_g\right)_{g=1}^\infty$ is consistent, when applied using the rules $\left(R_g\left(\mathbf{X}_n\right)\right)_{g=1}^\infty$, where $R_g\left(\mathbf{X}_n\right) = \mathbf{1}\left(P_g^1\left(\mathbf{X}_n\right) < \alpha_n\right)$.*

Proof The proof of this result appears in the Appendix.   □

We note that the modified STP resembles the time series order selection procedure of Potscher (1983). In fact, the conditions placed on the sequence $\left(\alpha_n\right)_{n=1}^\infty$ are

the same as those imposed in Potscher (1983, Thm. 5.7). Furthermore, we note that the conditions placed on $(\alpha_n)_{n=1}^{\infty}$ closely resemble the conditions that are required for the consistent application of information criteria methods; see Keribin (2000) and Baudry (2015). We can observe this resemblance by considering expression (8) and taking $\text{BIC}_{g+l} - \text{BIC}_g$, for any $g, l \in \mathbb{N}$. For the BIC procedure to be consistent, this expression must be negative, for large n, which requires that the difference in penalty $n^{-1} \log n (\dim_{g+1} - \dim_g)$ goes to zero. In the STP, if we set $n_1 = n/2$ (or as any fraction of n) and $\alpha_n = 1/n$, then we have the similar requirement (of the same rate in n) that $(2n)^{-1} \log n$ must go to zero.

5.2 Asymptotic tests

Throughout the manuscript, we have assumed that the p-values from which tests are constructed satisfy (1) for all n. This assumption is compatible with our application of the STP using the local tests proposed in Sect. 3. We note that the STP still provides guarantees for p-values that only satisfy (1) asymptotically, in the sense that

$$\limsup_{n \to \infty} \text{Pr}_f \left(P_g(\mathbf{X}_n) \leq \alpha \right) \leq \alpha \tag{9}$$

for all $f \in \mathcal{M}_g$. In such a case, we have the limiting version of the confidence statement (3):

$$\liminf_{n \to \infty} \text{Pr}_{f_0} \left(g_0 \geq \hat{g}_n \right) \geq 1 - \alpha. \tag{10}$$

To obtain (10), suppose that $f_0 \in \mathcal{M}_{g_0}$, for some finite $g_0 \in \mathbb{N}$. In the notation of Sect. 2, we can write $\mathbb{G}_0(f_0) = \mathbb{N} \backslash [g_0 - 1]$, and hence

$$\text{Pr}_{f_0} \left(f_0 \in \mathcal{M}_{\hat{g}_n - 1} \right) = \text{Pr}_{f_0} \left(\bigcup_{g \in \mathbb{N} \backslash [g_0 - 1]} \{ \bar{R}_g(\mathbf{X}_n) = 1 \} \right) = \text{Pr}_{f_0} \left(\bar{R}_{g_0}(\mathbf{X}_n) = 1 \right)$$

$$= \text{Pr}_{f_0} \left(\bigcap_{g \in [g_0]} \{ P_g(\mathbf{X}_n) \leq \alpha \} \right) \leq \text{Pr}_{f_0} \left(P_{g_0}(\mathbf{X}_n) \leq \alpha \right), \tag{11}$$

Then, since (11) holds for all n, we can apply Rudin (1976, Thm. 3.19) to obtain

$$\limsup_{n \to \infty} \text{Pr}_{f_0} \left(f_0 \in \mathcal{M}_{\hat{g}_n - 1} \right) \leq \limsup_{n \to \infty} \text{Pr}_{f_0} \left(P_{g_0}(\mathbf{X}_n) \leq \alpha \right) \leq \alpha,$$

as required. Using (10), we can justify the use of the STP with asymptotically valid tests, such as the procedure of Li and Chen (2010).

