Electronic Journal of Statistics
Vol. 16 (2022) 3910–3936
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2032

Determine the number of clusters by data augmentation

Wei Luo

Center for Data Science, Zhejiang University,
866 Yuhangtang Road, Hangzhou, China
e-mail: weiluo@zju.edu.cn

Abstract: Determining the number of clusters is crucial for the successful application of clustering. In this paper, we propose a new order-determination method called the data augmentation estimator (DAE), for the general model-based clustering. The estimator is based on a novel idea that augments data with an independently generated small cluster, which enables us to justify how the instability of clustering changes with the number of clusters assumed in clustering. The pattern of instability provides an alternative characterization of the true number of clusters to the commonly used goodness-of-fit measure. By combining the two sources of information appropriately, the proposed estimator reaches asymptotic consistency under general conditions and is easily implementable. It is also more efficient than the conventional BIC-type approaches that use the goodness-of-fit measure only. These properties are illustrated by the simulation studies and real data examples at the end.

Keywords and phrases: Data augmentation, instability of clustering, model-based clustering, order determination.

Received August 2021.

1. Introduction

Clustering has been an important research problem in unsupervised statistical learning for decades. It helps the researchers to detect the latent grouping pattern of data, so that specific sub-populations can be identified. In supervised learning procedures, the effect of the predictor may vary with the underlying sub-populations, so clustering of the predictor, if feasible, can also facilitate statistical modeling.

In the statistical literature, the clustering problem is commonly formulated as fitting a mixture model [13, 18, 24]. That is, suppose the observed random vector X is p-dimensional and is absolutely continuous with respect to the Lebesgue measure ζ_p on $\mathbb{R}^p$; we assume that $f_0(\cdot)$, the probability density function (pdf) of X, has the form

$$f_0(x) = \sum_{k=1}^{K} \pi_{k,0} f(x, \mu_{k,0}), \tag{1}$$

where $\pi_{1,0}, \ldots, \pi_{K,0}$ are positive weights with $\sum_{k=1}^{K} \pi_{k,0} = 1$, and each cluster $f(\cdot, \mu_{k,0})$ is a pdf with respect to ζ_p and belongs to the prefixed parametric family $\mathcal{H}_1 = \{f(x, \mu) : x \in \mathbb{R}^p, \mu \in \mathbb{R}^q\}$. To ensure the identifiability of (1), $\mu_{k,0}$'s must

be distinct from each other, and additional regularity conditions must be made on $\mathcal{H}_1$. These regularity conditions were built in 1960s and are applicable to general parametric families of continuous distributions, except for the family of uniform distributions with unspecified supports. A summary of relative details can be found in [19] and [31]. Under these regularity conditions, the parameter of interest in Model (1) can be formulated as the set $\omega_0 = \{(\pi_{k,0}, \mu_{k,0}) : k = 1, \ldots, K\}$, which is invariant of the rearrangements of its elements and thus is identifiable.

To fit the mixture model (1), numerous methods have been proposed, including the maximal likelihood estimator [MLE; 19], the method of moments [19], and the Group-Sort-Fuse procedure [GSF; 18], etc. Except for GSF, a common feature of these methods is that their consistency or efficiency hinges on a true specification of the number of clusters K: if K is underestimated, the working mixture model cannot recover (1) and the result loses consistency; if K is overestimated, the working mixture model will contain redundant parameters, which adds inefficiency and potential instability to the subsequent modeling. Hence, the order determination of K is a crucial problem in clustering.

Following the literature of model selection, order determination for clustering has often been conducted by information criteria, such as BIC and the integrated completed likelihood [ICL; 5], etc. The common strategy of these methods is to incorporate a penalty function of the working number of clusters into an appropriate goodness-of-fit measure. Commonly used goodness-of-fit measures include the log-likelihood [13, 17, 22, 24, 32], the gap statistic [30, 33], and the distortion statistic [26], etc. Because these measures convey an elbow shape asymptotically, with the elbow converging to K, the consistency of the information criteria directly follows [14, 19]. Based on the log-likelihood, a sequential testing procedure has also been constructed for order determination [19], but its validity highly relies on the true specification of $\mathcal{H}_1$. The aforementioned GSF [18] selects K automatically by fusing similar estimates of clusters into one when fitting (1). It requires arranging the clusters appropriately according to their similarities, whose feasibility, however, can be questionable in practice. We refer to [12] and [18] for a comprehensive review of the existing order-determination methods for mixture models.

As an alternative to the model-based clustering, distance-based clustering has also been widely studied, particularly in the literature of machine learning research. In this scenario, the research interest is to obtain a cluster assignment rule that divides the data into clusters or equivalently associates each outcome with a specific cluster. Representative distance-based clustering methods include the K-means method and its generalizations such as the power K-means method [34] and the convex clustering methods [23, 27, 28], etc. Order determination for the K-means method is often based on the instability of clustering, measured by the variation of cluster assignment rules derived from different samples [8, 11, 29]. With the conjecture that an overestimate of K would cause an instable clustering, researchers commonly estimate K by the largest working number of clusters that leads to a stable clustering. While this criterion is frequently used in practice, its underlying conjecture remains unjustified.

In this paper, we focus on order determination for the model-based clustering. However, we will borrow the concept of the instability measures originally designed for distance-based clustering, and combine it with the conventional goodness-of-fit measure to sharpen the estimation. Under Model (1), given an working number of clusters L and an estimate $\{(\widehat{\pi}_{l,0}, \widehat{\mu}_{l,0}) : l = 1, \dots L\}$ of ω_0, a cluster assignment rule is naturally derived by the Bayes theorem that associates an outcome x with the cluster that has the maximal $\widehat{\pi}_{l,0} f(x, \widehat{\mu}_{l,0})$ among all $l \in \{1, \dots, L\}$. Thus, the instability of a cluster assignment rule is equivalent to a measure of variation in estimating ω_0. While this term is generally difficult to be delineated for the original data, we will address the issue by adopting a novel data augmentation approach, where a small new cluster is generated independently and merged into the original data.

Compared with the goodness-of-fit measure, the clustering instability provides an alternative characterization of K and conveys a compensative pattern, so these two together can generate a consistent estimator of K that is more efficient than the BIC-type approaches. By its nature, we call this estimator the data augmentation estimator (DAE). As we shall see later, DAE does not involve asymptotic inference results, so it is more robust to the parametric assumption, i.e. the specification of $\mathcal{H}_1$, than the likelihood-based sequential testing procedure. In addition, its implementation requires no arrangement of clusters as in GSF. In these ways, DAE outperforms the existing order-determination methods. With the aid of bootstrap re-sampling, we also propose a variation of DAE, called DAE-Ⅱ, which shares the same properties.

The rest of the paper is organized as follows. In Section 2, we formulate the model fitting procedure for (1) in a general manner. The idea of data augmentation is introduced in Section 3, where we also study how it impacts the clustering results. We propose DAE and DAE-Ⅱ in Section 4, and discuss their implementations in Section 5. Section 6 and Section 7 are devoted to simulation studies and real data applications, respectively. A summary of the proposed work is presented in Section 8 at the end. For ease of presentation, we assume X to be continuous throughout the theoretical development. The proposed method also works for discrete X; see the simulation studies for an illustration.

2. A formulation of model fitting

We now formulate the estimation procedure for Model (1) based on the literature reviewed in Section 1. Let $X_1, \dots, X_n$ be the sample copies of X, and let $\widehat{F}_n$ be the corresponding empirical distribution function, i.e. $\widehat{F}_n(A) = \sum_{i=1}^n I(X_i \in A)/n$ for any measurable set $A \subset \mathbb{R}^p$. We assume independence or weak dependence between $X_1, \dots, X_n$, such that the general parametric analysis based on $\widehat{F}_n$ has the desired $n^{1/2}$-consistency. Let F_0 be the distribution function of X, i.e. that corresponds to the pdf f_0 in (1), and let $\mathcal{G}$ be the collection of all the probability distributions on $\mathbb{R}^p$, which includes both $\widehat{F}_n$ and F_0. For each positive integer L, let $\mathcal{H}_L$ be the family of the pdf of mixture models induced

from $\mathcal{H}_1$ with at most L clusters, i.e.

$$\mathcal{H}_L = \cup_{l=1}^{L}\{\sum_{i=1}^{l}\pi_i f(\cdot,\mu_i) : \pi_i > 0, \sum_{i=1}^{l}\pi_i = 1, f(\cdot,\mu_i) \in \mathcal{H}_1, \mu_i \neq \mu_j\}.$$

This definition complies with the previously defined $\mathcal{H}_1$ when $L = 1$. As L grows, $\mathcal{H}_L$ forms an increasing sequence, i.e. $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots$. In particular, $\mathcal{H}_L$ includes the true density f_0 whenever $L \geq K$, which means that K is the smallest integer such that $\mathcal{H}_K$ includes f_0. We equip the functional L_2 distance on $\mathcal{H}_L$, i.e.

$$\|f - g\|_2 \equiv \int_{\mathbb{R}^p}\{f(x) - g(x)\}^2 dx \tag{2}$$

for any $f, g \in \mathcal{H}_L$. When $L < K$, we regard $\mathcal{H}_L$ as a subset of $\mathcal{H}_K$. Then, under the regularity conditions on $\mathcal{H}_1$ that ensure the identifiability of f_0 (see Section 1), $\mathcal{H}_L$ has a positive distance with f_0.

For the most generality of our theory, given the working number of clusters L, we formulate the estimation of Model (1) as maximizing a stochastic functional $\phi_L : \mathcal{H}_L \times \mathcal{G} \mapsto \mathbb{R}$ over $\mathcal{H}_L \times \widehat{F}_n$, where $\times$ denotes the Cartesian product, and the pdf of mixture models, instead of the corresponding parameters, is used as the arguments to ease the presentation. As L varies, $\phi_L(\cdot,\cdot)$ differs only in its domain. This procedure includes the existing estimations such as MLE as special cases, where $\phi_L(\cdot,\widehat{F}_n)$ is the log-likelihood function $n^{-1}\sum_{i=1}^{n}\log\{\sum_{l=1}^{L}\pi_l f(X_i,\mu_l)\}$ for each working L.

For regularity, we assume that, for each positive integer L and each $F \in \mathcal{G}$, there always exists the unique maximizer of $\phi_L(\cdot,F)$. This imposes restrictions such as the boundedness of $\phi_L(\cdot,F)$, which exclude certain scenarios from our discussion, such as the log-likelihood for mixture of multivariate normal distributions with free covariance matrices and unrestricted parameter space. Generally, $\phi_L(\cdot,\cdot)$ can be treated as a measure of similarity between two distributions, in which sense maximizing $\phi_L(\cdot,\widehat{F}_n)$ over $\mathcal{H}_L$ amounts to finding the element of $\mathcal{H}_L$ that best approximates to the observed data. Because a rich literature has been developed for implementing MLE via the EM algorithm, etc. [20], we omit the relative details and consider the maximizer of $\phi_L(\cdot,\widehat{F}_n)$ as granted throughout the theoretical development.

In addition, for the consistency of estimation, we adopt the following regularity conditions.

(C1) If F is a continuous distribution with its pdf f falls in $\mathcal{H}_L$, then f is the unique maximizer of $\phi_L(\cdot,F)$.

