Our mixture model of consensus admits generalization for clustering ensembles with incomplete partitions. Such partitions can appear as a result of clustering of subsamples or resampling of a dataset. For example, a partition of a bootstrap sample only provides labels for the selected points. Therefore, the ensemble of such partitions is represented by a set of vectors of cluster labels with potentially missing components. Moreover, different vectors of cluster labels are likely to miss different components. Incomplete information can also arise when some clustering algorithms do not assign outliers to any of the clusters. Different clusterings in the diverse ensemble can consider the same point $\mathbf{x}_i$ as an outlier or otherwise, that results in missing components in the vector $\mathbf{y}_i$. Yet another scenario leading to missing information can occur in clustering combination of distributed data or ensemble of clusterings of non-identical replicas of a dataset.

It is possible to apply the EM algorithm in the case of missing data [14], namely missing cluster labels for some of the data points. In these situations, each vector $\mathbf{y}_i$ in $\boldsymbol{Y}$ can be split into observed and missing components $\mathbf{y}_i$ = ($\mathbf{y}_i^{\text{obs}}$, $\mathbf{y}_i^{\text{mis}}$). Incorporation of a missing data leads to a slight modification of the computation of E and M steps. First, the expected values $E[z_{im} |\mathbf{y}_i^{\text{obs}}, \Theta']$ are now inferred from the observed components of vector $\mathbf{y}_i$, i.e. the products in Eq. (14) are taken over known labels:

$$\prod_{j=1}^{H} \to \prod_{j:y^{obs}} .$$

Additionally, it is required to compute the expected values $E[z_{im}\ \mathbf{y}_i^{\text{mis}}|\ \mathbf{y}_i^{\text{obs}}, \Theta']$ and substitute them, as well as $E[z_{im}\ |\ \mathbf{y}_i^{\text{obs}}, \Theta']$, in the M-step for re-estimation of parameters $\vartheta_{jm}(k)$. More details on handling missing data can be found in [14, 24].

Though data with missing cluster labels can be obtained in different ways, we analyze only the case when components of $\mathbf{y}_i$ are missing completely at random [29]. It means that the probability of a component to be missing does not depend on other observed or unobserved variables. Note, that the outcome of clustering of data subsamples (e.g., bootstrap) is different from clustering the entire data set and then deleting a random subset of labels. However, our goal is to present a consensus function for general settings. We expect that experimental results for ensembles with missing labels are applicable, at least qualitatively, even for a combination of bootstrap clusterings.

The proposed ensemble clustering based on mixture model consensus algorithm is summarized below:

```
begin
  for $i$=1 to $H$   // $H$ - number of clusterings
    cluster a dataset $\pi \leftarrow k$-means($\boldsymbol{X}$)
    add partition to the ensemble $\Pi = \{\Pi,\pi\}$
  end
  initialize model parameters $\Theta = \{\alpha_1,\ldots, \alpha_M, \boldsymbol{\theta}_1,\ldots, \boldsymbol{\theta}_M\}$
  do until convergence criterion is satisfied
    compute expected values $E[z_{im}]$, $i$=1..$N$, $m$=1..$M$
    compute $E[z_{im}\ \mathbf{y}_i^{\text{mis}}]$ for missing data (if any)
    re-estimate parameters $\vartheta_{jm}(k)$, $j$=1..$H$, $m$=1..$M$, $\forall k$
  end
  $\pi_C(\mathbf{x}_i)$ = index of component of $\mathbf{z}_i$ with largest expected
    value, $i$=1..$N$
  return $\pi_C$  // consensus partition
end
```

Note that any clustering algorithm can be used to generate ensemble instead of the *k*-means algorithm shown in the above pseudocode.

The value of *M*, number of components in the mixture, deserves a separate discussion that is beyond the scope of this paper. Here, we assume that the target number of clusters is predetermined. It should be noted, however, that mixture models in unsupervised classification greatly facilitate estimation of the true number of clusters [10]. Maximum likelihood formulation of the problem specifically allows us to estimate *M* by using additional objective functions during the inference, such as the minimum description length of the model. In addition, the proposed consensus algorithm can be viewed as a version of Latent Class Analysis (e.g. see [3]), which has rigorous statistical means for quantifying plausibility of a candidate mixture model.

5 Empirical Study

The experiments were conducted with artificial and real-world datasets, where true natural clusters are known, to validate both accuracy and robustness of consensus via mixture model. We explored the datasets using five different consensus functions.

5.1 Datasets. Table 2 summarizes the details of the datasets. Five datasets have been used in the experiments. Two large real-world benchmarks: (i) The dataset of galaxies and stars, characterized by 14 features extracted from their images, with known classification provided by domain experts [26], (ii) Biochemical dataset of water molecules found in protein structures and categorized as