Journal of Data Science 23 (1), 188–207
January 2025
DOI: 10.6339/24-JDS1140
Statistical Data Science

A Two-Stage Classification for Dealing with Unseen Clusters in the Testing Data

Jung Wun Lee[1] and Ofer Harel[2,*]

[1]*Department of Biostatistics, Harvard University, Boston, MA, 02115, USA*
[2]*Department of Statistics, University of Connecticut, Storrs, CT, 06269, USA*

Abstract

Classification is an important statistical tool that has increased its importance since the emergence of the data science revolution. However, a training data set that does not capture all underlying population subgroups (or clusters) will result in biased estimates or misclassification. In this paper, we introduce a statistical and computational solution to a possible bias in classification when implemented on estimated population clusters. An unseen-cluster problem denotes the case in which the training data does not contain all underlying clusters in the population. Such a scenario may occur due to various reasons, such as sampling errors, selection bias, or emerging and disappearing population clusters. Once an unseen-cluster problem occurs, a testing observation will be misclassified because a classification rule based on the sample cannot capture a cluster not observed in the training data (sample). To overcome such issues, we suggest a two-stage classification method to ameliorate the unseen-cluster problem in classification. We suggest a test to identify the unseen-cluster problem and demonstrate the performance of the two-stage tailored classifier using simulations and a public data example.

Keywords *classification; cluster analysis; open set recognition; outlier detection*

1 Introduction

This paper focuses on a scenario in which the classification is implemented on estimated population clusters usually obtained from cluster analysis. One motivating example is a classification problem on electronic health records, on which researchers may be interested in estimating homogeneous groups of patients based on their features, such as demographic factors, biomarkers, medical history, or symptoms related to certain diseases. Estimated sample clusters can summarize patients' information and discover common characteristics. A future patient can be assigned to one of the estimated clusters so that this patient may receive more appropriate medical services or treatment. In such a way, a combination of cluster analysis and classification may play a key role in data science, such as public health research.

When a training data set fails to cover all existing population clusters, it is possible to have a new observation (i.e., testing data) from clusters not covered in the training data. In this paper, we denote this observation as *Unknown*. We define the unseen-cluster problem as a case in which a sample or training data fails to cover all existing population clusters, and a new observation from outside the sample may be inappropriately classified. Such unseen-cluster problems are likely to appear when (i) population clusters are unknown but estimated based

*Corresponding author. Email: jwlee@hsph.harvard.edu or ofer.harel@uconn.edu.

Received January 21, 2024; Accepted April 28, 2024

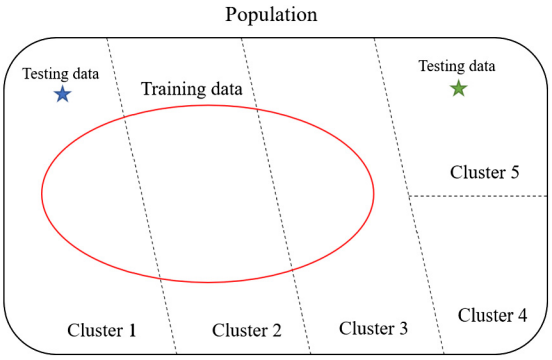

Figure 1: An illustration of unseen-cluster problem.

on a sample and (ii) a sample does not cover one or more underlying population clusters. For example, scenarios of selection bias (Bethlehem, 2010), capturing of rare clusters (Wankhade et al., 2018), or high dimension data (Klawonn et al., 2012) can all trigger such unseen-cluster problem. An illustration example is in Figure 1, where the training data (red oval) has three clusters, while two additional clusters exist in the population. It is clear that observations from clusters four or five will be assigned to the wrong cluster using the current methodology.

Dealing with population clusters that are uncovered in the training data can be understood as the open set recognition (OSR) problem (Geng et al., 2020). OSR aims to establish a classifier that may appropriately classify or identify covered and uncovered clusters. Numerous works have suggested classification methods embedded with the rejecting option to achieve this goal. For example, Bartlett and Wegkamp (2008) suggested using an extended discriminant function by embedding the user-specified rejection constant d into a binary classifier and rejecting new observations whose conditional probabilities fall into $[1/2 - d, 1/2 + d]$. Further, Bendale and Boult (2015) introduced the open world recognition (OWR) framework to deal with novel categories not covered in the training data. When new testing data appear, OWR performs (1) the open set recognition, (2) labeling on the testing data, and (3) tailoring the current classifier sequentially. For more details on OWR and advances, see (Bendale and Boult, 2015; Doan and Kalita, 2017; Lonij et al., 2017) and their references. These methods are worth acknowledging but lack theoretical justifications because statistical properties, such as type 1 error rates, power, and classification accuracies, are unstudied and thus unreliable for our motivating example and further empirical data analysis.

The out-of-distribution detection (ODD) framework (Hodge and Austin, 2004; Pimentel et al., 2014), which is one type of OSR method based on a single-cluster assumption on the population, can be another approach for identifying testing data sampled from uncovered clusters. Hodge and Austin (2004) suggested three types of ODD frameworks that differ in assumptions and prior information. Type 1 ODD does not use prior information to determine the outliers, and thus, it is equivalent to unsupervised learning with no outcome variable or labeled data. For example, literature on dealing with mixed labeled and unlabeled data sets is one stream of Type 1 ODD methodology (Miller and Browning, 2003). On the other hand, the Type 2 ODD approach models both normality and abnormality based on a sample (Hodge and Austin, 2004). This approach is analogous to supervised binary classification and requires pre-labeled data classified as normal or abnormal. For example, Schölkopf et al. (1999) used support vector algorithms for novelty detection. Finally, the Type 3 ODD method assumes only normality or,

in very few cases, model abnormality. In this sense, Type 3 ODD is generally named novelty detection or novelty recognition.

Some recent publications in novelty detection are classified as Type 3 ODD methods. For example, Bouveyron (2014) introduced Adaptive Mixture Discriminant Analysis (AMDA), a framework for model-based discriminant analysis that allows the testing data set to contain novel clusters not observed in the training data. Specifically, AMDA aggregates information from training and testing data to estimate both observed clusters covered by training data and unobserved clusters that may be included in testing data. Similarly, Cappozzo et al. (2020) introduces Robust and Adaptive Eigen Decomposition Discriminant Analysis (RAEDDA) based on a trimmed log-likelihood function. This paper aims to consider variable and cluster noises in the training data and achieve a parsimonious model with fewer parameters than the conventional Gaussian mixture model. In addition, Denti et al. (2021) suggested a two-stage Bayesian semi-parametric novelty detection model that employs prior information robustly extracted from a set of complete training data sets.

Since these model-based novelty detection methods use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) or the Dirichlet Process mixture model (DPMM) of Gaussian densities (Lo, 1984) to estimate the class/novelty membership, they require elaboration on prior distributions, initial values and a large number of iterations as well as diagnosing convergences. In this sense, they may be impractical in dealing with unseen-clusters in a classification problem where one needs to determine whether a testing observation is a novelty. Consequently, many simple diagnostic tests that may avoid high computational costs or the necessity of specifying details in prior distribution have been proposed, such as Mahalanobis distance-based method (Clifton et al., 2011; Lee et al., 2018; Liang et al., 2017), Manhattan distance-based method (Yong et al., 2012), Euclidian distance-based method (Feinman et al., 2017; Ma et al., 2018), Squeezer algorithm based approach (He et al., 2003), k-nearest neighborhood-based methods (Ma et al., 2018; Papernot and McDaniel, 2018), and Bootstrap-based method (Grosse et al., 2017). Since none of these papers provides theoretical justifications for their diagnostic tests (i.e., type 1 error calculations or power), we are motivated to propose a new diagnostic test for novelties in testing data.