(C2) For each $F \in \mathcal{G}$, $\phi_L(\cdot,F)$ is continuous everywhere on $\mathcal{H}_L$. For any compact set $A \subset \mathcal{H}_L$, $\phi_L(f,\widehat{F}_n) - \phi_L(f,F_0)$ converges to zero in probability uniformly for $f \in A$.

(C3) For any $L \geq K$ and any sequences $\{f_n \in \mathcal{H}_L : n = 1,\ldots\}$ converging to f_0, we have

$$\phi_L(f_n,\widehat{F}_n) = \phi_L(f_0,F_0) + \int_{\mathbb{R}^p} H_f(x)\{f_n(x) - f_0(x)\}^2 dx$$
$$+ \int_{\mathbb{R}^p} K_F(x)\{d\widehat{F}_n(x) - f_0(x)dx\} + O(n^{-1/2}\|f_n - f_0\|_2 + n^{-1}) + o(\|f_n - f_0\|_2^2),$$

where $K_F(\cdot)$ is square integrable under F_0, $\int_{\mathbb{R}^p} K_F(x)d\widehat{F}_n(x)$ denotes the sample mean $n^{-1}\sum_{i=1}^n K_F(X_i)$, and $H_f(\cdot) < -\delta$ for some $\delta > 0$ almost everywhere on $\mathbb{R}^p$.

Condition (C3) assumes a second-order von Mises expansion of $\phi_L(\cdot,\cdot)$ at (f_0, F_0), which essentially permits a quadratic approximation of $\phi_L(f_n, \widehat{F}_n)$. The von Mises expansion is a generalization of the usual Taylor expansion to the functionals, where the usual derivative is replaced with the Gateaux derivative; see [9, Chapter 3] and [25, Chapter 6] for more details. The term that corresponds to the first-order partial Gateaux derivative with respect to $f \in \mathcal{H}_L$ is missing in this expansion because, as indicated by (C1), f_0 is the unique maximizer of $\phi_L(\cdot, F_0)$. The generality of these conditions can also be seen if we set $\phi_L(\cdot, \widehat{F}_n)$ to be the log-likelihood function, in which case (C3) reduces to the commonly adopted smoothness condition for the consistency of MLE [21] if we instead use ω_0 as the working argument.

Let $\widehat{f}_L$ be the unique maximizer of $\phi_L(\cdot, \widehat{F}_n)$. When $L \geq K$, recall that f_0 falls in $\mathcal{H}_L$. Thus, (C1) implies that f_0 is the unique maximizer of $\phi_L(\cdot, F_0)$. The closeness between $\widehat{F}_n$ and F_0 implies the closeness between $\phi_L(\cdot, \widehat{F}_n)$ and $\phi_L(\cdot, F_0)$ under (C2), which then further implies the closenesses between their maximizers $\widehat{f}_L$ and f_0, and between their maxima $\phi_L(\widehat{f}_L, \widehat{F}_n)$ and $\phi_L(f_0, F_0)$, under (C3). Note that, in this case, $\phi_L(f_0, F_0)$ is invariant of L and is always equal to $\phi_K(f_0, F_0)$. By contrast, when $L < K$, the fact that f_0 has a positive distance with $\mathcal{H}_L$ implies that f_0 must also have a significant distance with $\widehat{f}_L$, which makes $\phi_L(\widehat{f}_L, \widehat{F}_n)$ significantly less than $\phi_K(f_0, F_0)$. These suggest an elbow shape of $-\phi_L(\widehat{f}_L, \widehat{F}_n)$, which conforms to the literature mentioned earlier. For convenience, we call $-\phi_L(\widehat{f}_L, \widehat{F}_n)$ a goodness-of-fit measure. To formulate these intuitions, we introduce the concept of $O_P^+(1)$ for a "large" stochastic sequence; that is, a sequence of random variables $\{Z_n : n = 1, \ldots, \}$ is $O_P^+(1)$ if $Z_n > \delta$ with probability converging to one for some $\delta > 0$. We refer to [15] for more details about this concept.

Theorem 1. *Under Conditions (C1)-(C3), we have*

(i) $\|\widehat{f}_L - f_0\|_2 = O_P^+(1)$ *and* $\phi_L(\widehat{f}_L, \widehat{F}_n) = \phi_K(f_0, F_0) - O_P^+(1)$ *when* $L < K$;
(ii) $\|\widehat{f}_L - f_0\|_2 = O_P(n^{-1/2})$ *and* $\phi_L(\widehat{f}_L, \widehat{F}_n) = \phi_K(f_0, F_0) + O_P(n^{-1/2})$ *when* $L \geq K$.

Proof. For any $L < K$, because $\min_{f \in \mathcal{H}_L} \|f - f_0\|_2 > 0$, we have $\|\widehat{f}_L - f_0\|_2 = O_P^+(1)$. Thus, (i) follows immediately by (C1) and (C2). For any $L \geq K$, by (C1), we have

$$\phi_L(f_0, F_0) > \phi_L(\widehat{f}_L, F_0), \tag{3}$$

and, by (C3), we have $\phi_L(f_0, F_0) - \phi_L(f_0, \widehat{F}_n) = -\int_{\mathbb{R}^p} K_F(x)\{d\widehat{F}_n(x) - f_0(x)dx\} + O_P(n^{-1})$. By the definition of $\widehat{f}_L$, we have $\phi_L(\widehat{f}_L, \widehat{F}_n) > \phi_L(f_0, \widehat{F}_n)$. These to-

gether imply

$$\phi_L(\widehat{f}_L, \widehat{F}_n) > \phi_L(f_0, F_0) + \int_{\mathbb{R}^p} K_F(x)\{d\widehat{F}_n(x) - f_0(x)dx\} + \epsilon_n, \tag{4}$$

where ϵ_n is a sequence of random variables with $\epsilon_n > 0$ and $\epsilon_n = O_P(n^{-1})$. Let $A_\delta = \{f \in \mathcal{H}_L : \|f - f_0\|_2 \le \delta\}$. Then A_δ is a compact set in $\mathcal{H}_L$ under the L_2 distance. Given $\widehat{f}_L \in A_\delta$, (C2) implies $\phi_L(\widehat{f}_L, \widehat{F}_n) = \phi_L(\widehat{f}_L, F_0) + o_P(1)$, which, together with (3) and (4), implies

$$\phi_L(\widehat{f}_L, F_0) = \phi_L(f_0, F_0) + o_P(1).$$

By (C2) again, this implies $\|\widehat{f}_L - f_0\|_2 = o_P(1)$ given $\widehat{f}_L \in A_\delta$. By letting $\delta \to \infty$, we have $\|\widehat{f}_L - f_0\|_2 = o_P(1)$. By (C3), this implies

$$\begin{aligned}\phi_L(\widehat{f}_L, \widehat{F}_n) &= \phi_L(f_0, F_0) + \int_{\mathbb{R}^p} H_0(x)\{\widehat{f}_L(x) - f_0(x)\}^2 dx \\ &+ \int_{\mathbb{R}^p} K_F(x)\{d\widehat{F}_n(x) - f_0(x)dx\} + O_P(n^{-1/2}\|\widehat{f}_L - f_0\|_2 + n^{-1}) + o(\|\widehat{f}_L - f_0\|_2^2)\end{aligned}$$

where $H_0(\cdot)$ satisfies $H_f(\cdot) < -\delta$ almost everywhere for some $\delta > 0$. Together with (4), we have

$$\delta\|\widehat{f}_L - f_0\|_2^2 - O_P(n^{-1/2}\|\widehat{f}_L - f_0\|_2 + n^{-1}) - o(\|\widehat{f}_L - f_0\|_2^2) \le \epsilon_n = O_P(n^{-1}), \tag{5}$$

which immediately implies $\|\widehat{f}_L - f_0\|_2 = O_P(n^{-1/2})$. By plugging this result in (C3), we have ($\ddot{u}$). This completes the proof. □

Let $\widehat{\omega}_L = \{(\widehat{\pi}_{L,l}, \widehat{\mu}_{L,l}) : l = 1, \ldots, L\}$ be the estimator of ω_0 induced from $\widehat{f}_L$. When $L > K$, the closeness between $\widehat{f}_L$ and f_0 means that either some $\widehat{\pi}_{L,l}$ is asymptotically negligible, potentially leading to a large variation of the corresponding $\widehat{\mu}_{L,l}$, or some $\mu_{k,0}$ is approximated by multiple $\widehat{\mu}_{L,l}$'s, where the index k is potentially random. In both cases, there will be a large variation of the cluster estimates $\{\widehat{\mu}_{L,1}, \ldots, \widehat{\mu}_{L,L}\}$. As mentioned in Section 1, this was heuristically used in [8] and [11], etc., to conduct order determination for distance-based clustering. We next seek for a data augmentation approach, under which a variation pattern of the cluster estimates can be rigorously built under the general settings.

3. Data augmentation and clustering instability

The concept of data augmentation has been employed in the literature of variable selection [2, 3] and dimension reduction [16], where augmentation means increasing the dimension of the predictor by generating new random variables, often called the knock-off variables. The augmentation in our context is fundamentally different in the sense that we enlarge n rather than enlarge p; that is, we artificially generate a new cluster and merge it with the original data.

For simplicity, we generate the sample observations in the new cluster independently and identically from some pdf $f(\cdot, \mu_{A,0})$ that falls in the same parametric family $\mathcal{H}_1$. Let m denote the sample size of the new cluster and $W_1, \ldots, W_m$ be the corresponding observations. Let $\pi_{A,0}$ denote m/n, where m and n are omitted from its subscript. For any observation in the augmented data set $\{X_1, \ldots, X_n, W_1, \ldots, W_m\}$, if we pretend not knowing whether it is from the original data or from the new cluster, then, by averaging out the cluster indexes, it will have pdf

$$f_0^*(x) = \sum_{k=1}^{K} \pi_{k,0}^* f(x, \mu_{k,0}) + \pi_{A,0}^* f(x, \mu_{A,0}), \tag{6}$$

where $\pi_{k,0}^* = \pi_{k,0}/(1 + \pi_{A,0})$ for $k = 1, \ldots, K$ and $k = A$. To guarantee the identifiability of Model (6), we emphasize here that $\mu_{A,0}$ must be chosen to differ from any of $\mu_{1,0}, \ldots, \mu_{K,0}$, or equivalently that the new cluster must be relatively separate from the main data clouds.

By definition, the augmented data have $K + 1$ clusters. If we let $\pi_{A,0}$ vanish to zero as n grows, then the new cluster is not as "important" as the original clusters; that is, the loss of information is asymptotically negligible if the newly added cluster is not recovered in clustering. This suggests that, if we still specify K to be the working number of clusters for clustering on the augmented data, then we will tend to end up with recovering the K original clusters as if the new cluster were not involved. By contrast, if we specify any $L > K$ to be the working number of clusters, then as long as the gain of recovering the new cluster dominates the gain of over-fitting the original data (though both are asymptotically negligible), we will end up with recovering both the original clusters and the new cluster. Intuitively, this will happen if $\pi_{A,0}$ vanishes at a relatively slow rate. Putting these together, it is natural to imagine that the data augmentation has a negligible impact on the estimation of individual clusters when $L = K$, and this impact will be elevated to be statistically significant when L increases to $K + 1$, i.e. by recovering the new cluster, and it will stay significantly large for all $L > K$.