5.3 Aggregated tests

Under the null hypothesis that $f_0 \in \mathcal{M}_g$, both the Split and Swapped statistics, $V_g^k(\mathbf{X}_n)$ and $\bar{V}_g(\mathbf{X}_n)$, are examples of e-values (which we shall write generically as E_g), as defined in Vovk and Wang (2021) (note that these values also appear as s-values in Grunwald et al., 2020, and as betting scores in Shafer, 2021), based on the defining feature that

$$\sup_{f \in \mathcal{M}_g} \mathrm{E}_f(E_g) \leq 1. \tag{12}$$

By Markov's inequality, (12) implies

$$\sup_{f \in \mathcal{M}_g} \mathrm{Pr}_f(E_g \geq 1/\alpha) \leq \alpha,$$

for any $\alpha \in (0, 1)$, from which we can derive the p-value $\min\{P_g = 1/E_g, 1\}$, which satisfies (1).

As discussed in Wasserman et al. (2020), any set of possibly dependent e-values $E_g^1, \ldots, E_g^m$ ($m \in \mathbb{N}$) can be combined by simple averaging to generate a new e-value $\bar{E}_g = m^{-1} \sum_{j=1}^m E_g^j$, which we shall call the aggregated e-value. As such, one may consider generating m different e-values based on either the Split or Swapped statistics, using different partitions of the data into subsequences $\mathbf{X}_n^1$ and $\mathbf{X}_n^2$. For any fixed n_1 and n_2, there are only a finite number of such partitions and thus one may imagine an aggregated e-value that averages over all such partitions. This hypothetical process was referred to as derandomization in Wasserman et al. (2020), since the resulting p-value is no longer dependent on any particular random partitioning of $\mathbf{X}_n$.

We further note that one can also aggregate the results from multiple instances of the Split and Swapped statistics via methods for aggregating over p-values. These methods are discussed at length in the works of Vovk and Wang (2020) who provide a detailed assessment of methods for combining arbitrarily dependent p-values, via generalized averaging operations.

6 Conclusions

In this work, we proved that the closed testing principle could be used to construct a sequence of null hypothesis tests that generates a confidence statement regarding the true number of mixture components of a finite mixture model. Further, we derive tests for each of the null hypotheses in the STP, using the universal inference framework of Wasserman et al. (2020), and proved that in the parametric case, under regularity conditions, such tests are consistent against fixed alternative hypotheses.

The performance of the STP for order selection of normal mixture models was considered via a comprehensive simulation study. We observe from the study that

the constructed confidence statements were conservative, as predicted by the theory, and we were also able to make recommendations regarding the different variants of the tests, for practical application. We also determined that the AIC and BIC point estimators provide accurate complements to the intervals provided by the STP. Example applications of the STP are further described to demonstrate the utility of our methods in practice. We recommend that our STP interval estimators be reported alongside an AIC or BIC point estimator to provide both an accurate and precise inference regarding the true order.

Extensions of the STP were also discussed, including the possibility of aggregating over multiple tests, and performing the STP with asymptotic tests. Of particular interest is a proof that the testing procedure could be modified to generate an order selection procedure that consistently determines the true number of mixture components, in the asymptotic sense. Our proof shows that such a procedure was essentially equivalent to other asymptotic model selection methods such as the Bayesian information criterion and its variants.

We note that our general order selection confidence result of Theorem 2 applies not only to finite mixture models, but also to any nested sequences of models. For example, we may consider the same STP to generate confidence statements regarding the number of factors in a factor analysis model or the degree of a polynomial fit. We leave the application of the STP to such problems for future work, along with the applications of our discussed variants on the testing procedures.

Appendix

Proof of Theorem 2

Firstly, since $\mathcal{M}_g \subset \mathcal{M}_{g+1}$, we have the fact that for any $g \in \mathbb{I} \subset \mathbb{N}$, $\bigcap_{g \in \mathbb{I}} \mathcal{M}_g = \mathcal{M}_{\min_{g \in \mathbb{I}} g}$ and thus the sequence $\left(H_g \right)_{g=1}^{\infty}$ is $\cap$-closed. Next, the sequential procedure rejects H_g if and only if $R_j(\mathbf{X}_n) = 1$ for each $j \in [g]$, or more compactly, H_g is rejected if and only if $\bar{R}_g(\mathbf{X}_n) = \min_{j \in [g]} R_j(\mathbf{X}_n) = 1$. Because $\mathcal{M}_g \subset \mathcal{M}_{g+1}$, we also have the fact that $\{j : \mathcal{M}_j \subseteq \mathcal{M}_g\} = [g]$, and thus $\left(\bar{R}_g(\mathbf{X}_n) \right)_{g=1}^{\infty}$ is exactly the sequence of closed tests for $\left(H_g \right)_{g=1}^{\infty}$, of form (4).