The rest of this article is as follows. In Section 2, we introduce *unseen-cluster problem* and explain how a testing observation belongs to an unseen cluster not observed in the training data. We also suggest a diagnostic test to identify the unseen-cluster problem. In addition, we propose a two-stage classification method that utilizes the proposed test in classification steps. In Section 3, we perform numerical studies to illustrate that the proposed test is appropriate by evaluating its Type 1 errors and power estimates. We also demonstrate the superiority of our two-stage classification method by comparing its classification accuracies with the conventional method under various scenarios. Applications of the proposed approach to public data sets are the topic of Section 4. Lastly, conclusions and further research goals are presented in Section 5.

2 Method

Suppose a population in interest consists of an unknown number of disjoint latent clusters, say, $[A_1, \ldots, A_G]$. Here, the number of latent clusters G is finite but cannot be observed. Next, a sample consists of K clusters $[A_1, \ldots, A_K]$, $K \leqslant G$, which is a set of clusters of the population but is not necessarily equal to the collection of all existing clusters. Now, consider a classification problem as two steps: (i) run cluster analysis on the collected sample and identify K number

of clusters, (ii) implement a classification process on a new observation by assigning it to one of the identified latent clusters from (i). Here, we define an *unseen-cluster problem*, which may appear in a conventional classification process.

Definition 1 (Unseen-cluster problem)**.** An unseen-cluster problem denotes a situation in which a testing observation is sampled from a latent population cluster that is not included in the training data.

An unseen-cluster problem occurs if $K < G$ and a new observation $\mathbf{x}$ is sampled from $[A_{K+1}, \ldots, A_G]$. This observation will be assigned to one of the clusters among $[A_1, \ldots, A_K]$ and thus be misclassified. Under such a situation, it is appropriate to label the new observation as "*unclassified*", instead of naively assigning it to one of the current clusters. In this sense, one goal of this paper is to propose an appropriate diagnostic test for the unseen-cluster problem. Define $\omega_{\mathbf{x}_{new}}$ to be a new observation $\mathbf{x}_{new}$'s the cluster membership, and let $\Omega_K = \{1, 2, \ldots, K\}$ be the list of population clusters that are included in the training data. Based on the definition and notations, we may say $\mathbf{x}_{new}$ is *unclassified* if $\omega_{\mathbf{x}_{new}} \notin \Omega_K$. Using a conventional hypothesis testing framework, our null and alternative hypotheses can be written as

$$H_0 : \omega_{\mathbf{x}_{new}} \in \Omega_K \quad vs \quad H_1 : \omega_{\mathbf{x}_{new}} \notin \Omega_K. \tag{1}$$

The calculation of type 1 error and finding a size α test can be challenging because it requires the computation of the events' probabilities that are not disjoint nor independent. Namely, the calculation of the type 1 error of the test requires additional assumptions that are often untestable based on the observed sample. To overcome such difficulties and dependencies on untestable assumptions, we propose a test for a single cluster (that is, $H_0^{(k)} : \omega_{\mathbf{x}_{new}} = k$ versus $H_1^{(k)} : \omega_{\mathbf{x}_{new}} \neq k$, $k = 1, \ldots, K$) then combine the results of K-single tests so that the overall type 1 error of the test does not exceed size α.

2.1 A Test for a Single Cluster

Let $\mathbf{x} \in \mathbb{R}^p$ be a sampled observation from a population that is a mixture of K clusters $[A_1, \ldots, A_K]$, where each component follows a p-dimensional multivariate normal distribution. Consequently, a conditional distribution of $\mathbf{x} \mid \omega_{\mathbf{x}} = k$ follows $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$. Conventional classification methods depend on the conditional probability of the cluster membership given observed values. In this sense, the proposed decision rule is based on the conditional probability $P(\mathbf{x} \in A_k \mid \omega_{\mathbf{x}} \in \Omega_K)$ can be written as

$$P(\mathbf{x} \in A_k \mid \omega_{\mathbf{x}} \in \Omega_K) = \frac{\phi(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \phi(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \quad k = 1, \ldots, K, \tag{2}$$

where $\phi(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the density function of $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. If an unseen-cluster problem does not occur, then it would be reasonable to assign the unit to the cluster with the highest conditional probability. Under this assumption, units belonging to cluster A_k will show high values of Eq. (2). In this sense, a high conditional probability $P(\mathbf{x} \in A_k \mid \omega_{\mathbf{x}} \in \Omega_K)$ may support the claim that $\mathbf{x}$ is indeed sampled from A_k. Consequently, a rejection region $RR_\alpha^{(k)}$ for $H_0^{(k)} : \omega_{\mathbf{x}} = k$ vs $H_1^{(k)} : \omega_{\mathbf{x}} \neq k$, $k = 1, \ldots, K$ can be constructed based on the quantity $D_k(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + \log |\boldsymbol{\Sigma}_k|$ as

$$RR_\alpha^{(k)} = \{\mathbf{x} \mid D_k(\mathbf{x}) \geqslant \xi_\alpha\}, \tag{3}$$

where ξ_α is a critical value. Since $P(\mathbf{x} \in A_k \mid \omega_\mathbf{x} \in \Omega_K)$ in Eq. (2) is a decreasing function of $D_k(\mathbf{x})$, a large value of $D_k(\mathbf{x})$ becomes an evidence of rejecting the null hypothesis. A critical value ξ_α can be determined based on the size of the test α and the distribution of $D_k(\mathbf{x})$ under the null hypothesis. We employ a well-known property of a quadratic form of the multivariate normal distribution as follows.

Lemma 1. *Let $\mathbf{x} \in \mathbb{R}^p$ be a random variable that follows $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, a quantity $Q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ follows χ^2-distribution with p degrees of freedom.*

To utilize Lemma 1 in practice, the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ true values are needed. In general, these parameter values are unknown, but we start with the simplest scenario, assuming that these parameters are known. Later, we gradually relieve these assumptions to make our proposed method more realistic and practical. When $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k]$ are known for all $k = 1, \ldots, K$ and $\mathbf{x} \mid \omega_\mathbf{x} = k$ follows multivariate normal distribution $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, then ξ_α can be determined as $\xi_\alpha = \chi^2_{p,1-\alpha} + \log|\boldsymbol{\Sigma}_k|$, where $\chi^2_{p,1-\alpha}$ is the $100(1-\alpha)\%$ quantile of Chi-square distribution with P degrees of freedom. Consequently, a size α test $\psi^{(k)}_\alpha$ for testing $H^{(k)}_0 : \omega_\mathbf{x} = k$ versus $H^{(k)}_1 : \omega_\mathbf{x} \neq k$, $k = 1, \ldots, K$ can be written as

$$\psi^{(k)}_\alpha(\mathbf{x}) = I(\mathbf{x} \in RR^{(k)}_\alpha), \quad k = 1, \ldots, K, \tag{4}$$

where $I(A)$ is an indicator function that has a value of 1 if A is a true event and 0 if not. The power of size α test in Eq. (4) increases as the magnitude of $D_k(\mathbf{x})$ under H_1 increases. For example, suppose that the alternative hypothesis is true in that $x \in A_m$, $m \neq k$ and thus $x \sim N(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}^{-1}_m)$. To simplify the example, we assume that $\boldsymbol{\Sigma}^{-1}_m = \boldsymbol{\Sigma}^{-1}_k$. Then $D_k(\mathbf{x}) - \log|\boldsymbol{\Sigma}_k| = (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}_k (\mathbf{x} - \boldsymbol{\mu}_k)$ follows a non-central Chi-squared distribution with non-central parameter $\delta = (\boldsymbol{\mu}_m - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}_k (\boldsymbol{\mu}_m - \boldsymbol{\mu}_k)/2$. Then the power of the test in Eq. (4) increases as $|\boldsymbol{\mu}_m - \boldsymbol{\mu}_k|$ increase because the magnitude of $D_k(\mathbf{x})$ becomes larger.

Now, suppose that parameters of the kth cluster $[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ are unknown. In such case, it would be natural to modify Eq. (3) by replacing $[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ with their consistent estimators $[\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k]$.