Based on the discussion in Section 2, the speculation above can be rigorously justified by a slight modification of Theorem 1(*ii*) for the augmented data. Let $\widehat{F}_{n+m}^*$ denote the empirical distribution induced by the augmented data $\{X_1, \ldots, X_n, W_1, \ldots, W_m\}$, which, in a rough sense, deviates from f_0 by $O_P(\pi_{A,0} + n^{-1/2})$ and deviates from the augmented pdf f_0^* by $O_P(n^{-1/2})$. Let $\widehat{f}_L^*(\cdot) = \sum_{l=1}^{L} \widehat{\pi}_{L,l}^* f(\cdot, \widehat{\mu}_{L,l}^*)$ be the unique maximizer of $\phi_L(\cdot, \widehat{F}_{n+m}^*)$. We call $\widehat{f}_L^*$ the augmented pdf estimate in contrast to the original pdf estimate $\widehat{f}_L$. A parallelization of Theorem 1 implies the following result, proof omitted.

Theorem 2. *Under Conditions (C1)-(C3), we have*

(i) $\|\widehat{f}_K^* - f_0\|_2 = O_P(\pi_{A,0} + n^{-1/2})$ *if* $\pi_{A,0} \to 0$ *as* $n \to \infty$.
(ii) $\|\widehat{f}_L^* - f_0^*\|_2 = O_P(n^{-1/2})$ *for each* $L > K$.

Theorem 2(*i*) implies that, when K is correctly determined and $\pi_{A,0}$ vanishes, each of the K clusters of the augmented estimate $\widehat{f}_K^*$ must consistently recover

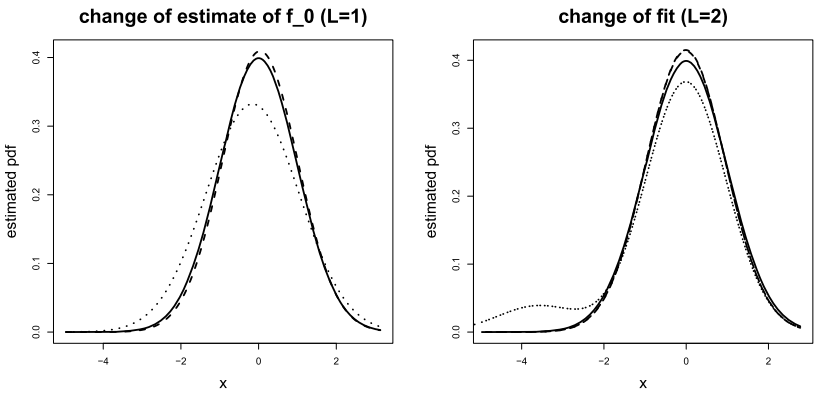

FIG 1. *The change of the impact of data augmentation on the estimation of f_0 as L varies: the original data are generated from the univariate standard normal distribution with sample size $n = 400$ and the new cluster is generated from $N(3.5, 1)$ with sample size 42. The left panel displays the case of $L = 1$ and the right panel is for $L = 2$. In each panel, the dashed curve displays $\widehat{f}_L$ derived from the original data, and the dotted curve displays $\widehat{f}^*_L$ derived from the augmented data.*

one of those of f_0, which means that $\widehat{f}^*_K$ must not recover the newly added cluster and behaves similarly to the original estimate $\widehat{f}_K$ derived before data augmentation. When X is univariate or bivariate, in which case f_0, $\widehat{f}_K$ and $\widehat{f}^*_K$ can be depicted, one will observe that $\widehat{f}_K$ and $\widehat{f}^*_K$ have the same shape and both are close to f_0, with $\widehat{f}^*_K$ being slightly more biased due to the data augmentation. This is revealed in the left panel of Figure 1, where we draw the graphs of f_0, $\widehat{f}_K$, and $\widehat{f}^*_K$ with the original data generated from the univariate standard normal distribution with sample size $n = 400$, i.e. with $K = 1$, and the new cluster generated from $N(3.5, 1)$ with sample size 42, i.e. $\pi_{A,0} \approx n^{-.375}$.

By contrast, whenever $L > K$, Theorem 2($\ddot{u}$) implies that the augmented estimate $\widehat{f}^*_L$ must recover the newly added cluster up to an error of $O_P(n^{-1/2})$, as long as $\pi_{A,0}$ vanishes at a slower rate than $n^{-1/2}$. Thus, $\widehat{f}^*_L$ would differ from the original estimate $\widehat{f}_L$ in the sense that it has at least one extra cluster that has small size but is away from the clusters of $\widehat{f}_L$. An example of this case is depicted in the right panel of Figure 1, where both the original and the augmented data are generated in the same way as in the previous paragraph but the working L for clustering is increased to two. If the newly added cluster has an excessively small size, i.e. $\pi_{A,0} = O(n^{-1/2})$, then $\widehat{f}^*_L$ needs not recover the new cluster, even though it still deviates from the augmented pdf f^*_0 by an error of $O_P(n^{-1/2})$. Overall, Theorem 2 justifies the change of behavior of $\widehat{f}^*_L$ when L increases from K to larger values.

We now construct a measure of deviation from $\widehat{f}^*_L$ to $\widehat{f}_L$ that amplifies their difference when $L > K$. For each working number of clusters L, let $\widehat{\omega}^*_L = \{(\widehat{\pi}^*_{L,l}, \widehat{\mu}^*_{L,l}) : l = 1, \ldots, L\}$ be the parameter estimate induced by $\widehat{f}^*_L$. Denote the element of $\{\widehat{\mu}^*_{L,1}, \ldots, \widehat{\mu}^*_{L,L}\}$ that is closest to $\mu_{A,0}$ by $\widehat{\mu}^*_{L,A}$, i.e. $\|\widehat{\mu}^*_{L,A} - \mu_{A,0}\|_2 =$

$\min_{l=1,\ldots,L} \|\widehat{\mu}^*_{L,l} - \mu_{A,0}\|_2$. We introduce

$$\tau(\widehat{f}_L, \widehat{f}^*_L) = \min\{\|\widehat{\mu}_{L,l} - \widehat{\mu}^*_{L,A}\|_2 : l = 1, \ldots, L, \widehat{\pi}_{L,l} > \pi_{A,0}\}. \tag{7}$$

The reason that we adopt the threshold $\widehat{\pi}_{L,l} > \pi_{A,0}$ in this measure is to exclude the noises caused by those cluster estimates in $\widehat{f}_L$ who have asymptotically negligible sizes, for which the behavior of the corresponding $\widehat{\mu}_{L,l}$ is intractable. To avoid the case that no cluster estimates satisfy $\widehat{\pi}_{L,l} > \pi_{A,0}$, which may occur if the sample size is limited and L is large, in practice, we set $\pi_{A,0}$ adaptively to L, e.g. proportional to L^{-1}, such that it is always less than L^{-1}; see Section 4 for a relative discussion and Section 5 for more details. Based on Theorem 2 and the discussions above, $\widehat{\mu}^*_{L,A}$ must consistently estimate $\mu_{A,0}$ when $L > K$, and it instead converges to one of $\mu_{1,0}, \ldots, \mu_{K,0}$ when $L = K$. Together with Theorem 1, $\tau(\widehat{f}_L, \widehat{f}^*_L)$ essentially detects whether $\widehat{\mu}^*_{L,A}$ recovers the newly added cluster or one of the original clusters, and it should be negligible when $L = K$ and jumps to be significantly positive when L is larger. These are justified in the next corollary.

Corollary 1. *Let* $\tau(f_0, f^*_0) = \min\{\|\mu_{k,0} - \mu_{A,0}\|_2 : k = 1, \ldots, K\}$. *Under Conditions (C1)-(C3), we have*

(i) $\tau(\widehat{f}_K, \widehat{f}^*_K) = O_P(\pi_{A,0} + n^{-1/2})$ *if* $\pi_{A,0} \to 0$ *as* $n \to \infty$*;*
(ïi) $\tau(\widehat{f}_L, \widehat{f}^*_L) = \tau(f_0, f^*_0) + O_P(n^{-1/2}/\pi_{A,0})$ *if* $L > K$ *and* $n^{-1/2}/\pi_{A,0} \to 0$ *as* $n \to \infty$.

Proof. Under the identifiability of the representation of f_0, Theorem 2(*i*) implies that, for each $k = 1, \ldots, K$, there must exist $l(k) \in \{1, \ldots, K\}$ such that $\|\widehat{\pi}^*_{K,k} f(\cdot, \widehat{\mu}^*_{K,k}) - \pi_{l(k),0} f(\cdot, \mu_{l(k),0})\|_2 = O_P(\pi_{A,0} + n^{-1/2})$, and that $(l(1), \ldots, l(K))$ is a permutation of $(1, \ldots, K)$. Equivalently, we have $\widehat{\pi}^*_{K,k} = \pi_{l(k),0} + O_P(\pi_{A,0} + n^{-1/2})$ and $\widehat{\mu}^*_{K,k} = \mu_{l(k),0} + O_P(\pi_{A,0} + n^{-1/2})$ for $k = 1, \ldots, K$. For simplicity, we assume that there exists the unique $\mu_{C,0} \in \{\mu_{k,0} : k = 1 \ldots, K\}$ that is closest to $\mu_{A,0}$ among all the $\mu_{k,0}$'s, i.e. $\|\mu_{C,0} - \mu_{A,0}\|_2 = \min_{k=1,\ldots,K} \|\mu_{k,0} - \mu_{A,0}\|_2$. These together indicate

$$\widehat{\mu}^*_{K,A} = \mu_{C,0} + O_P(\pi_{A,0} + n^{-1/2}), \tag{8}$$

By applying the same arguments to Theorem 1(*ïi*) with $L = K$, we have $\widehat{\pi}_{K,k} = \pi_{h(k),0} + O_P(n^{-1/2})$ and $\widehat{\mu}_{K,k} = \mu_{h(k),0} + O_P(n^{-1/2})$ for each $k = 1, \ldots, K$, where $(h(1), \ldots, h(K))$ is a permutation of $(1, \ldots, K)$. These together imply

$$\begin{aligned} &\min\{\|\widehat{\mu}_{K,k} - \mu_{C,0}\|_2 : k = 1, \ldots, K, \widehat{\pi}_{K,k} > \pi_{A,0}\} \\ &\quad = \min\{\|\mu_{h(k),0} - \mu_{C,0}\|_2 : k = 1, \ldots, K\} + O_P(n^{-1/2}) = O_P(n^{-1/2}). \end{aligned} \tag{9}$$

Together with (8), we have $\tau(\widehat{f}_K, \widehat{f}^*_K) = O_P(\pi_{A,0} + n^{-1/2})$, which is statement (*i*) of the theorem.