By Theorem 1, for each $f \in \mathcal{M}$, we have the inequality

$$\mathrm{Pr}_f \left(\bigcup_{g \in \mathbb{G}_0(f)} \left\{ \bar{R}_g(\mathbf{X}_n) = 1 \right\} \right) \le \alpha, \tag{13}$$

where the event $\left\{ \bigcup_{g \in \mathbb{G}_0(f)} \left\{ \bar{R}_g(\mathbf{X}_n) = 1 \right\} \right\}$ can be written as $\left\{ f \in \bigcup_{j \in [\hat{g}_n - 1]} \mathcal{M}_j \right\}$, since the sequential procedure first fails to reject hypothesis $H_{\hat{g}_n}$. Again, since $\mathcal{M}_g \subset \mathcal{M}_{g+1}$, $\bigcup_{j \in [\hat{g}_n - 1]} \mathcal{M}_j = \mathcal{M}_{\hat{g}_n - 1}$ and thus (13) can be written in form (2). This completes the proof.

Proof of Theorem 4

Write the event $\left\{ P_g^1\left(\mathbf{X}_n\right) < \alpha \right\}$ as

$$
\frac{\prod_{i=1}^{n_1} f\left(X_i^1; \tilde{\boldsymbol{\vartheta}}_n^{(g)}\right)}{\prod_{i=1}^{n_1} f\left(X_i^1; \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)}\right)} < \alpha,
$$

or equivalently

$$
\frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \tilde{\boldsymbol{\vartheta}}_n^{(g)}\right) - \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)}\right) - \frac{\log \alpha}{n_1} < 0. \tag{14}
$$

Thus, it suffices to show that the left-hand side converges in probability to a constant that is bounded above by zero.

By (A2) and (A3), we have the facts that (i): $\tilde{\boldsymbol{\vartheta}}_n^{(g)} \to \boldsymbol{\vartheta}_0^{(g)}$, in probability as $n_1 \to \infty$, where

$$
\boldsymbol{\vartheta}_0^{(g+l_g)} \in \left\{ \tilde{\boldsymbol{\vartheta}}^{(g+l_g)} \in \mathbb{T}_{g+l_g} : \mathrm{E}_{f_0} \log f\left(X; \tilde{\boldsymbol{\vartheta}}^{(g+l_g)}\right) \right.
$$

$$
\left. = \max_{\boldsymbol{\vartheta}^{(g+l_g)} \in \mathbb{T}_{g+l_g}} \mathrm{E}_{f_0} \log f\left(X; \boldsymbol{\vartheta}^{(g+l_g)}\right) \right\},
$$

and (ii):

$$
\frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \tilde{\boldsymbol{\vartheta}}_n^{(g)}\right) \to \mathrm{E}_{f_0} \log f\left(X; \boldsymbol{\vartheta}_0^{(g)}\right),
$$

in probability, as $n_1 \to \infty$, by application of Atienza et al. (2007, Lem. 1), which states that

$$
\left| \log f\left(X; \boldsymbol{\vartheta}^{(g)}\right) \right| \leq \sum_{z=1}^{g} \left| \log f\left(X; \boldsymbol{\theta}_z\right) \right|, \tag{15}
$$

and using the classic uniform weak law of large numbers of Jennrich (1969, Thm. 2). That is, (A2) permits the use of Potscher and Prucha (1997, Lem. 4.2) to prove result (i), by verifying the conditions for the uniform law, which can be done via the bound (15) and the existence of moments from (A3). Next, using (i), we show (ii) by considering the decomposition:

$$\left| \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \tilde{\boldsymbol{\vartheta}}_n^{(g)} \right) - \mathrm{E}_{f_0} \log f\left(X; \boldsymbol{\vartheta}_0^{(g)} \right) \right|$$

$$\leq \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \tilde{\boldsymbol{\vartheta}}_n^{(g)} \right) - \mathrm{E}_{f_0} \log f\left(X; \tilde{\boldsymbol{\vartheta}}_n^{(g)} \right) \right|$$

$$+ \left| \mathrm{E}_{f_0} \log f\left(X; \tilde{\boldsymbol{\vartheta}}_n^{(g)} \right) - \mathrm{E}_{f_0} \log f\left(X; \boldsymbol{\vartheta}_0^{(g)} \right) \right|,$$

where the first term on the right-hand side converges to zero in probability, by the uniform law, and using (A3), the second term is bounded from above by

$$\mathrm{E}_{f_0} \left| \log f\left(X; \tilde{\boldsymbol{\vartheta}}_n^{(g)} \right) - \log f\left(X; \boldsymbol{\vartheta}_0^{(g)} \right) \right| \leq 2g \mathrm{E}_{f_0} M_1(X) < \infty. \tag{16}$$

The continuity from (A2) and bound (16) then implies that the second term is continuous with respect to the argument $\tilde{\boldsymbol{\vartheta}}_n^{(g)}$ (cf. Makarov & Podkorytov, 2013, Thm. 7.1.3). The continuous mapping theorem then implies that the second term converges in probability to zero, as $n_1 \to \infty$.

Next, we write

$$\left| \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)} \right) - \mathrm{E}_{f_0} \log f\left(X; \boldsymbol{\vartheta}_0^{(g+l_g)} \right) \right|$$

$$\leq \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)} \right) - \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \boldsymbol{\vartheta}_0^{(g+l_g)} \right) \right|$$

$$+ \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \log f\left(X_i^1; \boldsymbol{\vartheta}_0^{(g+l_g)} \right) - \mathrm{E}_{f_0} \log f\left(X; \boldsymbol{\vartheta}_0^{(g+l_g)} \right) \right|.$$

Using (A3), the first term on the right-hand side can be bounded from above by

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \left| \log f\left(X_i^1; \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)} \right) - \log f\left(X_i^1; \boldsymbol{\vartheta}_0^{(g+l_g)} \right) \right|$$

$$\leq \frac{1}{n_1} \sum_{i=1}^{n_1} M_2\left(X_i^1 \right) \left\| \hat{\boldsymbol{\vartheta}}_n^{(g+l_g)} - \boldsymbol{\vartheta}_0^{(g+l_g)} \right\|.$$

Thus, the first term converges to zero in probability, as $n_1 \to \infty$, by the law of large numbers (since $\mathrm{E}_{f_0} M_2(X) < \infty$), and since $\hat{\boldsymbol{\vartheta}}_n^{(g+l_g)} \to \boldsymbol{\vartheta}_0^{(g+l_g)}$, in probability, as $n_2 \to \infty$. The second term converges to zero, in probability, as $n_1 \to \infty$, by the law of large numbers, since

$$\mathrm{E}_{f_0} \left| \log f\left(X; \boldsymbol{\vartheta}_0^{(g+l_g)} \right) \right| \leq 2(g + l_g) \mathrm{E} M_1(X) < \infty,$$

by application of bound (15).