Lemma 2. *Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$ be a random sample of size n from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{y} \in \mathbb{R}^p$ be a random variable from the same distribution and independent of $\mathbf{X}$. Also, let $[\hat{\boldsymbol{\mu}}(\mathbf{X}), \hat{\boldsymbol{\Sigma}}(\mathbf{X})]$ be a consistent estimator for $[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$. Then, a quantity $Q_n(\mathbf{X}, \mathbf{y}) = (\mathbf{y} - \hat{\boldsymbol{\mu}}(\mathbf{X}))^T \hat{\boldsymbol{\Sigma}}(\mathbf{X})^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}(\mathbf{X}))$ converges in probability to $Q(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^P$. Further, $Q_n(\mathbf{X}, \mathbf{y})$ converges in distribution to $Q(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^p$.*

Lemma 2 assures that one can use consistent estimators for mean and variances of clusters when choosing a critical value ξ_α. Several choices of consistent estimators $[\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k]$ are available depending on training data scenarios. For example, if the cluster memberships in the training data are available, then the sample mean and sample covariance matrix of each cluster can be used as consistent estimators for $[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$, $k = 1, \ldots, K$. If the cluster memberships of individuals in the training data are unknown, we can employ the EM algorithm to obtain the maximum likelihood estimates (MLE) $[\hat{\boldsymbol{\mu}}^{ML}_k, \hat{\boldsymbol{\Sigma}}^{ML}_k]$ of mean vectors and covariance matrices under the Gaussian mixture model framework (Dempster et al., 1977; Scrucca et al., 2016). Several researchers illustrated that the MLE obtained from the EM algorithm is consistent under certain regularity conditions (Wu, 1983; Redner and Walker, 1984).

2.2 A Test for Unseen-Cluster Problem

The goal of our proposed method is to diagnose an unseen-cluster problem. Let $[A_1, \ldots, A_K]$ be population clusters that are covered by the sample or training data. An unseen-cluster problem occurs when a new observation does not belong to the union of these clusters, $\bigcup_{k=1}^{K} A_k$. We can state a diagnosis of an unseen-cluster problem by testing Eq. (1). Rejecting the null hypothesis in Eq. (1) implies that the new observation $\mathbf{x}$ does not belong to any of A_k, which implies an occurrence of an unseen-cluster problem. In this sense, we suggest a size α test for Eq. (1) as

$$\psi_\alpha(\mathbf{x}) = \prod_{k=1}^{K} \psi_\alpha^{(k)}(\mathbf{x}) = I(\mathbf{x} \in RR_\alpha^{(1)}, \mathbf{x} \in RR_\alpha^{(2)}, \ldots, \mathbf{x} \in RR_\alpha^{(K)}), \tag{5}$$

where $RR_\alpha^{(k)}$ is defined as in Eq. (3). Intuitively, we perform a sequence of tests for each of the K clusters as suggested in Eq. (4), then we reject H_0 if all hypotheses $H_0^{(1)}, \ldots, H_0^{(K)}$ are rejected. If at least one $H_0^{(k)}$ is not rejected, then we do not reject H_0, and an unseen-cluster problem does not become an issue.

Theorem 1. *Suppose that $\psi_\alpha^{(k)}(\mathbf{x})$ is a size α test for $H_0^{(k)} : \omega_\mathbf{x} = k$ versus $H_1^{(k)} : \omega_\mathbf{x} \neq k$ for $k \in \Omega_K$. Then $\psi_\alpha(\mathbf{x}) = \prod_{k=1}^{K} \psi_\alpha^{(k)}(\mathbf{x})$ becomes size α test for $H_0 : \omega_\mathbf{x} \in \Omega_K$ versus $H_1 : \omega_\mathbf{x} \notin \Omega_K$.*

Theorem 1 requires a true number of clusters in the training data. In practice, the number of clusters K is unknown and should be estimated from training data. For example, the Bayesian information criterion (BIC) (Schwarz, 1978), Gap statistics (Tibshirani et al., 2001), and Silhouettes index (Rousseeuw, 1987), can be used to determine K. The theoretical validity of the proposed test is subject to the correctly estimated K, and the distributional assumption on each cluster. We refer to Xu et al. (2016) as a recent review on determining the number of clusters.

Theorem 2. *Suppose that $\psi_\alpha(\mathbf{x})$ is the size α test for $H_0 : \omega_\mathbf{x} \in \Omega_K$ versus $H_1 : \omega_\mathbf{x} \notin \Omega_K$ defined in Eq. (5), and define $\eta_k = P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \omega_\mathbf{x} \notin \Omega_K)$ as the probability of rejecting $H_0 : \omega_\mathbf{x} = k$ when $\omega_\mathbf{x} \notin \Omega_K$. Then, $\eta_{(1)} = \min\{\eta_1, \ldots, \eta_K\}$ becomes an upper bound of the power of the test $\psi_\alpha(\mathbf{x})$.*

Theorem 2 provides that the power of the proposed test is affected by the conditional probability of a correct diagnosis of an unseen-cluster problem, $\eta_k = P(\psi_\alpha^{(k)}(\mathbf{x}) = 1 \mid \omega_\mathbf{x} \notin \Omega_K)$. Here, η_k is the probability of accurately diagnosing an unseen-cluster problem, and $\eta_{(1)} = \min\{\eta_1, \ldots, \eta_K\}$. As discussed in Section 2.2, the probability of making an accurate decision depends on the magnitude of $D_k(\mathbf{x})$, which is a distance measure between a cluster A_k and a testing observation $\mathbf{x}$. Intuitively, the probability of a successful diagnosis of an unseen-cluster problem (i.e., rejecting all K sub-tests) is affected by the probability of the correct decision of concluding that $\mathbf{x}$ is *unclassified*, which involves the highest Type 2 error (that is, not rejecting $H_0 : \omega_\mathbf{x} = k$ even though an unseen-cluster problem appears) among K sub-tests.

2.3 A Two-Stage Classification

Based on the information above, and in particular Theorems 1 and 2, we suggest a two-stage classification method that combines the diagnosis test $\psi_\alpha(\mathbf{x})$ in Eq. (5) for the unseen-cluster

problem together with a conventional classification method. Consider a classification problem where a baseline classifier M_0 is determined based on a training data set, and the testing data that are independent of the training data are sampled from the population. The idea of the two-stage classification M_1 can be written as follows:

1. Implement the test $\psi_\alpha(\mathbf{x}_{new})$ to determine whether the new observation $\mathbf{x}_{new}$ triggers an unseen-cluster problem or not.
2. If $\psi_\alpha(\mathbf{x}_{new}) = 0$, it means an unseen-cluster problem does not occur, then proceed with a baseline classification M_0.
3. If $\psi_\alpha(\mathbf{x}_{new}) = 1$, we don't use M_0 to classify $\mathbf{x}_{new}$ into one of the clusters identified in the training data. Instead, we label $\mathbf{x}_{new}$ as "*unclassified*".

The proposed two-stage classification may improve classification accuracy while enjoying its well-established properties. Depending on data structure, such as distribution and/or dimension, users may choose which baseline classification method (M_0) to use. One drawback of our proposed classification method is that some individuals who should not trigger an unseen-cluster problem can be incorrectly classified as *unclassified* (which can be understood as a type 1 error of the test $\psi_\alpha(\mathbf{x}_{new})$). Consequently, the prediction accuracy of the two-stage classification is affected by the proportion of individuals with unseen-cluster memberships in the testing data set. The next theorem explains the relationship as follows.