For $L > K$, Theorem 2(*ïi*) implies that there is a subset $I(A)$ of $\{1, \ldots L\}$ such that $\| \sum_{l \in I(A)} \widehat{\pi}^*_{L,l} f(\cdot, \widehat{\mu}^*_{L,l}) - \pi_{A,0} f(\cdot, \mu_{A,0})\|_2 = O_P(n^{-1/2})$, which means

$$\textstyle\sum_{l \in I(A)} \widehat{\pi}^*_{L,l} = \pi_{A,0} + O_P(n^{-1/2}), \tag{10}$$

$$\sum_{l\in I(A)} \widehat{\pi}^*_{L,l}\|\widehat{\mu}^*_{L,l} - \mu_{A,0}\|_2 = O_P(n^{-1/2}). \tag{11}$$

Let $D \in I(A)$ be such that $\widehat{\pi}^*_{L,D} = \max_{l\in I(A)} \widehat{\pi}^*_{L,l}$. Then (10) implies $\widehat{\pi}^*_{L,D} \geq \pi_{A,0}/L$, which, together with (11), implies $\|\widehat{\mu}^*_{L,D} - \mu_{A,0}\|_2 = O_P(n^{-1/2}/\pi_{A,0})$. By the definition of $\widehat{\mu}^*_{L,A}$, we have

$$\|\widehat{\mu}^*_{L,A} - \mu_{A,0}\|_2 = O_P(n^{-1/2}/\pi_{A,0}). \tag{12}$$

Similarly to (10) and (11), Theorem 1($\ddot{u}$) implies the existence of $(I(1),\ldots,I(K))$ that partitions $(1,\ldots,L)$, such that, for each $k = 1,\ldots,K$,

$$\sum_{l\in I(k)} \widehat{\pi}_{L,l} = \pi_{k,0} + O_P(n^{-1/2}), \tag{13}$$

$$\sum_{l\in I(k)} \widehat{\pi}_{L,l}\|\widehat{\mu}_{L,l} - \mu_{k,0}\|_2 = O_P(n^{-1/2}). \tag{14}$$

For each $k = 1,\ldots,K$, (13) implies that there must exist $l \in I(k)$ such that $\widehat{\pi}_{L,l} \geq \pi_{k,0}/L + o_P(1)$, which, by $\pi_{A,0} = o(1)$, means $\widehat{\pi}_{L,l} > \pi_{A,0}$ in probability. Thus, without loss of generality, we can assume that, for each $k = 1,\ldots,K$, there always exists $l \in I(k)$ such that $\widehat{\pi}_{L,l} > \pi_{A,0}$. In addition, (14) implies that for each $l \in I(k)$, $\|\widehat{\mu}_{L,l} - \mu_{k,0}\|_2 = O_P(n^{-1/2})/\widehat{\pi}_{L,l}$, which is $O_P(n^{-1/2}/\pi_{A,0})$ if $\widehat{\pi}_{L,l} > \pi_{A,0}$. Together with (12) and that $(I(1),\ldots,I(K))$ is a partition of $(1,\ldots,L)$, we have

$$\begin{aligned}\tau(\widehat{f}_L, \widehat{f}^*_L) &= \min_{k=1,\ldots,K} \min\{\|\widehat{\mu}_{L,l} - \mu_{A,0}\|_2 + O_P(n^{-1/2}/\pi_{A,0}) : l \in I(k), \widehat{\pi}_{L,l} > \pi_{A,0}\} \\ &= \min\{\|\mu_{k,0} - \mu_{A,0}\|_2 + O_P(n^{-1/2}/\pi_{A,0}) : k = 1,\ldots,K\} \\ &= \tau(f_0, f^*_0) + O_P(n^{-1/2}/\pi_{A,0}),\end{aligned}$$

which is statement ($\ddot{u}$) of the theorem. This completes the proof. □

We emphasize here that $\tau(\widehat{f}_L, \widehat{f}^*_L)$ is not the only measure that one can use to represent the impact of data augmentation on clustering: theoretically, Corollary 1 also applies for other measures as long as they can detect the sudden change of cluster estimates caused by data augmentation when L jumps from K to larger values. The two statements in Corollary 1 together require $\pi_{A,0}$ to vanish but in a slower rate than $n^{-1/2}$, which complies with Theorem 2 and the relative discussions above. We will study its convergence rate in more details in the simulation studies.

It is important to note that the data augmentation can be regarded as a negligible perturbation of the original data, in the sense that the augmented empirical distribution $\widehat{F}^*_{m+n}$ differs from its original counterpart $\widehat{F}_n$ negligibly. In this spirit, Corollary 1 justifies the instability of clustering as a characterization of K; that is, the clustering is stable when K is truly specified and is instable when K is overestimated.

As mentioned in Section 1, the instability of clustering has only been used heuristically for distance-based clustering in the literature. From the discussion below Theorem 2 and also the proof of Corollary 1, we speculate that a data

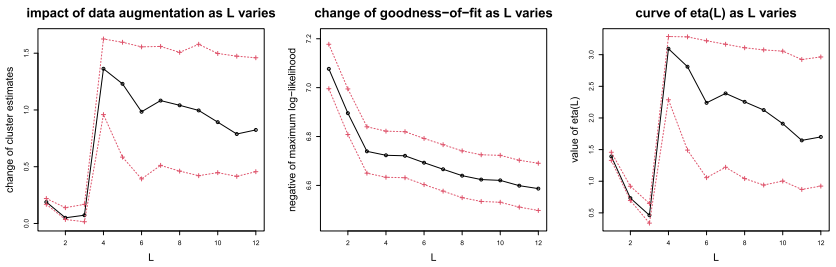

FIG 2. *The left panel depicts the change of the impact of data augmentation on clustering as L varies: the underlying model is Model 4 in Section 6 with $K = 3$; the solid curve displays the sample mean of (7) as L varies and the two dashed curves display its* 2.5% *and* 97.5% *sample quantiles, respectively, based on* 2000 *independent samples. The middle panel and the right panel depict the same curves for the negative maximum log-likelihood and for $\eta(L)$ as L varies, respectively, for the same model; see Section 4 later for detail.*

perturbation procedure that leads to a justifiable pattern of instability of clustering must satisfy two conditions. First, it perturbs the data by an larger error than $O(n^{-1/2})$. Second, the distribution for the perturbed data must have a significantly different parameter from the original distribution F_0. These conditions exclude the cross-validation and bootstrap approaches, and possibly make data augmentation exclusive for characterizing the instability of clustering.

To illustrate the results of Corollary 1, we generate the original data under Model 4 in the simulation studies later, which consist of three clusters, and generate the new cluster using the method suggested in Section 5. With n set at 400, the left panel of Figure 2 displays the summary curves of $\tau(\widehat{f}_L, \widehat{f}_L^*)$, i.e. its mean and its 95% pointwise confidence interval, approximated based on 2000 independent runs. Clearly, the function vanishes at three, i.e. the true value of K, and it suddenly jumps to a large value afterwards. For this model, $\tau(\widehat{f}_L, \widehat{f}_L^*)$ also vanishes when $L < K$.

Generally, when $L < K$, there is no simple answer on whether data augmentation will cause significant change of the cluster estimates. For example, if $\phi_L(\cdot, F_0)$ has the unique maximizer or equivalently that the clustering is asymptotically stable, then $\widehat{\omega}_L^*$ will only slightly modify $\widehat{\omega}_L$ due to the closeness between $\widehat{F}_{n+m}^*$ and $\widehat{F}_n$. In particular, if we set L at one, then $\widehat{\mu}_1$ and $\widehat{\mu}_1^*$ will be the same statistical functional of $\widehat{F}_n$ and $\widehat{F}_{n+m}^*$, respectively, which certainly approximate to each other. By contrast, $\phi_L(\cdot, F_0)$ can have multiple maximizers when certain symmetry exists in F_0. For example, suppose F_0 consists of four bivariate normal distributions with the covariance matrices uniformly being the identity matrix and the set of means $\{\theta_{i,0} : i = 1, \ldots, 4\}$ being invariant of a 90-degree rotation of the coordinate system. Then F_0 is invariant of a 90-degree rotation as well, which means that any maximizer of $\phi_L(\cdot, F_0)$ must still maximize this function after being rotated by 90 degrees. Since a mixture of two or three bivariate normal distributions must be different after being rotated by 90 degrees, the set of maximizers of $\phi_L(\cdot, F_0)$ must include multiple elements for both $L = 2$ and

$L = 3$. In these cases, $\widehat{\omega}_L$ may not be convergent, and the negligible disturbance we add to $\widehat{F}_n$ may cause a substantially different $\widehat{\omega}_L^*$ and thus deliver $\widehat{\mu}_{L,A}^*$ that differs dramatically from any of $\widehat{\mu}_{L,1}, \ldots, \widehat{\mu}_{L,L}$. Fortunately, as seen in the next section, the asymptotic behavior of $\tau(\widehat{f}_L, \widehat{f}_L^*)$ for $L < K$ is not needed in the proposed work.

Finally, it is only for the ease of theoretical development to generate the new cluster under the same parametric family as the original clusters. When it is inconvenient or infeasible to do so, the readers are free to generate the new cluster under other distributions. For example, when the original clusters have bounded or curved sample supports, the new cluster can still be generated from a multivariate normal distribution. Accordingly, $\phi_L(\cdot, \widehat{F}_{n+m}^*)$ needs to be adjusted, e.g. by allowing up to one cluster distribution to be multivariate normal if $\phi_L(\cdot, \widehat{F}_{n+m}^*)$ is the log-likelihood function. As long as $\pi_{A,0}$ is controlled to vanish at a slow rate, the same theory still applies following similar reasonings.

4. The data augmentation estimator

By Corollary 1, the change of cluster estimates before and after the data augmentation, as measured by $\tau(\widehat{f}_L, \widehat{f}_L^*)$, displays a pattern that characterizes the number of clusters K. As pointed out by an anonymous Referee, a variety of methods can be constructed based on $\tau(\widehat{f}_L, \widehat{f}_L^*)$ for consistent order determination. For example, one can estimate K as the largest L such that $\tau(\widehat{f}_L, \widehat{f}_L^*)$ is less than a prefixed threshold. Here, we choose to combine $\tau(\widehat{f}_L, \widehat{f}_L^*)$ with the goodness-of-fit measure $-\phi_L(\widehat{f}_L, \widehat{F}_n)$ to construct an objective function that tends to be minimized at K. The motivation is that, as $-\phi_L(\widehat{f}_L, \widehat{F}_n)$ has a sudden drop when L increases from $K-1$ to K and stays nearly flat afterwards, it conveys a scree plot that is compensative to the pattern of $\tau(\widehat{f}_L, \widehat{f}_L^*)$, so the two sources of information together can sharpen the order determination.

Let K_M be a prefixed upper bound of K. We define $\eta : \{1, \ldots, K_M\} \to \mathbb{R}$ as

$$\eta(L) = \frac{c \cdot \tau(\widehat{f}_L, \widehat{f}_L^*)}{\max_{l=1,\ldots,K_M} \tau(\widehat{f}_L, \widehat{f}_L^*)} + \frac{\phi_{K_M}(\widehat{f}_{K_M}, \widehat{F}_n) - \phi_L(\widehat{f}_L, \widehat{F}_n)}{\phi_{K_M}(\widehat{f}_{K_M}, \widehat{F}_n) - \phi_1(\widehat{f}_1, \widehat{F}_n)}, \tag{15}$$

where both terms on the right-hand side are normalized so that the function is invariant of stretches of data, and $c > 0$ controls the balance between the two terms. To address the case of $K = 1$, i.e. if the data only consist of one cluster, we further require $c > 1$. The explanation is deferred to later.

