Received: 10 August 2023 | Revised: 9 May 2025 | Accepted: 3 July 2025

DOI: 10.1111/sjos.70017

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

False selection rate control in mixture models

Ariane Marandon[1] | **Tabea Rebafka**[2] | **Etienne Roquain**[2] | **Nataliya Sokolovska**[3]

[1]The Alan Turing Institute, London, UK

[2]LPSM, Sorbonne Université, Université Paris Cité & CNRS, Paris, France

[3]LCQB, Sorbonne Université, CNRS, Paris, France

Correspondence
Ariane Marandon, The Alan Turing Institute, London, UK.
Email: amarandon-carlhian@turing.ac.uk

Funding information
DIM MATH INNOV; Turing-Roche partnership

Abstract

The clustering task consists in partitioning elements of a sample into homogeneous groups. Most datasets contain individuals that are ambiguous and intrinsically difficult to attribute to one or another cluster. However, in practical applications, misclassifying individuals is potentially disastrous and should be avoided. To keep the misclassification rate small, one can decide to classify only a *part* of the sample. In the supervised setting, this approach is well known and referred to as classification with an abstention option. In this paper, the approach is revisited in an unsupervised mixture-model framework. The purpose is to develop a method that guarantees the false selection rate (FSR) does not exceed a predefined level α. We propose a plug-in procedure and provide a theoretical analysis, quantifying the deviation of the FSR from the target α with explicit remainder terms. Bootstrap versions of the procedure are shown to improve the performance in numerical experiments.

KEYWORDS

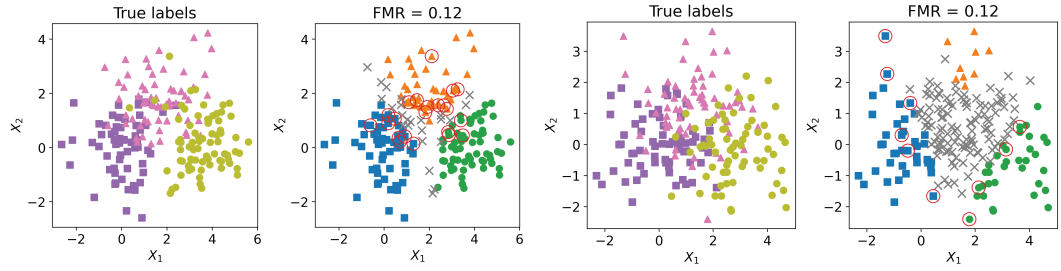
abstention option, bootstrap, clustering, false discovery rate, mixture models

1 | INTRODUCTION

1.1 | Background

Clustering is the standard statistical task that aims to group together items with similar features. However, it is common that datasets include ambiguous items that are inherently difficult to classify, which makes the clustering result potentially unreliable. To illustrate this point, consider a Gaussian mixture model with overlapping mixture components, see Figure 1. In this setting, it is impossible to construct a clustering method that consistently assigns the correct cluster label to data points that fall in the overlap of those clusters. That is, when the overlap is large, the misclassification rate (the proportion of elements which are erroneously clustered up to label switching, see (1) below for a formal definition) of a standard clustering method is inevitably elevated. This issue can be critical in applications where falsely clustered items induce high costs for the user or make the interpretation of the clusters difficult. In particular in an early-stage analysis of a new dataset, it might be much easier to interpret the clusters when they do not contain spurious observations.

To be more precise about the motivation of our unsupervised approach, let us consider the common situation where a medical doctor considers a clustering of a group of patients based on phenotyping, i.e. clusters are homogeneous groups of individuals in terms of their phenotypic profiles. The doctor may analyze each cluster to decide on a treatment for all patients in the same cluster. For instance, in Eckardt et al. (2023), a clustering method is applied to stratify leukemia patients into groups according to risk; these groups differ in genetics and clinical values, which leads to differences in patient survival and treatment. Another example concerns patients suffering from injuries who are to be clustered according to their physical and psychological profiles to identify optimal recovery for each group (Stoitsas et al., 2022). Our approach is motivated by the preliminary unsupervised stage, for which it is important to have patient clusters that we can clearly identify. Hence, patients with an ambiguous profile should not be considered at this stage and put aside, as taking a decision for these patients might require additional information. This corresponds to following an approach with abstention decision. In a supervised setting, classification with a reject (or abstention) option is a long-standing statistical paradigm, that can be traced back to Chow (1970), with more recent works including Herbei and Wegkamp (2006), Bartlett and



(a) Separated clusters (b) Ambiguous clusters

FIGURE 1 Data from Gaussian mixtures with three components ($n = 200$), in a fairly separated case (panel (a)) and an ambiguous case (panel (b)). In each panel, the left part displays the true clustering (colors/point characters correspond to the different true clusters), while the right part illustrates the new procedure (plug-in procedure at level $\alpha = 10\%$), that does not cluster all items (colors/point characters correspond to the different estimated clusters). The points not classified are depicted by grey crosses. Red circles indicate erroneous labels.

Wegkamp (2008), Wegkamp and Yuan (2011), Cortes et al. (Cortes et al., 2016), Ni et al. (2019), Cao et al. (2022), among others. In this line of research, rejection is accounted for by adding a term to the risk that penalizes any rejection (i.e., nonclassification).

Recently, still in the supervised setting, Geifman and El-Yaniv (2017) and Angelopoulos et al. (2025) have considered the problem of having a prescribed control of the classification error conditional on the item being classified (not rejected). In these works, the proposed method consists in thresholding the class probabilities estimated by a pretrained classifier, in a data-driven manner. The computation of the risk-controlling threshold is based on concentration inequalities that bound the true risk with respect to the empirical one. This criterion is also studied in Denis and Hebiri (2020) in a semisupervised setting. However, the control problem is different, because it rather aims at a type-II error control concerning the probability of classifying an item. Finally, a related line of research is conformal prediction (Angelopoulos & Bates, 2021; Lei, 2014; Romano et al., 2020; Sadinle et al., 2019; Vovk et al., 2005). This generic technique builds prediction sets that contain the true label with probability above a fixed confidence level. In that framework, one trades single-valued classification for set-valued predictions to have a prespecified prediction coverage, and finite-sample control is achieved using hold-out labeled data (calibration sample).

The supervised approaches can only be applied when classes are known and already have a specific interpretation, while our unsupervised method is an exploratory approach, where the goal is to identify unknown clusters of individuals that we wish to interpret for a better understanding of the dataset.

1.2 | Aim and approach

The goal of the present work is to propose a labeling guarantee on the classified items in the more challenging unsupervised setting, where no labeled training set is available (Eckardt et al., 2023; Stoitsas et al., 2022) and data are assumed to be generated from a finite mixture model (Grün, 2019; Ng, 2019). This is achieved by the possibility to refuse to cluster ambiguous individuals and by using the false selection rate (FSR), which is defined as the average proportion of misclassifications among the classified objects. Our procedures are devised to keep the FSR below some nominal level α, while classifying a maximum number of items.

It is important to understand the role of the nominal level α in our approach. It is chosen by the user and depends on their acceptance or tolerance for misclassified objects. Since the FSR is the misclassification risk that is allowed on the classified objects, the final interpretation of an FSR control at level α is clear: if, for instance, α is set to 5% and 100 items are finally chosen to be classified by the method, then the number of misclassified items is expected to be at most 5. This high interpretability is similar to the one of the false discovery rate (FDR) in multiple testing, which has known a great success in applications since its introduction by Benjamini and Hochberg (1995). This is a clear advantage of our approach for practical use compared to the methods with a rejection option that are based on a penalized risk.

In our framework, a procedure is composed of two components:

- a clustering method inferring the labels;
- a selection rule deciding which labels to keep.

Importantly, the selection rule is only applied *after* a clustering method is fitted on the (entire) sample. In other words, the procedure consists of two subsequent steps: a clustering step, that

assigns a cluster label to every item and a selection step, that chooses which items to classify—in which case, the label from the previous clustering step is assigned. For the items that are not selected, we discard the cluster label, that is, we effectively abstain to make a classification decision for those items. In particular, we emphasize that the clustering method is not fitted again after selection (which would in general lead to bias).

The quality of the selection heavily relies on the appropriate quantification of the uncertainty of the cluster labels. For this, our approach is model-based and can be viewed as a method that thresholds the posterior probabilities of the cluster labels with a data-driven choice of the threshold. The performance of the method will depend on the quality of the estimates of these posterior probabilities in the mixture model.

The adaptive character of our method is illustrated in Figure 1: when the clusters are well separated (panel (a)), the new procedure only discards few items and provides a clustering close to the correct one. However, when clusters are overlapping (panel (b)), to avoid a high misclassification error, the procedure discards most of the items and only provides few labels, for which the uncertainty is low. In both cases, the proportion of misclassified items among the selected ones is small and in particular close to the target level α (here 10%). Hence, by adapting the amount of labeled or discarded items, our method always delivers a reliable clustering result, inspite of the varying intrinsic difficulty of the clustering task.

1.3 | Presentation of the results

Let us now describe in more details the main contributions of the paper.

- We introduce three new data-driven procedures that perform simultaneously selection and clustering: the plug-in procedure (illustrated in Figure 1) and two bootstrap procedures (parametric and nonparametric), see Section 3.2.
- We provide a theoretical analysis of the plug-in procedure, quantifying the FSR deviation with respect to the target level α with explicit remainder terms, which become small when the sample size grows. In addition, this procedure is shown to satisfy the following optimality property: any other procedure that provides an FSR control necessarily classifies as many or less items than the plug-in procedure, up to a small remainder term (Theorem 2).
- Numerical experiments establish that the bootstrap procedures improve the plug-in procedure, and thus are more reliable for practical use, where the sample size may be moderate, see Section 5.1. In particular, the FSR control is shown to be valid in scenarios with various levels of difficulty.
- Our analysis also shows that a fixed threshold procedure that only clusters items with a maximum posterior probability larger than $1 - \alpha$ typically clusters less items than our procedures, see Section 5.1. To this extent, our procedures can be seen as refined algorithms that classify more individuals while maintaining the FSR control.
- The practical impact of our approach is demonstrated on a real dataset, see Section 5.2.

1.4 | Relation to previous work

1.4.1 | Misclassification criterion

In the unsupervised setting, defining a notion of error for a clustering procedure is not straightforward and there exists multiple criteria (Fahad et al., 2014). We adopt the proportion of

misclassifications with respect to predefined class labels (computed up to a permutation of the labels). This criterion takes a pointwise point of view where each data point is assigned a particular label independently of other data points, which can be motivated from a mixture model setup. Alternatively, many popular criteria are based on pairwise comparisons, assessing whether pairs of similar data points are placed in the same cluster and dissimilar data points in different clusters. In the binary case, Bao et al. (2020) established that the misclassification error is actually a monotonic function of the pairwise classification error, highlighting the strong relationship between these criteria.

1.4.2 | Clustering guarantees in unsupervised learning

While we provide a specific FSR control guarantee on the clustering, other criteria, not linked to a rejection option, have been previously proposed in an unsupervised setting. Previous works provided essentially two types of guarantees: while early works focused on the probability of exact recovery (Abbe, 2018; Arora & Kannan, 2005; Vempala & Wang, 2004), recent contributions rather considered minimizing the misclassification risk (Chretien et al., 2019; Giraud & Verzelen, 2018; Lei & Rinaldo, 2015; Lu & Zhou, 2016). Other criteria include the probability to make a different decision than the Bayes rule (Azizyan et al., 2013), or the fact that all clusters are mostly homogeneous with high probability (Najafi et al., 2020). In these works, the error rates can be small only in regions of the parameter space that make the clusters separated enough. By contrast, FSR control can be provided even when the clusters overlap by selecting only items that are not ambiguous.

1.4.3 | Comparison to Mary-Huard et al. (2021)

The recent work of Mary-Huard et al. (2021) also proposes a control of the FSR. However, the analysis therein is solely based on the case where the model parameters are known (thus corresponding to the oracle case developed in Section 3.1 here). Compared to Mary-Huard et al. (2021), the present work provides a number of new contributions, which are all outlined in Section 1.3. Let us also emphasize that we handle the label switching problem in the FSR, which seems to be overlooked in Mary-Huard et al. (2021).

1.4.4 | Relation to the false discovery rate

The FSR is closely related to the false discovery rate (FDR) in multiple testing, defined as the average proportion of errors among the discoveries. In fact, we can roughly view the problem of designing an abstention rule as testing, for each item i, whether the clustering rule correctly classifies item i or not. With this analogy, our selection rule is based on quantities similar to the local FDR values (Efron et al., 2001), a key quantity to build optimal FDR controlling procedures in multiple testing mixture models, see, e.g., Storey (2003), Sun and Cai (2007), Cai et al. (2019), Rebafka et al. (2022). In particular, our final selection procedure shares similarities with the procedure introduced in Sun and Cai (2007), also named cumulative ℓ-value procedure (Abraham et al., 2022). In addition, our theoretical analysis is related to the work of Rebafka et al. (2022), although the nature of the algorithm developed therein is different from here: they

use the q-value procedure of Storey (2003), while our method rather relies on the cumulative ℓ-value procedure.

1.5 | Organization of the paper

The paper is organized as follows: Section 2 introduces the model and relevant notation, namely the FSR criterion, with a particular care for the label switching problem. Section 3 presents the new methods: the oracle, the plug-in and the bootstrap approaches. Our main theoretical results are provided in Section 4, after introducing appropriate assumptions. Section 5 presents numerical experiments and an application to a real dataset, while a conclusion is given in Section 6. Proofs of the results and technical details are deferred to appendices.

2 | SETTING

This section presents the notation, model, procedures, and criteria that will be used throughout the manuscript.

2.1 | Model

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an observed random sample of size n. Each X_i is an i.i.d. copy of a d-dimensional real random vector, which is assumed to follow the standard mixture model:

$$\begin{aligned} Z &\sim \mathcal{M}(\pi_1, \ldots, \pi_Q), \\ X|Z = q &\sim F_{\phi_q}, \quad 1 \leq q \leq Q, \end{aligned}$$

where $\mathcal{M}(\pi_1, \ldots, \pi_Q)$ denotes the multinomial distribution of parameter π (equivalently, $\pi_q = \mathbb{P}(Z = q)$ for each q). The model parameters are given by

- the probability distribution π on $\{1, \ldots, Q\}$ that is assumed to satisfy $\pi_q > 0$ for all q. Hence, π_q corresponds to the probability of being in class q;
- the parameter $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_Q) \in \mathcal{U}^Q$, where $\{F_u, u \in \mathcal{U}\}$ is a collection of distributions on $\mathbb{R}^d$. Every distribution F_u is assumed to have a density with respect to the Lebesgue measure on $\mathbb{R}^d$, denoted by f_u. Moreover, we assume that the ϕ_q's are all distinct (and that the $F_u, u \in \mathcal{U}$ are all distinct).

The number $Q \geq 2$ of classes is assumed to be known and fixed throughout the manuscript (see Section 6 for a discussion). Thus, the overall parameter is $\theta = (\pi, \phi)$, the parameter set is denoted by Θ, and the distribution of (Z, X) is denoted by P_θ. The distribution family $\{P_\theta, \theta \in \Theta\}$ is the considered statistical model. We also assume that Θ is an open subset of $\mathbb{R}^K$ for some $K \geq 1$. In this mixture model, the latent vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is called the true latent clustering in the sequel. In what follows, the "true" parameter that generates (Z, X) is assumed to be fixed and is denoted by $\theta^* \in \Theta$. Furthermore, we use the common notation $\mathbb{P}_{\theta^*}(A)$ or $\mathbb{P}_{\mathbf{X} \sim P_{\theta^*}}(A)$ to denote the probability of an event A involving the sample $\mathbf{X} = (X_1, \ldots, X_n)$ (and possibly also the

corresponding $\mathbf{Z} = (Z_1, \ldots, Z_n)$), when this sample is iid and generated from P_{θ^*}. We also use $\mathbb{E}_{P_{\theta^*}}$ or $\mathbb{E}_{\mathbf{X}\sim P_{\theta^*}}$ in a similar way.

2.2 | Procedure and criteria

Our approach starts with a given clustering rule that aims at recovering the true latent clustering for all observed items. In general, a clustering rule $\widehat{\mathbf{Z}} = (\hat{Z}_i)_{1\le i\le n} \in \{1, \ldots, Q\}^n$ is defined as a (measurable) function of the observation $\mathbf{X}$ returning a vector of labels $\widehat{\mathbf{Z}}(\mathbf{X})$ for which the label q is assigned to individual i if and only if $\hat{Z}_i(\mathbf{X}) = q$. For better readability, we write $\widehat{\mathbf{Z}}$ instead of $\widehat{\mathbf{Z}}(\mathbf{X})$ in the sequel. The classification error of $\widehat{\mathbf{Z}}$, with respect to specific labels, is given by $\varepsilon(\widehat{\mathbf{Z}}, \mathbf{Z}) = \sum_{i=1}^n \mathbb{1}\{Z_i \neq \hat{Z}_i\}$.

In the unsupervised setting, only the partition of the observations is of interest, not the labels themselves. Switching the labels of $\widehat{\mathbf{Z}}$ does not change the corresponding partition. This is in stark difference to the supervised setting. An invariant label-switching error is the clustering risk of $\widehat{\mathbf{Z}}$ defined by

$$R(\widehat{\mathbf{Z}}) := \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(n^{-1}\varepsilon(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \mid \mathbf{X}\right)\right), \tag{1}$$

where $[Q]$ denotes the set of all permutations on $\{1, \ldots, Q\}$. The minimum over all permutations σ is the way to handle the aforementioned label-switching problem.

> *Remark* 1. The position of the minimum w.r.t. σ in the expression of the risk (1) matters: the permutation σ is allowed to depend on X but not on Z. Hence, this risk has to be understood as being computed up to a data-dependent label switching. This definition coincides with the usual definition of the misclassification risk in the situation where the true clustering is deterministic, see Lei and Rinaldo (2015), Lu and Zhou (2016). Hence, it can be seen as a natural extension of the latter to a mixture model where the true clustering is random.

Classically, we aim to find a clustering rule $\widehat{\mathbf{Z}}$ such that the clustering risk is "small". However, as mentioned above, whether this is possible or not depends on the intrinsic difficulty of the clustering problem and thus of the true parameter θ^* (see Figure 1). Therefore, the idea is to provide a selection rule S, that is, a (measurable) function of the observation $\mathbf{X}$ returning a subset of indices $S(\mathbf{X}) \subset \{1, \ldots, n\}$, such that the clustering risk *with restriction to* $S(\mathbf{X})$ is small. Again, we write S instead of $S(\mathbf{X})$ for short. Throughout the paper, a *procedure* refers to a couple $C = (\widehat{\mathbf{Z}}, S)$, where $\widehat{\mathbf{Z}}$ is a clustering rule and S is a selection rule.

> **Definition 1** (False selection rate). The false selection rate (FSR) of a procedure $C = (\widehat{\mathbf{Z}}, S)$ is given by

$$\mathrm{FSR}_{\theta^*}(C) := \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\frac{\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z})}{\max(|S|, 1)} \,\middle|\, \mathbf{X}\right)\right), \tag{2}$$

> where $\varepsilon_S(\widehat{\mathbf{Z}}, \mathbf{Z}) = \sum_{i\in S} \mathbb{1}\{Z_i \neq \hat{Z}_i\}$ denotes the misclassification error restricted to subset S.

In this work, the aim is to find a procedure C such that the false selection rate is controlled at a nominal level α, that is, $\mathrm{FSR}_{\theta^*}(C) \leq \alpha$. Obviously, choosing S empty implies $\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) = 0$ for any permutation σ and thus satisfies this control. Hence, while maintaining the control $\mathrm{FSR}_{\theta^*}(C) \leq \alpha$, we aim to classify as much individuals as possible, that is, to make $\mathbb{E}_{\theta^*}|S|$ as large as possible.

The definition of the FSR (2) involves an expectation of a ratio, which is more difficult to handle than a ratio of expectations. Hence, the following simpler alternative criterion will also be useful in our analysis.

Definition 2 (Marginal false selection rate). The *marginal* false selection rate (mFSR) of a procedure $C = (\widehat{\mathbf{Z}}, S)$ is given by

$$\mathrm{mFSR}_{\theta^*}(C) := \frac{\mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right)\right)}{\mathbb{E}_{\theta^*}(|S|)}, \tag{3}$$

with the convention $0/0 = 0$.

Note that the mFSR is similar to the criterion introduced in Denis and Hebiri (2020) in the supervised setting.

2.3 | Notation

We will use the following notation: for all $q \in \{1,\dots,Q\}$ and $\theta = (\pi, \phi) \in \Theta$, we let

$$\ell_q(X,\theta) := \mathbb{P}_\theta(Z = q \mid X) = \frac{\pi_q f_{\phi_q}(X)}{\sum_{\ell=1}^{Q} \pi_\ell f_{\phi_\ell}(X)}; \tag{4}$$

$$T(X,\theta) := 1 - \max_{q\in\{1,..,Q\}} \ell_q(X,\theta) \in [0, 1 - 1/Q]. \tag{5}$$

Hence, $\ell_q(X,\theta)$ is the posterior probability of belonging to class q given the measurement X under the distribution P_θ. The quantity $T(X,\theta)$ is a measure of the risk when classifying X: it is close to 0 when there exists a class q such that $\ell_q(X,\theta)$ is close to 1, that is, when X can be classified with large confidence.

3 | METHODS

In this section, we introduce new methods for controlling the FSR. We start by identifying an *oracle* method, that uses the true value of the parameter θ^*. Substituting the unknown parameter θ^* by an estimator in that oracle provides our first method, called the *plug-in* procedure. We then define a refined version of the plug-in procedure, that accounts for the variability of the estimator and is based on a *bootstrap* approach.

3.1 | Oracle procedures

Here, we proceed as if an oracle had given us the true value of θ^* and we introduce an oracle procedure based on this value. The oracle procedure C_α^{or} is defined in Algorithm 1 and described in detail below.