Theorem 3. *Let β_0 be the prediction accuracy of a baseline classifier M_0 on the testing data without unseen-cluster problems, and β_1 be the power of the diagnostic test $\psi_\alpha(\mathbf{x})$ and let M_1 be the two-stage classification method defined on M_0 as suggested in Section 2.3. Let α be the size of the test in Eq. (5) and δ be the proportion of unseen-cluster members in the testing data. Further, let $\zeta_0^{(\delta)}$ and $\zeta_1^{(\delta)}$ be the prediction accuracy of M_0 and M_1 on a testing data set with unseen-cluster proportion δ, respectively. Then we have $\zeta_1^{(\delta)} - \zeta_0^{(\delta)} = \delta\beta_1 - \alpha\beta_0$.*

Theorem 3 implies conditions of the prediction accuracy of the two-staged classifier (M_1) being higher than that of the baseline classifier (M_0) when the training and testing sets are given. The increment of prediction error by two-stage tailoring becomes higher as the proportion of observations with unseen-cluster memberships in the testing data (i.e., δ) is large, or the power of the diagnostic test β_1 increases. On the other hand, the two-stage tailored method loses its advantage when the proportion of observations with unseen-cluster memberships in the testing data is small or the power of the test $\psi_\alpha(\mathbf{x})$ is relatively smaller than β_0. When these situations are under concern, one may consider using a small value of α, the size of the diagnostic test, or increasing the power of the test β_1 by increasing the sample size of the training data.

3 Numerical Studies

In this section, we demonstrate the performance of our proposed test as a solution for the unseen-cluster problem via four different numerical studies. The first study investigates the type 1 error of the proposed test to ensure that type 1 error is controlled. In the second study, we investigate the power of the test to evaluate the accuracy of our proposed method to identify the unseen-cluster problem. In the third study, we evaluate the prediction errors of a conventional classifier and the two-stage classifier, which uses the conventional method. Lastly, we compare the performance of our proposed test and other existing tests in terms of type 1 error, power, and prediction accuracy when embedded in a classifier.

We consider scenarios with the least information on the training data in which cluster memberships of observations in the training data are unknown, but the number of clusters in

the training data is known. The number of clusters of the simulated population is known but only used for evaluating type 1 errors power of the tests and prediction errors of classification methods. Finally, since our proposed method is constructed under a mixture of normal distribution, we use *mclust* (Scrucca et al., 2016) as a conventional method and tailor it to the two-stage classification by embedding our proposed test as shown in Section 2.2.

3.1 Simulation I: Type 1 Errors

This simulation is designed to evaluate whether the type 1 error of the proposed test is being controlled appropriately. We generate data under the four-cluster Gaussian mixture model. Next, the simulated data are split into training and testing data, where cluster memberships are completely randomized. We implement a cluster analysis on the training data and obtain the size α test $\psi_\alpha(x)$ as suggested in Eq. (5) using the estimated mean vectors and covariance matrix. Finally, we implement the test on the testing data and calculate the proportion of observations that are classified as *unclassified*. Since both training and testing data cover all four clusters, classifying an observation as *unclassified* is considered a type 1 error. In this sense, we expect the estimated Type 1 error to be close to $\alpha = 0.05$. The course of the simulation study can be summarized as follows.

1. Generate a data set with four latent clusters and split it into training and testing data at random. This scenario is where an unseen-cluster problem does not occur.
2. Fit a four-cluster model on training data and define a size $\alpha = 0.05$ test $\psi_\alpha(x)$.
3. Implement the test on the testing data and obtain the empirical type 1 error as $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\psi_\alpha(\mathbf{x}_i^{new})$, where $\mathbf{x}_i^{new}$ denotes the ith subject in the testing data and n is the number of subjects in the testing data.
4. Repeat step 1. $\sim$ 3. 1,000 times and calculate $\hat{\alpha}_1 \ldots \hat{\alpha}_{1000}$.

Data are simulated under Gaussian mixture models with four latent clusters. We consider various types of underlying distributions by (i) manipulating a location parameter μ, which determines distances between centers of clusters (case I), (ii) manipulating covariance matrices to adjust the dispersion of components (case II), and (iii) manipulating both mean vectors and covariance matrices to adjust both location and dispersion of components (case III). We also evaluate different proportions of the four clusters, but no noticeable differences in the simulation result are discovered. For brevity, we illustrate the results of scenarios where the proportions of the four clusters are equal. Equation (6) provides the true parameter values of the simulated data by indicating both location and scale parameters $[\mu, \sigma]$. Nine scenarios are considered by combining values of $\mu = [4.0, 3.0, 2.0]$ and $\sigma = [1.2, 1.0, 0.5]$, respectively. In each type, large μ and small σ yield strong separation between all four clusters with high concentrations of densities near the mean vectors, while small μ and large σ provide overlapping clusters due to high dispersion. Specific, details of a single simulated data set are as follows:

$$\boldsymbol{X}_i^{(k)} = [X_{i1}, \ldots, X_{i4}]'^{(k)} \sim N_4(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad i = 1, \ldots, 1000,$$

$$\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3, \gamma_4] = [0.25, 0.25, 0.25, 0.25],$$

$$f(\boldsymbol{X}_i) = \sum_{k=1}^{4} \gamma_k \phi_k(\boldsymbol{X}_i^{(k)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

Table 1: Average type 1 errors and power of the size 0.05 test using Eq. (6).

Parameters	$\mu = 4$			$\mu = 3$			$\mu = 2$		
	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.2$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.2$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 1.2$
Type 1 error	0.053	0.046	0.045	0.046	0.044	0.044	0.045	0.046	0.045
Power	0.962	0.670	0.569	0.735	0.339	0.323	0.344	0.177	0.159

$$\boldsymbol{\mu}_1 = \mu \begin{bmatrix} 1.5 \\ 1.0 \\ 0.5 \\ 0 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \mu \begin{bmatrix} 0.0 \\ 1.5 \\ 1.0 \\ 0.5 \end{bmatrix}, \ \boldsymbol{\mu}_3 = \mu \begin{bmatrix} 0.5 \\ 0.0 \\ 1.5 \\ 1.0 \end{bmatrix}, \ \boldsymbol{\mu}_4 = \mu \begin{bmatrix} 1.0 \\ 0.5 \\ 0.0 \\ 1.5 \end{bmatrix}, \tag{6}$$

$$\boldsymbol{\Sigma}_1 = \sigma \begin{bmatrix} 4.50 & 1.50 & 0.75 & 0.38 \\ 1.50 & 4.50 & 1.50 & 0.75 \\ 0.75 & 1.50 & 4.50 & 1.50 \\ 0.38 & 0.75 & 1.5 & 4.50 \end{bmatrix}, \ \boldsymbol{\Sigma}_2 = \sigma \begin{bmatrix} 4.50 & -1.50 & -0.75 & -0.38 \\ -1.50 & 4.50 & -1.50 & -0.75 \\ -0.75 & -1.50 & 4.50 & -1.50 \\ -0.38 & -0.75 & -1.5 & 4.50 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_3 = \sigma \begin{bmatrix} 4.50 & -1.50 & 0.75 & 0.38 \\ -1.50 & 4.50 & -1.50 & 0.75 \\ 0.75 & -1.50 & 4.50 & -1.50 \\ 0.38 & 0.75 & -1.5 & 4.50 \end{bmatrix}, \ \boldsymbol{\Sigma}_4 = \sigma \begin{bmatrix} 4.50 & 1.50 & -0.75 & -0.38 \\ 1.50 & 4.50 & 1.50 & -0.75 \\ -0.75 & 1.50 & 4.50 & 1.50 \\ -0.38 & -0.75 & 1.5 & 4.50 \end{bmatrix}.$$

Table 1 provides summaries of the empirical type 1 error rates from 1,000 repetitions when the a simulated data set follows Eq. (6). We can observe that the empirical type 1 errors are well controlled in all scenarios in that the average type 1 error is close to $\alpha = 0.05$. Table 1 supports Theorem 1 which claims that Eq. (5) is a size α test.