We have thus established that the left-hand side of (14) converges in probability to

$$
\begin{aligned}
E_{f_0} \log f\left(X;\boldsymbol{\vartheta}_0^{(g)}\right) - E_{f_0} \log f\left(X;\boldsymbol{\vartheta}_0^{(g+l_g)}\right) &= D\left(f_0,f\left(\cdot;\boldsymbol{\vartheta}_0^{(g+l_g)}\right)\right) \\
&\quad - D\left(f_0,f\left(\cdot;\boldsymbol{\vartheta}_0^{(g)}\right)\right),
\end{aligned}
\tag{17}
$$

as $n_1, n_2 \to \infty$. Suppose, for contradiction, that (17) is equal to zero. Then, for all $f\left(\boldsymbol{x};\boldsymbol{\vartheta}^{(g+l_g)}\right) \in \mathcal{M}_{g+l_g}$,

$$
D\left(f_0,f\left(\cdot;\boldsymbol{\vartheta}^{(g+l_g)}\right)\right) - D\left(f_0,f\left(\cdot;\boldsymbol{\vartheta}_0^{(g)}\right)\right) \geq 0.
$$

In particular, for some $\boldsymbol{\theta} \in \mathbb{T}$ and $\varpi \in (0,1)$, we have

$$
\int_{\mathbb{X}} f_0(\boldsymbol{x}) \log \left\{ \frac{(1-\varpi)f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right) + \varpi f(\boldsymbol{x};\boldsymbol{\theta})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} \right\} d\boldsymbol{x} \leq 0.
$$

By Fatou's Lemma,

$$
\begin{aligned}
0 &\geq \int_{\mathbb{X}} f_0(\boldsymbol{x}) \liminf_{\varpi \to 0} \frac{1}{\varpi} \log \left\{ \frac{(1-\varpi)f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right) + \varpi f(\boldsymbol{x};\boldsymbol{\theta})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} \right\} d\boldsymbol{x} \\
&= \int_{\mathbb{X}} f_0(\boldsymbol{x}) \left\{ \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} - 1 \right\} d\boldsymbol{x},
\end{aligned}
$$

which implies that

$$
\int_{\mathbb{X}} f_0(\boldsymbol{x}) \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} d\boldsymbol{x} \leq 1.
\tag{18}
$$

Since $f_0 \in \mathcal{M}\backslash\mathcal{M}_g$, we have $f_0 = f\left(\cdot;\boldsymbol{\vartheta}_0^{(g_0)}\right) \in \mathcal{M}_{g_0}$, where $g_0 > g$ and $\boldsymbol{\vartheta}^{g_0}$ contains the pairs $\left(\pi_{0,z},\boldsymbol{\theta}_{0,z}\right)_{z=1}^{g_0}$. By taking the expectation of both sides of (18) with respect to the probability measure on $\boldsymbol{\theta}$, defined by

$$
\Pr\left(\boldsymbol{\theta} = \boldsymbol{\theta}'\right) = \sum_{z=1}^{g_0} \pi_{0,z} \mathbf{1}\left(\boldsymbol{\theta}' = \boldsymbol{\theta}_{0,z}\right),
$$

we have

$$\sum_{z=1}^{g_0} \int_{\mathbb{X}} f_0(\boldsymbol{x}) \frac{\pi_{0,z} f(\boldsymbol{x};\theta_{0,z})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} \mathrm{d}\boldsymbol{x} = \int_{\mathbb{X}} \frac{f_0^2(\boldsymbol{x})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} \mathrm{d}\boldsymbol{x} \leq 1.$$

Finally, by the fact that $\log a \leq a - 1$, for all $a > 0$, we have

$$\mathrm{D}\left(f_0, f\left(\cdot;\boldsymbol{\vartheta}_0^{(g)}\right)\right) = \int_{\mathbb{X}} f_0(\boldsymbol{x}) \log \left\{ \frac{f_0(\boldsymbol{x})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} \right\} \mathrm{d}\boldsymbol{x}$$

$$\leq \int_{\mathbb{X}} f_0(\boldsymbol{x}) \left\{ \frac{f_0(\boldsymbol{x})}{f\left(\boldsymbol{x};\boldsymbol{\vartheta}_0^{(g)}\right)} - 1 \right\} \mathrm{d}\boldsymbol{x} \leq 0,$$

which implies that $f\left(\cdot;\boldsymbol{\vartheta}_0^{(g)}\right) = f_0$, by (A1) and the definition of the Kullback–Leibler divergence (cf. Leroux, 1992, Lem. 1). Thus, we have the contradiction that $f_0 \in \mathcal{M}_g$, and hence

$$\mathrm{E}_{f_0} \log f\left(\boldsymbol{X};\boldsymbol{\vartheta}_0^{(g)}\right) - \mathrm{E}_{f_0} \log f\left(\boldsymbol{X};\boldsymbol{\vartheta}_0^{(g+l_g)}\right) < 0,$$

as required.