3.1.1 | MAP clustering

As the following lemma shows, the best clustering rule is well-known and given by the Bayes clustering $\widehat{\mathbf{Z}}^* = (\widehat{Z}_1^*, \ldots, \widehat{Z}_n^*)$, which can be written as the maximum a posteriori (MAP) rule

$$\widehat{Z}_i^* \in \underset{q \in \{1,..,Q\}}{\operatorname{argmax}} \ell_q(X_i, \theta^*), \quad i \in \{1, \ldots, n\}, \tag{6}$$

where $\ell_q(\cdot)$ is the posterior probability given by (4).

Lemma 1. *We have* $\min_{\widehat{\mathbf{Z}}} R(\widehat{\mathbf{Z}}) = R(\widehat{\mathbf{Z}}^*) = n^{-1} \sum_{i=1}^n \mathbb{E}_{\theta^*}(T_i^*) = \mathbb{E}_{\theta^*}(T_1^*)$, *for the Bayes clustering* $\widehat{\mathbf{Z}}^*$ *defined by* (6) *and for*

$$T_i^* := T(X_i, \theta^*) = \mathbb{P}_{\theta^*}(Z_i \neq \widehat{Z}_i^* | X_i), \quad i \in \{1, \ldots, n\}, \tag{7}$$

where $T(\cdot)$ *is given by* (5).

Lemma 1 is proved in Appendix B1. In words, it states that the oracle statistics T_i^* correspond to the posterior misclassification probabilities of the Bayes clustering. To decrease the overall misclassification risk, it is natural to avoid the classification of points with a high value of the test statistic T_i^*.

3.1.2 | Thresholding selection rules

In this section, we introduce the selection rule that decides which items are to be classified. From the above paragraph, it is natural to consider a thresholding-based selection rule of the form $S_t^* := \{i \in \{1, \ldots, n\} \ : \ T_i^* \leq t\}$, for some threshold t to be chosen suitably. The following result gives insights for the choice of such a threshold t.

Lemma 2. *For a procedure* $C = (\widehat{\mathbf{Z}}^*, S)$ *with Bayes clustering and an arbitrary selection* S,

$$\mathrm{FSR}_{\theta^*}(C) = \mathbb{E}_{\theta^*}\left(\frac{\sum_{i \in S} T_i^*}{\max(|S|, 1)} \right). \tag{8}$$

Lemma 2 is proved in Appendix B2. As a consequence, a first way to build an (oracle) selection with $\mathrm{FSR}_{\theta^*}(C) \leq \alpha$ is to set $S_\alpha^* = \{i \in \{1, \ldots, n\} \ : \ T_i^* \leq \alpha\}$ (an average of numbers smaller than α is also smaller than α). This procedure is referred to as the procedure *with fixed threshold* in the sequel. It corresponds to the following naive approach: to get a clustering with a risk of α, we only keep the items that are in their corresponding class with a posterior probability of at least $1 - \alpha$.

By contrast, the oracle selection rule considered here is given by $S^* := \{i \in \{1, \ldots, n\} \ : \ T_i^* \leq t(\alpha)\}$, for a threshold $t(\alpha) \geq \alpha$ maximizing $|S|$ under the constraint $\sum_{i \in S} T_i^* \leq \alpha |S|$. It has a threshold that adapts to θ^*, uniformly improves the procedure with fixed threshold and will in general lead to a (much) broader selection. This gives rise to the *oracle procedure*, that can be easily implemented by ordering the T_i^*'s, see Algorithm 1.

Algorithm 1. Oracle procedure

Input: Parameter θ^*, sample $(X_1, \dots, X_n)$, level α.
1. Compute the posterior probabilities $\mathbb{P}_{\theta^*}(Z_i = q|X_i)$, $1 \le i \le n$, $1 \le q \le Q$;
2. Compute the Bayes clustering $\widehat{Z}_i^*$, $1 \le i \le n$, according to (6);
3. Compute the probabilities T_i^*, $1 \le i \le n$, according to (7);
4. Order these probabilities in increasing order $T_{(1)}^* \le \cdots \le T_{(n)}^*$;
5. Choose k^* the maximum of $k \in \{0, \dots, n\}$ such that $\max(k, 1)^{-1} \sum_{j=1}^{k} T_{(j)}^*(\mathbf{X}) \le \alpha$;
6. Set S^* to the set of indices of the k^* smallest elements among the T_i^*'s.
Output: Oracle procedure $C_\alpha^{or} = (\widehat{\mathbf{Z}}^*, S^*)$.

Algorithm 2. Plug-in procedure

Input: Sample $(X_1, \dots, X_n)$, level α.
1. Compute an estimator $\hat{\theta}$ of θ based on $(X_1, \dots, X_n)$;
2. Run the oracle procedure given in Algorithm 1 with $\hat{\theta}$ in place of θ^*, returning a procedure that is denoted $\widehat{C}_\alpha^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$.
Output: Plug-in procedure $\widehat{C}_\alpha^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$.

3.2 | Empirical procedures

3.2.1 | Plug-in procedure

The oracle procedure cannot be used in practice since θ^* is generally unknown. A natural idea then is to replace θ^* by an estimator $\hat{\theta}$ and to plug this estimate into the oracle procedure. The resulting procedure, denoted $\widehat{C}^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$, is called the *plug-in procedure* and is implemented in Algorithm 2.

In Section 4, we establish that the plug-in procedure has convenient properties: when n tends to infinity, provided that the chosen estimator $\hat{\theta}$ behaves well and under mild regularity assumptions on the model, the FSR of the plug-in procedure is close to the level α, while it is nearly optimal in terms of the average selection number.

3.2.2 | Bootstrap procedure

Despite the favorable theoretical properties shown in Section 4, the plug-in procedure achieves an FSR that can exceed α in some situations, as we will see in our numerical experiments (Section 5). This is in particular the case when the estimator $\hat{\theta}$ is too rough. Indeed, the uncertainty of $\hat{\theta}$ is ignored by the plug-in procedure.

To take into account this effect, we propose to use a bootstrap approach. It is based on the following result.

Lemma 3. *For a given level $\alpha \in (0, 1)$, the FSR of the plug-in procedure $\widehat{C}_\alpha^{\mathrm{PI}}$ is given by*

$$\mathrm{FSR}_{\theta^*}(\widehat{C}_\alpha^{\mathrm{PI}}) = \mathbb{E}_{\mathbf{X}\sim P_{\theta^*}}\left(\min_{\sigma\in[Q]} \frac{\sum_{i=1}^{n} \{1 - \ell_{\sigma(\widehat{Z}_i^{\mathrm{PI}}(\mathbf{X}))}(X_i, \theta^*)\} \mathbb{1}\{i \in \widehat{S}_\alpha^{\mathrm{PI}}(\mathbf{X})\}}{\max(|\widehat{S}_\alpha^{\mathrm{PI}}(\mathbf{X})|, 1)} \right). \tag{9}$$

Algorithm 3. Bootstrap procedure

Input: Sample $(X_1, \dots, X_n)$, level α, number B of bootstrap runs.
1. Choose a grid of increasing levels $(\alpha(k))_{1\le k\le K}$;
2. Draw B bootstrap samples and apply Algorithm 2 to each of them.
3. Compute $\widehat{\mathrm{FSR}}^{B}_{\alpha(k)}$, $1 \le k \le K$, according to (10);
4. Choose $\tilde{k}$ according to (11).
Output: Bootstrap procedure $\widehat{\mathcal{C}}^{\mathrm{boot}}_{\alpha} = \widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha(\tilde{k})}$.

Lemma 3 is proved in Appendix B3. The general idea is as follows: since $\mathrm{FSR}_{\theta^*}(\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha})$ can exceed α, we choose α' as large as possible such that $\widehat{\mathrm{FSR}}_{\alpha'} \le \alpha$, where $\widehat{\mathrm{FSR}}_{\alpha'}$ is a bootstrap approximation of $\mathrm{FSR}_{\theta^*}(\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha'})$ based on (9).

The bootstrap approximation reads as follows: in the RHS of (9), we replace the true parameter θ^* by $\hat{\theta}$ and $\mathbf{X} \sim P_{\theta^*}$ by $\mathbf{X}' \sim \hat{P}$, where $\hat{P}$ is an empirical substitute of P_{θ^*}. This empirical distribution $\hat{P}$ is $P_{\hat{\theta}}$ for the parametric bootstrap and the uniform distribution over the X_i's for the nonparametric bootstrap. This yields the bootstrap approximation of $\mathrm{FSR}_{\theta^*}(\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha})$ given by

$$\widehat{\mathrm{FSR}}_{\alpha} := \mathbb{E}_{\mathbf{X}'\sim\hat{P}}\left(\min_{\sigma\in[Q]} \frac{\sum_{i=1}^{n}\{1-\ell_{\sigma(\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}'))}(X_i', \hat{\theta}(\mathbf{X}))\}\mathbb{1}\{i \in \hat{S}^{\mathrm{PI}}_{\alpha}(\mathbf{X}')\}}{\max(|\hat{S}^{\mathrm{PI}}_{\alpha}(\mathbf{X}')|, 1)} \,\middle|\, \mathbf{X} \right).$$

Above, $\mathbb{E}_{\mathbf{X}'\sim\hat{P}}$ denotes the expectation operator when $\mathbf{X}' = (X_1', \dots, X_n')$ is iid $\sim \hat{P}$. Classically, the latter is itself approximated by a Monte-Carlo scheme:

$$\widehat{\mathrm{FSR}}^{B}_{\alpha} := \frac{1}{B}\sum_{b=1}^{B} \min_{\sigma\in[Q]} \frac{\sum_{i=1}^{n}\{1-\ell_{\sigma(\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}^b))}(X_i^b, \hat{\theta}(\mathbf{X}))\}\mathbb{1}\{i \in \hat{S}^{\mathrm{PI}}_{\alpha}(\mathbf{X}^b)\}}{\max(|\hat{S}^{\mathrm{PI}}_{\alpha}(\mathbf{X}^b)|, 1)}, \tag{10}$$

with bootstrap samples $\mathbf{X}^1, \dots, \mathbf{X}^B$ i.i.d. $\sim \hat{P}$.

Let $(\alpha(k))_{1\le k\le K} \in (0,1)^K$ be a grid of increasing nominal levels (possibly with restriction to values slightly below the target level α). Then, the bootstrap procedure at level α is defined as $\widehat{\mathcal{C}}^{\mathrm{boot}}_{\alpha} = \widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha(\tilde{k})}$, where

$$\tilde{k} = \max\left\{ k \in \{1, \dots, K\} \;:\; \widehat{\mathrm{FSR}}^{B}_{\alpha(k)} \le \alpha \right\}. \tag{11}$$

This procedure is described in Algorithm 3.

Remark 2 (Parametric versus nonparametric bootstrap). The usual difference between parametric and nonparametric bootstrap also holds in our context: the parametric bootstrap is fully based on $P_{\hat{\theta}}$, while the nonparametric bootstrap builds an artificial sample (with replacement) from the original sample, which does not come from a P_{θ}-type distribution. This gives rise to different behaviors in practice: when $\hat{\theta}$ is too optimistic (that is, in cases where the estimated posterior probability to be in each class is biased towards 0 or 1, which will be typically the case when the estimation error is large), the correction brought by the parametric bootstrap (based on $P_{\hat{\theta}}$) is often weaker and insufficient compared to that of the nonparametric one. By contrast,

when $\hat{\theta}$ is close to the true parameter, the parametric bootstrap approximation is more faithful because it uses the correct model, see Section 5.

Remark 3 (Weighted likelihood-based nonparametric bootstrap). We explored another bootstrap approach in Appendix E1, which is suitable in the presence of small or overlapping clusters. It is based on weighting the likelihood which is conceptually different, see O'Hagan et al. (2019). However, it seems to provide only a minor improvement of the plug-in procedure compared to the parametric/nonparametric bootstrap approaches mentioned above.

4 | THEORETICAL GUARANTEES FOR THE PLUG-IN PROCEDURE

In this section, we derive theoretical properties for the plug-in procedure: we show that its FSR and mFSR are close to α, while its expected selection number is close to be optimal under some conditions.

4.1 | Additional notation and assumptions

We make use of an optimality theory for mFSR control, that will be developed in detail in Appendix A4. This approach extensively relies on the following quantities (recall the definition of $T(X,\theta)$ in (5)):

$$\mathrm{mFSR}_t^* := \mathbb{E}_{\theta^*}(T(X,\theta^*) \mid T(X,\theta^*) < t); \tag{12}$$

$$t^*(\alpha) := \sup\left\{t \in [0,1] \ : \ \mathrm{mFSR}_t^* \le \alpha\right\}; \tag{13}$$

$$\alpha_c := \inf\{\mathrm{mFSR}_t^* \ : \ t \in (0,1], \mathrm{mFSR}_t^* > 0\}; \tag{14}$$

$$\overline{\alpha} := \mathrm{mFSR}_1^*. \tag{15}$$

In words, mFSR_t^* is the mFSR of a procedure that selects the items with T_i^* smaller than some threshold t (Lemma 8). Then, $t^*(\alpha)$ is the optimal threshold such that this procedure has an mFSR controlled at level α. (This optimal procedure, studied in Appendix A4, is not the same as the oracle procedure C_α^{or} defined in Section 3.1, although these two procedures are expected to behave roughly in the same way for a large n.) Next, α_c and $\overline{\alpha}$ are the lower and upper bounds for the nominal level α, respectively, for which the optimality theory can be applied.

Now, we introduce our main assumption, which will be ubiquitous in our analysis.

Assumption 1. For all $\theta \in \Theta$ and $q \in \{1,\dots,Q\}$, under P_{θ^*}, the r.v. $\ell_q(X,\theta)$ given by (4) is continuous. In addition, the function $t \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) < t)$ is increasing on $(\alpha_c, \overline{\alpha})$, where $T(X,\theta)$ is given by (5).

Note that Assumption 1 implies the continuity of the r.v. $T(X,\theta)$. Indeed, $\mathbb{P}(T(X,\theta) = t) \le \sum_{q=1}^{Q} \mathbb{P}(\ell_q(X,\theta) = 1 - t)$. Hence, this assumption implies that $t \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) < t)$ is both continuous on $[0,1]$ and increasing on $(\alpha_c, \overline{\alpha})$. This is useful in several regards: first, it prohibits ties in the $T(X_i,\theta)$'s, $1 \le i \le m$, so that the selection rule (see Algorithm 1) can be truly formulated as a thresholding rule (see Lemma 9). Second, it entails interesting properties for function $t \mapsto \mathrm{mFSR}_t^*$,

see Lemma 8 (this in particular ensures that the supremum in (13) is a maximum). Also note that the inequality $0 \leq \alpha_c < \overline{\alpha} < 1 - 1/Q$ holds under Assumption 1.

The next assumption ensures that the density family $\{f_u, u \in \mathcal{U}\}$ is smooth, and will be useful to establish consistency results.

Assumption 2. For P_{θ^*}-almost all $x \in \mathbb{R}^d$, $u \in \mathcal{U} \mapsto f_u(x)$ is continuous.

Moreover, we can derive convergence rates under the following additional regularity conditions.

Assumption 3. There exist positive constants $r = r(\theta^*), C_1 = C_1(\theta^*), C_2 = C_2(\theta^*, \alpha), C_3 = C_3(\theta^*, \alpha)$ such that

(i) for $\mathbb{P}_{\theta^*}$-almost all $x \in \mathbb{R}^d$, $u \in \mathcal{U} \mapsto f_u(x)$ is continuously differentiable, and

$$\sum_{1 \leq q \leq Q} \mathbb{E}_{\theta^*} \sup_{\substack{\theta \in \Theta \\ \|\theta - \theta^*\|_2 \leq r}} \|\nabla_\theta \ell_q(X, \theta)\|_2 \leq C_1;$$

(ii) for all $t, t' \in [0,1]$, $|\mathbb{P}_{\theta^*}(T(X, \theta^*) < t) - \mathbb{P}_{\theta^*}(T(X, \theta^*) < t')| \leq C_2 |t - t'|$;
(iii) for all $\beta \in [(\alpha_c + \alpha)/2, (\alpha + \overline{\alpha})/2]$, $|t^*(\beta) - t^*(\alpha)| \leq C_3 |\beta - \alpha|$.

Example 1. In Appendix D, it is proved that Assumptions 1, 2 and 3 hold true in the homoscedastic two-component multivariate Gaussian mixture model, see Lemma 17.

Next, we consider the following complexity assumption to ensure concentration of the underlying empirical processes. It is given in terms of the Vapnik-Chervonenkis (VC) dimension of specific function classes involving ℓ_q. In the sequel, the VC dimension of a function set $\mathcal{F}$ is defined as the VC dimension of the set family $\{\{x \in \mathbb{R}^d : f(x) \geq u\}, f \in \mathcal{F}, u \in \mathbb{R}\}$, see, e.g., Baraud (2016). We denote

$$\mathscr{V} = \text{VC dimension of } \{\ell_q(., \theta), \theta \in \Theta, 1 \leq q \leq Q\}; \tag{16}$$

$$\mathscr{V}_- = \text{VC dimension of } \{\mathbb{1}\{\ell_q(., \theta) - \ell_{q'}(., \theta) \geq 0\}, \theta \in \Theta, 1 \leq q, q' \leq Q\}. \tag{17}$$

Assumption 4. The VC dimensions $\mathscr{V}$ and $\mathscr{V}_-$ are finite.

Example 2. In the two-component case, $Q = 2$ where P_θ belongs to an exponential family, we have that $\mathscr{V}, \mathscr{V}_- = O(k^2 \log(k))$ (see Lemma 13) with k the dimension of the sufficient statistic vector. For instance, $k = d + d^2$ for the Gaussian family, hence $\mathscr{V}, \mathscr{V}_- = O(d^4 \log(d))$ in that case. (For the specific case of the homoscedastic Gaussian family, we have that $\mathscr{V}, \mathscr{V}_- = O(d)$, see Lemma 18).

Let us now discuss conditions on the estimator $\hat{\theta}$ on which the plug-in procedure is based. We start by introducing the following assumption (used in the concentration part of the proof, see Lemma 11).

Assumption 5. The estimator $\hat{\theta}$ is assumed to take its values in a countable subset $\mathcal{D}$ of Θ.

This assumption is a minor restriction, because we can always choose $\mathcal{D} \subset \mathbb{Q}^K$ (recall $\Theta \subset \mathbb{R}^K$). Next, we additionally define a quantity measuring the quality of the estimator: for all $\epsilon > 0$,

$$\eta(\epsilon, \theta^*) = \mathbb{P}_{\theta^*}\left(\min_{\sigma \in [Q]} \|\hat{\theta}^\sigma - \theta^*\|_2 \geq \epsilon\right). \tag{18}$$

Above, $\hat{\theta}^{\sigma} = (\hat{\pi}^{\sigma}, \hat{\phi}^{\sigma})$ denotes the estimator $\theta^{\sigma} = (\hat{\pi}, \hat{\phi})$ after having permuted the labels according to the permutation σ, that is, with $\hat{\pi}^{\sigma}_q = \hat{\pi}_{\sigma(q)}$ and $\hat{\phi}^{\sigma}_q = \hat{\phi}_{\sigma(q)}$ for $1 \leq q \leq Q$.

Example 3. The literature provides several results regarding the estimation of Gaussian mixtures, see e.g. Regev and Vijayaraghavan (2017) for a review. Proposition 1 revisits some of these results, for the estimator derived from the EM algorithm (Balakrishnan et al., 2017; Dempster et al., 1977) and the constrained MLE (Ho & Nguyen, 2016).

4.2 | Results

We now state our main results, starting with the consistency of the plug-in procedure.

Theorem 1 (Asymptotic optimality of the plug-in procedure). *Let Assumptions 1, 2, and 4 be true. Consider an estimator $\hat{\theta}$ satisfying Assumption 5 that is assumed to be consistent in the sense that for all $\epsilon > 0$, the probability $\eta(\epsilon, \theta^*)$ given by (18) tends to 0 as n tends to infinity. Then the corresponding plug-in procedure $\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha}$ (Algorithm 2) satisfies the following: for any $\alpha \in (\alpha_c, \overline{\alpha})$, we have*

$$\limsup_n \mathrm{FSR}(\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha}) \leq \alpha, \quad \limsup_n \mathrm{mFSR}(\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha}) \leq \alpha, \tag{19}$$

and for any procedure $\mathcal{C} = (\hat{\mathbf{Z}}, S)$ that controls the mFSR at level α, we have

$$\liminf_n \{ n^{-1}\mathbb{E}_{\theta^*}(|\widehat{S}^{\mathrm{PI}}_{\alpha}|) - n^{-1}\mathbb{E}_{\theta^*}(|S|) \} \geq 0. \tag{20}$$

Next, we derive convergence rates under the additional regularity conditions given by Assumption 3.

Theorem 2 (Optimality of the plug-in procedure with rates). *Consider the setting of Theorem 1, where in addition Assumption 3 holds. Take $\alpha \in (\alpha_c, \overline{\alpha})$ and recall $\eta(\epsilon, \theta^*)$ defined by (18). With constants $A > 0$ and $B > 0$ only depending on $Q, C_1, C_2, C_3, \mathscr{V}$ (16), $\mathscr{V}\_(17)\alpha$ and θ^*, we have for any sequence $\epsilon_n > 0$ tending to zero, for n larger than a constant only depending on α and θ^*,*

$$\mathrm{FSR}(\widehat{\mathcal{C}}^{\mathrm{PI}}_{\alpha}) \leq \alpha + A\sqrt{\epsilon_n} + B\sqrt{\log n/n} + 5/n^2 + \eta(\epsilon_n, \theta^*) \tag{21}$$

$$n^{-1}\mathbb{E}_{\theta^*}(|\widehat{S}^{\mathrm{PI}}_{\alpha}|) - n^{-1}\mathbb{E}_{\theta^*}(|S|) \geq -A\sqrt{\epsilon_n} - B\sqrt{\log n/n} - \eta(\epsilon_n, \theta^*), \tag{22}$$

for any procedure $\mathcal{C} = (\hat{\mathbf{Z}}, S)$ that controls the mFSR at level α.