Clearly, the first term of $\eta(\cdot)$ is always non-negative. By definition, $\phi_L(\widehat{f}_L, \widehat{F}_n)$ is monotone increasing with L, so the second term of $\eta(\cdot)$ is also non-negative and bounded from above by one. By Theorem 1(i), when $L < K$, the second term of $\eta(L)$ is large, and so is $\eta(L)$. By Corollary 1($\ddot{u}$), when $L > K$, the first term of $\eta(L)$ is large, and so is $\eta(L)$ again. By contrast, Theorem 1($\ddot{u}$) and Corollary 1(i) together indicate the asymptotic negligibility of $\eta(K)$, as long as $K > 1$. Therefore, when $K > 1$, i.e. if there are indeed at least two clusters mixed in the original data, the minimizer of $\eta(\cdot)$ must converge to K

in probability. These are illustrated in Figure 2 in Section 3, where again the data are generated from Model 4 in the simulation studies with $n = 400$, and the same summary curves as for $\tau(\widehat{f}_L, \widehat{f}_L^*)$ in the left panel are depicted for $-\phi_L(\widehat{f}_L, \widehat{F}_n)$ and $\eta(L)$, respectively, in the middle and right panels. As $K = 3$ for this model, $\eta(\cdot)$ is clearly consistently minimized at K.

When $K = 1$, $\phi_L(\widehat{f}_L, \widehat{F}_n)$ is constantly negligible for all $L \in \{1, \ldots, K_M\}$, so both the numerator and the denominator of the second term of $\eta(L)$ are negligible, making this term intractable and potentially jeopardize the consistency of order determination. To address this issue, we need to modify the arguments above for $K = 1$. By Corollary 1, in this case, $\tau(\widehat{f}_L, \widehat{f}_L^*)$ will converge to zero when $L = 1$ and will converge to the positive constant $\tau(f_0, f_0^*)$ whenever $L > 1$. Thus, the first term of $\eta(L)$ will converge to zero when $L = 1$ and will converge to c whenever $L > 1$. By construction, the second term of $\eta(L)$ is always bounded by zero and one, which means that it is less than one when $L = 1$ and greater than zero when $L > 1$. These together imply

$$\eta(1) \leq 1 + \epsilon_n, \text{ and } \eta(L) \geq c + \epsilon_n' \text{ for all } L > 1,$$

for some random sequences $\{\epsilon_n : n = 1, \ldots\}$ and $\{\epsilon_n' : n = 1, \ldots\}$ that both are $o_P(1)$. Hence, $\eta(\cdot)$ tends to be minimized at $L = 1$ asymptotically as long as $c > 1$. For simplicity of implementation, we choose $c = 3$ uniformly in the numerical studies later in this paper. In practice, if the researchers can confirm that K must be greater than one before order determination, i.e. if there is clustering in the data, then they can confidently set c at one or a smaller positive value.

We summarize the discussions above in the next theorem. Its proof essentially resembles these discussions and thus is omitted.

Theorem 3. *Under Conditions (C1)-(C3), if we set $\pi_{A,0}$ appropriately such that $\pi_{A,0} + n^{-1/2}\pi_{A,0}^{-1} \to 0$ and set $c > 1$ in (15), then $\eta(\cdot)$ tends to be minimized uniquely at K; that is,*

$$\lim_{n\to\infty} P[\min\{\eta(L) : L = 1, \ldots, K_M, L \neq K\} > \eta(K)] = 1.$$

Let $\widehat{K}$ be the minimizer of $\eta(\cdot)$. By Theorem 3, $\widehat{K}$ is a consistent estimator of K. Because the estimation procedure is featured by data augmentation, we call $\widehat{K}$ the data augmentation estimator (DAE). Following the spirit of BIC, we can regard the first term of $\eta(\cdot)$ as a penalty function of L. Compared with its counterpart in BIC, this penalty function is data-driven, and has the fundamental advantage that it explores an intrinsic property of clustering and provides the unique characterization of K. For this reason, we expect DAE to be more effective than BIC in practice. In addition, because DAE does not involve asymptotic inference results such as the limiting distribution of $\widehat{\omega}_L$ or $\phi_L(\widehat{f}_L, \widehat{F}_n)$, it should be more robust to the parametric assumption $\mathcal{H}_1$ than the likelihood-based sequential tests. Finally, it is also easily implementable as it requires no additional calculation, e.g. arrangement of clusters for GSF in [18].

If we increase $\pi_{A,0}$, then there will be an increasing chance of recovering the new cluster for the augmented data or equivalently deriving a large value of $\tau(\widehat{f}_L, \widehat{f}_L^*)$, particularly when L is slightly less than K. Thus, the minimizer of $\eta(\cdot)$ will be shifted to the left probabilistically. This impact is significant if $\pi_{A,0}$ is non-vanishing, delivering an underbiased $\widehat{K}$. In practice, when n is limited and the working number of clusters L is relatively large, all the estimated clusters in the original data may have small sizes, which easily makes the size of the new cluster excessively large and the resulting $\widehat{K}$ underbiased even if $\pi_{A,0}$ vanishes asymptotically. Fortunately, with $\pi_{A,0}$ set adaptive to L as mentioned below (7) in Section 3 (also see Section 5 for more details), the size of the new cluster keeps shrinking as L increases, which helps stabilize the sample performance of $\widehat{K}$. The consistency of $\widehat{K}$ still applies for this adaptive choice of $\pi_{A,0}$.

A limitation of DAE is that its consistency can be jeopardized when the data are imbalanced, i.e. if the smallest cluster of the data is negligible compared with the largest. In this case, the desired pattern of $\tau(\widehat{f}_L, \widehat{f}_L^*)$ in Corollary 1 can easily fail, as it is implausible to find an appropriate size for the new cluster that is simultaneously small enough to be dominated by the smallest of the original clusters, and large enough to make the gain of fitting the new cluster dominate the gain of over-fitting the largest of the original clusters. By contrast, when the data are nearly balanced, we expect DAE to be consistent even when K is large, although all the clusters are small in this case. This is because, if we set $\pi_{A,0}$ to be n^{r-1}/L for some $r \in (1/2, 1)$, then it is dominated by K^{-1} when $L = K$ even if K is large, and it dominates $n^{-1/2}$ for $L > K$ as long as n is reasonably large with respect to L. Thus, $\tau(\widehat{f}_L, \widehat{f}_L^*)$ still conveys the desired pattern as in Corollary 1. These points will be re-visited in the simulation studies.

As mentioned below Theorem 2 in Section 3, when $\pi_{A,0}$ is too small, i.e. of order $O(n^{-1/2})$, the newly added cluster cannot make a notable change to the parameter estimation even when $L > K$. In this case, the first term in $\eta(\cdot)$ will display a flat curve and become useless to the order determination. To alleviate this issue, we incorporate the heuristics on the instability of clustering mentioned in Section 1 into the data augmentation procedure; that is, we construct the augmented data set by merging the new cluster with a bootstrap copy of the original data, i.e. $\{\widetilde{X}_1, \ldots, \widetilde{X}_n\}$ generated independently and identically from the empirical distribution $\widehat{F}_n$. Let $\widetilde{F}_{n+m}^*$ be the empirical distribution of the modified augmented data $\{\widetilde{X}_1, \ldots, \widetilde{X}_n, W_1, \ldots, W_m\}$, and, for each $L = 1, \ldots, K_m$, let $\widetilde{f}_L^*(\cdot) = \sum_{l=1}^{L} \widetilde{\pi}_{L,l}^* f(\cdot, \widetilde{\mu}_{L,l}^*)$ be the maximizer of $\phi_L(\cdot, \widetilde{F}_{n+m}^*)$ and $\widetilde{\mu}_{L,A}^*$ be the element of $\{\widetilde{\mu}_{L,1}^*, \ldots, \widetilde{\mu}_{L,L}^*\}$ such that $\|\widetilde{\mu}_{L,A}^* - \mu_{A,0}\|_2 = \min_{l=1,\ldots,L} \|\widetilde{\mu}_{L,l} - \mu_{A,0}\|_2$. We propose $\widetilde{K}$ as the minimizer of $\widetilde{\eta} : \{1, \ldots, K_M\} \to \mathbb{R}$ with

$$\widetilde{\eta}(L) = \frac{c \cdot \tau(\widehat{f}_L, \widetilde{f}_L^*)}{\max_{l=1,\ldots,K_M} \tau(\widehat{f}_l, \widetilde{f}_l^*)} + \frac{\phi_{K_M}(\widehat{f}_{K_M}, \widehat{F}_n) - \phi_L(\widehat{f}_L, \widehat{F}_n)}{\phi_{K_M}(\widehat{f}_{K_M}, \widehat{F}_n) - \phi_1(\widehat{f}_1, \widehat{F}_n)}, \tag{16}$$

and call this estimator DAE-Ⅱ. Under the conditions in Theorem 3, $\widetilde{K}$ also consistently recovers K. The heuristics on the instability of clustering suggests that the bootstrap re-sampling has a chance to provide additional deviation

between the cluster estimates for small samples when $L > K$, and thus sharpens the desired pattern of $\tau(\widehat{f}_L, \widehat{f}_L^*)$. Thus, DAE-Ⅱ may outperform DAE in terms of the robustness against excessively small $\pi_{A,0}$, when the sample size is limited.

5. Details of implementation

We now summarize the implementation of DAE as follows. DAE-Ⅱ can be implemented in the same way, except that the original data need to be replaced with a bootstrap sample generated independently from the empirical distribution $\widehat{F}_n$ in the following Step 2. Recall that $\pi_{A,0}$ is defined as m/n where m is the size of the new cluster. Hence, the requirement that $\pi_{A,0}$ vanishes in a slower rate than $n^{-1/2}$ is equivalent to that m diverges faster than $n^{1/2}$ but slower than n. For each working number of clusters L, we take m to be the nearest integer to n^r/L for some $r \in (1/2, 1)$.

Step 1 For each $L = 1, \ldots, K_M$, maximize $\phi_L(\cdot, \widehat{F}_n)$ over $\mathcal{H}_L$ to derive $\widehat{f}_L$. Calculate $\phi_L(\widehat{f}_L, \widehat{F}_n)$.

Step 2 For each $L = 1, \ldots, K_M$, set m to be nearest integer to n^r/L for some $r \in (1/2, 1)$; generate m independent observations from the pdf $f(\cdot, \mu_{A,0})$ and merge these observations into the original data to form the augmented empirical distribution $\widehat{F}^*_{n+m}$; maximize $\phi_L(\cdot, \widehat{F}^*_{n+m})$ over $\mathcal{H}_L$ to derive $\widehat{f}_L^*$ and calculate $\tau(\widehat{f}_L, \widehat{f}_L^*)$.

Step 3 Calculate $\eta(\cdot)$ as in (15) and minimize $\eta(\cdot)$ to derive $\widehat{K}$.

When $\phi_L(\cdot, \widehat{F}_n)$ is non-concave if regarded as a function of ω and thus is hard to maximize, we suggest running the numerical algorithm for multiple times with different initial values, and choosing the optimal result that delivers the maximal value of $\phi_L(\cdot, \widehat{F}_n)$. Theoretically, the proposed estimators must be robust against the selection of K_M, as long as $K_M \geq K$. This is because when L is too large, the excessive number of parameters will cause a large estimation bias in both $\widehat{f}_L$ and $\widehat{f}_L^*$, which naturally leads to a large $\tau(\widehat{f}_L, \widehat{f}_L^*)$ and consequently a large $\eta(L)$. To help avoid $K_M < K$, one can adopt an adaptive strategy in practice: for a given K_M, if the proposed estimators select K to be K_M, then we increase K_M until it is sufficiently large.