3.2 Simulation II: Power of the Test

The second numerical study is designed to evaluate the power behavior of the proposed test. Similar to the numerical studies in Section 3.1, we simulate a data set under a four-cluster Gaussian mixture model. Next, we divide the simulated data into training and testing data, where the training data consists of only cluster I $\sim$ III, and the testing data only contains cluster IV. Namely, this scenario represents a perfect representation of the unseen-cluster problem. Again, we build a size α test $\psi_\alpha(x)$ in the same manner as in Section 3.1 and implement the test in Eq. (5) on the testing data. The proportion of observations labeled as *unclassified* is the empirical power of the test. In this sense, we expect that the estimated power is close to 1. The course of the simulation study can be summarized as follows.

1. Generate a data set with four latent clusters and split it into training and testing data as follows: training data only consists of clusters I$\sim$III while the testing data only consists of cluster IV.
2. Fit a 3-cluster model using the training data and define a size $\alpha = 0.05$ test $\psi_\alpha(x)$.
3. Implement the test on the testing data and obtain empirical power as $\hat{\eta}_i = \frac{1}{n}\sum_{i=1}^{n} \psi_\alpha(\mathbf{x}_i^{new})$, where $\mathbf{x}_i^{new}$ denotes the ith subject in the testing data and n is the number of subjects in the testing data.
4. Repeat step 1. $\sim$ 3. 1,000 times and calculate $\hat{\eta}_1 \ldots \hat{\eta}_{1000}$.

We repeat the simulation procedure using the true values in Eq. (6). The summaries of the estimated power from different true values are shown in Table 1. As discussed in Theorem 2, the estimated power of the test increases for large μ or small σ, in which the magnitude of overlaps between clusters are small. On the other hand, the estimated power decrease in small values of μ or large values of σ, due to the large amount of overlap between clusters.

3.3 Simulation III: Comparison of Prediction Accuracies

In this Section, we design a set of numerical studies to compare the proposed two-stage classification method to the prediction accuracies of conventional classification. Since our synthetic data are generated under the Gaussian mixture model, we compare the performance of the Gaussian model-based classification method discussed in Scrucca et al. (2016) and its two-staged upgraded method. To reproduce a classification scenario with an unseen-cluster problem, we simulate data using the four-cluster Gaussian mixture model and divide the data into training and testing data so that a training data set consists of clusters I∼III while a testing data set contains all four clusters. For brevity, we only report results from $\sigma = 1.2$ and $\mu = [4.0, 3.0, 2.0]$ in Eq. (6). Details of the simulation study are as follows.

1. Generate a data set with four clusters and split the data into training (cluster I $\sim$ III) and testing data (cluster I $\sim$ IV).
2. Fit a three-cluster model using training data and define a size α test $\psi_\alpha(x)$.
3. Using the classification method based on training data, make predictions on the testing data and calculate prediction accuracies.
4. Repeat step 1. $\sim$ 3. 1,000 times and obtain the empirical prediction accuracies.

Prediction accuracies are calculated separately for predicting (a) clusters I∼III and (b) cluster IV, and then results are combined. Table 2 illustrates prediction accuracies of the proposed two-stage and conventional classification methods separately in clusters I∼III (i.e., clusters covered by training data) and cluster IV (i.e., uncovered cluster). For each sub-table, the first two rows illustrate the prediction accuracies of the proposed and conventional classification accuracies for the observations that belong to clusters I∼III in the testing data. These two methods employ the same algorithm except that the two-stage classification filters out observations identified as *unclassified*. Consequently, the conventional method shows higher prediction accuracies of classifying the observations in clusters I, II, and III covered by training data because the two-stage classification method may yield some false diagnosis of an unseen-cluster problem.

However, the conventional method completely fails to classify the observations that belong to cluster IV because it does not have cluster IV as a possible outcome. Still, our two-stage classification may filter out observations from cluster IV and correctly classify them as *unclassified* before the conventional classifier is applied, and thus eventually reduces the misclassification rates. The prediction accuracies of the two methods for the individuals in cluster IV are shown in the third and fourth rows. Finally, the prediction accuracy of the two-stage classification (i.e., *Accuracy (Two stage)*) and the amount of increase in its prediction accuracy (i.e., *Two stage - Conventional*) are shown in the fifth and sixth rows. *Accuracy (Two stage)* is a weighted sum of accuracies from training clusters (i.e., *Cluster I* $\sim$ *III*) and the testing cluster (*Cluster IV*), where weight is the proportion of *Cluster IV* in the testing data.

Table 2 illustrates the simulation results with different unseen-cluster proportions in the testing data. The prediction accuracy improvement by our two-stage classification is noticeable across all values of μ when the unseen-cluster proportion is 10%. This is because the unseen-cluster proportions in the testing data are larger than the size of the diagnostic test embedded in

Table 2: Prediction accuracies with diverse proportion of unseen clusters.

Parameters	$\mu = 4$		$\mu = 3$		$\mu = 2$	
Proportion of unseen clusters	10%	1%	10%	1%	10%	1%
Cluster I∼III (Proposed)	0.853	0.854	0.758	0.761	0.643	0.647
Cluster I∼III (Conventional)	0.873	0.873	0.769	0.772	0.644	0.648
Cluster IV(Proposed)	0.564	0.581	0.322	0.353	0.161	0.165
Cluster IV(Conventional)	0.000	0.000	0.000	0.000	0.000	0.000
Accuracy (Two stage)	0.824	0.851	0.714	0.757	0.595	0.642
Two stage - Conventional	0.039	−0.013	0.022	−0.008	0.015	0.000

the two-stage classification. Namely, using the diagnostic test for the unseen-cluster problem in the two-stage classification effectively improves the overall prediction accuracy when the unseen-cluster proportion is large. However, when the unseen-cluster proportion is smaller than the size of the diagnostic test (for example, 4% or 1%), the two-stage classification yields lower prediction accuracy than the conventional method. This is because the number of observations incorrectly diagnosed as *unclassified* becomes larger than those correctly identified as *unclassified*. In other words, the number of benefiting individuals becomes smaller than those penalized.

As discussed in Theorem 3, an increment in prediction accuracy is a linear combination of unseen-cluster proportion in the testing data, the size of the diagnostic test, and two types of errors. The two error types are (i) misclassification error, which fails to estimate a true cluster membership, and (ii) false diagnosis of unseen-cluster problem, which is equivalent to the type 1 error of a diagnostic test for the unseen-cluster problem. Simulation results in Table 2 support Theorem 3 in that the two-stage classification improves a conventional method's prediction accuracy when the testing data contains a large number of individuals that are exposed to the unseen-cluster problem (i.e., individuals who do not belong to clusters covered by training data), because our test may prevent these observations from being misclassified and thus increase the classification accuracy. On the other hand, implementing the two-stage classification may harm the classification accuracy if there is a relatively small proportion of subjects with unseen-cluster memberships.

3.4 Simulation IV: Comparison Study with Other ODD Methods

In this section, we compare the performance of our proposed diagnostic test with other existing ODD methods. To make a fair comparison, we only consider Type 3 based ODD methods (Hodge and Austin, 2004; Pimentel et al., 2014) and compare type 1 errors, power, and prediction accuracies using empirical simulation. Competing methods are (1) Mahalanobis distance-based method (Clifton et al., 2011; Lee et al., 2018; Liang et al., 2017), (2) Manhattan distance-based method (Yong et al., 2012), (3) Euclidian distance-based method (Feinman et al., 2017; Ma et al., 2018), (4) Squeezer algorithm based method (He et al., 2003), (5) k-nearest neighborhood-based method (Ma et al., 2018; Papernot and McDaniel, 2018), and (6) Bootstrap-based method (Grosse et al., 2017). Simulation data are generated under the same scenarios as shown in Eq. (6), and each scenario is repeated $1,000$ times to evaluate the performances of different methods.

Table 3 illustrates the estimated type 1 error of size $\alpha = 0.05$ ODD methods and our proposed test. We conclude that all methods except our proposed test fail to control type 1 errors

Table 3: Average type 1 errors of different diagnostic tests under Eq. (6).