The proofs of Theorems 1 and 2 are based on a more general nonasymptotical result, for which the remainder terms are more explicit, see Theorem 3 in Appendix A. A sketch of the proof is as follows: first, we identify the optimal procedure, which corresponds to invert the mFSR_{θ^*} functional with respect to the threshold of the procedure, which is possible thanks to Assumption 1. It is optimal in the sense that it maximizes the expected size of the selection among thresholding procedures that control the mFSR at level α (see Lemmas 6 and 8). Second, we quantify how the plug-in procedure (which can be written as a thresholding-based procedure, see Lemma 9) gets to

the optimal procedure when n grows. For this, we investigate how the expected quantities vary in θ (Lemma 10), by using a regularity assumption (specifically, Assumption 2 or Assumption 3) and how the empirical quantities approximate the expected ones (Lemma 11), by using Assumption 4. This allows to approximate the FSR (resp. expected selection size) of the plug-in procedure by the one of the optimal procedure, which leads to (19), (21) (resp. 20, 22). Overall, this proof uses techniques that share similarities with the work of Rebafka et al. (2022) developed in a different context. Here, an additional difficulty is to handle the new statistic $T(X_i, \hat{\theta})$ which is defined as an extremum, see (5).

Theorem 2 establishes that, given a model which is regular enough and a consistent estimator, the plug-in procedure controls the FSR and is asymptotically optimal up to remainder terms which are of the order of $\sqrt{\epsilon_n} + \sqrt{\log n/n} + \eta(\epsilon_n, \theta^*)$. Here, ϵ_n dominates the convergence rate of the parameter estimate, and is taken large enough to ensure that $\eta(\epsilon_n, \theta^*)$ vanishes.

For instance, in the multivariate Gaussian mixture model (with further assumptions) and by considering either the EM estimator or the constrained MLE, we have $\eta(\epsilon_n, \theta^*) \leq 1/n$ for $\epsilon_n = C\sqrt{\log(n)/n}$, see Proposition 1. This implies that the remainder terms in (21) and (22) are at most of order $((\log n)/n)^{1/4}$.

5 | EXPERIMENTS

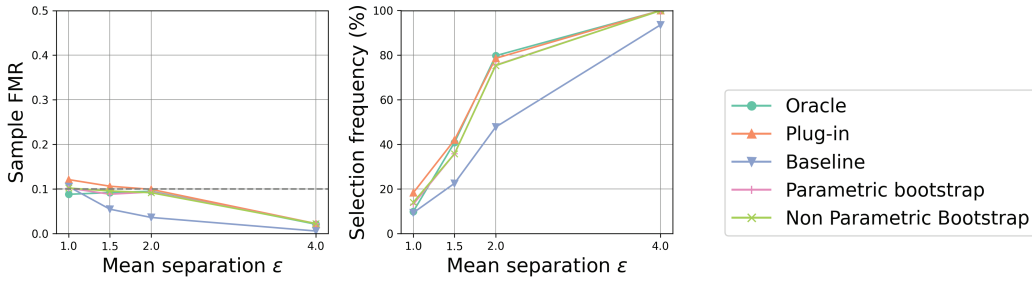
In this section, we evaluate the performance of the new procedures: plug-in (Algorithm 2), parametric bootstrap and nonparametric bootstrap (Algorithm 3). For this, we use both synthetic and real data. We publicly release the code of the numerical experiments at https://github.com/arianemarandon/fsrcontrol.

5.1 | Synthetic data

The performance of our procedures is studied via simulations in different settings with various difficulties. In this section, all mixtures are Gaussian, whereas Appendix E2 provides additional results for Student's t-mixture models. For parameter estimation, the classical EM algorithm is applied with 100 iterations and 10 starting points chosen with Kmeans++ (Arthur & Vassilvitskii, 2007). To help avoiding degenerate likelihood maxima when estimating covariance matrices (Chen, 2017), the covariances are regularized by adding 1e-6 to the diagonal after each M-step. In the bootstrap procedures, $B = 1000$ bootstrap samples are generated. The performance of all procedures is assessed via the *sample FSR*and the proportion of classified data points, which is referred to as the *selection frequency*. For every setting and every set of parameters, depicted results display the mean over 100 simulated datasets. As a baseline, we consider the fixed threshold procedure in which one selects data points that have a maximum posterior group membership probability that exceeds $1 - \alpha$. The oracle procedure (Algorithm 1) is also considered in our experiments for comparison.

5.1.1 | Degree of cluster separation

In the first setting, the true mixture proportions and covariance matrices are known and used in the EM algorithm. We consider the case $Q = 2$, $\pi_1 = \pi_2 = 1/2$ and $\Sigma_1 = \Sigma_2 = I_d$ with I_d the



(a) FSR (b) Selection frequency

FIGURE 2 Cluster separation: (a) FSR and (b) selection frequency as a function of the mean separation ϵ. Known mixture proportions and covariances setting with $Q = 2$, $n = 100$, $d = 2$, $\alpha = 0.1$.

$(d \times d)$-identity matrix. For the mean vectors, we set $\mu_1 = 0$ and $\mu_2 = (\epsilon/\sqrt{d}, \ldots, \epsilon/\sqrt{d})$. The quantity ϵ corresponds to the mean separation, that is, $\|\mu_1 - \mu_2\|_2 = \epsilon$ and accounts for the difficulty of the clustering problem.

Figure 2a displays the FSR for nominal level $\alpha = 0.1$, sample size $n = 100$, dimension $d = 2$ and varying mean separation $\epsilon \in \{1, \sqrt{2}, 2, 4\}$. Globally, our procedures all have an FSR close to the target level α (excepted for the very well separated case $\epsilon = 4$ for which the FSR is much smaller because a large part of the items can be trivially classified). In addition, the selection rate (b) is always close to the one of the oracle procedure. On the other hand, the baseline procedure is rather conservative: its FSR can be well below the nominal level and it selects up to 50% less than the other procedures. This is well expected, because unlike our procedures, the baseline has a fixed threshold and thus does not adapt to the difficulty of the problem.

We also note that the FSR of the plug-in approach is slightly inflated for a weak separation ($\epsilon = 1$). This comes from the parameter estimation, which is difficult in that case. This also illustrates the interest of the bootstrap methods, that allow to recover the correct level in that case, by appropriately correcting the plug-in approach.

5.1.2 | Unknown diagonal covariance matrices

In this setting, the true parameters are the same as in the previous paragraph, but the true mixture proportions and covariance matrices are unknown when fitting the mixture model. However, to help the estimation, we suppose a diagonal structure for Σ_1 and Σ_2, which is used in the EM algorithm.

Figure 3a displays the FSR and the selection frequency as a function of the separation ϵ. The conclusion is qualitatively the same as in the previous case, but with larger FSR values for a weak separation. Overall, it shows that the plug-in procedure is anticonservative and that the bootstrap corrections are able to recover an FSR and a selection frequency close to the one of the oracle. However, for a weak separation, namely $\epsilon = 1$, the parametric bootstrap correction is not enough and the latter procedure still overshoots the nominal level α. Indeed, in our simulations, it appears that $P_{\hat{\theta}}$ is often a distribution that is more favorable than P_{θ^*} from a statistical point of view (for instance, with more separated clusters). These conclusions also hold for varying sample size n, see Figure 3b.

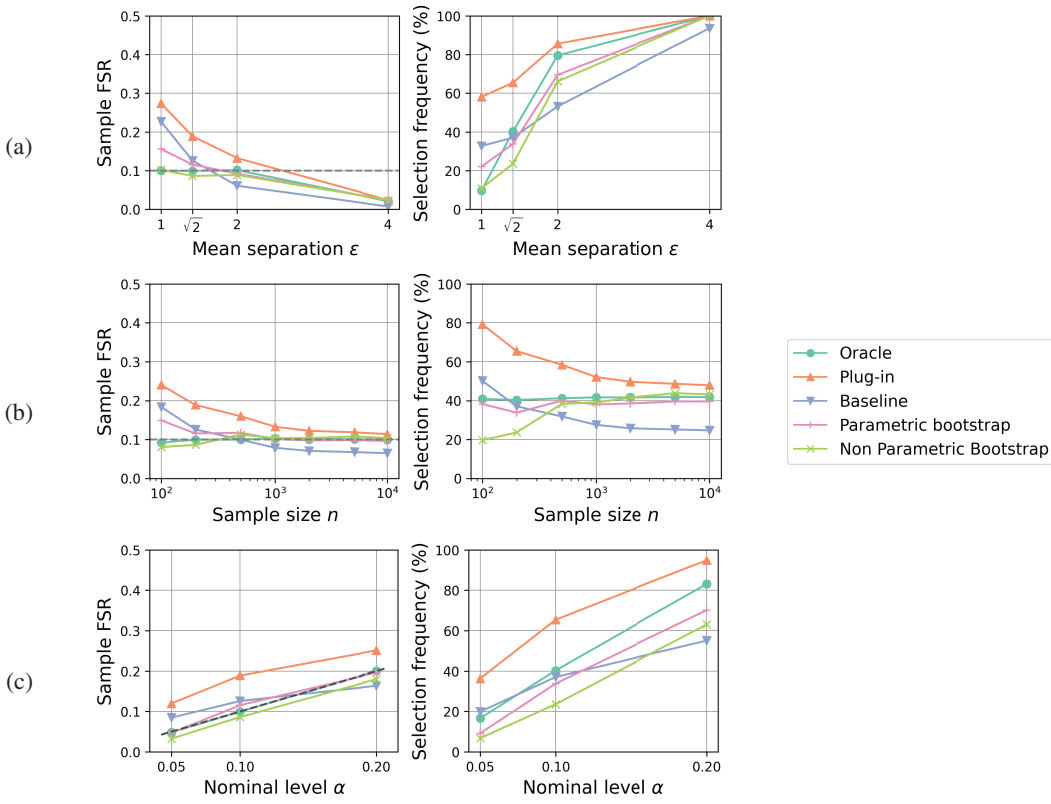
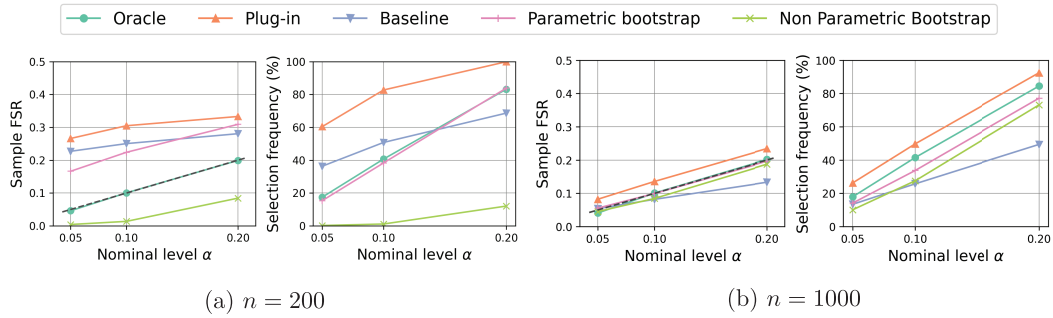


FIGURE 3 Unknown covariance matrices: FSR (left panel) and selection frequency (right panel) as a function of: (a) the mean separation; (b) the sample size n; and (c) the nominal level α. Diagonal covariance matrix setting with $Q = 2$, $d = 2$. Default settings are: $n = 200$, $\alpha = 0.1$, $\epsilon = \sqrt{2}$.

Figure 3c displays the FSR and the selection frequency for varying nominal level α, with $\epsilon = \sqrt{2}$ and $n = 200$. The plug-in is still anticonservative, while the bootstrap procedures have an FSR that is close to α uniformly on the considered α range. Moreover, we note that for all our procedures (including the plug-in), the gap between the FSR and the nominal level is roughly constant with α: this illustrates the adaptive aspect of our procedures. This is in contrast with the baseline procedure, for which this gap highly depends on α, and which may be either anticonservative or suboptimal depending on the α value.

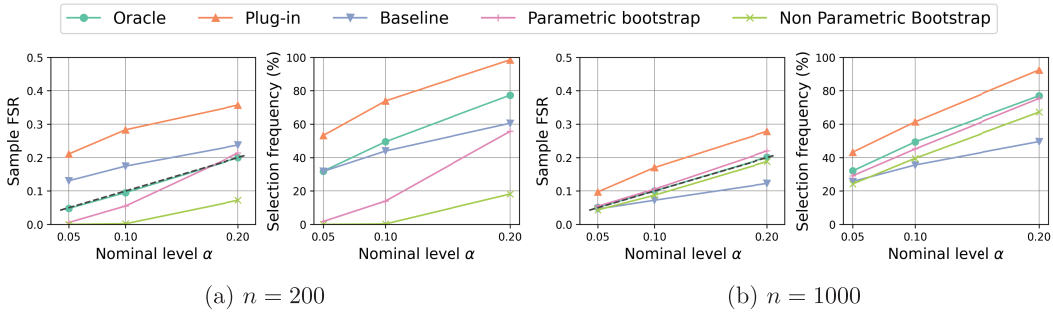
5.1.3 | Larger dimension

We now increase the dimension to $d = 10$. As in the previous setting, we consider a two-component mixture with balanced proportions and covariance matrices $\Sigma_q = I_d$, for any q, the diagonal structure being used as a constraint in the estimation. The mean separation $\|\mu_1 - \mu_2\| = \epsilon$ is set to $\sqrt{2}$. Figure 4 displays the FSR and the selection frequency for varying α. In that case, the parameter estimation is more challenging, because the maximum posterior probability for any point tends to be over-estimated. Hence, taking $n = 200$ is not enough to obtain



(a) $n = 200$ (b) $n = 1000$

FIGURE 4 Large dimension: FSR (left panel) and selection frequency (right panel) as a function of the nominal level α. Diagonal covariances setting with $Q = 2$, $d = 10$, $\epsilon = \sqrt{2}$.



(a) $n = 200$ (b) $n = 1000$

FIGURE 5 5-component mixture model: FSR (left panel) and selection frequency (right panel) as a function of the nominal level α. Diagonal covariances setting with $Q = 5$, $d = 2$, $\epsilon = 2$.

satisfactory performances. For $n = 1000$, however, the conclusions are qualitatively similar to dimension $d = 2$, which becomes clear by comparing Figures 4b and 3c.

5.1.4 | Five-component mixture

We next increase the number of classes to $Q = 5$, and set $\mu_1 = (0, 0)$, $\mu_2 = (\epsilon, 0)$, $\mu_3 = (0, \epsilon)$, $\mu_4 = (-\epsilon, 0)$, $\mu_5 = (0, -\epsilon)$ with $\epsilon = 2$. The mean separation is chosen so that the selection frequency of the oracle rule is approximately the same as in the previous cases. The mixture proportions are $\pi_q = 1/Q$, for any q and the covariance matrices are $\Sigma_q = I_d$, for any q, the diagonal structure being used as a constraint in the estimation. Figure 5 displays the FSR and the selection frequency for varying α. While the problem is more challenging in itself, the conclusions are qualitatively the same as in the previous settings.

5.2 | Real dataset

We consider the Vertebral Column dataset from the UCI ML repository. The data consists of $d = 6$ biomechanical features measured for $n = 310$ orthopaedic patients. Each attribute is derived from the shape and orientation of the pelvis and lumbar spine. Patients are classified

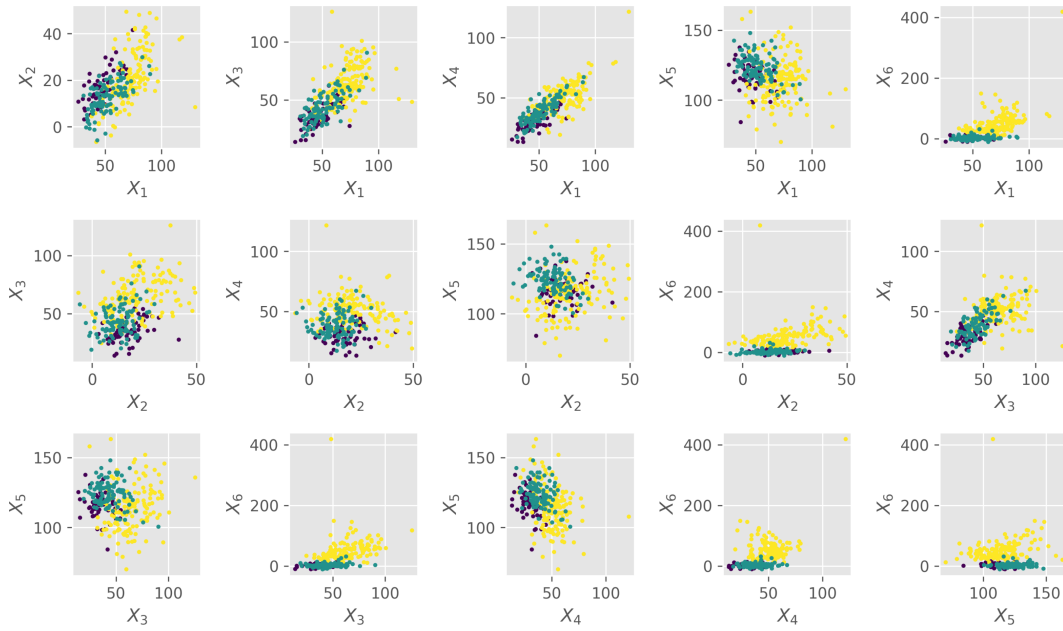


FIGURE 6 Relationship between each pair of features in the Vertebral Column dataset: plot of X_i as a function of X_j for each $i \neq j = 1,\ldots, 6$. Data points are colored according to which class they belong to: "normal" (green), "disk hernia" (purple), and "spondilolysthesis" (yellow).

into three classes: "normal" (100 patients), "disk hernia" (60 patients), or "spondilolysthesis" (150 patients). Figure 6 displays all pairs of variables (X_i, X_j) with $i \neq j$ for all six variables.

We choose to model the data as a mixture of Student's t-distributions as proposed in Peel and McLachlan (2000). Student's t-mixtures are appropriate for data containing observations with longer than normal tails or atypical observations leading to overlapping clusters. No constraints are imposed on the parameters, including the covariance matrices, for which no structural assumptions are made. The t-mixture is fit via the EM algorithm for t-mixtures (Peel & McLachlan, 2000) provided by the Python package `studenttmixture` (Parkinson, 2018). While the number of ground truth classes is $Q = 3$, the EM solution for 3 components leads to a cluster that is almost empty. Furthermore, one cluster essentially corresponds to the "spondilolysthesis" class, whereas members of the two classes "normal" and "disk hernia" are grouped into a unique cluster (which is clear from Figure 6, where the "normal" and "disk hernia" classes are hardly distinguishable). This leads us to run again the procedure for a mixture with only $Q = 2$ components.

For $Q = 2$ Table 1 reports the sample FSR and the selection frequency for varying α, where the FSR is computed with respect to the ground truth classes, by merging the classes "normal" and "disk hernia" into a single class. However, given that the assumption of a Student's t-mixture does not necessarily hold for these ground truth groups, the sample FSR value computed here should be interpreted loosely, and a precise control of this value is not expected. Here the ground truth groups are well separated and all procedures display a FSR value below the nominal level α. In terms of the selection frequency, the parametric bootstrap is close to the plug-in, as in the simulations, whereas the baseline procedure (which is necessarily more conservative than the plug-in) selects less data samples. In contrast to the other procedures, the nonparametric bootstrap displays adaptation to the level, selecting few examples when α is small, and selecting all examples when α is large.

TABLE 1 Vertebral Column dataset: sample FSR and selection frequency of each method for varying α.

α	Baseline	Plug-in	Parametric bootstrap	Nonparametric bootstrap
Sample FSR				
1%	0.009	0.017	0.015	0
5%	0.015	0.032	0.032	0.009
10%	0.018	0.032	0.032	0.032
Selection frequency				
1%	0.739	0.926	0.861	0
5%	0.848	1.000	1.000	0.745
10%	0.919	1.000	1.000	1.000

Note: The number of components Q is set to 2 and the sample FSR is computed with respect to the ground truth classes "normal/disk hernia" (merged into a single class), and "spondilolysthesis".

6 | CONCLUSION AND DISCUSSION

We have presented new data-driven methods providing both clustering and selection that ensure an FSR control guarantee in a mixture model. The plug-in approach was shown to be theoretically valid both when the parameter estimation is accurate and the sample size is large enough. When this is not necessarily the case, we proposed two second-order bootstrap corrections that have been shown to increase the FSR control ability on numerical experiments. Finally, applying our unsupervised methods to a supervised dataset, our approach has been qualitatively validated by considering the attached labels as revealing the true clusters.

We underline that the cluster number Q is assumed to be fixed and known throughout the study. In practice, the cluster number can be unknown, in which case it is estimated from the data. In that case, the interpretation of the FSR control is still meaningful, in the sense that it provides error control within a *specified* model. To obtain a more general guarantee, a solution could be to consider a definition of the clustering error that is able to compare partitions with different levels of granularity (and hence cluster numbers). There are several such measures in use in clustering (not restricted to a selection), such as the Adjusted Rand Index (Hubert & Arabie, 1985), among others (Rosenberg & Hirschberg, 2007; Nguyen et al., 2009). Building a selection rule that controls the clustering error in restriction to a part of the sample is expected to be much more challenging for such a criterion, but represents an interesting research direction.