To choose the value of $\mu_{A,0}$, we suggest two principles. First, the new cluster should be reasonably separate from the original data to ensure a large deviation between $\mu_{A,0}$ and $\{\mu_{1,0}, \ldots, \mu_{K,0}\}$, so that the change of cluster estimates can be fully represented in $\tau(\widehat{f}_L, \widehat{f}_L^*)$ when $L > K$. Second, $\mu_{A,0}$ should make the resulting $\widehat{f}_L^*$ robust to both the choice of $\pi_{A,0}$ and the randomness of the observations in the new cluster for all L, so that the proposed estimators are reliable to use in practice. When $\mu_{A,0}$ includes the mean $\theta_{A,0}$ of the new cluster, we speculate that these principles require $\theta_{A,0}$ to be distant from the original data in the directions that the latter vary the least. Namely, let $\bar{X}_n$ and S_n be the sample mean and sample covariance matrix of the original data, respectively, and let $U_n \in \mathbb{R}^{p \times (p-1)}$ be set of eigenvectors of S_n associated with its largest

$p-1$ eigenvalues and V_n be the eigenvector of S_n associated with its smallest eigenvalue σ_n^2. We suggest

$$\theta_{A,0} = U_n U_n^\intercal \bar{X}_n + V_n(V_n^\intercal \bar{X}_n + 2\sigma_n). \tag{17}$$

In particular, if the clusters are known (or assumed) to have multivariate normal distributions with common but unspecified covariance matrix Σ_0, then, for each working number of clusters L, we generate the new cluster from $N(\theta_{A,0}, \widehat{\Sigma}_L)$ with $\theta_{A,0}$ specified in (17) and $\widehat{\Sigma}_L$ being the estimate of Σ_0 in $\widehat{f}_L$.

6. Simulation studies

We now use simulation examples to illustrate the effectiveness of the proposed estimators, in comparison with the aforementioned GSF [18] equipped a variety of penalty functions, including SCAD [7], MCP [35], and Adaptive Lasso [36], as well as AIC and BIC [5] that were shown most effective among all the existing information criteria in [18]. These methods are available from the `R` packages `GroupSortFuse` [18], `mixtools` [4], and `mclust` [10], respectively. In addition, we will study how the performance of the proposed estimators is affected by both the size of the new cluster and the pre-fixed upper bound K_M.

For a fair comparison, we adopt the following simulation models listed in the ascending order of the number of clusters they induce, among which Models 3, 4, 5, 6, and 7 were also adopted in [18]. Let 0_p be the origin of $\mathbb{R}^p$ and I_p be the p-dimensional identity matrix, and, for any nonzero scalar a, let $(a^{|i-j|})_p$ be the p-dimensional square matrix whose (i,j)th entry is $a^{|i-j|}$. In Model 5, we generate discrete data where each cluster has a multinomial distribution with 50 trials in total, denoted by $M(50,\mu)$ with μ being the mean of a single trial. This case is not covered in the theoretical development above. For convenience, whenever $K>1$ and the data are continuous, we fix the mean of the first cluster, i.e. $\theta_{1,0}$, at 0_p.

Model 1: $X \sim N(0_5, (0.8^{|i-j|})_5)$.

Model 2: $X \sim (U_1^2, U_2^2, U_3^2)^\intercal$, each U_i generated independently from the uniform distribution on $(-1/2, 1/2)$.

Model 3. $X = \sum_{k=1}^{2} .5N(\theta_{k,0}, (0.5^{|i-j|})_2)$, $\theta_{2,0} = (2,2)^\intercal$.

Model 3*. Same as Model 3 but with $\Sigma_{1,0} = \mathrm{I}_2$ and $\Sigma_{2,0} = (0.5^{|i-j|})_2$.

Model 4: $X = \sum_{k=1}^{3}(1/3)N(\theta_{k,0}, \mathrm{I}_4)$, $\begin{cases}\theta_{2,0} = (2.5, 1.5, 2, 1.5)^\intercal \\ \theta_{3,0} = (1.5, 3, 2.75, 2)^\intercal\end{cases}$.

Model 5: $X = \sum_{k=1}^{3}(1/3)M(50, \mu_{k,0})$, $\begin{cases}\mu_{1,0}^\intercal = (.2,.2,.2,.2,.2) \\ \mu_{2,0}^\intercal = (.1,.3,.2,.1,.3) \\ \mu_{3,0}^\intercal = (.3,.1,.2,.3,.1)\end{cases}$.

Model 6: $X = \sum_{k=1}^{4}(1/4)N(\theta_{k,0}, (0.5^{|i-j|})_2)$, $\theta_{k,0} = (2k-2, 2k-2)^\intercal$.

Model 6*. Same as Model 6 but with $\Sigma_{1,0} = \Sigma_{3,0} = (0.5^{|i-j|})_2$ and $\Sigma_{2,0} = \Sigma_{4,0} = \mathrm{I}_2$.

Model 7: $X = \sum_{k=1}^{5}(1/5)N(\theta_{k,0}, \mathrm{I}_8)$, $\begin{cases}\theta_{2,0} = (1, 1.5, .75, 2, 1.5, 1.75, .5, 2.5)^\intercal \\ \theta_{3,0} = (2, .75, 1.5, 1, 1.75, .5, 2.5, 1.5)^\intercal \\ \theta_{4,0} = (1.5, 2, 1, .75, 2.5, 1.5, 1.75, .5)^\intercal \\ \theta_{5,0} = (.75, 1, 2, 1.5, .5, 2.5, 1.5, 1.75)^\intercal\end{cases}$.

Model 8: $X = \sum_{k=1}^{8}(1/8)t_3(\theta_{k,0}, 5)$ where $\theta_{2,0} = (0,0,2)^\tau$, $\theta_{3,0} = (0,2,0)^\tau$, $\theta_{4,0} = (0,2,2)^\tau$, $\theta_{5,0} = (2,0,0)^\tau$, $\theta_{6,0} = (2,0,2)^\tau$, $\theta_{7,0} = (2,2,0)^\tau$, $\theta_{8,0} = (2,2,2)^\tau$, and $t_3(\theta_{k,0}, 5)$ denotes the three-dimensional multivariate noncentral t distribution with five degrees of freedom and $\theta_{k,0}$ being the mean.

For clarity, in these models, the number of clusters K is 1, 1, 2, 2, 3, 3, 4, 4, 5, and 8, and the dimension of data is 5, 3, 2, 2, 4, 5, 2, 2, 8, and 3, respectively. These models represent a variety of numbers of clusters, dimensions of the data, and cluster distributions that one may meet in practice. To measure the degree of separation between clusters in each model, which would reveal the difficulty of clustering analysis and the corresponding order determination, we calculate the oracle Bayes error rate (OBER) defined as

$$1 - \int_{\mathbb{R}^p} \max_{k=1,\ldots,K}\{\pi_{k,0} f(x, \mu_{k,0})\}dx \tag{18}$$

for continuous X and defined similarly if X is discrete. Because the integrand in (18) is always bounded by zero and $f_0(x) = \sum_{k=1}^{K} \pi_{k,0} f(x, \mu_{k,0})$, OBER is always bounded by zero and one for any data. By construction, the more separate the clusters from each other, the smaller OBER is. For the models above with $K > 1$, OBER is .125, .103, .090, .128, .186, .143, .168, and .170, respectively, so the clustering analysis can be reasonably accurate but is nontrivial.

To implement DAE and DAE-Ⅱ in these models, we uniformly assume X to follow a homogeneous mixture multivariate normal distribution, i.e. with a common (but unknown) covariance matrix, whenever X is continuous. This parametric assumption is violated in Models 2, 3*, 6*, and 8, from which we can evaluate the corresponding robustness of the proposed estimators. In Model 5, we treat the family of multinomial distributions with 50 trials as known, and generate the new cluster from the same family. The choices of $\mu_{A,0}$ follow those suggested in Section 5. For all the models, we set $\phi_L(\cdot, \widehat{F}_n)$ to be the log-likelihood function under the working parametric assumption, which again differs from the true log-likelihood for Models 2, 3*, 6*, and 8. As suggested in Section 5, we maximize $\phi_L(\cdot, \widehat{F}_n)$ by running the EM algorithm multiple times with a variety of initial values, and choosing the optimal result. To study the robustness of the proposed estimators to $\pi_{A,0}$ or equivalently to r with $\pi_{A,0} = n^{r-1}/L$ for working L, we set r at .55, .625, and .70 sequentially with K_M set at 10. The robustness to K_M will be inspected later.

To make a dynamic comparison as the sample size n varies, we set n at 200, 400, and 800 sequentially for each continuous model with $K > 1$, and set n at 100, 200, and 400 for Models 1, 2 and 5. These settings comply with [18]. Table 1 records the performance of the aforementioned estimators, measured by the approximate percentage of correctly determining K based on 2000 independent runs for each model. For effective presentation, we only report the winner of AIC and BIC, and the winner of the GSF methods equipped with each of SCAD, MCP, and Adaptive Lasso penalties, specified for each model and each sample size after assessing the average performance over all the 2000 runs.

From Table 1, both the better of AIC and BIC and the best of GSF methods perform dramatically differently as the model varies: they are consistent in some

TABLE 1

The performance of the order-determination estimators: in Row 1, IC stands for the better performer of AIC and BIC per model and sample size after assessing the average performance across all runs, GSF stands for the best performer of the GSF methods when equipped with SCAD, MCP, and Adaptive Lasso penalties per model and sample size after assessing the average performance across all runs, and DAE and DAE-Ⅱ are the proposed estimators, with $r = .55$, $.625$, *and* $.70$ *sequentially; in each cell of Columns 3-10 is an approximation of percentage of correct order determination based on* 2000 *independent runs.*

Model	n	IC	GSF	DAE			DAE-Ⅱ		
1	100	0	100	35	30	33	92	82	72
	200	0	100	52	51	54	95	92	82
	400	0	100	54	64	74	99	98	94
2	100	0	15	33	28	35	89	81	69
	200	0	0	54	50	53	94	90	82
	400	0	0	60	70	75	99	97	93
3	200	56	59	68	81	90	68	81	89
	400	76	71	80	88	98	86	92	98
	800	96	99	99	99	99	99	99	99
3^*	200	0	48	89	97	91	91	95	86
	400	0	77	90	99	98	92	99	97
	800	0	96	96	99	99	98	99	99
4	200	45	43	78	87	96	90	92	94
	400	63	54	91	99	99	96	99	99
	800	93	96	92	100	100	98	99	100
5	100	80	86	61	71	72	52	52	62
	200	80	94	65	72	76	76	61	72
	400	82	97	67	75	79	70	74	88
6	200	18	10	47	59	52	30	39	25
	400	25	3	48	60	60	33	41	37
	800	32	4	49	61	64	40	45	50
6^*	200	3	1	64	48	8	48	30	4
	400	3	12	68	59	9	55	42	6
	800	0	12	70	77	15	53	65	12
7	200	20	6	43	25	3	11	7	3
	400	41	6	59	51	5	20	32	5
	800	66	74	91	85	14	51	57	8
8	200	5	7	58	58	54	56	53	55
	400	8	35	80	81	84	84	87	88
	800	11	60	99	98	98	97	97	95

models but complete fail in some others. By contrast, for all the three choices of r, both DAE and DAE-Ⅱ have much stabler performances across models, and they mostly have higher probabilities of correctly determining K, with Models 1 and 5 being the only exceptions where GSF is better. In particular, both DAE and DAE-Ⅱ are consistent in Model 8 where K is relatively large. When these estimators fail to truly estimate K in probability, i.e. in Models 1, 2, 5, 6, and 6^* for DAE and in Models 5, 6, 6^*, and 7 for DAE-Ⅱ, Table 2 records the percentages they mis-specify K to be each of $K-2$, $K-1$, $K+1$, and $K+2$, with $r = .625$ as an illustration. The results suggest that both estimators are mostly bounded by $K-2$ and $K+2$, so they can still give reasonable results when they are inconsistent. In Model 4, DAE-Ⅱ is more robust to the choice of r than DAE, which complies with the discussion in Section 4, but there is no

TABLE 2

The performance of the proposed estimators in more details: the number in each cell of Column 4-8 is the estimated percentage that DAE or DAE-Ⅱ equipped with $r = .625$ specifies K to be $K-2$, $K-1$, K, $K+1$, and $K+2$, respectively, based on 2000 independent runs. The number in Column 9 is the total of those in Columns 4-8.