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$\sigma = 0.5$	0.053	0.064	0.053	0.053	0.062	0.060	0.063
	$\sigma = 1.0$	0.046	0.064	0.053	0.053	0.062	0.059	0.063
	$\sigma = 1.2$	0.045	0.064	0.053	0.053	0.062	0.059	0.064
$\mu = 3.0$	$\sigma = 0.5$	0.046	0.064	0.053	0.053	0.062	0.059	0.063
	$\sigma = 1.0$	0.044	0.066	0.052	0.053	0.062	0.059	0.066
	$\sigma = 1.2$	0.045	0.068	0.052	0.052	0.062	0.059	0.067
$\mu = 2.0$	$\sigma = 0.5$	0.044	0.068	0.052	0.053	0.061	0.059	0.067
	$\sigma = 1.0$	0.046	0.070	0.052	0.053	0.062	0.058	0.070
	$\sigma = 1.2$	0.045	0.071	0.052	0.052	0.061	0.058	0.070

Table 4: Average power of different diagnostic tests under Eq. (6).

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$\sigma = 0.5$	0.962	0.967	0.869	0.940	0.966	0.955	0.966
	$\sigma = 1.0$	0.670	0.707	0.448	0.581	0.708	0.638	0.705
	$\sigma = 1.2$	0.569	0.614	0.361	0.481	0.617	0.542	0.612
$\mu = 3.0$	$\sigma = 0.5$	0.735	0.766	0.515	0.651	0.767	0.705	0.764
	$\sigma = 1.0$	0.339	0.394	0.261	0.326	0.388	0.338	0.392
	$\sigma = 1.2$	0.323	0.371	0.203	0.263	0.373	0.313	0.370
$\mu = 2.0$	$\sigma = 0.5$	0.344	0.393	0.214	0.281	0.396	0.332	0.391
	$\sigma = 1.0$	0.177	0.218	0.132	0.154	0.214	0.189	0.217
	$\sigma = 1.2$	0.159	0.196	0.120	0.132	0.185	0.167	0.194

in that their average type 1 error exceeds 0.05. In this sense, the proposed method is preferred to other methods. Next, Table 4 shows the estimated power under the alternative hypothesis as discussed in Section 3.2. Our proposed test shows higher power than the Manhattan, Euclidean distance method, and 5-NN for all scenarios. On the other hand, the Mahalanobis distance method, Squeezer algorithm, and Bootstrap algorithm show higher power than our proposed method. Such high power, however, is achieved by overusing type 1 error of the test and thus should be considered with caution.

Tables 5 and 6 show prediction accuracies of two-stage classifications using different diagnostic tests under 10% and 1% unseen-cluster proportions, respectively. The proposed method shows the highest prediction accuracy in all scenarios, followed by the methods using the Mahalanobis distance and the Squeezer algorithm. Even though the power of our proposed test is not the highest, it still achieves the highest prediction accuracy by having the lowest false-discovery rates. Other methods yield lower prediction accuracies due to high rates of false-discovery probabilities, even though their power is higher than our proposed test. This trend is noticeable when

Table 5: Average prediction accuracies of two-stage classifications with 10% unseen-clusters.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$\sigma = 0.5$	0.948	0.944	0.937	0.944	0.944	0.942	0.944
	$\sigma = 1.0$	0.858	0.856	0.830	0.844	0.856	0.847	0.856
	$\sigma = 1.2$	0.824	0.823	0.798	0.811	0.823	0.814	0.823
$\mu = 3.0$	$\sigma = 0.5$	0.878	0.876	0.852	0.866	0.876	0.868	0.876
	$\sigma = 1.0$	0.749	0.749	0.728	0.737	0.749	0.741	0.749
	$\sigma = 1.2$	0.714	0.715	0.697	0.703	0.714	0.707	0.715
$\mu = 2.0$	$\sigma = 0.5$	0.727	0.728	0.709	0.716	0.727	0.720	0.726
	$\sigma = 1.0$	0.615	0.617	0.606	0.609	0.616	0.612	0.612
	$\sigma = 1.2$	0.593	0.596	0.586	0.587	0.594	0.590	0.592

Table 6: Average prediction accuracies of two-stage classifications with 1% unseen-clusters.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$\sigma = 0.5$	0.948	0.943	0.944	0.945	0.943	0.942	0.943
	$\sigma = 1.0$	0.880	0.873	0.872	0.874	0.874	0.871	0.874
	$\sigma = 1.2$	0.853	0.846	0.845	0.846	0.847	0.844	0.847
$\mu = 3.0$	$\sigma = 0.5$	0.895	0.889	0.888	0.890	0.890	0.887	0.889
	$\sigma = 1.0$	0.789	0.784	0.781	0.783	0.784	0.781	0.784
	$\sigma = 1.2$	0.759	0.755	0.752	0.754	0.755	0.752	0.755
$\mu = 2.0$	$\sigma = 0.5$	0.769	0.765	0.762	0.763	0.764	0.761	0.765
	$\sigma = 1.0$	0.664	0.662	0.659	0.660	0.662	0.659	0.662
	$\sigma = 1.2$	0.642	0.641	0.638	0.639	0.641	0.638	0.641

the unseen-cluster proportion in the testing data is small because the number of false diagnoses increases.

In addition, we evaluate the performance of ODD methods when the data does not follow the Gaussian mixture distribution. To achieve this, we generate data sets from multivariate t distributions with different degrees of freedom while the mean vectors and covariance matrices are the same as in Eq. (6). Similar to the previous scenarios, we compare ODD methods' type 1 error, power, and prediction accuracies. When generating data, we consider several degrees of freedom as follows; $df = [5, 15, 30]$. As the degree of freedom becomes small, data distribution has a thicker tail than the normal distribution, while the overall shape is still symmetric. For brevity, we illustrate the results from $\mu = [4.0, 3.0, 2.0]$ with $\sigma = 1.2$.

Table 7 illustrates the empirical type 1 error from $1,000$ repetitions under multivariate t distribution with degrees of freedom 5, 15, and 30, respectively. When $df = 5$, our proposed method shows a significant decrease in the type 1 error, while the other nonparametric approaches, except for the Squeezer algorithm, retain their type 1 error close or higher. When

Table 7: Type 1 errors of the size 0.05 test under of t-distribution.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$df = 5$	0.017	0.057	0.052	0.052	0.085	0.055	0.056
	$df = 15$	0.058	0.064	0.052	0.052	0.058	0.056	0.063
	$df = 30$	0.053	0.064	0.053	0.053	0.058	0.057	0.064
$\mu = 3.0$	$df = 5$	0.013	0.054	0.050	0.050	0.254	0.053	0.054
	$df = 15$	0.054	0.065	0.053	0.053	0.060	0.056	0.065
	$df = 30$	0.050	0.066	0.051	0.052	0.061	0.057	0.065
$\mu = 2.0$	$df = 5$	0.012	0.054	0.050	0.050	0.474	0.053	0.053
	$df = 15$	0.046	0.065	0.052	0.052	0.064	0.055	0.065
	$df = 30$	0.050	0.070	0.052	0.052	0.061	0.056	0.069

Table 8: Average prediction accuracies of two-stage classifications under t-distribution.

Parameters		Proposed	Mahalanobis	Manhattan	Euclidean	Squeezer	5-NN	Bootstrap
$\mu = 4.0$	$df = 5$	0.769	0.766	0.753	0.757	0.767	0.760	0.767
	$df = 15$	0.808	0.808	0.785	0.795	0.809	0.798	0.809
	$df = 30$	0.821	0.820	0.795	0.807	0.821	0.811	0.821
$\mu = 3.0$	$df = 5$	0.696	0.694	0.688	0.689	0.695	0.691	0.695
	$df = 15$	0.720	0.720	0.706	0.710	0.721	0.713	0.721
	$df = 30$	0.729	0.729	0.712	0.718	0.730	0.721	0.730
$\mu = 2.0$	$df = 5$	0.620	0.618	0.615	0.615	0.619	0.617	0.618
	$df = 15$	0.630	0.630	0.623	0.625	0.631	0.627	0.632
	$df = 30$	0.633	0.634	0.626	0.627	0.635	0.630	0.636

$df = [15, 30]$, all methods show moderately inflated type 1 errors similar to 3. This is because the shapes of the simulated data are close to the normal curve, though their tails are thicker than those of the normal curve. From Table 7, we discover some failures of type 1 errors in our proposed method, but such failures do not occur only in our proposed method; other competing nonparametric methods also suffer from such failure when the data does not follow a mixture of normal distribution.