Concerning the pure task of controlling the FSR in the mixture model, our methods provide a correct FSR control in some region of the parameter space, leaving other less favorable parameter configurations with a slight inflation in the FSR level. This phenomenon is well known for FDR control in the two-component mixture multiple testing model (Roquain & Verzelen, 2022; Sun & Cai, 2007), and facing a similar problem in our framework is well expected. On the one hand, in some cases, this problem can certainly be solved by improving on parameter estimation: here the EM algorithm seems to over-estimate the extreme posterior probabilities, which makes the plug-in procedure too anticonservative. On the other hand, it could be hopeless to expect a robust FSR control uniformly valid over all configurations, while being optimal in the favorable cases. To illustrate that point, we refer to Roquain and Verzelen (2022) who show that such a procedure does not exist in the FDR controlling case, when the null distribution is Gaussian with

an unknown scaling parameter (which is a framework sharing similarities with the one considered here). Investigating such a "lower bound" result in the current setting would provide better guidelines for the practitioner and is therefore an interesting direction for future research.

Another avenue would be to assess the theoretical performances of the bootstrap procedures. While a consistency result for FSR/Power similar to Theorem 1 might be possible to obtain from classical theory (see, e.g., Politis et al., 1999)—but would require a fully devoted work—it seems even more challenging to theoretically justify why the corresponding remainder terms are smaller than those obtained for the plug-in procedure, hence to fully support our experimental findings.

Next, the continuity of the map $t \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) < t)$ (see Assumption 1) is a key property to apply our methodology (see in particular Lemma 6). For instance, Assumption 1 is not true for discrete data (e.g., when considering a Poisson model). However, one idea is to circumvent this difficulty by using a randomization technique, as for the original Neyman–Pearson test. Investigating this point further is left for future research.

Finally, to attempt to overcome problems of model misspecification related to strong distributional assumptions of the mixture model, one may explore the extension of our approach to semiparametric mixture models. In recent years, several new semiparametric mixtures have been proposed in the literature, see Xiang et al. (2019) for an overview. For univariate observations, there are mixtures over a location parameter (Bordes et al., 2006), and others that use shape constraints, namely log concavity, on the densities of the mixture components (Chang & Walther, 2007). For multivariate data, there also exist some approaches, namely mixtures where conditional independence of the covariates is assumed (Benaglia et al., 2009; Hall & Zhou, 2003). In most of these models, identifiability is an issue and the construction of consistent and computationally efficient estimators is challenging. Most often, inference is performed using some EM-type algorithm and kernel-based estimators for the component densities. Adapting our clustering approach to such approaches is challenging both from a practical point of view (in regard of the dimension of the observations and the constraints on the component densities) and from a theoretical point of view (given the limited existing literature proving consistency).

ACKNOWLEDGMENTS

This work has been supported by ANR-16-CE40-0019 (SansSouci), ANR-17-CE40-0001 (BASICS) and by the GDR ISIS through the "projets exploratoires" program (project TASTY). A. Marandon has been supported by a grant from Région Île-de-France ("DIM Math Innov"). We would like to thank Gilles Blanchard and Stéphane Robin for interesting discussions. We are also grateful to Eddie Aamari and Nhat Ho for their help for proving Lemma 16. Finally, we would like to thank anonymous referees and an associate editor for helpful comments.

ENDNOTE

[1]Here, $\lambda_1(\Sigma)$ (resp. $\lambda_d(\Sigma)$) denotes the smallest (resp. largest) eigenvalue of Σ.

ORCID

Ariane Marandon https://orcid.org/0000-0002-9114-2527
Etienne Roquain https://orcid.org/0000-0003-0214-5185

REFERENCES

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, *18*(177), 1–86.

Abraham, K., Castillo, I., & Roquain, É. (2022). Empirical bayes cumulative ℓ-value multiple testing procedure for sparse sequences. *Electronic Journal of Statistics*, *16*(1), 2033–2081.

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., & Lei, L. (2025). Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, *19*(2), 1641–1662.

Arora, S., & Kannan, R. (2005). Learning mixtures of separated nonspherical Gaussians. *Annals of Applied Probability*, *15*(1A), 69–92.

Arthur, D., & Vassilvitskii, S. (2007). *K-means++: The advantages of careful seeding*. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Azizyan, M., Singh, A., & Wasserman, L. (2013). *Minimax theory for high-dimensional gaussian mixtures with sparse mean separation*. In *Proceedings of the 26th international conference on neural information processing systems—volume 2*, NIPS'13 (pp. 2139–2147). Curran Associates Inc., USA.

Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, *45*(1), 77–120.

Bao, H., Shimada, T., Xu, L., Sato, I., & Sugiyama, M. (2020). Pairwise supervision can provably elicit a decision boundary. arXiv preprint arXiv:2006.06207.

Baraud, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) vc-major class. *Electronic Journal of Statistics*, *10*(2), 1709–1728.

Bartlett, P. L., & Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, *9*(59), 1823–1840.

Benaglia, T., Chauveau, D., & Hunter, D. R. (2009). An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, *18*(2), 505–526.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *57*(1), 289–300.

Bordes, L., Mottelet, S., & Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, *34*(3), 1204–1232.

Cai, T., Sun, W., & Wang, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *Journal of the Royal Statistical Society: Series B*, *81*(2), 187–234.

Cao, Y., Cai, T., Feng, L., Gu, L., Gu, J., An, B., Niu, G., & Sugiyama, M. (2022). Generalizing consistent multi-class classification with rejection to be compatible with arbitrary losses. *Advances in Neural Information Processing Systems*, *35*, 521–534.

Chang, G. T., & Walther, G. (2007). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, *51*(12), 6242–6251.

Chen, J. (2017). Consistency of the MLE under mixture models. *Statistical Science*, *32*(1), 47–63.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*(1), 41–46.

Chretien, S., Dombry, C., & Faivre, A. (2019). The Guedon-Vershynin semi-definite programming approach to low dimensional embedding for unsupervised clustering. *Frontiers in Applied Mathematics and Statistics*, *5*, 41.

Cortes, C., DeSalvo, G., & Mohri, M. (2016). *Learning with rejection*. In *Algorithmic learning theory: 27th international conference* (pp. 67–82). Springer.

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, Statistical Methodology*, *39*(1), 1–38.

Denis, C., & Hebiri, M. (2020). Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, *32*(1), 42–72.

Eckardt, J.-N., Röllig, C., Metzeler, K., Heisig, P., Stasik, S., Georgi, J.-A., Kroschinsky, F., Stölzel, F., Platzbecker, U., Spiekermann, K., Krug, U., Braess, J., Görlich, D., Sauerland, C., Woermann, B., Herold, T., Hiddemann, W., Müller-Tidow, C., Serve, H., … Middeke, J. M. (2023). Unsupervised meta-clustering identifies risk clusters in acute myeloid leukemia based on clinical and genetic profiles. *Communications Medicine*, *3*(1), 68.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, *96*(456), 1151–1160.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279.

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. (pp. 4885–4894). Curran Associates Inc. USA.

Giraud, C., & Verzelen, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Mathematical Statistics and Learning*, *1*(3), 317–374.

Grün, B. (2019). *Model-based clustering*. In *Handbook of mixture analysis* (pp. 157–192). Chapman and Hall/CRC.

Hall, P., & Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, *31*(1), 201–224.

Herbei, R., & Wegkamp, M. H. (2006). Classification with reject option. *Canadian Journal of Statistics*, *34*(4), 709–721.

Ho, N., & Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, *10*(1), 271–307.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Lei, J. (2014). Classification with confidence. *Biometrika*, *101*(4), 755–769.

Lei, J., & Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, *43*(1), 215–237.

Lu, Y., & Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. arXiv preprint arXiv:1612.02099.

Mary-Huard, T., Perduca, V., Blanchard, G., & Marie-Laure, M.-M. (2021). Error rate control for classification rules in multiclass mixture models.

Massart, P. (2007). *Concentration inequalities and model selection Ecole d'Eté de Probabilités de saint-flour XXXIII – 2003*Tex2Doc conversion from DLF (18-08-2025)-Batch-III (1st ed.). Springer.

Melnykov, V. (2013). On the distribution of posterior probabilities in finite mixture models with application in clustering. *Journal of Multivariate Analysis*, *122*, 175–189.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. The MIT Press.

Najafi, A., Motahari, S. A., & Rabiee, H. R. (2020). Reliable clustering of Bernoulli mixture models. *Bernoulli*, *26*(2), 1535–1559.

Ng, S.-K. (2019). *Mixture modelling for medical and health sciences*. Chapman and Hall/CRC.

Nguyen, X. V., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In A. P. Danyluk, L. Bottou, & M. L. Littman (Eds.), *Proceedings of the 26th Annual international conference on machine learning, ICML 2009 International Conference Proceeding Series* (Vol. *382*, pp. 1073–1080). ACM. https://doi.org/10.1145/1553374.1553511

Ni, C., Charoenphakdee, N., Honda, J., & Sugiyama, M. (2019). On the calibration of multiclass classification with rejection. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019* (pp. 2582–2592). NeurIPS. https://proceedings.neurips.cc/paper/2019/hash/571d3a9420bfd9219f65b643d0003bf4-Abstract.html

O'Hagan, A., Murphy, T. B., Scrucca, L., & Gormley, I. C. (2019). Investigation of parameter uncertainty in clustering using a gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, *34*(4), 1779–1813.

Parkinson, J. (2018). Python package studenttmixture. https://github.com/jlparki/mix_T

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, *10*(4), 339–348.

Politis, D. N., Romano, J. P., Wolf, M., Politis, D. N., Romano, J. P., & Wolf, M. (1999). *Subsampling in the IID case*. Springer.

Rebafka, T., Roquain, É., & Villers, F. (2022). Powerful multiple testing of paired null hypotheses using a latent graph model. *Electronic Journal of Statistics*, *16*(1), 2796–2858.

Regev, O., & Vijayaraghavan, A. (2017). On learning mixtures of well-separated Gaussians. In C. Umans (Ed.), *58th IEEE annual symposium on foundations of computer science, FOCS 2017* (pp. 85–96). IEEE Computer Society. https://doi.org/10.1109/FOCS.2017.17

Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. In H. Larochelle, M.' A. Ranzato, R. Hadsell, M.{.-}. F. Balcan, & H.{.-}. T. Lin (Eds.), *Advances in neural information*

processing systems 33: Annual conference on neural information processing systems 2020. NeurIPS. https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

Roquain, E., & Verzelen, N. (2022). False discovery rate control with unknown null distribution: Is it possible to mimic the oracle? *Annals of Statistics*, *50*(2), 1095–1123.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In J. Eisner (Ed.), *EMNLP-CoNLL 2007, Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 410–420). ACL. https://aclanthology.org/D07-1043/

Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, *114*(525), 223–234.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Stoitsas, K., Bahulikar, S., de Munter, L., de Jongh, M. A., Jansen, M. A., Jung, M. M., van Wingerden, M., & Van Deun, K. (2022). Clustering of trauma patients based on longitudinal data and the application of machine learning to predict recovery. *Scientific Reports*, *12*(1), 16990.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics*, *31*(6), 2013–2035.

Sun, W., & Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, *102*(479), 901–912.

Vempala, S., & Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, *68*(4), 841–860. Special Issue on FOCS 2002.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wegkamp, M., & Yuan, M. (2011). Support vector machines with a reject option. *Bernoulli*, *17*(4), 1368–1385.

Xiang, S., Yao, W., & Yang, G. (2019). An overview of semiparametric extensions of finite mixture models. *Statistical Science*, *34*(3), 391–404.

How to cite this article: Marandon, A., Rebafka, T., Roquain, E., & Sokolovska, N. (2025). False selection rate control in mixture models. *Scandinavian Journal of Statistics*, 1–47. https://doi.org/10.1111/sjos.70017

APPENDIX A. PROOF OF THEOREMS 1 AND 2

A1 A general result

In this section, we establish a general result, from which Theorems 1 and 2 can be deduced. It provides nonasymptotic bounds on the mFSR and the FSR of the plug-in procedure and on its average selection number, by relying only on Assumption 1. To state the result, we introduce some additional quantities measuring the regularity of the model which will appear in our remainder terms. Recall definitions (4), (5), and (13) of $\ell_q(X,\theta)$, $T(X,\theta)$, and $t^*(\alpha)$ respectively, and let for $\epsilon,\delta,\nu>0$,

$$\mathcal{W}_\ell(\epsilon)=\sup\left\{\sum_{q=1}^{Q}\mathbb{E}_{\theta^*}\left[|\ell_q(X,\theta^*)-\ell_q(X,\theta)|\right],\ \|\theta-\theta^*\|_2\le\epsilon,\ \theta\in\Theta\right\};\tag{A1}$$

$$\mathcal{W}_T(\delta)=\sup\{|\mathbb{P}_{\theta^*}(T(X,\theta^*)<t')-\mathbb{P}_{\theta^*}(T(X,\theta^*)<t)|,\tag{A2}$$

$$t,t'\in[0,1],|t'-t|\le\delta\};\tag{A3}$$

$$\Psi(\epsilon)=\mathcal{W}_T(\mathcal{W}_\ell(\epsilon)^{1/2})+\mathcal{W}_\ell(\epsilon)^{1/2};\tag{A4}$$

$$\mathcal{W}_{t^*,\alpha}(v) = \sup\{|t^*(\alpha+\beta) - t^*(\alpha)|\,,\ |\beta| \le v\}. \tag{A5}$$

Theorem 3. *Let Assumption 1 be true. For any $\alpha \in (\alpha_c, \overline{\alpha})$ and constants $s^* = s^*(\alpha, \theta^*) \in (0,1)$ and $e^* = e(\alpha, \theta^*) > 0$ depending only on α and θ^*, the following holds. Consider the plug-in procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$ introduced in Algorithm 2 and based on an estimator $\hat{\theta}$ satisfying Assumption 5, with $\eta(\epsilon, \theta^*)$ defined by (18). Then for $\epsilon \le e^*$ and $n \ge (2e)^3$, letting*

$$\Delta_n(\epsilon) = 2\Big(\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\delta_n + 8\Psi(\epsilon)/s^*)) + 4\Psi(\epsilon) + 2\delta_n\Big),$$

for $\delta_n = C\sqrt{(\log n)/n}/s^*$ where $C = 2 + 56Q\sqrt{\mathscr{V}} + 28Q^2\sqrt{\mathscr{V}_-}$ and with the quantities $\mathcal{W}_T$, $\mathcal{W}_\ell$, Ψ, $\mathcal{W}_{t^*,\alpha}$ defined by (A3), (A1), (A4), (A5), respectively, it holds:

- *The procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}$ controls both the FSR and the mFSR at level close to α in the following sense*:

$$\mathrm{FSR}(\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}) \le \alpha + \Delta_n(\epsilon)/s^* + 5/n^2 + \eta(\epsilon, \theta^*);$$
$$\mathrm{mFSR}(\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}) \le \alpha + \Delta_n(\epsilon)/s^* + s^{*-1}\big[50/n^2 + 10\eta(\epsilon, \theta^*)\big].$$

- *The procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}$ is nearly optimal in the following sense: for any other procedure $C = (\widehat{\mathbf{Z}}, S)$ that controls the mFSR at level α,*

$$n^{-1}\mathbb{E}_{\theta^*}(|\widehat{S}_\alpha^{\mathrm{PI}}|) \ge n^{-1}\mathbb{E}_{\theta^*}(|S|) - \Delta_n(\epsilon).$$

Before proving this result (which will be done in the next subsections, see Sections A.5 and A.6), let us first show that Theorem 3 implies Theorems 1 and 2.

A2 Proof of Theorem 1

By Lemma 4 below, $\Delta_n(\epsilon)$ tends to 0 when n tends to infinity and ϵ tends to 0. Moreover, by consistency of $\hat{\theta}$, $\eta(\epsilon, \theta^*)$ tends to 0 for all $\epsilon > 0$. This implies the result.

Lemma 4. *Under Assumption 1, we have $\lim_{\delta\to 0}\mathcal{W}_T(\delta) = 0$, $\lim_{v\to 0}\mathcal{W}_{t^*,\alpha}(v) = 0$. Under Assumption 2, we have $\lim_{\epsilon\to 0}\mathcal{W}_\ell(\epsilon) = 0$. Under both assumptions, we have $\lim_{\epsilon\to 0}\Psi(\epsilon) = 0$.*

Lemma 4 is proved in Section B.4.

A3 Proof of Theorem 2

Using Assumption 3 (with the notation therein) and Lemma 5, we have

$$\begin{aligned}\Delta_n(\epsilon) &= 2\Big(\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\delta_n + 8\Psi(\epsilon)/s^*)) + 4\Psi(\epsilon) + 2\delta_n\Big)\\ &\le 2C_2C_3\Big(2\delta_n + (8/s^*)\sqrt{C_1}(C_2+1)\sqrt{\epsilon}\Big) + 8\sqrt{C_1}(C_2+1)\sqrt{\epsilon} + 4\delta_n\\ &= 8\sqrt{C_1}(C_2+1)(1+2C_2C_3)\sqrt{\epsilon}/s^* + 4(C_2C_3+1)C\sqrt{\log n/n}/s^*,\end{aligned}$$

because $s^* \le 1$ and by definition of δ_n. This gives (21) and (22) with $A = 8\sqrt{C_1}(C_2+1)(1+2C_2C_3)/s^{*2}$ and $B = 4(C_2C_3+1)C/s^{*2}$.

Lemma 5. *Under Assumption 3, we have* $\mathcal{W}_{\ell}(\epsilon) \leq C_1\epsilon$, $\mathcal{W}_T(\delta) \leq C_2\delta$, $\mathcal{W}_{t^*,\alpha}(v) \leq C_3 v$ *and* $\Psi(\epsilon) \leq \sqrt{C_1}(C_2+1)\sqrt{\epsilon}$ *for* ϵ, δ, v *small enough.*

Lemma 5 is proved in Section B.5

A4 An optimal procedure

We consider in this section the procedure that serves as an optimal procedure in our theory. For $t \in [0,1]$, let $\mathcal{C}_t^* = (\widehat{\mathbf{Z}}^*, S_t^*)$ be the procedure using the Bayes clustering $\widehat{\mathbf{Z}}^*$ (6) and the selection rule $S_t^* = \{i \in \{1,\dots,n\} \ : \ T_i^* < t\}$. Let us consider the map $t \in [0,1] \mapsto \mathrm{mFSR}(\mathcal{C}_t^*)$ and note that $\mathrm{mFSR}(\mathcal{C}_t^*) = \mathrm{mFSR}_t^*$ as defined by (12). Lemma 8 below provides the key properties for this function.

Definition 3. The optimal procedure at level α is defined by $\mathcal{C}^*_{t^*(\alpha)}$ where $t^*(\alpha)$ is defined by (13).

Under Assumption 1, Lemma 8 entails that, for $\alpha > \alpha_c$, $\mathrm{mFSR}(\mathcal{C}^*_{t^*(\alpha)}) \leq \alpha$. Hence, $\mathcal{C}^*_{t^*(\alpha)}$ controls the mFSR at level α. In addition, it is optimal in the following sense: any other mFSR controlling procedure should select less items than $\mathcal{C}^*_{t^*(\alpha)}$.

Lemma 6 (Optimality of $\mathcal{C}^*_{t^*(\alpha)}$). *Let Assumption 1 be true and choose* $\alpha \in (\alpha_c, \overline{\alpha}]$. *Then the oracle procedure* $\mathcal{C}^*_{t^*(\alpha)} = (\widehat{\mathbf{Z}}^*, S^*_{t^*(\alpha)})$ *satisfies the following*:

(*i*) $\mathrm{mFSR}(\mathcal{C}^*_{t^*(\alpha)}) = \alpha$;

(*ii*) *for any procedure* $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ *such that* $\mathrm{mFSR}(\mathcal{C}) \leq \alpha$, *we have* $\mathbb{E}_{\theta^*}(|S|) \leq \mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|)$.

Lemma 6 is proved in Section B.6.

A5 Preliminary steps for proving Theorem 3

To keep the main proof concise, we need to define several additional notation. Let for $t \in [0,1]$ and $\theta \in \Theta$ (recall 5)

$$\widehat{\mathbf{L}}_0(\theta, t) = \frac{1}{n}\sum_{i=1}^n T(X_i,\theta)\mathbb{1}_{T(X_i,\theta)<t}; \tag{A6}$$

$$\widehat{\mathbf{L}}_1(\theta, t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{T(X_i,\theta)<t}. \tag{A7}$$

Denote $\widehat{\mathbf{L}} = \widehat{\mathbf{L}}_0/\widehat{\mathbf{L}}_1$, $\mathbf{L}_0 = \mathbb{E}_{\theta^*}\widehat{\mathbf{L}}_0$, $\mathbf{L}_1 = \mathbb{E}_{\theta^*}\widehat{\mathbf{L}}_1$, $\mathbf{L} = \mathbf{L}_0/\mathbf{L}_1$ (with the convention $0/0 = 0$). Note that for any $\alpha > \alpha_c$, the mFSR of the optimal procedure $\mathcal{C}^*_{t^*(\alpha)}$ defined in Section A.4 is given by $\mathrm{mFSR}(\mathcal{C}^*_{t^*(\alpha)}) = \mathbf{L}(\theta^*, t^*(\alpha)) = \alpha$.

Also, we denote from now on $\ell^*_{i,q} = \mathbb{P}_{\theta^*}(Z_i = q|X_i)$ for short and introduce for any parameter $\theta \in \Theta$ (recall 4 and 5)

$$\overline{q}(X_i,\theta) \in \operatorname*{argmax}_{q\in\{1,..,Q\}} \ell_q(X_i,\theta), \quad 1 \leq i \leq n; \tag{A8}$$

$$U(X_i,\theta) = 1 - \ell^*_{i,\overline{q}(X_i,\theta)}, \quad 1 \leq i \leq n; \tag{A9}$$

$$\widehat{\mathbf{M}}_0(\theta,t) = \frac{1}{n}\sum_{i=1}^n U(X_i,\theta)\mathbb{1}_{T(X_i,\theta)<t}, \quad t \in [0,1], \tag{A10}$$

Note that $\widehat{\mathbf{M}}_0(\theta^*, t) = \widehat{\mathbf{L}}_0(\theta^*, t)$ but in general $\widehat{\mathbf{M}}_0(\theta, t)$ is different from $\widehat{\mathbf{L}}_0(\theta, t)$. We denote $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}_0/\widehat{\mathbf{L}}_1$, $\mathbf{M}_0 = \mathbb{E}_{\theta^*}\widehat{\mathbf{M}}_0$ and $\mathbf{M} = \mathbf{M}_0/\mathbf{L}_1$ (with the convention $0/0 = 0$).