Model	Method	n	$K-2$	$K-1$	K	$K+1$	$K+2$	Total
1	DAE	100	–	–	30	5	23	58
		200	–	–	51	3	15	69
		400	–	–	64	28	6	98
2	DAE	100	–	–	28	13	9	50
		200	–	–	50	11	9	70
		400	–	–	70	2	4	76
5	DAE	100	0	3	71	17	6	97
		200	0	1	72	22	4	99
		400	0	0	75	23	2	100
	DAE-Ⅱ	100	0	45	52	2	0	99
		200	0	36	61	3	0	100
		400	0	26	74	0	0	100
6	DAE	200	24	10	59	7	0	100
		400	13	9	60	16	2	100
		800	38	0	61	1	0	100
	DAE-Ⅱ	200	31	23	39	7	0	100
		400	28	20	41	10	1	100
		800	42	10	45	3	0	100
6^*	DAE	200	22	30	48	0	0	100
		400	13	28	59	0	0	100
		800	3	20	77	0	0	100
	DAE-Ⅱ	200	28	42	30	0	0	100
		400	21	37	42	0	0	100
		800	4	31	65	0	0	100
7	DAE-Ⅱ	200	44	21	7	2	2	76
		400	41	27	32	0	0	100
		800	12	22	57	6	3	100

clear winner between the two estimators in general.

From the results for Model 2 in Table 1, when the assumption of homogeneous mixture normal distribution fails, DAE-Ⅱ still consistently estimates K, and DAE also truly specifies K with an increasing probability as n grows, both of which behave roughly the same as in Model 1 where the parametric assumption is satisfied and OBER is similar. The same phenomenon can be observed from Model 8 and also if we compare the performances of DAE and DAE-Ⅱ in Models 3 and 6 with those in Models 3^* and 6^* (except when $r = .70$ in Model 6^*), where the distributions of X differ only in the homogeneity of covariance matrices. These comply with our theoretical anticipation for the robustness of the proposed estimators to the parametric assumption adopted when fitting the mixture distribution.

Similarly, promising results about the robustness to the choice of r can be observed for both DAE and DAE-Ⅱ in most models, particularly as n grows. The only exceptions are Models 6^* and 7 where the proposed estimators almost completely fail for the choice of $r = 0.70$. Across different models, there is no universally optimal choice of r. We recommend using $r = .625$ as a conservative

TABLE 3

The sensitivity of the estimators to the imbalance of data: the pair of numbers in each cell of Column 1 is the value of $(\pi_{1,0}, \pi_{2,0})$ *that generates the original data from the modified Model 3 with* $n = 400$. *The meanings of the numbers in the other columns follow those in Table 1.*

$(\pi_{1,0}, \pi_{2,0})$	IC	GSF	DAE			DAE-Ⅱ		
$(0.1, 0.9)$	41	50	18	38	48	15	29	44
$(0.2, 0.8)$	65	65	29	50	87	48	69	89
$(0.3, 0.7)$	67	67	56	78	95	69	78	98
$(0.4, 0.6)$	70	70	73	85	98	85	91	98
$(0.5, 0.5)$	76	71	80	88	98	86	92	98

choice in practical applications.

As mentioned in Section 4, the proposed DAE and DAE-Ⅱ can be sensitive to the imbalance of the data. To illustrate this point, we modify Model 3 by changing $(\pi_{1,0}, \pi_{2,0})$ to be $(0.1, 0.9)$, $(0.2, 0.8)$, $(0.3, 0.7)$, $(0.4, 0.6)$, sequentially, with $n = 400$. The performances of all the aforementioned estimators are summarized in Table 3, with the case of balanced data, i.e. $\pi_{1,0} = \pi_{2,0} = 0.5$, copied from Table 1 and used as the reference.

From Table 3, compared with the existing estimators, the proposed DAE and DAE-Ⅱ are indeed more sensitive to the imbalance of the original data: when one cluster is much smaller than the other, both estimators are sub-optimal to the information criteria and GSF; when the two clusters have more comparable sizes, both DAE and DAE-Ⅱ have elevated performances, and they become the clear winners when the data are balanced.

Finally, we evaluate the robustness of DAE and DAE-Ⅱ to the prefixed upper bound K_M of the candidate choices of K. For efficiency of presentation, we choose $r = .625$ as suggested above when generating the new cluster. As a simple illustration, we first apply DAE and DAE-Ⅱ to Models 3 and 4 with $n = 400$, and draw the curve of the probability of true specification of K as K_M increases from 5 to 14, again based on 2000 independent runs. These curves are displayed in Figure 3, which shows promising results that are indicative for the robustness of both estimators to the choice of K_M. Table 4 illustrates the results for the other models and sample sizes, with n set the same as in Table 1 and K_M set sequentially at 6, 8, 10, 12, and 14. Generally, the same sign for robustness can be observed from this table, especially as n grows, with Models 1 and 2 being the only exceptions.

Comparing the performance of DAE and DAE-Ⅱ in all the simulation studies above, the use of bootstrap re-sampling has negligible impact on the proposed order-determination procedure for most models, and can be either advantageous or adverse otherwise. Thus, we make no recommendations on which of DAE or DAE-Ⅱ to use in practice. Nonetheless, we still speculate that a more delicate use of bootstrap re-sampling can polish the order determination at least in certain scenarios, which will be investigated in the future.

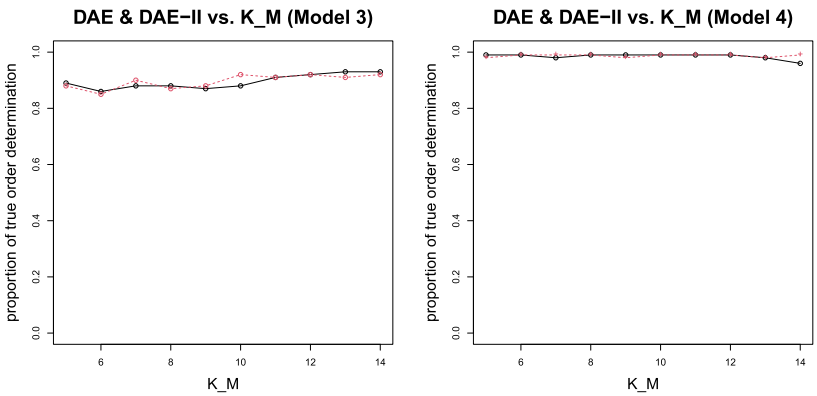

FIG 3. *The robustness of the proposed estimators to K_M: the left panel is for Model 3 and the right panel is for Model 4; in each panel, r is fixed at .625, the x-axis is K_M, and the y-axis is the probability of correct order determination. The solid line connected by $\circ$ and the dashed line connected by $+$ correspond to DAE and DAE-II, respectively.*

7. Real data examples

7.1. Seeds data

We first apply the proposed estimators to the seeds data studied in [18]. The data set is available in [6], where 210 seeds were collected from three varieties, and seven features were measured for each seed. Similarly to [6] and [18], we extract the first three principal components of the data and assume a mixture multivariate normal distribution with common but unknown covariance matrix subsequently. Under this assumption, we apply DAE and DAE-II with $\phi(\cdot, \widehat{F}_n)$ being the log-likelihood, and specify $K_M = 12$ and $r = .625$. The results are displayed in Figure 4. Because the true clusters, as indicated by the varieties the seeds come from, are actually known, we can estimate the oracle Bayes error rate (18) under the homogeneous mixture normality assumption, which is .428. Thus, the clustering problem is relatively difficult for this data set.

From the left panel of Figure 4, the data augmentation has a constantly large impact on clustering when $L > 3$, both with and without bootstrap re-sampling, and, as depicted in the middle panel of Figure 4, three is also the location of the elbow of the curve of $-\phi_L(\cdot, \widehat{F}_n)$. Thus, as shown in the right panel of Figure 4, both DAE and DAE-II suggest selecting $K = 3$. This conforms to the result of the GSF method equipped with the aforementioned multiple penalty functions and also to the results of AIC and BIC; see [18] for details.

7.2. Pen digits data

We next apply the proposed estimators to the pen-based recognition of handwritten digits data set [1], where 44 writers were collected and each created 250

TABLE 4

The robustness of the proposed estimators to K_M: the number in each cell of Columns 3-12 is the estimated percentage of correct order determination, based on 2000 *independent runs; those in Columns 3-7 are for DAE with K_M set at* 6, 8, 10, 12, *and* 14, *respectively; those in Columns 8-12 are for DAE-II likewise.*

Model	n	DAE					DAE-II				
1	100	64	40	30	25	12	83	82	82	57	14
	200	78	67	51	44	30	95	95	92	70	35
	400	80	70	64	44	32	99	98	98	79	54
2	100	54	37	28	19	18	82	81	81	38	34
	200	75	67	50	35	27	93	93	90	75	37
	400	76	70	70	37	27	98	97	97	76	37
3	200	80	80	81	86	58	83	81	81	79	52
	400	86	88	88	92	93	85	87	92	92	92
	800	99	99	99	98	95	99	99	99	98	97
3*	200	98	98	97	95	83	97	96	95	92	81
	400	99	99	99	98	98	99	99	99	98	95
	800	99	99	99	99	99	99	99	99	99	99
4	200	99	96	87	87	84	99	95	92	89	79
	400	99	99	99	99	96	99	99	99	99	99
	800	100	100	100	100	100	100	100	99	99	99
5	100	72	71	71	70	69	52	52	52	52	51
	200	74	72	72	72	70	61	61	61	60	60
	400	76	75	75	75	75	75	74	74	73	73
6	200	60	59	59	54	50	45	42	39	39	36
	400	62	60	60	59	57	48	46	41	40	39
	800	62	62	61	61	60	49	48	45	45	45
6*	200	52	50	48	39	33	32	30	30	25	18
	400	62	62	59	56	52	45	43	42	40	37
	800	80	78	77	75	73	65	65	65	64	64
7	200	28	27	25	25	24	12	9	7	7	6
	400	52	51	51	49	47	34	34	32	31	31
	800	85	85	85	85	85	58	58	57	57	57
8	200	−−	63	58	58	55	−−	55	53	53	53
	400	−−	85	81	80	80	−−	89	87	87	83
	800	−−	100	98	96	93	−−	99	97	95	95

handwritten digits ranged from zero to nine, and 16 features were extracted for each handwritten digit. Due to the limit of computational power, here we only analyze the sub-sample where the digits are one, seven, and nine, which has 3340 observations in total. These three digits are distinct in general but are also relatively similar to each other, so we can regard the corresponding groups of observations, which are of size 1143, 1142, and 1055, respectively, as three clusters that are partially overlapped. Based on an exploratory data analysis, each cluster can be considered as following an approximate multivariate normal distribution. Thus, if we ignore the labels, i.e. the indicator of the digits, and if we pretend not knowing $K = 3$, then we can fit a mixture multivariate normal distribution with equal but unknown covariance matrices on this sub-sample, with $\phi(\cdot, \widehat{F}_n)$ being the corresponding log-likelihood function.