Finally, Table 8 illustrates the prediction accuracies of two-stage classifications using different ODD tests under 10% unseen-cluster proportion. As shown in the table, prediction accuracies of the proposed model decrease as the data distribution deviates from the normal distribution. Our proposed method does not show the highest prediction accuracy across all scenarios. Still, other competing nonparametric methods also experience decreases in their prediction accuracies, and none show noticeably higher prediction accuracies than the proposed method. This implies that the nonparametric approach to the ODD test is not remarkably beneficial for dealing with data that does not follow a mixture of normal distributions. To increase the overall prediction

accuracy in practice, one should develop a specialized test tailored for data of interest and use a distribution-free classification method as a baseline classifier. Once a distribution-free method is suggested, we can improve it further by embedding it into our two-stage classification method. Such development is not the main focus of this paper, but we believe that it requires a thorough investigation of background knowledge on data-related fields, such as Sun et al. (2018).

4 Applications

In this section, we use the *Dry bean data* to mimic the unseen-cluster problem and demonstrate the performance of the two-stage classification (Koklu and Ozkan, 2020). The data set contains $13,611$ complete observations recorded from the grain images of seven different types of dry beans. The data set is available in the UC Irvine Machine Learning Repository (https://archive-beta.ics.uci.edu/). We divide a data set into training and testing data, build pairs of classifiers (i.e., conventional classifiers and their tailored versions) using the training data set, and then implement classification on the testing data. Prediction accuracies on the testing data are compared in pairs between a conventional classifier and its two-staged method to illustrate the contribution of two-stage tailoring. We repeat these processes $1,000$ times and compare prediction accuracies.

Table 9 illustrates the frequency table of seven bean clusters. Similar to the simulation studies discussed in Section 3, we use the smallest cluster (i.e., *Cluster 2* with 522 cases among $13,611$) as the target class, which is not covered by a training data set. We mimic the unseen-cluster problem by excluding observations belonging to *Cluster 2* from the training data. Next, we randomly divide the data set into seven clusters with equal sizes and choose one as the testing data set. Consequently, the training data contains $n = 10,000$ observations while testing data has at most $3,611$ observations, depending on the proportion of *Cluster 2*. Next, we use the smallest cluster (522 cases among $13,611$) as the target class, which is not covered by a training data set and causes the unseen-cluster problem. Namely, we split the data into training and testing data sets so that the training data set does not contain observations belonging to the target class. Finally, we consider four different scenarios in the proportion of the unseen-cluster in the testing data set.

1. *Scenario I*: all clusters are included in the training data (no unseen-cluster problem).
2. *Scenario II*: unseen-cluster proportion in the testing data set is 2.07%.
3. *Scenario III*: unseen-cluster proportion in the testing data set is 8.31%.
4. *Scenario IV*: unseen-cluster proportion in the testing data set is 14.4%.

We mimic a scenario in which the number of clusters in the training data is known, but their cluster memberships are not available. Consequently, we estimate a classifier based on the training data set and implement classification on the testing data set to evaluate prediction accuracies. When implementing classification, we use (i) the conventional classification method and (ii) the two-stage classification using the same classifier as in (i). Each scenario is repeated 100 times, and the prediction accuracies of the two methods are calculated.

Table 9: Frequency table of seven clusters of Dry beans.

Cluster	1	2	3	4	5	6	7
Frequency	1322	522	1630	3546	1928	2027	2636

Table 10: Prediction accuracies of *mclust* and the two-staged classification for Dry bean data.

Scenario I	Min	Q_1	Q_2	Mean	Q_3	Max
mclust	0.811	0.836	0.849	0.848	0.858	0.887
Two-staged *mclust*	0.814	0.837	0.848	0.846	0.855	0.878
Scenario II	Min	Q_1	Q_2	Mean	Q_3	Max
mclust	0.736	0.781	0.800	0.798	0.818	0.853
Two-staged *mclust*	0.740	0.785	0.802	0.801	0.819	0.855
Scenario III	Min	Q_1	Q_2	Mean	Q_3	Max
mclust	0.731	0.758	0.771	0.767	0.76	0.820
Two-staged *mclust*	0.737	0.772	0.800	0.778	0.784	0.823
Scenario IV	Min	Q_1	Q_2	Mean	Q_3	Max
mclust	0.583	0.605	0.612	0.610	0.616	0.633
Two-staged *mclust*	0.820	0.873	0.880	0.879	0.884	0.901

Table 10 illustrates the classification accuracies of *mclust* and its two-stage tailored method on the testing data from four different scenarios. The conventional classification method (i.e., *mclust*) illustrates higher prediction accuracies than the proposed two-stage classification method in Scenarios *I* and *II*, where the unseen-cluster problem does not occur, or its magnitude is very small. As discussed in Section 3, the conventional method is strictly better than our two-stage classification if no subject in the testing data belongs to unseen-cluster. Similar results occur in *Scenario II* in which the unseen-cluster proportion in the testing data is small (i.e., smaller than $\alpha = 0.05$). In such cases, the number of false diagnoses of unseen-cluster problem may exceed the number of *unclassified*. This implies that the overall prediction accuracy may decrease due to a relatively larger value of Type 1 error than that of unseen-cluster problem. On the other hand, our two-stage classification indicates noticeably higher prediction accuracies in *Scenarios III* and *IV*, because of the large number of individuals who need to be assigned to clusters not available in the training data. Since the proportion of individuals in clusters not available in the training data is higher in *Scenarios III* and *IV* compared with *Scenarios I* and *II*, the performances of the two-stage classification are much better than that of the conventional classifier (*mclust*).

5 Conclusions

In this paper, we introduce an unseen-cluster problem where training data fails to capture all underlying clusters of the population. As a solution for the misclassification due to the unseen-cluster problem, we suggest implementing a test before classifying a testing observation and determining whether it belongs to one of the covered clusters (by training data set) or not. Assuming that the population distribution is a finite mixture of normal distributions, we establish a diagnostic test of the unseen-cluster problem and propose a two-stage classification method. Using mean vectors and covariance matrices estimated from the training data, our two-stage classification method performs a diagnostic test to determine whether a new subject belongs to one of the estimated clusters or not. If the test result does not indicate an unseen cluster problem, the subject is assigned to one of the clusters based on its conditional probability.

If the test indicates an unseen-cluster problem, the subject is labeled as *unclassified* and is not assigned to one of the clusters. In such a way, our proposed method can also be employed when a classification is implemented on a set of estimated clusters in which the complete list of clusters is unavailable.

The proposed diagnostic test for the unseen-cluster problem can be considered a novelty detection problem in the classification problem. It also resembles statistical quality control (SQC), where statistical methods are employed to investigate whether a collected data set satisfies certain quality standards (Shewhart and Deming, 1986). In the unseen-cluster framework, we implement the diagnostic test by exploring all existing clusters and see if the new observation belongs to one or more clusters. Such a process can be considered as investigating the quality of a new instance, except for the fact that in the unseen-cluster problem, the numbers of clusters are unknown and needs to be estimated. In the unseen-cluster framework, a testing observation from an unseen cluster is treated as *unclassified* because it does not belong to any currently available clusters in the training data. From the SQC's perspective, such an observation can be considered a failure of quality control in that the new instance does not satisfy the current standards.