When $\alpha \in (\alpha_c, \overline{\alpha}]$ (recall 14 and 15), we also let

$$s^* = s^*(\alpha, \theta^*) = n^{-1}\mathbb{E}_{\theta^*}\left(|S^*_{t^*(\frac{\alpha+\alpha_c}{2})}|\right) = \mathbf{L}_1(\theta^*, t^*((\alpha+\alpha_c)/2)) > 0. \tag{A11}$$

We easily see that the latter is positive: if it was zero then $S^*_{t^*((\alpha+\alpha_c)/2))}$ would be empty which would entails that $\mathrm{mFSR}(\mathcal{C}^*_{t^*((\alpha+\alpha_c)/2)})$ is zero. This is excluded by definition (14) of α_c because $(\alpha+\alpha_c)/2 > \alpha_c$.

Also, we are going to extensively use the event

$$\Omega_\epsilon = \left\{\min_{\sigma\in[Q]}\|\hat{\theta}^\sigma - \theta^*\|_2 < \epsilon\right\}.$$

On this event, we fix any permutation $\sigma \in [Q]$ (possibly depending on X) such that $\|\hat{\theta}^\sigma - \theta^*\|_2 < \epsilon$. Now using Lemma 9, the plug-in selection rule can be rewritten as $\widehat{S}^{\mathrm{PI}}_\alpha = \{i \in \{1,\dots,n\} \ : \ \hat{T}_i < \hat{t}(\alpha)\}$ (denoted by $\widehat{S}$ in the sequel for short), where

$$\hat{t}(\alpha) = \sup\{t \in [0,1] \ : \ \widehat{\mathbf{L}}(\hat{\theta}, t) \le \alpha\}. \tag{A12}$$

With the above notation, we can upper bound what is inside the brackets of $\mathrm{FSR}(\widehat{\mathcal{C}}^{\mathrm{PI}})$ and $\mathrm{mFSR}(\widehat{\mathcal{C}}^{\mathrm{PI}})$ as follows.

Lemma 7. *For the permutation σ in Ω_ϵ realizing $\|\hat{\theta}^\sigma - \theta^*\|_2 < \epsilon$, we have on the event Ω_ϵ the following relations*:

$$|\widehat{S}| = \widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, \hat{t}(\alpha));$$
$$\min_{\sigma'\in[Q]}\mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}}(\sigma'(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right) \le \widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, \hat{t}(\alpha));$$
$$\min_{\sigma'\in[Q]}\mathbb{E}_{\theta^*}\left(\frac{\varepsilon_{\widehat{S}}(\sigma'(\widehat{\mathbf{Z}}), \mathbf{Z})}{\max(|\widehat{S}|, 1)} \,\middle|\, \mathbf{X}\right) \le \widehat{\mathbf{M}}(\hat{\theta}^\sigma, \hat{t}(\alpha)).$$

Lemma 7 is proved in Section B.7.

Finally, we make use of the concentration of the empirical processes $\widehat{\mathbf{L}}_0(\theta, t)$, $\widehat{\mathbf{L}}_1(\theta, t)$, and $\widehat{\mathbf{M}}_0(\theta, t)$, uniformly with respect to $\theta \in \mathcal{D}$ (where $\mathcal{D}$ is defined in Assumption 5). Thus, we define the following events, for $\delta > 0$ (recall s^* defined by A11):

$$\Gamma_{0,\delta,t} = \left\{\sup_{\theta\in\mathcal{D}}\left|\widehat{\mathbf{L}}_0(\theta, t) - \mathbf{L}_0(\theta, t)\right| \le \delta\right\};$$
$$\Gamma_{1,\delta,t} = \left\{\sup_{\theta\in\mathcal{D}}\left|\widehat{\mathbf{L}}_1(\theta, t) - \mathbf{L}_1(\theta, t)\right| \le \delta\right\};$$
$$\Gamma_{\delta,t} = \left\{\sup_{\theta\mathcal{D}, \mathbf{L}_1(\theta,t)\ge s^*}\left|\widehat{\mathbf{L}}(\theta, t) - \mathbf{L}(\theta, t)\right| \le \delta\right\};$$
$$\Upsilon_{0,\delta,t} = \left\{\sup_{\theta\in\mathcal{D}}\left|\widehat{\mathbf{M}}_0(\theta, t) - \mathbf{M}_0(\theta, t)\right| \le \delta\right\}.$$

Note that the following holds:

$$\Gamma_{0,\delta s^*/2,t} \cap \Gamma_{1,\delta s^*/2,t} \subset \Gamma_{\delta,t}. \tag{A13}$$

Indeed, on the event $\Gamma_{0,\delta s^*/2,t} \cap \Gamma_{1,\delta s^*/2,t}$, provided that $\mathbf{L}_1(\theta,t) \geq s^*$, we have

$$\begin{aligned}
&\left|\frac{\widehat{\mathbf{L}}_0(\theta,t)}{\widehat{\mathbf{L}}_1(\theta,t)} - \frac{\mathbf{L}_0(\theta,t)}{\mathbf{L}_1(\theta,t)}\right| \\
&\quad\leq \left|\frac{\mathbf{L}_0(\theta,t) - \widehat{\mathbf{L}}_0(\theta,t)}{\mathbf{L}_1(\theta,t)}\right| + \widehat{\mathbf{L}}_0(\theta,t)\left|\frac{1}{\widehat{\mathbf{L}}_1(\theta,t)} - \frac{1}{\mathbf{L}_1(\theta,t)}\right| \\
&\quad\leq (\delta s^*/2)/s^* + (\delta s^*/2)/s^* = \delta,
\end{aligned}$$

because $\widehat{\mathbf{L}}_0(\theta,t) \leq \widehat{\mathbf{L}}_1(\theta,t)$. This proves the desired inclusion.

A6 Proof of Theorem 3

Let us now provide a proof for Theorem 3.

A6.1 Step 1: Bounding $\hat{t}(\alpha)$ w.r.t. $t^*(\alpha)$

Recall (13), (A12) and (A11). In this part, we only consider realizations on the event Ω_ϵ. Let $\beta \in [\frac{2\alpha+\alpha_c}{3}, \frac{\alpha+\overline{\alpha}}{2}]$. By Lemma 10, we have

$$\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\beta)) \geq \mathbf{L}_1(\theta^*, t^*(\beta)) - \Psi(\|\hat{\theta}^\sigma - \theta^*\|_2) \geq \mathbf{L}_1(\theta^*, t^*((2\alpha+\alpha_c)/3)) - \Psi(\epsilon),$$

because $t^*(\beta) \geq t^*(\frac{2\alpha+\alpha_c}{3})$ since $t^*(\cdot)$ is nondecreasing by Lemma 8. Hence $\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\beta)) \geq s^*$ for ϵ smaller than a threshold only depending on θ^* and α. Hence, we have on $\Gamma_{\delta,t^*(\beta)}$ that

$$\mathbf{L}(\hat{\theta}^\sigma, t^*(\beta)) - \delta \leq \widehat{\mathbf{L}}(\hat{\theta}^\sigma, t^*(\beta)) \leq \delta + \mathbf{L}(\hat{\theta}^\sigma, t^*(\beta)).$$

By using again Lemma 10, we have

$$\mathbf{L}(\theta^*, t^*(\beta)) - 3\Psi(\epsilon)/s^* \leq \mathbf{L}(\hat{\theta}^\sigma, t^*(\beta)) \leq \mathbf{L}(\theta^*, t^*(\beta)) + 3\Psi(\epsilon)/s^*.$$

Given that $\mathbf{L}(\theta^*, t^*(\beta)) = \mathrm{mFSR}(C^*_{t^*(\beta)}) = \beta$ (see Lemma 6 (i)), it follows that for $\gamma = \gamma(\epsilon,\delta) = \delta + 4\Psi(\epsilon)/s^*$, on the event $\Gamma_{\delta,t^*(\alpha-\gamma)} \cap \Gamma_{\delta,t^*(\alpha+\gamma)}$,

$$\widehat{\mathbf{L}}(\hat{\theta}^\sigma, t^*(\alpha-\gamma)) \leq \alpha, \quad \widehat{\mathbf{L}}(\hat{\theta}^\sigma, t^*(\alpha+\gamma)) > \alpha,$$

where we indeed check that $\alpha - \gamma \geq \frac{2\alpha+\alpha_c}{3}$ and $\alpha + \gamma \leq \frac{\alpha+\overline{\alpha}}{2}$ for δ and ϵ smaller than some threshold only depending on θ^* and α. In a nutshell, we have established

$$\Gamma_{\delta,t^*(\alpha-\gamma)} \cap \Gamma_{\delta,t^*(\alpha+\gamma)} \cap \Omega_\epsilon \subset \left\{t^*(\alpha-\gamma) \leq \hat{t}(\alpha) \leq t^*(\alpha+\gamma)\right\}. \tag{A14}$$

A6.2 Step 2: Upper-bounding the FSR

Let us consider the event

$$\Lambda_{\alpha,\delta,\epsilon} := \Gamma_{0,\delta s^*/2,t^*(\alpha-\gamma)} \cap \Gamma_{1,\delta s^*/2,t^*(\alpha-\gamma)} \cap \Gamma_{0,\delta s^*/2,t^*(\alpha+\gamma)} \cap \Gamma_{1,\delta s^*/2,t^*(\alpha+\gamma)} \cap \Upsilon_{0,\delta,t^*(\alpha+\gamma)} \cap \Omega_\epsilon,$$

where the different events have been defined in the previous section.

Let us prove (21). Using Lemma 7 and (A14),

$$\begin{aligned}\mathrm{FSR}(\hat{\mathcal{C}}) &\le \mathbb{E}_{\theta^*}[\widehat{\mathbf{M}}(\hat{\theta}^\sigma, \hat{t}(\alpha))\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\le \mathbb{E}_{\theta^*}\left[\frac{\widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma))}{\widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma))}\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}\right] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c).\end{aligned}$$

Now using a concentration argument on the event $\Lambda_{\alpha,\delta,\epsilon} \subset \Gamma_{1,\delta,t^*(\alpha-\gamma)} \cap \Upsilon_{0,\delta,t^*(\alpha+\gamma)}$, we have

$$\begin{aligned}\mathrm{FSR}(\hat{\mathcal{C}}) &\le \mathbb{E}_{\theta^*}\left[\frac{\mathbf{M}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma)) + \delta}{\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma)) - \delta}\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}\right] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\le \frac{\mathbf{M}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &= \frac{\mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c), \qquad \text{(A15)}\end{aligned}$$

by using Lemma 10 and that $\mathbf{M}_0(\theta^*, t) = \mathbf{L}_0(\theta^*, t)$ for all t by definition. Now, using again Lemma 10, we have

$$\begin{aligned}\mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) &\le \mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) + \mathcal{W}_T(t^*(\alpha+\gamma) - t^*(\alpha-\gamma)) \\ &\le \mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) + \mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\gamma))\end{aligned}$$

This entails

$$\begin{aligned}\mathrm{FSR}(\hat{\mathcal{C}}) &\le \frac{\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) + \mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\gamma)) + 3\Psi(\epsilon) + \delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\le \frac{\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma))}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} \\ &\quad + (s^*/2)^{-1}\left(\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\gamma)) + 3\Psi(\epsilon) + \delta\right) + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),\end{aligned}$$

by choosing ϵ, δ smaller than a threshold (only depending on θ^* and α) so that $\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta \ge s^*/2$. Now using $\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) = (\alpha-\gamma)\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma))$, we have

$$\begin{aligned}\frac{\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma))}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} &= (\alpha-\gamma)\left(1 + \frac{\Psi(\epsilon)+\delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta}\right) \\ &\le \alpha\left(1 + (s^*/2)^{-1}(\Psi(\epsilon)+\delta)\right).\end{aligned}$$

This leads to

$$\mathrm{FSR}(\hat{\mathcal{C}}) \leq \alpha + (2/s^*)\left[\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\delta + 8\Psi(\epsilon)/s^*)) + 4\Psi(\epsilon) + 2\delta\right] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),$$

which holds true for δ, ϵ smaller than a threshold only depending on θ^* and α.

A6.3 Step 3: Upper-bounding the mFSR

We apply a similar technique as for step 2. Using Lemma 7 and (A14),

$$\begin{aligned}\mathrm{mFSR}(\hat{\mathcal{C}}) &\leq \frac{\mathbb{E}_{\theta^*}[\widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, \hat{t}(\alpha))\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbb{E}_{\theta^*}[\widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, \hat{t}(\alpha))\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}]} \\ &\leq \frac{\mathbb{E}_{\theta^*}[\widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma))\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbb{E}_{\theta^*}[\widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma))\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}]}.\end{aligned}$$

Now using a concentration argument on $\Lambda_{\alpha,\delta,\epsilon} \subset \Gamma_{1,\delta,t^*(\alpha-\gamma)} \cap \Upsilon_{0,\delta,t^*(\alpha+\gamma)}$, we have

$$\begin{aligned}\mathrm{mFSR}(\hat{\mathcal{C}}) &\leq \frac{\mathbb{E}_{\theta^*}[(\mathbf{M}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma)) + \delta)\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbb{E}_{\theta^*}[(\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma)) - \delta)\mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}]} \\ &\leq \frac{\mathbf{M}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)} \\ &= \frac{\mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)},\end{aligned}$$

by using Lemma 10 and that $\mathbf{M}_0(\theta^*, t) = \mathbf{L}_0(\theta^*, t)$ by definition. Letting $x = \mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta$, $y = \mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta$ and $u = \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$, we have obtained the bound $(x+u)/(y-u)$, which has to be compared with the FSR bound (A15), which reads $x/y + u$. Now, when $y \in [0,1]$, $x \geq 0$, $x/y \leq 2$, $u/y \leq 1/2$, $y - u \geq s^*/2$, we have

$$(x+u)/(y-u) \leq \frac{x/y}{1-u/y} + (2/s^*)u \leq x/y(1+2u/y) + (2/s^*)u \leq x/y + (10/s^*)u.$$

As a result, for ϵ, δ small enough, and $\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \leq s^*/4$, we obtain the same bound as for the FSR, with $\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$ replaced by $(10/s^*)\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$.

A6.4 Step 4: Lower-bounding the selection rate

In Step 3, when bounding the mFSR, we derived a lower bound for the denominator of the mFSR, that is, $\mathbb{E}_{\theta^*}(|\hat{S}|)$. It reads

$$\begin{aligned}n^{-1}\mathbb{E}_{\theta^*}(|\hat{S}|) &\geq \mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\geq \mathbf{L}_1(\theta^*, t^*(\alpha)) - \mathcal{W}_T(t^*(\alpha) - t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\geq n^{-1}\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) - \mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),\end{aligned}$$

by using (A3) and (A5). Now consider another procedure $\mathcal{C} = (\hat{\mathbf{Z}}, S)$ that controls the mFSR at level α, that is, $\mathrm{mFSR}(\mathcal{C}) \leq \alpha$. By Lemma 6, we then have $\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) \geq \mathbb{E}_{\theta^*}(|S|)$.

A6.5 Step 5: Concentration

Finally, we bound $\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$ by using Lemma 11 with $x = (1+2c)\sqrt{\frac{\log n}{n}}$ (with c defined in Lemma 11). This gives for $\delta = 2x/s^*$, and $n \geq (2e)^3$

$$\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \leq 5/n^2 + \mathbb{P}(\Omega_\epsilon^c).$$

APPENDIX B. PROOFS OF LEMMAS

B1 Proof of Lemma 1

The clustering risk of $\widehat{\mathbf{Z}}$ is given by

$$\begin{aligned} R(\widehat{\mathbf{Z}}) &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(n^{-1}\sum_{i=1}^n \mathbb{1}\{Z_i \neq \sigma(\hat{Z}_i)\} \,\middle|\, \mathbf{X}\right)\right) \\ &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \sigma(\hat{Z}_i) \mid \mathbf{X})\right) \\ &\geq \mathbb{E}_{\theta^*}\left(\min_{\hat{\mathbf{Z}}} n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \hat{Z}_i \mid \mathbf{X})\right), \end{aligned}$$

where, by independence, the minimum in the lower bound is achieved for the Bayes clustering. Thus, $R(\widehat{\mathbf{Z}}) \geq n^{-1}\sum_{i=1}^n \mathbb{E}_{\theta^*}(T_i^*)$. Moreover, $n^{-1}\sum_{i=1}^n \mathbb{E}_{\theta^*}(T_i^*) \geq R(\widehat{\mathbf{Z}}^*)$, since

$$\begin{aligned} R(\widehat{\mathbf{Z}}^*) &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \sigma(\hat{Z}_i^*) \mid \mathbf{X})\right) \\ &\leq \mathbb{E}_{\theta^*}\left(n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \hat{Z}_i^* \mid \mathbf{X})\right). \end{aligned}$$

Thus, $\min_{\hat{\mathbf{Z}}} R(\widehat{\mathbf{Z}}) = R(\widehat{\mathbf{Z}}^*)$ and the proof is completed.

B2 Proof of Lemma 2

Following the reasoning of the proof of Lemma 1, we have

$$\begin{aligned} \mathrm{FSR}_{\theta^*}(\mathcal{C}) &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\frac{\sum_{i\in S}\mathbb{1}\{Z_i \neq \sigma(\hat{Z}_i^*)\}}{\max(|S|,1)} \,\middle|\, \mathbf{X}\right)\right) \\ &= \mathbb{E}_{\theta^*}\left(\frac{\sum_{i\in S} T_i^*}{\max(|S|,1)}\right). \end{aligned}$$

B3 Proof of Lemma 3

By definition, we have

$$\mathrm{FSR}(\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}) = \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\frac{\sum_{i=1}^n \mathbb{1}_{Z_i\neq\sigma(\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}))}\mathbb{1}\{i\in\hat{S}^{\mathrm{PI}}(\mathbf{X})\}}{\max(|\hat{S}^{\mathrm{PI}}(\mathbf{X})|,1)} \,\middle|\, \mathbf{X}\right)\right),$$

so that (9) follows by a direct integration w.r.t. the latent variable Z.

B4 Proof of Lemma 4

The only nontrivial fact is for $\mathcal{W}_{t^*,\alpha}(v)$. Assumption 1 and Lemma 8 provide that $t \mapsto \mathrm{mFSR}_t^*$ is a one-to-one continuous increasing map from $(t^*(\alpha_c), t^*(\overline{\alpha}))$ to $(\alpha_c, \overline{\alpha})$. Hence, for $\alpha \in (\alpha_c, \overline{\alpha})$, $\beta \mapsto t^*(\alpha+\beta)$ is continuous in 0 and $\lim_{v\to 0} \mathcal{W}_{t^*,\alpha}(v) = 0$.

B5 Proof of Lemma 5

For all $\theta \in \Theta$ such that $\|\theta - \theta^*\|_2 \le \epsilon$, we have by the mean value theorem

$$\sum_{q=1}^{Q} \mathbb{E}_{\theta^*}\left[|\ell_q(X,\theta^*) - \ell_q(X,\theta)|\right] \le \sum_{q=1}^{Q} \mathbb{E}_{\theta^*} \sup_{\theta:\|\theta-\theta^*\|_2\le\epsilon} \|\nabla_\theta \ell_q(X,\theta)\|_2 \|\theta-\theta^*\|_2 \le C_1\epsilon,$$

by using Assumption 3 (i). This implies $\mathcal{W}_\ell(\epsilon) \le C_1\epsilon$. The fact $\mathcal{W}_T(\delta) \le C_2\delta$, $\mathcal{W}_{t^*,\alpha}(v) \le C_3 v$ are obvious by using Assumption 3 (ii) and (iii). The fact $\Psi(\epsilon) \le \sqrt{C_1}(C_2+1)\sqrt{\epsilon}$ easily follows.

B6 Proof of Lemma 6

By Lemma 8, we have that $\mathrm{mFSR}(C_t^*)$ is monotonous in t and continuous w.r.t. t on $(t^*(\alpha_c), 1]$, thus for $\alpha \in (\alpha_c, \overline{\alpha}]$, $\mathrm{mFSR}(C^*_{t^*(\alpha)}) = \alpha$ which gives (i). For (ii), let $C = (\widehat{\mathbf{Z}}, S)$ be a procedure such that $\mathrm{mFSR}(C) \le \alpha$. Let us consider the procedure C' with the Bayes clustering $\widehat{\mathbf{Z}}^*$ and the same selection rule S. Since C' is based on a Bayes clustering, by the same reasoning leading to $R(\widehat{\mathbf{Z}}^*) \le R(\widehat{\mathbf{Z}})$ in Section 3.1, we have that $\mathrm{mFSR}(C') \le \mathrm{mFSR}(C) \le \alpha$ with

$$\mathrm{mFSR}(C') = \frac{\mathbb{E}_{\theta^*}\left(\sum_{i\in S} T_i^*\right)}{\mathbb{E}_{\theta^*}(|S|)}.$$

Hence,

$$\mathbb{E}_{\theta^*}\left(\sum_{i\in S} T_i^*\right) \le \alpha \mathbb{E}_{\theta^*}(|S|). \tag{B1}$$

Now we use an argument similar to the proof of Theorem 1 in Cai et al. (2019). By definition of $S^*_{t^*(\alpha)}$, we have that

$$\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^* - t^*(\alpha)) \le 0$$

which we can rewrite as

$$\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^* - t^*(\alpha) + \alpha - \alpha) \le 0$$

and so

$$\mathbb{E}_{\theta^*}\left(\sum_{i=1}^{n}\left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}(\mathbf{X})}-\mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^*-\alpha)\right)$$

$$\leq (t^*(\alpha)-\alpha)\mathbb{E}_{\theta^*}\left(\sum_{i=1}^{n}\left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}(\mathbf{X})}-\mathbb{1}_{i\in S(\mathbf{X})}\right)\right)$$

$$= (t^*(\alpha)-\alpha)(\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|)-\mathbb{E}_{\theta^*}(|S|)).$$

On the other hand, $\mathrm{mFSR}(\mathcal{C}^*_{t^*(\alpha)})=\alpha$ together with (B1) implies that

$$\mathbb{E}_{\theta^*}\left(\sum_{i=1}^{n}\left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}(\mathbf{X})}-\mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^*-\alpha)\right)$$

$$= \mathbb{E}_{\theta^*}\left(\sum_{i\in S^*_{t^*(\alpha)}} T_i^*-\alpha|S^*_{t^*(\alpha)}|-\sum_{i\in S}T_i^*+\alpha|S|\right)\geq 0.$$

Combining, the relations above provides

$$(t^*(\alpha)-\alpha)(\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|)-\mathbb{E}_{\theta^*}(|S|))\geq 0.$$

Finally, noting that $t^*(\alpha)-\alpha>0$ since $\alpha=\mathrm{mFSR}(\mathcal{C}^*_{t^*(\alpha)})<t^*(\alpha)$ by (ii) Lemma 8, this gives $\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|)-\mathbb{E}_{\theta^*}(|S|)\geq 0$ and concludes the proof.