Proof of Corollary 1

It suffices to show that for each $\epsilon > 0$, there exists a $N(\epsilon) \in \mathbb{N}$, such that for all $n_1, n_2 \geq N(\epsilon)$, we have for any $f_0 \in \mathcal{M}$:

$$\mathrm{Pr}_{f_0}\left(g_0 = \hat{g}_n\right) \geq 1 - \epsilon.$$

Firstly, using the form of the sequential testing procedure, we can write

$$\mathrm{Pr}_{f_0}\left(g_0 = \hat{g}_n\right) = \mathrm{Pr}_{f_0}\left(\left(\bigcap_{g=1}^{g_0-1} \left\{ R_g^n = 1 \right\} \right) \cap \left\{ R_{g_0}^n = 0 \right\} \right)$$

$$= 1 - \mathrm{Pr}_{f_0}\left(\left(\bigcup_{g=1}^{g_0-1} \left\{ R_g^n = 0 \right\} \right) \cup \left\{ R_{g_0}^n = 1 \right\} \right)$$

$$\geq 1 - \sum_{g=1}^{g_0-1} \mathrm{Pr}_{f_0}\left(R_g^n = 0 \right) - \mathrm{Pr}_{f_0}\left(R_{g_0}^n = 1 \right).$$

By $n_1^{-1} \log \alpha_n \to 0$ and Theorem 4, and by $\alpha_n \to 0$ and Theorem 3, we have for any $\delta > 0$, there exist $N_{g_0}(\delta) \in \mathbb{N}$, such that

$$\Pr_{f_0}\left(R_g^n = 0\right) \leq \delta,$$

and

$$\Pr_{f_0}\left(R_{g_0}^n = 1\right) \leq \delta,$$

for all $n_1, n_2 \geq N_{g_0}(\delta)$ and $g \in [g_0 - 1]$. Thus, setting $\epsilon = g_0\delta$ and $N(\epsilon) = \max_{g \in [g_0]} N_g(\delta)$, we have

$$
\begin{aligned}
\Pr_{f_0}\left(g_0 = \hat{g}_n\right) &\geq 1 - (g_0 - 1)\delta - \delta \\
&= 1 - g_0\delta = 1 - \epsilon,
\end{aligned}
$$

as required.

References

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica, 69,* 683–734.

Atienza, N., Garcia-Heras, J., Munoz-Pichardo, J. M., & Villa, R. (2007). On the consistency of MLE in finite mixture models of exponential families. *Journal of Statistical Planning and Inference, 137,* 496–505.

Baudry, J.-P. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics, 9,* 1041–1077.

Bauer, P., & Kieser, M. (1996). A unified approach for confidence intervals and testing of equivalence and difference. *Biometrika, 83,* 934–937.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering wit the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 719–725.

Chen, J. (1998). Penalized likelihood-ratio test for finite mixture models with multinomial observations. *Canadian Journal of Statistics, 26,* 583–599.

Chen, J., & Khalili, A. (2009). Order selection in finite models with a nonsmooth penalty. *Journal of the American Statistical Association, 104,* 187–196.

Chen, J., Li, P., & Fu, Y. (2012). Inference on the order of a normal mixture. *Journal of the American Statistical Association, 107,* 1096–1105.

Cutler, A. & Windham, M. P. (1994) Information-based validity functionals for mixture analysis. In: *Proceedings of the First US/Japan conference on the frontiers of statistical modeling in informational approach*, Amsterdam. Kluwer.

DasGupta, A. (2008). *Asymptotic Theory of Statistics And Probability*. Springer.