To ease the clustering, we follow the similar strategy to the analysis of the seeds data above to only use the first six principal components of the original data, which count for about 85% of the variation of the data if each feature

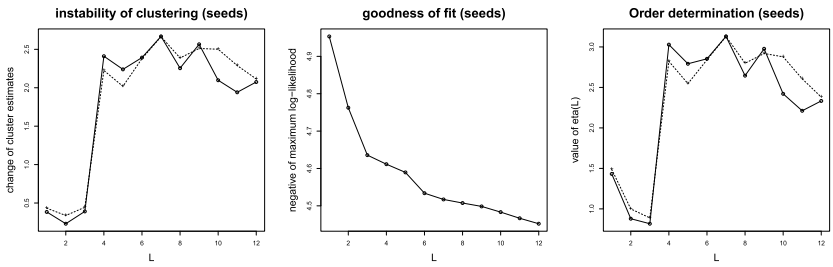

FIG 4. *The performance of DAE and DAE-Ⅱ when applied to the seeds data: in the left panel displays the change of cluster estimates caused by data augmentation as the working number of clusters varies, the solid line connected by* $\circ$ *for DAE and the dashed line connected by* $+$ *for DAE-Ⅱ; in the middle panel displays the corresponding change of the negative of the maximal log-likelihood function; in the right panel displays the curves of* $\eta(\cdot)$*, again the solid line connected by* $\circ$ *for DAE and the dashed line connected by* $+$ *for DAE-Ⅱ.*

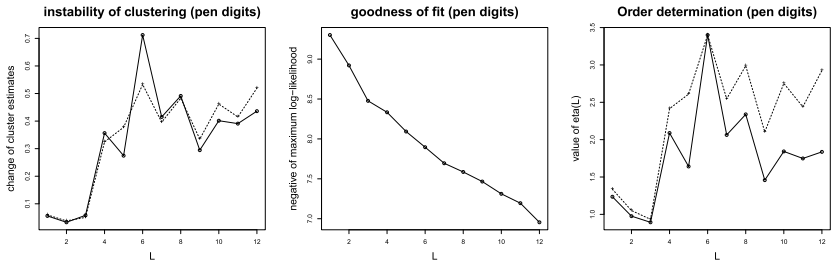

FIG 5. *The performance of DAE and DAE-Ⅱ when applied to the pen digits data with three digits* $\{1, 7, 9\}$*: in the left panel displays the change of cluster estimates caused by data augmentation as the working number of clusters varies, the solid line connected by* $\circ$ *for DAE and the dashed line connected by* $+$ *for DAE-Ⅱ; in the middle panel displays the corresponding change of the negative of the maximal log-likelihood function; in the right panel displays the curves of* $\eta(\cdot)$*, again the solid line connected by* $\circ$ *for DAE and the dashed line connected by* $+$ *for DAE-Ⅱ.*

is standardized to have the unit sample standard deviation. The oracle Bayes error rate (18) of the reduced data is approximately .438 under the mixture normality assumption, indicating that the clustering problem is relatively difficult. Figure 5 summarizes the results of DAE and DAE-Ⅱ with $K_M = 12$ and $r = .625$.

Similarly to the above, with bootstrap re-sampling or not, the curve for the impact of data augmentation always conveys a clear pattern that suggests three clusters in the data. The goodness-of-fit curve, however, becomes relatively linear and contributes little to order determination. For both DAE and DAE-Ⅱ, the curve of $\eta(\cdot)$ essentially resembles that for the impact of data augmentation and truly specifies K to be three. This suggests that the data augmentation can sometimes deliver a stronger signal than the commonly used goodness-of-fit measure, so it should be valued more in the future applications. For reference, K is mis-specified to be 12, 12, 9, 12, and 8 by AIC, BIC,

and GSF equipped with SCAD, MCP, and Adaptive Lasso, respectively, all of which are far biased. The failure of AIC and BIC is easily understood from the goodness-of-fit curve in Figure 5, which again shows no pattern for characterizing K.

8. Summary

In this paper, we introduce the novel idea of data augmentation to conduct consistent order determination for general model-based clustering. The essence of data augmentation is to perturb the data in a designed manner, so that it makes negligible change to clustering when the working number of clusters coincides with the true number of clusters, and it causes instability of clustering when the working number of clusters is larger. Although the instability of clustering has been used in multiple existing order-determination methods, as we are aware of, this is the first time it is justified rigorously under fairly general conditions.

As mentioned in Section 4, consistent order determination method can be delivered by using the pattern of the impact of data augmentation alone, i.e. $\tau(\widehat{f}_L, \widehat{f}_L^*)$, without the aid of goodness-of-fit measure. The key issue for this approach to find a uniform threshold that determines whether an outcome of $\tau(\widehat{f}_L, \widehat{f}_L^*)$ is negligible or not, especially when the sample size is limited. An alternative approach that may avoid this issue is to transform $\tau(\widehat{f}_L, \widehat{f}_L^*)$ appropriately in a similar way to the numerous existing transformations of the scree plot in other scenarios of order determination.

The proposed work still has several limitations. First, as mentioned in Section 4 and observed in Section 6, the proposed estimators tend to lose consistency if the data are severely imbalanced, i.e. if the clusters have dramatically different sizes. In this case, we suggest using the existing order-determination methods such as the information criteria and GSF. Second, as mentioned in Section 2, the proposed theory relies on the uniqueness of the maximizer of $\phi_L(\cdot, F)$ for each $F \in \mathcal{G}$ and candidate L. It remains unclear to us how to use the strategy of data augmentation for order determination in more general cases. Third, as mentioned in Section 6, the bootstrap re-sampling must be used more delicately to sharpen the instability pattern and improve the sample performance of the proposed method. Finally, the proposed estimators are not directly generalizable yet for distance-based clustering methods. Research towards these directions requires tremendous work, and will be investigated in more detail in the future.

Acknowledgments

The author thanks the editor, the associate editor and the referees for their helpful comments and constructive suggestions. Luo was supported in part by the National Science Foundation of China (12131006, 12001484).

References

[1] Alimoglu, F. and Alpaydin, E. (2001). Combining Multiple Representations for Pen-based Handwritten Digit Recognition. *Turkish Journal of Electrical Engineering and Computer Sciences* **9** 1–12.

[2] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085. MR3375876

[3] Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics* **47** 2504–2537. MR3988764

[4] Benaglia, T., Chauveau, D., Hunter, D. R. and Young, D. S. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* **32** 1–29.

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

[6] Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Lukasik, S. and Żak, S. (2010). Complete gradient clustering algorithm for feature analysis of X-ray images. *Advances in Intelligent Systems and Computing* 15–24. MR2914212

[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360. MR1946581

[8] Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Computational Statistics & Data Analysis* **56** 468–477. MR2853747

[9] Fernholz, L. T. (1983). *von Mises Calculus for Statistical Functionals*. Springer, New York. MR0713611

[10] Fraley, C. and Raftery, A. E. (1999). Mclust: software for model-based cluster analysis. *Journal of Classification* **16** 297–306. MR2019797

[11] Fujita, A., Takahashi, D. Y. and Patriota, A. G. (2014). A non-parametric method to estimate the number of clusters. *Computational Statistics & Data Analysis* **73** 27–39. MR3147972

[12] G. Celeux, S. F.-S. and Robert., C. (2019). Model selection for mixture models – perspectives and strategies. In *Handbook of Mixture Analysis* 117–154. MR3889692

[13] Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise Variable Selection for High-Dimensional Model-Based Clustering. *Biometrics* **66** 793–804. MR2758215

[14] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya A* **62** 49–66. MR1769735

[15] Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103** 875–887. MR3620445

[16] Luo, W. and Li, B. (2021). On order determination by predictor augmentation. *Biometrika* **108** 557–574. MR4298764

[17] Mai, Q., Zhang, X., Pan, Y. and Deng, K. (2021). A Doubly-Enhanced

EM Algorithm for Model-Based Tensor Clustering*. *Journal of the American Statistical Association* 1–15.

[18] Manole, T. and Khalili, A. (2020). Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *arXiv preprint arXiv:2005.11641* **0** 0–0. MR4352522

[19] McLachlan, G. J., Lee, S. X. and Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application* **6** 355–378. MR3939525

[20] McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. In *Wiley Series in Probability and Statistics*. MR1789474

[21] Millar, R. B. (2011). Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB.

[22] Pan, W. and Shen, X. (2007). Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Research* **8** 1145–1164.

[23] Panahi, A., Dubhashi, D. P., Johansson, F. D. and Bhattacharyya, C. (2017). Clustering by Sum of Norms: Stochastic Incremental Algorithm, Convergence and Cluster Recovery. In *Proceedings of the 34th International Conference on Machine Learning* **70** 2769–2777.

[24] Raftery, A. E. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association* **101** 168–178. MR2268036

[25] Serfling, R. (1980). Approximation Theorems of Mathematical Statistics.

[26] Sugar, C. A. and James, G. M. (2003). Finding the Number of Clusters in a Dataset. *Journal of the American Statistical Association* **98** 750–763. MR2012330

[27] Sun, D., Toh, K.-C. and Yuan, Y. (2021). Convex Clustering: Model, Theoretical Guarantee and Efficient Algorithm. *Journal of Machine Learning Research* **22** 1–32. MR4253702

[28] Tan, K. M. and Witten, D. (2015). Statistical properties of convex clustering. *Electronic Journal of Statistics* **9** 2324–2347. MR3411231

[29] Tibshirani, R. and Walther, G. (2005). Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics* **14** 511–528. MR2170199

[30] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of The Royal Statistical Society Series B-statistical Methodology* **63** 411–423. MR1841503

[31] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). Statistical analysis of finite mixture distributions. John Wiley & Sons.

[32] Wang, S. and Zhu, J. (2008). Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. *Biometrics* **64** 440–448. MR2432414

[33] Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**

713–726. MR2724855
[34] Xu, J. and Lange, K. (2019). Power k-Means Clustering. In *International Conference on Machine Learning* 6921–6931.
[35] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942. MR2604701
[36] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469