B7 Proof of Lemma 7

First, we have by definition $\ell_q(X_i,\theta^\sigma)=\ell_{\sigma(q)}(X_i,\theta)$ and thus $T(X_i,\hat\theta)=T(X_i,\hat\theta^\sigma)$ by taking the maximum over q. This gives $\widehat{S}^\sigma=\widehat{S}$ and yields the first equality. Next, we have on Ω_ϵ,

$$\min_{\sigma'\in[Q]}\mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}}(\sigma'(\widehat{\mathbf{Z}}),\mathbf{Z})\,\middle|\,\mathbf{X}\right)\leq \mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}}(\sigma(\widehat{\mathbf{Z}}),\mathbf{Z})\,\middle|\,\mathbf{X}\right)$$
$$\leq \mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}^\sigma}(\sigma(\widehat{\mathbf{Z}}),\mathbf{Z})\,\middle|\,\mathbf{X}\right),$$

still because $\widehat{S}^\sigma=\widehat{S}$. Now observe that,

$$\mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}^\sigma}(\sigma(\widehat{\mathbf{Z}}),\mathbf{Z})\,\middle|\,\mathbf{X}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta^*}(Z_i\neq\sigma(\overline{q}(X_i,\hat\theta))\,|\,\mathbf{X})\mathbb{1}_{T(X_i,\hat\theta^\sigma)<\hat t(\alpha)}$$

$$=\frac{1}{n}\sum_{i=1}^{n}(1-\ell^*_{i,\sigma(\overline{q}(X_i,\hat\theta))})\mathbb{1}_{T(X_i,\hat\theta^\sigma)<\hat t(\alpha)}$$

$$=\widehat{\mathbf{M}}_0(\hat\theta^\sigma,\hat t(\alpha)),$$

because $\sigma(\overline{q}(X_i,\hat\theta))=\overline{q}(X_i,\hat\theta^\sigma)$. This proves the result.

APPENDIX C. AUXILIARY RESULTS

Lemma 8. *Let us consider the procedure $\mathcal{C}_t^*$ defined in Section A.4, the* mFSR *criterion defined by (3) and the functional* mFSR_t^* *defined by (12). Then we have*

$$\mathrm{mFSR}_t^* = \mathrm{mFSR}_{\theta^*}(\mathcal{C}_t^*) = \frac{\mathbb{E}_{\theta^*}\left(\sum_{i=1}^n T_i^* \mathbb{1}_{T_i^* < t}\right)}{\mathbb{E}_{\theta^*}\left(\sum_{i=1}^n \mathbb{1}_{T_i^* < t}\right)}, \quad t \in [0,1]. \tag{C1}$$

Moreover, the following properties for the function $t \in [0,1] \mapsto \mathrm{mFSR}_{\theta^*}(\mathcal{C}_t^*)$:

(*i*) $\mathrm{mFSR}_{\theta^*}(\mathcal{C}_t^*)$ *is nondecreasing in* $t \in [0,1]$ *and, under Assumption 1, it is increasing in* $t \in (t^*(\alpha_c), t^*(\overline{\alpha}))$;
(*ii*) $\mathrm{mFSR}_{\theta^*}(\mathcal{C}_t^*) < t$ *for* $t \in (0,1]$;
(*iii*) *Under Assumption 1,* $\mathrm{mFSR}_{\theta^*}(\mathcal{C}_t^*)$ *is continuous w.r.t. t on* $(t^*(\alpha_c), 1]$, *where* $t^*(\alpha_c)$ *is given by (14).*

Proof. Write $\mathrm{mFSR}_{\theta^*}(\cdot)$ instead of $\mathrm{mFSR}(\cdot)$ for short. First, (C1) is obtained similarly than (8). For proving (i), let $t_1, t_2 \in [0,1]$ such that $t_1 < t_2$. We show that $\mathrm{mFSR}(\mathcal{C}_{t_1}^*) \leq \mathrm{mFSR}(\mathcal{C}_{t_2}^*)$. Remember here the convention $0/0 = 0$ and that $\mathrm{mFSR}(\mathcal{C}_t^*) = \mathbb{E}_{\theta^*}(T(X,\theta^*) \mid T(X,\theta^*) < t)$. First, if $\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1) = 0$ then the result is immediate. Otherwise, we have that

$$\begin{aligned}
&\mathrm{mFSR}(\mathcal{C}_{t_1}^*) - \mathrm{mFSR}(\mathcal{C}_{t_2}^*) \\
&\quad = (\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1))^{-1} \\
&\quad \cdot \mathbb{E}_{\theta^*}\left(T(X,\theta^*)\left\{ \mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1)}{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_2)} \mathbb{1}_{T(X,\theta^*)<t_2} \right\} \right),
\end{aligned}$$

where, given that $t_1 < t_2$, the quantity in the brackets is positive when $T(X,\theta^*) < t_1$ and is negative or zero otherwise. Hence,

$$\begin{aligned}
&T(X,\theta^*)\left\{ \mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1)}{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_2)} \mathbb{1}_{T(X,\theta^*)<t_2} \right\} \\
&\quad \leq t_1 \left\{ \mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1)}{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_2)} \mathbb{1}_{T(X,\theta^*)<t_2} \right\}.
\end{aligned}$$

Taking the expectation makes the right-hand-side equal to zero, from which the result follows. Now, to show the increasingness, if $\mathrm{mFSR}(\mathcal{C}_{t_1}^*) = \mathrm{mFSR}(\mathcal{C}_{t_2}^*)$ for $t^*(\alpha_c) < t_1 < t_2 < t^*(\overline{\alpha})$, then the above reasoning shows that

$$(T(X,\theta^*) - t_1)\left\{ \mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1)}{\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_2)} \mathbb{1}_{T(X,\theta^*)<t_2} \right\} \leq 0$$

and has an expectation equal to 0. Hence, given that $T(X,\theta^*)$ is continuous, we derive that almost surely

$$\mathbb{P}_{\theta^*}(T(X,\theta^*) < t_2)\mathbb{1}_{T(X,\theta^*)<t_1} = \mathbb{P}_{\theta^*}(T(X,\theta^*) < t_1)\mathbb{1}_{T(X,\theta^*)<t_2},$$

that is, $\mathbb{P}_{\theta^*}(t_1 \leq T_i^* < t_2) = 0$, which is excluded by Assumption 1. This entails $\mathrm{mFSR}(\mathcal{C}_{t_1}^*) < \mathrm{mFSR}(\mathcal{C}_{t_2}^*)$.

For proving (ii), let $t > 0$. If $\mathbb{P}_{\theta^*}(T(X,\theta^*) < t) = 0$ then the result is immediate. Otherwise, we have that $\mathrm{mFSR}(\mathcal{C}_t^*) - t = (\mathbb{P}_{\theta^*}(T(X,\theta^*) < t))^{-1}\mathbb{E}_{\theta^*}((T(X,\theta^*) - t)\mathbb{1}\{T(X,\theta^*) < t\})$. The latter is clearly not positive, and is moreover negative because $(T(X,\theta^*) - t)\mathbb{1}\{T(X,\theta^*) < t\} \leq 0$ and $\mathbb{P}_{\theta^*}(T(X,\theta^*) = t) = 0$ by Assumption 1.

For proving (iii), let $\psi_0(t) = \mathbb{E}_{\theta^*}(T(X,\theta^*)\mathbb{1}\{T(X,\theta^*) < t\})$ and $\psi_1(t) = \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)$, the numerator and denominator of $\mathrm{mFSR}(\mathcal{C}_t^*) = \mathrm{mFSR}_t^*$, respectively. $\psi_1(t)$ is nondecreasing in t, with $\psi_1(0) = 0$ and $\psi_1(1) > 0$. Moreover, ψ_0 and ψ_1 are both continuous under Assumption 1. Then denote by t_c the largest t s.t. $\psi_1(t) = 0$. ψ_1 is zero on $[0, t_c]$ then strictly positive and nondecreasing on $(t_c, 1]$, and we have that $t_c = t^*(\alpha_c)$. Hence, $\mathrm{mFSR}(\mathcal{C}_t^*)$ is zero on $[0, t_c]$ then strictly positive and continuous on $(t_c, 1]$. ∎

Remark 4. With the notation of the above proof, $t \mapsto \mathrm{mFSR}(\mathcal{C}_t^*)$ may have a discontinuity point at t_c since for $t_n \xrightarrow[t_n > t_c]{} t_c$, as $\psi_1(t_n) \to 0$, one does not necessarily have that $\mathrm{mFSR}(\mathcal{C}_t^*) \to 0$.

Lemma 9 (Expression of plug-in procedure as a thresholding rule). *For any $\alpha \in (0,1)$, let us consider the plug-in procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$ defined by Algorithm 2 and denote $K = |\widehat{S}_\alpha^{\mathrm{PI}}|$ the maximum of the $k \in \{0,\dots,n\}$ such that $\max(k,1)^{-1}\sum_{j=1}^{k}\hat{T}_{(j)} \leq \alpha$ for $\hat{T}_i = 1 - \max_q \ell_q(X_i, \hat{\theta})$, $1 \leq i \leq n$. Consider also $\hat{t}(\alpha)$ defined by (A12). Let Assumption 1 be true and consider an estimator $\hat{\theta}$ satisfying Assumption 5. Then it holds that $\hat{t}(\alpha) = \hat{T}_{(K+1)}$ and*

$$\widehat{S}_\alpha^{\mathrm{PI}} = \{i \in \{1,\dots,n\} \ : \ \hat{T}_i < \hat{t}(\alpha)\}.$$

Proof. If $\hat{T}_{(K)} < \hat{T}_{(K+1)}$ then the result is immediate. Thus it suffices to show that $\hat{T}_{(K)} = \hat{T}_{(K+1)}$ occurs with probability 0. From Assumption 5 (with the countable set D defined therein), we have

$$\mathbb{P}_{\theta^*}(\hat{T}_{(K)} = \hat{T}_{(K+1)}) \leq \mathbb{P}_{\theta^*}\left(\bigcup_{i \neq j}\{\hat{T}_i = \hat{T}_j\}\right) \leq \mathbb{P}_{\theta^*}\left(\bigcup_{\theta \in D}\bigcup_{i \neq j}\{T(X_i,\theta) = T(X_j,\theta)\}\right).$$

Now, the right term is a countable union of events which are all of null probability under Assumption 1. The result follows. ∎

Lemma 10. *We have for all $\theta \in \Theta$,*

$$\sup_{t \in [0,1]} |\mathbf{L}_1(\theta,t) - \mathbf{L}_1(\theta^*,t)| \leq \Psi(\|\theta^* - \theta\|); \tag{C2}$$

$$\sup_{t \in [0,1]} |\mathbf{L}_0(\theta,t) - \mathbf{L}_0(\theta^*,t)| \leq 2\Psi(\|\theta^* - \theta\|); \tag{C3}$$

$$\sup_{t \in [t^*((\alpha+\alpha_c)/2),1]} |\mathbf{L}(\theta,t) - \mathbf{L}(\theta^*,t)| \leq 3\Psi(\|\theta^* - \theta\|)/s^*; \tag{C4}$$

$$\sup_{t \in [0,1]} |\mathbf{M}_0(\theta,t) - \mathbf{M}_0(\theta^*,t)| \leq 3\Psi(\|\theta^* - \theta\|); \tag{C5}$$

$$\sup_{t \in [t^*((\alpha+\alpha_c)/2),1]} |\mathbf{M}(\theta,t) - \mathbf{M}(\theta^*,t)| \leq 4\Psi(\|\theta^* - \theta\|)/s^*; \tag{C6}$$

where $\alpha \in (\alpha_c, \overline{\alpha}]$ *and* $s^* > 0$ *is given by (A11). In addition, for all* $\theta \in \Theta$ *and* $t, t' \in [0,1]$,

$$|\mathbf{L}_0(\theta, t) - \mathbf{L}_0(\theta, t')| \leq 4\Psi(\|\theta^* - \theta\|) + \mathcal{W}_T(|t - t'|). \tag{C7}$$

Proof. Fix $\theta \in \Theta$ and $t \in [0,1]$. We have for any $\delta > 0$,

$$\begin{aligned}
&|\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \\
&\quad \leq (\mathbb{P}_{\theta^*}(T(X,\theta^*) < t + \delta) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)) \vee (\mathbb{P}_{\theta^*}(T(X,\theta^*) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t - \delta)) \\
&\qquad + \mathbb{P}_{\theta^*}(|T(X,\theta^*) - T(X,\theta)| > \delta) \\
&\quad \leq \mathcal{W}_T(\delta) + \mathbb{E}_{\theta^*}(|T(X,\theta^*) - T(X,\theta)|)/\delta.
\end{aligned}$$

In addition, by definition (5),

$$\begin{aligned}
|T(X,\theta^*) - T(X,\theta)| &\leq |\max_{1\leq q\leq Q} \ell_q(X,\theta^*) - \max_{1\leq q\leq Q} \ell_q(X,\theta)| \\
&\leq \max_{1\leq q\leq Q} |\ell_q(X,\theta^*) - \ell_q(X,\theta)|.
\end{aligned}$$

This implies by definition of $\mathcal{W}_\ell(\cdot)$ (see A1) that

$$\mathbb{E}_{\theta^*}(|T(X,\theta^*) - T(X,\theta)|) \leq \mathcal{W}_\ell(\|\theta - \theta^*\|).$$

Hence,

$$|\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \leq \inf_{\delta\in(0,1)} \{\mathcal{W}_T(\delta) + \mathcal{W}_\ell(\|\theta^* - \theta\|)/\delta\} \leq \Psi(\|\theta^* - \theta\|),$$

which establishes (C2).

Next, we have

$$\begin{aligned}
&\mathbf{L}_0(\theta, t) - \mathbf{L}_0(\theta^*, t) \\
&\quad = \mathbb{E}_{\theta^*}[T(X,\theta)(\mathbb{1}_{T(X,\theta)<t} - \mathbb{1}_{T(X,\theta^*)<t}) + \mathbb{1}_{T(X,\theta^*)<t}(T(X,\theta) - T(X,\theta^*))] \\
&\quad \leq t|\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \\
&\qquad + |\mathbb{E}_{\theta^*}[\mathbb{1}_{T(X,\theta^*)<t}(T(X,\theta) - T(X,\theta^*))]| \\
&\quad \leq |\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| + \mathbb{E}_{\theta^*}|T(X,\theta) - T(X,\theta^*)| \\
&\quad \leq 2\Psi(\|\theta^* - \theta\|)
\end{aligned}$$

By exchanging the role of θ and θ^* in the above reasoning, the same bound holds for $\mathbf{L}_0(\theta^*, t) - \mathbf{L}_0(\theta, t)$, which gives (C3). To prove (C4), we use for any $t \in [t^*(\frac{\alpha+\alpha_c}{2}), 1]$,

$$\begin{aligned}
&\left|\frac{\mathbf{L}_0(\theta,t)}{\mathbf{L}_1(\theta,t)} - \frac{\mathbf{L}_0(\theta^*,t)}{\mathbf{L}_1(\theta^*,t)}\right| \\
&\quad \leq \left|\frac{\mathbf{L}_0(\theta,t) - \mathbf{L}_0(\theta^*,t)}{\mathbf{L}_1(\theta^*,t)}\right| + \mathbf{L}_0(\theta,t)\left|\frac{1}{\mathbf{L}_1(\theta^*,t)} - \frac{1}{\mathbf{L}_1(\theta,t)}\right|
\end{aligned}$$

$$\leq 2\Psi(\|\theta^* - \theta\|)/s^* + \frac{1}{\mathbf{L}_1(\theta^*, t)} \frac{\mathbf{L}_0(\theta, t)}{\mathbf{L}_1(\theta, t)} |\mathbb{P}_{\theta^*}(T(X, \theta^*) < t) - \mathbb{P}_{\theta^*}(T(X, \theta) < t)|$$
$$\leq 3\Psi(\|\theta^* - \theta\|)/s^*,$$

because $\mathbf{L}_0(\theta, t) \leq \mathbf{L}_1(\theta, t)$ and $\mathbf{L}_1(\theta^*, t) \geq s^*$ by monotonicity. Similarly to the bound on $\mathbf{L}_0$, we derive

$$|\mathbf{M}_0(\theta, t) - \mathbf{M}_0(\theta^*, t)|$$
$$\leq |\mathbb{P}_{\theta^*}(T(X, \theta) < t) - \mathbb{P}_{\theta^*}(T(X, \theta^*) < t)| + \mathbb{E}_{\theta^*}|U(X, \theta) - U(X, \theta^*)|.$$

Define $\overline{q}(X, \theta) \in \operatorname{argmax}_{q \in \{1, \ldots, Q\}} \ell_q(X, \theta)$. Now, since $U(X, \theta^*) \leq U(X, \theta)$ by definition (A9), we have

$$\begin{aligned}
\mathbb{E}_{\theta^*}|U(X, \theta) - U(X, \theta^*)| &= \mathbb{E}_{\theta^*}[U(X, \theta) - U(X, \theta^*)] \\
&= \mathbb{E}_{\theta^*}[\ell_{\overline{q}(X,\theta)}(X, \theta^*) - \ell_{\overline{q}(X,\theta^*)}(X, \theta^*)] \\
&= \mathbb{E}_{\theta^*}[\ell_{\overline{q}(X,\theta)}(X, \theta^*) - \ell_{\overline{q}(X,\theta)}(X, \theta) \\
&\quad + \ell_{\overline{q}(X,\theta)}(X, \theta) - \ell_{\overline{q}(X,\theta^*)}(X, \theta^*)] \\
&\leq \mathbb{E}_{\theta^*}[\max_{1 \leq q \leq Q} |\ell_q(X, \theta^*) - \ell_q(X, \theta)|] \\
&\quad + \mathbb{E}_{\theta^*}[\max_{1 \leq q \leq Q} \ell_q(X, \theta) - \max_{1 \leq q \leq Q} \ell_q(X, \theta^*)] \\
&\leq 2\mathbb{E}_{\theta^*}[\max_{1 \leq q \leq Q} |\ell_q(X, \theta^*) - \ell_q(X, \theta)|] \leq 2\Psi(\|\theta^* - \theta\|).
\end{aligned}$$

This proves (C5) and leads to (C6) by following the reasoning that provided (C4).

Next, we have for $0 \leq t' \leq t \leq 1$, by (C3),

$$|\mathbf{L}_0(\theta, t) - \mathbf{L}_0(\theta, t')| \leq |\mathbf{L}_0(\theta^*, t) - \mathbf{L}_0(\theta^*, t')| + 4\Psi(\|\theta^* - \theta\|).$$

Moreover,

$$\begin{aligned}
|\mathbf{L}_0(\theta^*, t) - \mathbf{L}_0(\theta^*, t')| &= \mathbf{L}_0(\theta^*, t) - \mathbf{L}_0(\theta^*, t') = \mathbb{E}_{\theta^*}[T(X, \theta^*) \mathbb{1}_{t' \leq T(X,\theta^*) < t}] \\
&\leq \mathbb{E}_{\theta^*}[\mathbb{1}_{t' \leq T(X,\theta^*) < t}] \\
&= \mathbb{P}_{\theta^*}(T(X, \theta^*) < t) - \mathbb{P}_{\theta^*}(T(X, \theta^*) < t'),
\end{aligned}$$

which is below $\mathcal{W}_T(t - t')$ by (A3). This leads to (C7). ∎

Lemma 11 (Concentration of $\widehat{\mathbf{L}}_0$ (A6), $\widehat{\mathbf{L}}_1$ (A7), and $\widehat{\mathbf{M}}_0$ (A10)). *Let Assumption 1 be true. Recall $\mathscr{V}$, $\mathscr{V}_-$ defined by (16), (17) respectively, set $c := 14Q\sqrt{\mathscr{V}} + 7Q^2\sqrt{\mathscr{V}_-}$ and consider any countable set $\mathcal{D} \subset \Theta$. For all $t \in (0, 1]$ and for $n \geq (2e)^3$, we have*

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta \in \mathcal{D}} \left|\widehat{\mathbf{L}}_0(\theta, t) - \mathbf{L}_0(\theta, t)\right| > x\right) \leq n^{-2}; \tag{C8}$$

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta \in \mathcal{D}} \left|\widehat{\mathbf{L}}_1(\theta, t) - \mathbf{L}_1(\theta, t)\right| > x\right) \leq n^{-2}; \tag{C9}$$

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta \in \mathcal{D}} \left|\widehat{\mathbf{M}}_0(\theta, t) - \mathbf{M}_0(\theta, t)\right| > x\right) \leq n^{-2}, \tag{C10}$$

for any $x \geq (1+2c)\sqrt{\frac{\log n}{n}}$ *and provided that* $(1+2c)\sqrt{\frac{\log n}{n}} \leq 1$.