Dickhaus, T. (2014). *Simultaneous statistical inference: With applications in the life sciences*. Springer.

Donoho, D. L. (1988). One-sided inference about functionals of a density. *Annals of Statistics, 16,* 1390–1420.

Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis, 71,* 1054–1063.

Feng, Z. D., & McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society Series B, 58,* 609–617.

Ferrari, D., & Yang, Y. (2015). Confidence sets for model selection by F-testing. *Statistica Sinica, 25,* 1637–1658.

Finner, H., & Giani, G. (1996). Duality between multiple testing and selecting. *Journal of Statistical Planning and Inference, 54,* 201–227.

Goeman, J. J., & Solari, A. (2010). The sequential rejection principle of familywise error control. *Annals of Statistics, 38,* 3782–3810.

Gorman, K. B., Williams, T. D., & Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins. *PLoS One, 9,* e90081.

Grunwald, P., de Heide, R, & Koolen, W. M. (2020) Safe testing. In: *IEEE Information Theory and Applications Workshop (ITA)*.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica, 79,* 453–497.

Hardle, W. (1991). *Smoothing techniques with implementations in S*. Springer.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics, 40,* 633–643.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya A, 62,* 49–65.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., et al. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods, 16,* 1289–1296.

Leeb, H., & Potscher, B. M. (2009). Model selection. In T. G. Andersen, R. A. Davis, J.-P. Kreiss, & T. Mikosch (Eds.), *Handbook of financial time series* (pp. 889–925). Springer.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *Annals of Statistics, 20,* 1350–1360.

Li, P., & Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association, 105,* 1084–1092.

Makarov, B., & Podkorytov, A. (2013). *Real analysis: Measures, integrals and applications*. New York: Springer.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63,* 655–660.

Maurer, W., Hothorn, L. A., & Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: A priori ordered hypotheses. In J. Vollman (Ed.), *Biometrie in der Chemish-in-Pharmazeutischen Industrie*. Fischer-Verlag.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society Series C, 36,* 318–324.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. Wiley.

McLachlan, G. J., & Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *WIREs Data Mining and Knowledge Discovery, 4,* 341–355.

Melnykov, V., Chen, W.-C., & Maitra, R. (2012). MixSim: An R package for simulating data to study performance of clustering algorithms. *Journal of Statistical Software, 51,* 1–25.

Polymenis, A., & Titterington, D. M. (1998). On the determination of the number of components in a mixture. *Statistics and Probability Letters, 38,* 295–298.

Potscher, B. M. (1983). Order estimation in ARMA-models by Lagrangian multiplier tests. *B M Potscher, 11,* 872–885.

Potscher, B. M., & Prucha, I. R. (1997). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer.

Quinn, B. G., McLachlan, G. J., & Hjort, N. L. (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society B, 49,* 311–314.

R Core Team. (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Ritter, Gunter. (2014). *Robust cluster analysis and variable selection*. CRC Press.

Rudin, W. (1976). *Principles of mathematical analysis*. McGraw Hill.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society B, 184,* 407–431.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.

Vovk, V., & Wang, R. (2020). Combining p values via averaging. *Biometrika, 107,* 791–808.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination, and application. *Annals of Statistics, 49,* 1736–1754.

Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences, 117,* 16880–16890.

Windham, M. P., & Cutler, A. (1992). Information ratios for validating mixture analyses. *Journal of the American Statistical Association, 87,* 1188–1197.

Xu, C., & Chen, J. (2015). A thresholding algorithm for order selection in finite mixture models. *Communications in Statistics-Simulation and Computation, 44,* 433–453.

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

Yin, C., Lin, X. S., Huang, R., & Yuan, H. (2019). On the consistency of penalized MLEs for Erlang mixtures. *Statistics and Probability Letters, 145,* 12–20.

Zheng, C., Ferrari, D., & Yang, Y. (2019). Model selection confidence sets by likelihood ratio testing. *Statistica Sinica, 29,* 827–851.