The proposed two-stage classification has advantages because it implements the identification of an unseen-cluster problem without discarding the properties of well-established conventional classification methods. In Section 2, we illustrate the mathematical principles of how the proposed diagnostic test becomes a valid test for a given significance level. We also show that the increment of prediction accuracy of our two-stage classification method can be written as a function of the unseen-cluster proportion in the testing data, the power of a diagnostic test, and the power of a targeted conventional method. In this sense, prospective users may employ the proposed two-stage classification method when concerned with potential unseen-cluster problems due to diverse reasons such as sampling bias.

In Sections 2.3 and 3.3, we illustrate that a two-stage classification method shows higher prediction accuracy than its original classification method when unseen-cluster proportion in the testing data is higher than the size of the diagnosis test. Such a conclusion is based on the measure of prediction accuracy, where we regard the probability of false diagnosis and misclassification rate as equally important. When these two errors are distinguished in their importance, the measure of prediction accuracy becomes the weighted function of false diagnosis and misclassification rates. Consequently, users may choose a significance level of the diagnostic test for the unseen-cluster problem. For example, reducing the significance level of the test is needed if the false diagnosis of the unseen cluster problem is crucial. On the contrary, increasing the significance level will be preferred if the misclassification due to the unseen-cluster problem is more problematic.

The proposed two-stage classification is established by embedding a diagnostic test for the unseen-cluster into a conventional classifier. This paper employs a Gaussian-mixture model-based clustering/classification method in numerical studies and the public data example. As discussed in Section 2, our proposed diagnostic test assumes that each population cluster follows a normal distribution and the population follows a finite mixture of normal distributions. Consequently, the proposed diagnosis test is valid only if the assumption is acceptable, and users must check whether it is appropriate for their data. In addition, prospective users need to employ an appropriate classification method depending on their data distribution and the assumptions they are willing to make before they combine it with the diagnostic test so that the tailored classifier fits their data well.

Dealing with the unseen-cluster has substantial extensions that must be solved. The pro-

posed diagnostic test for the unseen-cluster has been designed to cover cross-sectional data with relatively small dimensions (that is, the sample size is sufficiently larger than the number of variables). In this sense, future research will examine the extension of the diagnosis of the unseen-cluster problem to complicated data structures such as high-dimensional or functional data. Similarly, the proposed method is established based on the mixture of multivariate normal distributions. Since such assumptions are not always acceptable, developing a robust test for the unseen-cluster problem is highly needed. Furthermore, an extension of the unseen-cluster problem to longitudinal data with missing values is also needed. Longitudinal studies are susceptible to unseen-cluster problems because the number of clusters contained in the sample may vary as samples are collected across time, especially in that some underlying clusters of the population may disappear or advent. In this sense, our proposed test for the unseen-cluster problem can be extended to incomplete data because missing values commonly occur in longitudinal studies for various reasons, such as drop-out.

Supplementary Material

- Supplementary document: The supplementary document provides the proofs of the Theorems 1, 2, and 3, and additional numerical study results.
- Software: R codes for the proposed methods and algorithms.

Funding

This work was partially supported by the National Science Foundation under grant DMS-2015320.

References

Bartlett PL, Wegkamp MH (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8): 1823–1840.

Bendale A, Boult T (2015). Towards open world recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1893–1902.

Bethlehem J (2010). Selection bias in web surveys. *International Statistical Review*, 78(2): 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

Bouveyron C (2014). Adaptive mixture discriminant analysis for supervised learning with unobserved classes. *Journal of Classification*, 31: 49–84. https://doi.org/10.1007/s00357-014-9147-x

Cappozzo A, Greselin F, Murphy TB (2020). Anomaly and novelty detection for robust semi-supervised learning. *Statistics and Computing*, 30(5): 1545–1571. https://doi.org/10.1007/s11222-020-09959-1

Clifton DA, Hugueny S, Tarassenko L (2011). Novelty detection with multivariate extreme value statistics. *Journal of Signal Processing Systems*, 65(3): 371–389. https://doi.org/10.1007/s11265-010-0513-6

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 39(1): 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Denti F, Cappozzo A, Greselin F (2021). A two-stage Bayesian semiparametric model for novelty detection with robust prior information. *Statistics and Computing*, 31(4): 42. https://doi.org/10.1007/s11222-021-10017-7

Doan T, Kalita J (2017). Overcoming the challenge for text classification in the open world. In: *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, 1–7. IEEE.

Feinman R, Curtin RR, Shintre S, Gardner AB (2017). Detecting adversarial samples from artifacts. arXiv preprint: https://arxiv.org/abs/1703.00410.

Geng C, Huang Sj, Chen S (2020). Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3614–3631. https://doi.org/10.1109/TPAMI.2020.2981604

Grosse K, Manoharan P, Papernot N, Backes M, McDaniel P (2017). On the (statistical) detection of adversarial examples. arXiv preprint: https://arxiv.org/abs/1702.06280.

He Z, Xu X, Deng S (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9–10): 1641–1650. https://doi.org/10.1016/S0167-8655(03)00003-5

Hodge V, Austin J (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2): 85–126. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9

Klawonn F, Höppner F, Jayaram B (2012). What are clusters in high dimensions and are they difficult to find? In: *Clustering High-Dimensional Data*, 14–33. Springer.

Koklu M, Ozkan IA (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174: 105507. https://doi.org/10.1016/j.compag.2020.105507

Lee K, Lee K, Lee H, Shin J (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: *Advances in Neural Information Processing Systems*, volume 31 (S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, eds.).

Liang S, Li Y, Srikant R (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint: https://arxiv.org/abs/1706.02690.

Lo AY (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1): 351–357.

Lonij V, Rawat A, Nicolae MI (2017). Open-world visual recognition using knowledge graphs. arXiv preprint: https://arxiv.org/abs/1708.08310.

Ma X, Li B, Wang Y, Erfani SM, Wijewickrema S, Schoenebeck G, et al. (2018). Characterizing adversarial subspaces using local intrinsic dimensionality. arXiv preprint: https://arxiv.org/abs/1801.02613.

Miller DJ, Browning J (2003). A mixture model framework for class discovery and outlier detection in mixed labeled/unlabeled data sets. In: *2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No. 03TH8718)*, 489–498. IEEE.

Papernot N, McDaniel P (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv preprint: https://arxiv.org/abs/1803.04765.

Pimentel MA, Clifton DA, Clifton L, Tarassenko L (2014). A review of novelty detection. *Signal Processing*, 99: 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026

Redner RA, Walker HF (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2): 195–239. https://doi.org/10.1137/1026034

Rousseeuw PJ (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

https://doi.org/10.1016/0377-0427(87)90125-7

Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J (1999). Support vector method for novelty detection. In: *Advances in Neural Information Processing Systems*, volume 12 (S Solla, T Leen, K Müller, eds.).

Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464.

Scrucca L, Fop M, Murphy TB, Raftery AE (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1): 289. https://doi.org/10.32614/RJ-2016-021

Shewhart WA, Deming WE (1986). *Statistical Method from the Viewpoint of Quality Control*. Courier Corporation.

Sun Z, Wang T, Deng K, Wang XF, Lafyatis R, Ding Y, et al. (2018). Dimm-sc: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, 34(1): 139–146. https://doi.org/10.1093/bioinformatics/btx490

Tibshirani R, Walther G, 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(2): 411–423. https://doi.org/10.1111/1467-9868.00293

Wankhade KK, Jondhale KC, Thool VR (2018). A hybrid approach for classification of rare class data. *Knowledge and Information Systems*, 56(1): 197–221. https://doi.org/10.1007/s10115-017-1114-5

Wu CJ (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1): 95–103.

Xu S, Qiao X, Zhu L, Zhang Y, Xue C, Li L (2016). Reviews on determining the number of clusters. *Applied Mathematics & Information Sciences*, 10(4): 1493–1512. https://doi.org/10.18576/amis/100428

Yong SP, Deng JD, Purvis MK (2012). Novelty detection in wildlife scenes through semantic context modelling. *Pattern Recognition*, 45(9): 3439–3450. https://doi.org/10.1016/j.patcog.2012.02.036