Proof. For a fixed $t \in (0,1]$, let $\mathscr{F}_{L_0} = \{T(.,\theta)\mathbb{1}\{T(.,\theta) \leq t\}, \theta \in \mathcal{D}\}$, $\mathscr{F}_{L_1} = \{\mathbb{1}\{T(.,\theta) \leq t\}, \theta \in \mathcal{D}\}$, and $\mathscr{F}_{M_0} = \{U(.,\theta)\mathbb{1}\{T(.,\theta) \leq t\}, \theta \in \mathcal{D}\}$. We apply Lemmas 14 and 15 for $\xi_i = X_i$, $1 \leq i \leq n$, $b = 1$, $a = 0$ and for each $\mathscr{F} \in \{\mathscr{F}_{L_0}, \mathscr{F}_{L_1}, \mathscr{F}_{M_0}\}$ to get that the corresponding probability in (C8), (C9) and (C10) is at most n^{-2} by taking

$$x \geq \sqrt{\frac{\log n}{n}} + 2\mathbb{E}\mathfrak{R}_n(\mathscr{F}),$$

where $\mathfrak{R}_n(\mathscr{F})$ denotes the Rademacher complexity of $\mathscr{F}$, see (C14). We now bound each $\mathfrak{R}_n(\mathscr{F})$ by using Lemma 12:

$$\begin{aligned}\mathbb{E}\mathfrak{R}_n(\mathscr{F}_{L_0}) &\leq \mathbb{E}\mathfrak{R}_n(\mathscr{F}_{L_1}) + \mathbb{E}\mathfrak{R}_n(\{T(.,\theta), \theta \in \Theta\}) \\ &\leq \mathbb{E}\mathfrak{R}_n(\mathscr{F}_{L_1}) + \sum_{q=1}^{Q}\mathbb{E}\mathfrak{R}_n(\{\ell_q(.,\theta), \theta \in \Theta\}); \qquad \text{(C11)}\\ \mathbb{E}\mathfrak{R}_n(\mathscr{F}_{L_1}) &\leq \sum_{q=1}^{Q}\mathbb{E}\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) < 1-t\}, \theta \in \Theta\}); \qquad \text{(C12)}\\ \mathbb{E}\mathfrak{R}_n(\mathscr{F}_{M_0}) &\leq \mathbb{E}\mathfrak{R}_n(\mathscr{F}_{L_1}) + \mathbb{E}\mathfrak{R}_n(\{U(.,\theta), \theta \in \Theta\}),\end{aligned}$$

where for (C11) and (C12), we used that $T(.,\theta) = 1 - \max_q \ell_q(.,\theta)$ and $\mathbb{1}\{T(.,\theta) \leq t\} = 1 - \prod_{q=1}^{Q}\mathbb{1}\{\ell_q(.,\theta) < 1-t\}$ and the fact that the variables $\ell_q(X_i,\theta)$ are continuous by Assumption 1. Similarly, we have $U(.,\theta) = \sum_{q=1}^{Q}\ell_q(.,\theta^*)\prod_{k\neq q}\mathbb{1}\{\ell_q(.,\theta) \geq \ell_k(.,\theta)\}$. Hence, Lemma 12 once again entails that

$$\begin{aligned}\mathbb{E}\mathfrak{R}_n(\{U(.,\theta), \theta \in \Theta\}) \leq \sum_{q=1}^{Q}\sum_{k=1,k\neq q}^{Q}\mathbb{E}\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) - \ell_k(.,\theta) \geq 0\}, \theta \in \Theta\}) \\ + \sum_{q=1}^{Q}\mathbb{E}\mathfrak{R}_n(\{\ell_q(.,\theta), \theta \in \Theta\}). \qquad \text{(C13)}\end{aligned}$$

To bound both $\mathbb{E}\mathfrak{R}_n(\{\ell_q(.,\theta), \theta \in \Theta\})$ and $\mathbb{E}\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) < 1-t\}, \theta \in \Theta\})$, we use the results of Baraud (2016) (more specifically the proof of Theorem 1 therein), to obtain that they are bounded by

$$\sqrt{\mathscr{V}\log\frac{2en}{\mathscr{V}}}\frac{\sqrt{2}}{\sqrt{n}} + 4\mathscr{V}\log\frac{2en}{\mathscr{V}}\frac{1}{n} \leq 7\sqrt{\mathscr{V}\frac{\log n}{n}},$$

provided that $\mathscr{V}(\log n)/n \leq 1$ and for $n \geq (2e)^3$. Similarly, $\mathbb{E}\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) - \ell_k(.,\theta) \geq 0\}, \theta \in \Theta\})$ is bounded by

$$\sqrt{\mathscr{V}_-\log\frac{2en}{\mathscr{V}_-}}\frac{\sqrt{2}}{\sqrt{n}} + 4\mathscr{V}_-\log\frac{2en}{\mathscr{V}_-}\frac{1}{n} \leq 7\sqrt{\mathscr{V}_-\frac{\log n}{n}},$$

$\mathscr{V}_-(\log n)/n \leq 1$ and for $n \geq (2e)^3$. Combining this with what is above entails

$$\mathbb{E}\mathfrak{R}_n(\mathcal{F}_{L_1}) \le 7Q\sqrt{\mathscr{V}\frac{\log n}{n}}$$
$$\mathbb{E}\mathfrak{R}_n(\mathcal{F}_{L_0}) \le 14Q\sqrt{\mathscr{V}\frac{\log n}{n}}$$
$$\mathbb{E}\mathfrak{R}_n(\mathcal{F}_{M_0}) \le 14Q\sqrt{\mathscr{V}\frac{\log n}{n}} + 7Q^2\sqrt{\mathscr{V}_-\frac{\log n}{n}}.$$

In particular, all expectations are upper-bounded by $c\sqrt{\frac{\log n}{n}}$, which leads to the result. ∎

Lemma 12. *If $\mathcal{F}$ is a class of indicator functions and $\mathcal{G}$ is a class of functions from $\mathbb{R}^d$ to $[0,1]$, we have*

$$\mathbb{E}\mathfrak{R}_n(\mathcal{F}\cdot\mathcal{G}) \le \mathbb{E}\mathfrak{R}_n(\mathcal{F}) + \mathbb{E}\mathfrak{R}_n(\mathcal{G})$$
$$\mathbb{E}\mathfrak{R}_n(\max(\mathcal{F},\mathcal{G})) \le \mathbb{E}\mathfrak{R}_n(\mathcal{F}) + \mathbb{E}\mathfrak{R}_n(\mathcal{G}),$$

where we denoted $\mathcal{F}\cdot\mathcal{G} = \{fg, f\in\mathcal{F}, g\in\mathcal{G}\}$ and $\max(\mathcal{F},\mathcal{G}) = \{f\vee g, f\in\mathcal{F}, g\in\mathcal{G}\}$.

Proof. We have

$$\begin{aligned}\mathbb{E}\mathfrak{R}_n(\mathcal{F}\cdot\mathcal{G}) &= \mathbb{E}\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i f.g(X_i)\right|\right)\\ &\le \mathbb{E}\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i (f(X_i)+g(X_i))\right|\right)\\ &\le \mathfrak{R}_n(\mathcal{F}+\mathcal{G}),\end{aligned}$$

because $fg = (f+g-1)_+ = 0.5(f+g-1+|f+g-1|)$ and by applying the contraction lemma of Talagrand (see e.g. Lemma 5.7 in Mohri et al., 2012) with $x\mapsto 0.5(x-1+|x-1|)$ which is 1-Lipchitz. Then we conclude by using the triangular inequality. For the max, we use $\max(f,g) = 0.5(f+g+|f-g|)$. ∎

Lemma 13. *Consider the case where $Q=2$ and $\{F_u, u\in\mathcal{U}\}$ is an exponential family, i.e. there exists some functions A,B,C,D such that $f(x,u) = \exp\big(A(u)^tB(x) - C(u) + D(x)\big)$. Let k be the dimension of the sufficient statistic vector $B(x)$. If $k\ge 3$, then $\mathscr{V}, \mathscr{V}_-$ defined by (16), (17) satisfy $\mathscr{V}, \mathscr{V}_- \le Qk(k+1)\big[3\log(k(k+1)) + 2(Q-1)\big]$. In addition, this bound still holds for $\mathscr{V}_-$ in the case $Q\ge 3$.*

Proof. Let us first bound $\mathscr{V}$. Given that, for $Q=2$, $\theta = (\pi_1,\pi_2,\phi_1,\phi_2)$, $\ell_1(x,\theta)\ge t$ is equivalent to $\pi_1 f(x,\phi_1)/\pi_2 f(x,\phi_2) \ge g(t)$ for some function g, we get that $\ell_1(x,\theta)\ge t$ if and only if $a(\theta)^tB(x) - b(\theta) \ge h(t)$ for some functions a,b,h. The set family is a subset of $\{\{x\in\mathbb{R}^d, a^tB(x)+b\ge 0\}, a\in\mathbb{R}^k, b\in\mathbb{R}\}$, whose VC dimension is bounded by $k(k+1)\big[3\log(k(k+1)) + 2\big]$ for $k\ge 3$, see Lemma 10.3 in Shalev-Shwartz and Ben-David (2014). By symmetry, this bound also holds for the VC dimension of $\{\ell_2(\cdot,\theta), \theta\in\Theta\}$. It follows that $\mathscr{V} \le Qk(k+1)\big[3\log(k(k+1)) + 2\big] + 2(Q-1)$ (see,

e.g., Exercice 3.24 in Mohri et al. (2012) on the VC dimension of the union of two classes with bounded VC dimension).

For $\mathscr{V}_-$, we have that for any $q \neq q' \in \{1,\dots,Q\}$, $\ell_q(x,\theta) - \ell_{q'}(x,\theta) \geq 0$ is equivalent to $\pi_q f(x,\phi_q)/\pi_{q'} f(x,\phi_{q'}) \geq 1$. The rest of the proof follows similarly as for $\mathscr{V}$. ■

Lemma 14 (Talagrand's inequality, Theorem 5.3. in Massart (2007)). *Let $\xi_1,\dots,\xi_n$ independent r.v., $\mathscr{F}$ a countable class of measurable functions s.t. $a \leq f \leq b$ for every $f \in \mathscr{F}$ for some real numbers $a \leq b$, and $W = \sup_{f\in\mathscr{F}} |\sum_{i=1}^n f(\xi_i) - \mathbb{E}(f(\xi_i))|$. Then, for any $x > 0$,*

$$\mathbb{P}(W - \mathbb{E}(W) \geq x) \leq e^{-\frac{2x^2}{n(b-a)^2}}.$$

Lemma 15 (Rademacher complexity bound, see, e.g., Lemma 1 in Baraud (2016)). *In the setting of Lemma 14 (and with the notation therein), we have*

$$\mathbb{E}(W) \leq 2\mathfrak{R}_n(\mathscr{F}),$$

where

$$\mathfrak{R}_n(\mathscr{F}) = \sup_{f\in\mathscr{F}} \left| \sum_{i=1}^n \varepsilon_i f(\xi_i) \right| \tag{C14}$$

is the Rademacher complexity of the class $\mathscr{F}$ (with $\varepsilon_1,\dots,\varepsilon_n$ being i.i.d. random signs).

APPENDIX D. AUXILIARY RESULTS FOR THE GAUSSIAN CASE

D1 Convergence rate for parameter estimation

The following result presents two situations where the parameter of a Gaussian mixture model can be consistently estimated, with an explicit rate.

Proposition 1. *Consider the mixture model (Section 2.1) in the d-multivariate Gaussian case with true parameter $\theta^* = (\pi^*, \phi^*)$, where $\phi_q^* = (\mu_q^*, \Sigma_q^*)$, $1 \leq q \leq Q$. Then $\eta(\epsilon,\theta^*)$ defined by (18) is such that $\eta(\epsilon_n,\theta^*) \leq 1/n$ for $\epsilon_n \geq C\sqrt{\log n/n}$, where $C > 0$ is a sufficiently large constant, in two following situations*:

(*i*) *$\hat{\theta}$ is the constrained MLE, that is, computed for $\phi_q = (\mu_q, \Sigma_q) \in \mathcal{U}$ with constrained parameter space $\mathcal{U} = [-a_n, a_n]^d \times \{\Sigma \in S_d^{++}, \underline{\lambda} \leq \lambda_1(\Sigma) \leq \lambda_d(\Sigma) \leq \overline{\lambda}\}$[1] where $a_n \leq L(\log n)^\gamma$ for some $L, \gamma > 0$ and S_d^{++} denotes the space of positive definite matrices, with $\underline{\lambda}, \overline{\lambda} > 0$. In that case, C only depends on θ^* and $L, \gamma, \underline{\lambda}, \overline{\lambda}$.*

(*ii*) *$\hat{\theta}$ is the estimator coming from EM algorithm (when the iteration number is infinite) for an initialization $\mu_1^{(0)}, \mu_2^{(0)}$ such that $\|(\mu_1^{(0)} - \mu_2^{(0)}) - (\mu_1 - \mu_2)\| \leq \Delta/4$, where $\Delta = \|\mu_1 - \mu_2\|_2$ is the separation between the true means. Here, we consider an homoscedastic model with $\Sigma_1 = \Sigma_2 = \Sigma = \nu I_d$ with known ν. The conclusion applies if the signal-to-noise ratio Δ/ν is large enough, and for a constant C of the form $c(\nu,\Delta)\sqrt{d}$.*

Proof. Since case (ii) is a direct application of Balakrishnan et al. (2017), we focus in what follows on proving case (i), by revisiting the result of Ho and Nguyen (2016). First, in the considered model, any mixture can be defined in terms of $\{f_u, u \in \mathcal{U}\}$ and a discrete mixing measure $G = \sum_{q=1}^{Q} \pi_q \delta_{\phi_q}$ with Q support points, as $\sum_{q=1}^{Q} \pi_q f_{\phi_q} = \int f_u(x) dG(u)$. As shown by Ho and Nguyen (2016), the convergence of mixture model parameters can be measured in terms of a Wasserstein distance on the space of mixing measures. Let $G_1 = \sum_{q=1}^{Q} \pi_q^1 \delta_{\phi_q^1}$ and $G_2 = \sum_{q=1}^{Q} \pi_q^2 \delta_{\phi_q^2}$ be two discrete probability measures on some parameter space, which is equipped with metric $\|.\|$. The Wasserstein distance of order 1 between G_1 and G_2 is given by

$$W_1(G_1, G_2) = \inf_{p} \sum_{q,l} p_{q,l} \|\phi_q^1 - \phi_l^2\|$$

where the infimum is over all couplings $(p_{q,l})_{1 \leq q,l \leq Q} \in [0,1]^{Q \times Q}$ such that $\sum_l p_{q,l} = \pi_q^1$ and $\sum_q p_{q,l} = \pi_l^2$. Let $G^*, \hat{G}_n$ denote the true mixing measure and the mixing measure that corresponds to the restricted MLE considered here, respectively. Theorem 4.2. in Ho and Nguyen (2016) implies that, with the notation of Ho and Nguyen (2016), for any $\epsilon_n \geq (\sqrt{C_1}/c)\delta_n$, and $\delta_n \leq C\sqrt{\log n/n}$, we have $\mathbb{P}_{\theta^*}(W_1(\hat{G}_n, G^*) \geq (c/C_1)\epsilon_n) \leq ce^{-n\epsilon_n^2}$. We apply this relation for $\epsilon_n = \max((\sqrt{C_1}/c)\delta_n, \sqrt{\log(cn)/n})$. In that case, we have still ϵ_n of order $\sqrt{\log n/n}$ and the upper-bound is at most $1/n$. On the other hand, if we have a convergence rate in terms of W_1, then we have convergence of the mixture model parameters in terms of $\|.\|$ at the same rate, see Lemma 16. This concludes the proof. ∎

Lemma 16. *Let $G_n = \sum_{q=1}^{Q} \pi_q^n \delta_{\phi_q^n}$ be a sequence of discrete probability measures on $\mathcal{U}$, and let G^*, W_1 be defined as in the proof of Proposition 1. There exists a constant C only depending on G^* such that if $W_1(G_n, G^*) \to 0$, then for sufficiently large n,*

$$W_1(G_n, G^*) \geq C \min_{\sigma \in [Q]} \|\theta_n^\sigma - \theta^*\|.$$

Proof. In what follows, we let $\{p_{q,l}^n\}$ denote the corresponding probabilities of the optimal coupling for the pair (G_n, G^*). We start by showing that $(\phi_q^n)_q \to (\phi_q^*)_q$ in $\|.\|$ up to a permutation of the labels. Let σ^n the permutation of the labels such that $\|\phi_q^n - \phi_l^*\| \geq \|\phi_{\sigma^n(l)}^n - \phi_l^*\|$ for all $q, l \in \{1, \ldots, Q\}$. Then, by definition,

$$\begin{aligned} W_1(G_n, G^*) &\geq \sum_{1 \leq q,l \leq Q} p_{q,l}^n \|\phi_{\sigma(l)}^n - \phi_l^*\| \\ &= \sum_l \pi_l^* \|\phi_{\sigma^n(l)}^n - \phi_l^*\|. \end{aligned}$$

It follows that each $\|\phi_{\sigma^n(l)}^n - \phi_l^*\|$ must converge to zero. Since $(\phi_q^n)_q \to (\phi_q^*)_q$ up to a permutation of the labels, without loss of generality we can assume that $\phi_q^n \to \phi_q^*$ for all q. Let $\Delta\phi_q^n := \phi_q^n - \phi_q^*$ and $\Delta\pi_q^n := \pi_q^n - \pi_q^*$. Write $W_1(G_n, G^*)$ as

$$W_1(G_n, G^*) = \sum_q p_{qq}^n \|\Delta\phi_q^n\| + \sum_{q \neq l} p_{ql}^n \|\phi_q^n - \phi_l^*\|$$

Define $C_{ql} = \|\phi_q^* - \phi_l^*\|$ and $C = \min_{q \neq l} C_{ql} > 0$. It follows from the convergence of ϕ^n that for $q \neq l$, $\|\phi_q^n - \phi_l^*\| \geq C/2$ for sufficiently large n. Thus,

$$W_1(G_n, G^*) \geq \frac{C}{2} \sum_{q \neq l} p_{ql}^n$$

We deduce that $\sum_{q \neq l} p_{ql}^n \to 0$. As a result, $p_{qq}^n = \pi_q^* - \sum_{l \neq q} p_{lq}^n \to \pi_q^*$, and so, $p_{qq}^n \geq (1/2)\pi_{\min}^*$ $:= \min_l \pi_l^*$ for sufficiently large n. On the other hand, $\sum_{q \neq l} p_{ql}^n = \sum_q \pi_q^n - p_{qq}^n = \sum_q \pi_q^* - p_{qq}^n$ where $p_{qq}^n \leq \min(\pi_q^n, \pi_q^*)$. Thus, $\sum_{q \neq l} p_{ql}^n \geq \sum_q \pi_q^n - \min(\pi_q^n, \pi_q^*) = \sum_{q, \pi_q^n \geq \pi_q^*} \pi_q^n - \pi_q^* = \sum_{q, \pi_q^n \geq \pi_q^*} |\pi_q^n - \pi_q^*|$ and similarly we have that $\sum_{q \neq l} p_{ql}^n \geq \sum_{q, \pi_q^* \geq \pi_q^n} |\pi_q^n - \pi_q^*|$. It follows that $2 \sum_{q \neq l} p_{ql}^n \geq \sum_q |\pi_q^n - \pi_q^*|$. Therefore, for sufficiently large n,

$$W_1(G_n, G^*) \geq \frac{1}{2} \pi_{\min}^* \sum_q \|\Delta \phi_q^n\| + \frac{C}{4} \sum_q |\Delta \pi_q^n|.$$

This gives the result. ∎

D2 Gaussian computations

The following lemma holds.

Lemma 17. *Let us consider the multivariate Gaussian case where* $\phi_q = (\mu_q, \Sigma_q)$, $1 \leq q \leq Q$, *with* $Q = 2$, $\Sigma_1 = \Sigma_2$ *is an invertible covariance matrix and* μ_1 *and* μ_2 *are two different vectors of* $\mathbb{R}^d$. *Then Assumptions 1, 2 and 3 hold true for* $\alpha_c = 0$ *and for a level* $\alpha \in (0, \overline{\alpha}) \setminus \mathcal{E}$ *for* $\mathcal{E}$ *a set of Lebesgue measure* 0.

Proof. Let us first prove that $\ell_q(X, \theta)$ is a continuous random variable under $\mathbb{P}_{\theta^*}$ (this is established below without assuming $\Sigma_1 = \Sigma_2$ for the sake of generality). We have

$$\begin{aligned}
&\mathbb{P}_{\theta^*}(\ell_1(X, \theta) = t) \\
&= \mathbb{P}_{\theta^*}\left(f_{\phi_1}(X)/f_{\phi_2}(X) = t\pi_2/\pi_1\right) \\
&= \mathbb{P}_{\theta^*}\left((X - \mu_1)^t \Sigma_1^{-1}(X - \mu_1) - (X - \mu_2)^t \Sigma_2^{-1}(X - \mu_2) = -2\log(t\pi_2/\pi_1) - \log(|\Sigma_1|/|\Sigma_2|)\right).
\end{aligned}$$

Now,

$$\begin{aligned}
&(X - \mu_1)^t \Sigma_1^{-1}(X - \mu_1) - (X - \mu_2)^t \Sigma_2^{-1}(X - \mu_2) \\
&\quad = (X - \mu_1)^t \Sigma_1^{-1}(X - \mu_1) - (X - \mu_1)^t \Sigma_2^{-1}(X - \mu_2) - (\mu_1 - \mu_2)^t \Sigma_2^{-1}(X - \mu_2) \\
&\quad = (X - \mu_1)^t (\Sigma_1^{-1} - \Sigma_2^{-1})(X - \mu_1) - (X - \mu_1)^t \Sigma_2^{-1}(\mu_1 - \mu_2) - (\mu_1 - \mu_2)^t \Sigma_2^{-1}(X - \mu_2) \\
&\quad = (X - \mu_1)^t (\Sigma_1^{-1} - \Sigma_2^{-1})(X - \mu_1) - (\mu_1 - \mu_2)^t \Sigma_2^{-1}(2X - \mu_2 - \mu_1).
\end{aligned}$$

Since the real matrix $\Sigma_1^{-1} - \Sigma_2^{-1}$ is symmetric, we can diagonalize it and we end up with a subset of $\mathbb{R}^d$ of the form

$$\left\{ y \in \mathbb{R}^d \ : \ \sum_{j=1}^d \left(\alpha_j y_j^2 + \beta_j y_j\right) + \gamma = 0 \right\},$$

for some real parameters $\alpha_j, \beta_j, \gamma$. The result follows because this set has a Lebesgue measure equal to 0 in any case.

Now, since $\Sigma_1 = \Sigma_2 = \Sigma$, we have for all $t \in (0,1)$,

$$\begin{aligned}\{T(X,\theta) > t\} &= \left\{ \forall q \in \{1,\dots,Q\}, \pi_q f_{\phi_q}(X) < (1-t)\sum_{\ell=1}^{Q} \pi_\ell f_{\phi_\ell}(X) \right\} \\ &= \left\{\pi_1 f_{\phi_1}(X) < (1/t-1)\pi_2 f_{\phi_2}(X)\right\} \cap \left\{\pi_2 f_{\phi_2}(X) < (1/t-1)\pi_1 f_{\phi_1}(X)\right\} \\ &= \left\{ (1/t-1)^{-1} < \frac{\pi_1 f_{\phi_1}(X)}{\pi_2 f_{\phi_2}(X)} < (1/t-1) \right\}.\end{aligned}$$

Applying $2\log(\cdot)$ on each part of the relation, we obtain

$$\{T(X,\theta) > t\} = \left\{-2\log(1/t-1) < a^t X + b < 2\log(1/t-1)\right\},$$

for

$$\begin{aligned} a &= a(\theta) = 2\Sigma^{-1}(\mu_1 - \mu_2) \in \mathbb{R}^d \setminus \{0\} \\ b &= b(\theta) = -(\mu_1-\mu_2)^t\Sigma^{-1}(\mu_1+\mu_2) + 2\log(\pi_1/\pi_2) \in \mathbb{R}^d. \end{aligned}$$

Since under P_{θ^*} we have $X \sim \pi_1^* \mathcal{N}(\mu_1^*, \Sigma^*) + \pi_2^* \mathcal{N}(\mu_2^*, \Sigma^*)$, we have $a^t X + b \sim \pi_1^* \mathcal{N}(a^t\mu_1^* + b, a^t\Sigma^* a) + \pi_2^* \mathcal{N}(a^t\mu_2^* + b, a^t\Sigma^* a)$. This yields for all $t \in (0,1)$,

$$\begin{aligned}\mathbb{P}_{\theta^*}(T(X,\theta) > t) = \pi_1 &\left[\Phi\left(\frac{2\log(1/t-1) - a^t\mu_1^* - b}{(a^t\Sigma^* a)^{1/2}}\right)\right. \\ &\left.-\Phi\left(\frac{-2\log(1/t-1) - a^t\mu_1^* - b}{(a^t\Sigma^* a)^{1/2}}\right)\right] \\ + \pi_2 &\left[\Phi\left(\frac{2\log(1/t-1) - a^t\mu_2^* - b}{(a^t\Sigma^* a)^{1/2}}\right)\right. \\ &\left.-\Phi\left(\frac{-2\log(1/t-1) - a^t\mu_2^* - b}{(a^t\Sigma^* a)^{1/2}}\right)\right]. \end{aligned} \tag{D1}$$

A direct consequence is that for all $t \in (0,1)$, we have $\mathbb{P}_{\theta^*}(T(X,\theta) > t) < 1$, that is, $\mathbb{P}_{\theta^*}(T(X,\theta) \le t) = \mathbb{P}_{\theta^*}(T(X,\theta) < t) > 0$. Hence, α_c defined in (14) is equal to zero. Moreover, from (D1), we clearly have that $t \in (0,1) \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) > t)$ is decreasing, so that $t \in (0,1) \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) \le t)$ is increasing. This proves that Assumption 1 holds in that case.

Let us now check Assumptions 2 and 3. Assumptions 2 and 3 (i) follow from Result 2.1 in Melnykov (2013).

As for Assumption 3 (ii), from (D1), we only have to show that the function $t \in (0,1) \mapsto \frac{\partial}{\partial t}\Phi\left(\frac{\log(1/t-1)-\alpha^*}{\beta^*}\right)$ is uniformly bounded by some constant $C = C(\alpha^*, \beta^*)$, for any $\alpha^* \in \mathbb{R}$ and $\beta^* > 0$. A straightforward calculation leads to the following: for all $t \in (0,1)$,

$$\left|\frac{\partial}{\partial t}\Phi\left(\frac{\log(1/t-1)-\alpha^*}{\beta^*}\right)\right| = \frac{e^{-(\frac{\log(1/t-1)-\alpha^*}{\beta^*})^2/2}}{\beta^*\sqrt{2\pi}}\frac{1}{t(1-t)}. \tag{D2}$$

Consider now $t_0 = t_0(\alpha^*, \beta^*) \in (0, 1/2)$ such that $(\frac{\log(1/t-1)-\alpha^*}{\beta^*})^2 \geq 2\log(1/t)$ for all $t \in (0, t_0)$. It is clear that the right-hand-side of (D2) is upper-bounded by $\frac{1}{\beta^*\sqrt{2\pi}(1-t_0)}$ on $t \in (0, t_0)$. Similarly, let $t_1 = t_1(\alpha^*, \beta^*) \in (1/2, 1)$ such that $(\frac{\log(1/t-1)-\alpha^*}{\beta^*})^2 \geq 2\log(1/(1-t))$ for all $t \in (t_1, 1)$. It is clear that the right-hand-side of (D2) is upper-bounded by $\frac{1}{\beta^*\sqrt{2\pi}t_1}$ on $t \in (t_1, 1)$. Finally, for $t \in [t_0, t_1]$, the upper-bound $\frac{1}{\beta^*\sqrt{2\pi}t_0(1-t_1)}$ is valid. This proves that Assumption 3 (ii) holds.

Let us now finally turn to Assumption 3 (iii). Lemma 8 ensures that $t \in (0, t^*(\overline{\alpha})) \mapsto \mathrm{mFSR}_t^*$ is continuous increasing. Hence, $t^* : \beta \in (0, \overline{\alpha}) \mapsto t^*(\beta)$ defined in (13) is the inverse of this function and is also continuous increasing. It is therefore differentiable almost everywhere in $(0, \overline{\alpha})$, so everywhere in $(0, \overline{\alpha}) \setminus \mathcal{E}$ where $\mathcal{E}$ is a set of Lebesgue measure 0. By taking α in $(0, \overline{\alpha}) \setminus \mathcal{E}$, this ensures that t^* is differentiable in α and thus that Assumption 3 (iii) holds. ∎

Lemma 18. *In the multivariate Gaussian case with $Q = 2$ and $\Sigma_1 = \Sigma_2$, we have that $\mathscr{V} \leq 2d + 4$ and $\mathscr{V}_- \leq 2d + 4$.*

Proof. In that case, we have that (see the proof of Lemma 17)

$$\{\ell_q(x, \theta) \leq u, x \in \mathbb{R}^d\} = \{a_\theta^t x + b_\theta \geq g(u), x \in \mathbb{R}^d\}.$$

Since the VC dimension of the vector space of real-valued affine functions is bounded by $d + 1$ (see, e.g., Exercice 3.19 in Mohri et al., 2012). We obtain the result by applying the usual bound on the VC dimension of the union of two classes with bounded VC dimension (see, e.g., Exercice 3.24 in Mohri et al., 2012). ∎

APPENDIX E. ADDITIONAL NUMERICAL EXPERIMENTS

E1 Additional bootstrap procedure

Here we evaluate an additional bootstrap procedure, that is based on the work of O'Hagan et al. (2019). O'Hagan et al. (2019) investigated different sampling methods for estimating standard errors of model parameters in the context of Gaussian mixture models. In particular, they considered the case where some of the clusters are small and/or overlapping and they proposed a so-called weighted likelihood bootstrap approach that is shown to be effective in that case. The method consists of sampling a weight vector $(w_i^b)_{1\leq i\leq n}$ (e.g. according to a uniform Dirichlet distribution), to then fit a bootstrap parameter estimate $\hat{\theta}^b$ by using the EM algorithm with the weighted log-likelihood $\sum_{i=1}^n \sum_{q=1}^Q \mathbb{1}_{Z_i=q} w_i^b (\log \pi_q + \log f_{\mu_q,\Sigma_q}(X_i))$. We adapt this approach by replacing $\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}^b)$, $\hat{S}_\alpha^{\mathrm{PI}}(\mathbf{X}^b)$ in (10) with the partition and selection $\hat{Z}_i^{\mathrm{PI}}(\mathbf{X})$, $\hat{S}_\alpha^{\mathrm{PI}}(\mathbf{X})$ of the plug-in procedure computed with $\hat{\theta}^b$ and the original sample $\mathbf{X}$. Formally, by denoting $\hat{S}_\alpha^{\mathrm{PI}}(\mathbf{X}) = \Psi_1(\mathbf{X}, \hat{\theta}(\mathbf{X}))$ and $\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}) = \Psi_2(\mathbf{X}, \hat{\theta}(\mathbf{X}))$ for some functions Ψ_1, Ψ_2, the *weighted nonparametric bootstrap* estimate of the FSR is given by

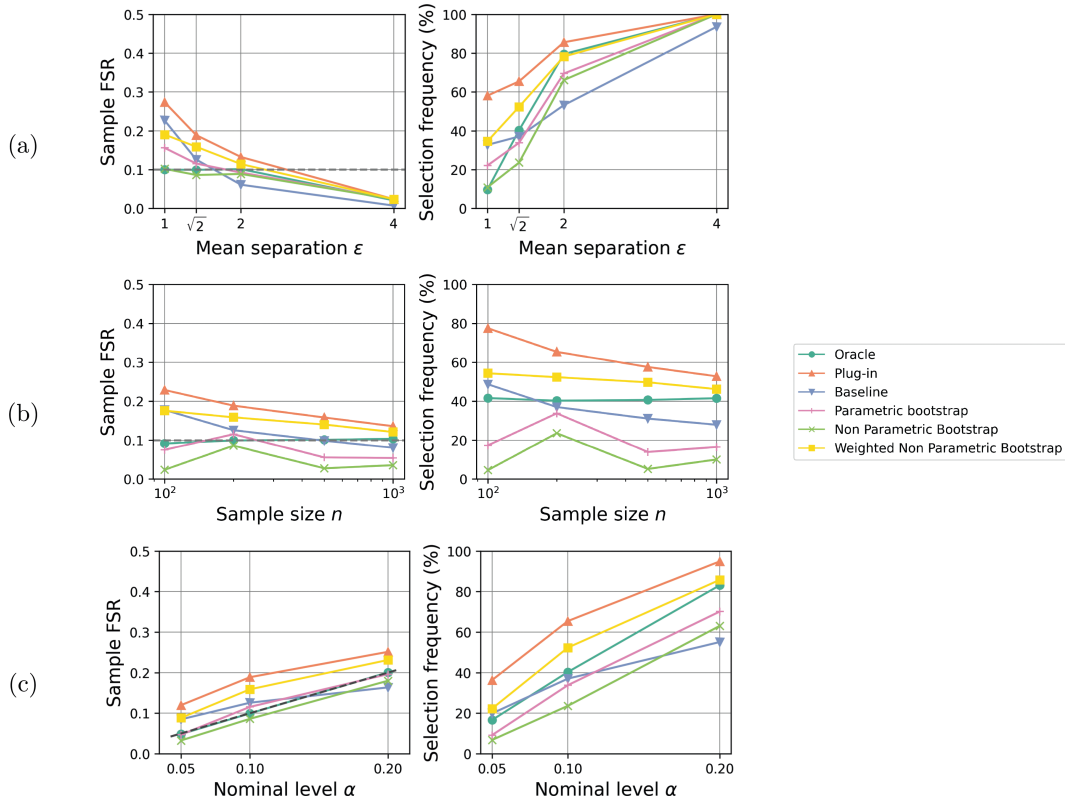


FIGURE E1 FSR (left panel) and selection frequency (right panel) as a function of: (a) the mean separation; (b) the sample size n; (c) the nominal level α. Diagonal covariances setting with $Q = 2$, $d = 2$. Default settings are: $n = 200$, $\alpha = 0.1$, $\epsilon = \sqrt{2}$.

$$\widehat{\mathrm{FSR}}_{\alpha}^{w,B} := \frac{1}{B}\sum_{b=1}^{B} \min_{\sigma\in[Q]} \frac{\sum_{i=1}^{n}\{1-\ell_{\sigma(\Psi_2(\mathbf{X},\hat{\theta}^b))}(X_i,\hat{\theta}(\mathbf{X}))\}\mathbb{1}\{i\in\Psi_1(\mathbf{X},\hat{\theta}^b)\}}{\max(|\Psi_1(\mathbf{X},\hat{\theta}^b)|,1)},$$

This is markedly different than the bootstrap approaches in (10) because the sample $\mathbf{X}$ is fixed once for all when computing the bootstrap estimates of the FSR.

We evaluate the procedure by reproducing the experiments of Figure 3, in which we considered a balanced Gaussian mixture model with $d = 2$, $Q = 2$ and the covariance matrices were assumed to be diagonal in the EM algorithm: Figure E1 displays the FSR and the selection frequency as a function of the mean separation ϵ (Figure E1a) and the nominal level α (Figure E1c). We observe that across all settings, the FSR of the weighted nonparametric bootstrap is between that of the plug-in and the parametric bootstrap. In addition, we evaluate the procedure in the presence of a very small component, considering as before a Gaussian mixture model with $d = 2$, $Q = 2$ and diagonal covariance matrices, with mixture proportions set to $\pi = (0.1, 0.9)$: Figure E2 displays the FSR and the selection frequency for varying α and $n \in \{200, 1000\}$. In this case the nonparametric bootstrap shows poor performance, which is consistent with the observations of O'Hagan et al. (2019). The parametric bootstrap still displays an FCR level close to the nominal level, which we explain by the accuracy of the specified model. The weighted nonparametric bootstrap gives results that are very close to the plug-in, as in the previous experiment, and so lacks

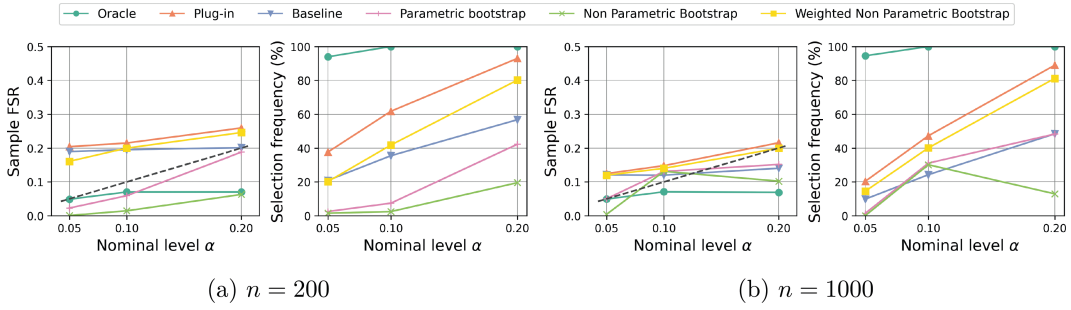


FIGURE E2 FSR (left panel) and selection frequency (right panel) as a function of the nominal level α. Gaussian mixture model with $Q = 2$, $\pi = (0.1, 0.9)$, $d = 2$, $\Sigma_1 = \Sigma_2 = I_2$, $\mu_1 = 0_2$ and $\mu_2 = (\epsilon/\sqrt{d}, \epsilon/\sqrt{d})$, $\epsilon = 2$.

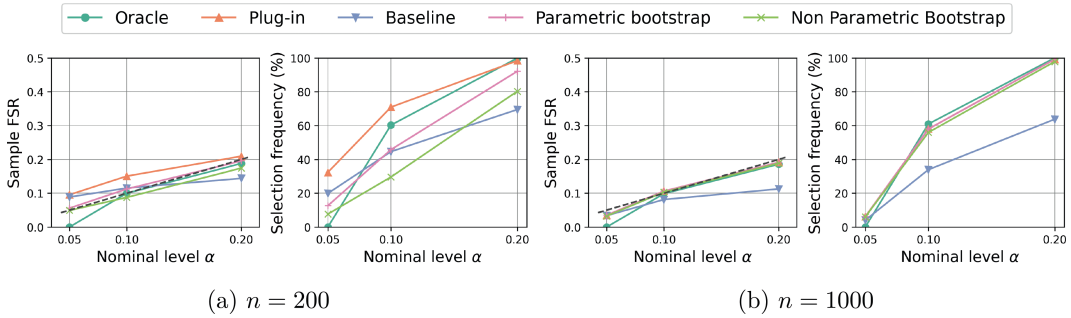


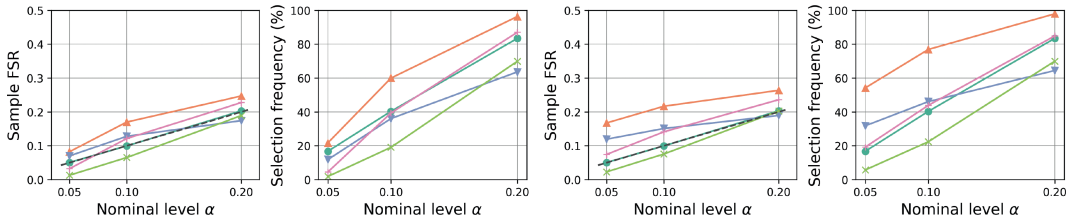
FIGURE E3 FSR (left panel) and selection frequency (right panel) as a function of the nominal level α. Here, P_{θ^*} is a t-mixture model with $Q = 2$, $\pi_1 = \pi_2 = 1/2$, $d = 2$, $\Sigma_1 = \Sigma_2 = I_2$, $\mu_1 = (0, 0)$ and $\mu_2 = (\sqrt{2}, \sqrt{2})$, and the degrees of freedom of each component is set to 4.

FCR control. We conjecture that this is due to evaluating the plug-in procedure on $\mathbf{X}$ in the computation of the FSR estimate instead of bootstrap samples $\mathbf{X}^b \sim \hat{P}$, resulting in reducing the variance of the FSP estimate and leading to an FSR estimate that is closer to the FSR of the plug-in in comparison to the other bootstrap procedures. We conclude that integrating the approach of O'Hagan et al. (2019) with our procedure presents nontrivial challenges that require further consideration and which we leave for future work.

E2 *t*-mixture models

In this section, we evaluate our procedures using t-mixture models, both on data generated from t-mixtures as well as on Gaussian mixtures. The t-mixture model is of particular interest in our context as it is appropriate for modeling data containing observations with longer than normal tails or atypical observations leading to overlapping clusters (Peel & McLachlan, 2000). In all experiments, the t-mixture is fit via the EM algorithm for t-mixtures (Peel & McLachlan, 2000) provided by the Python package `studenttmixture` (Parkinson, 2018) and no constraints are put on the parameters in the estimation procedure. In particular, all parameters are estimated—including degrees of freedom—without any assumption on the covariance structure.

We start with the case where the data is itself generated from a t-mixture. We consider $Q = 2$, $d = 2$, $\pi_1 = \pi_2 = 1/2$, $\Sigma_1 = \Sigma_2 = I_2$, $\mu_1 = (0, 0)$ and $\mu_2 = (\sqrt{2}, \sqrt{2})$, and the degrees of freedom of each component is 4. Figure E3 displays the FSR and the selection frequency for varying α



(a) $P_{\hat{\theta}}$: t-mixture (b) $P_{\hat{\theta}}$: Gaussian mixture

FIGURE E4 FSR (left panel) and selection frequency (right panel) as a function of the nominal level α. Here, P_{θ^*} is a Gaussian mixture model with $Q = 2$, $\pi_1 = \pi_2 = 1/2$, $d = 2$, $\Sigma_1 = \Sigma_2 = I_2$, $\mu_1 = (0, 0)$ and $\mu_2 = (1, 1)$. The sample size is $n = 200$.

and $n \in \{200, 1000\}$. The conclusions are qualitatively the same as in Section 5.1: the plug-in procedure has an FSR that exceeds the nominal level when the sample size is too small, whereas the bootstrap procedures allow for a more robust FSR control, the nonparametric one being the more robust and controlling the FSR in all settings.

Next, we evaluate the procedures with the use of t-mixture modeling on Gaussian-generated data. Specifically, P_{θ^*} is a Gaussian mixture model with $Q = 2$, $\pi_1 = \pi_2 = 1/2$, $\Sigma_1 = \Sigma_2 = I_2$, $\mu_1 = (0, 0)$ and $\mu_2 = (1, 1)$. However, concerning the specification of $P_{\hat{\theta}}$ we choose to model the generated data as a t-mixture. The idea is to fit a model that is more robust for parameter estimation in the context of overlapping clusters. Compared to Gaussian mixtures, Student's t-distributions are less concentrated and thus produce estimates of the posterior probabilities of class memberships that are less extreme, which is favorable for our selection procedures. Figure E4a displays the FSR and the selection frequency for varying α and $n = 200$. For comparison, Figure E4b displays the results obtained when using Gaussian mixture modeling, fit via the EM algorithm and with no parameter constraints. As expected, the use of t-mixture modeling leads to a lower FSR for all procedures. In particular, the FSR of the plug-in procedure and the parametric bootstrap procedure are closer to the nominal level α. In the case of the nonparametric bootstrap, since the procedure is already controlling the FSR when using Gaussian mixtures, the use of t-mixture modeling makes the procedure more conservative, but only very slightly.